Path to 5G: A Control Plane Perspective

Erik Guttman[1] and Irfan Ali[2]

[1]*Chairman of 3GPP TSG-SA, Samsung Electronics, Germany*
[2]*Senior 5G architect at Cisco Systems, USA*
E-mail: erik.guttman@samsung.com; irfaali@cisco.com

Received 31 March 2018;
Accepted 3 May 2018

Abstract

This paper provides an overview of some specific control plane functionality that has developed in the 3GPP architecture, from GPRS to EPC and now the 5G core network. Innovations of the 5G control plane are considered in the areas of selecting and maintaining the control plane topology, as well as the handling of state within the network.

Keywords: 3GPP, Telecommunications Core Networks, Control Signalling.

1 Introduction

This paper provides an overview of control plane functionality as it has developed in the 3GPP architecture. Successive generations broaden the set of services supported while maintaining compatibility with existing deployed telecommunication infrastructure and terminals. Every decade, a new set of standards are developed for the core network – 2.5G (which added packet data support to Global System for Mobile Telecommunications (GSM), developed by the European Technical Standards Institute (ETSI)) and more fully in 3G, 3GPP introduced the Generic Packet Radio Service (GPRS) [1]. A further evolution of this system occurred with the introduction of 4G: the Enhanced Packet System (EPS),

Journal of ICT, Vol. 6_1&2, 87–100. *River Publishers*
doi: 10.13052/jicts2245-800X.616

whose core network is called the Enhanced Packet Core (EPC) [2]. Now, the 5G architecture features a new 5G Core Network (5GC) [3]. Radio aspects and end-to-end interactions between terminals and services available in the network are not considered in this paper. Rather, the focus is the network that supports these functions and enables delivery of services. Specific innovations of the control plane in successive generations are introduced and briefly discussed.

2 What is the Control Plane

The purpose of the 3GPP system is to efficiently provide terminals, referred to as User Equipment (UE), with access to services (voice, text, data, etc.) available in data networks. The following figure shows that UE access to the Data Network involves two other distinct networking domains: the Access Network (e.g. Radio Access Network) and Core Network (GPRS, EPC or 5GC.)

Control plane aspects exist throughout the system. This paper concentrates on control plane aspects of the Core Network. A control plane exists also in the Access Network (which could be 3GPP radio access technology, non-3GPP access, etc.) as well as end to end, between the UE and the Service, though these aspects are not considered here.

The delivery of service, shown as the horizontal line in Figure 1, generally occurs via a data forwarding network or 'user plane.' The Core Network establishes and maintains this forwarding path, which requires the Core Network to support various capabilities. The mobile telecommunication system supports data forwarding even as the UE moves, transitions to and from the 'idle' state, intermittently becomes unreachable over the Access Network, and as services delivered to the UE change over time. The user plane is not a merely a packet data forwarding path: it supports many capabilities and constraints, for example monitoring, service level guarantees, charging and a wide range of network capabilities that require authorization.

The 'control plane' is the term used for all signalling used to support the functions in the mobile telecommunications system that establish and maintain

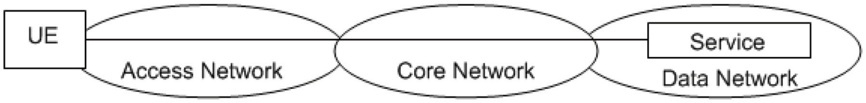

Figure 1 A simple model of service access using the 3GPP system.

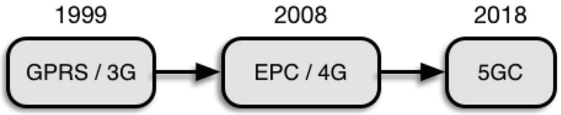

Figure 2 Core network evolution through generations.

the user plane. Signalling in this sense means exchange of information to enable but not to provide the end-to-end communication service itself. (In some cases, services are delivered in part by means of control plane mechanisms, e.g. SMS messages are delivered to the UE by means of control messages. This is not elaborated upon in this paper.) The control plane is itself a forwarding path to exchange information for operation of the service. As 'overhead' (it enables services but is not a service itself), the control plane must be efficient, scalable, reliable and suited to the needs of mobile network operators.

Once a mobile device can communicate using an access network, the UE can register with the network. Millions of these devices must be supported, even as they periodically cease communication or leave coverage, so that data and other services can be delivered at the first opportunity, both to the UE and from the UE. Within the Core Network, control plane interactions occur as needed, associated with each UE registered with the network. It is therefore imperative that the control plane interactions occur efficiently.

The Core Network supports several functions, most essentially access control, data packet routing and forwarding, mobility management, radio resource management and UE reachability functions. These functions are mentioned to illustrate the role of the control plane functions and are only elaborated upon further in the context of discussing areas in which the control plane has evolved over time.

Through successive generations, the Core Network has evolved and advanced with respect to how the above functions are supported.

The remainder of this paper considers specific capabilities and their development.

3 Control Plane and User Plane Separation

The following simplified representation of the architecture emphasizes the development of control plane/user plane separation. The entities shown include only a subset of those defined.

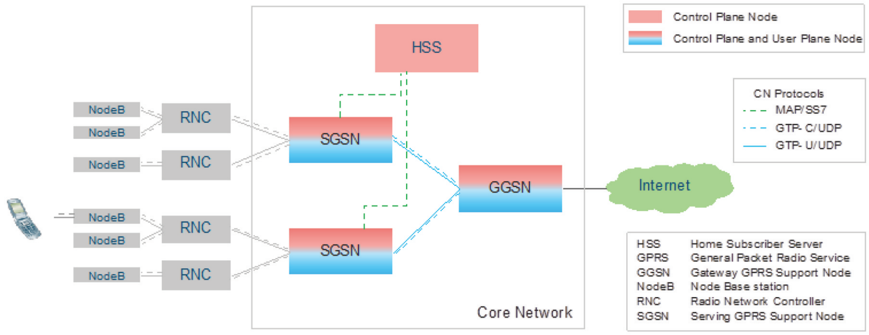
Control Plane Node
Control Plane and User Plane Node
HSS
NodeB
NodeB
NodeB
NodeB
NodeB
NodeB
RNC
RNC
RNC
RNC
SGSN
SGSN
GGSN
Internet
Core Network
CN Protocols
MAP/SS7
GTP-C/UDP
GTP-U/UDP
HSS Home Subscriber Server
GPRS General Packet Radio Service
GGSN Gateway GPRS Support Node
NodeB Node Base station
RNC Radio Network Controller
SGSN Serving GPRS Support Node

Figure 3 GPRS – the 3G core network.

In GPRS [1], the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN) terminate both user plane and control plane interfaces. The implementation and implicitly the deployment of these entities tightly couples the control and user planes.

In EPC [2], the mobility management (including authentication) functionality of the SGSN was separated out into the Mobility Management Entity (MME) and data-plane functionality of the SGSN separated into the Serving Gateway (SGW). This provides the opportunity to some extent to scale the control aspects in the MME independently of the session management and data forwarding aspects in the SGW and Packet Gateway (PGW). The GGSN functionality evolved into the PGW functionality. Also, shown in Figure 4 is the introduction of the PCRF to provide dynamic QoS and charging policies to the network. This was needed to support VoLTE and emergency IMS voice services. (Though Policy and charging architecture is also defined for GPRS, this has seldom been deployed).

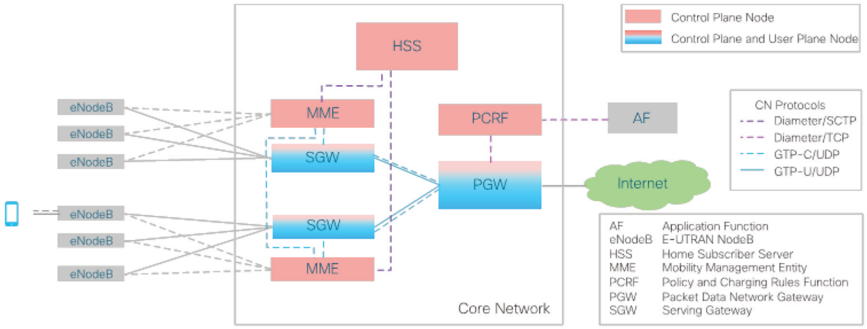

Figure 4 EPC – the 4G core network.

In Release 14, the architecture allowed a full separation of user plane and control plane [4], splitting the SGW and PGW into control and user plane aspects. This allows much more flexible, efficient and higher performance deployments of the user plane, e.g. to improve the placement, network control and resource management. Also, this enabled the centralization of the control functionality of the SGW and PGW as shown in Figure 5, where a single SGW-C controls both the SGW-U network elements.

The 5GC, as depicted in Figure 6, also separates the control plane and user plane. The Access and Mobility Management Function (AMF) provides mobility management functions, analogous to mobility management functions

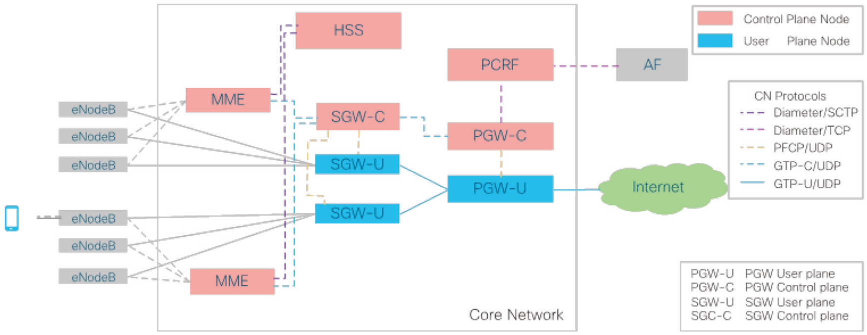

Figure 5 EPC with control plane user plane separation enhancement.

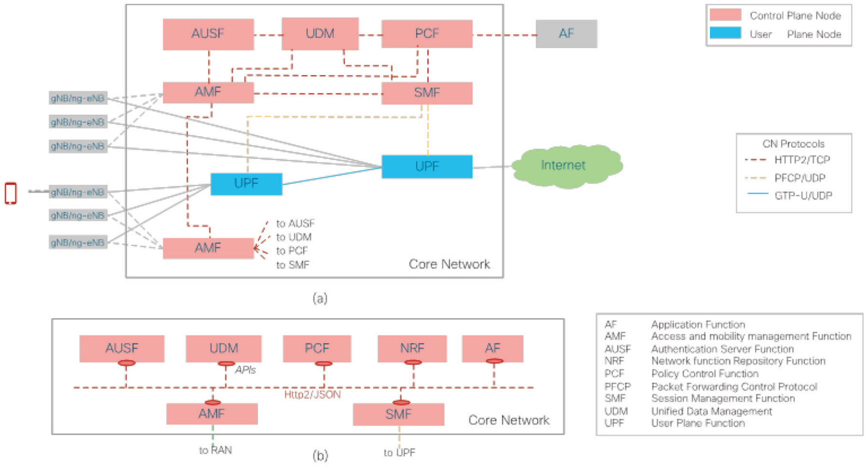

Figure 6 5G core network. (a) Interface representation, and (b) API level representation.

of the MME. The session management functions of the MME are separated out and combined with the data plane control functions of the SGW and GPW to create the Session Management Function (SMF). Thus the AMF, unlike the MME, does not include session management aspects. For example, in the 5GC, session management aspects of control messages from the UE are terminated by the SMF, whereas in the EPC, these would be terminated by the MME. One advantage of this mobility management and session management separation is that AMF can be adapted for non-3GPP access networks also. The session management aspects are very access specific and hence are specified initially for the Next Generation Radio Access Network (NG-RAN.)

Another important development in successive releases is a consolidation of the number of protocols used between functions in the control plane of the system. More importantly, in 5GC the protocol for interaction between all control-plane entities is HTTP, which is a protocol widely used in the Internet and not telecom-specific like dedicated Diameter applications or GTP-C.

4 Service Based Architecture

A key advance in the 5GC architecture is the introduction of the service based architecture. In GPRS and EPC control plane design, procedures defined all interactions between network functions as a series of message exchanges, carried out by protocol interactions. In the 5GC, network functions employing the Service Based Architecture offer and consume services of other network functions. Allowing any other network function to consume services offered by a network function enables direct interactions between network functions. In the past, several kinds of interactions piggybacked (or reused) messages exchanged along general purpose paths, since a direct interface does not exist between the consumer and producer network function. For example, the Policy Control Function (PCF) can directly subscribe to location change service offered by the AMF rather than having to have this event proxied via the SMF. In the EPC, by contrast, analogous information followed a hop by hop path from the MME, to the SGW, to the PGW and finally the Policy and Charging Rules Function (PCRF). This model also holds the promise of allowing services offered by network functions to be reused for other purposes than simply processing the control procedures defined for implementing the functions of the 5GC. There are other advantages at the protocol level, e.g. uniformity of network protocols leading to simpler implementations, use of modern transport and application protocol frameworks that are more extensible and efficient, etc., but this is not discussed further in this paper.

State management is an area where the 5GC has made significant advances. GPRS and EPC control entities defined state associated with a registered UE, called “context.” This information, both subscription information retrieved from the HSS, and dynamic information corresponding to the registered UE is stored in the SGSN and GGSN in the GPRS architecture and the MME, SGW and PGW in the EPC. As the UE moves, the SGSN (in GPRS) or MME and SGW (in EPC) may be relocated: new serving nodes may be selected. This procedure requires the ‘context’ to be transferred between the old and new entity, and additional state to be fetched, e.g. the subscription data to the new MME.

In the 5GC, state may be stored centrally. This can ease network function implementations in which state storage per network function and context transfer between network functions are not desirable. In Rel-15, procedures for AMF relocation specify context transfer procedures, as in 3G and 4G. In future, use of centralized storage may be defined to eliminate this requirement. Already in Rel-15, the centralized Unified Data Management (UDM) function is employed for some procedures for retrieval of state, for example, in the Registration with AMF-reallocation procedure. In this procedure, per slice subscriber data including access and mobility information is stored by the initial AMF and retrieved by the target AMF.

5 Slicing

Slicing is the concept of creating logically separated networks consisting of network elements dedicated to that slice. Slices can be created for different purposes. 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.

Slicing is a facility to support multiple instances of the same network function, associating each network function instance with a specific slice and then selecting a slice that serves a subscriber. The subscriber’s user and control plane is established and maintained by network functions of that slice. 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 GPRS through

EPS to 5GS. This section considers this evolution and highlights the key features introduced at each step of the evolution.

Before continuing, an important aspect to highlight is that in 3GPP networks, a UE has two types of connections with the core-network: a *signalling connection*, called Non-access Stratum (NAS), and one or more *data connections* – each associated with an IP address for transferring UE's IP traffic between the UE and a data network. This data connection is called a Packet Data Network (PDN) connection for GPRS and EPS and Packet Data Unit (PDU) session for 5GS. The NAS connection is between the UE and SGSN in GPRS, between the UE and MME in EPS and between the UE and AMF in 5GS. In 5GS the UE also communicates using NAS message with one or more SMFs (one for each PDU session). These messages are proxied via the single AMF that serves the UE.

In all 3GPP core networks, the selection of the node that terminates the UE's NAS connection occurs first, during the registration procedure. This is followed by the selection of the gateway for the UE's data connection (GGSN for GRPS, PGW for EPS and SMF+UPF for 5GS) during data connection setup. Both of these aspects are considered in this discussion of the evolution of slicing.

The evolution of the slicing concept is illustrated by the example of two subscribers, as shown in Figure 7. UE 1 communicates with servers in the Internet and UE 2 communicates with servers in two different IoT data networks, IoT-1 and IoT-2. 3GPP core networks enable this functionality by providing the UE with multiple IP addresses to a subscriber, with the subscriber using the data network specific IP address to access the servers in the appropriate data network. Access to these data networks requires the selection of gateways that serve the specific data networks and provide the UE with an address from that data network.

In GPRS networks, the selection of the SGSN during the registration procedure to terminate the UE's NAS traffic is not based on UE's subscription or data networks that the UE subsequently intends to connect. However, the selection of data gateway (GGSN) for UE's PDN connection is enabled by the use of Access Point Name (APN). APN is a string that the UE provides to the network during data connection setup, which identifies the data network that the UE wants to communicate with. Also, APNs may be part of subscription data or SGSN configuration. After applying a set of rules, an APN is identified for the UE's sessions. This process allows an operator to restrict the APNs that a subscriber is allowed access.

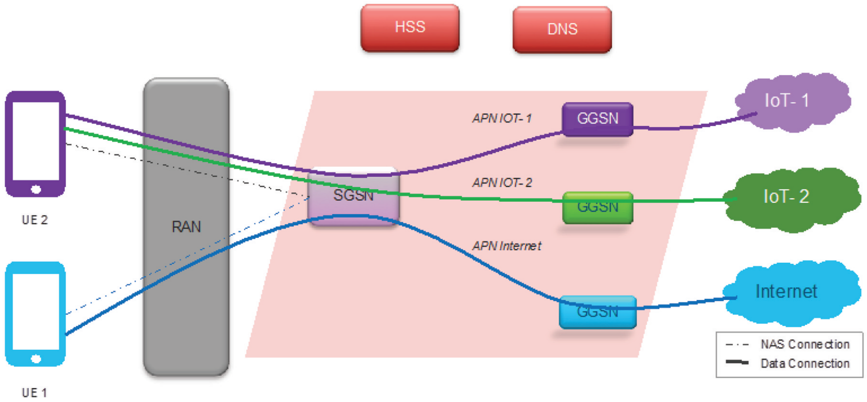

Figure 7 Use of APNs for selection of GGSN in GPRS.

APNs are used to obtain Domain Name Service (DNS) records of GGSNs' addresses. This enables the SGSN to select GGSN that serves a particular data network through DNS lookup during data connection setup. Hence, as shown in Figure 7, by using multiple APNs, the UEs' PDN connections are anchored at the GGSN that are gateways to the respective data networks. Note that both the UE 2's NAS connections are terminated by the same SGSN. In GPRS networks, the same DNS server is used for the lookup of GGSN for all the APNs.

For 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 simplified the task of selection of Core Network for the UE.

Figure 8 illustrates the application of DCNs to our example of the two UEs. In this example the two UEs are assumed to have different "UE Usage Types" and the network supports separate DCNs for the two UE Usage Type. For the UE's NAS connection, the UE 1 is assigned MME from DCN#1 and UE 2 is assigned MME from DNC#2. Note that this is not possible for GPRS networks (see Figure 7). In addition the SGWs for the two UEs are different in the two DCNs. The selection of the SGW and PGW is based on both the DCN-ID and APN of the PDN Connection. Similar to GPRS, there is a single DNS common to the two DCNs.

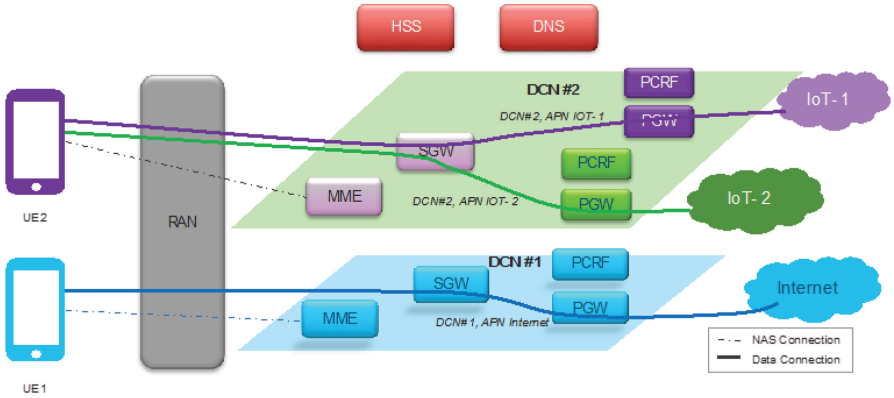

Figure 8 Dedicated core networks (DCNs) for EPS.

The introduction of Decor and eDecor was a major step forward towards slicing. The RAN is provided a DCN-ID but the UE and the RAN directs the UE's NAS connection towards the appropriate core network (MME). However, the network still needs to support pre-Release 13 UEs that do not provide DCN ID indication to the RAN. There is still the limitation that for a UE only a single SGW can be allocated for all UE's data connections. Additionally, the DNS is shared between all the DCNs in the operator's network. Missing too, were tools to configure the policies in the UE for use of DCNs, for example binding applications to specific DCNs and APNs. DCN was introduced as an add-on feature on an existing Core Network and had to work with the existing design of the network and UE.

Most of the limitations introduced in the preceding paragraph are resolved by slicing in the 5G System (5GS). This is depicted in Figure 9, which considers the same scenario as Figure 8. 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. Also, (not shown in Figure 9), UE can have multiple PDU sessions in a slice to different data networks, or multiple PDU sessions to the same data network via different slices, via the combination of slice identifier and APN.

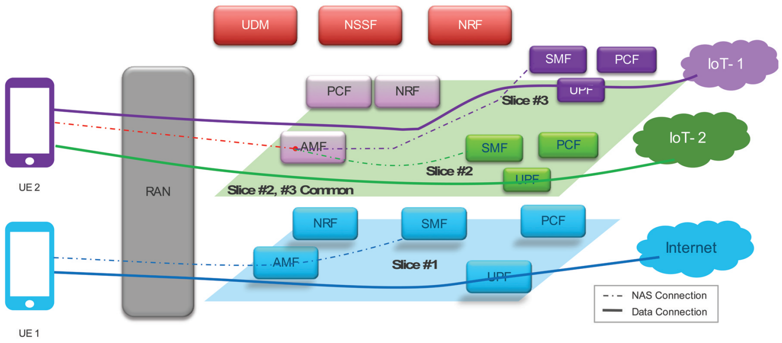

Figure 9 Network slicing applied to 5GS.

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

- Policies to bind applications to slices and APNs can be provided to the UE during registration or can be configured on the UE. 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.
- 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.
- The discovery of network functions (eg. SMF, UPF, PCF) is performed using a function called Network Function (NF) repository function (NRF). NRF can be slice-specific or shared across slices (both these options are depicted in Figure 9). Having slice-specific NRFs enables isolation between slices, with network configuration of one slice not being visible in another slice. This is not possible for EPS where the DNS is shared across DCNs.
- In 5GS there is support for RAN-slicing (not shown in Figure 9), 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.
- The 5G Core Network has been designed to take advantage of network orchestration mechanisms to instantiate, maintain and delete slices.

6 Summary

3GPP standards maintain backwards compatibility from release to release, even as the network architecture evolves. Each new core network generation evolves from the previous ones and at the same time introduces new features. This paper illustrated a few areas in which the control plane signalling architecture of the core network has advanced, e.g. by separating control and user plane and most recently, by introducing the notion of a Service Based Architecture and the support of network slicing. The evolution of the control plane is by no means over with the 5G core network in its first release.

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”.

Biographies

Erik Guttman, employed by Samsung Electronics, has been actively involved in networking and telecommunications standardization for over 20 years. He currently serves as the 3GPP Service and System Aspects Technical Specification Group Chairman. Preceding this, he held the position of 3GPP System Architecture working group for two terms. He has also chaired and actively contributed to numerous IETF working groups including

SVRLOC (Service Location Protocol) and ZEROCONF (Zero Configuration Networking). Erik's background includes leading research and product development projects that introduced emerging network application and system functions to operating environments. Erik developed frameworks and tools for distributed installation, testing and deployment. Erik served Chief Technical Officers as system architect and requirements researcher. Erik obtained a BA in Philosophy and Computer Science from the University of California, Berkeley and a MS in Computer Science from Stanford University.

Irfan Ali is an experienced engineer and researcher in telecommunications and networking. He is a technical expert on 5G, LTE, IoT and IMS systems through contributing to standards and as a systems engineer in leading cellular infrastructure companies. He has worked in the wireless industry for the past two decades in various roles for Motorola, Nokia and NTT Docomo. Currently, he is a senior 5G architect at Cisco Systems and represents Cisco in 3GPP for 5G standards. He has published several papers, a book and has been awarded more than twenty patents. He has also taught graduate level courses at Istanbul Technical University and Bosphorus University in Turkey. Irfan holds a Ph.D in Computer Engineering and a Masters in Computer & Electrical Engineering.