Technology Reports

CUPS | EPC/5GC | Traffic Characteristics

CUPS for Flexible U-Plane Processing Based on Traffic Characteristics

Core Network Development Department Yuya Miyazaki Kenzo Okuda
Kouichiro Kunitomo
DOCOMO Technology, Inc. Packet Network Division Takahiro Kaida

As 5G services become more widespread, it is imperative that EPC further address various traffic characteristics such as low latency communications and high-capacity user communications. NTT DOCOMO has introduced CUPS architecture that enables functional separation of C-Plane and U-Plane parts from SGW and PGW. This separation realizes flexible U-Plane handling based on traffic characteristics. This article provides an overview of CUPS technology and describes how facilities to process U-Plane are selected.

1. Introduction

At the initial deployment phase of 5th Generation mobile communication systems (5G), the 5G Non-Stand-Alone (NSA) architecture*1 was widely adopted to realize 5G services by connecting 5G base stations to the existing Evolved Packet Core (EPC)*2 [1]. As applications based on 5G become more widespread, the need for EPC to achieve higher speed and capacity communications, lower latency communications and simultaneous connection of many terminals than ever has become urgent [2] [3]. Specifically, it is necessary to increase the number of high-capacity gateway devices capable of processing hundreds of Gbps to several Tbps to achieve high-speed, high-capacity communications,

*1 5G NSA architecture: A 5G system configuration in which 5G radio base stations and EPCs are linked. This enables a low barrier to deployment because it allows 5G to be provided without installing 5GC (See *38).

*2 EPC: The fourth-generation IP-based core network specified by 3GPP for LTE and other access technologies.

to distribute gateway devices near base station facilities to achieve even lower latency communications, and to improve session*3 processing performance for connecting massive numbers of terminals simultaneously.

Conventional single gateway devices have both Control Plane (C-Plane)*4 functions to manage communication sessions and control communications, and User Plane (U-Plane)*5 functions to handle communications traffic [4]. Therefore, if the previously assumed balance between the number of sessions and communications capacity is disrupted, either the C-Plane or the U-Plane will have excess processing capacity. In high-speed, high-capacity communications, the C-Plane has excess processing power, and in multiple terminal simultaneous connections, the U-Plane has excess processing power because the volume of communications is small compared to the number of sessions. If the C-Plane and U-Plane can be scaled*6 independently, these issues can be resolved, and efficient facility design can be expected. In addition, low-latency communications require distributed deployment of the U-Plane function near the base station facilities to reduce propagation delay. However, in the distributed deployment of conventional devices with integrated C-Plane and U-Plane functions, the number of sessions and communication volume are unevenly distributed among the gateway devices, resulting in a decrease in the efficiency of facility utilization. Since there is no need for distributed deployment of C-Plane functions, if the C-Plane and U-Plane functions can be separated and the way they are deployed changed according to their characteristics, the loss of facility utilization efficiency related to C-Plane processing capacity could be greatly reduced.

From above background, NTT DOCOMO has been planning to deploy Control and User Plane Separation (CUPS)*7 architecture to realize the separation of C-Plane and U-Plane functions as specified in 3rd Generation Partnership Project Technical Specification (3GPP TS) 23.214. Separating the C-Plane and U-Plane functions of gateway devices with CUPS architecture makes it possible to scale the C-Plane and U-Plane independently and balance the centralized deployment of C-Plane functions with the distributed deployment of U-Plane functions, thereby enabling the deployment and development of a flexible and efficient core network*8 [5]. In addition to solving the aforementioned issues, CUPS will also enable independent equipment upgrades for C-Plane and U-Plane functions, and the adoption of U-Plane devices specialized for specific traffic characteristics.

In the user perspective, the introduction of CUPS can be expected to dramatically improve the user experience through the operation of facilities specializing in various requirements, and enable further increases in facilities and lower charges to pursue user benefits by improving the efficiency of core network facilities.

Regarding the CUPS architecture, a source of value for both operators and users, this article includes an overview of the architecture, additional control protocols, U-Plane control schemes based

*3 Session: A generic term that includes a virtual communications channel for exchanging data in the U-Plane, data exchanged in the communications channel, and metadata such as management information exchanged in the C-Plane about the communication channel.

*4 C-Plane: A series of control processes that are exchanged to establish communications and authentication.

*5 U-Plane: The process of sending and receiving user data, which are the main signals for communication.

*6 Scaling: Designing the capacity of processing resources and planning the increase or decrease of communications facilities based on the performance characteristics of devices, traffic models reflecting communication characteristics, expected number of subscribers, and spare capacity for burst traffic.

*7 CUPS: An architecture in which the C-Plane and U-Plane functions of SGW and PGW are separated and specified as separate devices.

*8 Core network: A network consisting of gateway devices, location management devices, subscriber information management devices, and mobility management functions. The core part of the network that constitutes a mobile communication system. A mobile device communicates with the core network via a radio access network consisting of radio base stations.

on traffic characteristics, and future developments toward a 5G Stand-Alone (5G SA) architecture*9.

2. CUPS

2.1 CUPS Overview

1) Concept and Architecture

The architecture of the EPC with CUPS is shown in **Figure 1**. CUPS is an architecture defined in 3GPP TS23.214 that separates the Serving GateWay (SGW)*10/Packet data network GateWay (PGW)*11 configuration of the EPC into the C-Plane and U-Plane. The CUPS architecture is designed so that there is no difference in the interface between the existing architecture and the CUPS architecture - even with CUPS architecture deployed in SGW/PGW, opposing devices such as a Mobility Management Entity (MME)*12, Policy and Charging Rules Function (PCRF)*13, evolved NodeB (eNB)*14/next generation NodeB (gNB)*15, and SGWs/PGWs of other networks such as Mobile Virtual Network Operator (MVNO)*16 and roaming*17 are not affected. For C-Plane, SGW Control plane function (SGW-C)*18/PGW Control plane function (PGW-C)*19, and for U-Plane, SGW User plane function (SGW-U)*20/PGW User plane function (PGW-U)*21 are equipped with call processing functions. By introducing CUPS, C-Plane/U-Plane capacities can be expanded individually as needed. Combined SGW-C/PGW-C*22 and Combined SGW-U/PGW-U*23 can

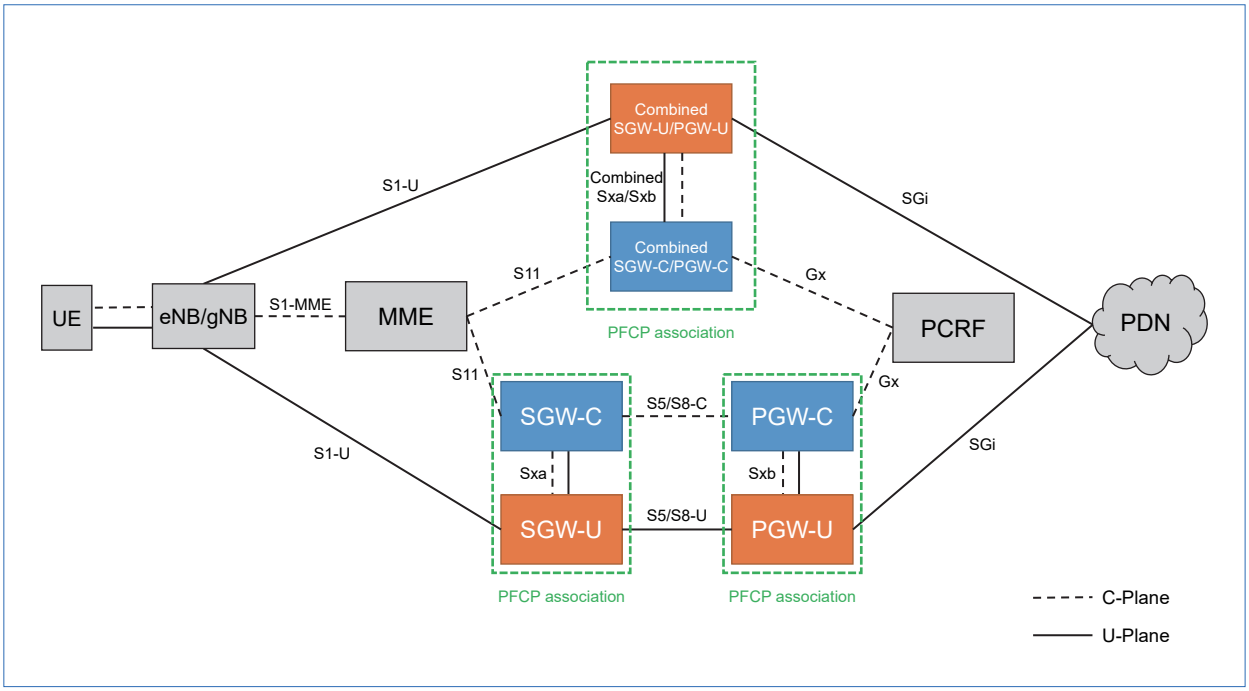

Figure 1 EPC architecture with CUPS introduced

*9 5G SA architecture: One of the system configurations of 5G. 5G is provided by 5G radio base stations and 5GCs. It is referred to as Stand-Alone configuration because it is an all-5G system that makes 5G-specific function available.

*10 SGW: The gateways (Serving GW and PDN GW) deal with the user plane. They transport the IP data traffic between the User Equipment (UE) and the external networks. The Serving GW is the point of interconnect between the radio-side and the EPC. As its name indicates, this gateway serves the UE by routing the incoming and outgoing IP packets. (source: 3GPP)

*11 PGW: The PDN GW is the point of interconnect between the EPC and the external IP networks. PDN stands for "Packet Data Network". The PDN GW routes packets to and from the PDNs (source: 3GPP).

*12 MME: The MME (Mobility Management Entity) deals with the control plane. It handles the signaling related to mobility and security for E-UTRAN access (source: 3GPP).

*13 PCRF: A logical device that manages and controls user billing policies.

*14 eNB: A radio base station that supports the LTE radio system.

*15 gNB: A radio base station that supports the 5G radio system.

handle the functions of SGW and PGW in common devices. In the standard specification, in addition to SGW/PGW, the Traffic Detection Function (TDF)*24 can also be separated into TDF-C and TDF-U, but the details are omitted in this article.

SGW-C/PGW-C have the C-Plane interfaces (S5/S8-C, S11, Gx, etc.) to handle the C-Plane and SGW-U/PGW-U have the U-Plane interfaces (S1-U, S5/S8-U, SGi, etc.) to handle the U-Plane. Interfaces that handle both C-Plane and U-Plane are separated into different interfaces for C-Plane and U-Plane. For example, for S5, SGW-C/PGW-C and SGW-U/PGWU have S5-C/S5-U, respectively. Between SGW-C/PGW-C and SGWU/PGW-U there are Sx interfaces as defined in 3GPP TS29.244 [6]. The interface between SGW-C and SGW-U is called Sxa, between PGW-C and PGW-U is called Sxb, and between Combined SGWC/PGW-C and Combined SGW-U/PGW-U is called Combined Sxa/Sxb.

For the Sx interface, the Packet Forwarding Control Protocol (PFCP)*25 specified in 3GPP TS29.244 is used. PFCP is described in detail later. It is also possible to send U-Plane over the Sx interface with GPRS Tunneling Protocol for User Plane (GTP-u)*26.

2) Standard Call Processing in CUPS

The call processing sequence for SGW/PGW with CUPS architecture is designed so that it does not affect existing architecture. The only difference between CUPS and non-CUPS architectures is more signaling between SGW-C/PGW-C and SGW-U/PGW-U. An example of LTE Attach*27 is shown in **Figure 2**. The MME that receives the Attach request selects the SGW-C/PGW-C respectively, and sends a session establishment request to the selected SGW-C (Fig. 2 (1)). SGW-C selects the SGW-U for establishing the session (Fig. 2 (2)). The SGW-C sends the PFCP session establishment request to the SGW-U for establishing the session (Fig. 2 (3)). The SGW-C then sends a session establishment request to the PGW-C (Fig. 2 (4)). Upon receiving this, PGW-C obtains the billing policy and other information for the session to be established from PCRF as well as non-CUPS architecture. The PGW-C that establishes the session next selects the PGW-U and sends the PFCP session establishment request to establish the session (Fig. 2 (5)). SGW-C sends the U-Plane path information about PGW-U obtained from the response to the session establishment request from PGW-C to SGW-U, and establishes the U-Plane path between SGW-U and PGW-U by changing the PFCP session (Fig. 2 (6)). After the MME completes the configuration of the eNB side, the SGW-C receives the eNB U-Plane path information. SGW-C notifies SGW-U of the eNB-side U-Plane path information, and SGW-U establishes the U-Plane path with the eNB side (Fig. 2 (7)).

2.2 Control between CUs

1) PFCP Overview

PFCP is a protocol used by SGW-C/PGW-C that controls packet forwarding between SGW-U/PGW-U with a set of rules. The protocol enables SGW-C/PGW-C to assign a set of rules to SGW-U/SGW-U on a per session basis.

PFCP signals are classified into two types: Node

*16 MVNO: A mobile communication service provider that does not own the radio access network infrastructure over which it provides services to customers.
*17 Roaming: A mechanism that enables mobile subscribers to receive services from visited networks when traveling outside the geographical coverage area of their home networks.
*18 SGW-C: The functional section that performs SGW C-Plane processing in CUPS architecture.
*19 PGW-C: The functional section that performs PGW C-Plane processing in CUPS architecture.
*20 SGW-U: The functional section that performs SGW U-Plane processing in CUPS architecture.
*21 PGW-U: The functional section that performs PGW U-Plane processing in CUPS architecture.
*22 Combined SGW-C/PGW-C: A single C-Plane processing functional section that combines SGW-C and PGW-C and combines the functions of both.
*23 Combined SGW-U/PGW-U: A single U-Plane processing functional section that combines SGW-U and PGW-U and combines the functions of both.
*24 TDF: A functional section that identifies traffic, detects applications, and notifies the PCRF. It is not implemented by NTT DOCOMO.

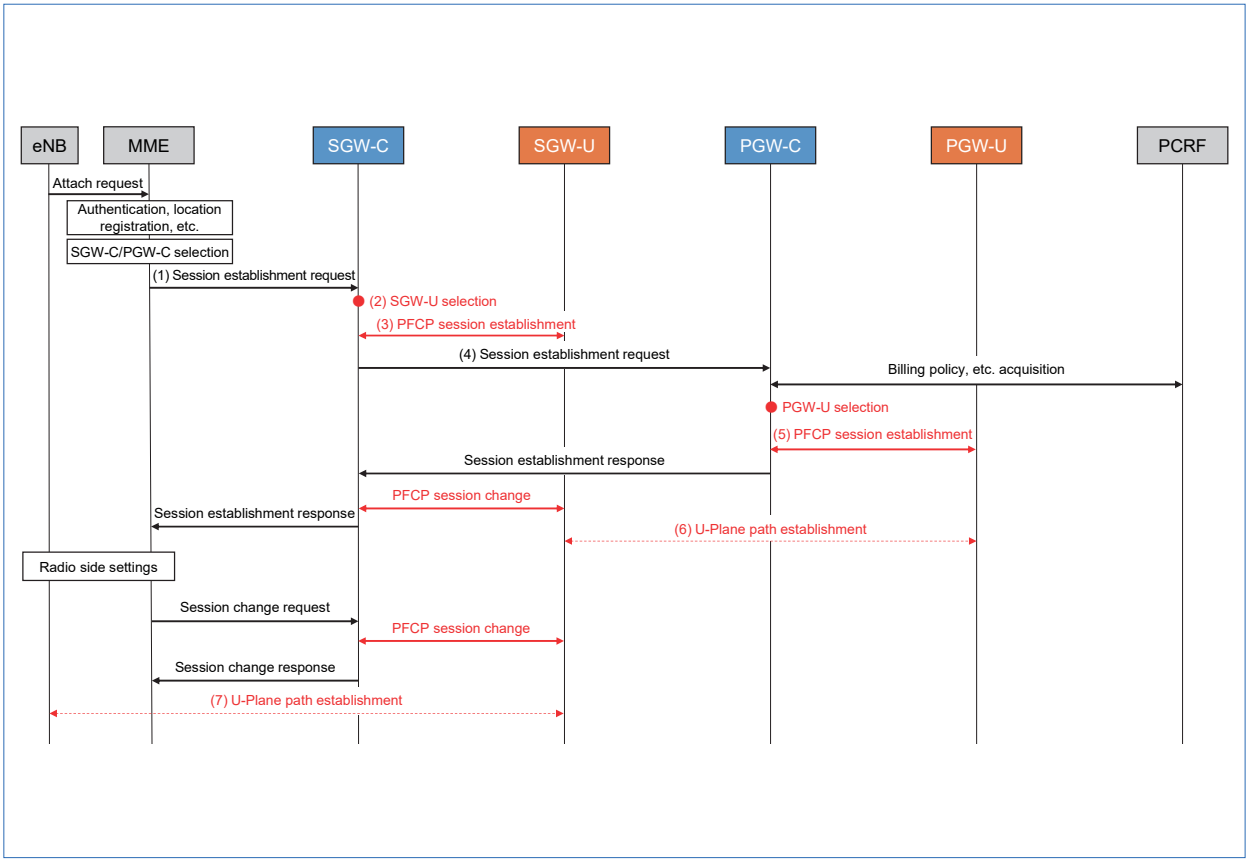

Figure 2 Attach sequence

Related Procedures/Messages that specify inter-device control and Session Related Procedures/Messages that specify PFCP session control. The SGW-C/PGW-C and SGW-U/PGW-U pair is called a PFCP association (Fig. 1). SGW-C/PGW-C can only control SGW-U/PGW-U in a PFCP association.

2) Packet Processing Model

In PFCP, multiple rules are combined to control packet processing. The sets of rules used in PFCP are shown in **Table 1**. There are five types of rules: Packet Detection Rule (PDR), Forwarding Action Rule (FAR), Buffering Action Rule (BAR), Quality of Service (QoS)*28 Enforcement Rule (QER), and Usage Reporting Rule (URR). In PFCP, packet processing is achieved by combining various rules around PDR. In PDR, the receiving interface and 5-tuple information*29 to be monitored are defined. When a packet is received, the SGW-U/PGW-U judges whether the 5-tuple information associated with the packet meets the PDR conditions. If the conditions are met, packet processing is performed according to the specified set of rules. For PDR,

*25 PFCP: C-Plane protocol used at the Sx reference point, where SGW-C/PGW-C instructs SGW-U/PGW-U on the packet control method using PFCP.

*26 GTP-u: A tunneling protocol used by radio base stations and devices in the core network to transmit user data.

*27 Attach: The processing of registering a mobile UE with a network when UE power is turned on, or the state of being registered.

*28 QoS: A technology for properly managing communications quality on a network by marking packets and giving them priority in processing, such as giving priority over data transfer to avoid interruptions in voice calls, etc.

*29 5-tuple information: A generic term for five pieces of information stored in the IP header and Transmission Control Protocol (TCP)/User Datagram Protocol (UDP) header: the destination IP address, destination port number, source IP address, source port number, and protocol number.

Table 1 PFCP rule groups

Rule name	Role
PDR	Specifies judgment conditions for received packets and the set of rules to be applied to the packets.
FAR	Specifies whether or not the packet is to be forwarded by discarding or buffering it, in addition to operations related to forwarding such as the tunneling header information to be added to the packet and the interface to be used.
BAR	Specifies the maximum number of packets to be retained and the dwell time between the arrival of a downlink packet and the notification to the SGW-C, etc. for buffering.
QER	Specifies the QoS of forwarded packets, such as the allowed bandwidth and the DSCP value assignment.
URR	Specifies how to count the packets detected by the associated PDR and when to notify the C-Plane device of the count status.

settings are made individually, such as PDR for detecting downlink packets and PDR for detecting uplink packets. Also, it is possible to change the rules to be applied for each destination IP address or protocol, for example, it is possible to specify that packets related to Dynamic Host Configuration Protocol for IP version 6 (DHCPv6)*30 are forwarded from PGW-U to PGW-C. Except PDR, it is possible to associate each rule with multiple PDRs. For example, by associating one QER with multiple PDRs, it is possible to have the same Differentiated Services Code Point (DSCP)*31 value assigned by SGW-U/PGW-U.

Now we describe the actual flow of applying rules. At first, SGW-C/PGW-C informs the multiple PDRs and the respective rules associated with them to SGW-U/PGW-U. Second, the SGW-U/PGW-U judges the group of PDRs notified when packets are received in order, and searches for matching PDRs. If a matching PDR is found, the received packet is processed based on the FAR associated with the PDR. If the content of the FAR is buffering, buffering is performed based on BAR. When forwarding packets, DSCP marking and bandwidth controls are performed based on QER. Lastly, the packet count and volume count are performed using the methods specified in URR, and if it is necessary to notify the SGW-C/PGW-C, the count information and other information is notified. In this way, PFCP rules are combined to perform packet processing in SGW-U/PGW-U.

3) U-Plane Device Management

SGW-C/PGW-C establish PFCP associations for SGW-U/PGW-U and manage these associations before actual U-Plane processing. The SGW-U/PGW-U that is the candidate for selection at the time of session establishment is the device that established the PFCP association. When establishing a PFCP association, it is possible to choose whether to share the optional functions of the standard specification between SGW-C/PGW-C and SGW-U/PGW-U. SGW-C/PGW-C and SGW-U/PGW-U exchange their

*30 DHCPv6: A protocol for distributing DNS server information, address information, and other information necessary for connecting to a network using IPv6.

*31 DSCP: A value that indicates the priority of a packet when controlling the QoS priority of IP packets. It is represented by the first six bits of the Type of Service in the IP header. 64 levels of priority can be specified.

capabilities in Node Related Procedures/Messages and decide which one will support the functions depending on how optional functions are handled.

SGW-C/PGWC and SGW-U/PGW-U that establish a PFCP association send heartbeat packets*32 periodically to each other for alive monitoring and confirming restart time. Heartbeat packets can be sent and received in both directions between these devices.

3. U-Plane Control Based on Traffic Characteristics

3.1 GW Selection after CUPS Introduction

1) CUPS GW Selection Method

For EPC, the MME selects the SGW/PGW using Domain Name System (DNS)*33. The MME selects an appropriate SGW/PGW using DNS based on key information such as Access Point Name (APN)*34 and location of the terminal. Before the introduction of CUP architecture, the MME determines the SGW/PGW by considering both the C-Plane perspective, such as number of sessions, and the U-Plane perspective, such as expected traffic volume and physical distance. In contrast, after the introduction of CUPS, the MME and SGW-C/PGW-C will share the responsibility of selecting U-Plane path. The MME will select SGW-C/PGW-C, and SGW-C/PGW-C will select SGW-U/PGW-U. Two types of SGW-U/PGW-U selection methods are specified in the standard specification. The first is that SGW-C/PGW-C selects SGW-U/PGW-U alone. The second is that SGW-C/PGW-C selects SGW-U/PGW-U cooperating with DNS. In the standard specification, the second method is optional, so this article describes the first method.

The three main parameters that can be used for SGW-U/PGW-U selection as specified in the standard specification are as follows:

(1) C-Plane information received in the session establishment request, including APN and terminal location information.
(2) Static information about the performance and functions of SGW-U/PGW-U.
(3) Dynamic information such as the load status of SGW-U/PGW-U.

Basically, the selection of SGW-U/PGW-U is the same as selection by a conventional MME, using APN and location information, but (3) leads to the realization of functions that were difficult to achieve with the selection by the MME. An example of (3) is notification of the current load status as a parameter for load balancing, although even in the past there have been specifications for notifying the MME of the load status from the SGW/PGW and utilizing it. However, since the necessity of utilizing the notified SGW/PGW load information depends on the implementation status of the MME, there was a possibility that the information would not be properly utilized by roaming and MVNOs. After the introduction of CUPS, load balancing for SGW-U/PGW-U and selection according to traffic is done by SGW-C/PGW-C on the operator's network so devices can be selected based on the load information.

*32 Heartbeat packet: A packet for survival confirmation or its survival response sent to monitor the life and death of an opposing device. Timestamps and other information are sent and received in PFCP heartbeats.
*33 DNS: A function that resolves domain names and IP addresses on a network. In the core network, it is used for service discovery for gateway devices, etc.
*34 APN: An identifier that specifies the connection destination for the UE (See *36), used by the UE as an identifier to specify the PDN to connect to when requesting a connection to the core network.

2) Device Deployment after CUPS Installation

A **Figure 3** shows an example of device deployment after CUPS installation. After the introduction of CUPS, SGW-C/PGW-C and SGW-U/PGW-U can be deployed independently. In the standard specification, SGW-C/PGW-C and SGW-U/PGW-U can be connected N to N. For example, SGW-C/PGW-C can be redundantly distributed in consideration of disasters and congestion, so that even if one SGW-C/PGW-C is out of order, SGW-U/PGW-U can be selected from another SGW-C/PGW-C, thus making effective use of resources.

With the introduction of CUPS, SGW-C/PGW-C have an architecture that does not connect to eNB/gNB. This is advantageous because it eliminates the need to consider the transmission distance or the number of eNBs/gNBs that can be connected. Also, since geographical considerations are no longer necessary and centralized management can be realized, maintenance efficiency can be expected to increase with a reduction in the number of maintenance sites.

SGW-U/PGW-U can be deployed according to the requirements of the services to be accommodated and their locations. For example, this will realize the concepts of deploying a large number

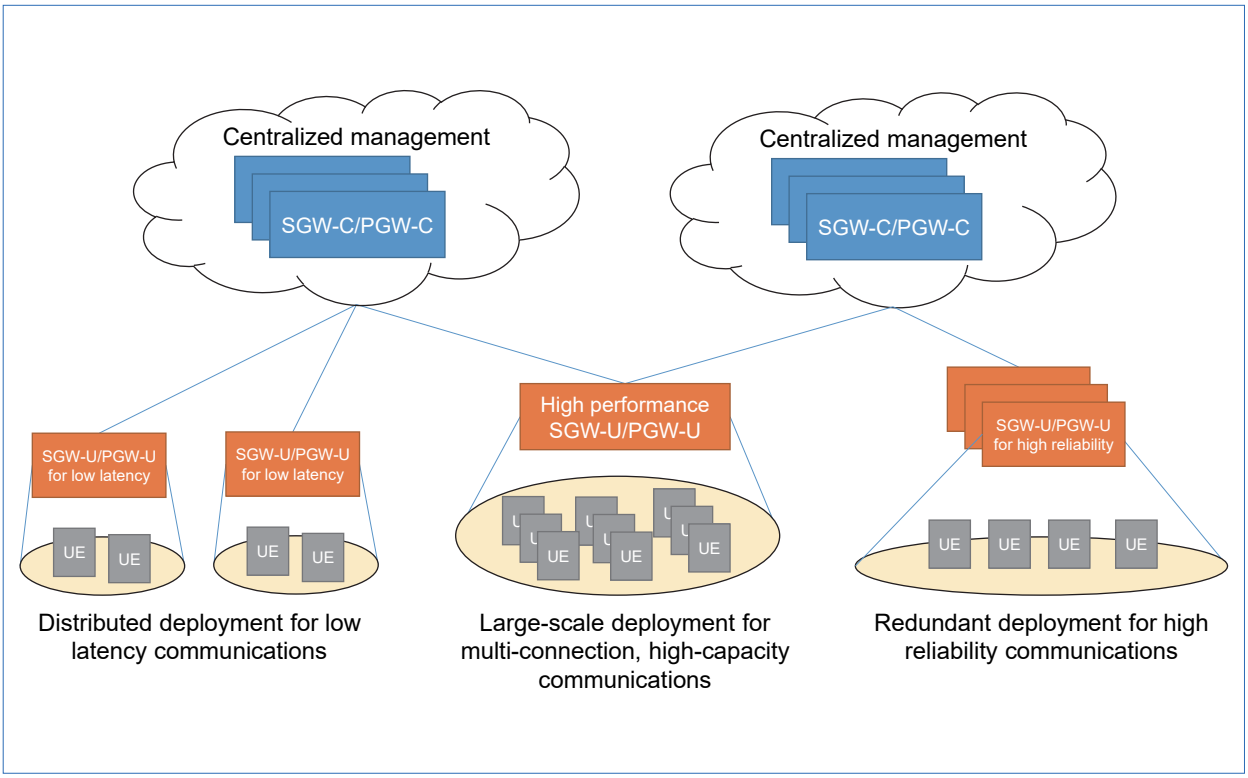

Figure 3 Example of device deployment after CUPS installation

of devices in detail nationwide for low latency data communication services, deploying high performance SGW-U/PGW-U in urban areas where there is a large population and high speed and large capacity communications are required based on regional characteristics, and deploying devices with high redundancy and reliability for disaster resilient services based on service requirements (Fig. 3).

3) Improving the Efficiency of Traffic Routes

In this section, we explain the concept and issues of traffic route efficiency before the introduction of CUPS, and how these issues can be solved with CUPS.

- Issues before the introduction of CUPS

An example of a traffic route before the introduction of CUPS is shown in **Figure 4**. Generally, in Evolved Packet System (EPS)*35, multiple sessions are set up using a multiple Packet Data Network (PDN)*36 per User Equipment (UE). An example is the use of IP Multimedia Subsystem (IMS)*37 sessions for voice services and sessions for data communication services. The two axes of traffic route efficiency are the accommodation of facilities according to service requirements and the reduction of transmission paths.

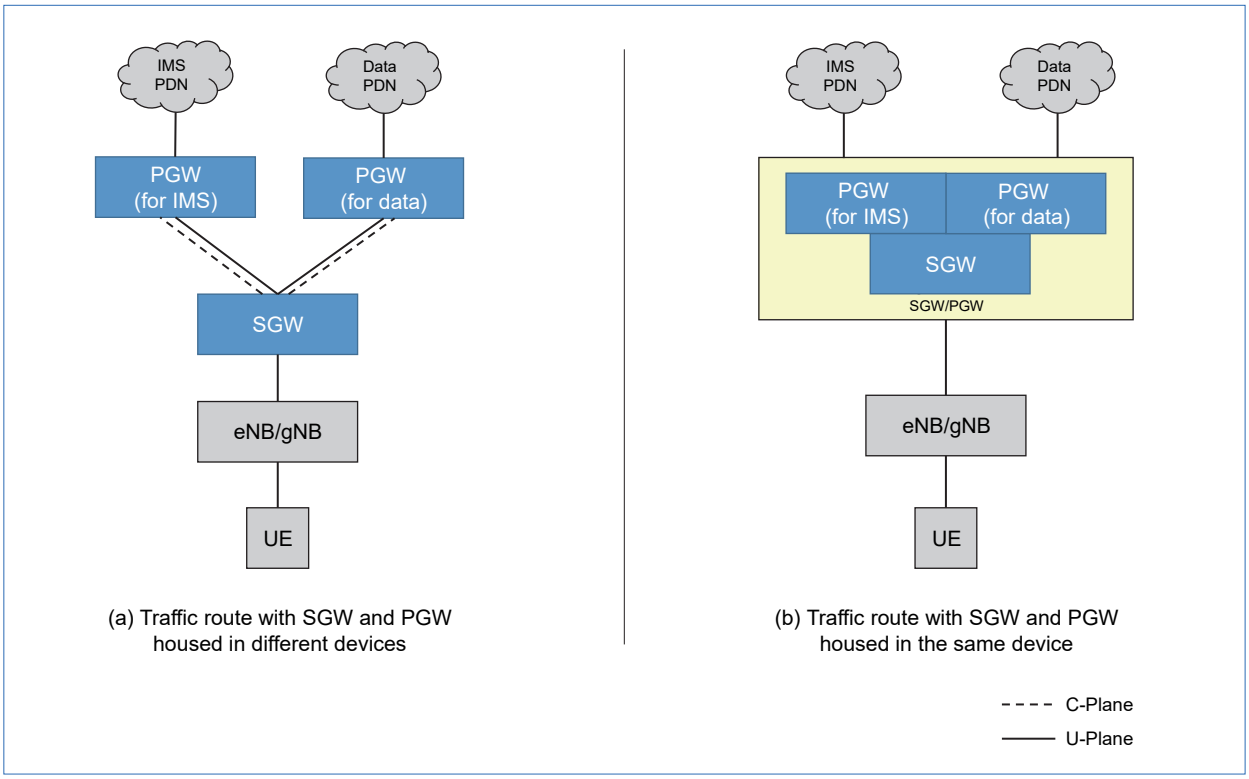

Figure 4 Example of a traffic route before the introduction of CUPS

*35 EPS: The generic name for IP-based packet networks specified by 3GPP for LTE and other access technologies.

*36 PDN: The external packet network to which the mobile core network connects.

*37 IMS: A standardized system for providing multimedia services, including voice communications, over packet communications networks.

When selecting a device according to service requirements, only one SGW per UE can be selected (Fig. 4 (a)), although it was possible to select several different PGWs according to the session with APN as the key. This led to the issue of the same SGW device requirements being applied even though service requirements differ from session to session.

Also, when considering transmission path reduction, it is possible to reduce the path of the S5 reference point by selecting SGW and PGW to be the same device by the MME. However, since only one SGW can be selected, these must be housed in the same SGW/PGW for S5 route reduction for voice and data communication services (Fig. 4 (b)). Therefore, even though PGWs can be housed in different devices depending on service requirements, the need to house them in the same device was an issue.

Thus, before the introduction of CUPS, it was difficult to both accommodate equipment to meet service requirements and reduce the transmission path.

- Improving traffic route efficiency after introducing CUPS

An example of a traffic route after the introduction of CUPS is shown in **Figure 5**. After the introduction of CUPS, SGW-U/PGW-U can be selected for each session, enabling the construction of facilities to meet service requirements.

In cases where SGW-C and PGW-C are selected from different components, each function selects U components based on independent selection rules (Fig. 5 (a)). However, since the current standard specification stipulates that APNs cannot be used when selecting SGW-U, there is a concern that traffic routes cannot be constructed using appropriate devices. For example, to simultaneously provide a voice service with high reliability requirements and a data communications service with low latency requirements to a single UE, it would be ideal to construct a route by accommodating the voice service in a device with high reliability and the low latency data communication service in a device with a short transmission distance. However, if the SGW-U cannot be selected according to the APN, there is a concern that the SGW-U for the voice service will be used to build the traffic route for the low latency data communication service.

This issue can be solved by selecting SGW-C and PGW-C as Combined SGW-C/PGW-C. SGW-C/PGW-C are deployed centrally because geographical considerations are not required, and selection is made by the MME so that SGW-C and PGW-C become Combined SGW-C/PGW-C (Fig. 5 (b)). When selected, SGW-U/PGW-U become controllable as a Combined SGW-U/PGW-U, and PGW-U selection rules can thus be applied to SGW-U/PGW-U selection parameters. As

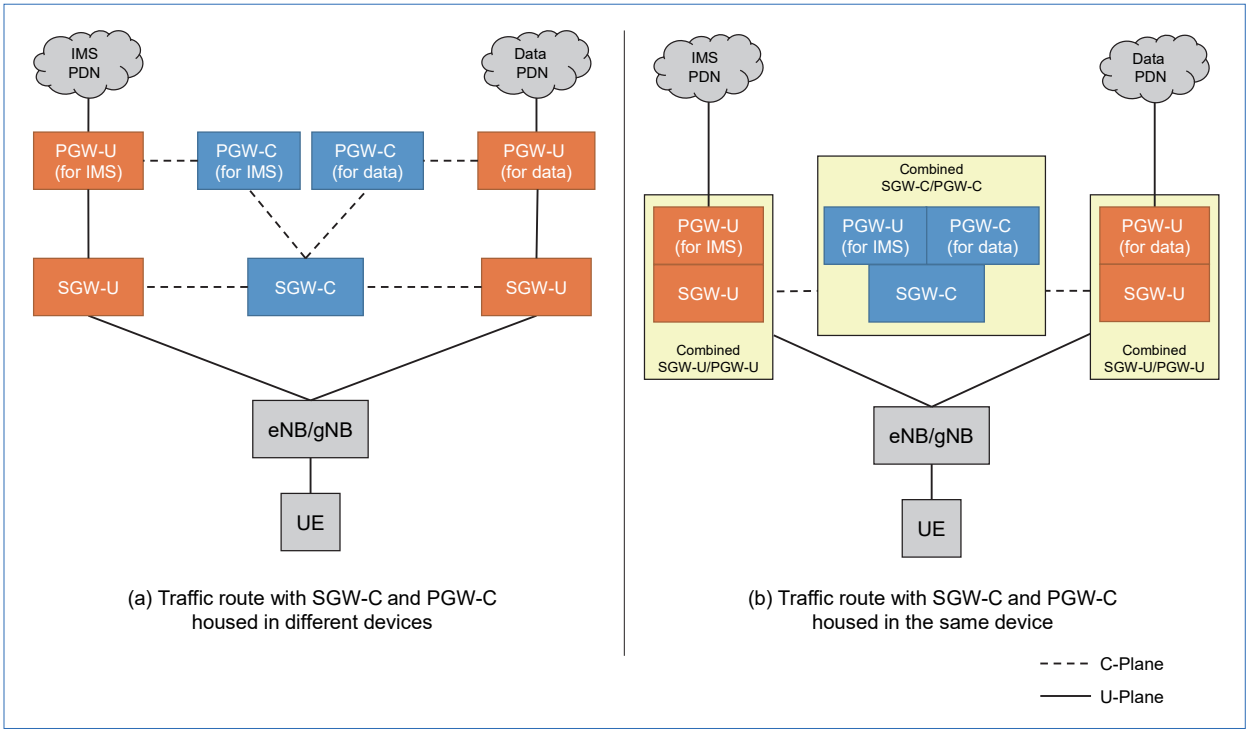

Figure 5 Example of a traffic route after the introduction of CUPS

a result, the SGW-U is selected according to the APN. From the transmission path perspective, the path of the S5-U reference point can be reduced in Combined SGW-U/PGW-U in the same way as before the introduction of CUPS. Furthermore, placing SGW-U/PGW-U closer to the eNB/gNB makes it possible to reduce the transmission distance of the S1-U reference point.

3.2 Expansion of SGW-U/PGW-U

1) Core Network Development with CUPS

SGW/PGW faces various network nodes such as eNB/gNB, MME and PCRF. Increasing the number of SGW/PGW variations was difficult because of the wide range of technical considerations to oppose existing devices. After the introduction of CUPS, SGW-U/PGW-U counterparts will be more limited than before the introduction of CUPS because the effects related to C-Plane can be closed by SGW-C/PGW-C. This will make it easier to increase the number of variations of SGW-U/PGW-U. If the new SGW-U/PGW-U to be introduced can be controlled from existing SGW-C/PGW-C, the only other consideration is the U-Plane, which has the advantage of a lowered threshold for introduction. In addition, development of SGW-U/PGW-U itself can be expected. For example, it will be

possible to provide SGW-U/PGW-U to meet required service levels and equipment requirements, such as products suitable for simultaneous connection of many terminals for event venues and low-cost products for distributed deployment for low latency services. Operators will be able to develop flexible core networks by deploying SGW-U/PGW-U to suit their use cases.

2) Precautions for Device Expansion

To expand SGW-U/PGW-U, it is necessary to consider interconnectivity with SGW-C/PGW-C. PFCP clearly specifies division of functions between SGW-C/PGW-C and SGW-U/PGW-U. However, some functions may be implemented by both SGW-C/PGW-C and SGW-U/PGW-U. Therefore, it is necessary for operators to evaluate the method to be adopted.

We explain end markers as an example that needs to be evaluated in interconnectivity. The end marker is a function that notifies the end of forwarded packets to the old route when the route is switched due to handover. In the standard specification, the end marker is defined to be generated by SGW-C/PGW-C. Generation by SGW-U/PGW-U is specified as an option. However, comparing the two, the latter generation by SGW-U/PGW-U is considered to be more suitable. In terms of the amount of signal, the method generated by SGW-C/PGW-C requires an end marker transmission procedure separate from the handover control, while the method generated by SGW-U/PGW-U does not increase the amount of signal because the end marker transmission direction is performed in the handover control. In addition, in the SGW-C/PGW-C scheme, an independent session for end marker transfer between SGW-C/PGW-C and SGW-U/PGW-U needs to be established, and technical studies on the interconnection of independent sessions are required. For these reasons, we believe that the specifications generated by SGW-U/PGW-U are superior for the end marker. As such, there are cases where optional methods are used, hence, care is required when expanding the variations of SGW-U/PGW-U.

In Release 17, the functions that can be performed by both SGW-C/PGW-C and SGW-U/PGW-U, including the aforementioned end marker, were again discussed and recommended methods were clarified. NTT DOCOMO plans to follow the recommended regulations and expand the variations of SGW-U/PGW-U.

4. Development towards 5G Interworking

The 5th Generation Core network (5GC)*38 that is currently being introduced for the commercialization of the 5G SA system uses the CUPS architecture as same as in EPC. In interconnection with EPC, the SMF+PGW-C, a single device combining the 5GC Session Management Function (SMF)*39 and PGW-C has the PGW-C function of EPC. In the same way, the UPF+PGW-U, a single device combining the 5GC User Plane Function (UPF)*40 and PGW-U has the PGW-U function of EPC (**Figure 6**) [7].

SMF+PGW-C has a high degree of similarity to

*38 5GC: The core network specified by 3GPP for fifth-generation mobile telecommunications systems.
*39 SMF: The functional section that manages sessions in 5GC. Equivalent to SGW-C/PGW-C in EPC.
*40 UPF: The functional section that relays and terminates the U-Plane in the 5GC. Equivalent to SGW-U/PGW-U in EPC.

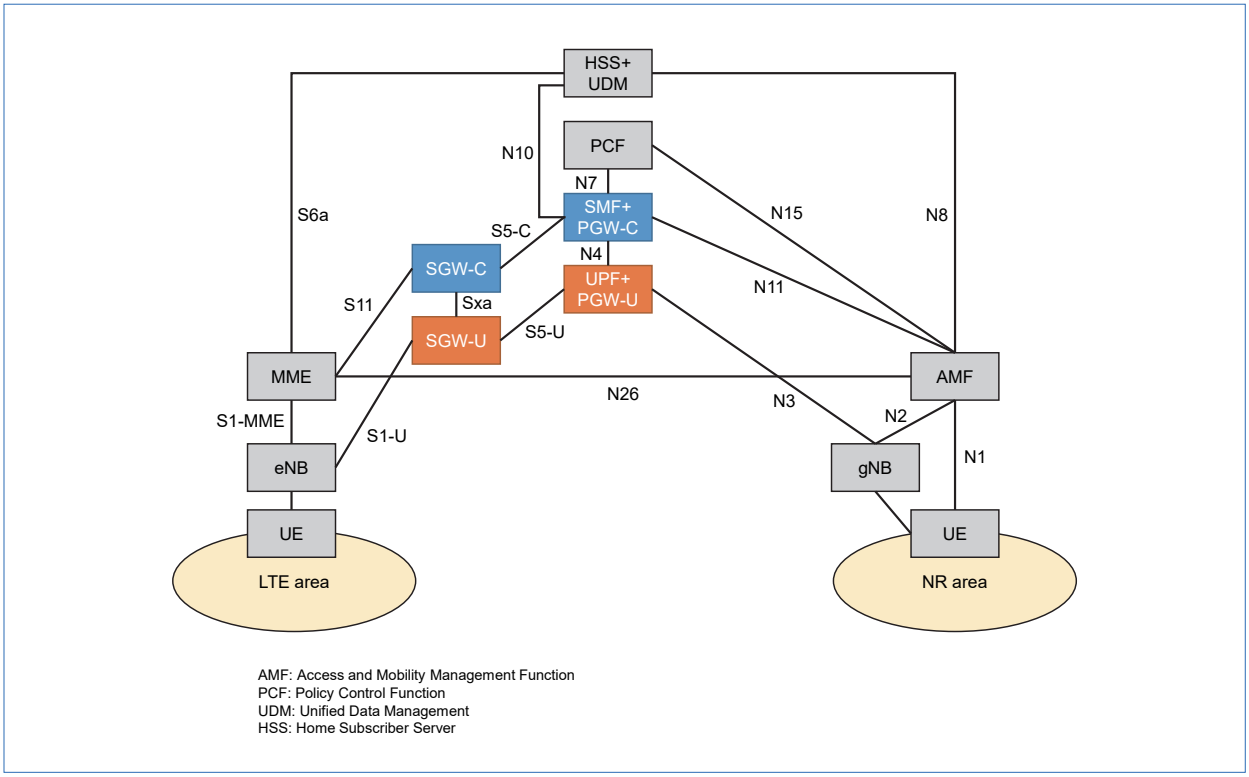

Figure 6 Introduction of SMF+PGW-C and UPF+PGW-U in EPC-5GC interwork

EPC CUPS architecture since it uses the aforementioned PFCP to control UPF+PGW-U. If CUPS architecture can be applied to EPC prior to the introduction of 5GC, a seamless transition from EPC to 5GC can be expected. Specifically, if SGW-C and SGW-U functions can be equipped in SMF+PGW-C and UPF+PGW-U, respectively, a transition plan can be considered in which either or both SGW-C and SGW-U are rolled into 5GC in stages.

5. Conclusion

This article described CUPS architecture in mobile core networks, the control scheme in CUPS architecture, the control protocol PFCP, and the GW selection scheme for flexibly selecting various SGW-U/PGW-U according to use case, and introduced the advantages of CUPS for both mobile operators and users.

NTT DOCOMO has been applying CUPS architecture to EPC to flexibly and optimally accommodate the increased traffic resulting from the introduction of 5G NSA and achieve smooth interconnection with the soon-to-be-introduced 5GC. In the future, we plan to expand the variation of SGW-U/PGW-U devices and introduce and deploy 5GC networks.

REFERENCES

[1] Y. Sagae et al.: "5G Network," NTT DOCOMO Technical Journal, Vol.22, No.2, pp.23–39, Oct. 2020.
[2] E. Endo et al.: "Overview of 5G Pre-commercial Service," NTT DOCOMO Technical Journal, Vol.21, No.3, pp.4–8, Jan. 2020.
[3] Y. Kojo et al.: "Overview of 5G Commercial Service," NTT DOCOMO Technical Journal, Vol.22, No.1, pp.4–9, Jul. 2020.
[4] 3GPP TS23.401 V15.12.0: "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," Sep. 2020.
[5] 3GPP TS23.214 V15.5.0: "Architecture enhancements for control and user plane separation of EPC nodes; Stage 2," Dec. 2018.
[6] 3GPP TS29.244 V15.10.0: "Interface between the Control Plane and the User Plane nodes," Sep. 2020.
[7] 3GPP TS23.501 V15.12.0: "System architecture for the 5G System (5GS)," Dec. 2020.