Building a cloud-native 5G core will be essential

David Nowoswiat - April 23, 2018

As we get closer to a full 5G 3GPP specification, it is clear that the cloud will not simply be part of the 5G network, it will be a foundational element; the network will be “cloud native.” This became apparent in later releases of the 4G spec, including release 14, in which certain cloud and virtual architectural options began pointing toward a preferred direction on the road to 5G.

As I highlighted in a previous article, 5G services such as augmented and virtual reality, HD video, IoT, and critical machine-to-machine communications (MTC, or sometimes just M2M) such as autonomous vehicles and industrial automation, will be more varied than traditional network services. Some services will have more stringent requirements in terms of bandwidth capacity, latency, and reliability, while also being less predictable in terms of demand, location, revenues, etc. Existing call models simply won’t apply.

Access networks are those that connect subscribers to their communications providers; the core is comprised of the networks that connect the providers. Access networks are being virtualized so that they are programmable. In order to respond quickly and adapt to changing service demands and requirements, the new 5G core must likewise be programmable.

A key attribute of this programmable core is support for network slicing – the ability to logically isolate segments of the networks. This will enable network resources for access, transport/backhaul and core to be dynamically allocated either vertically or horizontally.

Vertical slices will support a specific service or set of common service characteristics (e.g., narrowband-IoT, non-IP data, power-savings mode devices). Horizontal slices can be reserved for specific enterprise customers, for instance a public utility (think smart grid or intelligent transport systems), or a private company’s extended intranet. This sharing of a common pool of network resources through programmability and network slicing is more efficient and cost effective than building traditional overlay networks or over-engineering the network capacity to account for unpredictable spikes in demand.

In order to meet these new service requirements, the 5G core must be built from a cloud-native design. That involves taking the best practices, concepts and technologies from the IT and webscale environment and applying them to the design of the core network. A simple software upgrade of the existing packet core into a virtualized environment won’t do.

User and control plane separation
In today’s mobile network, the mobile gateway (GW) performs several functions in the data path between the applications/services and the end user device. It performs as the serving gateway (SGW) and/or packet data network gateway/gateway GPRS support node (PGW/GGSN) — as the user
device anchor point and the packet forwarding function. It can also perform the traffic-detection function (TDF) for packet inspection and traffic classification.

These functions will be instantiated in software – they’ll be virtualized. Virtualization enables one of the key features of the cloud-native core architecture – the complete separation of the control and user plane functions within the mobile gateway. In 3GPP release 14, this feature is specified as control/user plane separation (CUPS) in the evolved packet core (EPC) nodes (TS 23.214).

With this feature (Figure 1), the mobile gateway function essentially splits into two: the gateway control (GW-C) function, which handles session/bearer and IP address management for end-user devices, as well as signaling communications between itself and the other core network functions; and gateway user (GW-U) functions, which do the packet forwarding, marking, deep packet inspection, policy control, lawful intercept, and charging. In a virtualized environment, the GW-C and GW-U would be separate virtual network function (VNF) instances.

With the CUPS architecture, the GW-C and GW-U can be independently scaled and located wherever they are needed. Depending upon the architecture and requirements of the core network, these functions can also be separate or combined within the same virtual instance.

![Figure 1](image1.png) CUPS principles illustrated for Sx interface between GW-C and GW-U functions

This facilitates a distributed mobile gateway architecture where, for example, the user plane can be placed in a data center at the cloud edge, close to the end-user or IoT device, and the control plane can be placed in a central data center to handle the control signaling. This distributed configuration would be necessary to support low-latency applications such as autonomous vehicles or mobile gaming.

As shown in Figure 1, the GW-C can interface with more than one GW-U function. The communication service provider (CSP) determines which GW-U is selected by the GW-C. It will either be dynamically selected, based upon load, or can be statically configured, based upon various parameters such as capacity, location, or access point name (APN) that match the specific user-device session. This gives the CSP the deployment flexibility to select, from the available pool of GW-U resources, one that meets the requirements of that service, which could be either a virtual or a physical function (Figure 2).
Figure 2 Illustrating the pool of GW-U resources available to the GW-C.

With 3GPP Release 15, the new 5G core reference architecture is defined (TS 23.501). In this release (Figure 3), a new session management function (SMF) is defined, which combines the session management of the mobility management entity (MME) and the GW-C in the Release 14 CUPS architecture, plus additional features to support Ethernet PDU sessions and determine the session service continuity mode and other policy-related functions.

Likewise, in Release 15, the user plane function (UPF) is the anchor point for mobility and external session connectivity to the data network. It also handles the packet forwarding and QoS (quality of service), packet inspection, policy rules enforcement, traffic usage reporting, and lawful intercept.

Key differences between 5G and 4G
Just as in the CUPS gateway architecture, the 5G core design specifies that the SMF can establish sessions with multiple UPFs depending upon the service. It can also select or reselect a UPF based upon the dynamic load or a static configuration chosen by the CSP. That configuration could be based upon a number of different criteria, including: UPF capacity supporting a specific data network name (DNN), UPF location, user equipment (UE) location, access or session type.

One of the key differences between 5G and 4G is session and service continuity (SSC), which defines three different modes for how a UE packet data unit (PDU) session is treated and whether the UE
preserves its IP address for that PDU session. In Mode 1, the network preserves service connectivity and the UE IP address is also preserved for that PDU session. In Mode 2, the network may release the service connectivity as well as the UE IP address for that PDU session. In Mode 3, the network ensures that there is no loss of service connectivity, but the UE IP address and the UPF anchor point changes for that PDU session.

SSC Mode 3 in 5G is a useful mechanism lacking in 4G. 4G networks follow a “break before make” procedure in which a service is often dropped when moving from one access type to another (e.g., video session between cellular and Wi-Fi). This is because the PDU session for that service was terminated (the “break”) on the mobile gateway that anchored it before a new PDU session and the anchor point on a different mobile gateway is established (the “make”).

SSC Mode 3 is a “make before break” process. It establishes (makes) a new PDU session and anchor point for the service before it terminates (breaks) the previous session, thus providing service continuity.

Whereas in the past, traditional networks tended to be purpose-built for one kind of traffic (voice, data, mobile), 5G represents the ultimate convergence of our previous network technologies and applications into a single, multi-purpose, multi-access network. The only way it can handle such diversity is by being completely flexible, scalable, and programmable.

To accomplish this it must adopt fully a cloud-native architecture and move beyond the simple virtualization of existing platforms. Innovation in the mobile core will play a significant role in creating a global interconnected fabric that can actually support the full range of exciting applications and services that 5G makes possible.

David E. Nowoswiat is a Senior Product and Solutions Marketing Manager with over 25 years of telecom industry experience in both wireless and wireline technologies. He is currently in Nokia’s IP and Optical Business Group supporting cloud and 5G next generation packet core Marketing.

Related articles:
- Virtualization and the cloud—A tutorial
- Building 5G networks from the core up
- 5G reaches milestone with New Radio specification
- The 5 best 5G use case