MOBILE DIVISION

STUDY PAPER ON

5G CORE NETWORK

TELECOMMUNICATION ENGINEERING CENTRE
KHURSHID LAL BHAWAN, JANPATH
NEW DELHI - 110001
INDIA

Disclaimer: This study paper is intended purely for academic purpose and does not represent official view of DoT. The information present in the study paper is based on facts which were current at the time of writing of the paper.

CONTENT

1 BACKGROUND

In the field of Cellular Mobile communication, a new generation of technological advancement has come up almost every decade. Starting from first generation, it has completed the journey up to fourth generation i.e. 4G (LTE) crossing the decade gap with development of 2G(GSM) and 3G(UMTS) in between. At present, the base specifications for 5G have been frozen and further improvements are underway.

Broadly, cellular mobile communication system can be divided into two parts namely, Radio Access Network (RAN)/Radio Access and the Core Network (CN). The core network is the core part of the cellular communication system, which is behind the scene from the users' perspective unlike the radio part but is instrumental in providing services to the users. Some of the basic and crucial functions that the core network offers are authentication and authorisation and to maintain the location of the users so that services can be delivered to them. Though overall implementation of such functions have remained the same, the ways in which these functions have been realised in the core network have differed from generation to generation. Further, the core functions have to be delivered within certain timeframes as specified in the technology and these timelines have become more stringent over the technology generations.

In GSM, the architecture relied on circuit-switching (CS). This means that circuits are established between the calling and called parties throughout the telecommunication network (radio, core network of the mobile operator, fixed network). In GSM, mainly voice services are transported over circuit-switched telephony network. However, some data was also possible with very low data rate.

The first step towards an Internet Protocol (IP) based packet switched solution was taken with the evolution of GSM (Circuit switched solution) to General Packet Radio Service (GPRS), using the same air interface and access method. As, in GPRS, packet-switching (PS) is added to the circuit-switching, data is transported in packets without the establishment of dedicated circuits. This offers more flexibility and efficiency. Therefore, the core network was composed of two domains: circuit switched and packet switched. In UMTS (3G), this dual-domain concept remained on the core network side.

When designing the evolution for the 3G system to move towards 4G (4^{th} Generation) the 3GPP community decided to use IP (Internet Protocol) as the key protocol to transport all services. Accordingly, in 4G core network, called the Enhanced Packet Core (EPC), data is transported in packets only. Due to this the network architecture and the way that the services are provided has significantly changed over previous generations. For example, implementations for voice and short messages were required to be changed to IP-based solutions.

Though 4G technology provided high speed data with peak data rate of 1 Gbps, SMS services and other telecom facilities at reasonably good scale with latency around 10ms, however, with the passage of time, the expectations and requirements for the services to be provided by the telecom network have changed drastically. With the requirements that the telecom network be able to support vehicle to vehicle communication, augmented reality/virtual reality, remote health care etc, the existing 4G network were not in a position to fulfil the connectivity requirements for these services. Latency needs of around 1-2 ms for V2V communication, optimisation for connecting the millions of IOT devices to handle such kind of traffic generated by them, use of AR/VR in disaster recovery, remote monitoring etc. which needs support both for high data rate and low latency were not possible in 4G systems. Therefore, need was felt to design a next generation network, which ensures the fulfilment of data availability with even higher rates compared to 4G network and the availability of some more advance features, which are not available with the 4G network. Accordingly, 5G (5th Generation) network have been developed, which supports AR, VR, real time operations with support for smart farming, smart health, smart transportation and smart cities which were not available in the contemporary network facilities. While previous generations of mobile networks were purpose built for delivering communications services such as voice and messaging between persons (e.g. 2G) or mobile broadband (e.g. 4G), 5G will have flexibility and configurability at the heart of its design to enable mobile operators to serve IoT (Internet of Things) use cases, ultra-reliable, low latency connections as well as use cases pertaining to enhanced mobile broadband. These use cases have been grouped into sets namely mMTC (massive Machine Type Communication), URLLC (Ultra Reliable Low Latency Communication), eMBB (enhanced Mobile Broadband) respectively in 5G. Particularly, new use cases designed to support smart cities, smart agriculture, logistics and public safety agencies will have deep impact on the betterment of every aspect of our lives. 5G technology is expected to fundamentally transform the role that telecommunications technology plays in society. 5G is also expected to enable further economic growth and pervasive digitalisation of a hyper connected society, where not only are all people connected to the network whenever needed, but also many other devices/things are connected, creating the society where everything is connected (i.e. Internet of Everything). Introduction of 5G technology in China, EU, Japan, Korea, USA and India in their respective markets are being accelerated.

2 INTRODUCTION

The 5G is designed to meet the requirements of IMT-2020 set by the ITU-R specification M.2083 with the aim of providing more advanced and enhanced capabilities compared to 4G LTE (IMT-Advanced), as shown in Figure 1 below.

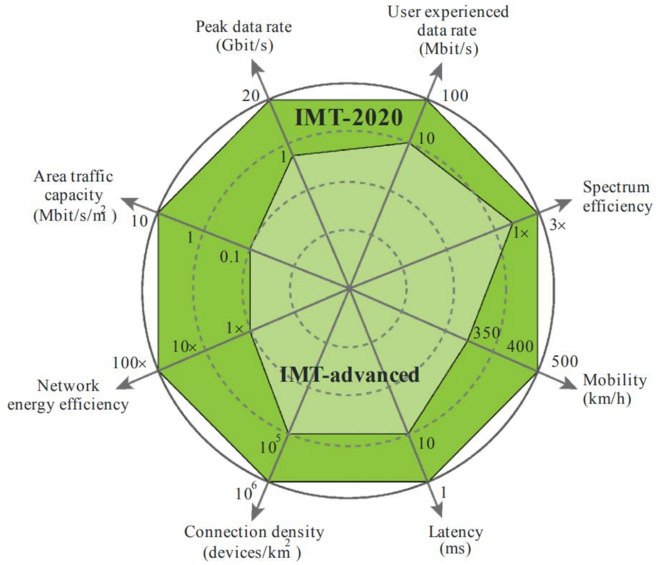

FIGURE 1 ENHANCEMENT OF KEY CAPABILITIES FROM IMT-ADVANCED (4G) TO IMT-2020 (5G)
Source: ITU R Recommendation M.2083-0

The technical performance requirements for IMT-2020 have been specified in Report ITU-R M.2410-0 "Minimum requirements related to technical performance for IMT-2020 radio interface(s)" as under.

TABLE 1

Parameter	**Value**
Downlink peak data rate	20 Gbit/s
Uplink peak data rate	10 Gbit/s
Downlink peak spectral efficiency	30 bit/s/Hz
Uplink peak spectral efficiency	15 bit/s/Hz
Downlink user experienced data rate	100 Mbit/s
Uplink user experienced data rate	50 Mbit/s
Area traffic capacity (Indoor hotspot eMBB only)	10 Mbit/s/m^2
Minimum requirements for user plane latency	4 ms for eMBB, 1 ms for URLLC
Minimum requirement for control plane latency	20 ms
Connection density (mMTC)	1 000 000 devices per km^2

Energy Efficiency	High sleep ratio and long sleep duration
Reliability	$1\text{-}10^{-5}$
Mobility	Stationary up to 500Kmph

To support these capabilities as given in the Table I, the 5G architecture features a new 5G Core Network (5GC) also known as the 5G Next Generation Core (NG-Core or NGC) which will be the heart of the network and act as an enabler for the myriad 5G applications and services. In the subsequent sections, the 5G core design principles have been outlined followed by 5G Architecture. The 5G core concepts like network slicing, service based architecture etc. that allow the 5G core to be agile, scalable and flexible has also been enumerated.

3 5G CORE ATTRIBUTES

As 5G system architecture has to support use cases like AR, VR, real time operations, and use cases like smart farming, smart health, smart transportation and smart cities etc. which all have varied requirements from the network, several concepts have been introduced into design of the 5G Core. Further, as scalability and protection of investment in equipment have gained in importance, softwarisation of the network has also increased. Therefore the 5G core network is aimed to be:

a) Flexible: 5G Core should be able to flexibly add new services and configure the network on a shared infrastructure.

b) Agile: The 5G Core should shorten the service implementation time from hours to minutes and delivers faster time-to-market (TTM) for new services from months to days.

c) Scalable: 5G Core should be scalable rapidly and highly with telco-grade reliability.

d) Tunable: 5G Core should be able to quickly adapt and optimise the network according to operating conditions. It should also support cost-efficient migration from 4G to 5G based on access agnostic common core.

To achieve these aims, the key 5GC design principles are:

a. Network slicing: This feature enables independent scalability and decoupled technical evolution and flexible deployments & configuration of the network as per the needs of the different services.

b. Modular function design: This is a form of functional disaggregation such that a function composed of multiple modules can be created according to the use case's requirements.

c. Unified authentication framework : This is useful in multi-access core, for efficiency and to enable operators to offer "follow-the-user" services, independently of access method.

d. “Stateless" network functions: With this feature, the "compute" resource is decoupled from the "storage" resource. This concept is derived from cloud applications. It enables much more efficient creation and consumption of network resources.

e. Support for cloud native applications: This is further step from NFV (Network Function Virtualisation) and improves the scalability and efficient creation and consumption of network resources.

f. Network capability exposure: Exposing information about the network's capabilities to internal and external applications is very important where operators want to integrate 5G with vertical industry processes. Standardizing this interface makes integration easier for vertical customers, especially those with international operations and multi-operator relationships.

g. Support for mobile edge computing: This is to support access to low- latency services hosted in local data centres. Typically, user-plane functions might be deployed remotely i.e. near to the user, while the control plane is centralized. In very low-latency, mission- critical applications, the control plane may also be distributed.

4 5G CORE NETWORK ARCHITECTURE

The 5G core network architecture standardised by 3GPP, enables support for increased throughput demand, reduced latency and increased reliability as per requirements of various applications and services that 5G must support. The new 5G core, as defined by 3GPP, utilizes cloud-aligned, service-based architecture (SBA) that spans across all 5G functions and interactions including authentication, security, session management and aggregation of traffic from end devices. The 5G core further emphasizes NFV as an integral design concept with virtualized software functions capable of being deployed in the network. Figure 2 shows the bus based 5G core network architecture. The 5G core network architecture is specified in 3GPP Technical Specification 23.501.

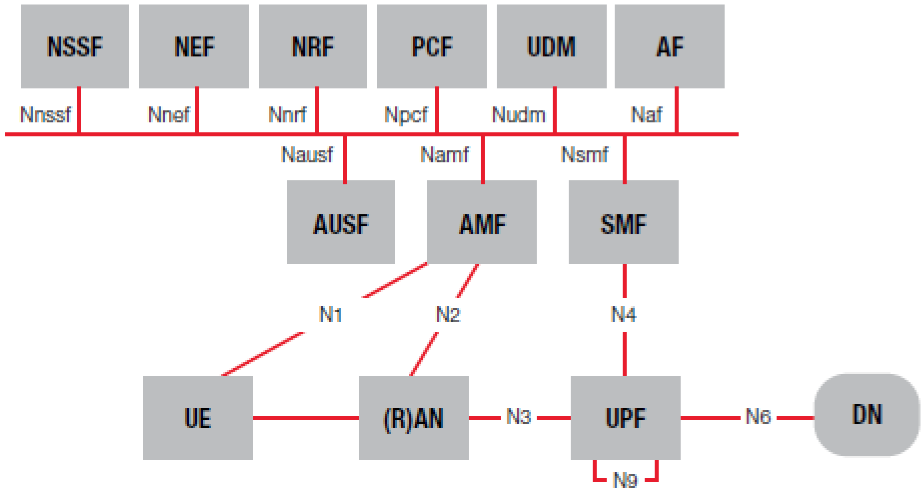

FIGURE 2 5G CORE NETWORK ARCHITECTURE (NON ROAMING) WITH REFERENCE POINT REPRESENTATION
Source: 3GPP Technical Specification 23.501

The primary Network Functions (NFs) and their capabilities as they are defined in the standards process today are as below:

i. Authentication Server Function (AUSF):

This acts as an authentication server. It contains mainly the EAP authentication server functionality and acts as storage for keys and provides keying material to the requester NF.

ii. Access and Mobility Management Function (AMF)

It carries out termination of NAS signalling, NAS ciphering & integrity protection, registration management, connection management, mobility management, access authentication and authorization, security context management. The AMF also includes the Network Slice Selection Function (NSSF) and acts as the termination point for RAN CP interfaces (N2).

iii. Session Management Function (SMF)

It carries out session management (session establishment, modification and release), UE IP address allocation & management, DHCP functions, termination of NAS signalling related to session management, DL data notification and traffic steering configuration for UPF for proper traffic routing.

iv. User Plane Function (UPF)

It carries out packet routing & forwarding, packet inspection, QoS handling, acts as external PDU session point of interconnect to Data Network (DN), and is an anchor point for intra- & inter-RAT mobility.

v. Network Exposure Function (NEF)

It supports exposure of capabilities and events, secure provision of information from external application to 3GPP network and translation of internal/external information. It acts as an API gateway that allows external users, such as enterprises or partner operators, the ability to monitor, provision and enforce application policy, for users inside the operator network. Thus, it

a. Provides security when services or Application Functions (AF) access 5G Core nodes

b. Acts as a proxy, or API aggregation point, or translator into the Core Network

vi. NF Repository Function (NRF)

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. It maintains/supports

a. Profiles of Network Function (NF) instances and their supported services within the network

b. Service-Based Interfaces, Management & Maintenance

In the SBA, the NF Repository Function (NRF) provides service discovery between individual network functions. It maintains profiles of network function instances and their supported services (for example, function ID, function type, network slice identifiers, capacity information, supported services, and endpoint information such as IP addresses). In this sense, it is an important "pivot" in the SBA.

The NRF plays an important role in the establishment of a new session. Here, the SMF discovery and selection request is initiated by the AMF when a request to establish a data session is received from the UE. The NRF is used to assist the discovery and selection of the correct SMF. In a network slice context, the same process occurs: the AMF queries the NRF to select an SMF that is part of a Network Slice instance based on S-NSSAI, UE subscription profile and operator policy, when the UE requests a session to be set up.

Control-plane functions communicate with one another, via the NRF, over service-based interfaces (using HTTP 2.0 transport). These are self-contained software modules that are reusable independently of each other and can be thought of as micro services. The network function is either a producer or consumer of services (top part of Figure 2), with two modes of interaction: either a consumer NF can request a response from a producer NF – for example, to request subscriber policy information; or it can subscribe to a producer and be notified if needed – for example, if a subscriber's state changes to inactive mode.

vii. Policy Control Function (PCF)

It carries out unified policy framework, providing policy rules to CP functions, access subscription information for policy decisions in UDR. This provides a policy framework incorporating network slicing, roaming and mobility management. It has similarities with the existing Policy and Charging Rules Function - PCRF of 4G.

viii. Unified Data Management (UDM)

It stores subscriber data and profiles and carries out generation of Authentication and Key Agreement (AKA) credentials, user identification handling, access authorization, subscription management.

ix. Application Functions (AF)

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. For services considered to be trusted by the operator, the AF can access Network Functions directly whereas untrusted or third-party AFs would access the Network Functions through the NEF.

x. Data Network (DN)

This refers to the external data network through which operator services, 3rd party services etc. can be accessed.

This new representation of the 5G Core Architecture in terms of network functions (for example, Access Management Function (AMF)) within the control plane enables other authorized network functions to access their services. Communications with these nodes leverage HTTP based APIs, replacing protocols like Diameter. This represents a major shift in the telecom world with the adoption of stateless functions, interfaces based on APIs and signals a move towards the software world. The overall design and their potential benefits include:

- **a.** Flexible and extensible architecture
- **b.** Easier integration with third party software using Application Programming Interfaces (APIs)
- **c.** Multi-slice User Equipment (UE), single UE simultaneously connecting to multiple services over multiple slices with optimized access and mobility signalling.
- **d.** Improved QoS

5 DESIGN PRINCIPLES OF 5G CORE

The 5G core network has been designed from scratch to include SBA or Service Based Architecture, network slicing, mobile edge computing etc. that bring flexibility, easy integration with third party software, easy provision of services belonging to different verticals, improved QOS etc. These design principles are discussed in detail below.

5.1 SERVICE BASED ARCITECTURE

"Service-Based Architecture" (SBA) is a modular architecture which is centred on services that can register themselves to provide certain services and also can subscribe to other services and using these as the building blocks, more complex services can be provided. 3GPP has adopted a service-based architecture of the 5G core system. This is significantly different from 4G, where such modularity in the network elements that constitute the core network was not included. SBA helps in the adoption of virtualisation, provides increased flexibility and adaptability to the core network. The architecture shown in Figure 3 is composed of network functions (NFs) like NSSF, NEF, SMF etc. and reference points that connects NFs. All these interfaces are candidates to be REST (Representational State Transfer) APIs and the architecture reference points can be simply replaced by a "message bus" in a logical diagram, with each NF communicating using HTTP2 protocol.

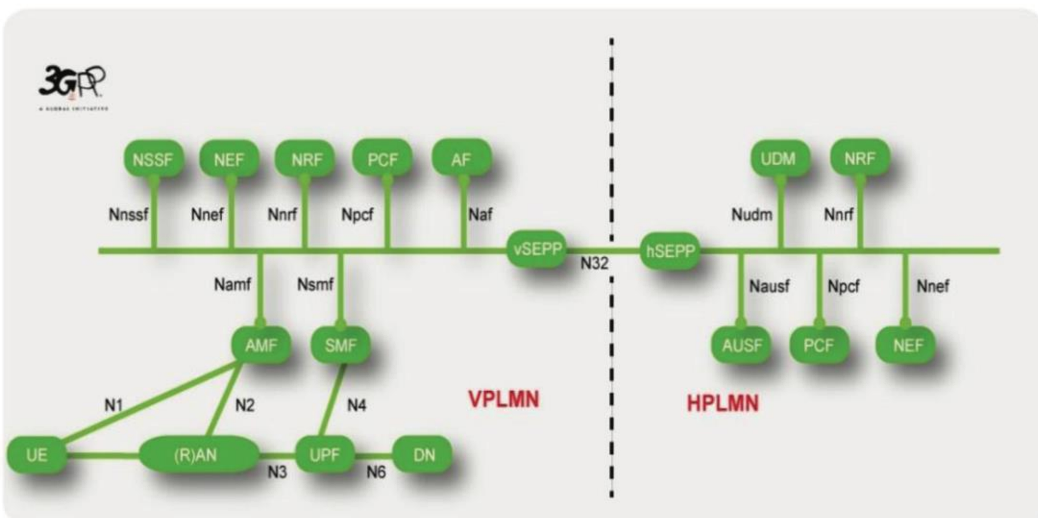

FIGURE 3 SERVICE BASED ARCHITECTURE (SBA)
Source: 3GPP Technical Specification 23.501

To make the applications more flexible, scalable and dynamic, the applications have been deconstructed into smaller and smaller components which can be reused for other applications and can also developed, tested and deployed independently. These units are called micro services. Micro services are an architectural design pattern where the system is composed of small granularity, highly cohesive and loosely-coupled services. Each of these services fulfils a specific functionality and is self-contained. Interactions between services implement standard light-weight interfaces (e.g. Restful principles etc.). This concept from the software world has been adopted into the telecom domain where each network element, which carries out certain functions has been broken down into the constituent parts and each network function can register, discover and use the services offered by other functions. The network functions and the interfaces are designed to have minimum impact on others so that introduction of new network functions, upgradation of existing network functions etc. does not impact the whole network. This makes possible continuous updation, upgradation, and introduction of new services. This is the Service Based Architecture approach.

The 5GC control plane (CP) design is based on services exposed by network functions (NFs) using new service-based interfaces (SBIs). Once a 5GC function registers its services with the new 5G Core Network Functions Repository Function (NRF), it then simply exposes services that any authorized consumer can consume, rather than having to define a new point-to-point interface and the procedures between the two network functions as an EPC requires. This offers operators greater flexibility and more efficiency by decoupling the service consumer from the service producer. The Service-based architecture brings following benefits to 5G:

a. Easy update of Network:
   - Finer granularity allows individual services to be upgraded with minimal impact to other services.
   - Facilitated continuous integration reduces the time-to-market for installing bug fixes and rolling out new network features and operator applications.
b. Extensibility:
   - Light-weighted service-based interfaces are needed to communicate across services.
c. Modularity & Reusability:
   - The network is composed of modularized services reflecting the network capabilities and provides support to key 5G features such as network slicing.
   - A service can be easily invoked by other services (with appropriate authorization), enabling each service to be reused as much as possible.
d. Openness

Together with some management and control functions (i.e. authentication, authorization, accounting), the information about a 5G network can be easily exposed to external users such as 3rd parties through a specific service without complicated protocol conversion

5.2 NETWORK SLICING

Slicing is the concept of creating logically separated customised networks consisting of network elements dedicated to that slice. Slices can be created for different purposes. A network slice is viewed as a logical end-to-end network that can be dynamically created. A given User Equipment (UE) may access to multiple slices over the same Access Network (e.g. over the same radio interface). Each slice may serve a particular service type with agreed upon Service-level Agreement (SLA). For example, to serve different traffic types: a slice designed for enhanced Mobile Broadband (eMBB) traffic is able to handle very high per- user throughput. Another slice, for massive IoT (mIoT), rather serves large number of subscribers that transmit small data infrequently but however generate significant signalling traffic due to idle to active state transitions. Slices can also be created to serve subscribers belonging to different enterprises, e.g. a slice dedicated to subscribers for each Mobile Virtual Network Operator (MVNO) hosted by the operator. Figure 4 shows network slicing as applied to 5G pictorially.

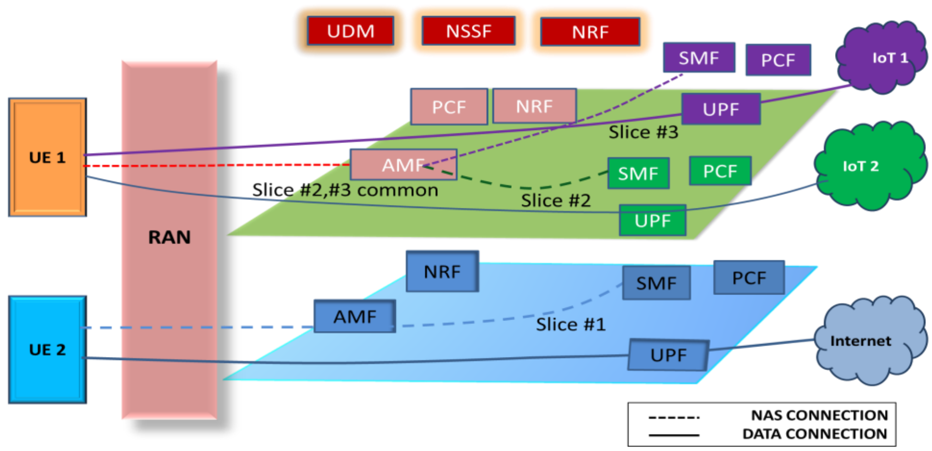

FIGURE 4 NETWORK SLICING

In the figure, UE1 accesses two different types of services represented by IOT1 and IOT2 both with different requirements from the network. Correspondingly, two network slices, slice#2 and slice #3 exist, which network resources have been configured to satisfy the requirements. Slice #2 and #3 share the AMF, while the session management is carried out separately. Further, the user plane function for both the slices is different. UE2 accesses the internet and a separate network slice #1 is configured appropriately to support the service.

Network Slicing as defined in 3GPP TS 23.501 is defined within a Public Land Mobile Network (PLMN) and includes the Core Network Control Plane and User Plane Network Functions as well as the 5G Access Network (AN). Though slicing as a term is new and used specifically with the advent of 5G networks, variants of this functionality have existed and evolved from EPS through to 5GS (5G System).

Network slicing was initially introduced to 4G in 3GPP Release 13. In EPS, 3GPP Release 13 added a feature to support Dedicated Core Networks (DCNs) called 'Decor.' The selection of the MME was based in part on UE's subscription, specifically a "UE Usage Type" parameter in the UE's subscription. In 3GPP Release 14, an enhancement (called Enhanced Decor, or eDecor) to DCNs further added the capability of UE to store the selected DCN ID and provide that to the RAN and core network during attach. This simplifies the task of selection of Core Network for the UE. However it also increases the dependency on UE support which is a disadvantage.

Most of the limitations introduced in the preceding paragraph are resolved by the slicing in the 5G System (5GS). All 5GS capable UEs and networks are required to support network slicing. In the user plane, each data connection of the UE is served by an SMF+UPF belonging to the same assigned slice. A UE can have data connections to different slices. However, there is a single AMF allocated to terminate the UE's NAS connection, which proxies session management messages to and from SMFs in the different slices.

3GPP TS 23.501 defines Network Function, Slice, and Slice Instance as follows:

a) Network Function: A 3GPP adopted or 3GPP defined processing function in a network, which has defined functional behavior and 3GPP defined interfaces. (Note: A network function can be implemented either as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, or as a virtualized function instantiated on an appropriate platform, e.g. on a cloud infrastructure.)
b) Network Slice: A logical network that provides specific network capabilities and network characteristics.
c) Network Slice instance: A set of Network Function instances and the required resources (e.g. compute, storage and networking resources) which form a deployed Network Slice.

The identification of a Network Slice is done via the Single Network Slice Selection Assistance Information (S-NSSAI). The NSSAI (Network Slice Selection Assistance Information) is a collection of up to eight (8) S-NSSAIs, meaning that a single UE may be served by at most eight Network Slices at a time. The S-NSSAI signaled by the UE to the network, assists the network in selecting a particular Network Slice instance. An S-NSSAI is comprised of:

a) A Slice/Service type (SST), which refers to the expected Network Slice behaviour in terms of features and services;
b) A Slice Differentiator (SD), which is an optional information that complements the Slice/Service type(s) to differentiate amongst multiple Network Slices of the same Slice/Service type.

The S-NSSAI may be associated with a PLMN (e.g., PLMN ID) and have network-specific values or have standard values. The SST values have been standardized in 3GPP TS 23.501 and reflect the most commonly used Slice/Service Types to assist with global interoperability for slicing.

The following are some highlights of network slicing feature supported in 5GS:

a. Policies to bind applications to slices and APNs can be provided to the UE during registration or can be configured on the UE. If provided during registration, AMF that receives the registration request from the UE retrieves the slices that are allowed by the user subscription and interacts with the Network Slice Selection Function (NSSF) to select the appropriate Network Slice instance (e.g., based on Allowed S-NSSAIs, PLMN ID, etc.). This could result in a change of AMF if needed. These policies can be subsequently updated at a later time, using NAS procedures. All 5GS UEs support these procedures. Such procedures do not exist for EPS or GPRS and rely on, eg. Open Mobile Alliance Device Management (OMA DM) procedures which are not supported by all UEs or networks.

b. In the network, operator policies for selection of network slices can be centralized in a network function called the Network Slice Selection Function (NSSF) or can be configured in each AMF. The centralization of network policies for slice selection in NSSF improves the operability of the network.

c. The discovery of network functions (eg. SMF, UPF, PCF) is performed using a function called Network Function (NF) repository function (NRF) – the detailed procedures are specified in 3GPP TS 23.502. NRF can be slice-specific or shared across slices. Having slice-specific NRFs enables isolation between slices, with network configuration of one slice not being visible in another slice.

d. In 5GS there is support for RAN-slicing, where the slice IDs of PDU session is provided to the RAN and the RAN can, via scheduling and radio resource management algorithms, share both uplink and downlink radio resources amongst the slices based on operator configuration.

Managing a complete Network Slice Instance (NSI) is not only managing all the functionalities but also the resource necessary to support certain set of communications service and these management and orchestration aspects are covered in 3GPP TS23.501.

5.3 CLOUD NATIVE ARCHITECTURE

In practice, there are three possible ways to implement NFs: either as a network element on a dedicated hardware, or as a software instance running on a dedicated hardware, or as a virtualized / cloud-native function instantiated on an appropriate platform, e.g., a cloud infrastructure. Leveraging cloud software technologies, 3GPP NF architecture leads to higher flexibility, programmability, automation and significant cost/energy reduction.

With the direction taken by 3GPP, it is expected that 5G core NFs become cloud- and virtualization-based applications. 5G core platform will be more programmable and allow many different functions to be built, configured, connected and deployed at needed scale.

5G promises support for various new use cases and services but for operators to offer these services and to get commercial benefit requires managing increased network complexity and the mobile traffic. One of the ways is to have a network in which new services can be easily deployed, managed and configured. This requirement is forcing companies to design "cloud-native" functions. “Cloud-Native” is the name of an approach to designing, building and running applications/virtual functions which lead to more agility and scalability in the system. This approach supports a very high degree of automation. Cloud-Native strategy is about reducing technical risk. In the past, our standard approach to avoiding danger was to move slowly and carefully. The Cloud-Native approach is about moving quickly but taking small, reversible and low-risk steps.

The Cloud Native Computing Foundation defines cloud-native as container-packaged, dynamically managed by a central orchestrator, and micro services-oriented. Container technologies, including Docker, Kubernetes, Mesosphere etc., are in fact ideal for cloud-native applications as containers offer a lightweight atomic unit of computing.

a. Observability to help collect, store and visualize logs, metrics, traces and other data points, is a prerequisite for seamless monitoring and operations
b. Security is vital to deal with storage and provisioning, as well as the handling of identities, certificates and keys.
c. To break down and implement business logic using stateless microservices, Cloud Native Applications typically need to rely on stateful backing services to store their data. A multitude of open-source projects aims to address these needs, including various database technologies.

While the goal of large-scale software reuse and utilization of open-source projects existed well before the emergence of the cloud-native paradigm, it is much more likely to be achieved now. This is because container technology isolates the different services from each other to a very high degree.

Applications built and deployed with the cloud-native pattern have some common characteristics mentioned below:

a. Based on microservices: Each application is a collection of small services that can be operated independently of one another. Microservices are often owned by individual development teams that operate on their own schedule to develop, deploy, scale and upgrade services.
b. Based on Containers: Cloud native applications are packaged in containers, aiming to provide isolation contexts for microservices. Containers are highly accessible, scalable, and easily portable from one environment to another and fast to create or teardown. This flexibility makes them ideal for building and running applications composed of microservices.
c. Based on continuous delivery model: Cloud-native applications are built and run on a continuous delivery model supporting fast cycles of build, test, deploy, release, and develop. This helps software service developers and infrastructure IT operations teams to collaborate with one another for building, testing and releasing software updates as soon as they are ready without affecting end-users or developers of other teams. This follows the DevOps principles continuous building, testing and releasing services happen. Cloud-native applications are dynamically managed, often built and run on modern platforms such as Kubernetes etc. Adoption of this principle in telecom sector can bring

unprecedented speed, agility and resilience in service development and management process. It greatly increases developer's productivity and simplifies operations. This allows new features to be pushed. New application features and services can be pushed live for customer use whenever they are ready with no need to worry about other teams work and zero impact on the end user experiences.

5.4 MOBILE EDGE COMPUTING (MEC)

Currently, use cases under URLLC (Ultra Reliable Low Latency Communication) involving image recognition, speech analysis, and large-scale use of sensors, specifically video security surveillance, self-driving cars, connected industrial robots, traffic flow and congestion prediction for smart city etc. are gaining in importance. In such use cases a processing delay or latency between the device and the remote servers used for online computations and content storage, typically situated far away from the end user and their smart device, can be very critical and will negatively affect the outcome. With Multi-access Edge Computing (MEC), the application server or content storage will be integrated into local cellular base stations, and will be brought closer to the end user and their smart device so that the latency constraints for critical use cases can be met.

MEC was first introduced in a 2015 white paper on the subject by the European Telecommunications Standards Institute (ETSI) according to which MEC "represents a key technology and architectural concept to enable the evolution to 5G". In addition to reducing the load on the core network and the backhaul, MEC plays a major part in reducing latency for 5G networks and achieving a reduced latency of the order of 1ms for the data plane for URLLC applications and services.

5G networks based on the 3GPP specifications are a key future environment for Multi-access Edge Computing (MEC) deployments, as both the 3GPP system - with its Service Based Architecture (SBA) - and the ETSI Industry Specification Group (ISG) MEC work similarly to leverage interactions between different network functions, aligning system operations with the network virtualization and Software Defined Networking paradigms.

3GPP deals with edge computing in a more direct way, in Technical Specification (TS) 23.501 (Clause 5.13) on the architecture for 5G Systems, where a set of new functional enablers are given for the integration of MEC in 5G networks. Figure 5 shows integrated MEC deployment in 5G network with MEC deployed on the N6 reference point, i.e. in a data network external to the 5G system. This is enabled by flexibility in locating the UPF. The User Plane Function (UPF) has a key role in an integrated MEC deployment in a 5G network. UPFs can be seen as a distributed and configurable data plane from the MEC system perspective. The control of that data plane, i.e. the traffic rules configuration, now follows the NEF-PCF-SMF route. The MEC orchestrator is a MEC system level functional entity that, acting as an AF, can interact with the Network Exposure Function (NEF), or in some scenarios directly with the target 5G NFs.

The key components are the ability of MEC, as a 5G Application Function, to interact with the 3GPP 5G system to influence the routing of the edge applications' traffic and the ability to receive notifications of relevant events, such as mobility events, for improved efficiency and end user experience. Some of the enablers as included in 3GPP specifications are:

a. User plane (re)selection: the 5G Core Network (re)selects UPF to route the user traffic to the local Data Network
b. Local Routing and Traffic Steering: the 5G Core Network selects the traffic to be routed to the applications in the local Data Network;
c. An Application Function may influence UPF (re)selection and traffic routing via PCF or NEF
d. Network capability exposure: 5G Core Network and Application Function to provide information to each other via NEF

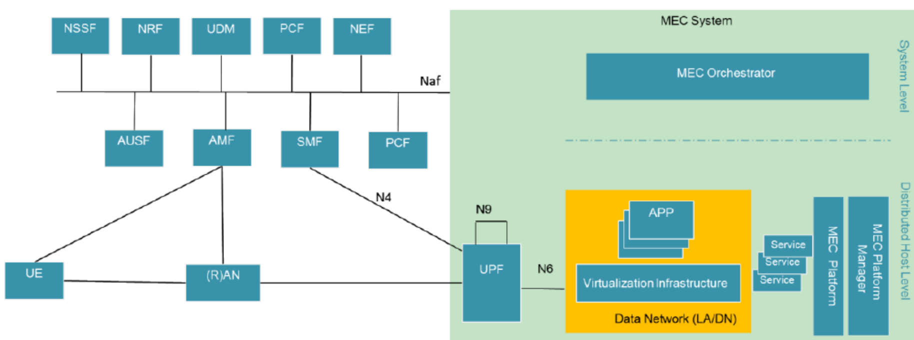

FIGURE 5 INTEGRATED MEC DEPLOYMENT IN 5G NETWORK
Source: ETSI MEC Whitepaper

Currently, 5G network with MEC capability is being provided by Verizon in collaboration with Amazon AWS in Chicago in USA which enables HD video, real time high quality gaming etc. Another example is by two operators- KT and SK Telecom in Korea where 5G network with MEC capability has been deployed for specific vehicle to vehicle (V2X) and Industrial IOT applications for certain companies.

These edge locations are suited to 5G because of the extreme low-latency requirements of some service types (such as URLLC applications) and to scale to meet the growing traffic demands. Operators have a physical footprint in the form of base station controller sites, and transport aggregation sites that they can convert into micro distributed data centres. Depending on the service requirements, these data centres can be used to terminate access connections from the 5G RAN and become the obvious place to deploy 5G core functions, especially user-plane functions, and to host latency-sensitive content and applications. Converged fixed and mobile operators have an even greater range of edge locations, such as central offices and local exchanges where they could host 5G network functions.

5.5 ENABLING SUPPORT FOR AI/ML

The 3GPP as part of Release 15 has developed a network function, the network data analytics function (NWDAF) that can enable machine learning and allow 5G operators to monitor the status of a network slice or third-party application performance and become the central point for analytics in the 5G core network. The 3GPP does not plan on standardizing the algorithms that will be used but rather the types of raw information the NWDAF will examine.

3GPP TR 23.791 has currently listed the following formula-based/AI-ML analytics use cases for 5G using NWDAF.

- Load-level computation and prediction for a network slice instance
- Service experience computation and prediction for an application/UE group
- Load analytics information and prediction for a specific NF
- Network load performance computation and future load prediction
- UE Expected behaviour prediction
- UE Abnormal behaviour/anomaly detection
- UE Mobility-related information and prediction
- UE Communication pattern prediction
- Congestion information – current and predicted for a specific location
- Quality of service (QoS) sustainability which involves reporting and predicting QoS change

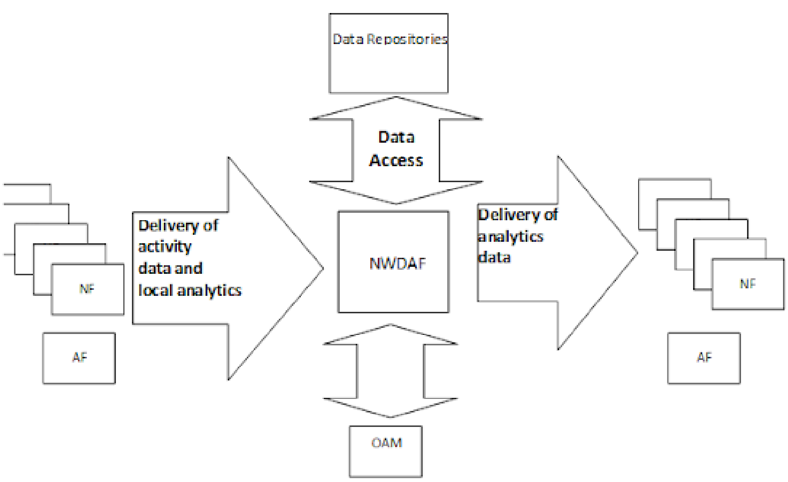

FIGURE 6 GENERAL FRAMEWORK FOR 5G NETWORK AUTOMATION (3GPP TS 23.503)

The NF, AF and the OAM can be the data sources for the NWDAF and communication happens using Service Based Interfaces as defined by 3GPP, shown in above figure. It also clearly provides correlation, services, the procedure to discover the right node for data needed to do Analytics. This inclusion of the NWDAF part of 5G Core, makes interworking complexities simplified for the Analytics solution vendors. In Release 17, studies are ongoing for network automation phase 2.

6 5G QOS IN CORE NETWORK

The QCI in 4G defined to be a scalar value that is used as a reference by the core network elements while setting up appropriate bearers and controlling the scheduling/forwarding of packets to support the service. It consisted of Resource Type (GBR or non-GBR), Priority Level (PL), Packet Delay Budget (PDB) and Packet Error Loss Rate (PELR).

In case of 5G, the concept of QCI is the same but its parameters have been extended so as to support new services that fall under URLLC/MCPTT use cases like Augmented reality, V2X, smart grid, automation etc.

The 5G equivalent of the QCI parameter as per 3GPP Technical Specification TS23.501is the 5G QoS Identifier (5QI), which includes –

- Resource Type (GBR or non-GBR)
- Priority Level (PL)
- Packet Delay Budget (PDB)
- Packet Error Loss Rate (PELR)
- Maximum Burst Size
- Data Rate Averaging Window

A new category of 5QI, called Delay-Critical GBR has been added that aids in the support of delay critical services. Of particular interest are the QCI values from 80 – 85, all of which have packet delay budgets of 30 milliseconds or less (remember that this delay budget is from UE to PCEF/P-GW). Also, the next generation services supported by QCIs 82 – 85 require the additional information provided by the maximum burst size and data rate averaging window for aiding the eNodeB scheduler to provide satisfactory service for URLLC/MCPTT use cases.

7 SECURITY CONSIDERATIONS

With virtualisation, softwarisation and cloudification as the basic design principles of 5G networks, security which earlier was based on the fact of a physically separate network operated and in control of solely the network operator, has to be expanded to include the fact that now the 5G core network elements may be implemented on shared infrastructure spanning multiple geographies. Keeping these issues in mind, security now includes- resilience, communication security, identity management, privacy and security assurance. These properties make the 5G system a trustworthy platform that enables many new services to be created. 3GPP has developed the 3GPP TS 33.501 "Security architecture and procedures for 5G system" as the foundation 5G security document in Release 16. 3GPP's 5G security architecture is designed to integrate 4G equivalent security and also includes the reassessment of other security threats such as attacks on radio interfaces, signalling plane, user plane, masquerading, privacy, replay, bidding down, man-in-the-middle and inter-operator security issues 5G leading to further security enhancements. The 5G system has secure identity management for identifying and authenticating subscribers through the 5G authentication and key agreement (5G AKA) protocol and the extensible authentication protocol (EAP) framework. A key feature of EAP is the flexible way in which different authentication protocols and credential types can be used without affecting intermediate nodes. Under security assurance, 3GPP Security specification is 3GPP TS 33.511 "Security Assurance Specification (SCAS) for the next generation Node B (gNodeB) network product class", which is part of Release 16. With the addition of new features with every release, the 3GPP security specifications also evolve correspondingly.

8 LATEST FEATURES ADDED BY 3GPP

While in the previous sections, the basic 5G core network architecture and its features have been discussed, these mainly pertain to 3GPP Release 15. With the on-going standardisation work in 3GPP, more features have been added to the 5G core network as given in 3GPP Release 16 which was finalised in June 2020 (Stage 3 freeze). Work on 3GPP Release 17 is under progress and is targeted for completion in Q2 2022. These features are discussed in this section.

8.1 3GPP RELEASE 16

3GPP Release 16 is a major release because it brings the IMT-2020 submission - for an initial full 3GPP 5G system - to its completion. The overall trend in 3GPP Release 16 is to make the 3GPP 5G System (5GS) a communication-enabling platform suitable for a wide range of industries ("verticals"), such as e.g. transportation (autonomous driving V2X, Railways, Maritime), automated factories, healthcare, public safety and many more, with enhancements to Ultra-Reliable Low Latency Communications (URLLC), Network Slicing, Edge Computing, Cellular IoT (Internet of Things), Non-Public Networks, Positioning Services and LAN-type services. In addition, the use of 5G as an underlying communication network (i.e. to be used transparently by applications external to the network) has been enhanced, mostly under the work on "Northbound APIs". The 3GPP Release 16 includes many features, out of which some of the important features are described below.

a) **The enhancements for Service Based Architecture (eSBA) in 5GC** enables indirect communication and delegated discovery through Service Communication Proxy. This allows flexible architecture and NF communication via Network Repository Function (NRF) and Service Communication Proxy (SCP). By defining NF Service Set and NF Set, the network functions and services defined can be deployed and selected modularly and provide redundancy for the NF services.

b) **The enhancement of Location Services (LCS)** uses 3GPP Release-15 baseline architecture and then adds support for roaming for location services. Some of the additional functions include: NEF enabled exposure of location (at GMLC and AMF/UDM (Cell Identity)), control plane positioning for non-3GPP accesses, Location privacy subscription, temporary blocking of positioning consent from the UE or from Application Function, low powered devices optimized location reporting using Cellular IOT (CIoT) operations for LTE and bulk location reporting operation.

c) **The network slicing function** is being further improved with interworking support with EPS to 5GS mobility. It introduces a new procedure for reallocation of AMF and V-SMF as part of topology enhancement work in Connected mode, and also by reallocation to a new AMF during Idle mode mobility. Another added feature known as Network Slice Specific Authentication and Authorization (NSSAA) enables separate authentication and authorization per Network Slice. The trigger of NSSAA is based on subscription information from UDM and operator policy and may be performed when UE indicates support for the feature.

d) **Enhancements to the Network Automation (eNA) architecture compared to 3GPP Release 15 adds** data collection and network analytics exposure features. Network Analytics ID allows specific analytics data collection based on the source and type of information to Network Data Analytics Function (NWDAF). It defines the output analytics information based on statistics and prediction from the data collected. Some examples of Analytics ID include: Slice Load Level information, Network Performance Information, UE Mobility Information, QoS sustainability, etcetera. A network may deploy multiple NWDAFs, so Analytics ID type support are available either via NRF or via local configuration.

e) **5G_CIoT introduces support for CIoT** (similar to EPS) in 5GS for NB-IoT and LTE-M devices with 5GS support. The following features are included: Support for infrequent small data transmission (DoNAS), Frequent small data communication (UP optimization), High latency communication, Power saving functions, Management of Enhanced Coverage, Support of common north-bound APIs for EPC-5GC interworking, Monitoring, Network parameter configuration API via NEF, Overload Control for small data, Inter-RAT mobility support to/from NB-IoT, Support for Expected UE behavior, QoS Support for NB-IoT, Core Network selection for Cellular IoT, Group communication and messaging, Support of the Reliable Data Service and MSISDN-less Mobile Originated SMS.

f) **Non-Public Network (NPN) support** that enables the 5G architecture to deliver two types of NPNs: Stand-Alone Non-Public Networks – SNPN or access to a non-public network

through a separate logical access network, and Public Network Integrated NPN – PNI NPN or access through an operators public access network. The architecture also provides support for service continuity and support access of NPN services via PLMN and vice versa. SNPN has certain restrictions such as no roaming between SNPNs, no interworking with EPS and UEs supporting SNPN need to be configured with specific data such as PLMN and Network ID (NID). PNI NPN operation may optionally make use of the concept known as Closed Access Group (CAG) which enables the control of UEs' access to PNI NPN on a per cell basis (CAG cells). A UE may be configured with CAG information on a per PLMN basis.
Further enhancements to support NPNs are studied for subsequent releases including enabling support for UE onboarding and provisioning for an NPN, enabling support for SNPN along with subscription/credentials owned by an entity separate from the SNPN.

g) **Time Sensitive Networking architecture** enables 5GS to provide time synchronization of packet delivery in each hop, to support Time Sensitive Networking (TSN) and industrial control via 5GS. For 3GPP Release 16, the 5G System is integrated with an external network providing TSN services as a TSN bridge. Currently, the release is supporting a centralized TSN model and with specific subset of IEEE specifications with support enabled for the integration.

h) **Architecture enhancements were done in the 5G system** for redundant transmission support to enable better reliability for services that require URLLC communications support. Three different variants of solutions are available in 3GPP Release 16. Operators can choose any or multiple mechanisms depending on the deployment and applications requiring URLLC services.
These variants are: dual connectivity based end to end redundant PDU sessions for the service associated with URLLC profile; redundant user planes between NG-RAN and UPF (redundant N3/N9 interfaces) for the same PDU session where only that link is considered need to be redundant; underlying transport network redundancy where UPF transmits packets utilizing two different redundant transport link and NG-RAN eliminates redundant packets and vice versa.

i) **Northbound APIs related items**: Further enhancements and changes to the 3GPP Northbound APIs (i.e. SCEF (Service Capability Exposure Function) Northbound APIs, NEF Northbound (Network Exposure Function) APIs, CAPIF (Common API Framework) APIs and xMB API) are necessary and the enhancements specified are NEF/SCEF Northbound APIs registration with CAPIF Core Function in order to enable the discovery of the Northbound APIs by 3rd party Application Servers, handling of communication failure case etc. Enhancements for Common API Framework for 3GPP Northbound APIs have also been included in this Release.

8.2 3GPP RELEASE 17

Release-17 core-network-standard features is built on Release 15 and Release 16 specifications. It provides further system enhancements and additional features that were not included in the previous releases. Current schedule for Release 17 complete date is targeted at Q2 2022. Some of the features under study are:

a) **Architectural enhancements for 5G multicast-broadcast services (5MBS):** This feature is important for multiple use cases in 5G such as CIoT, V2X, mission critical services, and others. In addition, with 5G service-based architecture, there is an opportunity to define a modular and simpler architecture.

b) **System enhancement for Proximity based Services in 5GS:** proximity services (also referred to as Device to Device (D2D) direct communications) was also defined in LTE and is an important feature for public safety, V2X, and other commercial services for 5G services. This study will define a framework in coordination with the corresponding RAN work for supporting such features in 5G. There are two sets of objectives for this framework: one is to support public safety use cases, and the other is to support commercial use cases.

Other features include enhancement of network slicing phase II, network automation for 5G Phase II, enhancement of support for Edge computing in 5GC etc.

9 GLOBAL DEPLOYMENT

As per data available by January 2021 from Ookla, 21863 5G networks are commercially available while 191 are in the process of rolling out of network, trialling etc. The first country to adopt 5G on a large scale was South Korea, in April 2019. Major South Korea telecom service providers like SK Telecom has 38,000 base stations, KT Corporation has 30,000 and LG U plus has 18,000; of which 85% are in six major cities. As per current information, 3.5 GHz (sub-6) spectrum in non-standalone (NSA) mode is being used and tested speeds were from 193 to 430 Mbit/s down.

All initial launches of 3GPP 5G services have been based on the non-standalone specifications (NSA), which uses 5G Access network with existing EPC (LTE Core) Network. Beyond mobile operator networks, 5G is also expected to be widely used for private networks with applications in industrial IoT, enterprise networking, and critical communications.

10 SUMMARY & CONCLUSION

The 5G core network is fundamental to the provisioning of required services for commercial success of 5G. In addition to the basic functions needed to set up and manage user sessions, it enables new and diverse services across many different vertical industries. The service-based architecture for 5GC so far developed in the 3GPP with further enhancements offers many attributes that make it suitable for progressive operators seeking to accelerate deployment of cloud-native 5G networks and services. Technologies like MEC and Network slicing are also crucial to the success of 5G and are being adopted in the 5G core network.

ACRONYNMS

3GPP: Third Generation Partnership Project
5G: Fifth Generation Radio system.
5QI: 5G QoS Identifier
CIoT: Cellular IOT
CP (also C-Plane): Control Plane
D2D: Device to Device
E2E: End to end
eMBB: Enhanced Mobile Broadband
EPC: Evolved Packet Core
ETSI: European Telecommunications Standards Institute
HTTP: Hyper Text Transfer Protocol
IoT: Internet of Things.
KPI: Key Performance Indicator
L2: Layer 2
LTE: Long Term Evolution
M2M: Machine to Machine
MAC: Medium Access Control
MBMS: Multimedia Broadcast Multicast Service
mMTC: massive Machine Type Communication
MNO: Mobile Network Operator
NB-IOT: Narrowband Internet of Things
NEF: Network Exposure Function
NF: Network Function
NFV: Network Function Virtualization
NGMN: Next Generation Mobile Network
NMO: Network Management and Orchestration
NMS: Network Management System
PCC: Policy and Charging Control
QOE: Quality of Experience
QoS: Quality of Service
RAN: Radio Access Network
RAT: Radio Access Technology
RF: Radio Frequency
RLC: Radio Link Control
RRC: Radio Resource Control
SCEF: Service Capability Exposure Function
SDN: Software Defined Network
SDO: Standards Development Organization

SMF: Session Management Function
S-NSSAI: Single – Network Slice Selection Assistance Information
TDM: Time Division Multiplexed
UP (also U-Plane): User data Plane
VNF: Virtual Network Function

REFERENCES

[1] 3GPP TS 23.060, "General Packet Radio Service (GPRS); Service description; Stage 2"
[2] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access".
[3] 3GPP TS 23.501, "System Architecture for the 5G System".
[4] 3GPP TS 23.214, "Architecture enhancements for control and user plane separation of EPC nodes".
[5] 5G Network Transformation, 2017 5G Americas
[6] 5G Core and EPC interworking variant, Heavy Reading
[7] ETSI Mobile Edge Computing White Paper
[8] 5G Core Vision, Samsung
[9] Road to 5G: Introduction and Migration, GSMA, April 2018
[10] 5G-Nework-Architecture, Huawei
[11] 5G-deployment-considerations, Ericson
[12] LTE and 5G in the Middle East and Africa –a snapshot, GSA
[13] View on 5G Architecture, 5GPPP
[14] Global Progress to 5G - Trials, Deployments and Launches, GSA
[15] 5G Deployment, State of Play in Europe, USA and Asia, EU, April 2019