JOURNAL OF ICT STANDARDIZATION

Volume 6, No. 1–2 (January and May 2018)

JOURNAL OF ICT STANDARDIZATION

Chairperson: Ramjee Prasad, CTIF, Aarhus University, Denmark
Editor-in-Chief: Anand R. Prasad, NEC, Japan
Associate Editor: Sivabalan Arumugam, NEC, India
Advisors: Bilel Jamoussi, ITU, Switzerland
Jesper Jerlang, Dansk Standard, Denmark

Editorial Board
Signe Annette Bøgh, Dansk Standard, Denmark
Hermann Brand, ETSI, France
John Buford, JP Morgan Chase Bank, USA
Rajarathnam Chandramouli, Stevens Institute of Technology, USA
Sunghyun Choi, Seoul National University, South Korea
Debabrata Das, IIIT-Bangalore, India
Ashutosh Dutta, Johns Hopkins University Applied Physics Lab, Maryland, USA
Vijay K. Gurbani, Bell Labs, Nokia, USA
Kiritkumar Lathia, Independent ICT Consultant, UK
Julien Laganier, Cisco Meraki, USA
Yoichi Maeda, TTC Japan
Monique Morrow, Hycon, South Korea
Kohei Satoh, ARIB, Japan
Henk J. de Vries, Rotterdam School of Management,
Erasmus University, The Netherlands
Alf Zugenmaier, University of Applied Sciences Munich, Germany

Objectives

- Bring papers on new developments, innovations and standards to the readers
- Cover pre-development, including technologies with potential of becoming a standard, as well as developed/deployed standards
- Publish on-going work including work with potential of becoming a standard technology
- Publish papers giving explanation of standardization and innovation process and the link between standardization and innovation
- Publish tutorial type papers giving new comers a understanding of standardization and innovation

Aims

- The aim of this journal is to publish standardized as well as related work making "standards" accessible to a wide public – from practitioners to new comers.
- The journal aims at publishing in-depth as well as overview work including papers discussing standardization process and those helping new comers to understand how standards work.

Scope

- Bring up-to-date information regarding standardization in the field of Information and Communication Technology (ICT) covering all protocol layers and technologies in the field.

Published, sold and distributed by:
River Publishers
Alsbjergvej 10
9260 Gistrup
Denmark

River Publishers
Lange Geer 44
2611 PW Delft
The Netherlands

www.riverpublishers.com

Journal of ICT Standardization is published three times a year. Publication programme, 2018: Volume 6 (3 issues)

ISSN: 2245-800X (Print Version)
ISSN: 2246-0853 (Online Version)

All rights reserved © 2018 River Publishers

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

JOURNAL OF ICT STANDARDIZATION

Volume 6, No. 1–2 (January and May 2018)

Foreword v–viii

JOE BARRETT / Outlining the Roadmap to 5G 1–14

TOON NORP / 5G Requirements and Key Performance Indicators 15–30

BALAZS BERTENYI, SATOSHI NAGATA, HAVISH KOOROPATY, XUTAO ZHOU, WANSHI CHEN, YOUNSUN KIM, XIZENG DAI AND XIAODONG XU / 5G NR Radio Interface 31–58

BALAZS BERTENYI, RICHARD BURBIDGE, GINO MASINI, SASHA SIROTKIN AND YIN GAO / NG Radio Access Network (NG-RAN) 59–76

FRANK MADEMANN / The 5G System Architecture 77–86

ERIK GUTTMAN AND IRFAN ALI / Path to 5G: A Control Plane Perspective 87–100

GEORG MAYER / RESTful APIs for the 5G Service Based Architecture 101–116

FRÉDÉRIC GABIN, GILLES TENIOU, NIKOLAI LEUNG AND IMRE VARGA / 5G Multimedia Standardization 117–136

ANAND R. PRASAD, SIVABALAN ARUMUGAM, SHEEBA B AND ALF ZUGENMAIER / 3GPP 5G Security 137–158

THOMAS TOVINGER, JEAN-MICHEL CORNILY, MARYSE GARDELLA, CHEN SHAN, CHEN AI, ANATOLY ANDRIANOV, JOEY CHOU, JAN GROENENDIJK, ZHANG KAI, ZOU LAN, XIAOWEN SUN, WEIXING WANG, ZHU WEIHONG AND YIZHI YAO / Management, Orchestration and Charging in the New Era 159–178

Foreword

Every decade, telecommunications standards developers aspire to a new generation. Member companies in 3rd Generation Partnership Project (3GPP) identify and work towards an expansion of the telecommunication industry. Collectively, 3GPP[*] takes on the task, drawing on the aspirations and insights of individual experts, member companies, partner Standard Developing Organizations (SDOs) and external organizations. The formal focus of this activity is the ITU-R IMT program. With IMT-2020, IMT-Advanced and now IMT-2020, a vision and expression of the aspiration of the coming generation is established. The IMT goals are both commercial and more general; as telecommunications have become central to social and cultural development across the world. These programs establish criteria for the radio network which are met by qualifying submissions. 3GPP's 3rd generation was accepted for IMT-2000, and 4th generation for IMT-Advanced. 3GPP is hard at work now developing standards that will be submitted for the IMT-2020 program. The context and aspiration for 5G are well described in "Outlining the Roadmap to 5G" in this issue.

3GPP is a partnership project, bringing together SDOs from around the world. These SDOs formally transpose and release the standards specifications that 3GPP completes. This activity continually proceeds – maintaining existing standards through corrections and producing both new specifications and adding features to existing specifications to provide new features and functions. Different parts of the 3GPP organization work on the standards from different perspectives. Some working groups are further ahead, identifying requirements for new services and functions, others following by defining these features technically – maintaining compatibility with existing capabilities as necessary and not creating any incompatibilities, and still other working groups who complete the specifications at the protocol level. The partnership project therefore comprises committees at all phases of the work, simultaneously. Features are completed together as 'releases.' A release, beyond offering attractive new complete capabilities to the market, eases

[*] Information regarding 3GPP is available at http://www.3gpp.org/

coordination of standards development activities and makes comprehensive testing possible. In the end, a major aspect of what 3GPP standards' success lies in the emergence of equipment that is both compatible and compliant with the standard according to the specifications, initially when deployed and for years to come.

The first releases of 5G, leading to the IMT-2020 submission, are Release 15 (which is termed 5G Phase 1) and Release 16 (which is also called 5G Phase 2.) Release 16's functional freeze date – also termed 'stage 3 freeze' will occur at the time that the IMT-2020 submission will be sent. At the time of the publication of this special issue on 3GPP 5G Standardization, 3GPP will have frozen Release 15 and already have determined the focus for work on Release 16.

5G standardization is an extended program whose goals extend from the short to the long term. The article "5G Requirements and key performance indicators" provides the background and outcome of the extensive work 3GPP completed on normative standards and identifying the goals for 3GPP's 5G standardization program. 5G Phase 1 and Phase 2 description can also be found in this article.

The standards under development include the entire range of activity of 3GPP – across all TSGs and working groups. 3GPP will fully complete the standard sufficient to realize the IMT-2020 vision, though this will not occur in 5G Phase 1 and 2. The work will continue for several releases, extending the standard in compatible ways to satisfy ever more requirements. It is not yet fully clear what coming releases, past 5G Phase 2, will include. 3GPP continually re-examines and takes current market conditions and opportunities into account.

The articles in this special issue on "3GPP 5G Specifications" reflect the diverse progress and substance of the 5G standards. Some of these standards have already been completed and frozen – others are currently under development and will be finished over the coming months.

One area that greatly distinguishes each generation of 3GPP standards from the last is the innovation brought to the radio access network. "5G NR radio interface" provides an overview of this exciting new standard and how it relates to the IMT-2020 radio requirements. In addition, coexistence of NR and LTE is presented. The 5G Access Network (5G-AN) includes both LTE and NR radio access technologies. "NG Radio Access Network (NG-RAN)" takes a broader look at the 5G access system in which NR operates and in particular describes how the NG-RAN can be deployed in different scenarios to realize migration from LTE-based networks to 5G and NR. This paper also considers the overall 5G radio access architecture and its key interfaces and protocols.

Another area of significant development is the 5G Core Network (5GC) and overall system. “The 5G System Architecture” offers an introduction and explains some of its key characteristics. In “Path to 5G: A Control Plane Perspective” some of the most significant system developments from earlier generations to 5G are considered in more detail. “RESTful APIs for the 5G Service Based Architecture” explores how the 5GC internal communication has been specified using a RESTful design, and explains how the protocols and interfaces developed provide opportunities for future integration with other systems.

Finally, three articles concern 5G standards development that involves most aspects of the system – including radio, network, and end-to-end service delivery aspects. “5G Multimedia Standardization” covers evolution of streaming services and media delivery architecture for 5G – including Virtual Reality 360° video streaming, real-time speech and audio communications VR evolution and user generated multimedia content. “3GPP 5G Security” presents security aspects as they differ from the 4G (LTE) system – including several major security enhancements achieved, applied to every aspect of the NG-RAN, 5G-AN and 5GC, as well as end to end communication made possible by the 5G system. “Management, Orchestration and Charging in the New Era” surveys the range of standards under development for management and orchestration of the access network and core network, with particular attention on management of network slicing.

Though the articles in this volume cover much of the standards included in Release 15 and to be included and enhanced in Release 16, this does not provide a complete overview of all work. The reader will learn essential aspects and advances to the 3GPP standard and be in an excellent position to follow work as it proceeds in the years to come.

Erik Guttman
Frank Mademann
Anand R. Prasad

Outlining the Roadmap to 5G

Joe Barrett

CEO, Global mobile Suppliers Association, UK
E-mail: joe.barrett@gsacom.com

Received 9 April 2018;
Accepted 3 May 2018

Abstract

This paper provides an overview of the background to the roadmap and evolution to 5G and the path the industry is taking to realise the benefits that will come from a mass industry deployment of this standardised mobile technology. 4G LTE network rollouts are continuing and new LTE features are being deployed to support both industry and user expectations for what 5G will deliver once the technology has been deployed.

Keywords: 3GPP, 5G, 4G, Trials, LTE, Ecosystem.

1 Introduction

Of all the technologies that have created the widest impact over the past thirty years, mobile communication could be argued to be the most prominent. 2017 was a breakthrough year for the mobile industry with mobile connections, including licensed cellular Internet of Things, passing 8.5 billion and the number of net mobile subscribers surpassing 5 billion globally [1].

Operator revenue exceeded 1 $trillion in 2017 and mobile technologies and services generated 4.5% of GDP, a contribution that amounted to $3.6 trillion of economic value. According to the Global mobile Suppliers Association (GSA) expectations and predictions across the mobile industry forecast 4G will continue to be a leading technology and that 5G will herald a shift to

Journal of ICT, Vol. 6_1&2, 1–14. *River Publishers*
doi: 10.13052/jicts2245-800X.611
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

massive connectivity, massive bandwidth, massive user experiences, as the amount of data we all consume continues to rise exponentially [2].

This level of growth brings its own challenges, and the mobile industry is not taking a back seat in dealing with the economics of delivering more and more data per user. New harmonized spectrum is needed for 5G and GSA – via is Spectrum Group – and in cooperation with the GSMA (GSM Association), is working to promote the global benefits of freeing up spectrum in different frequency bands for early 5G deployments, the first of which will happen in 2018.

The ecosystem is also mobilizing. Chipsets and devices supporting Category-16 (Cat-16) are now appearing [3] and Cat-18 capability will emerge in 2018. Gigabit LTE will be a common theme as more operators roll out Release 14 and Release 15 features [4].

5G is probably the most ambitious mobile generation technology envisaged to-date. Expectations on what 5G will deliver are high – suggesting new levels of performance, efficiency, and connectivity as well as better user experiences. There is however alignment in the industry on the potential solutions for 5G and agreement on the immense impact 5G will have across all aspects of industry and society.

2 LTE to LTE-Advanced Pro

The mass adoption of the 3GPP standard known as LTE (Long Term Evolution), has been the first globally accepted mobile technology, and arguably the fastest adopted mobile technology ever. It took just 5 years for LTE to cover 2.5 billion people, compared to 8 years for WCDMA/HSPA. By the end of 2017 there were 2.8 billion LTE subscriptions globally, as seen in Figure 1, with 832 million new subscriptions added during the year [5]. This equated to a 42.4% Year-on-Year growth with LTE now accounting for 35.7% of all global subscriptions. GSA forecasts LTE will account for more than 50% of all subscriptions by 2020 and 60% by 2022. By comparison GSA estimates 5G subscriptions will reach 400 million by 2022 [6].

There are currently very few countries that do not have at least one LTE network deployed [7] and these countries are mostly in Africa or are islands. According to GSA data in its Networks, Technologies & Spectrum database (GAMBoD), as of 10th April 2018 there were 855 operators investing in LTE as detailed in Figure 2. 667 LTE networks have been launched, according to data published by GSA with a further 127 networks either in deployment, planned, are in a testing/trialling phase or a license for LTE deployment has been granted. 59 LTE networks still need confirmation of their status.

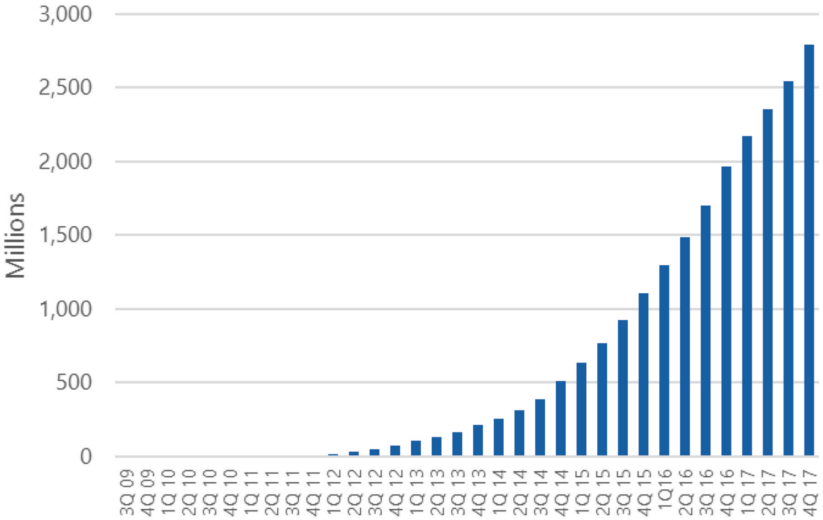


Figure 1 LTE Subscription Growth.

Source: Ovum WCIS.

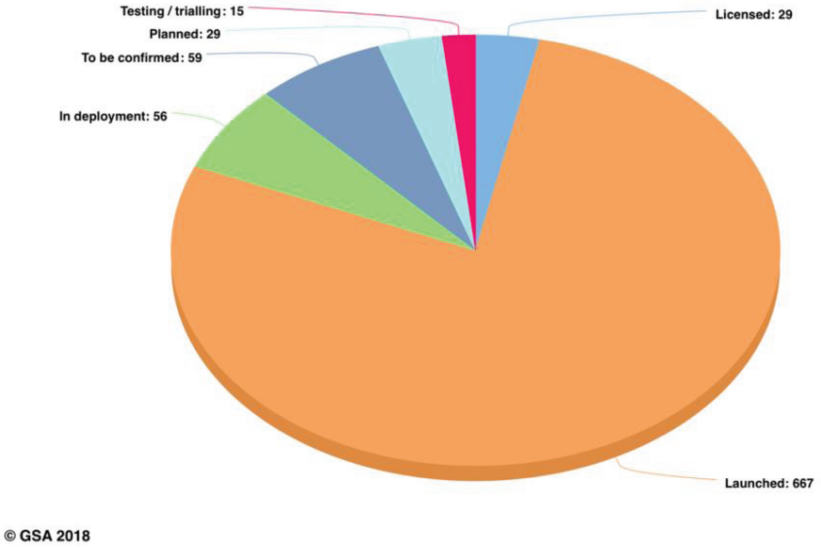


Figure 2 LTE network investments and launches.

Source: GSA NTS database.

The progress of the LTE standard to LTE-Advanced and LTE-Advanced Pro has also brought new features such as Carrier Aggregation, VoLTE, Mission Critical Communications, 4x4 MIMO and support for unlicensed frequency bands.

As of 10th April 2018 the following LTE deployment facts have been reported by GSA:

• LTE-Advanced	239 networks launched
• LTE-Advanced Pro	123 networks launched or traillng
• LTE-TDD	109 networks launched
• Carrier Aggregation	290 networks launched or trialling
• 4x4 MIMO	114 networks launched or trialling
• 256QAM	96 operators have launched or using in DL
• eMBMS	LTE Broadcast – 45 operators evaluating – 3 have launched
• VoLTE	221 investing – 143 launched

As is evidenced by the figures on the deployment of LTE features, LTE has enabled the mobile telecommunications industry to deliver a totally new mobile user experience with 88 operators in 55 countries investing in a Gigabit LTE service and 39 deployed or commercial network in 29 countries using Carrier Aggregation and 4 × 4 MIMO or above and 256QAM in the downlink [8].

There has been rapid deployment of the three key enabling technologies for Gigabit LTE networks, and many operators are deploying them in combination. Commercial network peak downstream speeds are dependent to a great extent on the maximum aggregated bandwidth available, and for users, very high (near-Gigabit, or above-Gigabit) speeds can only be realised with devices meeting downlink User Equipment (UE) Cat-16 specifications. Nonetheless, momentum is building strongly behind these advanced network technologies. Figure 3 shows the number of networks GSA has tracked using Carrier Aggregation (any number of aggregated carriers, including those in unlicensed spectrum), 4x4 MIMO (or higher-order), and 256QAM.

3 Ecosystem

The LTE ecosystem of devices has continued to expand over the GSA tracking period of the past 8–9 years [9]. As of February 2018, there were 10,655 LTE user devices including frequency and operator variants from 602 suppliers [10], – a 12% increase over the previous 3-month period.

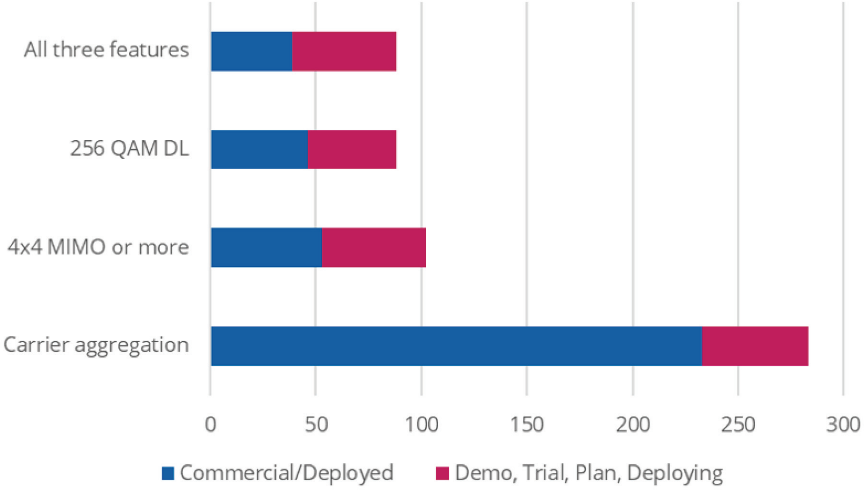


Figure 3 Commercial and demo/trial/planned networks using three core LTE-Advanced features.

Source: gsacom.com

The phone form factor has the largest ecosystem with 7,038 phones announced, including operator and frequency variants, giving a 66.1% share of all LTE devices. The LTE-connected tablet PC segment (761 devices) is also large, and the module segment (703 devices) is growing fast. See Figure 4. Most devices operate in the FDD mode while the number of terminals that support LTE TDD (TD-LTE) continues to grow and gain market share: 4,371 (41%) of LTE devices support the LTE TDD (TD-LTE) mode.

There has been a clear move from the ecosystem suppliers to bring higher order devices to market. This can be seen in the increase in devices supporting Cat-12 and above and the number of Cat-16 devices now available. The following data is extracted from the GSA LTE Ecosystem report – February 2018:

- 96 Cat-9 devices are launched (450/50 Mbit/s)
- 2 Cat-10 devices launched (450/100 Mbit/s)
- 53 Cat-11 devices are launched (600/50 Mbit/s)
- 72 Cat-12 devices launched (600/100 Mbit/s)
- 37 Cat-13 devices are launched (390/150 Mbit/s)
- 4 Cat-15 devices are launched (up to 750 Mbit/s DL)
- 42 Cat-16 devices are launched (up to 1 Gbit/s DL), up from 26 in November 2017.

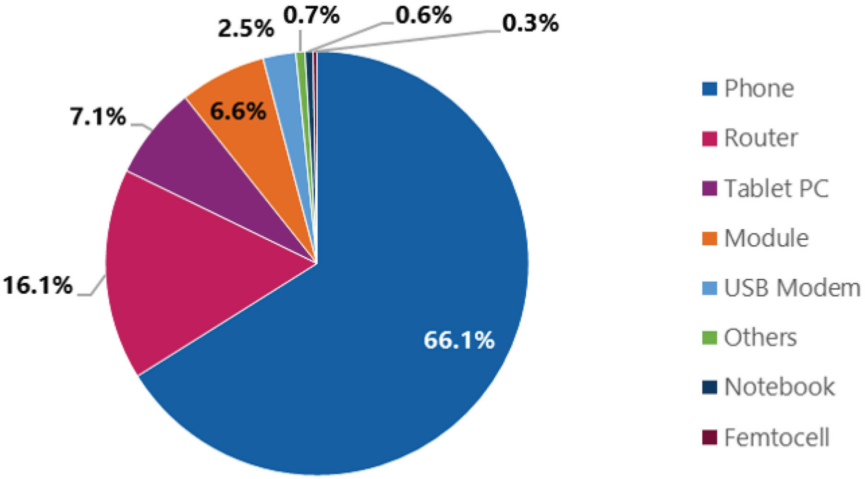


Figure 4 LTE Ecosystem by form factor February 2018.

Source: GSA LTE Ecosystem Report February 2018 – gsacom.com

Note to the above figures – that not all vendors publish details of UE category or up/downlink speeds.

The first Cat-18 devices (up to 1.174 Gbit/s DL) are starting to appear and will be tracked by GSA during 2018.

4 Chipsets

The ecosystem is dependent on the silicon vendors bringing chipsets to the market in a timely manner. There are at least 22 cellular LTE modem chipsets available separately [11], from six vendors: Hi-Silicon, Intel, Qualcomm, Samsung, Sanechips (formerly ZTE Microelectronics) and Spreadtrum. Other modems are MediaTek's WorldMode modem integrated within many of MediaTek's platforms.

The largest category of chipsets is mobile processor platforms: GSA has counted 101 commercially available mobile processors/platforms (other than those specifically designed for IoT applications) from 13 vendors. The total includes some market-specific variants such as chipsets designed to meet automotive industry standards.

Higher order chipsets (supporting Cat-20 and above) have been announced that are pre-empting the move to 5G. The chipset status is shown in Figure 5 and is taken from the GSA Chipset report – February 2018.

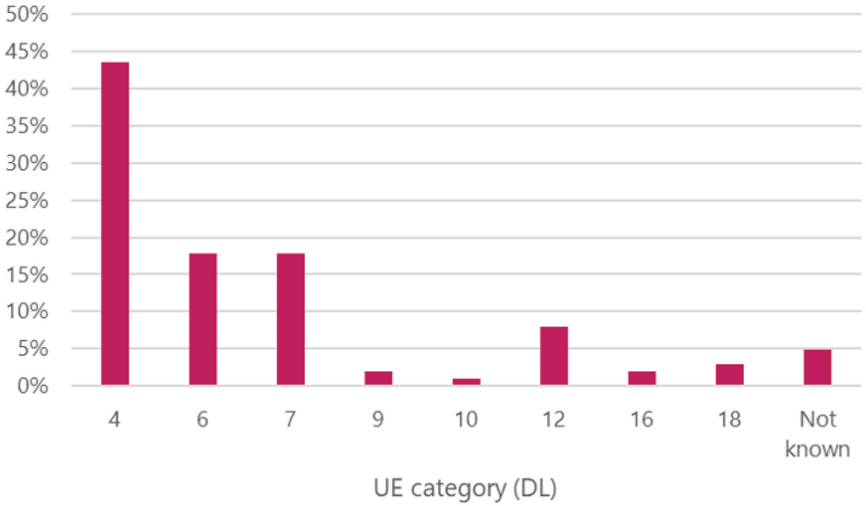


Figure 5 Percentage of mobile processors/platforms supporting specific UE categories.

Chipsets are increasingly coming to market supporting the latest 3GPP features that enable device manufacturers to meet the market need for more capacity and higher speeds. Smartphones are now more like media centres and applications like Augmented and Virtual Reality will push the boundaries of LTE technology to where 5G will be needed to achieve the required level of user experience.

The unlicensed bands are also a prime focus for LTE and technologies like MulteFire are due to impact the market in 2019 with Qualcomm and Nokia both promoting the technology.

At the other end of the chipset spectrum is the narrow-band segment. As of February 2018, there were 24 chipsets (modem chipsets and integrated processors/platforms) designed specifically to address M2M and IoT applications and supporting LTE Cat-1, Cat-M1 and Cat-NB1 user equipment. In the last quarter there were chipsets announced from Nordic Semiconductor, CEVA and ARM, and a second chipset from Neul (Huawei).

5 The Need for 5G

Mobile data increased by 65% in the 12 months between 3Q 2016 and 3Q 2017 [12], and the forecast is that global mobile data traffic as shown will grow from 14 ExaBytes per month in 2017 to110 ExaBytes per month with video accounting for 75% of data traffic. See Figure 6.

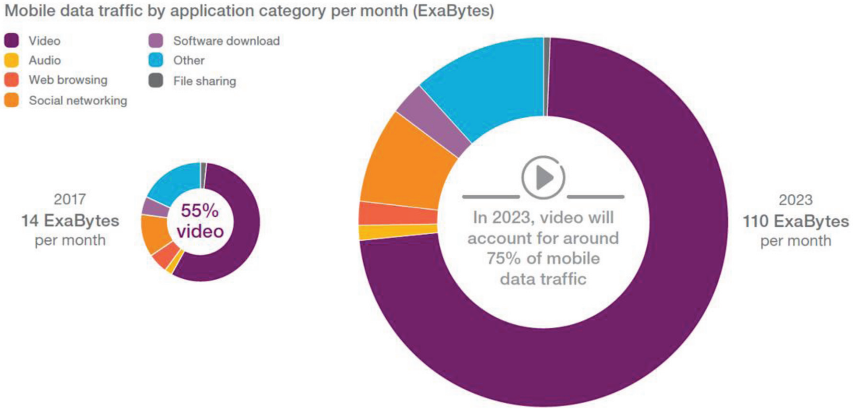


Figure 6 Mobile data traffic forecast.

Source: Ericsson Mobility Report – November 2017[12]

The need for 5G is driven by multiple factors, not least by what is often referred to as the Fourth Industrial Revolution (Industry 4.0); depicted as a fusion of technologies that is merging the physical, digital, and biological worlds. 5G will bring enhanced capabilities to support Industry 4.0 some of which are already being deployed or are planned, including The Internet of Things (IoT), factory automation, robotics, smart cities, connected drones and autonomous vehicles.

Worldwide monthly data traffic per active smartphone is predicted to increase from 2.9 GB to 17 GB [12]. This will only be realised with new spectrum – and lots of it. This means spectrum bands below 6 GHz need to be utilized as well as mmWave bands. 3GPP has defined a number of 5G/NR (New Radio) frequency bands and these can be seen in Tables 1 to 3 [13].

Table 1 5G/NR – mmWave bands

5G/NR – mmWave			
Band	Frequencies [GHz]	BW [MHz]	Duplex mode
n257	26.5–29.5	50–400	TDD
n258	24.25–27.5	50–400	TDD
n260	37.0–40.0	50–400	TDD
TBD	37.0–43.5	50–400	TDD

Table 2 5G/NR – spectrum below 6 GHz

5G/NR – Below 6 GHz			
Band	Frequencies [MHz]	BW [MHz]	Duplex mode
n77	3300–4200	10–100	TOD
n78	3300–3800	10–100	TOD
n79	4400–5000	40–100	TOD
n80	1710–1785/N/A	5–30	SUL
n81	880–915/N/A	5–20	SUL
n82	832–862/N/A	5–20	SUL
n83	703–748/N/A	5–20	SUL
n84	1920–1980/N/A	5–20	SUL

Table 3 5G/NR – re-farmed spectrum

5G/NR – Refarmed			
Band	Identifier	Frequencies [MHz]	BW [MHz]
n1	IMF Core Band	1920–1980/2110–2170	5–20
n2	PCS 1900	1850–1910/1930–1990	5–20
n3	1800	1710–1785/1805–1880	5–30
n5	850	824–849/869–894	5–20
n7	IMF Extension	2500–2570/2620–2690	5–20
n8	900	880–915/925–960	5–20
n13	US 700 Upper C	777–787/746–756	tbd
n20	CEPT800	832–862/791–821	5–20
n25	PCS1900G	1850–1915/1930–1995	tbd
n26	E850 Upper	814–849/859–894	tbd
n28	APT 700	703–748/758–8035–20	5–20
n34	TDD 2000 Upper	2010–2025	tbd
n38	IMF Extension Gap	2570–2620	5–20
n39	China TDD 1900	1880–1920	tbd
n40	TDD 2300	2300–2400	tbd
n41	TDD 2600	2496–2690	10–100
n50	TDDL-band	1432–1517	5–80
n51	TDDL-band, local	1427–1432	5
n66	AWS Extension	1710–1780/2110–2200	5–40
n70	AWS-3/4	1695–1710/1995–2020	5–25
n71	US 600	663–698/617–652	5–20
n74	FDDL-band	1427–1470/1475–1517	5–20
n75	Extended SDL L-band	N/A/1432–1517	5–20
n76	Extended SDL L-band, local	N/A /1427–1432	5

6 5G Use Cases

5G will usher in a new level of use cases as bandwidth requirements, latency, coverage, capacity and the economics of mobile deployments deliver a better user experience. A number of use cases have been discussed in the industry and some of them are covered briefly here [14].

Cloud Virtual and Augmented Reality: The bandwidth requirements needed for VR/AR to operate effectively are considerable and rendering can take up a huge amount of processing power in the device. Much of this rendering could be carried out in the cloud, but there is still the need to deliver high quality imaging with some applications needing in excess of 100 Mbps.

Connected Automotive: The automotive industry is moving quickly to support and test autonomous driving and in some cases autonomous cars will require ultra-low latency communications (ULLC) to support V2X (Vehicle to Everything).

Smart Manufacturing – including Robotics: Smart robotics and lean engineering are at the heart of Industry 4.0 and mobility is taking a foothold in the workplace in areas such as manufacturing, supply and asset management/tracking. Mobility is enabling real-time access to mission critical data and Artificial Intelligence is being used to speed up processes, improve industry performance and increase productivity.

Connected Energy: Understanding energy needs and distribution is paramount for the energy companies to effectively manage their business. Outage management and even video surveillance of sub-stations will all be part of the 5G network energy solution.

Connected Drones: Unmanned Aerial Vehicles (UAV) are ideal products for 5G with often a need for real time video to support traffic surveillance, crime prevention or emergency support – in case of a major fire for instance. UAVs will need a 5G connection to validate parcel deliveries, potentially using facial recognition to ensure delivery to the right location and person.

Smart Cities: A 5G connected city will without doubt be highly efficient and able effectively manage parking, lighting, traffic flow, refuse collection, floods, pollution monitoring and fly tipping. High megapixel IP cameras will be at the heart of the Smart City and many will need the bandwidth of 5G to deliver high resolution imaging.

7 Network-as-a-Service

Mobile architecture is becoming more IT centric with virtualisation, cloud native software, including a virtualised evolved packet core (EPC).

The need to improve scalability, resource utilisation and the economics of Radio Access Network operation means the mobile industry is also considering ways to reduce network operating expenditure, including reducing energy consumption, which can account for up to 15% of network OPEX in mature markets [15]. The by-product of this initiative will also have a positive impact on CO_2 emission reduction.

By virtualizing the EPC functionality, mobile networks can be customised to satisfy the different requirements of each customer by creating a tailored service solution either fully in the cloud, partially in the cloud with some local component – to ensure low latency for instance – or as a fully deployed private mobile network. Virtualisation is enabling the deployment of compact standalone 4G networks on oil rigs or in mining sites and 5G will further extend the -as-a-service capabilities from the full 5G network-as-a-service into areas like IoT-as-a-service or drone-as-a-service type solutions. As an example of the scale virtualisation can enable, Telefonica have demonstrated a complete mobile network on a drone [16].

8 5G Trials

Telecom operators from all continents have announced involvement with 5G demonstrations, lab tests and field trials. GSA has identified 134 operators, in 62 countries that have demonstrated, are testing or trialling, or have been licensed to conduct, field trials of 5G-enabling and candidate technologies [17]. Over 326 separate demonstrations, tests or trials have been announced confirming the high interest in bringing 5G technology to market.

At least 61 projects have involved testing Massive MIMO in the context of 5G (i.e. MIMO trials involving 64 or more transmitters, or lower order MIMO used on new high frequency spectrum bands or involving some other 5G aspect such as New Radio characteristics). This figure is up from 54 at the end of 2017.

There have been at least 73 demos, tests or trials of New Radio technologies (up from 42 during the last three months of 2017). GSA has also identified 19 projects that have explicitly featured network slicing.

One use case that has gained prominence is use of 5G to delivery fixed wireless broadband services. At least 15 tests so far that have specifically focused on the fixed wireless access (FWA) use case.

A variety of spectrum bands have been used for 5G Trials and the main bands are shown in Figure 7.

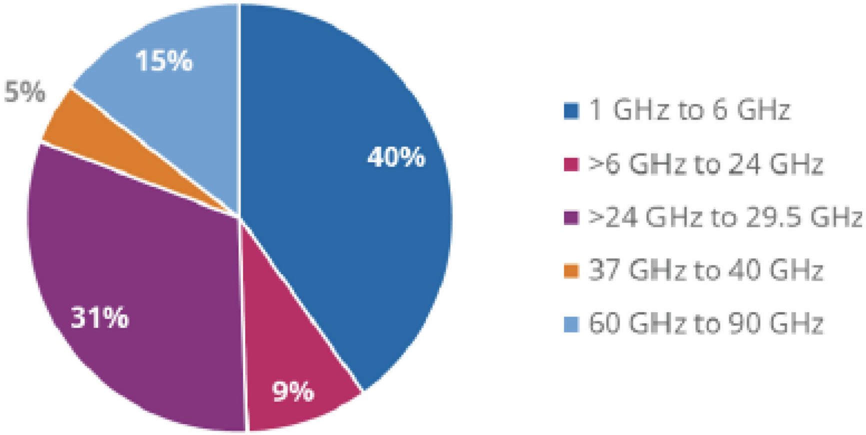


Figure 7 Distribution of 5G demonstrations and trials by broad spectrum ranges.

9 Conclusion

The road to 5G is clearly defined by the impact and success of 4G/LTE in the mobile industry over the past nine years – specifically in the last 5 years as LTE has dominated the growth in mobile devices and networks. Many of the features of 5G are being tested in LTE-Advanced Pro networks around the world creating a strong foundation for 5G to grow out of.

New spectrum for 5G will bring in massive capacity and the capability to deliver bandwidth intense services to meet the ongoing explosion of data usage that is being predicted. Video and VR/AR will reside at one end of this spectrum, but ULLC and lower bandwidth IoT services will reside at the other end of the services mapping. Gaming though may need both ULLC and high bandwidth.

References

[1] GSMA: The Mobile Economy 2018.
[2] Ericsson Mobility Report – November 2017.
[3] GSA Ecosystem Report – February 2018.
[4] GSA Gigabit LTE Networks Report – May 2018.
[5] Ovum WCIS report March 2018.
[6] GSA Report: LTE Subscriptions to 4Q 2017 – March 2018.
[7] GSA Report: LTE Not-Spots Map – October 2017.
[8] GSA Report: Progress to Gigabit LTE networks – February 2018.

[9] GAMBoD database tracking database setup in 2009.
[10] GSA Report: LTE ecosystem report – February 2018.
[11] GSA Report: LTE, 5G and 3GPP IoT Chipsets: Status Update – February 2018.
[12] Ericsson Mobility Report – November 2017.
[13] Ericsson 3GPP Spectrum Bands – Release-15 – January 2018.
[14] Huawei White Paper: 5G unlocks a world of opportunities – January 2018.
[15] Nokia White Paper: 5G Network Energy Efficiency – 2016.
[16] Telefónica LTE network running in just 40 grams of hardware.
[17] GSA Report: 5G Update – Global Market Trials – April 2018.

Biography

Joe Barrett is President of the Global mobile Suppliers Association (GSA), the leading industry association representing industry mobile suppliers worldwide. GSA promotes and advocates the family of 3GPP technologies covering the evolution of 3G, 4G/LTE and 5G technologies as well as standardised Low Power Wireless Access technologies. The GSA Spectrum Group advocates the harmonisation of spectrum for mobile use. Prior to becoming President of GSA, Joe worked at both Nokia and Qualcomm representing each company on the GSA Executive Board and helped drive the mobile technology agenda to position GSA as the trusted global supplier's voice in the industry.

5G Requirements and Key Performance Indicators

Toon Norp

Senior Business Consultant at TNO, 3GPP SA1 chairman at KPN,
The Netherlands
E-mail: toon.norp@tno.nl

Received 5 April 2018;
Accepted 3 May 2018

Abstract

This paper presents an overview of 5G requirements as specified by 3GPP SA1. The main drivers for 5G were the requirement to provide more capacity and higher data rates and the requirement to support different 'vertical' sectors with ultra-reliable and low latency communication. The paper discusses basic requirements that are new for 5G and provides 5G performance requirements. The paper also discusses a number of vertical sectors that have influenced the 5G requirements work (V2X, mission critical, railway communication) and gives an overview of developments in 3GPP SA1 that will likely influence 5G specifications in the future.

Keywords: Mobile telecommunications, 5G, requirements, performance.

1 Introduction

Work in 3GPP on 5G started in 2015 with a "Feasibility Study on New Services and Markets Technology Enablers" in the 3GPP Working Group SA1, which is responsible within 3GPP for services and feature requirement specification. The study first collected a large number of use cases – 74 in total – that illustrate what new capabilities are required from 5G [1]. The study took inspiration from a number of whitepapers from NGMN [2], the European 5G-PPP [3],

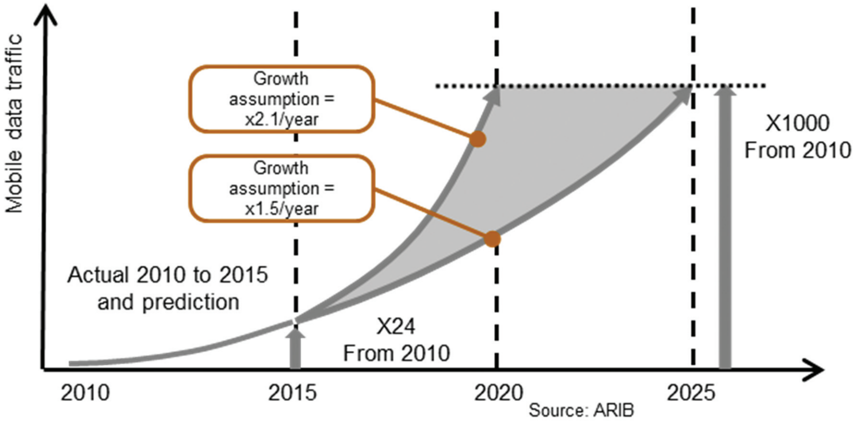
Journal of ICT, Vol. 6_1&2, 15–30. *River Publishers*
doi: 10.13052/jicts2245-800X.612
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

the China IMT2020 project [4], 4G Americas [5], the GSMA [6] and the Japanese standardisation development organisation ARIB [7].

A clear driver for 5G that emerged from all the whitepapers was the enormous growth of mobile data. Figure 1 shows mobile data growth projections from ARIB [7] which indicate that between 2020 and 2025 at the start of 5G mobile data traffic will have grown 1000x from 2010, at the start of 4G. 5G technology will have to be able to support network operators to cope with that data growth. It is clear that a 1000 times growth of data volume cannot lead to a similar growth of costs of energy consumption; 5G will have to be more efficient than earlier generations.

Not only the data volumes are expected to rise further, also the number of devices continues to grow. Expectations for the number of connected devices in the Internet-of-Things vary significantly; but all projections agree that in the 5G era several billions of devices will be connected.

The growth in mobile data volumes is largely related to the popularity of video applications. This trend is expected to continue in 5G. With the advance of technology the applications become ever more bandwidth demanding. Where users used to enjoy watching a funny video of a cat in low resolution; in the future they want to see similar videos in 3D, Virtual Reality, or Ultra High Definition. If also large TV screens are connected wirelessly, then 5G will need to deliver much higher data rates compared to 4G in order to support the demand for video applications.



Source: ARIB

Figure 1 Projected growth of mobile data [7].

The inspiration for 5G was not only in providing a more efficient and higher data rate version of 4G. As is shown in the name of the 3GPP SA1 study – “New Services and Markets Technology Enablers” – 5G is also very much about enabling new types of applications. In a digital society, consumers, governments, corporations and industries will make use of mobile telecommunications to improve all kinds of processes. These so called ‘vertical’ sectors, often use specific applications, with a diverse set of requirements on mobile telecommunications. Rather than designing specific wireless technology for each of these vertical applications, it is expected that 5G technology is flexible enough to support all kinds of applications, even together on a single network.

A common characteristic of vertical applications is that they have higher demands on reliability, availability and coverage. It is a nuisance when consumers cannot make use of mobile telecommunications to watch videoclips or update their social media. However, when payment terminals, trains, or the control of the electricity network become reliant on mobile telecommunications, then an outage of a mobile network has more far reaching consequences. For applications such as e-health, excellent coverage – both indoor and outdoor – is essential. It is not really possible to send a patient home with a hearth-monitor, if that has to come with a warning not to go into cellars or not to go camping.

Another aspect that 5G is improving, to cater for vertical applications, is the end-to-end latency. Many applications use control loops that will not work if the latency of data communication between sensors, controllers and actuators is too high. With current 4G networks, worst case latency in case of congestion can be up to a second. It is clear that this is not good enough if you are relying on mobile telecommunications for example to control an electricity network.

The 74 use cases that resulted from the first phase of the 3GPP SA1 feasibility study were consolidated in four different Technical Reports [8–11], each of which focuses on a different use case category (see Figure 2). The categories enhanced Mobile Broadband, Critical Communications, and massive Machine Type Communications are loosely based on similar categories introduced by the ITU. The main difference is that 3GPP SA1 also defined a category Network Operations. This category does not focus on performance requirements of specific services, but more on operational requirements that 5G networks have to fulfil.

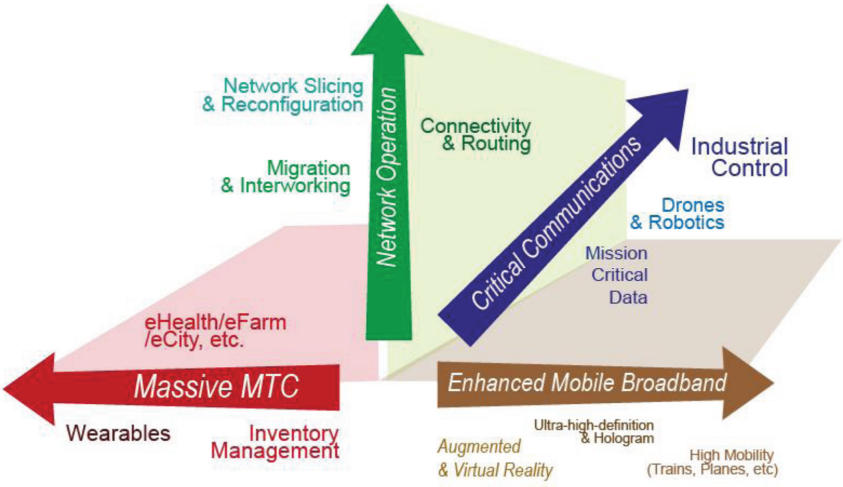


Figure 2 The four use case categories.

The four Technical Reports were subsequently consolidated in a normative 3GPP Technical Specification 22.261 "Service requirements for next generation new services and markets" [12]. This specification forms a basis for the 5G work in other 3GPP Working Groups. Note that not all requirements in [12] will be implemented in 5G Phase 1 (Release 15). 3GPP has concluded that it is unrealistic to implement all functionality for 5G in a single 3GPP release. However, this phased implementation has not yet been taken into account in the definition of requirements in [12].

In the remainder of this paper, first the 5G functional requirements from [12] are introduced in Section 2. Then Section 3 focuses on the 5G performance requirements. Section 4 discusses a number of vertical sectors that had a strong influence on the development of 5G requirements in 3GPP. Section 5 concludes with some future developments in 3GPP SA1 that are aimed at 5G Phase 2 (Release 16).

2 Basic Capabilities

5G is partly an evolution of existing mobile technologies. That implies that a lot of the functionality commonplace in earlier releases is also provided in 5G. The 5G requirement specification [12] only lists new functionality that is added in 5G; for existing functionality [12] refers to requirements specifications for

3G and 4G. The remaining part of Section 2 provides an overview of what functional requirements are specified for 5G.

2.1 Network Slicing

With the concept of network slicing (see Figure 3), operators can customise their network for different applications and customers. Slices can differ in functionality (e.g., priority, policy control, and security), in performance requirements (e.g., latency, availability, reliability and data rates), or they can serve only specific users (e.g., Public Safety users, corporate customers, or industrial users). A network slice can provide the functionality of a complete network, including radio access network and core network functions. One network can support one or several network slices.

A 5G user equipment can provide assistance information to enable the network to select one or more network slices. The network operator controls what slices should be available to a specific device and associated subscription.

Given the multitude of use cases for new verticals and services, a specific network may support only a subset of the vertical industries and services. However, this should not prevent an end-user from accessing all new services and capabilities. Therefore, the 5G system shall enable users to obtain services from more than one network simultaneously on an on-demand basis.

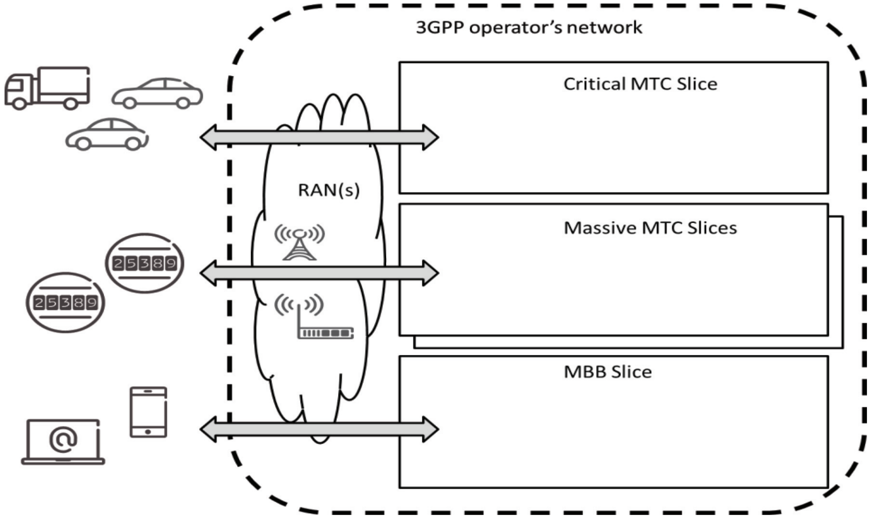


Figure 3 The concept of slicing [1].

2.2 Efficiency

5G will support diverse devices and services with different performance (e.g., high throughput, low latency and massive connection densities) and data traffic models (e.g., IP data traffic, non-IP data traffic, and short data bursts). In order to do this efficiently, 5G needs to be optimized for these different requirements.

For Internet of Things based applications, optimizations are needed to handle very large numbers of devices. Configuration, deployment, and use of IoT devices may benefit from optimizations such as bulk provisioning, resource efficient access, and optimization for device originated data transfer.

Sensors send data packages ranging in size from a small status update to streaming video. Smart phones similarly generate widely varying amounts of data. Where 4G was designed mostly with large amounts of data in mind; 5G will also have to efficiently support short data bursts without the need for lengthy signalling procedures before and after sending a small amount of data.

Cloud applications like cloud robotics perform computation in the network rather than in a device. This requires low end-to-end latencies and high data rates. The 5G system optimizes the user plane resource efficiency for such scenarios by locating operator or third party provided applications in a service hosting environment close to the end user. Video-based services (e.g., live streaming, Virtual Reality) and personal data storage applications have generated a massive growth in mobile broadband traffic. In-network content caching, provided by the operator, third party or both, can improve user experience, reduce backhaul resource usage and utilize radio resource efficiently for such applications. These optimization efforts also contribute to achieving higher reliability.

Energy efficiency is a critical issue in 5G. It is clear that a 1000 times increase of mobile data traffic from 4G to 5G cannot imply a similar increase of energy usage. Mobile operators are already one of the most significant users of electricity in many countries. For devices, energy efficiency translates directly into battery standby time. Small form factor devices also typically have a small battery and this not only puts constrains on general power usage but also implies limitations on both the maximum peak power and continuous current drain.

2.3 Diverse Mobility Management

The flexible nature of 5G will support different mobility management methods that minimize signalling overhead and optimize access for user equipment with different mobility management needs. Devices may be;

- stationary during their entire usable life (e.g., sensors embedded in infrastructure),
- stationary during active periods, but nomadic between activations (e.g., fixed access),
- mobile within a constrained and well-defined space (e.g., in a factory), or
- fully mobile.

Furthermore, different applications have varying requirements for the network to hide the effects of mobility. Applications such as voice telephony rely on the network to ensure seamless mobility. Applications such as video streaming on the other hand have application layer functionality (e.g. buffering) to handle service delivery interruptions during mobility. These applications will still require the network to minimize the interruption time.

Because of the much more distributed nature of 5G networks, mobility also has an impact in the network. With IP traffic offload or service hosting close to the network edge, mobility of a device also implies that the anchor node in the network may need to be updated. Internet peering and service hosting will have to follow the device when it is travelling across the network coverage area.

2.4 Multiple Access Technologies

The 5G system will support multiple 3GPP access technologies; next to one or more 5G New Radio (NR) variants, it will also include 4G radio interface technology (E-UTRA). Furthermore, 5G will support various non-3GPP access technologies.

Interoperability and integration among the various access technologies is important. 5G will select the most appropriate 3GPP or non-3GPP access technology for a service, taking into account e.g., service, traffic characteristics, radio characteristics, and the speed at which a device is moving. A single device can simultaneously use multiple access technologies, adding or dropping the various access connections as and when appropriate.

To support coverage even at sea or in remote areas and to improve availability in disaster scenarios, 5G shall also be able to provide services via satellite access. Service continuity is required between land based 5G access and satellite based access owned by the same operator or based on an agreement between the operators.

5G will also support fixed broadband access. A fixed access residential gateway can be a relay device, forwarding 5G connectivity to other end user devices within the premises. Alternatively, one or more home base stations are connected to the fixed access, which can then provide 5G coverage within the home.

2.5 Priority, QoS and Policy Control

The 5G network will support many commercial services and regulatory services (e.g., Public Safety communication) that need priority treatment. Some of these services share common QoS characteristics such as latency and packet loss rate, but may have different priority requirements. Mobile telephony and voice based services for Public Safety share common QoS characteristics, yet may have different priority requirements. The 5G network will have to be able to decouple the priority of a particular communication from the associated QoS characteristics such as latency and reliability. The traffic prioritisation may be enforced by adjusting resource utilization or pre-empting lower priority traffic.

As 5G is expected to operate in a heterogeneous environment with multiple access technologies, multiple types of devices, etc., it should support a harmonised QoS and policy framework that applies to multiple accesses. Furthermore, 5G QoS needs to be end-to-end (including radio access, backhaul, core network, and network to network interconnect) to achieve the 5G user experience (e.g., ultra-low latency).

2.6 Connectivity Models

5G will support different connectivity models. Next to direct network connections, 5G will also support indirect network connections (see Figure 4). With indirect network connections, a remote device can connect to the network via a relay device. Indirect network connections can be used to connect wearables via a mobile phone, but can also be used to improve indoor coverage, connecting e.g. printers or consumer electronic devices. The relay device can access the network using 3GPP or non-3GPP access technologies (e.g., WLAN access, fixed broadband access). Also for the connection between remote device and relay device 3GPP or non-3GPP radio technologies can be used.

When a remote device attempts to establish an indirect network connection, there can be multiple relay devices in proximity to choose from. A discovery and selection mechanism needs to be supported to select an optimal relay device for the remote device.

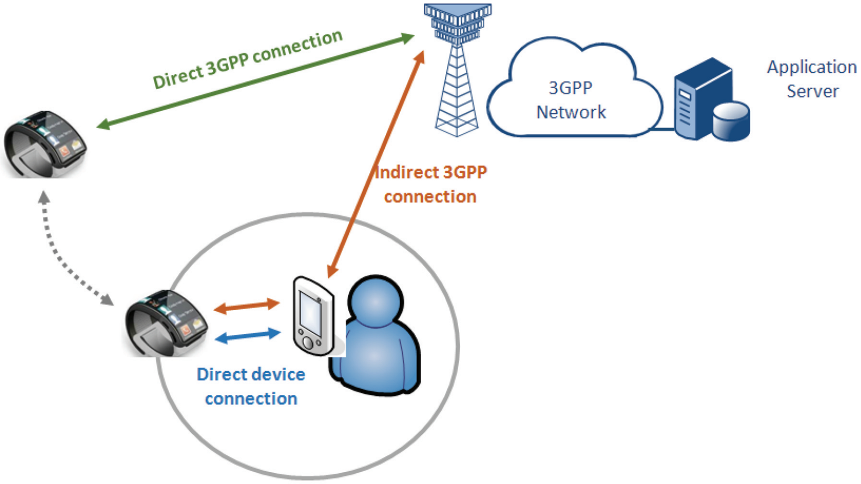


Figure 4 Connectivity modes for devices [8].

2.7 Network Capability Exposure and Context Awareness

Network capability exposure enables third party providers to interact with an operator network. With the advent of 5G, new network capabilities need to be exposed to the third party (e.g., to allow the third party to customize a dedicated network slice or to allow the third party to manage an application in a service hosting environment).

If network conditions can be provided to applications through network capability exposure, the applications can adjust resource usage to suit the network. At times when resources are scarce the application can be frugal with its resource usage. When resources are plenty the applications can compensate and use extra resources. Operators may provide incentives for applications to make their resource usage more network friendly, thus sharing the advantage of reduced network investments with the third party providers.

Applications may also provide the network with context information. For example, radio resource management can be optimised if the network is informed about application characteristics (e.g. expected traffic over time). Other characteristics of the device such as mobility, speed, battery status can be used to optimize allocation of functionality and content in the network.

2.8 Flexible Broadcast/Multicast Service

The proliferation of video services, ad-hoc multicast/broadcast streams, software delivery over wireless, group communications and broadcast/multicast IoT applications have created a need for a flexible broadcast/multicast service. Such a flexible broadcast/multicast service should allow flexible and dynamic allocation of radio resources between unicast and multicast services within a network, but also the deployment of stand-alone broadcast networks. It should be possible to stream multicast/broadcast content efficiently over wide geographic areas as well as target the distribution of content to very specific geographic areas spanning only a limited number of base stations.

3 5G Performance Requirements

Performance requirements highly depend on traffic scenarios. In an indoor hotspot scenario (e.g. for an office), the focus is on providing high data rates and high capacity. In a rural scenario, the focus is more on providing coverage. The data rates for a rural scenario will be lower, but 5G should ensure that a minimum data rate is available also in urban and rural macro scenarios. Table 1, shows the different performance requirements for the basic scenarios from indoor, dense urban, urban macro to rural macro. There are also scenarios for specific situations, such as broadband access in a crowd, broadcast, and connectivity in high speed trains, vehicles and airplanes.

Network access also needs to be supported in more extreme scenarios, with long range coverage, or in low end market scenarios, where access to power and backhaul facilities are not a given. Very large cell coverage areas of more than 100 km radius shall be supported with 1 Mbps downlink at cell edge. For constrained circumstances, and even larger areas, 5G shall be able to support a minimum user experience with 100 kbps, end-to-end latency of 50 ms, and a lower availability of 95%.

For vertical applications, other performance requirements are more important than data rates. For industry applications, the end-to-end latency is crucial. Motion control will not work if the time it takes to send information from a sensor to a controller is too long. Reliability – the percentage of packets successfully delivered within the time constraint – and communication service availability – the percentage of time the end-to-end communication service is delivered according to an agreed QoS – are crucial requirements for many industrial applications. Table 2 shows performance indicators for a number of vertical scenarios with low latency and high reliability requirements.

Table 1 5G performance requirements for high data rate and traffic density scenarios [12]

Scenario	Experienced Data Rate (Downlink)	Experienced Data Rate (Uplink)	Area Traffic Capacity (Downlink)	Area Traffic Capacity (Uplink)	Overall User Density	UE Speed
Indoor hotspot	1 Gbps	500 Mbps	15 Tbps/km^2	2 Tbps/km^2	250 000/km^2	Pedestrians
Dense urban	300 Mbps	50 Mbps	750 Gbps/km^2	125 Gbps/km^2	25 000/km^2	Pedestrians and users in vehicles (up to 60 km/h)
Urban macro	50 Mbps	25 Mbps	100 Gbps/km^2	50 Gbps/km^2	10 000/km^2	Pedestrians and users in vehicles (up to 120 km/h
Rural macro	50 Mbps	25 Mbps	1 Gbps/km^2	500 Mbps/km^2	100/km^2	Pedestrians and users in vehicles (up to 120 km/h
Broadband in a crowd	25 Mbps	50 Mbps	3,75 Tbps/km^2	7,5 Tbps/km^2	500 000/km^2	Pedestrians
Broadcast-like services	Maximum 200 Mbps (TV channel)	Modest (e.g., 500 kbps per user)	N/A	N/A	15 TV channels of 20 Mbps	Stationary to in vehicles (up to 500 km/h)
High-speed train	50 Mbps	25 Mbps	15 Gbps/train	7,5 Gbps/train	1000/train	Users in trains (up to 500 km/h)
High-speed vehicle	50 Mbps	25 Mbps	100 Gbps/km^2	50 Gbps/km^2	4000/km^2	Users in vehicles (up to 250 km/h)
Airplanes connectivity	15 Mbps	7,5 Mbps	1,2 Gbps/ plane	600 Mbps/ plane	400/plane	Users in airplanes (up to 1000 km/h)

Table 2 5G performance requirements for low latency and high reliability scenarios [12]

Scenario	End-to-End Latency	Communication Service Availability	Reliability	User Experienced Data Rate	Connection Density	Service Area Dimension
Discrete automation – motion control	1 ms	99,9999%	99,9999%	1 Mbps to 10 Mbps	100 000/km^2	100 × 100 × 30 m
Process automation – remote control	50 ms	99,9999%	99,9999%	1 Mbps to 100 Mbps	1000/km^2	300 × 300 × 50 m
Process automation – monitoring	50 ms	99,9%	99,9%	1 Mbps	10 000/km^2	300 × 300 × 50
Electricity distribution – medium voltage	25 ms	99,9%	99,9%	10 Mbps	1000/km^2	100 km along power line
Electricity distribution – high voltage	5 ms	99,9999%	99,9999%	10 Mbps	1000/km^2	200 km along power line
Intelligent transport – infrastructure backhaul	10 ms	99,9999%	99,9999%	10 Mbps	1000/km^2	2 km along a road

Low latency requirements come with specific service area dimensions as low latency communication is only possible when source and destination are nearby. The speed of light in optical communication – approximately 1 ms per 200 km – makes that network transport is not negligible, certainly not when latency caused by routers, switches and servers is taken into account. For the very low latency requirements of motion control, Table 2 specifies a service area dimension of a factory (100 × 100 m). This implies communication from a sensor in the factory to a controller that is also located within the factory; a much more distributed network topology than what is commonplace in 4G networks.

4 Vertical Applications

The support for vertical applications is one of the main goals of 5G. However, 3GPP has seen the introduction of vertical applications already in the 4G era. Though these early vertical applications can be supported in 4G, they have played an important role in the definition of 5G as well.

The first vertical in 3GPP SA1 was Mission Critical Communication. Targeted mainly at the Public Safety community, services like Mission Critical Push to Talk [13], Mission Critical Video [14] and Mission Critical Data [15] have been defined. For mission critical services, it is very important to ensure that communication remains available even in congestion situations. It cannot be that a fireman cannot communicate with other firemen because spectators are uploading videos of a fire. This implies that prioritization of mission critical services has to be provided. Recently, the mission critical community has explicitly indicated that requirement specifications for mission critical services shall also apply for 5G.

V2X (Vehicle to Anything) communication is a vertical application that has influenced 5G requirements from the start. The first V2X requirements are intended to be supported already by LTE. However, at the same time as the 5G requirements specification were completed, 3GPP SA1 also completed a 3GPP Technical Specification on enhancements of 3GPP support for V2X scenarios [16]. These functional and performance requirements in [16] are part of the overall 5G requirements.

Railway communications are the latest vertical application that introduced 5G Phase 1 requirements. European operators are looking for a replacement of their GSM-R networks in a timeframe up to 2030. Also in other countries, notably Korea, cellular network based technology for railway communication is planned. Use cases for railway communications, together with potential requirements are identified in [17]. Most of the resulting requirements find their way in updates of the mission critical requirements [13–15].

5 Outlook to Future Requirements

After the completion of Release 15 requirements specifications in June 2017, 3GPP SA1 has been working on further enhancements of 5G in the Release 16 time frame. A number of studies and work items that are intended to be incorporated in 5G Phase 2 are listed below:

- *LAN support in 5G* aims to support 5G LAN-type services over the 5G system. In this context, 5G LAN-type services allow a restricted set of devices to communicate amongst each other using Ethernet style data transport.
- *Communication for automation in vertical domains* identifies key performance indicators for various vertical use cases. Communication for vertical sectors may take place in separate, privately owned networks.
- *Using satellite access in 5G* is a study on how to integrate satellite communication with land based 5G networks. An example is how to support network selection when multiple land based mobile networks share a common satellite based access.
- *5G message service for massive IoT* aims to specify a light weight message service that can be used between (groups of) devices or between devices and application servers.
- *Positioning use cases* identifies new use cases, their scope and environment of use along with the related key performance indicators. Requirements can be achieved with a combination of 3GPP and non-3GPP positioning technologies.
- *Enhancements to IMS for new real time communication services* identifies a number of use cases (e.g. Augmented and Virtual Reality telepresence), where IMS and/or mission critical specifications need to be enhanced for new 5G real time communication services.
- *Layer for centric identifiers and authentication* aims to enable network operators to become identity providers. Use cases include how to use the new user identifier within the 3GPP system e.g. to provide customized services and how to provide this identifier to external parties to enable authentication for systems and services outside 3GPP.

References

[1] 3GPP TR 22.891, “Study on New Services and Markets Technology Enablers”

[2] NMGN 5G Whitepaper v1.0

[3] European Commission’s 5G PPP “5G Vision”, Feb. 2015. www.5g-ppp.eu.

[4] China IMT2020 (5G) Promotion Group white paper “5G Concept” February 2015.

[5] 4G Americas’ Recommendations on 5G Requirements and Solutions, October 2014.

[6] GSMA, “Understanding 5G: perspectives on future technological advancements in mobile”, Dec., 2014.

[7] ARIB 2020 and Beyond Ad Hoc Group White Paper, October 2014.

[8] 3GPP TR 22.861, “Feasibility Study on New Services and Markets Technology Enablers for Massive Internet of Things; Stage 1”.

[9] 3GPP TR 22.862, “Feasibility Study on New Services and Markets Technology Enablers – Critical Communications; Stage 1”

[10] 3GPP TR 22.863, “Feasibility Study on New Services and Markets Technology Enablers – Enhanced Mobile Broadband; Stage 1”

[11] 3GPP TR 22.864, “Feasibility Study on New Services and Markets Technology Enablers – Network Operation; Stage 1”

[12] 3GPP TS 22.261, “Service requirements for next generation new services and markets”

[13] 3GPP TR 22.179, “Mission Critical Push to Talk (MCPTT); Stage 1”

[14] 3GPP TR 22.281, “Mission Critical Video services”

[15] 3GPP TR 22.282, “Mission Critical Data services”

[16] 3GPP TR 22.186, “Service requirements for enhanced V2X scenarios”

[17] 3GPP TR 22.889, “Study on Future Railway Mobile Communication System”

Biography

Toon Norp is a Senior Business Consultant at TNO. Toon Norp joined TNO (former KPN Research) in 1991, where he has since been working on network aspects of mobile communications. Toon advises European operators on strategy, and architecture related to mobile core network, M2M/IoT, and 5G. He has been involved in standardisation of mobile networks for more than 20 years, and is the chairman of the 3GPP SA1 service aspects working group. Toon is member of the 5G-PPP association, a joint initiative between the European ICT industry and the European Commission to support the research and development of 5G infrastructures. Toon holds a Master’s Degree in Electrical Engineering from the Eindhoven University of Technology, The Netherlands.

5G NR Radio Interface

Balazs Bertenyi[1], Satoshi Nagata[2], Havish Kooropaty[3], Xutao Zhou[4], Wanshi Chen[5], Younsun Kim[6], Xizeng Dai[7] and Xiaodong Xu[8]

[1] *Chairman of 3GPP RAN, Hungary*
[2] *3GPP TSG-RAN – Vice Chairman, Japan*
[3] *Master Researcher at Ericsson Inc., 3GPP RAN1 Vice-Chairman,
San Francisco, USA*
[4] *Enginner at Beijing Samsung R&D Center, Chaoyang District, Beijing, China*
[5] *System Engineer at Ericsson, Greater San Diego, USA*
[6] *Head of Samsung RAN1 standards team and vice-chairman of 3GPP RAN1, Korea*
[7] *Engineer at Huawei, Beijing City, China*
[8] *Principal Researcher with CMCC, China*

Received 10 April 2018;
Accepted 3 May 2018

Abstract

This paper presents an overview of the 5G NR radio interface as specified by 3GPP. Specifically, the paper covers 5G NR in IMT2020 context, key design criteria and requirements, fundamental technology components of 5G NR, RF requirements and spectrum bands, Radio Resource Management and Link Monitoring and co-existence/sharing of 5G NR and LTE.

Keywords: 5G, NR, cellular radio technology, 3GPP, IMT2020, RRM, spectrum bands, mMTC, URLLC, eMBB.

1 3GPP 5G NR Standards Process in the Context of IMT2020

Mobile communications have become an integral part of daily life across the world: cellular technology developments are changing the society to a

Journal of ICT, Vol. 6_1&2, 31–58. *River Publishers*
doi: 10.13052/jicts2245-800X.613
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

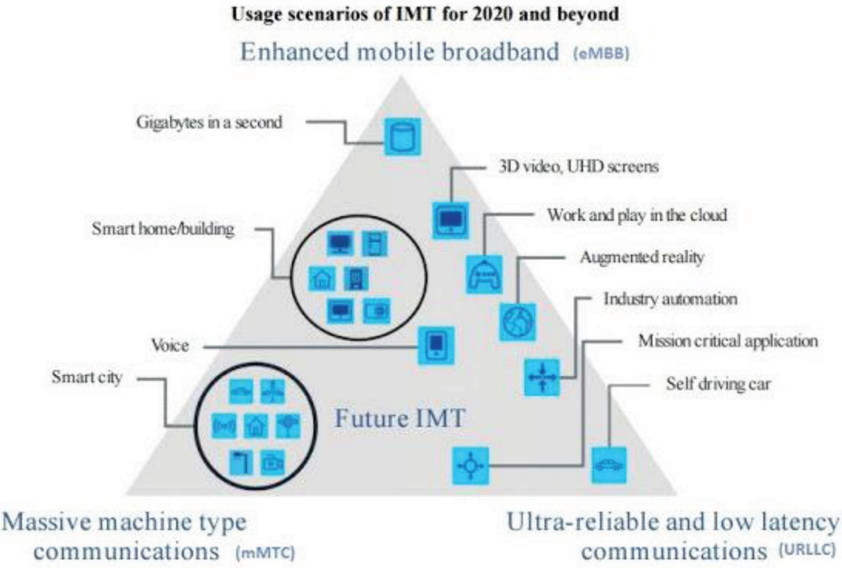

Usage scenarios of IMT for 2020 and beyond
Enhanced mobile broadband (eMBB)
Gigabytes in a second
3D video, UHD screens
Work and play in the cloud
Smart home/building
Augmented reality
Industry automation
Voice
Mission critical application
Smart city
Self driving car
Future IMT
Massive machine type communications (mMTC)
Ultra-reliable and low latency communications (URLLC)

Figure 1 5G use case landscape.

fully connected world. Cellular technology evolution has reached the 5th generation, 5G networks are expected to be the predominant choice for communications in 2020 and beyond. To this end, back in 2015 ITU-R established the 5G vision and described it in Recommendation ITU-R M.2083. In essence, 5G technology is expected to be applied to a diverse range of usage scenarios including enhanced mobile broadband (eMBB), massive machine type communication (mMTC) and ultra-reliable and low latency communication (URLLC) – see Figure 1.

Starting with an all-encompassing workshop in September of 2015 3GPP has set out to deliver the technology standards to fulfil the communication needs of the next 20 years. Roadmaps and plans were put in place to deliver on an extremely ambitious schedule, see Figure 2.

3GPP is set out to complete the first version of 5G technology standards in June 2018 (with ASN.1 protocol freeze in September 2018). As an intermediate step 3GPP is delivering an intermediate set of 5G standards 6 months ahead of this schedule to meet the high industry demand for rapid availability of 5G-based Mobile Broadband capacity expansion.

It shall be noted, however, that the realization of the full 5G vision will take several evolutionary steps after the initial launch, and will take 3GPP several more standards releases over the coming decade to deliver all the capabilities required.

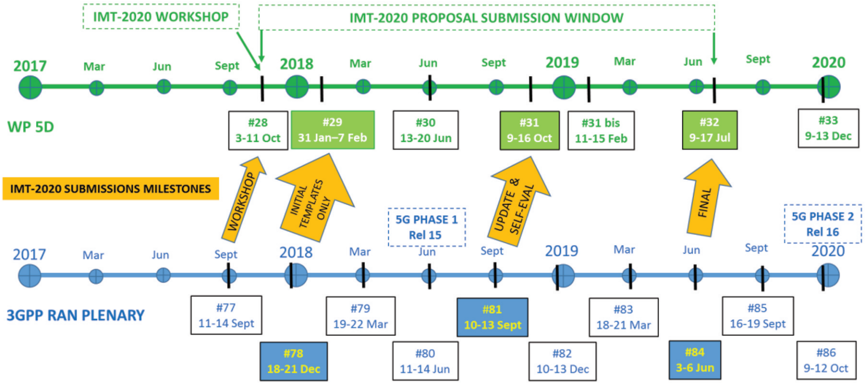


Figure 2 Overall time plan for 3GPP technology submissions to IMT2020.

2 Design Criteria and Requirements

The key capabilities of a 5G network are defined in ITU-R as shown in Figure 3.

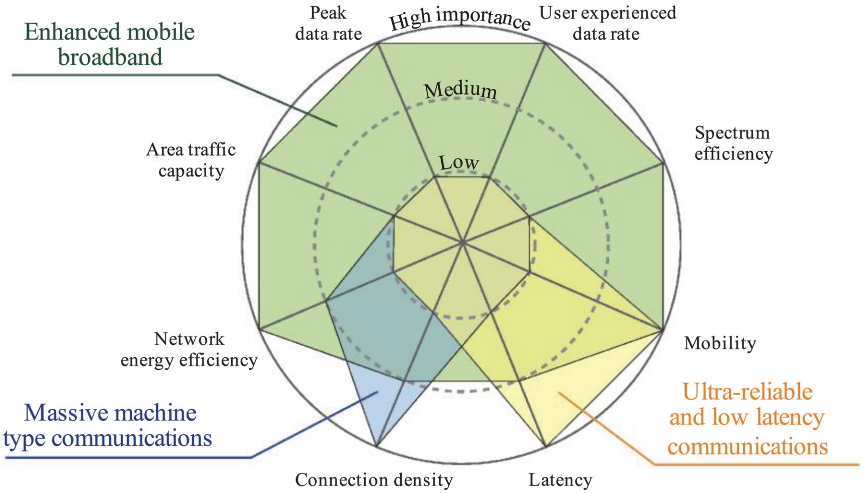
For Enhanced Mobile Broadband (eMBB) usage scenario, the 100 Mbps user experience data rate and area traffic capacity of 10 Mbps/m2 are expected with the support of large bandwidth and 3 times spectral efficiency improvement as compared to 4G systems. These capabilities should be reached while retaining sustainable energy consumption levels. Mobility is also important and should be improved to support devices moving with speeds as high as 500 km/h.

For Massive Machine Type Communications (mMTC) usage scenario, connection density is expected to reach 1,000,000 devices per km^2 due to the demand of connecting vast number of devices over the next decade.

For Ultra Reliable Low Latency (URLLC) usage scenario, the 1 ms latency with very high (99.999%) reliability has been put forward as a design goal.

To reach the 5G vision defined by ITU-R, 3GPP further studied the deployment scenarios and the related requirements associated with the three usage scenarios. The 3GPP requirements complement the ITU requirements defining relevant metrics to the usage scenario. For example, 3GPP also defines targets of low power consumption and deep coverage for mMTC usage scenario.

In general, both ITU and 3GPP requirements imply that 5G networks should deliver diverse capabilities depending on the type of services and applications. Furthermore, the unforeseen future services should be supported in a smooth manner. Meanwhile, these capabilities should be provided subject to the constraint of spectrum, energy consumption, and affordable cost. Therefore, 5G needs to be flexible with a unified radio interface of high


Enhanced mobile broadband
Peak data rate
High importance
User experienced data rate
Medium
Spectrum efficiency
Area traffic capacity
Low
Network energy efficiency
Mobility
Massive machine type communications
Ultra-reliable and low latency communications
Connection density
Latency

M.2083-04

Figure 3 Key capabilities of 5G networks.

spectrum utilization efficiency as well as energy efficiency. All this calls for a higher degree of innovation on the different technical components of the 5G system. In terms of cellular radio technology 3GPP has answered the call by designing a new radio interface, called NR.

3 5G NR Radio Interface Technology Components

3.1 Physical Layer Structure

In NR, similar to LTE, a radio frame is fixed to be 10 ms, which consists of 10 subframes each of 1ms. However, different from LTE which has a fixed subcarrier spacing (SCS) for 15 kHz, NR supports scalable numerology for more flexible deployments covering a wide range of services and carrier frequencies. In particular, NR supports the following SCSs (f_0):

- $f_0 = 15$ kHz * 2^m, where $m = \{0, 1, 2, 3, 4\}$, i.e., $f_0 = \{15, 30, 60, 120, 240\}$ kHz

Note that 15 kHz, 30 kHz and 60 kHz are applicable to carrier frequencies of 6 GHz of lower (sub-6), where 60 kHz, 120 kHz and 240 kHz are applicable to above 6 GHz carrier frequencies.

The subframe duration of 1ms is based on 15 kHz reference numerology with 14 symbols per subframe for the case of normal cyclic prefix (NCP). It is

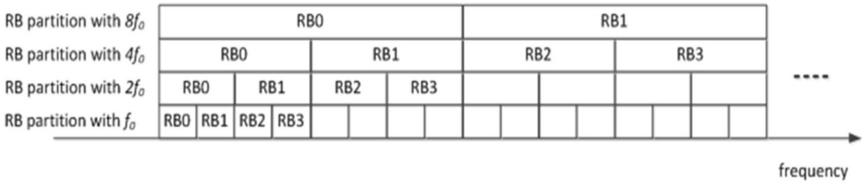


Figure 4 Illustration of nested RB-structure across numerologies.

also called a slot for 15 kHz SCS. For other SCSs, 14-symbol per slot is always assumed for NCP (except for 240 kHz, where 28-symbol per slot is assumed for NCP), resulting in SCS-dependent slot duration and nested slot structure across numerologies. As an example, a 30 kHz SCS has a slot duration of 0.5 ms, which can be mapped to two slots (each of 0.25 ms) for a 60 kHz SCS. Moreover, frequency-alignment within the channel is also achieved via nested resource blocks (or RBs, each of 12 frequency-consecutive tones) structure across numerologies, as illustrated below. Such nested slot structure and nested RB-structure facilitates multiplexing of different numerologies in a same cell or for a same UE, see Figure 4.

In addition to slots, NR frame structure supports *slot aggregation* and *mini-slots*. Slot aggregation refers to the case when a transmission can span two or more slots in order to achieve improved coverage and/or reduced overhead. Mini-slots (also known as *non-slot-based scheduling*) refer to the case when a transmission can span a number of symbols significantly less than the number of symbols in a slot (14), e.g., as small as 1-symbol. This provides more flexible resource management for a cell and possibilities to achieve low latency (LL), which when combined with ultra-reliability (UR) readily brings URLLC (URLL communications) services.

Flexible slot structure is one essential component for NR, not only for flexible resource management for current deployments but also necessary for future compatibility. To that end, NR supports up to two DL/UL switching points in a slot, particularly:

- Zero switching point within a slot, which implies 14 ‘DL’ symbols, 14 ‘flexible’ symbols, or 14 ‘UL’ symbols. The flexible symbols can be dynamically and UE-specifically indicated for DL or UL symbols based on actual need.
- One switching point within a slot, which starts with zero or more DL symbols and ends with zero or more UL symbols, with necessary ‘flexible’ symbols in between.

- Two switching points within a slot, where the first (or second) 7 symbols start with zero or more DL symbols and ends with at least one UL symbol at symbol #6, with zero or more 'flexible' symbols in between.

The maximum channel bandwidth supported by NR is 100 MHz for sub-6 and 400 MHz otherwise. Note that the maximum supported UL/DL channel bandwidth in the same band can be different. The minimum channel bandwidth is 5 MHz for sub-6 and 50 MHz otherwise. New maximum channel bandwidths, if necessary, can be added in future releases as NR is designed to ensure forward compatibility. The channel bandwidth of a cell that can be utilized for communications is as high as 98%.

3.2 Initial Access and Mobility

NR supports up to 1008 physical cell identities, twice as many as that of LTE. It follows a similar two-step cell identification procedure as in LTE, via detection of primary synchronization signal (PSS) and secondary synchronization signal (SSS). Time synchronization (in terms of symbol-level and slot-level) and frequency synchronization are also realized via PSS/SSS.

Master information block (MIB) of a cell is detected via a channel called primary broadcast channel (PBCH). System frame number (SFN) synchronization is acquired accordingly. In addition, PBCH demodulation enables reception of subsequent physical downlink control channels (PDCCH) and physical downlink shared channels (PDSCH), which schedule remaining minimum system information (RMSI), other system information (OSI), and paging messages.

For initial access, an essential building block called SS Block (SSB) is defined. A 4-symbol SSB consists of a 1-symbol PSS, a 1-symbol SSS, and a 2-symbol (and a bit extra) PBCH, as illustrated in Figure 5. The SCS for PSS/SSS depends on different frequency ranges, particularly:

- For sub-6 GHz: 15 kHz or 30 kHz for SSB
- For above-6 GHz: 120 kHz or 240 kHz for SSB

A SS burst set is comprised of a set of SS blocks (see Figure 5), each of potentially different beams necessary particularly for high carrier frequencies for initial access. Each SS burst set is limited to a 5 ms window regardless of the periodicity, which can be {5, 10, 20, 40, 80, 160} ms as indicated in RMSI, configured for SS burst sets. For initial cell selection, the SS burst set periodicity is default at 20 ms for all frequency range. Both the number of SS

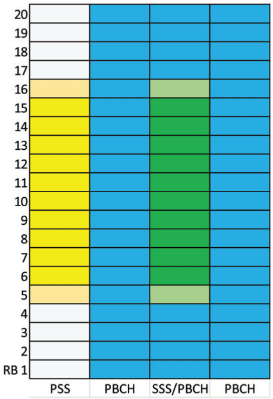


Figure 5 Illustration of SS block.

blocks (L) within a SS burst set and the location of SS burst set within the 5 ms window depend on the carrier frequency range. As an example,

- For carrier frequency range up to 3 GHz, $L = 4$
- For carrier frequency range from 3 GHz to 6 GHz, $L = 8$
- For carrier frequency range from 6 GHz to 52.6 GHz, $L = 64$

The number of possible PSS sequences is 3, each of a frequency-domain BPSK length-127 M-sequence. SSS sequence also has a length of 127 and it is a scrambled M-sequence. Both PSS and SSS are mapped to 127 consecutive tones within 12 RBs, where among the 144 tones, 8 tones and 9 tones are reserved on the two sides respectively. A 56-bit payload PBCH (including CRC) is mapped to a total of 240 tones. PBCH has a transmit-time-interval (TTI) of 80 ms. In other words, PBCH contents, including information such as SFN, SSB index, raster offset, default DL numerology, RMSI configuration, DM-RS location, etc., are updated every 80 ms. PSS, SSS, and PBCH are all one port only and share the same port.

PDSCH, scheduled by PDCCH, carries RMSI. The configuration of PDCCH for RMSI is provided by PBCH. The COntrol REsource SET (CORESET) configuration for RMSI is associated with a SS block in a SSB burst set. A one-bit information field in PBCH signals the SCS of RMSI, as well as OSI and other messages in random access procedures for initial access. The possible SCS combinations are:

- {SSB SCS, RMSI SCS} = {{15, 15}, {15, 30}, {30, 15}, {30, 30}, {120, 60}, {120, 120}, {240, 60}, {240, 120}} kHz

The RMSI PDCCH monitoring window is associated with an SSB and recurs periodically. The TTI for RMSI is 160 ms. Multiplexing of SSB and RMSI can be TDM or FDM. However, the pattern of multiplexing depends on and is restricted for a given SCS combination. As an example, for a {30, 30} SCS combination, only TDM pattern is allowed.

Similarly, for OSI, it is also carried by PDSCH, which is scheduled by PDCCH. For broadcast OSI CORESET configuration, the same configuration for RMSI CORESET is reused. The monitoring window configuration for OSI, e.g., time offset, duration, periodicity, etc., is explicitly signalled in a corresponding RMSI. In addition, for connected mode UEs, non-broadcast and on-demand (i.e., dedicated) OSI transmission is supported.

For paging, its subcarrier spacing of control and data channels is the same as that of RMSI. A UE is explicitly signalled paging occasion configuration, e.g., time offset, duration, periodicity, etc. Paging CORESET reuses the same configuration for RMSI CORESET. Two paging mechanisms are supported:

- Paging is done via PDSCH scheduled PDCCH, both channels in the same slot
- Paging is done via PDCCH only, useful for short paging messages

Random access (RA) enables a UE to access a cell, and it is performed by a 4-step procedure, similar to LTE:

- Message 1 (RA channel preamble): UE → gNB
    - It is based on Zadoff-Chu sequence with two sequence lengths, called long sequences and short sequences
    - Both contention-based RA (CBRA) and contention-free based RA (CFRA) are supported
    - One or multiple SSBs can be mapped to one PRACH transmission occasion
- Message 2 (Random access response or RAR): gNB → UE
    - It carriers information such as TA commands, temporary ID, etc.
- Message 3 (first PUSCH transmission): UE → gNB
    - It is scheduled by the UL grant in RAR
- Message 4 (PDCCH/PDSCH): gNB → UE

Radio resource management (RRM) in NR is based on measurements of SSB or CSI-RS, and can be reported with metrics such as reference signal received power (RSRP), reference signal received quality (RSRQ), and signal-to-interference-noise-ratio (SINR). Similarly, for radio link monitoring (RLM), both SS block based RLM and CSI-RS based RLM are supported. A hypothetical PDCCH block-error-rate (based on RLM-RS SINR) is the metric for determining in-sync (IS) and out-of-sync (OOS) with the cell.

3.3 Channel Coding and Modulation

In NR, new channel coding mechanism were chosen (LTE has used turbo codes and tail-biting convolutional codes).

NR uses LDPC codes for data which is transmitted on the physical downlink and uplink shared channels (PDSCH and PUSCH). Polar codes are used for downlink control information (DCI) that is transmitted on the physical downlink control channel (PDCCH) and for the master information block (MIB) which is transmitted on the physical broadcast channel (PBCH). Polar codes, repetition codes, simplex codes or the LTE Reed-Muller code are used for uplink control information (UCI) that is transmitted on the physical uplink control channel (PUCCH) or the PUSCH.

In general, LDPC codes are defined by a sparse parity check matrix which defines a set of linear equations (parity checks) that must be satisfied by any valid codeword. A parity check matrix is defined by a base graph along with a lifting size and cyclic shifts for the edges of the graph. For NR, two base graphs are defined, along with eight sets of lifting sizes which cover a wide range of information block sizes and code rates. For a given base graph, eight sets of parity check matrices are defined by providing eight different sets of cyclic shifts, one for each set of lifting sizes. The choice of base graph depends on the size and code rate of the initial transmission.

The operations applied to data that is LDPC-coded and transmitted on the physical downlink and uplink shared channels (PDSCH and PUSCH) are described below. Data is sent in units called transport blocks (TB) to which a CRC is attached so that the receiver may detect whether the TB was received correctly. The CRC is 24 bits when the TB size is larger than 3824 bits, and a 16 bit CRC is used in all other cases. The transport blocks (including the 24 bit CRC) are segmented into multiple code blocks when base graph 1 is used with a transport block size greater than 8424 bits or when base graph 2 is used with a transport block size greater than 3824 bits. When code block segmentation is applied, each code block is appended with its own 24-bit CRC. When no segmentation is used, the code block is the transport block. The one or more code blocks are then individually coded using the LDPC code. Rate matching is performed for each code block. Rate matching is a process that adjusts the number of coded bits of the code blocks to fit the resources available for the transport blocks. The available resources are dependent on the resources being used for other purposes including reference signals, system information, control channels and reserved resources. The coded bits selected in the rate-matching process, both for initial transmissions and re-transmissions, are chosen from a circular buffer into which the LDPC encoder output is written. Incremental redundancy is employed for Hybrid-ARQ re-transmissions by selecting different sets of coded bits for different transmissions. This is achieved by using different starting points in the circular buffer to generate the different sets of consecutive coded bits. After rate matching, bit-level interleaving is applied to each code block prior. The code blocks are then concatenated, scrambled and then modulated.

Polar codes strive to transform a set of noisy channels into noiseless and completely noisy synthetic channels when the code size approaches infinity. List decoding is used for decoding of polar codes with assistance from the CRC, and from additional parity check bits for uplink control, to choose candidates in the list.

For the DCI transmitted on the PDCCH, a block of 24 CRC bits is distributed among the DCI bits using an inter-leaver. The distributed CRC allows list decoding in the UE to potentially terminate early. Scrambling of the CRC bits at the end with the RNTI enables the UE to determine if the message is intended for it. Polar coding is used to generate the coded bits which are then rate matched and modulated. Rate-matching is performed using a circular buffer by using shortening, puncturing or repetition of the coded bits. The maximum polar code size is 512 bits. The polar code defined for PDCCH is reused for PBCH with a fixed input and output size where a 24-bit CRC is attached to the 32-bit MIB content before Polar encoding.

For uplink control information (UCI), repetition coding, (3,2) simplex coding or LTE Reed-Muller coding is used if the UCI length is 1 bit, 2 bits or 3–11 bits respectively. When the UCI length is 12–19 bits, polar coding is used with a 6-bit CRC attached at the end and three additional parity check bits inserted to assist with list decoding. When the UCI length is greater than 19 bits, polar coding is used with a 11-bit CRC attached at the end and a part of the CRC being useable for assistance to the list decoder. Rate-matching and bit-level interleaving are then performed. The maximum polar code size for UCI is 1024 bits. UCI of 360 bits or greater may be evenly segmented into two blocks which are individually encoded as described above, interleaved and concatenated prior to modulation.

In NR, QPSK, 16-QAM, 64-QAM and 256-QAM modulations are supported for the PDSCH and PUSCH. In addition, $\pi/2$-BPSK, where a $\pi/2$ shift is applied to successive modulated symbols, is supported when DFT-spread OFDM is used in the uplink. The PDCCH and the PBCH use QPSK. The PUCCH uses sequence selection, BPSK or QPSK depending on the PUCCH format and the number of bits with $\pi/2$-BPSK available as a configurable option.

3.4 Scheduling and Hybrid ARQ

For NR, the physical downlink control channel (PDCCH) is used for dynamic scheduling to deliver downlink control information (DCI), which includes the information required for the UE to process the scheduled data. For downlink data scheduling, the scheduling DCI also includes radio resource/timing information of the Hybrid ARQ-acknowledgement (HARQ-ACK) feedback. After receiving the downlink data, the UE reports its HARQ-ACK by the physical uplink control channel (PUCCH) at the radio resource/timing where the scheduling DCI indicates. For uplink data scheduling, there is no specific channel to inform HARQ-ACK. Once gNB fails to decode the uplink data, it schedules re-transmission of the uplink data, while if gNB successfully decodes the uplink data, it can schedule new uplink transmission data.

For NR, one of the key differences from LTE is its highly symmetric properties in the downlink and uplink scheduling and HARQ. In LTE, radio resource allocation schemes are different between downlink and uplink due to different multi access schemes, and downlink HARQ is basically asynchronous and adaptive while uplink HARQ is synchronous and non-adaptive. On the other hand, in NR, almost all scheduling and HARQ mechanisms are common between downlink and uplink such as: (1) radio resource allocation schemes, (2) Rank/modulation/coding adaptations, and (3) asynchronous and adaptive Hybrid ARQ.

Another key difference from LTE is its high flexibility in the time-domain. In LTE, time-domain radio resources for scheduled data and/or HARQ-feedback are basically not informed by the scheduling DCI, and it is determined by the frame structure and the UL-DL configuration. In NR, as shown in Figure 6, the scheduling DCI basically includes time-domain information of the scheduled data (and time-domain information of HARQ-ACK feedback in case of downlink) where the time-domain information here refers to the combination of the scheduled slot, the start symbol position, and the

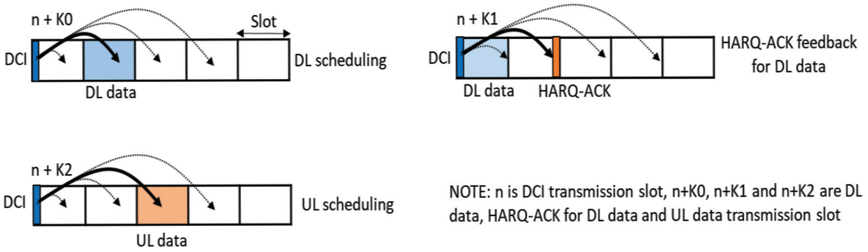

n + K0
Slot
DCI
DL scheduling
DL data
n + K1
DCI
HARQ-ACK feedback for DL data
DL data
HARQ-ACK
n + K2
DCI
UL scheduling
UL data
NOTE: n is DCI transmission slot, n+K0, n+K1 and n+K2 are DL data, HARQ-ACK for DL data and UL data transmission slot

Figure 6 Dynamic radio resource allocation timing for NR scheduling and HARQ.

transmission duration. By this, NR can easily realize various operations e.g., full/half duplex FDD, dynamic/semi-static TDD, and unlicensed operation *etc*. and satisfy different UE's requirements, e.g., lower latency, higher data rates. See Figure 6.

Regarding HARQ-ACK feedback for downlink data, and for uplink data transmission, UE requires processing time. In LTE, the minimum processing time is 3 ms. NR significantly reduces this processing time; it is subcarrier-spacing and demodulation reference signal mapping dependent, but overall the minimum processing time for downlink data is 0.2–1 ms and for uplink data is 0.3–0.8 ms. Together with enabling shortened data transmission duration, NR can realize lower U-plane latency compared to LTE.

3.5 MIMO

The use of multi-antenna technology in NR is focused on two objectives. First objective is to ensure sufficient coverage for NR deployment in over-6 GHz spectrum where propagation loss over wireless channels is significantly higher than that of sub-6 GHz spectrum. For example, compared to 2∼3 GHz where many of today's LTE networks are deployed, transmission over 28 GHz spectrum is expected to experience signal attenuation that is 100 times stronger. The second objective is to achieve a spectral efficiency that is 3 times that of LTE. This spectral efficiency improvement is especially important for sub-6 GHz spectrum since NR needs to compete against LTE in this spectrum.

Overcoming the large propagation loss is achieved in NR with multi-beam operation where the transmitter and receiver utilize multiple highly directional beams using a large number of antenna elements. At a given time instance, data transmission to or from a base station is made using one of the multiple beams that can provide sufficient signal quality. Support for multi-beam operation in NR includes beam quality measurement, beam quality reporting, beam assignment, and recovery mechanism in case the assigned beam quality is not good enough. NR provides support for multi-beam operation at every stage of the radio operation: initial/random access, paging, data/control transmission/reception, and mobility handling.

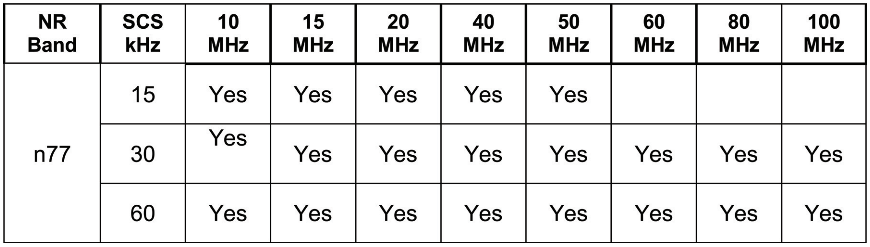
Improving the spectral efficiency over LTE is achieved in NR with the utilization of additional antenna ports in combination with an accurate channel status information (CSI). For example, compared to basic LTE which supports up to 4 transmit antenna ports are supported for a base station and 2 receive antennas are mandated for a terminal, NR supports up to 32 transmit antenna ports for a base station and 4 receive antennas are mandated for a terminal (in certain frequency bands).

An accurate CSI is essential in order for the base station to effectively separate the transmission signals to or from multiple terminals in the spatial domain. For the uplink, sounding reference signal (SRS) can be used for CSI acquisition. For the downlink of time-division duplexing (TDD) bands, when downlink-uplink channel reciprocity is available, channel measurement via UL signals can be used. For the downlink of frequency-division duplexing (FDD) bands or TDD bands where channel reciprocity is not available, NR supports efficient CSI reporting with high-resolution spatial channel information well beyond what LTE supports. High-resolution spatial channel information in NR is provided via a two-stage high-resolution precoding where the first stage selects a basis subset, and the second stage selects a set of coefficients for approximating a channel eigenvector with a linear combination of the basis subset.

4 RF Requirements and Spectrum Bands

NR brings new spectrum opportunities which allow the operating bands to extend up to 52.6 GHz in Release 15. Considering different testing methods to verify the Radio Frequency (RF) and Radio Resource Management (RRM) requirements in different frequency range, i.e., Over the Air (OTA) testing or conductive testing, two frequency ranges are categorized, i.e., Frequency Range 1 (FR1) 450 MHz–6 GHz and Frequency Range 2 (FR2) 24.25 GHz–52.6 GHz.

For operating bands within FR1 and FR2, prefix “n” with Arabic numerals is used to label the NR bands to differentiate the LTE bands labelled in Arabic numerals and UTRA bands labelled in Roman numerals. In FR1, prefix “n” with the same LTE band number are used for NR band with exactly same frequency range as LTE band. In addition to these bands, the range of



NR Band	SCS kHz	10 MHz	15 MHz	20 MHz	40 MHz	50 MHz	60 MHz	80 MHz	100 MHz
n77	15	Yes	Yes	Yes	Yes	Yes			
	30	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	60	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Figure 7 NR bands, sub-carrier spacing, channel bandwidth.

n65∼n256 and range n257–n512 are reserved for new bands in FR1 and FR2 respectively.

For each operating band, limited number of channel bandwidth has been specified for each subcarrier spacing (SCS). Taking UE channel bandwidth for Band n77 (TDD band with frequency range 3300 MHz – 4200 MHz) as example, the UE channel bandwidth is defined in a table manner in 3GPP specification. In the table, “Yes” indicates whether the channel bandwidth is supported for certain SCS of certain band. See Figure 7.

For UE channel bandwidth in Release 15, it is further specified that all channel bandwidth listed in current version of specification shall be mandatory supported by UE with a single component carrier in FR1, and all channel bandwidth below 200 MHz shall be mandatory supported by UE with a single component carrier in FR2.

To be noted, for some operating band in FR1 (n77 and n78 in current version of specification), additional channel bandwidth comparing with UE channel bandwidth, i.e., 70 MHz and 90 MHz are specified for BS channel bandwidth to allow more flexible deployment scenario.

More than 90% spectrum utilization has been specified in NR (except certain SCS) as maximum transmission configuration for each SCS and each channel bandwidth in implementation agonistic manner. In order to meet the relative emission requirements, the minimum guard band for each UE channel bandwidth and SCS has been also defined.

To locate the frequency position of RF channel and synchronization block, both channel raster and synchronization raster have been numbered as NR Absolute Radio Frequency Channel Number (NR-ARFCN) and Global Synchronization Channel Number (GSCN). To further assist the UE to find frequency position of RF channel and synchronization block in certain band, the applicable NR-ARFCN and GSCN are specified as the range of NR-ARFCH/GSCN and different step size for each operating band.

For FR2 NR UE and some NR BS types, due to highly integrate antenna implementations, physical conductive testing interface may not exist anymore. To specify the radiated requirements has to consider both RF performance and also the test methods. Overall, directional requirements, e.g., EIRP/EIS and non-directional requirements, e.g., TRP have been specified for corresponding RF requirements for UE and BS.

For UE RF requirements, not only the requirements for UE operating with single NR carrier but also the requirements for UE operating with Carrier Aggregation (CA), E-UTRAN-NR Dual Connectivity (EN-DC) and Supplementary uplink (SUL) are specified. For EN-DC operation, different set of requirements have been specified for intra-band EN-DC configuration and inter-band EN-DC configuration.

Different sets of requirements have been specified for different type BSs. In Rel-15, according to the applicable requirements, i.e., conductive, OTA or Hybrid, and also operating frequency range, four BS types are specified which are BS type 1-C, BS type 1-H, BS type 1-O and BS type 2-O. For example, BS type 1-C is defined as BS operating at FR1 with requirements set consisting of conductive requirements.

5 Radio Resource Management (RRM) and Demodulation

The new synchronization signals for initial access and mobility are designed in 5G NR to provide more flexibility and better trade-off between the system performance and UE power consumption for measurement. 5G NR supports both standalone (SA) and non-standalone (NSA including LTE-NR DC). Compared to LTE, multiple numerologies, the wider channel bandwidths, the more flexible uplink-downlink slot configurations, and the wider frequency ranges covering sub-6 GHz (Frequency range 1, FR1 for short) and mmWave (Frequency range 2, FR2 for short) are supported. The 5G NR RRM/RLM and demodulation performance requirements have been specified considering all these aspects.

5.1 Overview of RRM Core Requirements

The SS/PBCH block (SSB) burst consists of multiple SSB-s, which are associated with the different SSB indices and potentially with the different transmission beams. Besides, the CSI-RS signals can also be configured for beam management and measurement. The SSB-based measurement timing configuration (SMTC) with a certain duration and periodicity is used to restrict the UE measurement on the certain resources to reduce the UE power consumptions. Within SMTC period and on the configured SSB and/or CSI-RS, UE will conduct the RLM/RRM measurement.

For LTE-NR DC the initial access and mobility are done on LTE PCell (Primary Cell). NR CC will be added or released as SCell (Secondary Cell), while in SA mode all the operations will be done directly on NR PCell. Thus, the RRM requirements for LTE-NR DC (NSA) and SA are partially different. Table 1 below summarizes the RRM requirements. Those requirements guarantee the initial access and mobility performance for the LTE-NR DC, Supplemental Uplink (SUP), and NR-NR Carrier Aggregation (CA).

Table 1 RRM Core requirements for E-UTRA-NR DC (NSA) and SA

Core Requirements	SA and NSA Common	NSA Specific	SA Specific
RRC_IDLE state mobility	–	–	Cell selection
	–	–	Cell re-selection and interruption in paging reception
RRC_INACTIVE	–	–	Cell re-selection
	–	–	RRC_INACTIVE mobility control
RRC Connection Mobility Control	–	–	Handover
	–	–	RRC Re-establishment
	–	Random access for NR PSCell	Random access for NR PCell
	–		RRC connection release with redirection
Timing	UE transmit timing	–	–
	UE timer accuracy	–	–
	timing advance	–	–
	Cell phase synchronization accuracy for NR TDD BS	–	–
	–	Maximum transmission timing difference (MTTD) for E-UTRA-NR DC	–

(*Continued*)

Table 1 Continued

Core Requirements	SA and NSA Common	NSA Specific	SA Specific
	–	Maximum receive timing difference (MRTD) for E-UTRA-NR DC	Maximum receive timing difference (MRTD) for NR CA
Signal characteristics	RLM	–	–
	–	Interruption on NR PSCell due to the operations for E-UTRA PCell/SCell or UL carrier RRC reconfiguration	–
	–	Interruption on E-UTRA PCell/SCell due to the operation for NR PSCell/SCell or UL carrier RRC reconfiguration	–
	–	Activation/deactivation SCell in SCG for E-UTRA-NR DC	–
	UE UL carrier RRC reconfiguration delay	–	–
RRM measurement	–	Measurement capability for E-UTRA-NR DC	Measurement capability for SA NR
	Intra-frequency measurement with gap	–	–
	Intra-frequency measurement without gap	–	–
	Inter-frequency measurement	–	–
	–	–	Inter-RAT measurement

5.2 Definition of Intra-Frequency and Inter-Frequency Measurement

The first step for NR RRM is the definition of inter-frequency and intra-frequency measurement. Unlike LTE NR may utilize the different numerologies and configure the frequency location of SSB in a more flexible way. Thus, the definitions of inter-frequency and intra-frequency measurement for NR are different from LTE. For LTE, if the center frequency between the serving cell and the targeting cell is the same, the measurement can be viewed as intra-frequency measurement. Otherwise, it is viewed as the inter-frequency. For NR, a measurement is defined as a SSB-s based intra-frequency measurement provided the centre frequency of the SSB of the serving cell indicated for measurement and the centre frequency of the SSB of the neighbor cell are the same, and the subcarrier spacing of the two SSB-s are also the same.

5.3 Measurement Capability

In contrast to LTE the detected beam (SSB) number as well as the frequency layer number and cell number will be defined for NR capability by taking into account the support of multiple Tx beams. Besides, because there is difference between FR1 RF and FR2 RF chains, e.g., separate RF chains will be used for FR1 and FR2, and the analog Rx beamforming will be conducted for FR2, the capabilities for FR1 and FR2 are defined separately. When deriving the concrete numbers of carriers, cells and SSBs, the trade-off between the high capability and UE complexity is considered, and it is desirable not to increase NR UE capability too much compared to LTE capability.

The LTE-NR DC capable UE is required to measure the frequency layers of NR, LTE FDD, LTE TDD, UTRA FDD, UTRA TDD and GSM. But in the first version of the NR specifications (Release 15) a Standalone (SA) NR UE is required to measure LTE and NR only.

5.4 Measurement Gap

For measurement gap design there are totally 24 gap patterns defined to match the different SMTC durations, which correspond to different SSB burst lengths caused by different numerologies and different uplink-downlink transmission configurations. Given that different SCS-s are supported in FR1 and FR2, the specified lengths of SMTC between FR1 and FR2 are different. Thus, the different gap patterns apply to FR1 (roughly pattern #0$\sim$11) and FR2 (roughly pattern #12$\sim$23) separately.

As explained above, it is expected to be common that the separate RF chains is utilized for FR1 and FR2. Hence, the gap pattern for NR could be configured per UE or per-RF range.

In addition, the gap sharing between intra-frequency measurement with gap and inter-frequency measurement is specified to keep the lower UE power consumption. And the measurement gap timing advance mechanism will be utilized to improve the measurement performance for the case where the measurement gap and SMTC window duration are not aligned.

5.5 Measurement Requirements

The measurement requirements for NR include SSB based measurements and CSI-RS based measurements.

For SSB based measurement, UE will conduct intra-frequency/inter-frequency RSRP, RSRQ and RS-SINR measurement with or without gap. For CSI-RS based measurement, the CSI-RS based beam measurement will be conducted and UE will report the physical layer RSRP. The CSI-RS based RSRP, RSRQ and RS-SINR are also supported.

From measurement perspective, FR2 UE will utilize the analog and/or digital receiver beamforming for the measurement. Hence, longer measurement time is needed for FR2 to allow that the FR2 UE sweeps the whole space. The typical intra-frequency measurement period for FR1 consists of cell identification time, SSB timing index detection time and RSRP/RSRQ measurement period. At the same time, an FR2 UE is required to decode PBCH payload and thus the longer time is needed for intra-frequency measurement. Hence, the measurement requirements are specified for FR1 and FR2 separately.

The period of inter-frequency or intra-frequency measurement with gaps will be scaled by the gap periodicity based on the intra-frequency measurement period. The requirements in DRX mode will be derived by using the similar approach as for LTE.

5.6 Radio Link Monitoring (RLM)

There are two sets of target Block Error Rates (BLER) for NR RLM measurement: one corresponding to generic data service and one corresponding to voice service. Accordingly, two sets of configurations are specified. Both SSB-based RLM and CSI-RS based RLM are specified.

For NR the measurement period (for Signal-to-Noise Ratio (SNR)) is defined for each RLM-RS resource.

It is specified that a UE shall be able to monitor up to 2 RLM-RS resources for frequency range equal to or less than 3 GHz, 4 RLM-RS resources for frequency range larger than 3 GHz, and 8 RLM-RS resources for FR2.

5.7 Demodulation Performance Requirements

In NR the demodulation performance requirements will be specified covering single carrier, LTE-NR DC, and NR-NR CA schemes. This includes the baseline uplink-downlink transmission configuration, channel model, the number of Rx and Tx, etc. For UE both 2Rx and 4Rx based demodulation requirements are specified.

6 Coexistence and Sharing of 5G NR and LTE

The success of any new generation of wireless technologies depends on the ability of operators and users to migrate from currently deployed wireless systems to the new system. Each generation of wireless technologies has traditionally been introduced along with new spectrum that is made available for the deployment of that technology. As the number of users using the new technology gradually increases, spectrum can be migrated from the older to the newer technology. However, in the case of 5G NR, an additional constraint is that the new spectrum being considered for NR initially is generally not at low frequencies and therefore does not allow for the same level of coverage as the spectrum in which LTE is currently deployed. Hence, there is a motivation to consider techniques beyond the traditionally provided ability to handover users between older and newer systems. Many deployment options and techniques for system operation are defined as part of NR to achieve efficient coexistence and migration.

Dual connectivity between NR and LTE is a deployment option supported by NR where LTE, deployed typically in a lower frequency band and acting as an anchor carrier, can be used to ensure coverage while NR can be used at higher frequencies, including milli-meter wave frequencies, to provide very high capacity. The NR gNB and LTE eNB may or may not be co-located. While such deployments can allow NR to provide a capacity boost while running in such a non-standalone mode of operation, for a full transition to NR, NR will be deployed to operate in a standalone mode as well where an LTE anchor carrier is not needed. Eventually, it is desirable for NR to be deployed at lower frequencies as well. Given the paucity of spectrum in low frequency bands, providing new spectrum or re-farming LTE spectrum at lower frequency bands is difficult without an adverse effect to current users of LTE. Therefore, a finer granularity in allocation of radio resources between NR and LTE is needed. To achieve this, NR supports a carrier that is overlapped in frequency with LTE.

NR provides the option of only deploying the uplink (either independent or shared with LTE) in the lower frequency band while the downlink continues to operate in a higher frequency band. Another deployment option for NR is to operate a supplementary uplink in addition to a downlink and uplink operating in a higher frequency band. In cases where the UL coverage is the limiting factor at higher frequencies, these deployment options may allow operation of NR with a lower site density. When the UE transmits on two uplink carriers simultaneously, intermodulation distortion can affect the receiver sensitivity on the downlink carrier frequency. To avoid this, NR supports cases where the UE transmits only on one uplink at a time. When a supplementary uplink is used, uplink control and data are always transmitted on the same carrier in a slot with the sounding reference signal (SRS) being the only signal that may be transmitted on a different carrier from data and control in a slot although the transmissions don't occur simultaneously. For initial access, the UE selects between the supplementary uplink and the non-supplementary uplink based on the measured received signal strength on the downlink and a decision threshold that is broadcast by the network.

When the spectrum occupancy of the NR and LTE carriers overlaps, the system must ensure that all the signals and channels necessary for normal operation of both NR and LTE can be received in the downlink. The sharing of time-frequency resources can happen dynamically through scheduling or in a semi-static manner. For example, MBSFN (Multicast-broadcast Single-Frequency Network) subframes can be semi-statically configured in LTE and part of the symbols in the subframes can be used for the NR downlink. The design of NR also allows for forward compatibility with the use of reserved resources that are configurable for specific OFDM symbols, physical resource blocks and subframes. The NR signals and channels operate without the use of these resources. This mechanism could be used to protect signals such as the PSS, SSS and PBCH in LTE during an ongoing NR transmission. NR also is designed to be able to avoid transmissions in the resource elements corresponding to the cell-specific reference signals (CRS) in LTE. For instance, a specific pattern for NR synchronization and PBCH signal with 30 kHz subcarrier spacing is supported in order to avoid collision with LTE CRS symbols. These mechanisms facilitate coexistence of NR and LTE on the same carrier in the downlink.

Coexistence in the uplink is enabled mainly by using the scheduling flexibility integral to NR while minimizing the changes to LTE. When the NR and LTE uplinks are on the same carrier, or on separate carriers but with restrictions that force the UE to only transmit on a single uplink at a time, NR and LTE transmissions must be multiplexed in time. To enable such operation for LTE, UL/DL reference configurations corresponding to one of the existing LTE TDD reference configurations can be reused. Such operation can already be configured for an LTE FDD SCell that is carrier-aggregated with an LTE TDD PCell. For coexistence of NR-LTE these configurations have been extended to the LTE PCell. With this approach, the UE only transmits LTE uplink in the UL subframes defined by the reference configuration while NR can be transmitted in other subframes. For an LTE FDD carrier, it is desirable to ensure that all subframes are useable. For the downlink, this is possible and the HARQ-ACKs are transmitted on the uplink based on the UL/DL reference configuration. For the uplink, while transmissions from a single UE are restricted to a subset of subframes, separate offsets to the UL/DL configuration may be defined for different UEs to improve overall uplink utilization.

7 Summary and Outlook

3GPP have fully committed to delivering technology standards to provide the foundation for the 5G era. This article has described the basic technology components and characteristics of the NR radio interface. It is expected that NR will constitute the foundation for all 5G radio networks in the future, however, LTE will remain to be an integral part of operator networks providing an ever-improving mobile broadband experience.

Whilst initial 5G radio and system specifications are becoming available in 2018 with Release 15, it is expected that the delivery of all technology capabilities fulfilling the entire 5G vision will span over an evolutionary period of several years. 3GPP Release 16, 17 and beyond will continue to add functionality that enables efficient support of an ever wider ranging set of use cases and services.

Abbreviations

BS	Base Station
CORESET	Control Resource Set
E-UTRA	Evolved Universal Terrestrial Radio Access
eMBB	Enhanced Mobile Broadband
FEC	Forward Error Correction
HARQ	Hybrid Automatic Repeat Request
LDPC	Low Density Parity Check
MAC	Medium Access Control
mMTC	Massive Machine Type Communications
MIMO	Multiple Input Multiple Output
NCP	Normal Cyclic prefix
OFDM	Orthogonal Frequency Division Multiplexing
CP	Cyclic Prefix
FDD	Frequency Division Duplex
TDD	Time Division Duplex
PDSCH	Physical Downlink Shared Channel
PDCCH	Physical Downlink Control Channel
PBCH	Physical Broadcast Channel
PRACH	Physical Random Access Channel
PSS	Primary Synchronization Signal
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
RB	Resource Block
RRC	Radio Resource Control
RRM	Radio Resource Management
RLC	Radio Link Control
SCS	Sub Carrier Spacing
SSS	Secondary Synchronization Signal
UE	User Equipment
URLLC	Ultra Reliable Low Latency

Biographies

Balazs Bertenyi received an M.Sc. Degree in Computer Science and Telecommunications in 1998 at the Technical University of Budapest.

Balazs joined Nokia in 1998 and started to work on circuit switched mobile switching.

In 1999 he joined the research group on IMS (IP Multimedia Subsystem) and soon started to work in 3GPP standardization on IMS architecture.

- 2000–2003 Representative of Nokia in 3GPP SA2 (Architecture Working Group)
- 2003–2007 Head of delegation for Nokia in 3GPP SA2
- 2007–2011 Chairman of 3GPP SA2
- 2011–2015 Chairman of 3GPP TSG-SA (Technical Specification Group – Services and System Aspects)

In 2015 Balazs moved to the radio standardization group within Nokia to focus on 5G matters. He started attending TSG-RAN in 3GPP in 2015 covering 5G topics.

In 2017 Balazs was elected as Chairman of TSG-RAN for a 2-year term with a potential to be re-elected for one additional term.

Balazs has led various key projects for Nokia on Mobile Broadband architecture and standards strategy.

Satoshi Nagata received his B.E. and M.E. degrees from Tokyo Institute of Technology, Tokyo, Japan, in 2001 and 2003, respectively. In 2003, he joined NTT DOCOMO, INC. He worked for the research and development for wireless access technologies for LTE, LTE-Advanced. He is currently a senior research engineer working for 5G and 3GPP standardization. He had contributed to 3GPP over 10 years, and contributed 3GPP TSG-RAN WG1 as a vice chairman during November 2011 to August 2013, and also contributed as a chairman during August 2013 to August 2017. He is currently a vice chairman of 3GPP TSG-RAN since March 2017.

Havish Koorapaty (Havish.Koorapaty@ericsson.com) received his B.S., M.S., and Ph.D. degrees in Electrical and Computer Engineering from North Carolina State University in 1991, 1993 and 1996 respectively. He has been with Ericsson Research since 1996, where he has worked in the general area of wireless communications systems including cellular and satellite systems. His work spans a wide range of topics including error control coding, location determination and tracking, mobile phone systems engineering, 4G broadband wireless system design, wireless backhaul solutions, energy efficiency, spectrum sharing and small cells. He has over 200 technical papers and patents in these areas. He has represented Ericsson in standardization efforts in the TIA, IEEE and ETSI standardization bodies. Recently, he has worked on LTE evolution and 5G wireless systems. He has been involved in standardization efforts in 3GPP for 5G NR and LTE including serving as the rapporteur for the licensed-assisted access study and work items. He is currently serving as a vice-chairman in 3GPP RAN1.

Xutao Zhou received his Master Degree in Telecommunications form University of Warwick, UK. Mr. Zhou has been employed with Samsung Research China Beijing in 2007 and has been working in standard team, focusing on standardization of mobile communications. Mr. Zhou has been elected as 3GPP RAN4 Chairman in 2015. Since then, Mr. Zhou has been chairing the 3GPP RAN4 meetings for defining the radio frequency requirements for 5G NR and LTE.

Wanshi Chen is currently 3GPP TSG RAN1 Chairman, where under this position, he has successfully managed a wide range of 3GPP TRG RAN1 Long Term Evolution (LTE) and New Radio (NR) sessions. He has over 18 years of experiences in telecommunications in leading telecom companies including operators, infrastructure vendors, and chipset vendors. He has been with Qualcomm since 2006 and is responsible for LTE and NR research, design, and standardization. From 2000 to 2006, he was with Ericsson for 3GPP2 related system design, integration, and standardization. He also worked for China Mobile between 1996 and 1997 for wireless network maintenance and optimization. He received a Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA.

Younsun Kim received B.S. and M.S. degrees in electronic engineering from Yonsei University, and his Ph.D. degree in electrical engineering from the University of Washington, in 1996, 1999, and 2009, respectively. He joined Samsung Electronics in 1999 and has since worked on the physical layer standardization of cdma2000, HRPD, LTE, and recently NR. Currently, he is serving as the vice-chairman of 3GPP RAN1(physical layer) working group.

Xizeng Dai works with Huawei Technologies Co. Ltd since 2008. From October 2008 he began attending 3GPP RAN WG4 meeting as the regular delegate and made contributions to LTE demodulation and RRM topics from Release 8 onwards. Now he is the Vice Chairman of 3GPP RAN WG4 and chaired the RRM and performance session covering NR and LTE topics. He was elected as Vice Chairman in August 2015 and re-elected in August 2017 for the second term. He received his Ph.D of Electrical engineering from Tsinghua University in 2008. He received his MSEE from China Academy of Telecommunication Technology in 2004 and during that period worked on IS-95 and CDMA2000 system development.

Xiaodong XU is a Principal Researcher with CMCC. He has spent more than 10 years working on 3GPP standardization and LTE field network, ranging from physical layer design, higher layer design, and cooperative networking among 2G&3G&4G. And now, he mainly focuses on 3GPP's 5G technology. He is currently serving as the vice chairman of the 3GPP TSG RAN that is responsible for the specification of radio interface for 2G and onwards. Xiaodong XU received his PhD degree in communication and information system from the Southeast University, Nanjing, China, in 2007.

NG Radio Access Network (NG-RAN)

Balazs Bertenyi[1], Richard Burbidge[2], Gino Masini[3], Sasha Sirotkin[4] and Yin Gao[5]

[1]*Chairman of 3GPP RAN, Hungary*
[2]*Senior Wireless Systems Architect at Intel Corporation, Shrivenham, Oxfordshire, UK*
[3]*Systems Manager, Concepts and Standards, at LM Ericsson AB, Stockholm, Sweden*
[4]*Vice-Chair of 3GPP RAN3 at Intel Corporation, Israel*
[5]*Vice Chairman 3GPP, ZTE Corporation, China*

Received 10 April 2018;
Accepted 3 May 2018

Abstract

This paper presents an overview of the NG Radio Access Network (NG-RAN) architecture and key protocols. NG-RAN is the new RAN defined in conjunction with 5G by 3GPP. The paper presents the overall architecture, migration path options, the 5G base station architecture and key protocol components.

Keywords: 5G, NR, radio architecture, NG RAN, CU, DU, Xx, Xn, RRC, MAC, PDCP, RLC.

1 Overview of the NG-RAN Architecture

The NG-RAN represents the newly defined radio access network for 5G. NG-RAN provides both NR and LTE radio access. An NG-RAN node (i.e. base station) is either:

- a gNB (i.e. a 5G base station), providing NR user plane and control plane services;
  or,
- an ng-eNB, providing LTE/E-UTRAN services towards the UE.

Journal of ICT, Vol. 6_1&2, 59–76. *River Publishers*
doi: 10.13052/jicts2245-800X.614
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

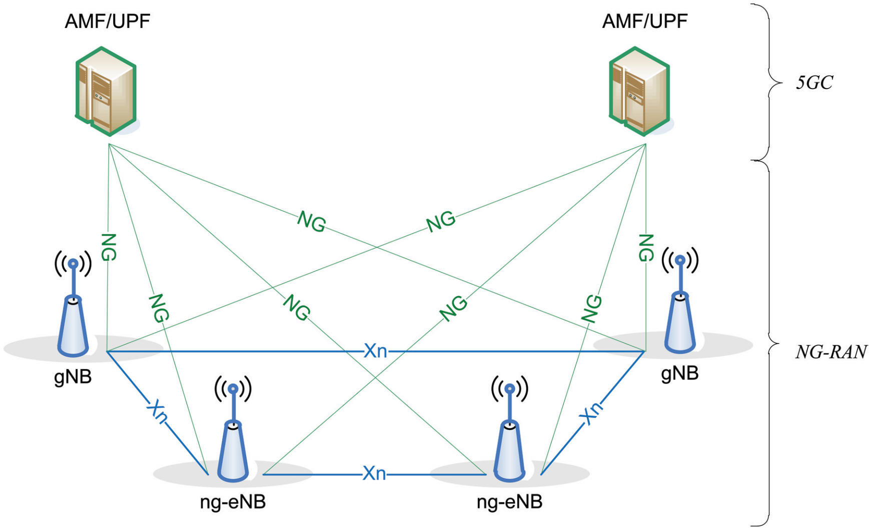


Figure 1 NG-RAN in relation to the 5G system.

The gNBs and ng-eNBs are interconnected with each other by means of the Xn interface. The gNBs and ng-eNBs are also connected by means of the NG interfaces to the 5G Core (5GC), more specifically to the AMF (Access and Mobility Management Function) by means of the NG-C interface and to the UPF (User Plane Function) by means of the NG-U interface.

The overall relation of NG-RAN in relation to the overall 5G system is shown in Figure 1.

Both the user plane and control plane architectures for NG-RAN follow the same high-level architecture scheme, as depicted in Figure 2 below. For further details of the protocol stacks see Section 4.

2 Architecture Options and Migration Paths

One of the distinctive features of NG RAN is the capability to operate in both so-called "Stand-Alone" (SA) operation and "Non-Stand-Alone" (NSA) operation. In SA operation, the gNB is connected to the 5G Core Network (5GC); in NSA operation, NR and LTE are tightly integrated and connect to the existing 4G Core Network (EPC), leveraging Dual Connectivity (DC) toward the terminal. In a Dual Connectivity architecture, a Master Node (MN) and a Secondary Node (SN) concurrently provide radio resources towards the terminal for enhanced end-user bit rates. Both NSA and SA architecture options are specified as part of the phase-1 5G standards of 3GPP in Release 15.

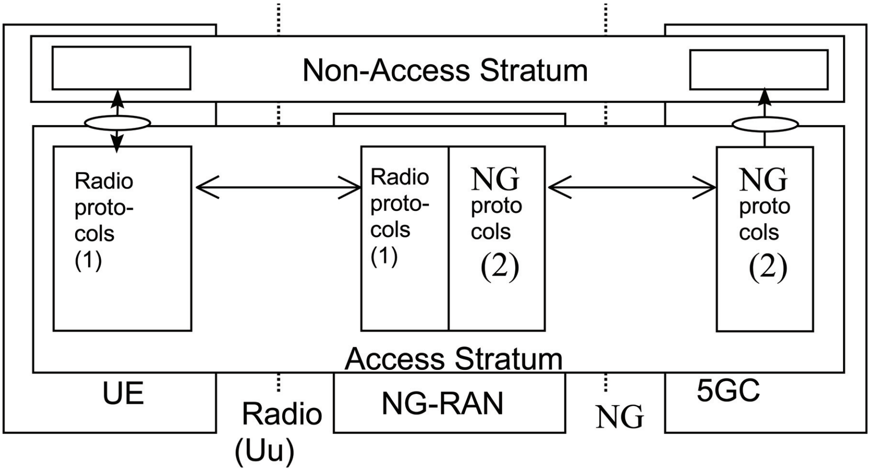


Figure 2 Overall NG-RAN architecture.

One can derive several different configuration options from the overall architecture, each of these options represent a viable deployment option for network operators. These architecture options are depicted in the subsections below. The numbering of these architecture options does not bear any particular logic or significance, it is purely historical.

2.1 NR gNB Connected to the 5GC (Option 2)

In this option, the gNBs are connected to the 5G Core Network (5GC) through the NG interface. The gNBs interconnect through the Xn interface.

2.2 Multi-RAT DC with the EPC (Option 3)

In this option, commonly known as EN-DC (LTE-NR Dual Connectivity), a UE is connected to an eNB that acts as a MN and to an en-gNB that acts as a SN, see Figure 3. An en-gNB is different from agNB in that it only implements part of the 5G base station functionality that is required to perform SN functions for EN-DC.

The eNB is connected to the EPC via the S1 interface and to the en-gNB via the X2 interface. The en-gNB may also be connected to the EPC via the S1-U interface and to other en-gNBs via the X2-U interface. The resulting architecture is shown in Figure 3 below. Notice that the en-gNB may send UP to the EPC either directly or via the eNB.

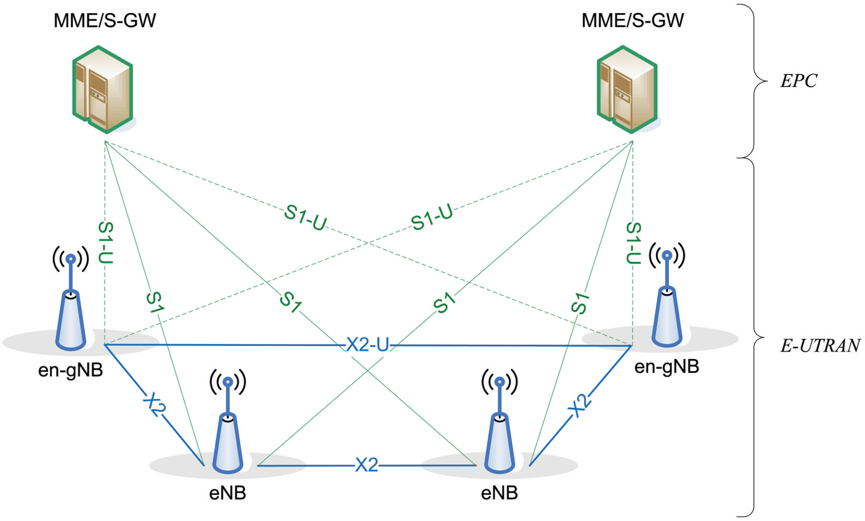


Figure 3 Overall LTE (E-UTRAN)-NR DC architecture.

2.3 Multi-RAT DC with the 5GC, NR as Master (Option 4)

In this option, a UE is connected to a gNB that acts as a MN and to an ng-eNB that acts as an SN. This option requires the 5G Core to be deployed. The gNB is connected to 5GC and the ng-eNB is connected to the gNB via the Xn interface. The ng-eNB may send UP to the 5G Core either directly or via the gNB.

2.4 LTE ng-eNB Connected to the 5GC (Option 5)

In this option, the ng-eNBs are connected to the 5G Core Network (5GC) through the NG interface. The ng-eNBs interconnect through the Xn interface. Essentially this option allows the existing LTE radio infrastructure (through an upgrade to the eNB) to connect to the new 5G Core.

2.5 Multi-RAT DC with the 5GC, E-UTRA as Master (Option 7)

In this option, a UE is connected to an ng-eNB that acts as a MN and to a gNB that acts as an SN. The ng-eNB is connected to the 5GC, and the gNB is connected to the ng-eNB via the Xn interface. The gNB may send UP to the 5GC either directly or via the ng-eNB.

2.6 Migration Considerations

When 5G is first rolled out with NR, a likely scenario is to deploy NR on higher frequencies than for LTE. In this case, NR coverage is typically much smaller than LTE coverage, especially with frequencies above 6 GHz. Then, it is desirable to leverage the existing LTE coverage to provide continuous nationwide coverage and mobility, while boosting User-plane capacity with NR in target areas with high traffic load. Option 3 enables operators to launch the NR service in this way, building on top of their existing investments for E-UTRAN and EPC.

When operators decide to introduce the 5G Core, this will “unlock” a new set of possible deployment scenarios, among which the support for NR as a stand-alone Radio Access Technology (RAT) (Option 2), while at the same time leveraging the deployed LTE nodes as secondary nodes through Dual Connectivity (DC) (Option 4). Another possibility for the introduction of 5G Core is to keep LTE as the main “anchor”, connecting it to the 5G Core (Option 5) while still leveraging NR as secondary node through DC (Option 7).

The choice between deploying NR with 5GC as “anchor” and keeping LTE as “anchor” with the new 5GC, will typically be a business decision by each operator. It will typically depend on factors including deployed LTE network density, availability of new frequencies, rate of increase for end-user traffic demand, and relative “weight” in the business case of new functionality (such as e.g. slicing) which only the new networks can provide.

3 5G NR Base Station (gNB) Architecture

The 4G RAN architecture was based on a “monolithic” building block, the eNB. This resulted in a very simple RAN architecture, where few interactions between logical nodes need to be specified. Since the earliest phases of the NR study, however, it was felt that splitting up the gNB (the NR logical node) between Central Units (CUs) and Distributed Units (DUs) would bring additional benefits. Some benefits in this regard were in fact identified already in the early study phase, including:

- A flexible hardware implementation allows scalable cost-effective solutions.
- A split architecture allows coordination of performance features, load management and real-time performance optimization. It also enables virtualized deployments.

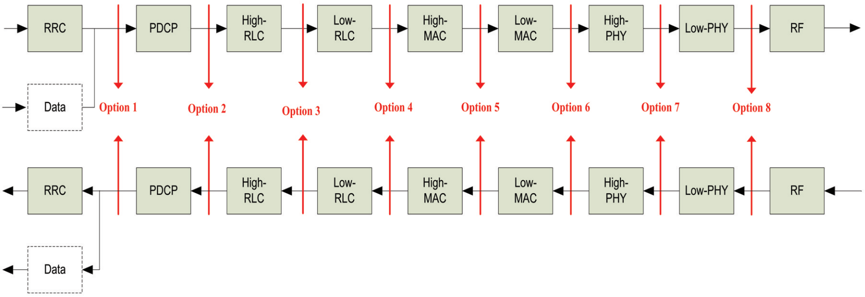


Figure 4 Function Split alternatives.

- Configurable functional splits enable adaptation to various use cases, such as variable transport latency.

The choice of how to split NR functions in the architecture depends on radio network deployment scenarios, constraints and envisaged services. For example, it depends on the need to support specific QoS settings per offered services (e.g. low latency, high throughput, specific user density and load demand per given geographical area (which may influence the level of RAN coordination), or the need to interoperate with transport networks having different performance levels: from ideal to non-ideal.

Several possible CU-DU split options, shown in Figure 4, were considered during the study phase. The E-UTRA protocol stack, which includes PHY, MAC, RLC, PDCP, and RRC, was taken as a basis for this investigation. The investigation analyzed the possible split points across the protocol stack, these possible split points are depicted in Figure 4 as different enumerated Options.

After detailed comparison 3GPP decided to take Option 2 (based on centralised PDCP/RRC and decentralised RLC/MAC/PHY) as a basis for normative specification work. The prime reason for selecting this option was the close similarity to the protocol stack split applied in Dual Connectivity: in a DC configuration the Master Node (MN) and the Secondary Node (SN) are “split” along the same point as Option 2.

3.1 Higher Layer Split (HLS) of the gNB

The overall NG-RAN architecture with a split gNB is shown in Figure 5 below.

As shown in Figure 5, in NG-RAN a set of gNBs is connected to the 5G Core Network (5GC) through the NG interface, and they can be interconnected through the Xn interface.

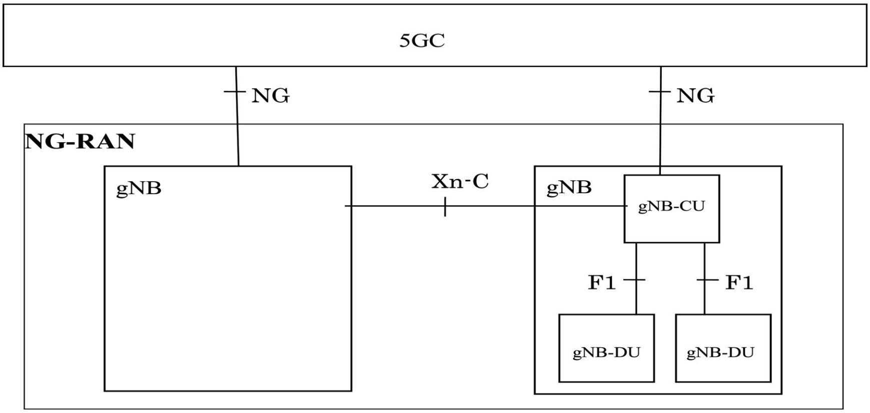


Figure 5 Higher Layer split of the gNB.

A gNB may then consist of a gNB-CU and one or more gNB-DU(s), and the interface between gNB-CU and gNB-DU is called F1. The NG and Xn-C interfaces for a gNB terminate in the gNB-CU. The maximum number of gNB-DUs connected to a gNB-CU is only limited by implementation. In 3GPP standard, one gNB-DU connects to only one gNB-CU, but implementations that allow multiple gNB-CUs to connect to a single gNB-DU e.g. for added resiliency, are not precluded. One gNB-DU may support one or more cells. The internal structure of the gNB is not visible to the core network and other RAN nodes, so the gNB-CU and connected gNB-DUs are only visible to other gNBs and the 5GC as a gNB.

The F1 interface supports signaling exchange and data transmission between the endpoints, separates Radio Network Layer and Transport Network Layer, and enables the exchange of UE-associated and non-UE-associated signaling. In addition, F1 interface functions are divided into F1-C and F1-U functions.

F1-C (Control Plane) Functions:

- F1 Interface Management Functions: These consist of F1 setup, gNB-CU Configuration Update, gNB-DU Configuration Update, error indication and reset function.
- System Information Management Functions: The gNB-DU is responsible for the scheduling and broadcasting of system information. For system information broadcasting, the encoding of NR-MIB and SIB1 is performed by the gNB-DU, while the encoding of other SI messages is performed by the gNB-CU. The F1 interface also provides

signaling support for on-demand SI delivery, enabling UE energy saving.

- F1 UE Context Management Functions: These functions are responsible for the establishment and modification of the necessary UE context. The establishment of the F1 UE context is initiated by the gNB-CU, and the gNB-DU can accept or reject the establishment based on admission control criteria (e.g., the gNB-DU can reject a context setup or modification request in case resources are not available). In addition, an F1 UE context modification request can be initiated by either gNB-CU or gNB-DU. The receiving node may accept or reject the modification. The F1 UE context management function can be also used to establish, modify and release Data Radio Bearers (DRBs) and Signaling Radio Bearers (SRBs).
- RRC Message Transfer Function: This function is responsible for the transferring of RRC messages from the gNB-CU to the gNB-DU, and vice versa.

F1-U (User Plane) Functions:

- Transfer of User Data: This function allows to transfer user data between gNB-CU and gNB-DU.
- Flow Control Function: This function allows to control the downlink user data transmission towards the gNB-DU. Several functionalities are introduced for improved performance on data transmission, like fast retransmission of PDCP PDUs lost due to radio link outage, discarding redundant PDUs, the retransmitted data indication, and the status report.

The following connected-mode mobility scenarios are supported in the case of CU-DU split:

- Inter-gNB-DU Mobility: The UE moves from one gNB-DU to another within the same gNB-CU.
- Intra-gNB-DU inter-cell mobility: The UE moves from one cell to another within the same gNB-DU, supported by UE Context Modification (gNB-CU initiated) procedure.
- EN-DC Mobility with Inter-gNB-DU Mobility using MCG SRB: The UE moves from one gNB-DU to another within the same gNB-CU when only MCG SRB is available during EN-DC operation.
- EN-DC Mobility with Inter-gNB-DU Mobility using SCG SRB: The UE moves from one gNB-DU to another when SCG SRB is available during EN-DC operation.

3.2 Separation of CP and UP with Higher Layer Split (HLS)

To optimize the location of different RAN functions according to different scenarios and performance requirements, the gNB-CU can be further separated into its CP and UP parts (the gNB-CU-CP and gNB-CU-UP, respectively).

The interface between CU-CP and CU-UP is called E1 (purely a control plane interface). The overall RAN architecture with CU-CP and CU-UP separation is shown in Figure 6.

The gNB-CU-CP hosts the RRC and the control plane part of the PDCP protocol; it also terminates the E1 interface connected with the gNB-CU-UP and the F1-C interface connected with the gNB-DU. The gNB-CU-CP hosts the user plane part of the PDCP protocol of the gNB-CU for an en-gNB, and the user plane part of the PDCP protocol and the SDAP protocol of the gNB-CU for a gNB. The gNB-CU-UP terminates the E1 interface connected with the gNB-CU-CP and the F1-U interface connected with the gNB-DU.

A gNB may consist of a gNB-CU-CP, multiple gNB-CU-UPs, and multiple gNB-DUs. The gNB-CU-CP is connected to the gNB-DU through the F1-C interface, and gNB-CU-UP is connected to the gNB-DU through the F1-U interface. One gNB-CU-UP is connected to only one gNB-CU-CP, but implementations allowing a gNB-CU-UP to connect to multiple gNB-CU-CPs e.g. for added resiliency, are not precluded. One gNB-DU can be connected to multiple gNB-CU-UPs under the control of the same gNB-CU-CP. One gNB-CU-UP can be connected to multiple DUs under the control of the same gNB-CU-CP.

The basic functions of the E1 interface include E1 interface management function and E1 bearer context management function.

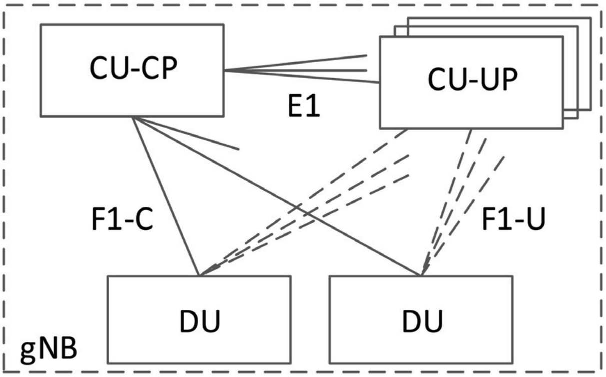


Figure 6 Overall RAN architecture with CU-CP and CU-UP separation.

4 NG-RAN Key Interfaces and Protocols

4.1 Xx Interface Family

NG-RAN nodes can be interconnected by means of the horizontal Xn and X2 interfaces, which are primarily used for three purposes: mobility (i.e. handover), multi-connectivity and SON (Self Optimized Networks). X2 was originally defined as the interface between two E-UTRAN nodes, later on extended to support EN-DC (i.e. as the interface between eNB and en-gNB) and will be further extended to support NR-DC. Xn follows similar design, interconnecting two gNBs.

Xn and X2 protocol stacks are similar. In the user plane, it relies on GTP-U running on top of UDP and IP. In the control plane, SCTP is used. This is illustrated in Figure 7:

The Xn-U/X2-U interface provides non-guaranteed delivery of user plane PDUs between two NG-RAN nodes to support dual/multi connectivity or mobility operation. Additionally, it supports the flow control function through Downlink Data Delivery Status procedure.

The Xn-C/X2-C interface uses Xn-AP/X2-AP protocols respectively for interface maintenance, mobility (handover, UE context retrieval, etc.) and dual/multi connectivity operation.

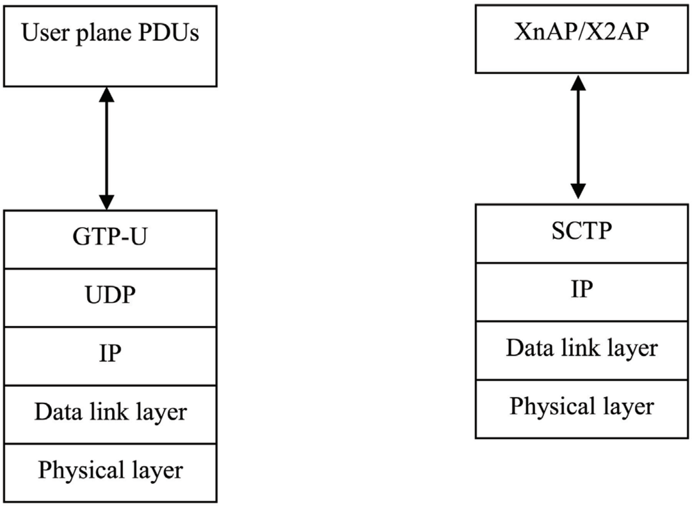


Figure 7 Xn and X2 protocol stacks.

4.2 NR Radio Interface Protocol

This section provides an overview of the architecture of the radio interface protocols that operate between NG-RAN and the UE, and then gives some details of the features of each protocol. The protocols will have some familiarity to those already knowledgeable of 4G LTE radio protocols and the description will identify some of the key differences and reasons for them.

The radio protocol architecture consists of a user plane, used for the transfer of the user data (IP packets) between the network and the UE, and a control plane that is used for control signalling between NG-RAN and the UE.

4.2.1 User plane

Figure 8 shows the user plane protocols stack within the UE and the gNB.

4.2.1.1 *Service data adaptation protocol (SDAP)*

The SDAP protocol is a notable difference in the user plane architecture compared to that of LTE, and it is introduced to support the new flow based QoS model of the 5G core network. With this new QoS model, the core network can configure different QoS requirements for different IP flows of a PDU session. The SDAP layer provides mapping of IP flows with different QoS requirements to radio bearers that are configured appropriately to deliver that required QoS. The mapping between IP flows and radio bearers may be configured and reconfigured by RRC signalling but it can also be changed more dynamically without the involvement of RRC signalling though a reflective mapping process.

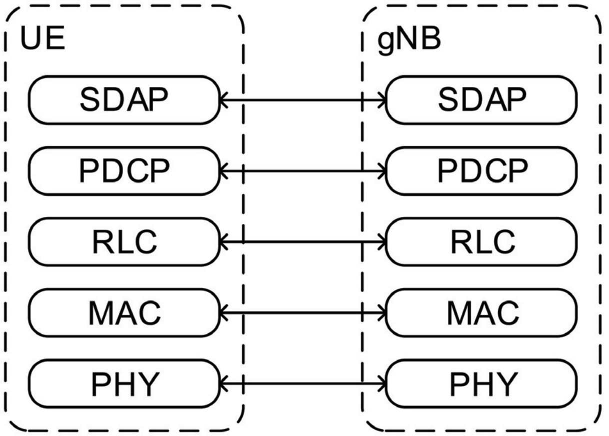


Figure 8 User plane protocol stack.

4.2.1.2 *Packet data convergence protocol (PDCP)*

The main functions of the PDCP protocol are to provide header compression and decompression through the use of RoHC (Robust Header Compression), security functions including ciphering/deciphering and integrity protection, duplication of transmitted PDCP PDUs, and reordering and duplicate detection of received PDCP PDUs. The most significant differences in NR PDCP compared to LTE are the introduction of the data duplication over different transmission paths in order to achieve extremely high reliability for URLLC (Ultra Reliable Low Latency) applications, and the introduction of integrity protection for user plane data.

4.2.1.3 *Radio link control protocol (RLC)*

NR RLC is very similar in functionality to LTE RLC, with the main functions being to provide segmentation, in order to match the transmitted PDU size to the available radio resources, and error correction through ARQ. One difference compared to LTE RLC is that it does not provide concatenation of RLC SDUs, with equivalent functionality now provided by the MAC layer, and does not provide reordering, with the protocol stack instead relying only on the reordering within PDCP.

4.2.1.4 *Medium access control (MAC)*

Similar to LTE MAC, the functionality provided includes multiplexing and demultiplexing of data from different radio bearers to the transport blocks that are carried by the physical layer, priority handling between data from different radio bearers, and error correction through Hybrid ARQ. A notable addition compared to LTE is that the MAC protocol carries control signalling used for the purpose of beam management within the physical layer.

4.2.2 Control plane

Figure 9 shows the control plane protocol stack. The Non Access Stratum (NAS) protocols terminate in the UE and the AMF of the 5G core network and are used for core network related functions such as registration, authentication, location updating and session management. The Radio Resource Control (RRC) protocol terminates in the UE and the 5G-RAN and is used for control and configuration of the radio related functions in the UE.

A significant difference in NR RRC compared to LTE RRC is the introduction of a 3-state model with the addition of the RRC INACTIVE state, as shown in the figure below. RRC Inactive provides a state with battery efficiency similar to RRC Idle but with a UE context remaining stored within

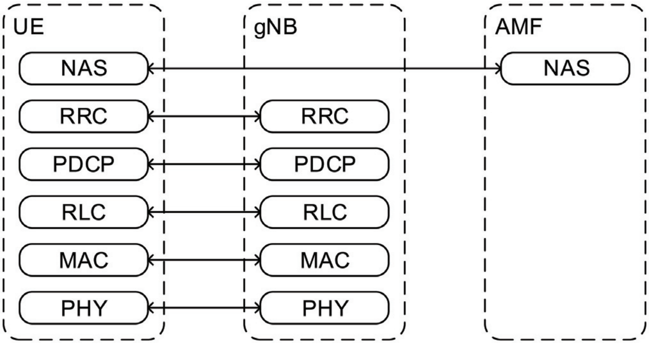


Figure 9 Control plane protocol stack.

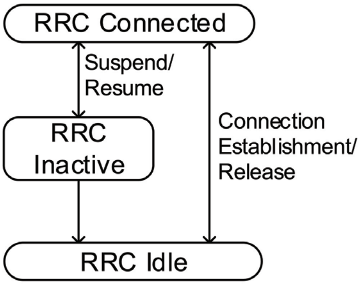


Figure 10 NR RRC state model.

the NG-RAN so that transitions to/from RRC Connected are faster and incur less signalling overhead. See Figure 10 above.

The other significant additions relative to LTE RRC are the support of an 'on demand' system information mechanism that enables the UE to request when specific system information is required instead of the NG-RAN consuming radio resources to provide frequent periodic system information broadcast, and the extension of the measurement reporting framework to support beam measurements for handover within a high frequency beam based deployment.

5 Summary

As described in this paper, the NG-RAN architecture builds on the success of the 4G LTE radio architecture, while introducing a number of key, revolutionary and forward looking concepts both on the overall architecture front as well as in protocols.

3GPP has taken several steps to specify interfaces and protocols that ease the migration of LTE-based cellular networks to 5G and NR. It is expected that these steps will help the uptake of NR and 5GC while making it easier to evolve networks in the most cost-efficient manner possible. Enhancements beyond phase-1 will address requirements and functions needed for industries beyond cellular mobile broadband: automated driving, industry automation, e-health services, etc. The 5G platform is promising to deliver the foundation for the next decade in the digital age.

Abbreviations

5GC	5G Core Network
AMF	Access and Mobility Management Function
CP	Control Plane
CU	Central Unit
DC	Dual Connectivity
DU	Distributed Unit
EN-DC	LTE-NR Dual Connectivity
E-UTRA	Evolved Universal Terrestrial Radio Access
MAC	Medium Access Control
NG-RAN	NG Radio Access Network
NGAP	NG Application Protocol
NR	NR Radio Access
PDCP	Packet Data Convergence Protocol
RRC	Radio Resource Control
RLC	Radio Link Control
SDAP	Service Data Adaptation Protocol
SMF	Session Management Function
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable and Low Latency Communications
XnAP	Xn Application Protocol
Xn-C	Xn-Control plane
Xn-U	Xn-User plane

Biographies

Balazs Bertenyi received an M.Sc. Degree in Computer Science and Telecommunications in 1998 at the Technical University of Budapest.

Balazs joined Nokia in 1998 and started to work on circuit switched mobile switching.

In 1999 he joined the research group on IMS (IP Multimedia Subsystem) and soon started to work in 3GPP standardization on IMS architecture.

- 2000–2003 Representative of Nokia in 3GPP SA2 (Architecture Working Group)
- 2003–2007 Head of delegation for Nokia in 3GPP SA2
- 2007–2011 Chairman of 3GPP SA2
- 2011–2015 Chairman of 3GPP TSG-SA (Technical Specification Group – Services and System Aspects)

In 2015 Balazs moved to the radio standardization group within Nokia to focus on 5G matters. He started attending TSG-RAN in 3GPP in 2015 covering 5G topics.

In 2017 Balazs was elected as Chairman of TSG-RAN for a 2-year term with a potential to be re-elected for one additional term.

Balazs has led various key projects for Nokia on Mobile Broadband architecture and standards strategy.

Richard Burbidge is a Senior Principal Engineer with Intel Corporation. With 25 years of experience in the wireless industry, he has been involved with the development of cellular standards since 1998, and has been a 3GPP delegate since its inception in 1999. Over this time he has contributed to 3 generations of cellular technology, firstly UMTS, then LTE, and now NR, 3GPP's 5G technology. He is currently serving as the chairman of the 3GPP TSG RAN 2 committee that is responsible for the specification of the layer 2 and 3 radio interface protocols for LTE and 5G/NR, having previously served as the vice chairman from 2005–2009. Richard received a Bachelor's degree in Electrical and Information Sciences from the University of Cambridge in 1993.

Gino Masini, MBA received his M.Sc. degree in Electronics Engineering from Politecnico di Milano and his MBA from SDA Bocconi School of Management in Milano, Italy. He started his career as a researcher in the Department of Electronics at Politecnico di Milano, working on microwave propagation and satellite telecommunications for the European Space Agency and the Italian Space Agency. He joined Ericsson in 1999, working at first on microwave antennas and network planning for microwave radio links, and later with MMIC design. Since 2009 he works with 4G and 5G radio network architecture, interfaces, and protocols. He is active in standardization since 2001: he attended ETSI, ITU, CEPT, and occasionally the Small Cell Forum.

He is active in 3GPP since 2009, and he is currently serving as 3GPP RAN WG3 Chairman. He has more than 30 patents granted, he authored or co-authored more than a dozen scientific publications, and he holds a “Six Sigma” certification.

Sasha Sirotkin is a senior wireless architect in the Next Generation and Standards group at Intel Corporation. Sasha has over 15 years of experience in wireless communications, in product development, system architecture, standardization and research. Sasha is currently serving as the vice chairman of 3GPP TSG RAN 3 committee that is responsible Radio Access Network architecture. Sasha holds B.Sc. degrees in Computer Science and Physics and M.Sc. degree in Applied Statistics and Astrophysics from Tel-Aviv University, Israel.

Yin Gao received the Master degree in Circuit and System from Xidian University, Xi’an, China, in 2005. She attended the Xidian University where she received his B.Sc. in Electronic Engineering in 2002. Since 2005 she has been with research center of ZTE and is engaged in the study of 3G/4G/5G technology, especially in both LTE and 5G RAN architecture and higher layer signaling procedures. From August, 2017, she was elected as 3GPP RAN3 Vice Chairman.

The 5G System Architecture

Frank Mademann

Chairman of 3GPP SA2, Huawei Technologies, Germany
E-mail: frank.mademann@huawei.com

Received 28 April 2018;
Accepted 3 May 2018

Abstract

This article provides an introduction to the 3GPP 5G system architecture and highlights its key features and characteristics.

Keywords: 3GPP, System Architecture, 5G.

1 Introduction

The milestone of defining the 3GPP 5G system architecture was achieved at the end of 2017. Within two years the 3GPP 5G architecture work progressed from the study period in 2016 to the delivery of a complete set of stage 2 level specifications. By achieving this milestone in 3GPP Release 15 the 5G system architecture has been defined – providing 3GPP's 5G phase 1 set of features and functionality needed for deploying a commercially operational 5G system. The overall 5G system architecture details features, functionalities and services including dynamic system behaviour defined by information flows.

This article offers a brief introduction to the 5G system architecture, highlighting some of its main characteristics. It was first published at 3GPP.org in December 2017. The complete description is provided by the 3GPP specifications TS 23.501 [1], TS 23.502 [2] and TS 23.503 [3].

Journal of ICT, Vol. 6_1&2, 77–86. *River Publishers*
doi: 10.13052/jicts2245-800X.615
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

These 5G stage 2 level specifications include the overall architecture model and principles, support of broadband data services, subscriber authentication and service usage authorization, application support in general, but also specifically support for applications closer to the radio as with edge computing. Its support for 3GPP's IP Multimedia Subsystem includes also emergency and regulatory services specifics. Further, the 5G system architecture model uniformly enables user services with different access systems, like fixed network access or interworked WLAN, from the onset. The system architecture provides interworking with and migration from 4G, network capability exposure and numerous other functionalities.

2 Service Based Architecture

Compared to previous generations the 3GPP 5G system architecture is service based. That means wherever suitable the architectural elements are defined as network functions that offer their services via interfaces of a common framework to any network functions that are permitted to make use of these provided services. Network Repository Functions (NRF) allow every network function to discover the services offered by other network functions. This architectural model, which further adopts principles like modularity, reusability and self-containment of network functions, is chosen to enable deployments to take advantage of the latest virtualization and software technologies. The related service based architecture figures of TS 23.501 [1] depict those service based principles by showing the network functions, primarily Core Network functions, with a single interconnect to the rest of the system. Reference point based architecture figures are also provided by the stage 2 specifications, which represent more specifically the interactions between network functions for providing system level functionality and to show inter-PLMN interconnection across various network functions. In the context of 3GPP specifications Public Land Mobile Network (PLMN) is typically denoting a network according to the 3GPP standard. The various architecture figures can be found in [1].

Figure 1 shows one of the service based architecture figures, which is for a roaming scenario with local breakout, i.e. the roaming UE interfaces the Data Network (DN) in the visited network (VPLMN) and the home network (HPLMN) enables it with subscription information from Unified Data Management (UDM), Authentication Server Function (AUSF) and UE specific policies from Policy Control Function (PCF). Network Slice Selection

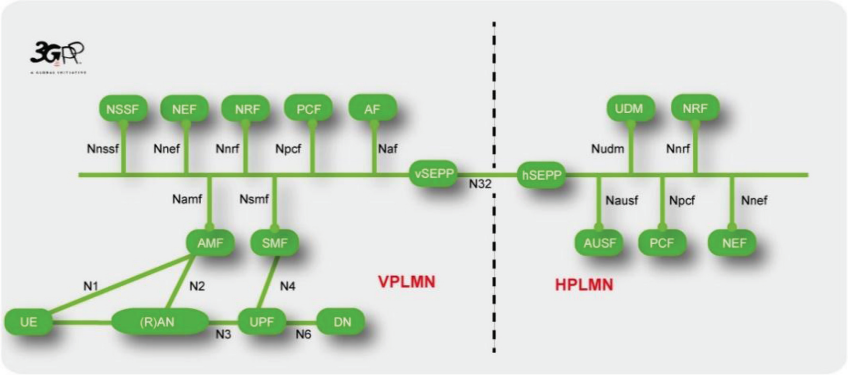


Figure 1 A service based architecture figure [1].

Function (NSSF), Access control and Mobility management Function (AMF), data Session Management Function (SMF) and Application Functions (AF) are provided by the VPLMN. The user plane provided via User Plane Functions (UPF) is managed following a model of control and user plane separation similar to what was already introduced in the latest 3GPP 4G release. Security Edge Protection Proxies (SEPP) protect the interactions between PLMNs. For more details and other scenarios see [1].

In the local breakout scenarios a UE receives the services of a PLMN typically completely from the serving operator's administrative domain. Home-routed data services are the alternative for roaming scenarios, which have also network functions from the home operator's administrative domain involved and the UE interfaces the DN in the HPLMN.

Service based principles apply between the control plane network functions of the Core Network. Further, the 5G system architecture allows network functions to store their contexts in Data Storage Functions (DSF). Functionality for releasing the UE specific Access Network – Core Network transport associations from one AMF and re-binding with another AMF enables separating such data storage also for the AMF. Earlier system architectures had more persistent UE specific transport associations, which made it more complex to change the UE's serving node that compares to an AMF. The new functionality simplifies changing the AMF instance that serves a UE. It also supports increasing AMF resilience and load balancing as every AMF from a set of AMFs deployed for the same network slice can handle procedures of any UE served by the set of AMFs.

3 Common Core Network

The generalised design of the functionalities and a forward compatible Access Network – Core Network interface enable the 5G common Core Network to operate with different Access Networks. In 3GPP Release 15 these are the 3GPP defined NG-RAN and the 3GPP defined untrusted WLAN access. Studies on other access systems that may be used in future releases started already. The 5G system architecture allows for serving both Access Networks by the same AMF and thereby also for seamless mobility between those 3GPP and non-3GPP accesses. The separated authentication function together with a unified authentication framework are for enabling customization of the user authentication according to the needs of the different usage scenarios, e.g. using different authentication procedures per network slice. Most of the other 5G system architecture functionality introduced by this article is common for different Access Networks. Some functionality provides variants that are more suitable for specific Access Networks, like certain Quality of Service (QoS) functionality described later.

4 Network Slicing

A distinct key feature of the 5G system architecture is network slicing. The previous generation supported certain aspects of this with the functionality for dedicated Core Networks. Compared to this 5G network slicing is a more powerful concept and includes the whole PLMN. Within the scope of the 3GPP 5G system architecture a network slice refers to the set of 3GPP defined features and functionalities that together form a complete PLMN for providing services to UEs. Network slicing allows for controlled composition of a PLMN from the specified network functions with their specifics and provided services that are required for a specific usage scenario.

Earlier system architectures enabled typically rather a single deployment of a PLMN to provide all features, capabilities and services required for all wanted usage scenarios. Much of the capabilities and features provided by the single, common deployment was in fact required for only a subset of the PLMN's users/UEs. Network slicing enables the network operator to deploy multiple, independent PLMNs where each is customized by instantiating only the features, capabilities and services required to satisfy the subset of the served users/UEs or a related business customer needs.

The very abstract representation in Figure 2 shows an example of a PLMN deploying four network slices. Each includes all what is necessary to form a

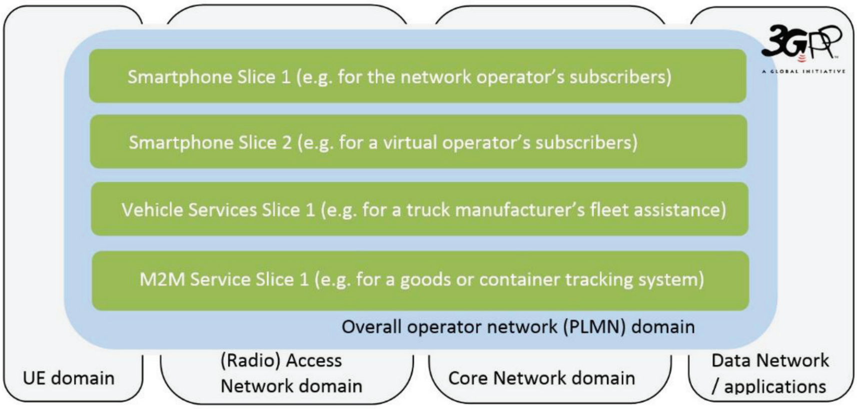

Smartphone Slice 1 (e.g. for the network operator's subscribers)
Smartphone Slice 2 (e.g. for a virtual operator's subscribers)
Vehicle Services Slice 1 (e.g. for a truck manufacturer's fleet assistance)
M2M Service Slice 1 (e.g. for a goods or container tracking system)
Overall operator network (PLMN)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

Figure 2 Abstract representation of a network deploying network slices.

complete PLMN. The two network slices for smart phones demonstrate that an operator may deploy multiple network slices with exactly the same system features, capabilities and services, but dedicated to different business segments and therefore each possibly providing different capacity for number of UEs and data traffic. The other slices present that there can be differentiation between network slices also by the provided system features, capabilities and services. The M2M network slice could, for example, offer UE battery power saving features unsuitable for smartphone slices, as those features imply latencies not acceptable for typical smart phone usages.

The service based architecture together with softwarization and virtualization provides the agility enabling an operator to respond to customer needs quickly. Dedicated and customized network slices can be deployed with the functions, features, availability and capacity as needed. Typically, such deployments will be based on a service level agreement. Further, an operator may benefit by applying virtualization, platforms and management infrastructure commonly for 3GPP-specific and for other network capabilities not defined by 3GPP, but that a network operator may need or want to deploy in his network or administrative domain. This allows for a flexible assignment of the same resources as needs and priorities change over time.

Deployments of both the smaller scope of the 3GPP defined functionality and the larger scope of all that is deployed within an operator's administrative domain are both commonly termed a "network". Because of this ambiguity and as the term "slicing" is used in industry and academia for slicing of virtually any kind of (network) resources, it is important to emphasize that the 3GPP system architecture specifications define network slicing only within the scope of 3GPP specified resources, i.e. that what specifically composes a PLMN. This doesn't hinder a PLMN network slice deployment from using e.g. sliced transport network resources. Please note, however, that the latter is fully independent of the scope of the 3GPP system architecture description. Pursuing the example further, PLMN slices can be deployed with as well as without sliced transport network resources.

Figure 3 presents more specifics of 3GPP network slicing. In that figure, network slice #3 is a straightforward deployment where all network functions serve a single network slice only. The figure also shows how a UE receives service from multiple network slices, #1 and #2. In such deployments there are network functions in common for a set of slices, including the AMF and the related policy control (PCF) and network function services repository (NRF). This is because there is a single access control and mobility management instance per UE that is responsible for all services of a UE. The user plane services, specifically the data services, can be obtained via multiple, separate network slices. In the figure, slice #1 provides the UE with data services for Data Network #1, and slice #2 for Data Network #2. Those slices and the data services are independent of each other apart from interaction with common access and mobility control that applies for all services of the user/UE. This makes it possible to tailor each slice for e.g. different QoS data services or different application functions, all determined by means of the policy control framework.

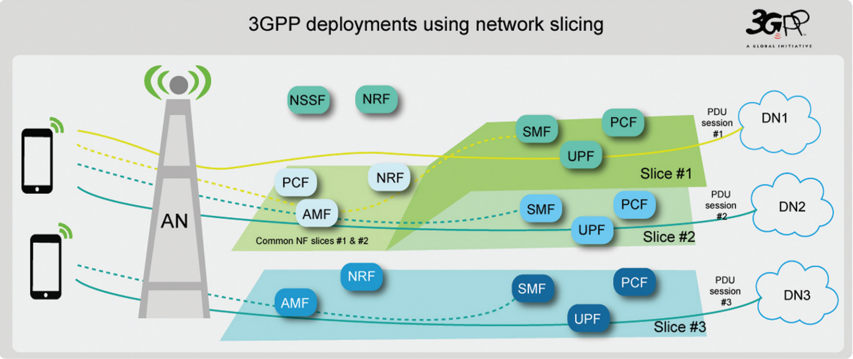


Figure 3 Network functions composing network slices.

5 Application Support

Data services are the basis of the application support. Compared to earlier generations data services offer considerably more flexibility for customization. A main part of this is the new QoS model of the 3GPP 5G system architecture, shown in Figure 4, which enables differentiated data services to support diverse application requirements while using radio resources efficiently. Service Data Flows (SDF) denote user plane data that certain QoS rules apply to. The actual description of SDFs is using SDF templates. Further, the QoS model is designed to support different Access Networks, including fixed accesses where QoS without extra signaling may be desirable. Standardized packet marking informs QoS enforcement functions what QoS to provide without any QoS signaling. While the option with QoS signaling offers more flexibility and QoS granularity. Furthermore, symmetric QoS differentiation over downlink and uplink is supported with minimal control plane signaling by the newly introduced Reflective QoS.

A large part of the functionality providing data connectivity is for supporting flexible deployment of application functions in the network topology as needed for edge computing, which is supported, for example, via three different Session and Service Continuity (SSC) modes or via the functionality of Uplink Classifiers and Branching Points.

The different SSC modes are sketched in Figure 5. The SSC modes include the more traditional mode (SSC 1), where the IP anchor remains stable to

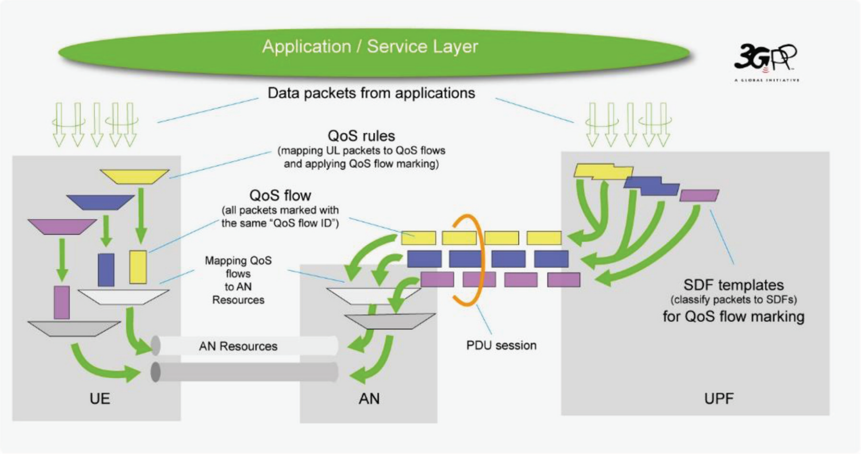


Figure 4 QoS model.

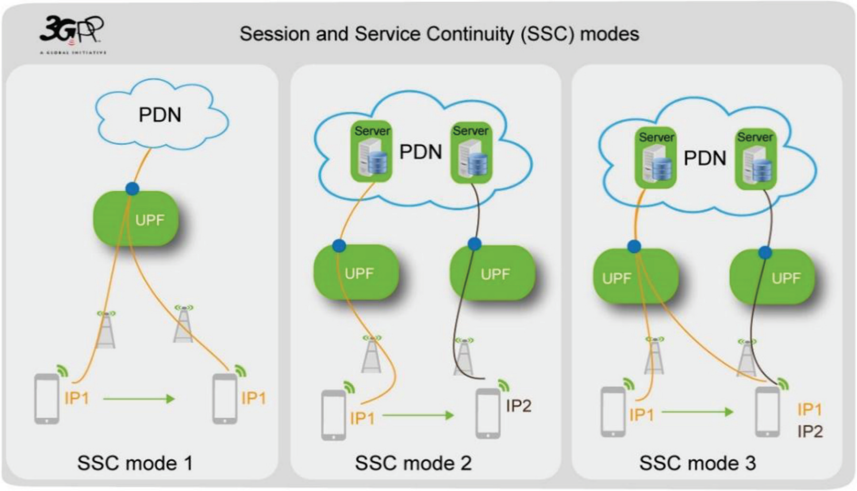


Figure 5 Session and Service Continuity modes.

provide continual support of applications and maintenance of the path towards the UE as its location is updated. The new modes allow for relocating the IP anchor. There are two options, make-before-break (SSC mode 3) and break-before-make (SSC mode 2). The architecture enables applications to influence selection of suitable data service characteristics and SSC mode.

As 5G network deployments are expected to serve huge amounts of mobile data traffic, an efficient user plane path management is essential. The system architecture defines in addition to the SSC modes the functionality of Uplink Classifiers and Branching Points to allow for breaking out and injecting traffic selectively to and from application functions on the user plane path before the IP anchor. Also, as permitted by policies, application functions may coordinate with the network by providing information relevant for optimizing the traffic route or may subscribe to 5G system events that may be relevant for applications.

6 Continuation of the work

The delivered stage 2 level specifications define the 3GPP 5G system from an overall, architectural perspective. The related work in the RAN, security, OAM and CT working groups continued with some specific stage 2 level aspects and with delivering stage 3 level specifications until June 2018.

This article has highlighted some of the most important advances of the 3GPP system architecture introduced with Phase 1 of 5G. Further advances and enhancements will be introduced in coming releases. Studies concerning Phase 2 functionality of 5G have already begun.

Specification work in 3GPP is a continuous process. More and up-to-date information can be found at 3GPP.org.

References

[1] 3GPP TS 23.501 – System Architecture for the 5G System; Stage 2.
[2] 3GPP TS 23.502 – Procedures for the 5G System; Stage 2.
[3] 3GPP TS 23.503 – Policy and Charging Control Framework for the 5G System; Stage 2.

Biography

Frank Mademann started his career with research and development on GSM circuit switched data services in 1991. This was also his initial work in ETSI SMG, which changed with the begin of standardization for GPRS. Since then he was involved in design and definition of all packet domain architectures and services that were specified by SMG and 3GPP. This includes GPRS from the very beginning as well as the packet domains of UMTS and LTE/SAE.

Frank has been working in the telecom and mobile industries for more than twenty years. He has been actively involved in the Architecture Working Group of 3GPP since 1999, where he is recognized as a key contributor to technical aspects of all packet domain architectures that were specified by SMG and 3GPP and also contributing to leadership and organizational matters. Having held earlier the position of a Vice Chair, he is the Chairman of 3GPP’s Architecture Working Group since 2015.

Path to 5G: A Control Plane Perspective

Erik Guttman[1] and Irfan Ali[2]

[1]*Chairman of 3GPP CT, Samsung Electronics, Germany*
[2]*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
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

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

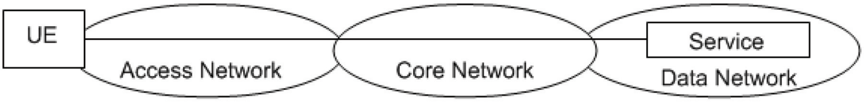


Figure 1 A simple model of service access using the 3GPP system.

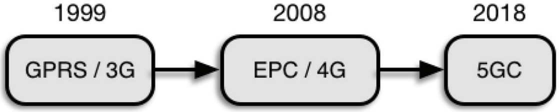


Figure 2 Core network evolution through generations.

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.

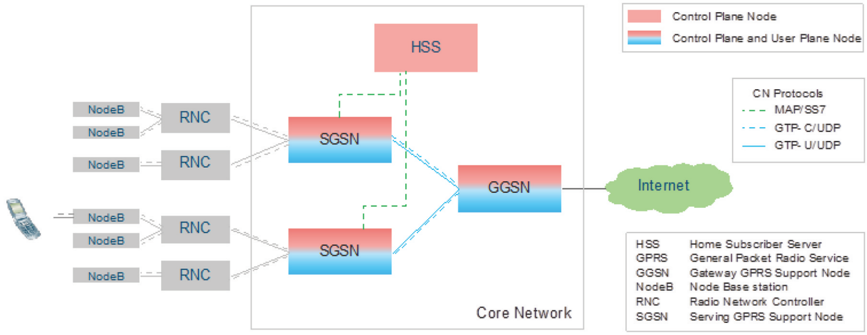

Control Plane Node
Control Plane and User Plane Node
HSS
NodeB
NodeB
NodeB
NodeB
NodeB
NodeB
RNC
RNC
RNC
RNC
SGSN
SGSN
GGSN
Internet
CN Protocols
MAP/SS7
GTP-C/UDP
GTP-U/UDP
HSS Home Subscriber Server
GPRS General Packet Radio Service
GGSN Gateway GPRS Support Node
NodeB Node Base station
RNC Radio Network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

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).

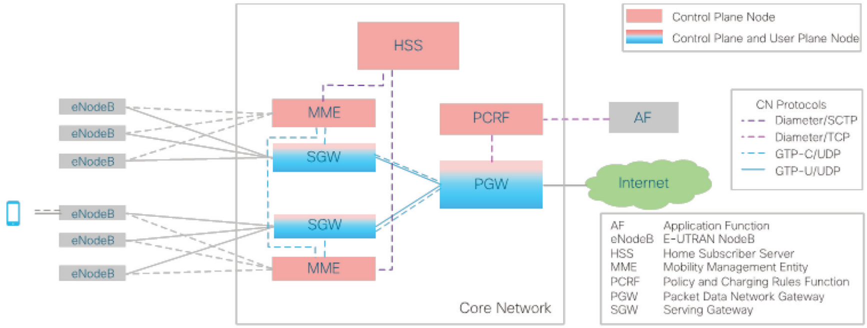


Figure 4 EPC – the 4G core network.

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

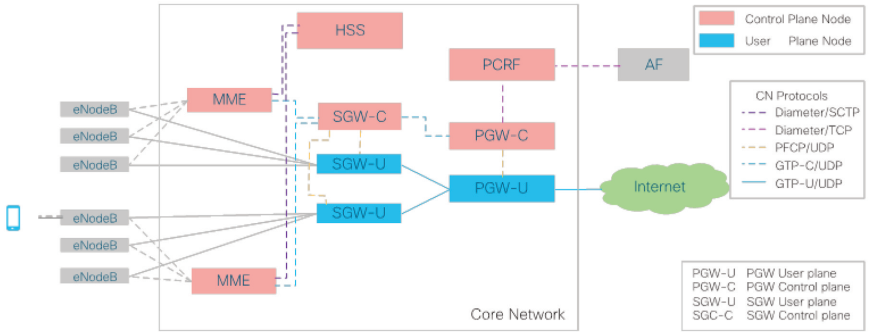
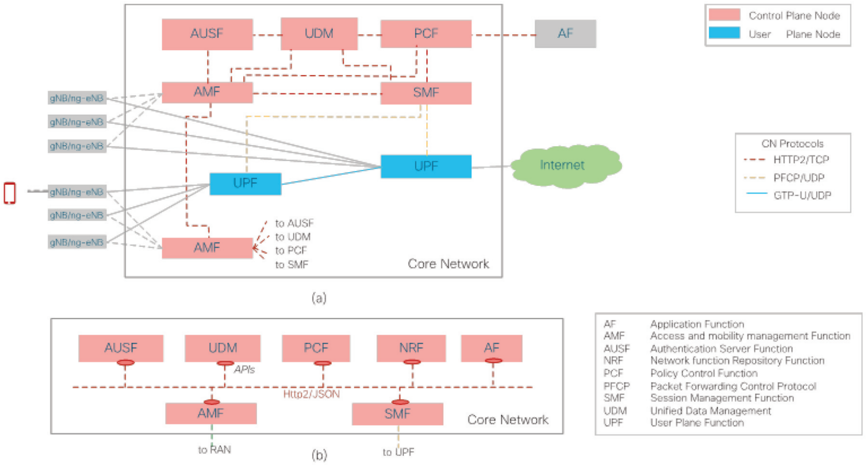


Figure 5 EPC with control plane user plane separation enhancement.



(a)

(b)

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.

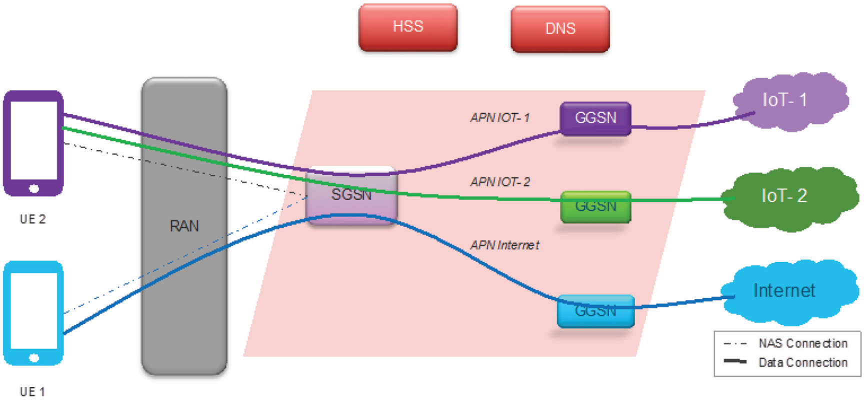


Figure 7 Use of APNs for selection of GGSN in GPRS.

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.

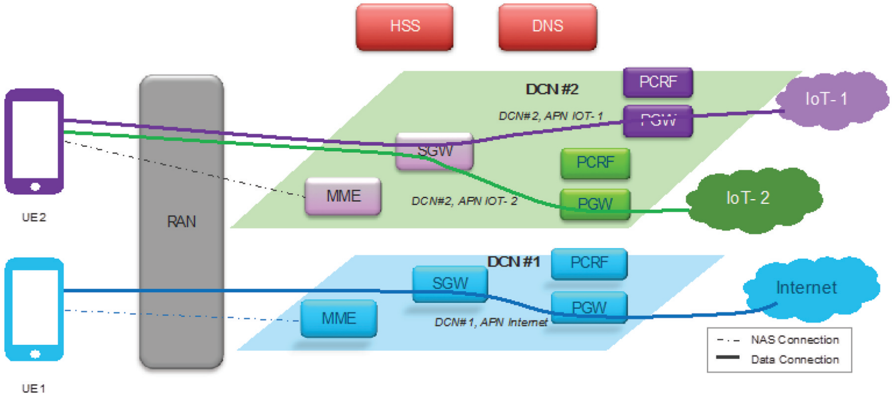


Figure 8 Dedicated core networks (DCNs) for EPS.

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.

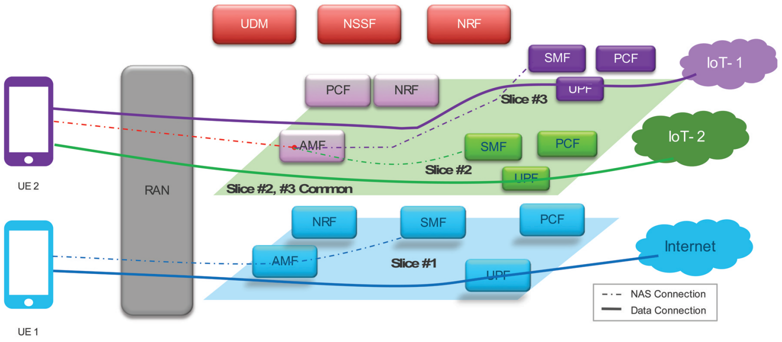


Figure 9 Network slicing applied to 5GS.

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.

RESTful APIs for the 5G Service Based Architecture

Georg Mayer

Chairman of 3GPP CT, Huawei, Vienna, Austria
E-mail: georg.mayer.huawei@gmx.com

Received 31 March 2018;
Accepted 3 May 2018

Abstract

5G sets out to be the global connectivity and integration platform for a broad variety of industries in the upcoming decade. In order to do so it not only needs to fulfil the requirements of these industries but must also ensure its tight integration into the digital infrastructure of the 2020s by embracing key technologies. This article shows how one of these key technologies, the RESTful design of Application Programming Interfaces (APIs), is used in the 5G Service Based Architecture (SBA). The basic principles of modern API development are explained and it is shown how those integrate into the specific needs of the 5G Core Network.

Keywords: 5G, REST, application programming interfaces, 3GPP, service based architecture, SBA, northbound APIs, NAPS, HTTP.

1 Introduction

5G aims to become the global connectivity enabler and service platform for a broad variety of industries in the upcoming decade and beyond. To achieve this goal the 5G system not only opens up towards a whole new group of customers, the so-called verticals, it also embraces the technologies which are and will be used by these verticals and more commonly in the digital landscape of the 2020s.

Journal of ICT, Vol. 6_1&2, 101–116. *River Publishers*
doi: 10.13052/jicts2245-800X.617
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

Technologies and concepts such as virtualization, cloud computing, internet of things, functionality exposure and self-organizing networks are central building blocks of 5G and will allow seamless communication as well as enable synergies between different industries.

For the purpose of exposure of functionality to 3rd parties as well as other types of system internal communication 3GPP chose to make use of the widely established REST architecture design paradigm, which describes the design of distributed applications and more specifically of Application Programming Interfaces (APIs). This article explains why the REST paradigm was chosen for certain aspects of the 5G system and how the different REST principles are applied in the 5G Service Based Architecture.

The so-called RESTful APIs can be understood as an example of 3GPP's commitment to tightly integrate 5G in the current and upcoming digital ecosystem of their customer.

2 New Communication for the Mobile Core

Capability exposure, i.e. making 5G Core Network functionalities available to 3rd parties such as service providers and vertical industries outside the operator's domain, is provided by the Network Exposure Function (NEF). The interface provided by the NEF to 3rd parties can be regarded as one of the essential membranes through which 5G communicates more closely towards vertical industries than mobile networks of earlier generations did. It was therefore a key requirement that 3GPP defines this interface in way that it would fully align with widely accepted and future proof principles for the design of such exposure interfaces.

Exposure of functionality is a common concept used by modern software design, especially for web-services which are offered over the internet. The use of APIs for this purpose is practically without competition and is applied from simple temperature sensors in automated home environments to large-scale cloud providers enabling near real-time content access. All these different APIs are defined along a number of common principles which are referred to as REST architectural style, which will be described in the following sections.

3GPP decided that 5G service exposure by the NEF should be based on RESTful APIs, as shown in Figure 1.

Nevertheless, the APIs offered to 3rd parties, also known as "northbound APIs" are only applicable to a single interface of the 5G system, whilst the NEF is one of many Network Functions within the completely redesigned 5G Core

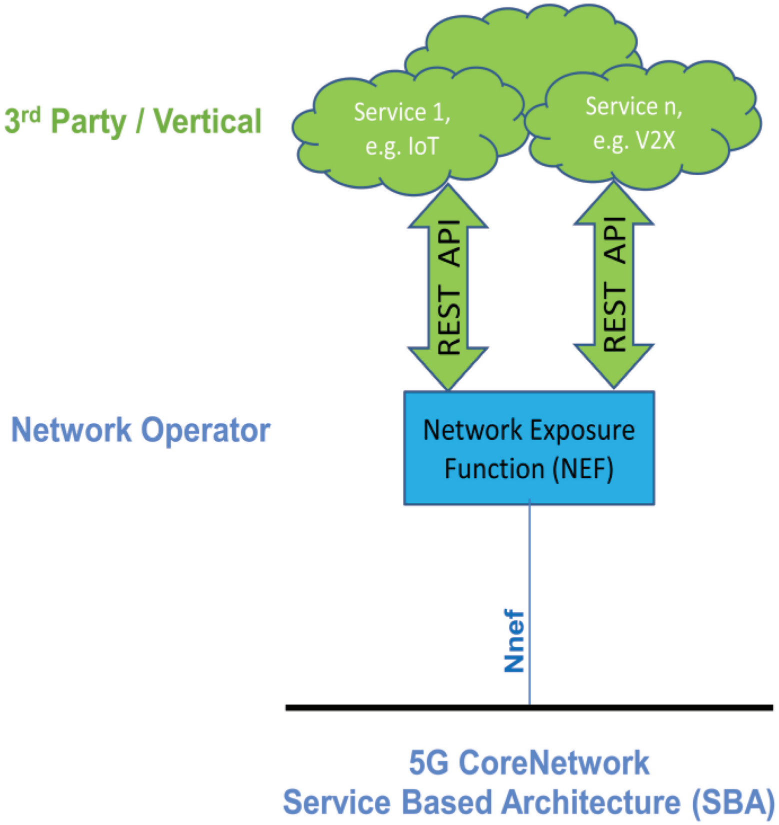


Figure 1 NEF (as part of the 5G SBA) providing services to 3rd parties via RESTful APIs.

Network. This redesigned core is the architectural and technical realization of the service-based change design of the 5G system and therefore was named Service Based Architecture (SBA). The Network Functions (NFs) forming the SBA communicate with each other via Service Based Interfaces (SBI), as shown in Figure 2.

3GPP took the forward-looking decision to use RESTful APIs not only for 3rd party functionality exposure but also for via the SBIs. Therefore the 5G Core Network internal communication obeys the same principles as the functional exposure, thus allowing a harmonized and holistic technological approach of the complete 5G system, fully in-line with the progressive paradigms which are at the heart of a wide range of services used by end-customers as well as for the automation of whole industries.

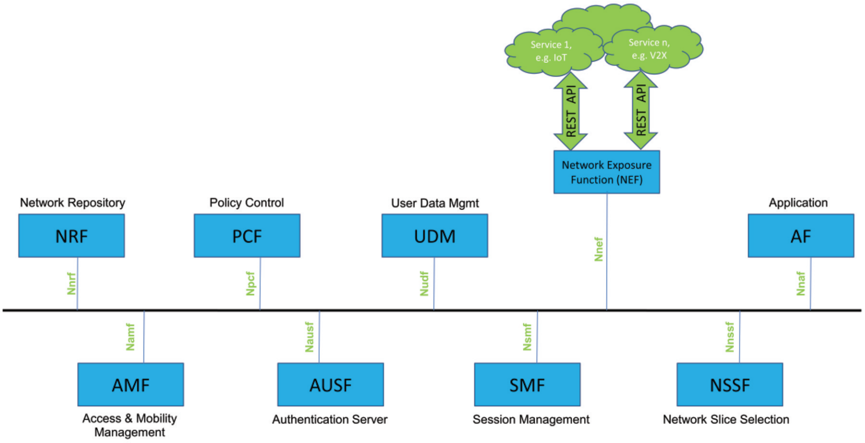


Figure 2 RESTful APIs for the service based interfaces and northbound communication.

But 3GPP didn't stop there. Once this decision of using RESTful APIs over SBI was taken, the CT4 Working Group came up with 3GPP TS 29.501 [5] which states guidelines for API creation within 3GPP. These guidelines are now not only used for northbound APIs and SBA but will also be used for e.g. the orchestration APIs. Other 5G functions are expected to be aligned to these principles during upcoming 3GPP releases.

The use of RESTful APIs throughout the system perfectly exemplifies how 5G sets out to become an open and integrated communication enabler for the technological convergence foreseen in the 2020s. 3rd party services have already widely adopted RESTful paradigms and with 5G these services will be enabled to seamlessly communicate first individually and subsequently also amongst each other. Thereby the choices taken by 3GPP will set free currently unforeseen synergies amongst services and sectors.

3 The RESTful Ecosystem

Roy Fielding described what he called the REST architectural style in his dissertation [1] which was published in the year 2000. REST stands for *REpresentational State Transfer* and is not a protocol or description language, it is also not a specific architecture. It is usually described as a set of principles or paradigm. Whilst this view is correct, the term RESTful is nowadays used not only for the related principles themselves but also for deployed applications and software environments following these principles.

In chapter 6 of his dissertation Fielding describes in detail how the principles of REST can be used within the World Wide Web, i.e. by making use of Uniform Resource Indicators (URIs), the Hypertext Transfer Protocol (HTTP), different data description languages and how such technologies can be used in a "RESTful" way for real world deployments.

In the 18 years since the publication of Fielding's dissertation, the REST paradigm has fundamentally re-shaped the way how software applications are designed, implemented and deployed. It is used throughout the IT industry, there exist countless tools as well as books, articles and web pages to support its use and a huge developer community is experienced with REST principles. The paradigm itself as well as the related technologies, protocols and tools are further developed not only by software companies and universities, but also by the open source community and by global standards organizations such as the W3C (World Wide Web Consortium) and IETF (Internet Engineering Task Force) [1].

REST has proven to be a reliable and future proof way for developing distributed applications. It's therefore safe to say that there is thriving ecosystem which is built on the REST principles.

4 Example RESTful SBA Procedures

This section describes an example scenario consisting of three API calls within the 5G SBA which are meant to exemplify how RESTful principles are used by 3GPP. The given examples are not complete and were only chosen to give the reader an initial overview. Further information can be found in the given references. Section 5 will explain how the REST principles are applied in 5G SBA, based on the examples of this section.

4.1 Example Scenario

The functional split chosen for 5G SBA Network Functions includes e.g. an Access and Mobility Management Function (AMF), which serves as the single-entry point for a user equipment (UE) for all its communication. Once the user decides to use one of the services, e.g. to browse the web, the AMF needs to assign a Session Management Function (SMF) which manages the users session context. As in 5G virtual network functions (VNF) can be instantiated and deleted at any time, the AMF first needs to discover an available and suitable SMF, which is achieved via the Service Discovery procedure performed between the AMF and the Network Repository Function

(NRF). To allow for successful discovery, the SMF must have registered beforehand with the NRF.

Thus we look at a dedicated example, consisting of three different procedures, which are depicted in Figure 3.

1. *Service Registration*: the SMF registers the services it provides with the NRF.
2. *Service Discovery*: the AMF queries the NRF for a suitable SMF and in return receives the address of the SMF which registered in step 1.
3. *Session Establishment*: the requested session is established on the control plane level by the AMF via the SMF.

3GPP TS 23.501 [2] defines the roles of service consumer and service producer. The service consumer is the NF which requests the related service, whilst the NF which exposes the requested service is the service producer. Services exposed by a NF are further structured into service operations, as defined in 3GPP TS 23.501 [2] and 3GPP TS 29.501 [5].

The basic procedures as outline here are defined in 3GPP TS 23.502 [3], whilst the more detailed flows and protocol elements which are exchanged are specified in the related NF specifications, i.e. for the NRF in 3GPP TS 29.510 [7] and for the SMF in 3GPP TS 29.502 [6].

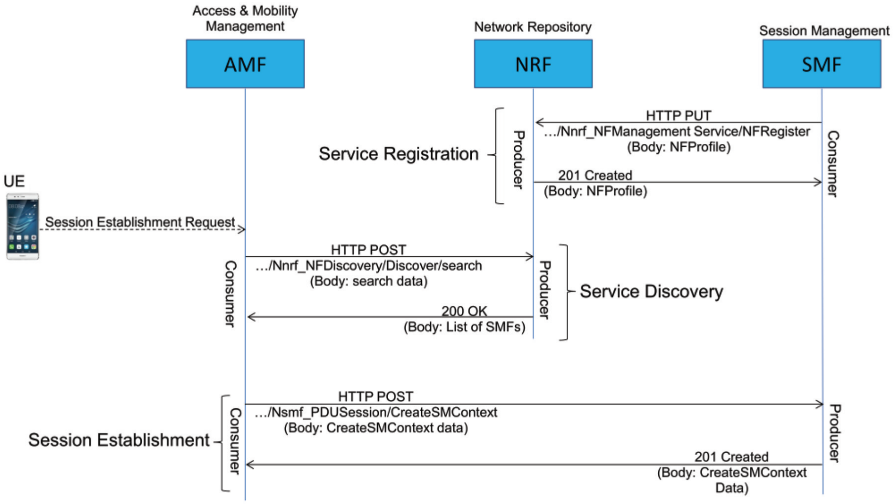


Figure 3 Simplified API calls for example SBA procedures.

4.2 Service Registration

During Service Registration the SMF takes the role of the service consumer and the NRF takes the role of the service producer, exposing the Nnrf_NFManagement service with the NFRegister service operation (see 3GPP TS 29.510 [7] and TS 23.501 [2]).

The SMF sends a HTTP PUT request towards the NRF. The Uniform Resource Locator (URI) addresses the profile information which is intended to be created at the NRF, e.g. "`https://nrf7.slice-v2x.opx.3gpp/nrf-nfm/v1/nf-instances/smf5-slicev2x`". Note that the URI directly addresses the Service Profile of the SMF as it will be stored at the NRF, i.e. the URI is not just the address of the NRF but of the resource which it is requested to create and host.

Within the Service Registration procedure the SMF sends the NFProfile information, which is encoded as a structured data type in a JSON (short for JavaScript Object Notation, as defined by RFC 8259 [10] and ECMA-404 [11]) document in the body of a HTTP PUT request. This NFProfile document includes information about the SMF, such as

- the Type of NF (in this case the SMF);
- the network identification (PLMN ID) and slice identifiers (S-NASSI, NSI ID) to which the SMF belongs to;
- the address or addresses (IP Address or FQDN) of the NF (e.g. "`https://smf5.slice-v2x.opx.3gpp`");
- a list of service names which the NF supports (e.g. "Nsmf_PDUSession" and "Nsmf_EventExposure" in the case of SMF);
- for each service name a list of the supported service operations (e.g. for the Nsmf_PDUSession the service operations "Create SM Context", "Update SM Context", "Release SM Context", "Notify SM Context Status", "Query SM Context", etc).

Once the NRF receives the HTTP PUT request it authenticates and verifies it. Assuming the received HTTP PUT request passes all tests the NRF then uses the URI, i.e. the address to which the HTTP PUT request is sent to, to create a new local resource, including the SMF NFProfile.

It then acknowledges the creation of the resource by returning a HTTP 201 Created response, in which the NFProfile is included again in the body.

Due to the Service Registration procedure a new resource with the unique address "`https://nrf7.slice-v2x.opx.3gpp/nrf-nfm/v1/nf-instances/smf5-slicev2x`" is created at the NRF. This resource includes the above described structured information about the SMF.

4.3 Service Discovery

During Service Discovery the AMF takes the role of the service consumer and the NRF takes the role of the service producer, exposing the Nnrf_NFDiscovery service with the NFDiscover service operation (see 3GPP TS 29.510 [7]).

The AMF sends a HTTP POST request to the NRF, querying for a list of SMFs which registered with a specific set of supported services. The list of services which the SMF's need to support are included as a structured data type (JSON document) in the body of the HTTP POST request.

The NRF authenticates and validates the incoming HTTP POST request. Once it has passed all checks the NRF searches its local resources, i.e. the registered and stored profiles of NFs, for matches against the query list received from the AMF.

The list of matches is sent from the NRF to the AMF in the body of the 200 OK response to the HTTP POST request.

4.4 Session Establishment (AMF Creates Session Context at SMF)

Once the AMF receives the response it chooses "`https://smf5.slice-v2x.opx.3gpp`" as the SMF for the intended session establishment.

During Session Establishment the AMF takes the role of the service consumer and the SMF takes the role of the service producer, exposing the Nsmf_PDUSession service with the Create SM Context service operation (see 3GPP TS 29.502 [6]).

The AMF sends a HTTP POST request to the SMF, addressing the exposed service of Session Establishment which was found in the list of supported services, i.e. "`https://smf5.slice-v2x.opx.3gpp/nsmf-pdusession/v1/sm-contexts`". This indicates that the AMF wants to create a Session Management context at the SMF. The details of the context under creation are included as the SmContextCreateData structured data type in a JSON document within the body of the HTTP POST request. The SmContextCreateData information includes all information necessary to create a Session Management Context, e.g. the IDs of the requesting UE, slice identifiers and so forth.

The SMF authenticates and validates the incoming HTTP POST request. Once it has passed all checks the SMF, if it has the necessary resources available, creates the requested Session Management Context.

All information related to the created Session Management Context is then returned by the SMF to the AMF in the body of the 201 Created response to the HTTP POST request.

Note that especially for the case of Session Establishment this article only treats the very first SBA interaction between the AMF and the SMF. It leaves out all subsequent interactions e.g. with the Unified Data Management (UDM), the Policy Control Function (PCF) and any other NF. It also does not describe interactions between the SMF and a User Plane Function (UPF). More details can be found in 3GPP TS 23.502 [3] Section 4.3.2.2.

5 REST Principles and Technologies Adopted by 5G SBA

At the core of RESTful service architecture design and service development as described by Fielding [1] lay six principles which all aim to make the creation and deployment of distributed services flexible, coherent and scalable.

This section first explains a number of essential terminologies used to describe REST principles, then details all six principles and evaluates how they are implemented in 5G SBA. Finally this section gives a short overview how HTTP is used within RESTful deployments and 5G SBA specifically.

5.1 Resource, Serialization and Representation

In the context of REST, a **resource**, e.g. the SMF Service Profile as used during Service Registration, can be stored in any form on a server (here: NRF) and can also take on different internal states. It only exists under a unique Uniform Resource Location (URI), which is used to address it for different purposes – e.g. creation, modification and deletion.

During these procedures it is often necessary to exchange either all or part of the information related to the resource between a client (here: SMF) and a server (here: NRF). To allow this exchange the resource gets **serialized**, i.e. all information related to it is written into a document. In the case of 5G SBA objects are serialized as JSON documents. These documents are individual **representations** of the resource, i.e. they are not the resource themselves, which still is only available via the URI, but only a snapshot of it.

5.2 Client/Server Principle

Client/Server is the REST principle which demands that a dedicated split of responsibilities between the entity which requests a resource, the client, and the entity which provides access to the resource, the server.

In the 5G SBA example Service Registration procedure outlined in the previous section the SMF acts as a client towards the NRF during Service

Registration, whilst the NRF acts as a server. 3GPP alternatively uses the terms service consumer for the client/SMF and service producer for the server/NRF.

Also the Service Discovery example shows the clear split between client (AMF, service consumer) and server (NRF, service producer).

5.3 Stateless Principle

Stateless is the REST principle which demands that the server does not keep any state related information concerning the communication with specific clients. This in turn means that the requests from the client need to include all necessary information which enables the server to perform the desired service. This not only frees resources at the server, but also allows for load balancing and resilience in a distributed environment, as requested services can be made available by a set of servers, to which load can be distributed on a per-request basis.

In the example scenarios both the NRF as well as the SMF, when acting as service producers, are simply returning the requested information to the respective clients, e.g. the AMF. Neither NRF nor SMF keep any additional state about the communication with the AMF.

Going further, in order to achieve full statelessness it is also necessary that all information that is assumed to be "local" on the server is indeed kept within a storage unit (database server), which is accessible from all servers offering a particular service. In the example scenarios the service profiles are still stored locally at the NRF after the example Service Registration procedure. This means that access to this Service Profile can only be provided via a single NRF, i.e. "nrf7.slice-v2x.opx.3gpp", which acted upon the Service Registration request from the SMF. If this specific NRF would e.g. to be shut down, the stored Service Profile would not be available any longer within the network.

In order to create a truly distributed system 3GPP therefore created a data storage architecture (see e.g. 3GPP TS 23.501 [2]), where structured data and unstructured data can be stored at central repositories. The Unstructured Data Storage Function (UDSF) is part of this data storage architecture and offers services for data storage, manipulation and retrieval to every NF within the 5G SBA.

In the current 3GPP specifications it is specifically foreseen that the UDSF is used for achieving a fully stateless AMF (see 3GPP TS 23.501 [2] section 5.21.2 and 3GPP TS 29.500 [4] section 6.5.2). Nevertheless the UDSF can be used by any other NF to store local data within the centralized repository, so that services related to specific users or session can be handled not only by a single instance of a specific NF type, but also multiple instances, thus achieving higher reliability and load-distribution by statelessness.

5.4 Cacheable Principle

Cacheable is the REST principle which demands that clients get indication from servers whether the received information can locally be cached at the client, i.e. it can be re-used by the client at a later time. This of course is only possible if the information is not supposed to change. Therefore responses which contain serialized representations of resources must indicate whether the contained information can be cached by the client or not. 5G SBA follows this principle.

5.5 Layered System Principle

Layered system is the REST principle which demands that the client is unaware to which specific server it is connected to and if parts of the communication (e.g. authentication) are performed by different servers than other types of requests (e.g. mobility management).

Such a decoupling of services and servers which offer them is implemented into the 5G SBA e.g. via the NRF. As shown in the example procedures, the Service Registration and Service Discovery procedures enable a server, e.g. the AMF, to search for specific services provided within the system, without looking for dedicated servers.

More generally, the function split chosen by 3GPP for the 5G SBA was guided by the idea to create a truly layered system. This is visible in the different tasks assigned to the NFs. Whilst, for example, the AMF offers services access and mobility management, the SMF offers those for session management, the Network Slice Selection Function (NSSF) for network slice selection and so forth. All of these functionalities are required to make features like mobility, roaming or security seamlessly available throughout the global system to its users.

5.6 Code on Demand Principle

Code on demand is the REST principle which allows pieces of code to be downloaded from a server to a client to allow for much more dynamic and flexible service operation. This principle is currently not implemented by 5G SBA, as at currently the 5G SBA services do not require this flexibility.

5.7 Uniform Interface Principle

The uniform interface is a REST principle which encompasses four essential building blocks for the creation of distributed services to allow flexible and dynamic applications.

Resource identification in requests means that the resource or service needs to have a unique address, usually an URI. It is important to note that it is not a specific server which is addressed by the URI but the resource. As outlined above it is possible in distributed systems that more than one server can offer the same resource, in such a case the resource would have the same URI, regardless on which server it would be offered during a specific API call.

Within the example procedures we used the URI "`https://nrf7.slice-v2x.opx.3gpp/nrf-nfm/v1/nf-instances/smf5-slicev2x`" to address the Service Registration of the SMF at the NRF. This URI is constructed based on strict rules which are defined in 3GPP TS 29.510 [7] for NRF addresses. 3GPP defined related URI construction rules for every NF within the 5G SBA.

The URI includes the following elements:

- scheme ("https:"), i.e. the protocol used for message transport;
- the API route ("nrf7.slice-v2x.opx.3gpp"), i.e. the address of the NF to which this request is sent to;
- the api-name ("nrf-nfm"), i.e. the specific service which is triggered, in this case;
- the api-version ("v1"), as specified in the related 3GPP specifications;
- NF specific resource information ("nf-instances"), indicating that the conveyed information relates to a specific NF instance; and
- the instance ID of the NF which registers its profile ("smf5-slicev2x").

Note that in the given examples fixed server addresses (e.g. "nrf7.slice-v2x.opx.3gpp") were chosen, which is not in general necessary. 5G SBA allows that more generic API roots (e.g. "nrf.opx.3gpp") can be used, as long as different NF instances all offer the same resources.

Resource manipulation through representation requires that a resource on a server can be modified or replaced by a client by sending a related request which includes a representation of either the whole or part of the resource.

5G SBA fully supports this principle. For example an already stored Service Profile of an SMF at an NRF can be modified with e.g. a HTTP PATCH request from that SMF. HTTP PATCH request needs to include a

JSON document which represents either the complete Service Profile or at least those parts of it, which are intended to be modified.

Self-descriptive messages are important for the client/server and statelessness concepts of REST. It demands that requests and response sent needs to include enough information indicating how the message should be processed. This was already outlined in section 5.1.3 above.

Hypermedia as the engine of application state (HATEOAS) allows the client to learn by means of hyperlinks included in the response which further actions can be taken towards the resource in the given situation. The server provides the hyperlinks in the response based on the situation a specific resource is in.

This principle is intended to be followed within 5G SBA, but so far has not been applied to any dedicated service.

A simple example for future use is the list communication means by which a user is reachable. A user registered from a UE to the 5G system is e.g. available for calls and for receiving short messages. If an NF needs to get aware of the available communication means for the specific user it will query e.g. the AMF. The AMF, acting as service producer or server, then returns the options (hyperlinks) within the body of the response to the querying NF. In the described case these would be one hyperlink pointing to the service for establishing a call and a second hyperlink pointing to the service for sending short messages to the user.

If the user would not be registered to receive voice calls, then the AMF should, in the body of the response, only return a single hyperlink, pointing to the service for sending short messages to the user.

5.8 HTTP

RESTful principles were right from the beginning meant to evolve the way services are developed and deployed in the world wide web [1]. At that time the Hypertext Transfer Protocol (HTTP) was available as version 1.1 (as defined in IETF RFC 2068 [8]) and had a number of shortcomings which complicated the deployment of distributed services. HTTP/1.1 was designed for straightforward web-browsing, i.e. in its basic operation a browser (client) would send a HTTP request to a server, identifying a web page in the URI of the request. All going well the server then would send back the related web page in the 200 OK response to the HTTP request.

Meanwhile HTTP/2 is available (defined in RFC 7540 [9]), which solves a number of problems and especially allows for higher performance than its predecessor due to new features such as e.g. multiplexing and binary framing. The Internet Engineering Task Force (IETF) developed HTTP/2 to allow full compatibility with HTTP/1.1 but also to allow for better handling of (RESTful) API operations and general web-service development.

The HTTP methods or request types (sometimes also called “verbs”) trigger different types of functionalities within a RESTful environment, e.g.

- the HTTP POST request creates a new resource which can be addressed by the URI to which the request is sent;
- the HTTP GET request lists/retrieves the resource(s) addressed by the URI;
- the HTTP PUT request replaces the resource addressed by the URI with the representation contained in the request;
- the HTTP PATCH request updates a resources addressed with the representation contained in the request;
- the HTTP DELETE request deletes the resource addressed by the URI.

5G SBA makes use of all these different HTTP request types in the described manner. The example procedures showed how POST and PUT requests are used in 5G SBA.

6 Conclusion and Outlook

3GPP decided to not only base its northbound APIs on the principles of RESTful design, but also the 5G Core Network internal communication, i.e. to make use of RESTful APIs over all Service Based Interfaces. Already in the first 5G release, i.e. 3GPP Release 15, the principles of RESTful design have been followed strictly and thus guarantee a high level of flexibility in future service creation and enhancement.

Many details concerning the usage of RESTful APIs in 5G SBA could not be handled in this article in detail, e.g. the Richardson Maturity Levels, subscriptions to NF services, interactions with the transport layer protocol and also the openness of 5G SBA towards Remote Procedure Calls (RPC). All this would require a level of detail which cannot be covered by an article of this limited length.

Another aspect which was not handled here is the strong cooperation between 3GPP and IETF for ensuring 5G APIs to be aligned with the latest internet protocol and RESTful design developments. This coordination also allows 3GPP to push for IETF standardized solutions of the specific 5G requirements in these areas. There are many actors in both 3GPP and IETF which work to enable the alignment and integration of 5G and internet technologies.

3GPP currently (May 2018) is in the process of completing the first 5G release. This is also the first 3GPP release which saw the serious implementation of RESTful design principles. By December 2019 the second 5G release will be completed and currently the discussions are ongoing on how to further evolve the service concept of the SBA and which additional functionalities could make use of RESTful design. But already with the decisions taken so far it is clear that RESTful APIs play a central role for the technical realization and integration of 5G.

References

[1] R. Fielding, 'Architectural Styles and the Design of Network-based Software Architectures', Dissertation, University of California, Irvine, https://www.ics.uci.edu/∼fielding/pubs/dissertation/fielding_dissertation.pdf, 2000.

[2] 3GPP TS 23.501, 'System Architecture for the 5G System; Stage 2 (Rel-15)', March 2015.

[3] 3GPP TS 23.502, ' Procedures for the 5G System; stage 2 (Rel-15)', March 2015.

[4] 3GPP TS 29.500, '5G System; Technical Realization of Service Based Architecture; Stage 3', March 2015.

[5] 3GPP TS 29.501, '5G System; Principles and Guidelines for Services Definition; Stage 3', March 2015.

[6] 3GPP TS 29.502, 'Session Management Services; Stage 3', March 2015.

[7] 3GPP TS 29.510, 'Network Function Repository Services; Stage 3', March 2015.

[8] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, T. Berners-Lee, 'RFC 2068 – Hypertext Transfer Protocol – HTTP/1.1', January 1997.

[9] M. Belshe, R. Peon, M. Thomson, 'RFC 7540 - Hypertext Transfer Protocol Version 2 (HTTP/2)', May 2015.

[10] T. Bray, 'RFC 8259 - The JavaScript Object Notation (JSON) Data Interchange Format', December 2017.

[11] ECMA-404, 'The JSON Data Interchange Syntax', December 2017.

Biography

Georg Mayer is the Chairman of the Core Network and Terminals, Technical Specification Group of 3GPP (TSG CT). His current focus is on the coordination of 5G related work inside and outside 3GPP. He participates in the IETF and works closely with several of the new stakeholders in 5G, such as public safety, railways, autonomous systems and IoT service providers.

He is a published author and has co-authored a book on IMS, amongst other titles.

Georg Mayer holds an MSc in Computer Science from the University of Hagen, Germany and he works for Huawei Technologies.

5G Multimedia Standardization

Frédéric Gabin[1], Gilles Teniou[2],
Nikolai Leung[3] and Imre Varga[4]

[1] *Chairman of 3GPP SA4, Ericsson, France*
[2] *Vice Chairman of 3GPP SA4, Orange, France*
[3] *Vice Chairman of 3GPP SA4, Qualcomm, Philippines*
[4] *Chairman of EVS SWG, Qualcomm, Germany*
E-mail: frederic.gabin@ericsson.com; gilles.teniou@orange.com; nleung@qti.qualcomm.com; ivarga@qti.qualcomm.com

Received 30 March 2018;
Accepted 3 May 2018

Abstract

In the past 10 years, the Smartphone device and its 4G Mobile Broadband Connection supported the now well-established era of video multimedia services. Future mass market multimedia services are expected to be highly immersive and interactive. This paper presents an overview of 5G multimedia aspects as specified by 3GPP for various services that will be provisioned over the 5G network. Specifically, we cover the evolution of streaming services for 5G, Virtual Reality 360° video streaming, real-time speech and audio communication services VR evolution and user generated multimedia content.

Keywords: AR, VR, Audio, Video, Codec, Immersive, Live, Streaming, Multimedia.

1 Introduction

In the past 10 years, the Smartphone device and its 4th generation (4G) Mobile Broad-Band (MBB) connection supported the now well-established era of video multimedia services. Future mass market multimedia services

Journal of ICT, Vol. 6_1&2, 117–136. *River Publishers*
doi: 10.13052/jicts2245-800X.618
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

are expected to be highly immersive and interactive. Multimedia services evolve rapidly, require capable devices with flexible Application programming interfaces (APIs) and scalable distribution networks. The 5th Generation (5G) system enhances the support of MBB applications that consume bandwidth and require low latencies.

This paper presents an overview of 5G multimedia aspects as specified by the Third Generation Partnership project (3GPP) for various services that will be provisioned over the 5G network. Specifically, we cover the evolution of streaming services for 5G, Virtual Reality 360° video streaming, real-time speech and audio communication services VR evolution and user generated multimedia content.

Content providers, broadcasters and operators intend to leverage 5G systems using enhanced Mobile Broadband (eMBB) slices to deliver on-demand and live multimedia content to their subscribers. 3GPP currently studies the required evolution of media delivery services specifications.

Virtual Reality (VR) is currently the hottest topic in the field of new audio-visual experiences. It is a rendered version of a delivered audio and visual scene, designed to mimic the sensory stimuli of the real world as naturally as possible to an observer as he moves within the limits defined by the application and the equipment. Providing a 360° experience relies on a new set of representation formats for both audio and video signals that 3GPP intends to specify.

3GPP launched the new Work Item on Enhanced Voice Service (EVS) Codec Extension for Immersive Voice and Audio Services (IVAS) in September 2017 meeting. IVAS is the next generation 3GPP codec for 4G/5G, built upon the success of the EVS codec. It intends to cover use cases on real-time conversational voice, multi-stream teleconferencing, VR conversational and user generated live and non-live content streaming. In addition to address the increasing demand for rich multimedia services, teleconferencing applications over 4G/5G will benefit from this next generation codec used as an improved conversational coder supporting multi-stream coding (e.g., channel, object and scene-based audio).

User generated content, especially video, has become recently some of the leading content viewed by Internet users, surpassing the popularity of branded videos and movies. Slightly preceding this trend has been the rapid increase in video traffic uploaded to popular streaming sites, with surveys showing that most Internet users upload or share a video at least once a month. The initial 3GPP version of the Framework for Live Uplink Streaming focused on a fast time to market solution leveraging IP Multimedia System (IMS) Multimedia Telephony-based implementations and allowing Hyper Text Transfer protocol (HTTP [21]) Representational State Transfer (RESTful) interface for both control signalling 3rd party services specific user plane protocol stacks like e.g. RTMP streaming protocol.

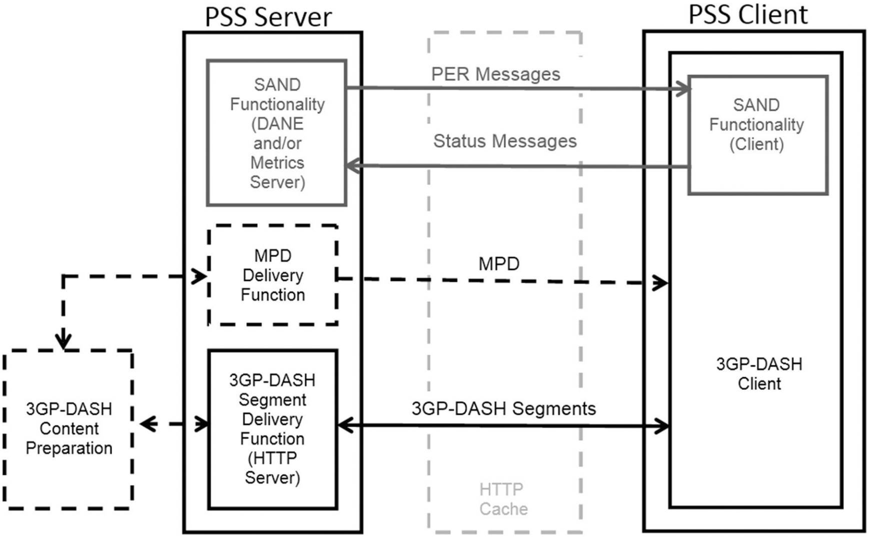
2 Streaming Service Evolutions for 5G

2.1 Introduction

3GPP Packet Switched Streaming (PSS) services have been specified and maintained by 3GPP since its Release 4 (Rel-4). While initially most deployments were operator managed audio-visual services, the specifications evolved into a set of enablers for streaming application, network and User Equipment (UE) capabilities.

2.2 Architecture

The latest PSS architecture as defined in TS 26.233 [1] in Rel-15 is depicted in Figure 1:



PSS Server
PSS Client
SAND Functionality (DANE and/or Metrics Server)
PER Messages
Status Messages
SAND Functionality (Client)
MPD Delivery Function
MPD
3GP-DASH Client
3GP-DASH Content Preparation
3GP-DASH Segment Delivery Function (HTTP Server)
3GP-DASH Segments
HTTP Cache

Figure 1 Rel-15 PSS architecture.

The PSS architecture maps to 3G, 4G and 5G systems with PSS server mapped to Application Function (AF) and PSS Client in the UE. The PSS Client supports the 3GPP Dynamic Adaptive Streaming over HTTP (3GP-DASH) [6] protocol and format for acquisition of Media Presentation Description playlist (MPD) and acquisition and playback media segments. The UE supports a set of audio and video decoders and subtitling formats as specified in TS 26.234 [14], the video profiles in TS 26.116 [2] and the yet to be published virtual reality profiles in TS 26.118 [3]. It also supports Scene Description format as specified in TS 26.307 [4]. The PSS server handles content acquisition and delivery according to the same set of codecs and formats. The server may also act as Application Function and interact with the Policy and Charging Rules Function (PCRF) via the Rx reference point for Quality of Service (QoS) control as specified in TS 29.214 [5]. The Server And Network Assisted DASH (SAND) protocol as specified in TS 26.247 [6] enables network assistance, proxy caching and consistent Quality of Experience (QoE) and QoS operations.

2.3 Foreseen Streaming Evolutions for 5G

Content providers, broadcasters and operators intend to leverage 5G systems using eMBB slices (see TS 23.501 [7]) to deliver on-demand and live multimedia content to their subscribers.

3GPP currently studies the required evolution of media delivery services specifications. The findings are documented in TR 26.891 [8].

Figure 2 illustrates the current work in progress with regards to media functions and interfaces on a 5G system architecture.

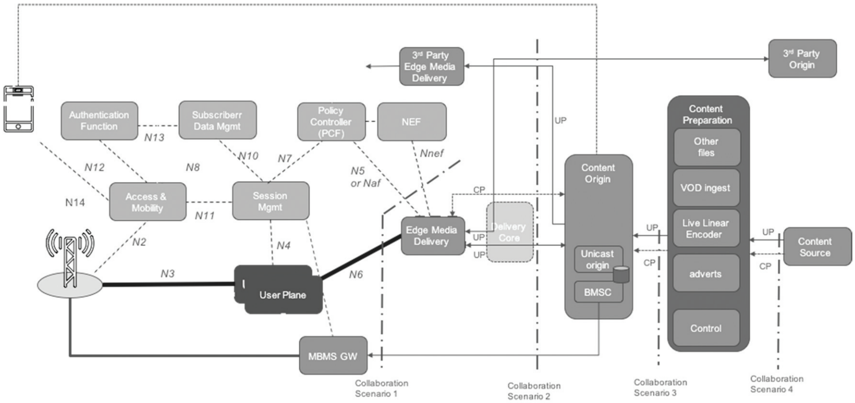


Figure 2 Media on 5G system architecture [8].

Most media distribution on 5G are expected to be based on Adaptive Bit Rate (e.g. 3GP-DASH) streaming with HTTP 1.1 to deliver file-based video content. The expected common video container format is fragmented MPEG 4 (fMP4) which is based on International Organization for Standardization Based Media File Format (ISO-BMFF) as currently used in 3GP-DASH profile. The recently specified Motion Picture Expert Group (MPEG) Common Media Application Format (CMAF) format [9] is a profile of fMP4 and can be used with various manifest formats. Typically, media segments are addressed with Uniform Resource Locators (URLs) where the domain name indicates the content provider name, i.e. the domain name of Content Origin.

The major components of media distribution are Content Preparation, Content Origin, and Delivery. The media delivery network elements and their functions are currently being studied and mapped to the 5G system. Media delivery functions are: Playlist and Media segment acquisition/delivery, Capability exchange, QoE metrics collection, Digital Rights Management (DRM) protection, Scene description, Network assistance, domain Name Server (DNS) address resolution and Load balancing, Content Distribution Network and QoS management.

For example, the streaming server may reside within the Mobile Network Operator's (MNO's) network as a dedicated application function or it may reside externally and interact with the network through the Network Exposure Function (NEF, see [7]). As another example, the media data flows through the Data Network (DN, see [7]) to the User Plane Function (UPF, see [7]) directly or through one or more AFs. In the latter case, the AF may act as if it were the origin server.

TR 26.891 "5G enhanced mobile broadband; Media distribution" [8] is planned to be completed in June 2018 and it is expected to become the foundation for Rel-16 normative work on the 3GPP 5G streaming evolved specification.

3 VR 360° Video Streaming in 5G

3.1 Introduction

Virtual Reality is certainly the hottest topic in the field of new audio-visual experiences these days. It is a rendered version of a delivered audio and visual scene, designed to mimic the sensory stimuli of the real world as naturally as possible to an observer as he moves within the limits defined by the application and the equipment.

Virtual reality usually, but not necessarily, requires a user to wear a Head Mounted Display (HMD), to completely replace the user's field of view with a simulated visual component, and to wear headphones, to provide the user with the accompanying audio. Some form of head and motion tracking of the user in VR is usually also necessary to allow the simulated visual and audio components to be updated to ensure that, from the user's perspective, items and sound sources remain consistent with the user's movements.

Apart from the complexity aspects on how to produce VR contents, the demanding required distribution bitrate and end-to-end latency put 5G as the most appropriate access network technology for ensuring the quality of experience.

3.2 Media Formats of VR 360°

Providing a 360° experience relies on a new set of representation formats for both audio and video signals. Wherever the user looks, he shall receive a comprehensive set of images and sounds for him to understand where the "objects" are around him.

This is achieved with 360° video representations for which each pixel corresponds to a particular viewing orientation. At the production side, camera systems acquire pictures from every direction that are stitched altogether at a later stage depending on the projected surface selected by the transmission system.

The most popular projections are the EquiRectangular Projection (ERP) map where the 360° picture is mapped to a sphere, and the Cube Map Projection (CMP) where the picture is mapped to a cube as illustrated on Figure 3.

These video signals can then be encoded with existing codecs such as MPEG-4 Advance Video coding (AVC) and High Efficiency video coding

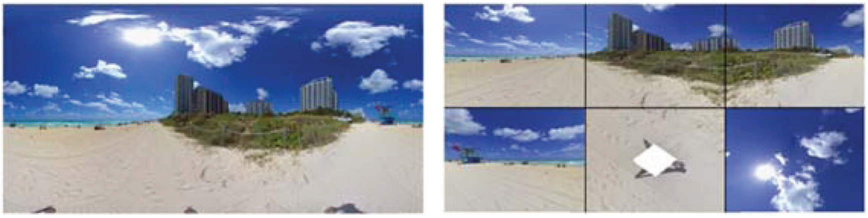


Figure 3 Equirectangular and cube-map projections.

(HEVC), together with some metadata describing the 360° nature of the content required for the correct rendering.

On audio aspects, the sound coming from the 360° scene needs to be rendered accordingly to the viewer instantaneous orientation. It means that a spatial audio representation is required together with a binaural renderer to the user's headphones.

The audio capture can be achieved with an appropriate microphone array capturing the surrounding audio field and/or using a multiple microphones configuration associated to each audio source in the scene.

There are 3 common representation models for spatial audio:

- *Channel-based audio* where each signal is associated to a rendering configuration (e.g. mono, stereo and 5.1),
- *Object-based audio* where each signal is associated to an audio source,
- *Scene-based audio* where the signal representing the entire scene can be associated to various speaker configurations.

3.3 Use Cases for VR 360°

3GPP Service and Architecture Working Group #4 (SA4, the media and codec working group) has conducted a study on Virtual Reality, documented in the technical report TR 26.918 [10]. The group has investigated the possible relevant VR 360° use cases impacting the 3GPP ecosystem (access networks and user equipment).

- **Event broadcast**: The coverage of a live event where the user can experience various viewing positions including those that are not accessible to the public (e.g. being on stage at a concert...).
- **VR Streaming**: The immersive experience provided in unicast of a live or on demand 360° content.
- **Social VR**: This is the combination of the VR streaming case with the multipoint conversational service in which the viewer can chat with friends who are physically in different locations.
- **E-Learning**: An area of applicability of VR 360° is education. With the remote class participation, a viewer can be virtually present in a classroom for a live or on-demand session.
- **VR calls**: Although it is challenging to solve the issue of the viewer representation in the virtual world, these video conferencing cases address the different possibilities of point-to-point and multi-point conversational services in VR.

- **2D legacy on HMD**: The private consumption of so-called Live Television (TV) and on-demand contents within the VR environment, such as in a virtual movie theater, including scene compositing for graphics elements like Electronic Program Guide (EPG), info bannersetc.

3.4 VR 360° Streaming

The Rel-15 normative work of 3GPP consists of defining the technical enablers for the delivery of an audiovisual 360° scene. The specification TS 26.118 3GPP Virtual Reality profiles for streaming applications [3] defines interoperability points for VR 360° streaming services.

Similarly to what has been achieved in the past with the TV profiles [2] this specification, still under development at the time of writing, will provide a set of operation points describing the media formats for VR 360° together with their mapping to the DASH delivery.

To achieve this, a reference architecture of the 3GPP client has been described as depicted in Figure 4.

This reference client architecture defines 3 interoperability points highlighted by the vertical red bars on the figure:

- **3GPP VR Media Profile**: Operation points defining the audio and video codecs and their constraints on the bitstream format and signaling together with the associated requirements on the decoding capabilities for the User Equipment.
- **3GPP VR DASH Profile**: Mapping of the operation points to the 3GPP DASH delivery format for which constraints are defined.

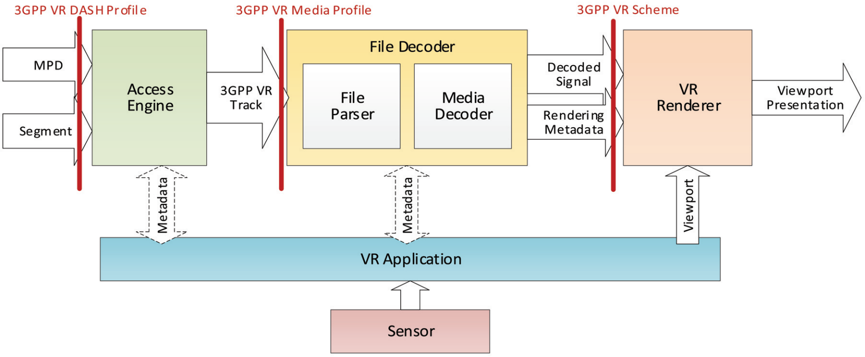


Figure 4 Client reference architecture for VR DASH streaming applications.

- **3GPP VR Scheme**: Suitable for post-processing of decoder output signals together with rendering metadata.

3.5 VR 360° Services in 5G

The suitability of the 5G for VR 360° relies many on 2 main features of this radio access network.

- First the ability to support significantly higher bit rates than previous generations allows for considering the delivery high quality services with a guaranteed quality of service.
- Also, the extreme low transmission delay of 5G, particularly within managed control of the network, enables the efficient use of viewport dependent streaming approaches.

Operation points defined in TS 26.118 [3] will for sure enable VR 360° services in 4G but the 5G capacity will allow the service to be scaled to a larger population.

4 Real-Time Audio and Video Communication Services VR Evolution

4.1 Introduction

3GPP launched the new Work Item on EVS Codec Extension for Immersive Voice and Audio Services (IVAS Codec) at the TSG-SA September 2017 meeting.

IVAS is the next generation 3GPP codec for 4G/5G, built upon the success of the EVS codec. The 3GPP real-time Enhanced Voice Services (EVS) codec has delivered a highly significant improvement in user experience with the introduction of super-wideband (SWB) and full-band (FB) speech and audio coding, together with improved packet loss resiliency.

The basic idea behind the IVAS codec work item is to cover use cases on real-time conversational voice, multi-stream teleconferencing, VR conversational and user generated live and non-live content streaming. In addition to address the increasing demand for rich multimedia services, teleconferencing applications over 4G/5G will benefit from this next generation codec used as an improved conversational coder supporting multi-stream coding (e.g., channel, object and scene-based audio).

4.2 Motivation

The introduction of 4G/5G high-speed wireless access to telecommunications networks, combined with the availability of increasingly powerful hardware platforms, will enable advanced communications and multimedia services to be deployed more quickly and easily than ever before.

Immersive services and applications, as envisioned in 3GPP TR 22.891 [11] and especially VR services and applications described in TR 26.918 [10], are expected to provide an immersive user experience which, when compared to existing media services, will deliver a quantum leap in the quality of experience. An immersive audio-visual experience implies, for the audio component, that a spatial sound impression is convincingly consistent with the presented visual scene. In addition, the user should be able to move, within certain limits defined by the application, throughout the scene, and the audio component will adjust to reflect the user's spatial orientation/ position.

3GPP TR 22.891 [11] and TR 26.918 [10] identify various immersive use cases and application scenarios that may be broadly subdivided into either UE-originated (user generated) or professionally generated content.

The approach proposed is to build upon the EVS codec with the goal of developing a single codec with attractive features and performance (e.g. excellent audio quality, low delay, spatial audio coding support, appropriate range of bit rates, high-quality error resiliency, practical implementation complexity). In the scope of 3GPP, the predominant audio rendering instrument is envisaged to be headphones but configurations with e.g. tablet speaker playback may also be of relevance.

4.3 Objectives

The overall objective of the IVAS work item is to develop a single general-purpose audio codec for immersive 4G and 5G services and applications including the VR use cases envisioned in 3GPP TR 26.918 [10].

The objectives of the standardization work are detailed below:

- Handling of encoding/decoding/rendering of speech, music and generic sound.
    - It is expected to support encoding of channel-based audio (e.g. mono, stereo or 5.1) and scene-based audio (e.g., higher-order ambisonics) inputs including spatial information about the sound

field and sound sources. The solution is expected to provide support for diegetic and non-diegetic input.
  - It is expected to provide a decoder for the encoded format and a renderer with sufficiently low motion to sound latency.

- Operation with low latency to enable conversational services over 4G/5G.
- Providing high error robustness under various transmission conditions from clean channels to channels with packet loss and delay jitter and to be optimized for 4G/5G.
- Support for a range of service capabilities, e.g., from mono to stereo to fully immersive audio encoding/decoding/rendering.
- Support of implementation on a wide range of UEs to address various needs in terms of balancing user experience and implementation complexity/cost.
- Inclusion into Multimedia Telephony Service over IMS (MTSI) services and potentially support of Multimedia Broadcast and Multicast (MBMS) and PSS services through the definition of a new immersive audio media component. Support for MTSI services is also accomplished by the provision of bit-exact EVS operation as part of the solution.

The developments under this work item should lead to a set of new specifications defining among others textual description, fixed-point C code, floating-point C code and associated test vectors of the IVAS codec, also including Real Time Transport protocol (RTP) payload format, Session Description Protocol (SDP) parameter definitions, jitter buffer management, rendering and packet loss concealment methods. It is envisioned that subsequent work outside this work item will address suitable acoustic send and receive end requirements enabling immersive user experience.

4.4 IVAS Standardization Process

The standardization process consists of two major parts: settings requirements that successful IVAS codec candidates must fulfil and defining the rigorous testing and selection framework to ensure the selection and subsequent standardization of the most attractive candidate based on well-known technical characteristics. Several permanent project documents are being prepared to support the standardization process. The current target for IVAS codec is to become part of Rel-16.

5 User Generated Multimedia Content

5.1 User Generated Content (UGC)

In the recent years, user generated content, especially video, has become some of the leading content viewed by Internet users, surpassing the popularity of branded videos and movies. Slightly preceding this trend has been the rapid increase in video traffic uploaded to popular streaming sites, with surveys showing that most Internet users upload or share a video at least once a month. The latest statistics report that more video content is uploaded in 30 days than the major U.S. television networks have created in 30 years [12]. It is expected that consumption will become even more compelling as user generated media becomes richer in quality, resolution, timeliness, and immersiveness.

5.2 Mobile Network Operator vs. Over-the-Top Services?

As the revenue shift from MNO-managed services to Over-The-Top (OTT) services continues, traditional companies in the 3GPP and wireless ecosystem are often challenged between resisting this trend or finding a way to generate revenue working with these new OTT business models. With their very low bandwidth requirement allowing operation over best-effort QoS, OTT speech services have commoditized the voice services market. For higher bandwidth applications such as video streaming, some MNOs have tried to provide their own content offerings, while others work with existing content providers in reduced- or zero-rate subscription models that have slightly more success but with unclear monetization models for the MNO beyond subscriber growth. Such efforts are often questioned as taking steps in a race to the bottom with their competitors.

Live Uplink Streaming has the potential to address this competition between these two segments with a collaborative model that is necessary for successfully addressing the user generated content market. Neither can manage without the other: QoS support by the MNO becomes more relevant for uploading richer media in a timely manner (a.k.a. "Live") while the existing OTT user base is necessary for widespread use and commercial adoption.

5.3 Overview of UGC

After voice, the earliest form of user generated content could be considered to be the Short Messaging Service (SMS) with its 140-character limit still providing value in today's political climate. SMS evolved into the Multimedia Messaging Service (MMS) which upgraded the media formats to include speech, audio, synthetic audio, still images, bitmap graphics, video, vector graphics, etc. [13]. While not as ubiquitous as SMS, MMS was well-adopted in some markets where photos and video clips were used primarily for mobile advertising.

A less well-known but also standardized messaging service was the IMS Messaging Services [15–17] which supports Immediate messaging, Session based messaging, and provides descriptions of using Combining CS and IMS Services (CSI) which allows the sharing of video or still images during a CS voice call.

The support of multiple media types (e.g., speech/audio, video, and timed text) in a single Multimedia Message sent from the MMS client to the MMS proxy and MMS servers is provided by the 3GPP File format [18] which is an instance of the ISO base media file format. This specification also provides the timing, structure, and media data for multimedia streams that are used by PSS [14] and MBMS [19]. HTTP streaming extensions are also defined for use with DASH [6].

Aside from defining a structure for integration of speech/audio codecs (including Adaptive Multirate Wideband codec (AMR-WB), EVS, Enhanced aacPlus) and video codecs (including AVC/H.264 and HEVC/H.265), the 3GPP File Format also integrates location timed metadata which, along with camera orientation information, could be leveraged for immersive media experiences that are captured from multiple perspectives using mobile devices (e.g., multiple drones filming an event, see Figure 5).

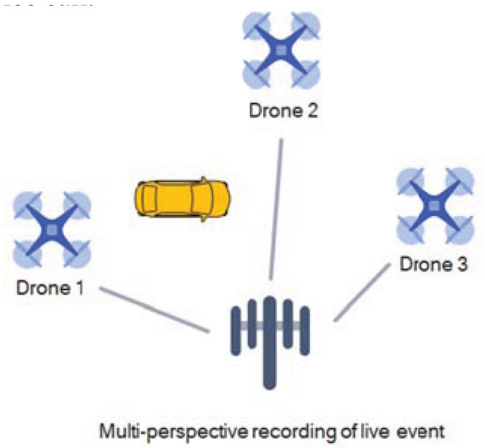


Figure 5 Multi-perspective recording of live event.

5.4 Live Uplink Streaming

Personal and semi-professional “Live” broadcasting of user generated content has become more popular, especially among users of social networks and streaming services. However current applications that operate OTT using best-effort QoS can only provide low resolution (480p, sometimes 720p) streams with quality already considered unacceptable. Semi-professionals in the field have reportedly resorted to streaming High Definition (HD) and Ultra High Definition (UHD) video uploads over multiple simultaneous links via multiple mobile devices and across different radio access technologies (see Figure 6). Some form of guaranteed QoS is needed to provide a viable service with practical capturing and transmitting devices that could support wider adoption and usage.

Tests in commercial 4G Long Term Evolution (LTE) networks demonstrate that uplink transmission of high-quality video requires QoS and can only support 2-3 users per cell. With its ability to provide even higher data rates at lower latencies, 5G has the potential to provide a QoS level that supports multiple live HD and UHD video streams in the same geographic area.

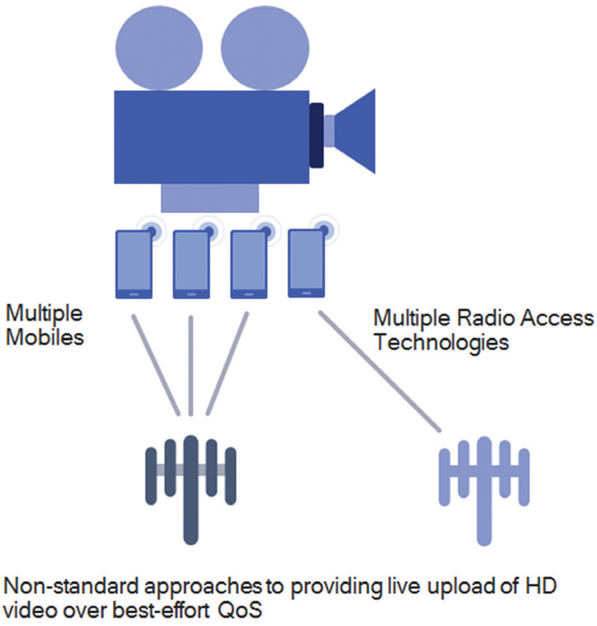

Multiple Mobiles
Multiple Radio Access Technologies
Non-standard approaches to providing live upload of HD video over best-effort QoS

Figure 6 Non-standard approaches.

Live Streaming typically does not require the same low latencies as conversational media, even when there is some interaction between the viewers and the operator of the capture device. The additional latency budget provides flexibility for uplink schedulers to improve cell capacity when supporting these high data rate streams.

The initial Release 15 version of the Framework for Live Uplink Streaming (FLUS) TS [20] focused on a fast time to market solution leveraging IMS MTSI-based implementations. The architecture (see Figure 7) re-used the IMS session control and MTSI protocol stack to support delivering an uplink stream to a server from which it could be forwarded onto viewers. It also supported live streaming directly to another MTSI client to provide a richer form of the "See What I See" CSI service.

SA4 also recognized that support for 3rd party service providers (e.g., social networks and streaming sites for UGC), with their large user base, would be important for a successful service. Initial support for 3rd party services was provided in Release 15 by defining the non-IMS framework that enabled use of the more web-friendly HTTP RESTful interface for control signalling (F-C) and allowing a flexible user plane (F-U) that allows 3rd party services to continue using their user plane protocol stacks over the 3GPP radio interface. For example, the widely used Real Time Messaging Protocol (RTMP) streaming protocol can be used within the Live Uplink Streaming framework. The specific codec formats for this feature were unspecified in Release 15 to allow for maximum flexibility and left interoperability to be handled by the 3rd party between its servers and clients.

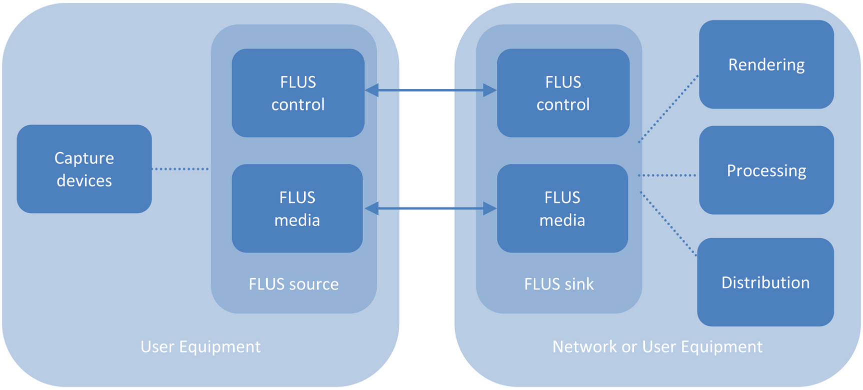


Figure 7 FLUS architecture and sub-functions.

The plans for Release 16 include further enhancing the support for 3rd party service providers by providing a QoS network API that would enable 3rd parties to request the necessary QoS for the uplink. Along with this, SA4 plans to investigate developing new QCIs that would provide other latency operating points to trade-off between latency, bandwidth, and capacity. APIs between the terminal/application and the uplink server/sink will also be developed to enable control of network-based processing such as stitching, transcoding, and how the media is to be distributed (e.g., over PSS, DASH, MBMS).

6 Conclusions

The area of multimedia technology is highly dynamic and advances rapidly. The adoption and growth of new services requires high performances, reliability and scalability of the 5G systems and its multimedia enablers. 3GPP relies solely on the contributions of its members to define those enablers. 5G commercial launches are around the corner and the standardization work on 5G multimedia has started and will continue in the years to come to support new advanced immersive and interactive services offered by operators and third-party to their subscribers.

References

[1] 3GPP TS 26.233 Transparent end-to-end Packet-switched Streaming service (PSS); General description.

[2] 3GPP TS 26.116 : "Television (TV) over 3GPP services; Video profiles".

[3] 3GPP TS 26.118: "3GPP Virtual reality profiles for streaming applications".

[4] 3GPP TS 26.307: "Presentation layer for 3GPP services".

[5] 3GPP TS 29.214: "Policy and Charging Control over Rx reference point".

[6] 3GPP TS 26.247: "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[7] 3GPP TS 23.501: "System Architecture for the 5G System".

[8] 3GPP TR 26.891: "5G enhanced mobile broadband; Media distribution".

[9] MPEG CMAF: ISO/IEC CD 23000-19 Common Media Application Format

[10] 3GPP TR 26.918: "Virtual Reality (VR) media services over 3GPP".

[11] 3GPP TR 22.891: “Study on New Services and Markets Technology Enablers”.

[12] https://www.wordstream.com/blog/ws/2017/03/08/video-marketing-statistics

[13] 3GPP TS 26.140: “Multimedia Messaging Service (MMS); Media formats and codecs”.

[14] 3GPP TS 26.234 : “Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs”.

[15] 3GPP TS 26.141: “IP Multimedia System (IMS) Messaging and Presence; Media formats and codecs”.

[16] 3GPP TS 22.340: “IP Multimedia Subsystem (IMS) messaging; Stage 1”.

[17] 3GPP TS 22.141: “Presence service; Stage 1”.

[18] 3GPP TS 26.244: “Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)”.

[19] 3GPP TS 26.346: “Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs”.

[20] 3GPP TS 26.238: “Uplink streaming”.

[21] IETF RFC 2616, Hypertext Transfer Protocol – HTTP/1.1

Biographies

Frédéric Gabin received his M.Sc. degree in electronics and digital telecommunication systems in 1997 from Telecom ParisTech. He worked in the areas of speech and radio signal processing as research and standardization engineer in Nortel Networks, as systems engineer, standardization project manager and then research manager for NEC terminal division and as standardization manager for Ericsson Mobile Platform and Ericsson. He is Standardization Manager for Media area at Ericsson. Frédéric Gabin served as delegate and chairman of several standardization groups at ETSI, 3GPP, DVB, GSMA and IMTC. He is the Chairman of the 3GPP SA4 Codec and Multimedia Working Group.

Gilles Teniou received the Master's degree in Engineering Computer Vision, from the Education and Research department in Computer Science and Electrical Engineering, University of Rennes, France. He has been Head of video coding standardization activities at Orange. Gilles is currently Senior Standardization Manager on Content and TV services at Orange. He is in charge of managing the technical and operational standardization activities related to TV & Audiovisual Services including service architecture, technologies used for audiovisual streams (media formats and protocols) as well as the application environment used for TV. In 3GPP, Gilles is the Vice Chair of the 3GPP SA4 Working Group and the chairman of the Video Sub-Working Group.

Nikolai Leung received his B.S. Degree in Electrical Engineering from the University of the Philippines and his M.S. Degree in Electrical Engineering Communication Systems from the University of Michigan. He has been responsible for leading different engineering teams at QUALCOMM Technologies Incorporated and is currently a Director of Technical Standards. He is also serving as the Vice Chair of the 3GPP SA4 Working Group and the Chair of the Multimedia Telephony Services Sub-Working Group.

Imre Varga received his M.Sc. and Ph.D. degrees in electrical engineering and worked in various positions (R&D, project lead, line manager, department head for multimedia) on signal processing for professional audio, multimedia communication, speech coding and transmission, acoustics pre-processing, video applications, and command-and-control systems. He is Director of Technical Standards at QUALCOMM Technologies Incorporated responsible for speech and audio standardization.

Imre Varga served as delegate and chairman of standardization groups for speech and audio coding at ITU-T, 3GPP and IMTC. He also serves as the Chairman of the Enhanced Voice Services Sub-Working Group of 3GPP SA4.

3GPP 5G Security

Anand R. Prasad[1], Sivabalan Arumugam[2],
Sheeba B[3] and Alf Zugenmaier[4]

[1] *Chairman of 3GPP SA3, NEC Corporation, Japan*
[2,3] *NEC Technologies India Pvt. Ltd., India*
[4] *Vice Chairman of 3GPP SA3 & Rapporteur, Munich University of Applied Sciences, Germany*
E-mail: anand@bq.jp.nec.com; sivabalan.arumugam@india.nec.com; sheeba.mary@india.nec.com; alf.zugenmaier@hm.edu

Received 30 March 2018;
Accepted 3 May 2018

Abstract

5G is the next generation of mobile communication systems. As it is being finalized, the specification is stable enough to allow giving an overview. This paper presents the security aspects of the 5G system specified by the 3rd Generation Partnership Project (3GPP), especially highlighting the differences to the 4G (LTE) system. The most important 5G security enhancements are access agnostic primary authentication with home control, security key establishment and management, security for mobility, service based architecture security, inter-network security, privacy and security for services provided over 5G with secondary authentication.

Keywords: LTE, 5G, 5G Core, NR, Authentication, Services, Security, Privacy.

1 Introduction

The 5G system is an evolution of the 4G mobile communication system, i.e. System Architecture Evolution/Long Term Evolution (SAE/LTE). Accordingly, the 5G security architecture has been designed to integrate

Journal of ICT, Vol. 6_1&2, 137–158. *River Publishers*
doi: 10.13052/jicts2245-800X.619
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

4G equivalent security into the 5G system. In addition, reassessment of other security threats such as attacks on radio interfaces, signalling plane, user plane, masquerading, privacy, replay, bidding down, man-in-the-middle, service based interfaces (SBI), and inter-operator network security have led to integration of further security mechanisms. This paper gives an overview of the security in phase 1, also called release 15 in 3GPP, and highlights the security features and security mechanisms offered by the 5G system, and the security procedures performed within the 5G System including the 5G Core (5GC) and the 5G new radio (NR), i.e. the 5G radio interface.

The paper starts by laying out the underlying trust models in 5G system considering roaming and non-roaming cases in Section 2 along with a brief summary on 5G key hierarchy. The enhancements in authentication and privacy are dealt with in Section 3. Section 4 discusses the multiple registration scenarios of User Equipment (UE) considering different cases such as same Public Land Mobile Network (PLMN) and different PLMN scenarios. The mobility procedures and intra-/inter-network security are discussed in Sections 5 and 6 respectively. The role of secondary authentication in services security is briefed in Section 7. Section 8 discusses the security aspects of network interconnects and Section 9 elaborates the migration and interworking security. Finally the paper is concluded in Section 10.

2 Evolution of the Trust Model

In the new 5G system, trust within the network is considered as decreasing the further one moves from the core. This has impact on decisions taken in 5G security design thus we present the trust model in this section, at the beginning fo the paper, together with the 5G key hierarchy.

2.1 Trust Model

The trust model in the user equipment (UE) is reasonably simple: there are two trust domains, the tamper proof universal integrated circuit card (UICC) on which the the Universal Subscriber Identity Module (USIM) resides as trust anchor. Mobile Equipment (ME) and the USIM together form the UE.

The network side trust model for roaming and non-roaming cases are shown in Figures 1 and 2 respectively, which shows the trust in mulitple layers, like in an onion.

The Radio Access Network (RAN) is separated into distributed units (DU) and central units (CU) – DU and CU together form gNB the 5G base-station.

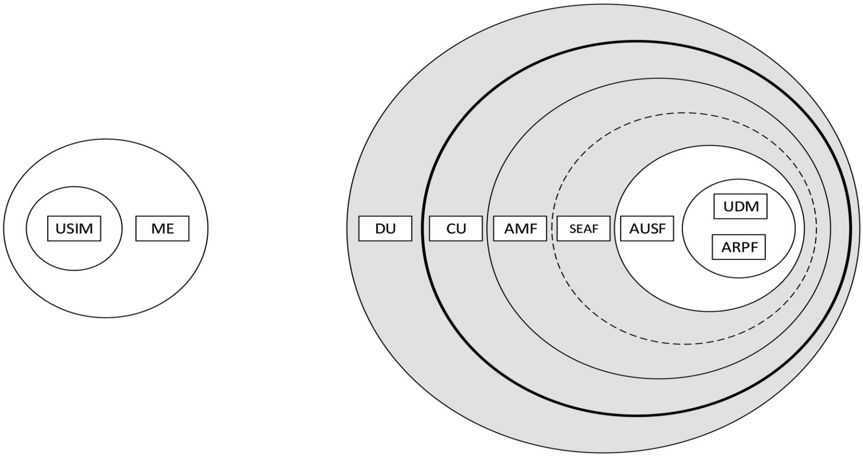


Figure 1 Trust model of non-roaming scenario.

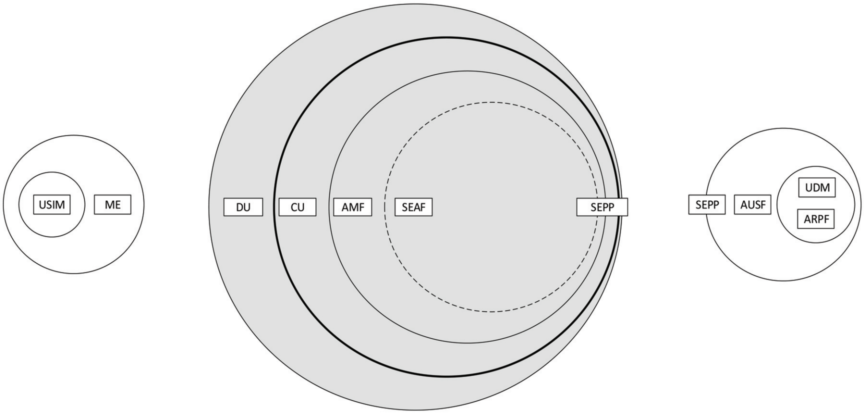


Figure 2 Trust model of roaming scenario.

The DU does not have any access to customer communications as it may be deployed in unsupervised sites. The CU and Non-3GPP Inter Working Function (N3IWF – not shown in the figures), which terminates the Access Stratum (AS) security, will be deployed in sites with restricted access to maintenance personnel.

In the core network the Access and Mobility Management Function (AMF) serves as termination point for Non-Access Stratum (NAS) security. Currently, i.e. in the 3GPP 5G Phase 1 specification [2], the AMF is collocated with the SEcurity Anchor Function (SEAF) that holds the root key (known as anchor key) for the visited network. The security architecture is defined in a future proof fashion, as it allows separation of the security anchor from the mobility function in a future evolution of the system architecture.

In the roaming architecture, the home and the visited network are connected through SEcurity Protection Proxy (SEPP) for the control plane of the internetwork interconnect. This enhancement is done in 5G because of the number of attacks coming to light recently such as key theft and re-routing attacks in SS7 [16] and network node impersonation and source address spoofing in signalling messages in DIAMETER [17] that exploited the trusted nature of the internetwork interconnect [18]. Authentication Server Function (AUSF) keeps a key for reuse, derived after authentication, in case of simultaneous registration of a UE in different access network technologies, i.e. 3GPP access networks and non-3GPP access networks such as IEEE 802.11 Wireless Local Area Network (WLAN). Authentication credential Repository and Processing Function (ARPF) keeps the authentication credentials. This is mirrored by the USIM on the side of the client, i.e. the UE side. The subscriber information is stored in the Unified Data Repository (UDR). The Unified Data Management (UDM) uses the subscription data stored in UDR and implements the application logic to perform various functionalities such as authentication credential generation, user identification, service and session continuity etc. Over the air interface, both active and passive attacks are considered on both control plane and user plane. Privacy has become increasingly important leading to permanent identifiers being kept secret over the air interface.

2.2 Key Hierarchy

The long term secret key (K) provisioned in the USIM and the 5G core network acts as the primary source of security context in the same way as in of an 4G system. Different to LTE, in 5G there are 2 types of authentication, primary authentication that all devices have to perform for accesing the mobile network services, and secondary authentication to an external data network (DN), if so desired by the external data network.

After a successful primary authentication between the UE and the network, the serving network specific anchor key (K_{SEAF}) is derived from K. From the anchor key, confidentiality and integrity protection keys are derived for NAS signalling and the AS consisting of control plane (CP), ie. radio resource control (RRC) messages, and user plane (UP). The key hierarchy of 5G is shown in Figure 3. The key hierarchy includes K, Cipher Key (CK) and

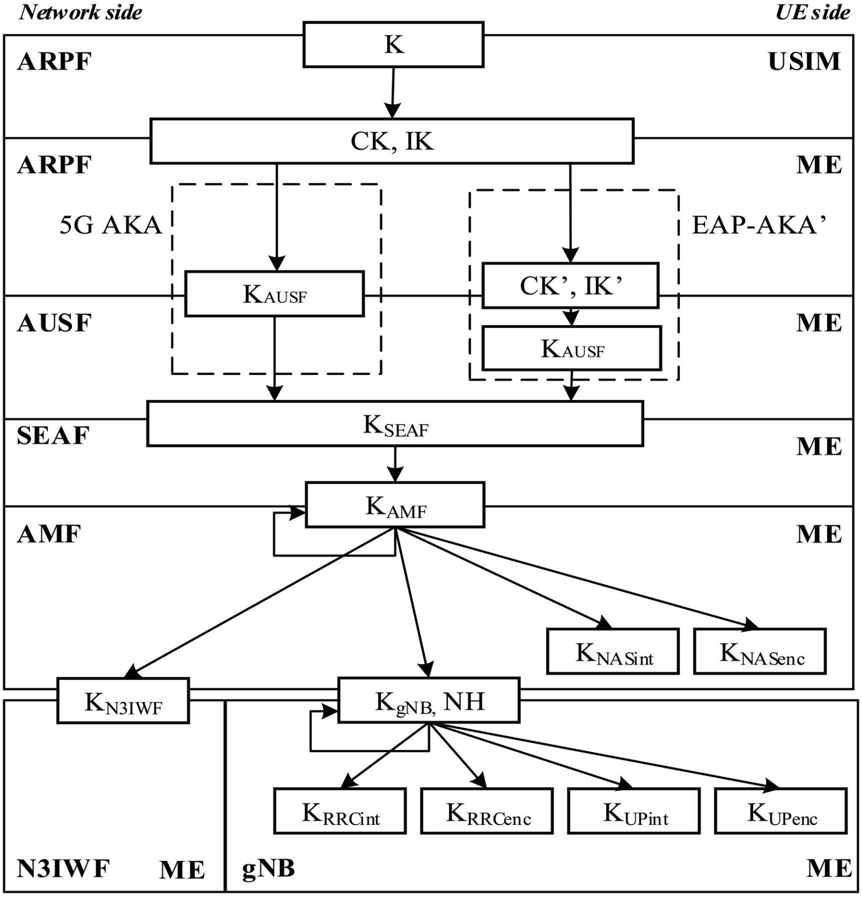


Figure 3 Key hierarchy.

Integrity Key (IK), K_{AUSF}, K_{SEAF}, K_{AMF}, K_{NASint}, K_{NASenc}, K_{N3IWF}, K_{gNB}, K_{RRCint}, K_{RRCenc}, K_{UPint} and K_{UPenc}.

The K_{AUSF} is derived by ME and ARPF from CK and IK during 5G Authentication and Key Agreement (AKA). If the 3GPP credential K is used for authentication over a radio access technology supporting the extensible authentication protocol EAP, K_{AUSF} is derived by ME and AUSF according to the EAP AKA' specification. From K_{AUSF}, the AUSF and ME derive the anchor key K_{SEAF} that is then used to derive the K_{AMF} by ME and SEAF. The K'_{AMF} is a key that can be derived by ME and AMF from previous K_{AMF} when the UE moves from one AMF to another during inter-AMF mobility.

The integrity and confidentiality keys, K_{NASint} and K_{NASenc} respectively, are derived by ME and AMF from K_{AMF} for the NAS signalling protection. The K_{gNB} is derived by ME and AMF from K_{AMF}. The K_{gNB} is also derived by ME and source gNB using a intermediary key, $K_{\mathrm{gNB}}*$, during mobility that can lead to, what is known as, horizontal or vertical key derivation. The integrity and confidentiality keys for AS, i.e. UP (K_{UPint} and K_{UPenc}) and RRC (K_{RRCint} and K_{RRCenc}), are derived by ME and gNB from K_{gNB}. UP integrity protection is another enhancement in 5G that is valuable for the expected Internet of Things (IoT) services. The intermediate key NH is derived by ME and AMF to provide forward secrecy during handover.

3 Access and Authentication

Network access in 5G network supports privacy of the permanent identifier to attackers on the air interface. This was not available in past generations of mobile neworks. In this section we discuss privacy in 5G and authentication. We do not discuss Security Mode Command (SMC) for AS and NAS.

3.1 Overview

Access to the network requires subscriber authentication, which is done by primary authentication mechanism in 5G system. So that the network can identify the subscriber, the UE has to send the subscription permanent identifier (SUPI in 5G). This permanent subscription identifier was sent in clear until 4G leading to various privacy related attacks.

In 5G privacy is achieved, even before authentication and key agreement, by encrypting the SUPI before transmitting using a HN public key which is stored in the USIM. Although specified, privacy enablement is under the control of the home network of the subscriber. Privacy in 5G is elaborated in Section 3.2. Up to 4G, the home network had to trust the visited network through which the authentication took place. Subsequent procedures such as location updates or submission of Customer Data Records (CDRs) would need to be taken at face value. This lead to fraud cases impacting operator's revenue. Another case is the fraudulent registration attempt by an attacker to register the subscriber's serving AMF in UDM when UE is not present in the serving AMF. To resolve these issues, in 5G the concept of increased home control was introduced, where the home network receives proof of UE participation in a successful authentication.

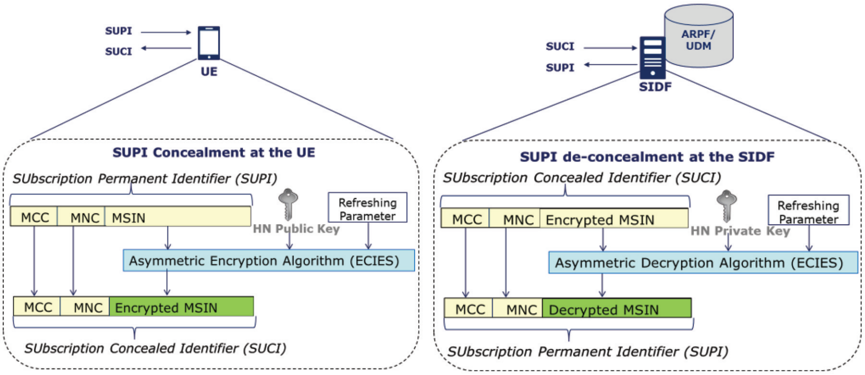

SUPI
SUCI
UE
SUPI Concealment at the UE
SUbscription Permanent Identifier (SUPI)
MCC
MNC
MSIN
HN Public Key
Refreshing Parameter
Asymmetric Encryption Algorithm (ECIES)
MCC
MNC
Encrypted MSIN
SUbscription Concealed Identifier (SUCI)
SUCI
SUPI
SIDF
ARPF/ UDM
SUPI de-concealment at the SIDF
SUbscription Concealed Identifier (SUCI)
MCC
MNC
Encrypted MSIN
HN Private Key
Refreshing Parameter
Asymmetric Decryption Algorithm (ECIES)
MCC
MNC
Decrypted MSIN
SUbscription Permanent Identifier (SUPI)

Figure 4 SUPI structure and concealed sensitive information.

3.2 Privacy

The subscription identifier SUPI, see Figure 4, contains sensitive subscriber as well as subscription information thus it should not be transferred in clear text except for parts necessary for proper functioning of the system, i.e. routing information in the form of Mobile Country Code (MCC) and Mobile Network Code (MNC). As explained in 3.1, the subscriber privacy enablement is under the control of the home network of the subscriber. Note that in case of unauthenticated emergency calls, privacy protection is not required. So as to provide privacy the UE generates and transmits the Subscription Concealed Identifier (SUCI[1]) using a protection scheme, i.e. one of the Elliptic Curve Integrated Encryption Scheme (ECIES) profiles, with the public key that was securely provisioned in control of the home network.

The UE constructs the SUCI from the protection scheme identifier, the home network public key identifier, the home network identifier and the protection scheme-output that represents the output of a public key protection scheme. The SUCI will contain routing information in the clear, which is the mobile network and mobile country code of the home network, as well as potentially some routing information within the home network, where the home network is so large that it needs to be segmented. At the home network de-conealment of the SUPI from SUCI is done by the Subscription Identifier De-concealing Function (SIDF) that is located at the ARPF/UDM. To meet the LI requirements along with privacy, binding of SUPI to the derivation of the K_{AMF} is done.

[1] SUCI is pronounced sushi.

3.3 Authentication Procedure In 5G System

EAP-AKA' and 5G AKA are mandatory 5G primary authentication methods. Other EAP based authentication methods can be used optionally as well. For the purpose of explanation we have divided the authentication steps in two phases, see Figure 5. Phase 1 is initiation of 5G authentication and authentication method selection. Phase 2 is mutual authentication between the UE, subscription, and the network.

During phase 1, the UE sends a registration request (N1 message) to the SEAF that contains a concealed identifier SUCI or 5G-Globally Unique Temporary UE Identity (5G-GUTI) where, as the name suggests, 5G-GUTI

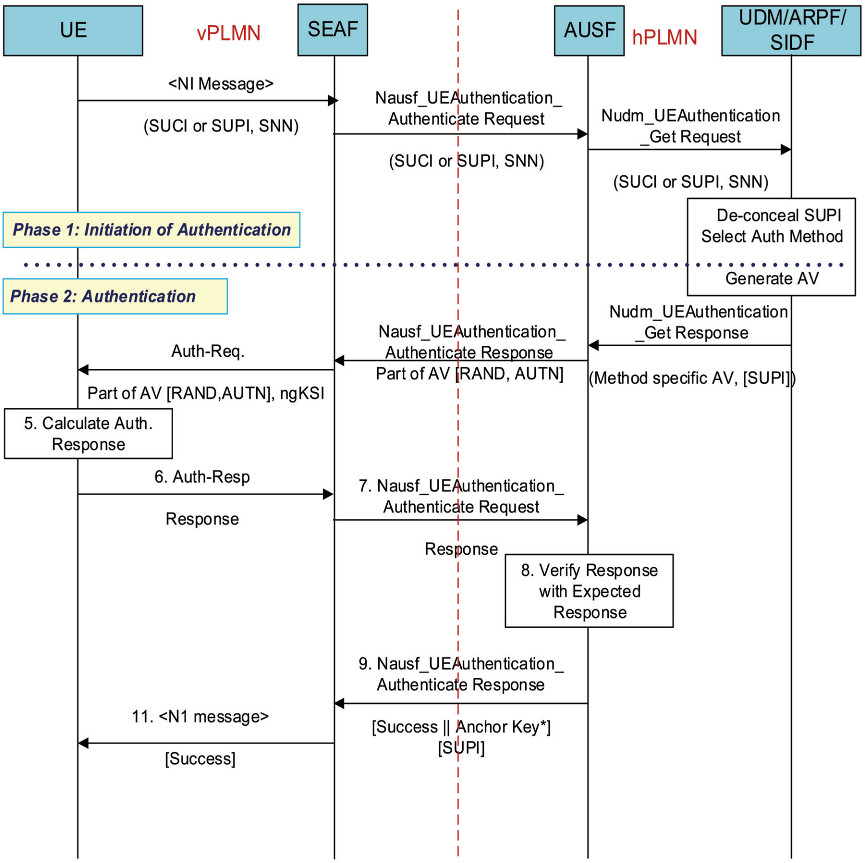

UE
vPLMN
SEAF
AUSF
hPLMN
UDM/ARPF/ SIDF
<NI Message>
(SUCI or SUPI, SNN)
Nausf_UEAuthentication_ Authenticate Request
(SUCI or SUPI, SNN)
Nudm_UEAuthentication _Get Request
(SUCI or SUPI, SNN)
De-conceal SUPI
Select Auth Method
Generate AV
Phase 1: Initiation of Authentication
Phase 2: Authentication
Nudm_UEAuthentication _Get Response
(Method specific AV, [SUPI])
Nausf_UEAuthentication_ Authenticate Response
Part of AV [RAND, AUTN]
Auth-Req.
Part of AV [RAND,AUTN], ngKSI
5. Calculate Auth. Response
6. Auth-Resp
Response
7. Nausf_UEAuthentication_ Authenticate Request
Response
8. Verify Response with Expected Response
9. Nausf_UEAuthentication_ Authenticate Response
[Success || Anchor Key*]
[SUPI]
11. <N1 message>
[Success]

Figure 5 Message exchanges involved in 5G authentication procedure.

is a temporary identity assigned by the network during a previous session. On receiving a registration request from the UE the SEAF sends an authentication request (Nausf_UEAuthentication_Authenticate Request) message to the AUSF with the serving network name (SNN) [2] and either SUPI, if available and 5G-GUTI is valid, or SUCI. The SNN is a concatenation of service code and the Serving Network Identity (SN Id). Upon receiving the authentication request, the AUSF checks whether the requesting SEAF is authorized to use the SNN which is a form of home control in 5G. If the serving network is not authorized to use the SNN, the AUSF respond with "serving network not authorized" in the authentication response (Nausf_UEAuthentication_Authenticate Response). The authentication information request (Nudm_UEAuthentication_Get Request) from AUSF to UDM/ARPF/SIDF includes the SUCI or SUPI and the SNN. SIDF is invoked to de-conceal the SUPI from SUCI. Based on SUPI and the subscription data, the UDM/ARPF choose the authentication method to be used.

In phase 2, on selection of authentication methods, mutual authentication takes place. The authentication procedure involved in 5G, see Figure 5, is briefly explained in the following steps for both EAP-AKA' and 5G AKA.

- Authentication Vector (AV) generation:
  EAP-AKA': The authentication procedure is followed as discussed in RFC 5448 [9] except the authentication vector (AV) derivation at the UDM/ARPF. The UDM/ARPF first generates an AV with AMF separation bit = 1 [8] and generates CK' and IK' from CK, IK and SNN. The UDM/ARPF subsequently sends this transformed AV (RAND, AUTN, XRES, CK', IK') to the AUSF with an indication that the AV is to be used for EAP-AKA'.

  5G AKA: The UDM/ARPF derives the K_{AUSF} from CK, IK and SNN and generates the 5G Home Environment AV (5G HE AV) where the 5G HE AV contains the RAND, AUTN, XRES*, and K_{AUSF}. 5G HE AV is sent to the AUSF in the authentication information Request Response (Auth-info Resp) message. The AUSF stores the K_{AUSF} and XRES* until expiry.
- The AUSF derives the K_{SEAF} (anchor key) from K_{AUSF} and sends the Challenge message to the SEAF in a Nausf_UEAuthentication_Authenticate Response message with K_{SEAF}, AUTN and RAND. In case of 5G AKA HXRES* is also sent.

- At receipt of the RAND and AUTN, the USIM computes a response RES and returns RES, CK, IK to the UE. In case of 5G AKA additionally the ME compute RES* from RES. The UE then sends the Challenge Response message to the SEAF in a NAS message Auth-Resp message.
- The SEAF forwards the Response Challenge message to the AUSF in Nausf_UEAuthentication_Authenticate Request message. In case of 5G AKA the SEAF computes HRES* from RES*, and compares HRES* with HXRES*. If the values are same, the SEAF considers the authentication as successful and sends the received RES*, in a Nausf_UEAuthentication_Authenticate Request message containing the SUPI or SUCI and the SNN, to the AUSF.
- The AUSF verifies the message to support increased home control and if the verification is successfull, the AUSF acts according to the authentication method as explained below. Note that if the AUSF received SUCI from the SEAF, then the AUSF also includes the SUPI in 5G-Authentication Confirmation Answer message.

  EAP-AKA': The AUSF and UDM in the home network obtains confirmation that the UE has been successfully authenticated when the EAP-Response/AKA'-Challenge received by the AUSF has been successfully verified. The AUSF derives EMSK from CK' and IK' as described in RFC 5448. The AUSF then uses the first 256 bits of EMSK as the K_{AUSF} and derives the anchor key K_{SEAF} from K_{AUSF}. The AUSF sends EAP Success message to the SEAF inside Nausf_UEAuthentication_Authenticate Response along with the K_{SEAF}.

  5G AKA: The AUSF compares the received RES* with the stored XRES* and if they are equal, the AUSF considers the confirmation message as successfully verified and indicates this to the SEAF. The AUSF and UDM in the home network obtains confirmation that the UE has been successfully authenticated.
- The SEAF sends the Success message to the UE in the N1 message.
- The SEAF then derives the K_{AMF} from the K_{SEAF} and the SUPI and send it to the AMF. On receiving the Success message, the UE derives K_{AUSF} and K_{SEAF} in the same way as the AUSF and derives the K_{AMF} from the K_{SEAF} and the SUPI. The SEAF provide the ngKSI and the K_{AMF} to the AMF.

4 Multiple Registrations

There are two cases as shown in Figure 6 where the UE can be registered in both a network accessed through 5G NR and simultaneously in network accessed through a non-3GPP radio access technology like WLAN. This can be in the same PLMN or in the different PLMN's serving networks. The UE will establish two NAS connections with the network in both cases. This is called multiple registration

The first case is where the UE is registered with the same AMF in the same PLMN serving network over both 3GPP and non-3GPP accesses. A common NAS security context is created during the registration procedure over the first access type. In order to realize cryptographic separation and replay protection, the common NAS security-context will have parameters specific to each NAS connection. The connection specific parameters include a pair of NAS COUNTs for uplink and downlink and unique NAS connection identifier. The value of the unique NAS connection identifier is set to "0" for 3GPP access and set to "1" for non-3GPP access. The second case is when the UE is registered in one PLMN over a certain type of access (e.g. 3GPP) and is registered to another PLMN over the other type of access (e.g. non-3GPP). The UE independently maintains and uses two different 5G security contexts, one per PLMN. Each security context is established separately via a successful primary authentication procedure with the Home PLMN. All the NAS and AS security mechanisms defined for single registration mode are applicable independently on each access using the corresponding 5G security context.

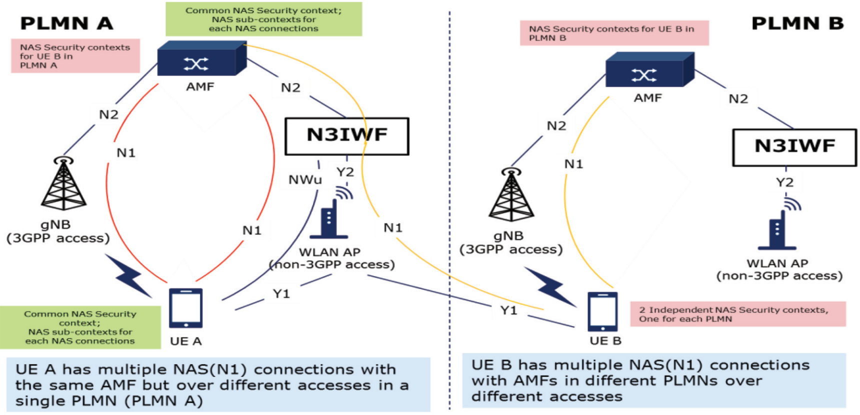


Figure 6 5G supporting multiple NAS connections.

5 Mobility

Depending on an operator's security requirements, the operator can decide whether to have Xn or N2 handovers for a particular gNB according to the security characteristics of a particular gNB. Where Xn handover is handover over Xn interface without involvement of AMF and N2 handover involves the AMF. The 5G mobility scenarios are depicted in Figure 7 is briefed as follows.

Xn-handover: The handover of UE from a source gNB to a target gNB over Xn is referred to as Xn-handover. The source gNB includes the UE 5G security capabilities in the handover request message containing the ciphering and integrity algorithms used in the source cell. The target gNB selects the algorithm with highest priority from the received 5G security capabilities of the UE according to the prioritized locally configured list of algorithms. The chosen algorithms are indicated to the UE in the Handover Command message if the target gNB selects different algorithms. If the UE does not receive any selection of integrity and ciphering algorithms, it continues to use the same algorithms as before the handover [2, 4]. In the Path-Switch message, the target gNB sends the UE's 5G security capabilities received from the source gNB to the AMF. The AMF will verify that the UE's 5G security capabilities received from the target gNB are the same as the UE's 5G security capabilities that the AMF has locally stored. If there is a mismatch, the AMF will send its locally stored 5G security capabilities of the UE to

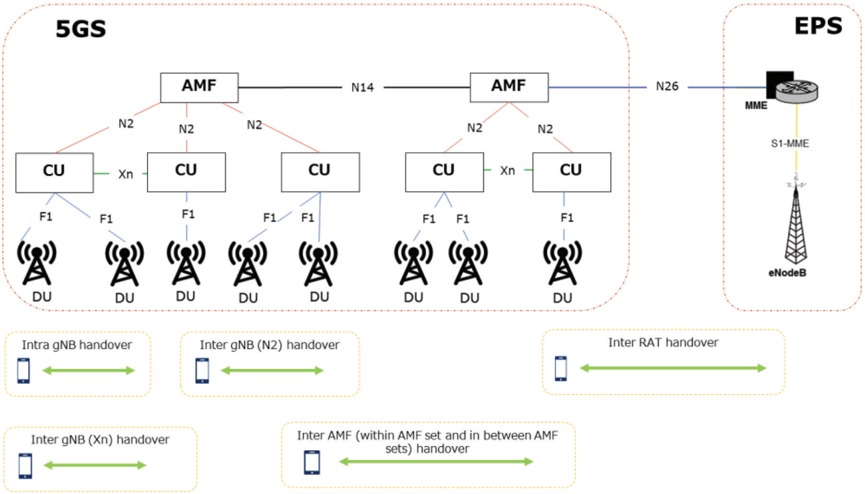


Figure 7 Oveview of 5G mobility scenario.

the target gNB in the Path-Switch Acknowledge message. Additionally, the AMF may log the event and may take additional measures, such as raising an alarm.

N2-handover: The handover from a source gNB to a target gNB over N2 interface possibly including an AMF change is referred to as N2-handover or inter-AMF handover. For N2-handover, the source gNB includes AS algorithms used in the source cell in the source to target transparent container that is sent to the target gNB. The AS algorithms used in the source cell are provided to the target gNB so that it can decipher and integrity verify the RRCConnectionReestablishmentComplete message on Signalling Radio Bearer 1 (SRB1) in the potential RRC Connection Re-establishment procedure. The AMF should not initiate any of the N2 procedures including a new key towards a UE if a NAS Security Mode Command (SMC) procedure is ongoing with the UE. The AMF will not initiate a NAS SMC towards a UE if one of the N2 procedures including a new key is ongoing with the UE.

Intra-gNB-CU handover: This type of handover occurs in gNBs with split DU-CU, where the UE performs handover between DUs within a gNB-CU. It is not required to change the AS security algorithms during intra-gNB-CU handover as the security termination point remains the same. If the UE does not receive an indication of new AS security algorithms during an intra-gNB-CU handover, the UE can continue to use the same algorithms as before.

6 DU-CU Interface Security

The F1 interface [5, 6] between DU and CU could also be protected by NDS/IP [11, 12]. Messaging over F1 interface include control-plane (F1-C), management traffic and user-plane (F1-U). The security requirements for the F1 interface includes support of confidentiality, integrity and replay protection. It is expected that F1-U security is independent of F1-C or management traffic security, i.e. one could configure F1-U security differently than F1-U and management traffic security.

7 Services Security – Secondary Authentication

5G supports optional EAP based secondary authentication between the UE and an external data network (DN). Session Management Function (SMF) performs the role of the EAP Authenticator [14] and relies on an external DN-AAA server to authenticate and authorize the UE's request for the establishment of a PDU sessions. See Figure 8 for secondary authentication procedure with the external DN-AAA server.

As a pre-condition the UE is registered with the network performing primary authentication with the AUSF/ARPF and establishes a NAS security context with the AMF. The UE initiates establishment of a new PDU Session by sending an SM NAS message containing a PDU Session Establishment Request message to the AMF. The UE includes slice information (identified by S-NSSAI) and the PDN it would like to connect to (identified by DNN). The AMF sends the request to the SMF for PDU session establishment (Nsmf_PDUSession_CreateSMContext Request message) with SM NAS message, SUPI, the received S-NSSAI, and the DNN. The SMF sends an Nsmf_PDUSession_CreateSMContext Response message to the AMF. The SMF then obtains subscription data from the UDM for the given SUPI and

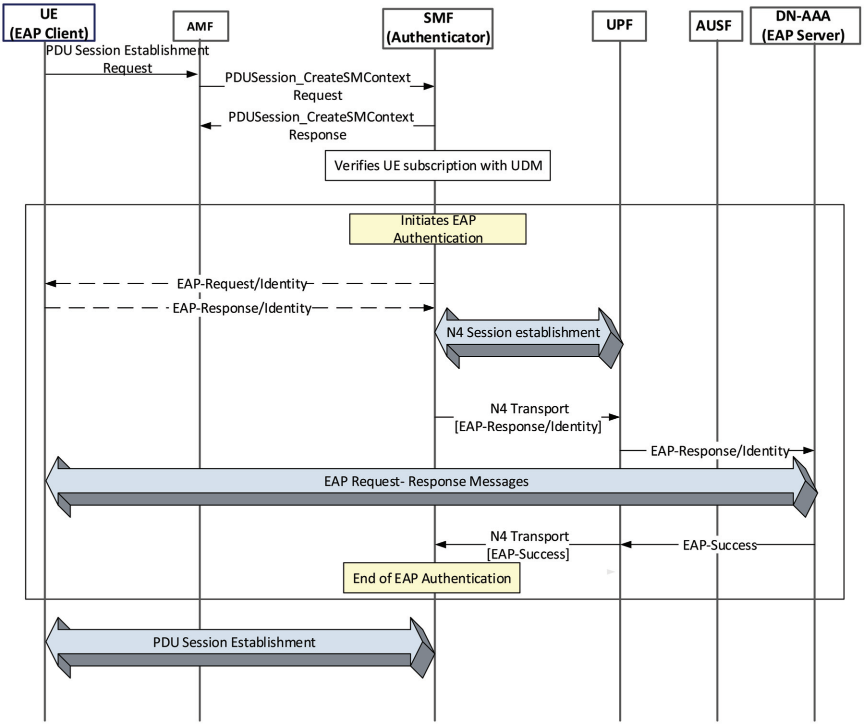

UE
(EAP Client)
AMF
SMF
(Authenticator)
UPF
AUSF
DN-AAA
(EAP Server)
PDU Session Establishment Request
PDUSession_CreateSMContext Request
PDUSession_CreateSMContext Response
Verifies UE subscription with UDM
Initiates EAP Authentication
EAP-Request/Identity
EAP-Response/Identity
N4 Session establishment
N4 Transport [EAP-Response/Identity]
EAP-Response/Identity
EAP Request- Response Messages
N4 Transport [EAP-Success]
EAP-Success
End of EAP Authentication
PDU Session Establishment

Figure 8 Secondary authentication.

verifies whether the UE request is compliant with the user subscription and with local policies. The SMF may also verify whether the UE has been authenticated and/or authorized by the same DN, as indicated DNN, or the same AAA server in a previous PDU session establishment. The SMF can skip the rest of the procedure if the verification is successful.

If the SMF finds that the UE has not been authenticated with the external DN-AAA server, then the SMF will trigger EAP Authentication to obtain authorization from an external DN-AAA server and sends an EAP Request/Identity message to the UE. The UE then send an EAP Response/Identity message with its DN-specific identity complying with Network Access Identifier (NAI) format. The DN AAA server and the UE can exchange EAP messages as required by the EAP method. EAP messages are sent in the SM NAS message between the UE and the SMF; The SMF communicates with the external DN-AAA via UPF using N4 and N6 interface [2]. After the completion of the authentication procedure, DN AAA server will send EAP Success message to the SMF. The SMF may save the UE ID and DNN (or DN's AAA server ID if available) in a list for successful authentication/authorization between the UE and SMF. Alternatively, the SMF may update the list in UDM. If the authorization is successful, PDU Session Establishment proceeds according to TS 23.502 [10].

In case of roaming scenario, two SMFs such as visitor SMF (V-SMF) and home SMF (H-SMF) are involved, where H-SMF acts as the authenticator. Following the PDU Session Establishment Request message from the UE via AMF as discussed above, the V-SMF sends an Nsmf_PDUSession_Create Request to the H-SMF. To establish the requested PDU session after a successful EAP based secondary authentication, the H-SMF sends an Nsmf_PDUSession_Create Response to V-SMF with EAP Success and this message is inturn sent to the UE by the V-SMF.

8 Inter Operator Network Security

N32 interface provides inter operator network connectivity (see Figure 9) that might traverse over Internetwork Packet Exchange (IPX). To ensure interconnect security, the SEPP is introduced as an entity that resides at the perimeter of the PLMN. The SEPP implements application layer security for all the service layer information exchanged between two Network Functions (NFs) across two different PLMNs. On receiving service layer messages from a given NF, the SEPP protects the messages before sending them over the N32 interface. Similarly, on receiving a message over N32

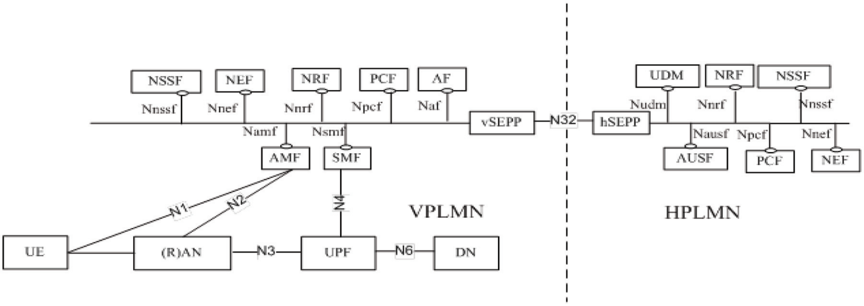


Figure 9 Interconnect security and SEPP.

interface the SEPP forwards the message to approproiate NF after security verification. The SEPP provides integrity protection, confidentiality protection of parts of message and replay protection. Mutual authentication, authorization, negotiation of cipher suites and key management are also parts of SEPP security functions. It also performs topology hiding and spoofing protection.

9 Interworking Security

Since ubiquitous coverage of 5G will not be available from day-one, it is essential to provide 4G to 5G interworking solutions that give a migration path to stand-alone 5G network. There are two cases of interworking we discuss in this Section 1 Non-Stand Alone (NSA) case, this is discussed in detail here and 2 case where 5G stand-alone and 4G networks are connected to each other and UE moves between the networks, this case is briefly discussed.

Several interworking scenarios are defined for 5G [7, 15]. These scenarios are also know as dual-connectivity since the UE connects with both NR and LTE. The starting step defined by 3GPP is NSA as depicted in Figure 10, this is also known as Option 3 or E-UTRA-NR Dual Connectivity (EN-DC), where both LTE (E-UTRA) and NR connects to the 4G core network. In case of NSA the UE and the Master eNB (MeNB) establish the RRC connection. The MeNB computes and delivers the K_{SgNB} to the Secondary gNB (SgNB) along with the UE NR security capabilities in the SgNB addition request message. The UE also derives the same K_{SgNB}. The MeNB checks whether the UE has 5G NR capabilities and access rights to SgNB. The SgNB selects the algorithm and notifies to the MeNB in SgNB addition

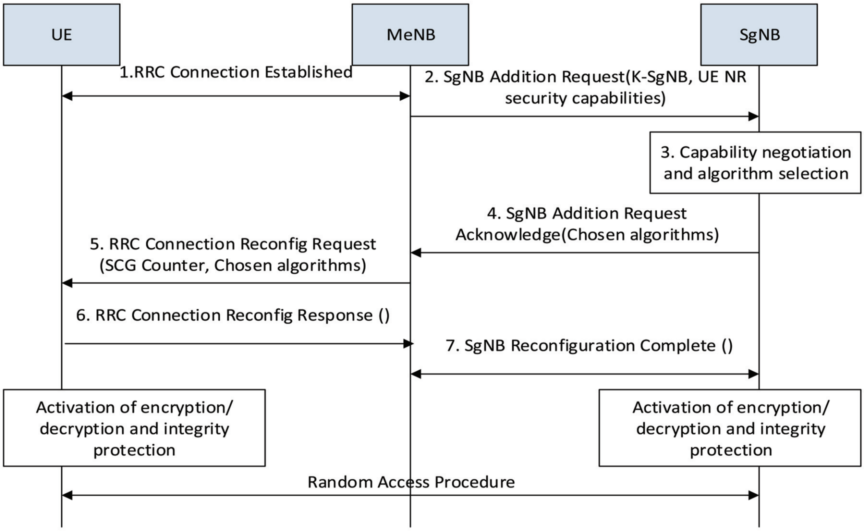

UE
MeNB
SgNB
1.RRC Connection Established
2. SgNB Addition Request(K-SgNB, UE NR security capabilities)
3. Capability negotiation and algorithm selection
4. SgNB Addition Request Acknowledge(Chosen algorithms)
5. RRC Connection Reconfig Request (SCG Counter, Chosen algorithms)
6. RRC Connection Reconfig Response ()
7. SgNB Reconfiguration Complete ()
Activation of encryption/ decryption and integrity protection
Activation of encryption/ decryption and integrity protection
Random Access Procedure

Figure 10 EN-DC procedure with SgNB encryption/decryption and integrity protection activation.

request acknowledgement message. The MeNB sends the RRC Connection Reconfiguration Request with SCG Counter parameter to the UE instructing it to configure the new DRBs and/or SRB for the SgNB and compute the S-$\mathrm{K}_{\mathrm{gNB}}$. The UE computes the S-$\mathrm{K}_{\mathrm{gNB}}$ and sends the RRC Reconfiguration Complete to the MeNB activating encryption/decryption and integrity protection. The MeNB then sends SgNB Reconfiguration Complete to the SgNB over the X2-C to inform the configuration result and following this, the SgNB can activate the chosen encryption/decryption and integrity protection with UE. Unlike dual connectivity in 4G network, RRC messages are exchanged between UE and SgNB, thus keys such as $\mathrm{K}_{\mathrm{SgNB-RRC-int}}$ as well as $\mathrm{K}_{\mathrm{SgNB-UP-enc}}$ used for integrity and confidentiality protection of RRC messages as well as UP are derived. Integrity protection for UP will not be used in EN-DC case. Use of confidentiality protection is optional for both UP and CP.

Security solution for mobility between 4G and 5G networks is similar to that specified for 4G [1, 19]. There are various situations such as state of device and security contexts available. Handover will happen between 4G and 5G incase UE is in active state. Identity, be it SUCI or temporary identity, and key identity of security context will be used to locate the security context in the network and derive a mapped security context ($\mathrm{K}_{\mathrm{AMF}}$ to $\mathrm{K}_{\mathrm{ASME}}$ for 5G to 4G or vice versa) for secure service continuity. For idle mode mobility mapped context could be used else existing context, if existing, will be activated. Mapped context is basically derivation of say 4G key from 5G.

10 Near Future

NSA and 5G Phase-1 gives us a taste of the new generation with mobile broadband. The next step will be solutions for IoT covering several scenarios in the form of massive Machine Type Communication (mMTC) and Ultra-Reliable and Low Latency Communications (URLLC). Where mMTC relates to very large number of devices transmitting a relatively low volume of non-delay-sensitive data and URLLC relates to services with stringent requirements for capabilities such as throughput, latency and availability.

For (mMTC) very low data-rates, going down to few bits per day, we will have to consider the extent of security (be it authentication, confidentiality, integrity or otherwise) that can be provisioned. Several IoT or Machine-to-Machine (M2M) services and devices fall under this category, examples are temperature sensors giving hourly updates, sensors on farm animals giving vital signature couple of times a day etc. Such devices will also be resource constrained in terms of battery, computation and memory. This brings us to several requirements on security like complete security related message sequence, e.g. authentication, should not run for every communication and even when run, they should be performed with minimum payload and round-trip. Other requirement will be to reduce security related bits, e.g. integrity, for every communication. Security and cryptographic algorithms must be energy efficient and optimized to work for resource constrained devices.

On the other end (URLLC) are high data-rate devices with potentially higher battery and computational resources; examples include cars, Industrial IoT (IIoT) devices like machineries in factories and virtual or augmented reality (VR or AR) devices used for gaming or real-time services. Provisioning of higher data rates also means that complexity of security functions should be considered to avoid processing delay. At the same time, higher data rates are provisioned by decreasing the overhead bits in radio interface that in turn has implications on bits that can be budgeted for security.

11 Conclusion

Overview of 5G Phase-1 security requirements and solutions is presented in this paper. Major differences from 4G security are the trust model, key hierarchy, security for inter-operator network, privacy and service based architecture security. Current specification supports security for 4G to 5G migration and interworking with 4G. The 5G phase 2 specifications will provide enhanced security for scenarios covered by mMTC and URLLC.

References

[1] 3GPP TS 33.401, "Technical Specification Group Services and System Aspects: 3GPP System Architecture Evolution (SAE) Security architecture", Release 15, v 15.3.0, March 2018.

[2] 3GPP TS 33.501, "Security architecture and procedures for 5G system", Release 15, v 15.0.0, March 2018.

[3] 3GPP TS 24.501, "Non-Access-Stratum (NAS) protocol for 5G System (5GS)", Release 15, v 1.0.0, March 2018.

[4] 3GPP TS 38.331, "NR-Radio Resource Control (RRC) protocol specification", Release 15, v 15.0.0, March 2018.

[5] 3GPP TS 38.470, "NG-RAN: F1 general aspects and principles", Release 15, v 15.0.0, March 2018.

[6] 3GPP TS 38.472, "NG-RAN: F1 signalling transport", Release 15, v 15.0.0, December 2017.

[7] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) - Overall description", Release 15, v 15.0.0, March 2018.

[8] 3GPP TS 33.102, "3G Security - Security architecture", Release 14, v 14.1.0, March 2017.

[9] RFC 5448, "Improved Extensible Authentication Protocol Method for 3rd Generation Authentication and Key Agreement (EAP-AKA')", Nokia, May 2009.

[10] 3GPP TS 23.502, "Procedures for the 5G System", Release 15, v 15.1.0, March 2018.

[11] 3GPP TS 33.210, "3G security; Network Domain Security (NDS); IP network layer security", Release 14, v 14.0.0, December 2016.

[12] RFC-7296, "Internet Key Exchange Protocol Version 2 (IKEv2)".

[13] RFC-7321: "Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)".

[14] RFC-3748: "Extensible Authentication Protocol (EAP)".

[15] NEC White paper, "Making 5G a Reality", 2018, https://www.nec.com/en/global/solutions/nsp/5g_vision/doc/wp2018ar.pdf.

[16] Tobias Engel. (December 2014). "SS7: Locate. Track. Manipulate", http://berlin.ccc.de/∼tobias/31c3-ss7-locate-track-manipulate.pdf

[17] GSMA RIFS: "Diameter Roaming Security – Proposed Permanent Reference Document".

[18] 3GPP TS 33.899, "Study on the security aspects of the next generation system", Release 14, v 1.3.0, August 2017.

[19] Anand R. Prasad and Seung-Woo Seo, Security in Next Generation Mobile Networks: SAE/LTE and WiMAX, River Publishers, September 2011.

Biographies

Anand R. Prasad, Dr. & ir., Delft University of Technology, The Netherlands, is Chief Advanced Technologist, Executive Specialist, at NEC Corporation, Japan, where he leads the mobile communications security activity. Anand is the chairman of 3GPP SA3 (mobile communications security standardization group), a member of the governing body of Global ICT Standardisation Forum for India (GISFI), founder chairman of the Security & Privacy working group and a governing council member of Telecom Standards Development Society, India. He was chairman of the Green ICT working group of GISFI. Before joining NEC, Anand led the network security team in DoCoMo Euro-Labs, Munich, Germany, as a manager. He started his career at Uniden Corporation, Tokyo, Japan, as a researcher developing embedded solutions, such as medium access control (MAC) and automatic repeat request (ARQ) schemes for wireless local area network (WLAN) product, and as project leader of the software modem team. Subsequently, he was a systems architect (as distinguished member of technical staff) for IEEE 802.11 based WLANs (WaveLAN and ORiNOCO) in Lucent Technologies, Nieuwegein, The Netherlands, during which period he was also a voting member of IEEE 802.11. After Lucent, Anand joined Genista Corporation, Tokyo, Japan, as a technical director with focus on perceptual QoS. Anand has provided business and technical consultancy to start-ups, started an offshore development center based on his concept of cost effective outsourcing models and is involved in business development.

Anand has applied for over 50 patents, has published 6 books and authored over 50 peer reviewed papers in international journals and conferences. His latest book is on “Security in Next Generation Mobile Networks: SAE/LTE and WiMAX”, published by River Publishers, August 2011. He is a series editor for standardization book series and editor-in-chief of the Journal of ICT Standardisation published by River Publishers, an Associate Editor of IEEK (Institute of Electronics Engineers of Korea) Transactions on Smart Processing & Computing (SPC), advisor to Journal of Cyber Security and Mobility, and chair/committee member of several international activities.

He is a recipient of the 2014 ITU-AJ “Encouragement Award: ICT Accomplishment Field” and the 2012 (ISC)2 Asia Pacific Information Security Leadership Achievements (ISLA) Award as a Senior Information Security Professional. Anand is Certified Information Systems Security Professional (CISSP), Fellow IETE and Senior Member IEEE and a NEC Certified Professional (NCP).

Sivabalan Arumugam received Ph.D in Electrical Engineering from Indian Institute of Technology Kanpur, India in 2008 and M.Tech degree from Pondicherry University, India, in 2000. He has 14 years of experience in Academic teaching and Research. Presently he works as Assistant General Manager for Research at NEC Mobile Network Excellence Center (NMEC), NEC Technologies India Pvt Ltd, Chennai. Prior joining NECI he was associated with ABB Global Services and Industries Limited, Bangalore as Associate Scientist. He has published more than 25 papers in various International Journals and Conferences and also participated in many National and International Conferences. In his current role, he is representing NEC for Global ICT Standards forum of India (GISFI). His research interest includes Next Generation Wireless Networks.

Sheeba Backia Mary Baskaran received her Ph.D. in Faculty of Information and Communication Engineering from Anna University, Chennai in 2017. She received her M.E. degree in Computer science and engineering from Anna University, Coimbatore and received the B.Tech. degree in Information Technology from Anna University, Chennai. She was a member of NGNLabs Anna University and was a recipient of Maulana Azad National Fellowship from 2013–2016. She has 19 months of experience in Research and Development of mobile communication networks and security standardization. She is carrying out her research in Security Solutions for 5G, Internet of Things, Public Safety network and Common API Framework. Her research interest includes LTE, LTE-Advanced, 5G, IoT Security and MAC layer protocol design. She contributes to 3GPP SA3 standard Specifications and applied for more than 5 patents in next generation network security. She has authored over 10 publications in international journals (IEEE Access, ACM, Elsevier & Springer) and conferences. She is also a reviewer for IEEE Access and Elsevier journals.

Alf Zugenmaier is teaching mobile networks and security at the Munich University of Applied Sciences. He also represents NTT DOCOMO at the 3GPP security working group of which he is vice chair. He has been contributing to security standardization in 3GPP for ten years, supporting 4G and 5G security standardization. Prior to joining the University, he worked at DOCOMO Euro-labs in Munich, Germany, and Microsoft Research in Cambridge, UK. His areas of interest are network and systems security as well as privacy.

Management, Orchestration and Charging in the New Era

Thomas Tovinger[1], Jean-Michel Cornily[2], Maryse Gardella[3], Chen Shan[4], Chen Ai[5], Anatoly Andrianov[6], Joey Chou[7], Jan Groenendijk[8], Zhang Kai[9], Zou Lan[9], Xiaowen Sun[5], Weixing Wang[10], Zhu Weihong[11] and Yizhi Yao[7]

[1]*Chairman of 3GPP SA5, Ericsson, Sweden*
[2]*Vice Chairman of 3GPP SA5 & Rapporteur, Orange, France*
[3]*Chairman of 3GPP SA5 SWG & Rapporteur, Nokia, France*
[4]*Vice Chairman of 3GPP SA5 SWG & Rapporteur, Huawei, China*
[5]*Rapporteur of 3GPP SA5, China Mobile, China*
[6]*Rapporteur of 3GPP SA5, Nokia, USA*
[7]*Rapporteur of 3GPP SA5, Intel, USA*
[8]*Rapporteur of 3GPP SA5, Ericsson, Ireland*
[9]*Rapporteur of 3GPP SA5, Huawei, China*
[10]*Rapporteur of 3GPP SA5, Nokia, China*
[11]*Rapporteur of 3GPP SA5, ZTE, China*

Received 23 April 2018;
Accepted 3 May 2018

Abstract

In December 2017, 3GPP passed two major milestones for 5G by approving the first set of 5G New Radio (NR) specifications and by putting in place the 5G Phase 1 System Architecture. These achievements have brought about the need for new management standards, as 5G adds to the ever-growing size and complexity of telecom systems.

3GPP management standards from Working Group (WG) SA5 are approaching another major milestone for 5G. With our studies on the 5G management architecture, network slicing and charging completed last year,

doi: 10.13052/jicts2245-800X.6110
This is an Open Access publication. © 2018 *the Author(s). All rights reserved.*

we are now well under way with the normative work for the first phase in 3GPP Release 15, which includes building up a new service-oriented management architecture and all the necessary functionalities for management and charging for 5G networks.

SA5's current work also includes several other work/study items such as management of Quality of Experience (QoE) measurement collection and new technologies for RESTful management protocols. However, this article will focus on the new 5G Rel-15 architecture and the main functionalities, including charging.

1 5G Networks and Network Slicing

Management and orchestration of 5G networks and network slicing is a feature that includes the following work items: management concept and architecture, provisioning, network resource model, fault supervision, assurance and performance management, trace management and virtualization management aspects. With the output of these work items, SA5 provides specified management interfaces in support of 5G networks and network slicing. An operator can configure and manage the mobile network to support various types of services enabled by 5G, for example eMBB (enhanced Mobile Broadband) and URLLC (Ultra-Reliable and Low Latency Communications), depending on the different customers' needs. The management concept, architecture and provisioning are being defined in TS (Technical Specification) 28.530 [4], 28.531 [5], 28.532 [6] and 28.533 [7].

Network slicing is seen as one of the key features for 5G, allowing vertical industries to take advantage of 5G networks and services. 3GPP SA5 adopts the network slice concept as defined in WG SA2 and addresses the management aspects. Network slicing is about transforming a Public Land Mobile Network (PLMN) from a single network to a network where logical partitions are created, with appropriate network isolation, resources, optimized topology and specific configuration to serve various service requirements.

As an example, a variety of communication service instances provided by multiple Network Slice Instances (NSIs) are illustrated in Figure 1 below. The different parts of an NSI are grouped as Network Slice Subnets (e.g. Radio Access Network (RAN), 5G Core Network (5GC) and Transport) allowing the lifecycle of a Network Slice Subnet Instance (NSSI) to be managed independently from the lifecycle of an NSI.

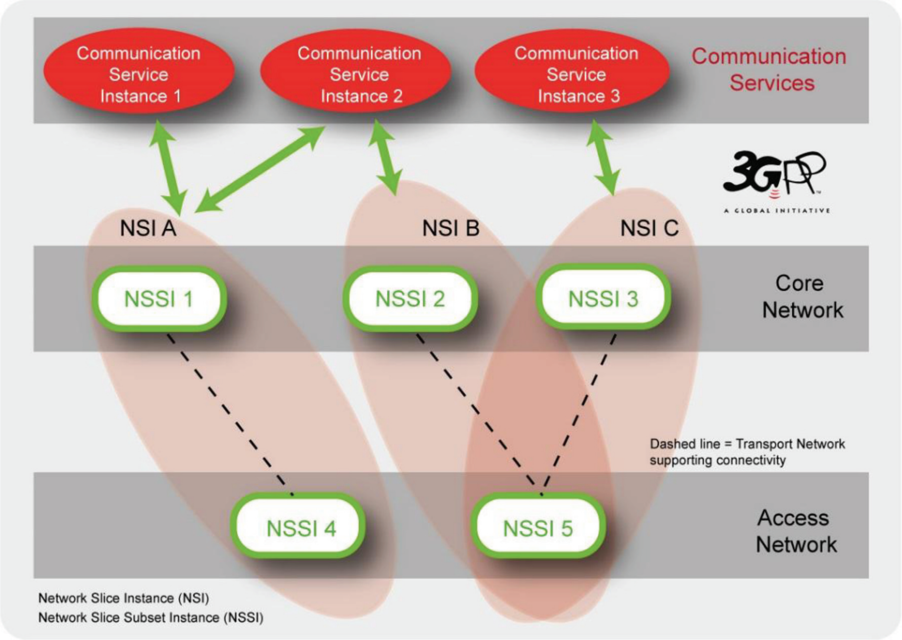

Communication Service Instance 1
Communication Service Instance 2
Communication Service Instance 3
Communication Services
NSI A
NSI B
NSI C
NSSI 1
NSSI 2
NSSI 3
Core Network
Dashed line = Transport Network supporting connectivity
NSSI 4
NSSI 5
Access Network
Network Slice Instance (NSI)
Network Slice Subset Instance (NSSI)

Figure 1 Communication service instances provided by multiple NSIs.

Provisioning of network slice instances

The management aspects of a network slice instance can be described by the following four phases and is depicted in Figure 2:

1. Preparation: in the preparation phase the network slice instance does not exist. The preparation phase includes network slice template design, network slice capacity planning, on-boarding and evaluation of the

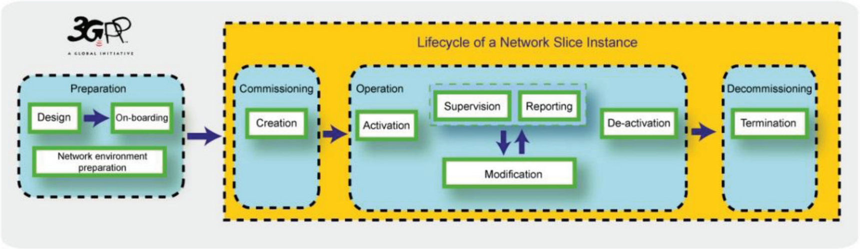


Figure 2 Lifecycle of a network slice instance.

network slice requirements, preparing the network environment and other necessary preparations required to be done before the creation of a network slice instance.

2. Commissioning: provisioning in the commissioning phase includes creation of the network slice instance. During network slice instance creation all needed resources are allocated and configured to satisfy the network slice requirements. The creation of a network slice instance can include creation and/or modification of the network slice instance constituents.
3. Operation: includes the activation, supervision, performance reporting (e.g. for Key Performance Indicator (KPI) monitoring), resource capacity planning, modification, and de-activation of a network slice instance. Provisioning in the operation phase involves activation, modification and de-activation of a network slice instance.
4. Decommissioning: network slice instance provisioning in the decommissioning phase includes decommissioning of non-shared constituents if required and removing the network slice instance specific configuration from the shared constituents. After the decommissioning phase, the network slice instance is terminated and does not exist anymore.

Similarly, provisioning for a Network Slice Subnet Instance (NSSI) includes the following operations:

- Create an NSSI;
- Activate an NSSI;
- De-active an NSSI;
- Modify an NSSI;
- Terminate an NSSI.

Roles related to 5G networks and network slicing

The roles related to 5G networks and network slicing management are depicted in Figure 3 and include: Communication Service Customer, Communication Service Provider (CSP), Network Operator (NOP), Network Equipment Provider (NEP), Virtualization Infrastructure Service Provider (VISP), Data Centre Service Provider (DCSP), NFVI (Network Functions Virtualization Infrastructure) Supplier and Hardware Supplier.

Depending on actual scenarios:

- each role can be played by one or more organizations simultaneously;
- an organization can play one or several roles simultaneously (for example, a company can play CSP and NOP roles simultaneously).

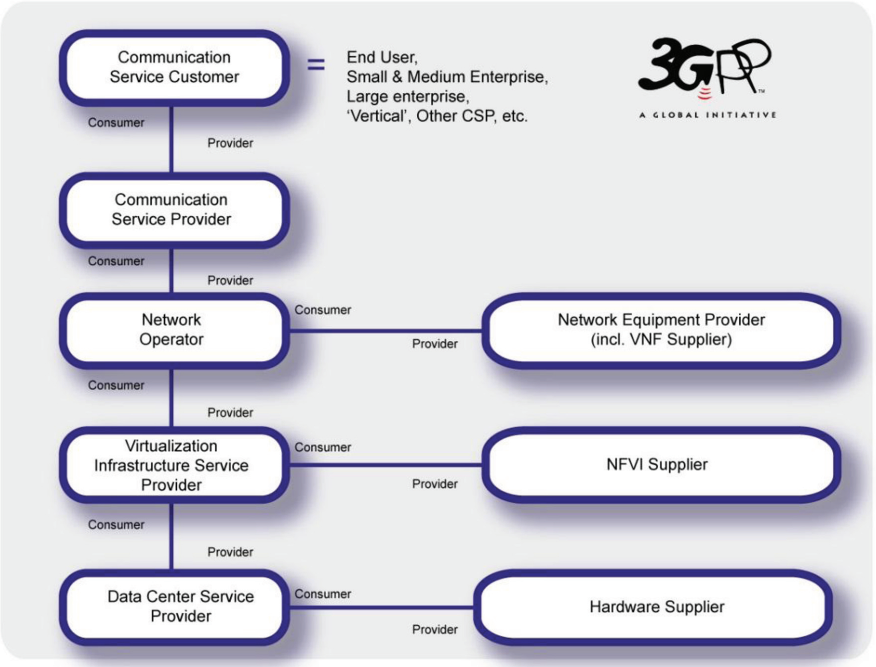

Communication Service Customer
=
End User,
Small & Medium Enterprise,
Large enterprise,
'Vertical', Other CSP, etc.
Consumer
Provider
Communication Service Provider
Consumer
Provider
Network Operator
Consumer
Provider
Network Equipment Provider (incl. VNF Supplier)
Consumer
Provider
Virtualization Infrastructure Service Provider
Consumer
Provider
NFVI Supplier
Consumer
Provider
Data Center Service Provider
Consumer
Provider
Hardware Supplier

Figure 3 Roles related to 5G networks and network slicing management.

Management models for network slicing

Different management models can be used in the context of network slicing and are depicted in Figure 4:

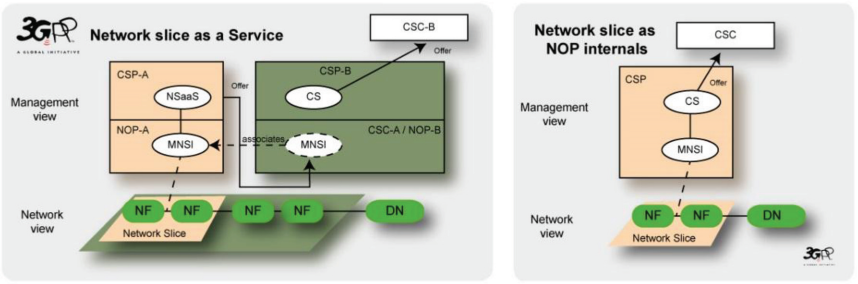


Figure 4 Management models for network slicing.

1. Network Slice as a Service (NSaaS): NSaaS can be offered by a CSP to its CSC in the form of a communication service. This service allows CSC to use the network slice instance as the end user or optionally allows CSC to manage the network slice instance as manager via management exposure interface. In turn, this CSC can play the role of CSP and offer their own services (e.g. communication services) on top of the network slice instance. The MNSI (Managed Network Slice Instance) in the figure represents a network slice instance and CS represents a communication service.
2. Network Slices as NOP internals: network slices are not part of the CSP service offering and hence are not visible to CSCs. However, the NOP, to provide support to communication services, may decide to deploy network slices, e.g. for internal network optimization purposes. This model allows CSC to use the network as the end user or optionally allows CSC to monitor the service status.

Management architecture

The 3GPP SA5 management architecture will adopt a service-oriented management architecture which is described as interaction between management service consumer and management service producer. For example, a management service consumer can request operations from management service producers on fault supervision service, performance management service, provisioning service and notification service, etc. A management service offers management capabilities. These management capabilities are accessed by management service consumers via standardized service interfaces, depicted in Figure 5, composed of individually specified management service components. The basic elements of a management service include a group of management operations and/or notifications agnostic of managed entities (Management service component type A), management information

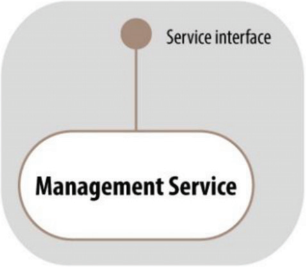


Figure 5 Management service and service interface.

represented by an information model of managed entities (Management service component type B), and performance information of the managed entity and fault information of the managed entity (Management service component type C).

Management services may reside in different management layers. For example, a network provisioning service may reside at the network and network slice management layer, and a subnetwork provisioning service may reside at the subnetwork and network slice subnet management layer.

SA5 recognizes the need for automation of management by introducing new management functions such as a Communication Service Management Function (CSMF), Network Slice Management Function (NSMF) and a Network Slice Subnet Management Function (NSSMF) to provide an appropriate abstraction level for automation.

2 Network Resource Model (NRM) for 5G Networks and Network Slicing

To support management and orchestration of 5G networks, the Network Resource Model (NRM) representing the manageable aspects of 5G networks needs to be defined, according to 5G network specifications from other 3GPP working groups as well as considering requirements from 5G management architecture and operations.

The 5G NRM specifications family includes 4 specifications: TS 28.540 [8] and TS 28.541 [9] for NRM of NR and NG-RAN (Next Generation Radio Access Network), TS 28.542 [10] and TS 28.543 [11] for NRM of 5G core network.

According to content categorization, 5G NRM specifications can be divided into 3 parts:

- Requirements, also known as stage 1,
- Information Model definitions also known as stage 2, and
- Solution Set definitions also known as stage 3.

Identified in the specifications of 5G NRM requirements (TS 28.540 [8] and TS 28.542 [10]), the NRM of 5G network comprises NRM for the 5G core network (5GC) and NRM for 5G radio access network (i.e. NR and NG-RAN). The 5GC NRM definitions support management of 5GC Network Functions, respective interfaces as well as AMF Set and AMF Region. The NR and NG-RAN NRM definitions cover various 5G radio networks connectivity options (standalone and non-standalone radio node deployment options) and architectural options (NR nodes with or without functional split).

The 5G Information Model definitions specify the semantics and behavior of information object class attributes and relations visible on the 5G management interfaces, in a protocol and technology neutral way (UML (Universal Modeling Language) as protocol-neutral language is used). The 5G Information Model is defined according to 5GC, NR and NG-RAN specifications. For example, in 3GPP TS 38.401, the NR node (gNB) is defined to support three functional split options (i.e. non-split option, two split option with CU (Central Unit) and DU (Distributed Unit), three split option with CU-CP (Control Plane), CU-UP (User Plane) and DU), so in the NR NRM Information Model, corresponding Information Object Class (IOC) is defined for each network function of gNB specified, and different UML diagrams show the relationship of each gNB split option respectively. Further, in the 5G Information Model definitions, the existing Generic NRM Information Service specification (TS 28.622 [14]) is referenced to inherit the attributes of generic information object classes, and the existing EPC (Evolved Packet Core) NRM Information Service specification (TS 28.708 [15]) is referenced for 5GS (5G System)/EPS (Evolved Packet System) interworking relationships description.

Besides 5G networks NRM definitions in the abovementioned four specifications, the information model of network slice and network slice subnet is specified in TS 28.532 [6].

Finally, NRM Solution Set definitions map the Information Model definitions to a specific protocol definition used for implementations. According to recommendation from TR (Technical Report) 32.866 [22] (Study on RESTful based Solution Set), JSON (JavaScript Object Notation) is expected to be chosen as data modelling language to describe one 5G NRM Solution Set.

3 Fault Supervision of 5G Networks and Network Slicing

Fault Supervision is one of the fundamental functions for the management of a 5G network and its communication services. For the fault supervision of 5G networks and network slicing, the following 3GPP TSs are being specified:

1. TS 28.545 [12] “Management and orchestration of networks and network slicing; Fault Supervision (FS); Stage 1”, which includes:
    - The use cases and requirements for fault supervision of 5G networks and network slicing.
    - The definitions of fault supervision related management services

2. TS 28.546 [13] "Management and orchestration of networks and network slicing; Fault Supervision (FS); Stage 2 and stage 3", which includes the definition of:

    - Operations of the fault supervision related management services (e.g. getAlarmList, subscribeAlarmNotify, unsubscribeAlarmNotify, acknowledgeAlarms, clearAlarms, unacknowledgeAlarms, etc.); (Stage 2)
    - Notifications (notifyNewAlarm, notifyClearedAlarm, notifyAlarmListRebuilt, notifyAckStateChanged, notifyChangedAlarm, etc.); (Stage 2)
    - Alarm related information models (e.g. alarmInformation, alarmList, etc.); (Stage 2)
    - Solution set(s) (e.g. RESTful HTTP-based solution set for Fault Supervison); (Stage 3)
    - New event types and probable causes if necessary.

4 Assurance Data and Performance Management for 5G Networks and Network Slicing

The 5G network is designed to accommodate continuously fast increasing data traffic demand, and in addition, to support new services such as IoT (Internet of Things), cloud-based services, industrial control, autonomous driving, mission critical communications, etc. Such services may have their own performance criteria, such as massive connectivity, extreme broadband, ultra-low latency and ultra-high reliability.

The performance data of the 5G networks and NFs (Network Functions) are fundamental for network monitoring, assessment, analysis, optimization and assurance. For the services with ultra-low latency and ultra-high reliability requirements, any faults or performance issues in the networks can cause service failure which may result in serious personal and property losses. Therefore, it is necessary to be able to collect the performance data in real-time (e.g., by performance data streaming), so that the analytic applications (e.g., network optimization, Self-Organizing Networks (SON), etc.) could use the performance data to detect any network performance problems, predict the potential issues and take appropriate actions quickly or even in advance.

For network slicing, the communication services are provided on top of the end-to-end network slice instances, so the performance needs to be monitored from end-to-end point of view.

The end to end performance data of 5G networks (including sub-networks), NSIs (Network Slice Instances) and NSSIs (Network Slice Subnet Instances) are vital for operators to know whether they can meet the communication service requirement.

The performance data may be used by various kinds of consumers, such as network operator, SON applications, network optimization applications, network analytics applications, performance assurance applications, etc.

To facilitate various consumers to get their required performance data, the following items are being pursued by this WI:

- performance management services for managing the measurement jobs for collecting the NF/NSSI/NSI/network performance data (the network performance data is not specific to network slicing);
- performance management services for reporting the NF/NSSI/NSI/ network performance data, including performance data file reporting and performance data streaming;
- performance measurements (including the data that can be used for performance assurance) for 3GPP NFs;
- end to end KPIs, performance measurements (including the data that can be used for performance assurance) for NSIs, NSSIs and networks (where the performance data is not specific to network slicing).

5 Management and Virtualization Aspects of 5G Networks

For 5G networks, it is expected that most of the network functions will run as software components on operators' telco-cloud systems rather than using dedicated hardware components. Besides the virtualization for Core Network (including 5GC, EPC and IMS (IP Multimedia Subsystem)), the NG-RAN architecture is being defined with functional split between central unit and distributed unit, where the central unit can also be virtualized.

SA5 conducted a study on management aspects of the NG-RAN that includes virtualized network functions, and has concluded in TR 32.864 [21] that the existing specifications (related to management of mobile networks that include virtualized network functions) need some enhancements for 5G. The enhancements are mainly on the interactions between 3GPP management system and external management systems (e.g., ETSI NFV

(Network Functions Virtualization) MANO (Management and Orchestration)) for the following aspects:

- Management requirements and architecture;
- Life Cycle Management (e.g., PNF management);
- Configuration Management;
- Performance Management;
- Fault Management.

There are gaps identified between 3GPP SA5 requirements and ETSI ISG NFV solutions to support the required enhancements, 3GPP SA5 is in cooperation with ETSI ISG NFV to solve these gaps.

Although the need for enhancements found in TR 32.864 [21] is to target 5G, SA5 generally agreed that these enhancements can be used for 4G as well. So the specifications for management of mobile networks that include virtualized network functions are being made generally applicable to both 4G and 5G networks. However, as 5G management will be based on a new service-oriented management architecture, the management and virtualization aspects of 5G networks need to be updated to adapt to the new architecture.

6 5G Charging System Architecture and Service Based Interface

Commercial deployment of the Rel-15 5G System will not be possible without capabilities for Operators to be able to monetize the various set of features and services which are specified in TS 23.501 [25], TS 23.502 [26] and TS 23.503 [27]. This is defined under the charging framework, which includes e.g. real-time control of subscriber's usage of 5G Network resources for charging purpose, or per-UE (User Equipment) data collection (e.g. for Charging Data Record (CDR) generation) which can also be used for other purposes e.g. analytics.

SA5 has investigated, during a study period in 2017, on how charging architecture should evolve, which key features should be specified as part of charging capabilities, and which alternative amongst charging solutions should be selected, to better support the first commercial 5G system deployment. Based on the study results, the charging architecture evolution and selected Rel-15 key functionalities for 5G system are under ongoing normative phase through development of a complete set of specifications (architecture, functionalities and protocols) A brief overview of the charging coverage for the Rel-15 5G system is provided in this article.

Service Based Interface

One key evolution of the charging architecture is the adoption of a service based interface integrated into the overall 5G system service based architecture, enabling deployments of charging functions in virtualized environment and use of new software techniques. The new charging function (CHF) and service based interface Nchf are introduced in the 5G system architecture, as shown in Figure 6 below, allows charging services to be offered to authorized network functions. The "converged online and offline charging" service will be defined. In addition to charging services, the CHF also exposes the "Spending Limit Control" service for the PCF (Policy Control Function) to access policy counter(s) status information.

While offering the service based interface to the 5G system, the overall converged charging system will be able to interface the billing system as for the existing system (e.g. 4G) to allow Operators to preserve their billing environment. These evolutions are incorporated in the TS 32.240 [16] umbrella architecture and principles charging specification. The services, operations and procedures of charging using Service Based Interface will be specified in a new TS 32.290 [18], and TS 32.291 [19] will be the stage 3 for this interface.

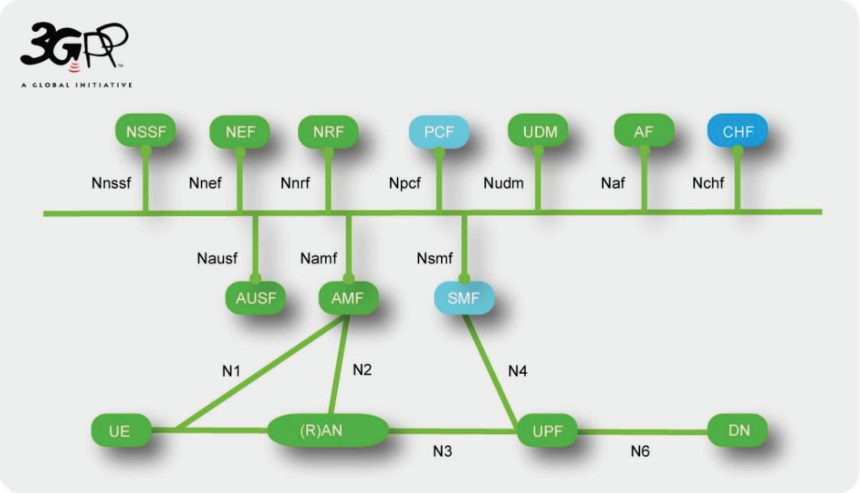


Figure 6 New charging function (CHF) and service based interface Nchf.

5G Data connectivity charging

The “5G Data connectivity charging”, achieved by the Session Management Function (SMF) invocation of charging service(s) exposed by the charging function (CHF), will be specified in a new TS 32.255 [17], encompassing the various configurations and functionalities supported via the SMF, which are highlighted below.

For 3GPP network deployments using network slicing, by indicating to the charging system which network slice instance is serving the UE during the data connectivity, the Operator will be able to apply business case charging differentiation. Further improvements on flexibility in charging systems deployments for 5G network slicing will be explored in future releases.

The new 5G QoS (Quality of Service) model introduced to support requirements from various applications in data connectivity, is considered to support QoS based charging for subscriber’s usage. 5G QoS-based charging is also defined to address inter-Operator’s settlements (i.e. between VPLMN (visited PLMN (Public Land Mobile Network)) and HPLMN (Home PLMN)) in roaming Home-routed scenario.

All charging aspects for data services in Local breakout roaming scenarios will be further considered.

In continuation with existing principles on Access type traffic charging differentiation, the two Access Networks (i.e. NG-RAN and untrusted WLAN access) supported in Rel-15 are covered.

Charging capabilities encompass the various functionalities introduced in the 5G system to support flexible deployment of application functions (e.g. edge computing), such as the three different Session and Service Continuity (SSC) modes and the Uplink Classifiers and Branching Points.

Charging continuity for interworking and handover between 5G and existing EPC is addressed.

In 5G Multi-Operator Core Network sharing architecture (i.e. shared RAN), identification of the PLMN that the 5G-RAN resources were used to convey the traffic, allows settlements between Operators.

The stage 3 for “5G data connectivity charging” will be available in TS 32.298 [20] for the CDRs’ ASN.1 (Abstract Syntax Notation 1) definition and in TS 32.291 [19] for the data type definition in the protocol used for the service based interface.

7 5G Trace Management

Subscriber and Equipment Trace can provide detailed information at session level on one or more specific users or devices. The collected information is useful in various use cases: e.g. troubleshooting triggered by an end user complain, or network performance monitoring and optimization.

SA5 is mandated to take the lead on Trace related normative work in cooperation with RAN and CT WGs, and SA5 is now specifying management and signalling trace activation mechanisms for 5GC and NG-RAN ensuring that subscriber and equipment trace capabilities are supported in 5G on par with UMTS (Universal Mobile Telecommunications System) and LTE (Long Term Evolution) systems.

The 5G Trace activation mechanisms specified by SA5 have been communicated to the RAN and CT WGs ensuring that 5G signaling specifications will support this important feature.

This 5G system Trace specifications comprise the following aspects:

- 5G Trace use case and requirements
- 5G Trace session activation and deactivation mechanism (including both management based and signalling based Trace activation and deactivation).
- 5G Trace control and configuration parameter definitions
- 5G Trace record data definitions and trace data collection mechanism
- 5G Trace management requirements and interface specifications in alignment with the Management and Orchestration of 5G networks and network slicing work items

8 Study on Energy Efficiency of 5G Networks

Following the conclusions of the study on Energy Efficiency (EE) aspects in 3GPP Standards, TSG SA#75 recommended initiating further follow-up studies on a range of energy efficiency control related issues for 5G networks including the following aspects:

- Definition and calculation of EE KPIs in 3GPP Systems
- Energy Efficiency control in 3GPP Systems
- Coordinated energy saving in RAN and other subsystem in 3GPP Systems
- Power consumption reduction at the site level
- Energy Efficiency in 3GPP systems with NFV
- Energy Efficiency in Self-Organizing Networks (SON).

TR 32.972 [23] (Study on system and functional aspects of energy efficiency in 5G networks) aims to:

- Identify EE KPI definitions made by ETSI TC (Technical Committee) EE, ITU-T SG5, ETSI NFV ISG (Industry Specification Group), etc., which are relevant for 5G networks, in addition to definitions made in SA TR 21.866 [24]. Such EE KPIs can be defined at various levels, incl. network and equipment levels (potentially, at virtualized network function and virtualized resource level), and per deployment scenario (dense urban, rural, etc.). With 5G, potentially, EE KPIs can be defined at network slice level;
- Identify metrics to be defined by 3GPP so as to be able to calculate the above EE KPIs for 5G networks. Such metrics might relate to data volumes, coverage area or energy consumption;
- Assess whether existing OA&M (Operation, Administration and Maintenance) mechanisms enable to control and monitor the identified metrics. In particular, check if the Integration Reference Point (IRP) for the control and monitoring of Power, Energy and Environmental (PEE) parameters for Radio Access Networks (RAN) (TS 28.304 [1], 28.305 [2], 28.306 [3]) can be applied to 5G networks. If not, identify potential new OA&M mechanisms;
- Elaborate further on the EE control framework defined in TR 21.866 [24] and identify potential gaps with respect to existing management architectures, incl. SON and NFV based architectures;
- Examine whether new energy saving functionalities might enable the 3GPP management system to manage energy more efficiently. In particular, the applicability of ETSI ES 203–237 [28] (Green Abstraction Layer; Power management capabilities of the future energy telecommunication fixed network nodes) to the management of 5G networks is to be evaluated;
- Identify potential enhancements in existing standards which could lead to achieving improved 3GPP system-wide energy efficiency.

This study requires interactions with other 3GPP working groups and SDOs (Standards Development Organisations) working on related topics, including ITU-T SG5, ETSI TC EE, ETSI NFV ISG.

9 Conclusion

The road to 5G lies ahead of us and we are moving along it swiftly, towards the next milestone for SA5; the approval of the first phase of standard service-oriented 5G management specifications in 3GPP Release 15.

The journey will not end there. Release 16 will build on the achievements of phase 1 with more services, extended information models and new measurements that will specify the management and charging of the evolving 3GPP 5G eco-system.

References

[1] 3GPP TS 28.304: "Control and monitoring of Power, Energy and Environmental (PEE) parameters Integration Reference Point (IRP); Requirements".

[2] 3GPP TS 28.305: "Control and monitoring of Power, Energy and Environmental (PEE) parameters Integration Reference Point (IRP); Information Service (IS)".

[3] 3GPP TS 28.306: "Control and monitoring of Power, Energy and Environmental (PEE) parameters Integration Reference Point (IRP); Solution Set (SS) definitions".

[4] 3GPP TS 28.530: "Management and orchestration of networks and network slicing; Concepts, use cases and requirements".

[5] 3GPP TS 28.531: "Management and orchestration of networks and network slicing; Provisioning; Stage 1".

[6] 3GPP TS 28.532: "Management and orchestration of networks and network slicing; Provisioning; Stage 2 and stage 3".

[7] 3GPP TS 28.533: "Management and orchestration of networks and network slicing; Management and orchestration architecture".

[8] 3GPP TS 28.540: "Management and orchestration of networks and network slicing; NR and NG-RAN Network Resource Model (NRM); Stage 1".

[9] 3GPP TS 28.541: "Management and orchestration of networks and network slicing; NR and NG-RAN Network Resource Model (NRM); Stage 2 and stage 3".

[10] 3GPP TS 28.542: "Management and orchestration of networks and network slicing; 5G Core Network (5GC) Network Resource Model (NRM); Stage 1".

[11] 3GPP TS 28.543: "Management and orchestration of networks and network slicing; 5G Core Network (5GC) Network Resource Model (NRM); Stage 2 and stage 3".

[12] 3GPP TS 28.545: "Management and orchestration of networks and network slicing; Fault Supervision (FS); Stage 1".

[13] 3GPP TS 28.546: "Management and orchestration of networks and network slicing; Fault Supervision (FS); Stage 2 and stage 3".

[14] 3GPP TS 28.622: "Telecommunication management; Generic Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)".

[15] 3GPP TS 28.708: "Telecommunication management; Evolved Packet Core (EPC) Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)".

[16] 3GPP TS 32.240: "Telecommunication management; Charging management; Charging architecture and principles".

[17] 3GPP TS 32.255: "Telecommunication management; Charging management; 5G Data connectivity domain charging; stage 2".

[18] 3GPP TS 32.290: "Telecommunication management; Charging management; 5G system; Services, operations and procedures of charging using Service Based Interface (SBI)".

[19] 3GPP TS 32.291: "5G system; Charging service, stage 3".

[20] 3GPP TS 32.298: "Telecommunication management; Charging management; Charging Data Record (CDR) parameter description".

[21] 3GPP TR 32.864: "Telecommunication management; Study on management aspects of virtualized network functions that are part of the New Radio (NR)".

[22] 3GPP TR 32.866: "Telecommunication management; Study on a REST(REpresentational State Transfer)-ful HTTP-based Solution Set (SS)".

[23] 3GPP TR 32.972: "Telecommunication management; Study on system and functional aspects of energy efficiency in 5G networks".

[24] 3GPP TR 21.866: "Study on Energy Efficiency Aspects of 3GPP Standards".

[25] 3GPP TS 23.501: "System Architecture for the 5G System".

[26] 3GPP TS 23.502: "Procedures for the 5G System".

[27] 3GPP TS 23.503: "Policy and Charging Control Framework for the 5G System; Stage 2".

[28] ETSI ES 203 237 V1.1.1 (2014-03): "Environmental Engineering (EE); Green Abstraction Layer (GAL); Power management capabilities of the future energy telecommunication fixed network nodes".

Biographies

Thomas Tovinger received a Master of Science degree in Engineering Physics with Computer Science specialization at Chalmers University of Technology, Gothenburg, Sweden, 1980. He joined Ericsson the same year and since 1993 he has been a standardization delegate representing Ericsson as an OSS expert in ETSI and 3GPP. He has held numerous rapporteurship and leadership positions in 3GPP SA5 since 1999, particularly Vice Chair from 2007 to 2015 and Chair from 2015 until present date. He has also co-authored a number of articles in 3GPP News and IEEE Communications Magazine.

J-M. Cornily received his Ph.D. degree in Computer Science from the University of Rennes, France, in 1988. As an Operations, Administration & Maintenance (OA&M) architect and standards expert, he actively participated in ITU-T SG15, ETSI TM and OMG. He co-authored ‘Achieving Global Information Networking’ (Artech House). As an Orange representative to 3GPP since 2008 and Vice-Chairman of 3GPP/SA5 since 2013, he has been rapporteur for topics such as e.g. 3GPP networks energy efficiency, and integration of ONAP for the management of 5G networks.

M. Gardella holds the 3GPP SA5 Charging SWG Chairman position since 2013. She is graduated from Ecole Centrale de Lyon - France (Master of Science, Engineer's degree in Telecommunications). She started her career (working for Alcatel) in GSM Mobile Industry from the early beginning, and was further involved in Mobile Networks Architectures & Protocols evolution activities. She has over 15 years of experience in the field of standardization, focusing from 2008, on "Charging Architecture and protocols" by joining the 3GPP SA5 Charging SWG. During this period, she actively contributed to support Operator's business models in their network evolution towards 3GPP new technologies (e.g. from GSM, to GPRS, UMTS, LTE and now 5G) and services. Representing Nokia, she is currently playing a key role in shaping the 3GPP 5G charging architecture and solutions to the new service based concept, for the industry to benefit from new software technologies.

Chen Shan received her Bachelor's Degree in Electronic Information Engineering from Wuhan Institute of Technology and Master degrees in Intelligent Control from Dalian University of Technology, CHINA. She joined Huawei in 2007 and she is a standardization delegate of Huawei in 3GPP (5G WI rapporteur and SA5 SWG Charging Vice chair from 2016), ITU-T and IEEE.