Functional Architecture

Andrea Detti

1 Introduction

Before 5G, the design of cellular network architecture was centered on the end users. The end users were the first-class citizens of the network and, generation by generation, cellular network designers strove to offer them an always-increasing capacity, thereby transforming the architecture into a full-IP one, as in the case of 4G.

The network architecture was made up of *nodes*, such as eNB, SGW, P-GW, or MME for 4G. Each node was usually composed of proprietary hardware and software, coupled in a single device.

5G is a sort of breakthrough in this thirty-year design pattern. Besides expected improvements in speed and time (20 Gb/s and 1 ms delay over the air), 5G architecture introduced a new category of first-class citizens: the verticals. 5G provides communication services not only for end users, but also for different vertical markets, such as automotive, energy, city management, government, healthcare, manufacturing, and intelligent transport systems.

Such heterogeneity creates demand for a level of service agility typical of a software environment, rather than an "ossified" hardware one. For this reason, 5G architecture has been designed to allow (and foster) a possible *softwarization* of network functions. Consequently, software defined networking (SDN), network function virtualization (NFV) and cloud computing are fundamental technologies for making full use of the power of a 5G network.

Andrea Detti
CNIT - Electronic Eng. Dept., University of Rome Tor Vergata, e-mail: `andrea.detti@uniroma2.it`

1.1 5G network architecture and services

A 5G network is composed of a 5G access network (AN) and a 5G core network (5GC) [1] (fig. 1). The access network itself is made up of a new-generation radio access network (NG-RAN) [3], which uses the 5G new radio interface (NR) [4], and/or a non-3GPP AN (e.g. WiFi, xDSL, etc.) connecting to a 5G core network. The different network entities are connected by an underlying TCP/IP transport network, which supports diff-serv QoS.

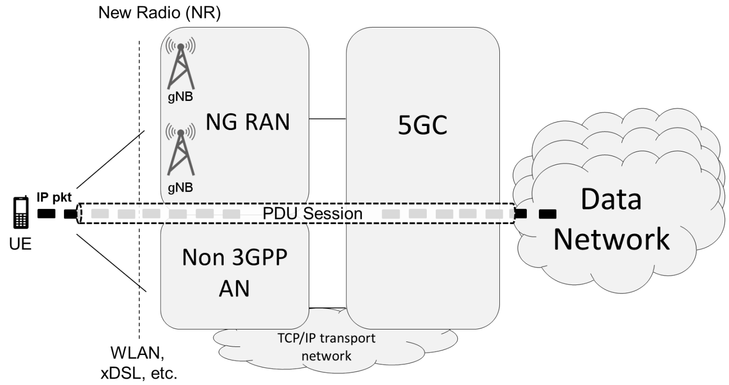


Fig. 1 5G Network Achitecture

Like previous generations, a 5G network connects user equipment (UE) to external data networks. The 5G connectivity service is named *PDU Session*. From a transport point of view, a PDU session is made by a sequence of NG tunnels in 5GC, and of one or more radio bearers on the radio interface. This set of "pipes" eventually connects the UE to its control functions and to the external data network for user traffic exchange (fig. 2). A major task of the mobile network is to establish and release the tunnels and the bearers dynamically, so as to follow user movements and states (idle, connected, etc.).

A PDU session is very similar to an EPS bearer in LTE, except for the QoS model and the supported user data units. Indeed, a PDU session can transport not only user plane IP packets, but also Ethernet or unstructured frames, thus allowing layer-2 communication among groups of UE. The 5G QoS model is based on the new concept of QoS flow [1], where a flow is the finest granularity of QoS differentiation. Different QoS flows may belong to a single PDU Session[1].

Fig. 3 shows the splits between the 5G functions executed in the NG-RAN and in the 5G core. In broad terms, the NG-RAN takes care of establishing, maintaining and releasing the parts of the PDU sessions that cross the radio interface. It copes

[1] We note that in the case of LTE, the finest QoS granularity is the EPS bearer. Different QoS services require different EPS bearers

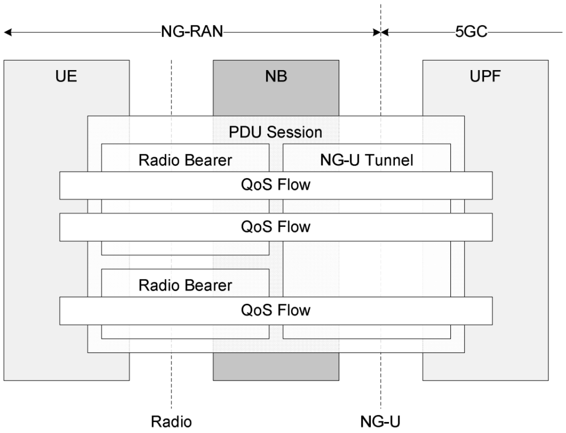


Fig. 2 PDU Sessions and QoS Flows: User Plane (source [4])

with physical impairments (e.g., fading, interference, power reduction); inter-gNB handovers; and session multiplexing (scheduling). The 5GC functions manage the remaining parts of the PDU sessions and take care of all the other processes not related to radio access (e.g., mobility management, security, IP address allocation, etc.).

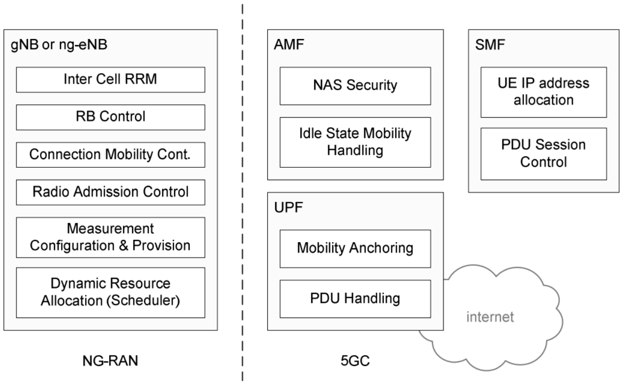


Fig. 3 Functional Split Between NG-RAN and 5GC (source [4])

1.1.1 New Generation Radio Access Network (NG-RAN)

As shown in fig. 4, the NG-RAN consists of a set of 5G base stations, called gNBs, which are connected to the 5GC through a set of logical interfaces. As in LTE, gNBs can be interconnected through the Xn interface to improve mobility (e.g., handover) and management functions (e.g., inter-cell interference coordination).

The functionality of a gNB is sometimes distributed. In that case, the resulting architecture is formed by a central unit (gNB-CU) that controls one or more distributed units (gNB-DU) through the F1 interface. A distributed unit is connected to a remote radio head (RRH), i.e., the actual radio transceiver. The central unit is again split in two parts, one for control plane functions (gNB-CU-CP) and one for user plane functions (gNB-CU-UP), following the control and user plane separation (CUPS) / SDN approach already introduced in the latest LTE releases.

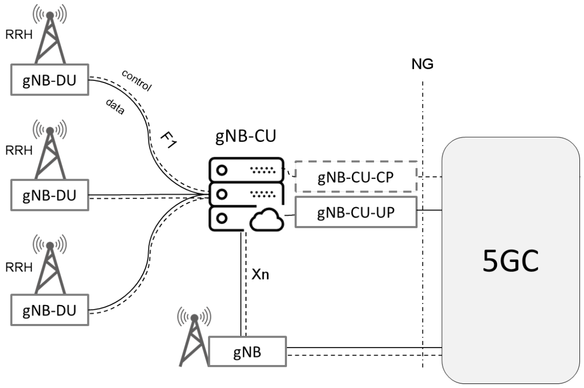


Fig. 4 Overall NG-RAN Architecture

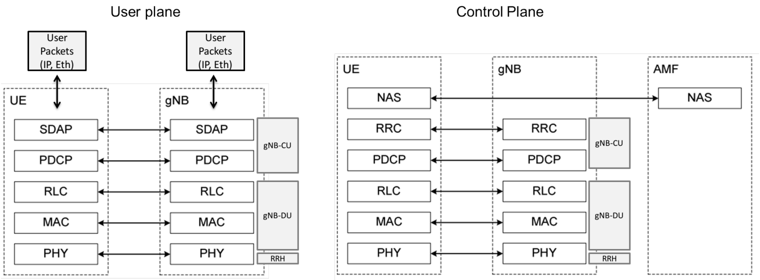


Fig. 5 NG-RAN Protocol Stack (original source [4])

Fig. 5 shows the stack of the protocols crossing the radio interface and their placement on the aforementioned gNB units. The stack is almost the same as the LTE one, except for the service data adaptation protocol (SDAP) of the user plane. The main functionalities of the different layers are as follows:

- The physical layer (PHY) contains the digital and analogue signal processing functions that the mobile and base station use to send and receive information. It is based on OFDMA, with adaptive carrier spacing (15,30,60,120,240 kHz) and an adaptive modulation/coding scheme (e.g., from $\pi/2$ BPSK to 256 QAM)[2].

- The medium access control (MAC) protocol provides low-level control of the physical layer, primarily by scheduling data transmissions between the mobile and the gNB.
- The radio link control (RLC) protocol ensures reliable delivery of data streams that need to arrive intact (HARQ). It also handles segmentation.
- The packet data convergence protocol (PDCP) carries out higher-level transport functions related to header compression and security.
- The service data adaptation protocol (SDAP) maps the interaction between the packet of a QoS flow and a data radio bearer (due to the new QoS framework) by marking the user data packets properly.
- The radio resource control (RRC) is the signaling protocol used in "access stratum" procedures involving the mobile and the gNB. It includes connection establishment and release functions; the broadcast of system information; radio bearer establishment, reconfiguration and release; RRC connection mobility procedures; paging; and power control.
- The non-access stratum (NAS) protocol is the signaling protocol used between the UE and the 5GC for PDU session management, security, mobility management, etc. The 5GC entity that takes care of controlling the UE is the access and mobility management function (AMF), which is similar to the LTE MME.

1.1.2 5G Core Network (5GC)

To some extent, the NG-RAN architecture, as well as its protocol stack, is similar to the LTE one. However, the architecture of the 5G core network is unique in many ways.

The decomposition of the functions executed by the network nodes of the previous generations led to a 5G architecture completely defined in terms of network functions (NF) that are exposed as services. Accordingly, as we can see in fig. 6, every block name ends with the letter "F": function.

As occurs in the NG-RAN, we have a control and user plane separation. In the user plane, we have one or more user plane functions (UPFs), which mainly carry out packet forwarding between the different NG-U tunnels (fig. 2) that form the PDU session. All other network functions belong to the control plane.

Another radical change from the previous generations is the interface modeling, which has moved from "bit-oriented point-to-point" to "web-oriented service-based." Indeed, 5G core is said to have a *service-based architecture*; wherever applicable, procedures (i.e., the sets of interactions between network functions) are defined as services, so that it is possible to reuse them.

There is a standardized point-to-point interface (real or logical) between any pair of interacting 2G, 3G and 4G network entities, and this interface uses a specific bit-oriented protocol. In the 5GC, the interactions among control plane entities use

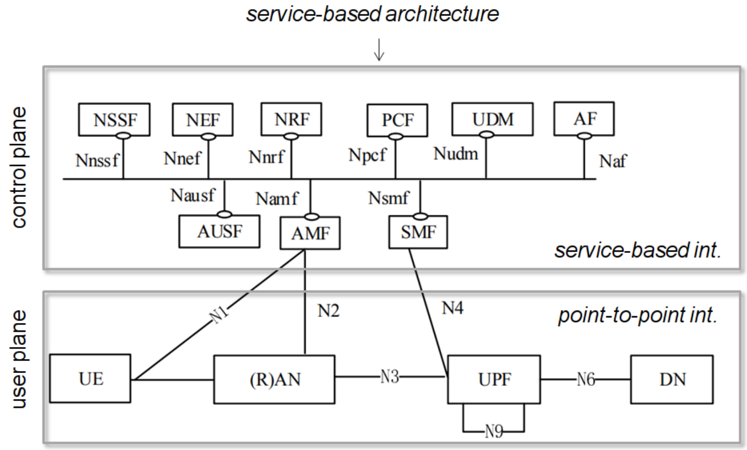


Fig. 6 5G System Architecture, Non-roaming (original source [1])

service-based interfaces, supported by web-oriented tools such as HTTP/2, REST and JSON.

Where are the differences? Whereas a point-to-point interface connects two well-defined entities (e.g., LTE S6a is strictly between MME and HSS), a service-based interface exposed by an entity is actually an API that any other entity could use: it is a one-for-all product.

Service-based modeling strongly improves the agility of the network in evolving or adapting itself to unforeseen needs. In the point-to-point interface model, if the system designers wish to add a new network entity and connect it to a set of N old network entities, they need to standardize N new interfaces and create related protocols. This complexity often leads to an ossification of the network. With the service-based interface model, the designers just have to standardize the API of the new network entity. Similarly, let us assume that there is a chain of network functions, namely NFa-NFb. Let us now assume that the designers wish to introduce another network function, NFc, in the middle of the chain, creating the sequence NFa-NFc-NFb. With the point-to-point model, they would need to standardize two new interfaces: NFa-NFc and NFc-NFb. By using the service-based model, they only need to standardize the API of NFc, and that is only if NFc is a new entity. If NFc is already standardized, they just need to reconfigure the sequence of functions.

The upper part of fig. 6 shows the set of network functions that form the 5G control plane. All of them expose service-based interfaces. For this reason, they are depicted as being connected by a network bus rather than by point-to-point links. The interface name is equal to the function name, with an "N" used as prefix. In this arrangement, one NF queries a network repository function (NRF) to discover and enable communication with other NFs. The insertion of a new network function, including a third-party one, is merely the insertion of a record in the NRF database. Under suitable security controls (i.e., authentication and authorization), a subset of service-based interfaces can be easily exposed to external users, such as third-party application providers, to enable them to optimize their services.

In the lower part of fig. 6, we have the set of network entities belonging to the user plane. As we can see, we still have point to point-interfaces there, identified by an "N" plus a number.

Let us conclude this section by describing the main roles of 5G NFs and their relation to 4G.

- The user plane function (UPF) handles the NG-U tunnel forwarding and the related data path services, such as anchoring for handover, QoS, and traffic policy enforcement. There can be multiple UPFs associated with a UE; these UPFs can be located in a single slice or in multiple ones. The UPF contains parts of the 4G SGW and PGW functionalities.
- The session management function (SMF) is the control part of a PDU session. That is, it configures NG tunnels, allocates IP addresses with DHCP, and configures traffic steering (e.g., towards a third party or an edge cloud). There can be multiple SMFs associated with a UE, though only one per slice. The SMF contains parts of the 4G MME and PGW functionalities.
- The access and mobility management function (AMF) handles all the 5GC signaling coming from and going to the UE. Unlike the SMF, it is a single function that is present in multiple slices. It supports user access to the network and manages mobility by interacting with the UE and with other NFs (e.g., SMF, AUSF, etc.). The AMF contains part of the 4G MME functionality.
- The authentication server function (AUSF) supports authentication for 3GPP and non-3GPP access. It contains part of the 4G HSS functionality.
- The unified data management (UDM) function can be considered a repository for UE-related information, such as credentials, identifiers, AMF details, and SMF assignments for the current session. The underlying idea of the UDM is to create, wherever possible, a central database for UE configuration information, so that the NFs can be designed as *stateless* services, improving architectural agility. The UDM contains part of the 4G HSS functionality.
- The policy control function (PCF) is a unified entity providing policy rules (QoS, filtering, charging, etc.) to other control plane functions, such as SMF. The PCF contains part of the 4G PCRF functionality.
- The network slice selection function (NSSF) selects the set of network slice instances serving the UE, along with the best AMF for that purpose. It is not present in 4G.
- The network exposure function (NEF) exposes the capabilities of networks and network/UE events for third-party, application function, edge computing, and other purposes. It is not present in 4G.
- The network repository function (NRF) discovers network function instances. When it receives an NF discovery request from a NF instance, it provides the discovered NF instances. It is not present in 4G.
- The application function (AF) resembles an application server that can interact with the other control-plane NFs. AFs can exist for different application services, and can be owned by the network operator or by trusted third parties. For instance,

the AF of an over-the-top application provider can influence routing, steering its traffic towards its external edge servers.

1.1.3 Network Slicing in the 5G Architecture

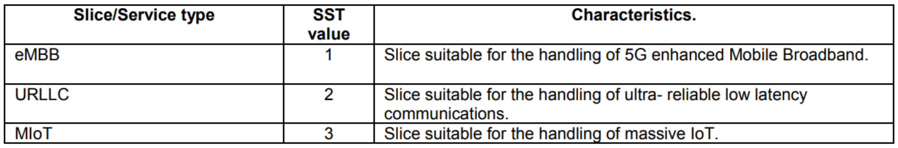
The 5G core architecture is made of network functions. This structure enables its immediate deployment with software and cloud tools. Indeed, 5GC is a *cloud-native* architecture.

Network functions can be implemented as pieces of software embedded in light virtual machines (e.g., Docker or Unikernel) and executed using a cloud infrastructure whose servers are spread all over the 5G network and are interconnected by an agile SDN. This allows for the easy reconfiguration of virtual network connectivity among "virtual" NFs. By using such a cloud-based deployment, there is a complete decoupling of the NFs from both the execution hardware and the interconnecting network infrastructure.

Cloud-based deployment of the 5G network also makes it possible for a tenant to create an isolated ICT environment, formed by specific instances of control and user plane NFs, supported by a dedicated 5GC virtual network and customized radio bearers. Such an isolated environment is actually a *5G slice*, i.e., a network-as-a-service offered to the different verticals.

As shown in fig. 8, a network operator can deploy multiple network slices with different features, or with the same features but for different groups of UE. For example, a slice for its customers can be equal to another slice for the customers of a virtual operator.

Each slice has a unique identifier, which includes the slice/service type (SST), referring to the expected behavior of the slice in terms of features and services. Currently, there are three standardized SST values (fig. 7). These are used to support the roaming use cases for the most commonly used slice/service types more efficiently.



Slice/Service type	SST value	Characteristics.
eMBB	1	Slice suitable for the handling of 5G enhanced Mobile Broadband.
URLLC	2	Slice suitable for the handling of ultra- reliable low latency communications.
MIoT	3	Slice suitable for the handling of massive IoT.

Fig. 7 Standardised Slice/Service Types (source [1])

As reported in fig. 9, NFs in different slices can be used in different configurations, or be placed further from or closer to the UE, depending on the vertical application using the slice. For instance, an eMBB slice could use a high-capacity radio bearer and have two UPFs, one in the edge and one in the cloud, to better support user mobility (two anchor points). A vehicular slice could have a radio bearer with low delay and medium capacity and many control functions moved to the edge to further reduce latency. An IoT slice could have a low-bit-rate radio bearer, a single UPF assuming low mobility, and most control NFs in the core, provided that latency is not important. Although they are not included in the figure, some slices can share the same instances of NFs, and there are some NFs, such as NSSF, that are common to all slices.

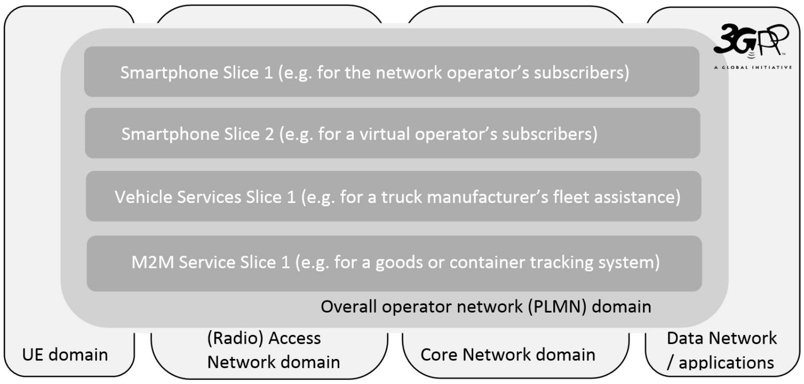

Smartphone Slice 1 (e.g. for the network operator's subscribers)
Smartphone Slice 2 (e.g. for a virtual operator's subscribers)
Vehicle Services Slice 1 (e.g. for a truck manufacturer's fleet assistance)
M2M Service Slice 1 (e.g. for a goods or container tracking system)
Overall operator network (PLMN) domain
UE domain
(Radio) Access Network domain
Core Network domain
Data Network / applications
3GPP
A GLOBAL INITIATIVE

Fig. 8 5G Slicing Example (source [5])

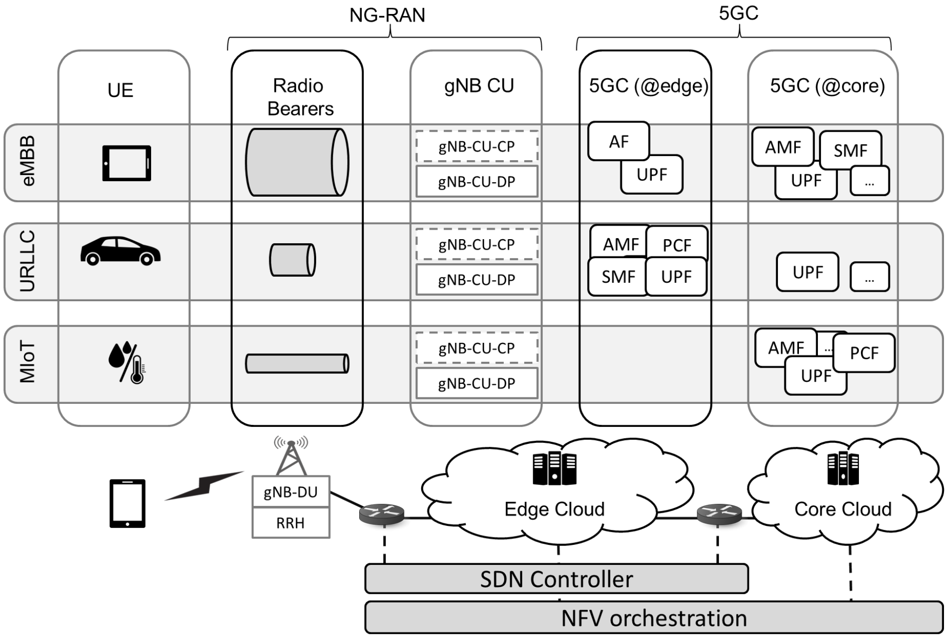


Fig. 9 5G NFs Deployment in Slices

References

1. 3GPP TS 23.501: "System Architecture for the 5G System; Stage 2", version 15.3.0 Release 15
2. 3GPP TS 38.221: " NR; Physical channels and modulation ", version 15.3.0 Release 15
3. 3GPP TS 38.401: "NG-RAN; Architecture description",version 15.3.0 Release 15
4. 3GPP TS 38.300: "NR; Overall description; Stage-2", version 15.3.1 Release 15
5. Frank Mademann, "System architecture milestone of 5G Phase 1 is achieved", 3GPP news, available at http://www.3gpp.org/NEWS-EVENTS/3GPP-NEWS/1930-SYS_ARCHITECTURE