# 5G Radio Access Network Architecture

# 5G Radio Access Network Architecture

The Dark Side of 5G

*Edited by*

*Sasha Sirotkin*

WILEY

IEEE PRESS

This edition first published 2021


*Registered Offices*
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

*Editorial Office*
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.



*Library of Congress Cataloging-in-Publication Data*

Names: Sirotkin, Alexander, 1975- editor.
Title: 5G radio access network architecture : the dark side of 5G / Alexander Sirotkin, editor.
Description: Hoboken, NJ, USA : Wiley-IEEE Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020020729 (print) | LCCN 2020020730 (ebook) | ISBN 9781119550884 (hardback) | ISBN 9781119550891 (Adobe PDF) | ISBN 9781119550914 (ePub)
Subjects: LCSH: 5G mobile communication systems. | Computer network architectures.
Classification: LCC TK5103.25 .A148 2021 (print) | LCC TK5103.25 (ebook) | DDC 621.39/81–dc23
LC record available at https://lccn.loc.gov/2020020729
LC ebook record available at https://lccn.loc.gov/2020020730

Cover Design: Wiley
Cover Image: © Paul Cooklin/Getty Images

Set in 9.5/12.5pt STIXTwoText by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

*To my parents, Natalia and Arkadiy, for, among the many things you've given me, the best gifts of all – aspiration for knowledge and critical thinking – which to a very large extent define who I am.*

*To my children, Jonathan, Maya, Ron, and Tom – given the pace of the world we live in, in a few years when you are old enough to read this book, 5G is likely to become a thing of the past. Don't get discouraged by this in your aspirations to undertake any project you may think of, but do remember – time flies, so use it wisely.*

*To my wife Tatyana, for the understanding that, despite not saying this as often as I should, I love you dearly, respect you deeply, and value everything that you do.*

# Contents

# Preface

This is a different kind of book about 5G.

Most books on this subject (5G in particular, or wireless technologies in general) focus on the physical layer. While the physical layer (together with the access stratum protocol stack) is extremely important and is arguably the key aspect of any wireless technology responsible for most of its performance characteristics, curiously enough it is not necessarily the most important factor when determining how successful a certain wireless technology would be in the market.

The second largest category of books on wireless technologies typically focus on the core network, as it is often the core network features and design that determine the kind of services that a given technology would provide to operators and users. Without questioning the importance of the core network, we note that when it comes to the deployment of a new wireless technology by an operator, the core network is perhaps the most critical component as failures in the core may (and often do) affect the whole network and all the users. Nevertheless, in terms of deployment complexity and ultimately cost, the core network is in no way the biggest contributor to operator's efforts when deploying a network.

In terms of deployment and development complexity and cost, the biggest component of a network is actually the one that is often overlooked in literature – that is the Radio Access Network (RAN). The RAN is a collection of base stations, interconnected by a transport network, which also connects it to the core. That collection of base stations, if deployed and configured properly, is ultimately responsible for providing coverage and capacity to the network users. As the number of base stations deployed by an operator is huge (and is expected to grow substantially in 5G), the RAN is (together with spectrum acquisition) by far the biggest contributor to the cost of deploying and running a cellular network.

Unlike the other network components, design of the RAN is more art than science. That is because it is not feasible to analyze or simulate the RAN in its entirety and, therefore, there are very few objective measures of what constitutes a good RAN design. This inevitably leads to a multitude of different designs (or architectures) – some competing, some complementing each other. In this book we try to lead the reader through this maze of different RAN architectures, technical and business considerations that led to their design, and practical considerations affecting the choice of the proper architecture and deploying it successfully and in a cost-efficient manner.

Welcome to the "dark side" of 5G – one of the most important 5G aspects, which is not in the spotlight as much as it should be.

This book is accompanied by the website: www.darksideof5g.com

## Acknowledgments

This book is the result of the joint work of many contributors who used the vast domain expertise in their respective areas to make it possible. I would like to thank them all.

Furthermore, special thanks go to all the reviewers for helping to ensure correctness and consistency of the material presented in the book: Apostolos Papathanassiou, Intel Corporation; Jaemin Han, Intel Corporation; Krzysztof Kordybach, Nokia; and Markus Dominik Mueck, Intel Corporation.

## List of Contributors

***Alexander (Sasha) Sirotkin*** is a senior engineer with 20 years of experience in telecommunications, international standardization, intellectual property, machine learning, real-time systems, and open source.

Currently his primary focus areas are 4G/LTE and 5G/NR Radio Access Network (RAN) Architecture, and licensed and unlicensed spectrum integration and co-existence. In standards, Sasha contributed to 3GPP RAN2, RAN3, and RAN plenary, where he served as rapporteur for multiple specifications, as well as work and study items. Currently Sasha serves as the 3GPP RAN3 vice chairman and leads the Intel's RAN3 delegation.

In addition to 3GPP, Sasha has contributed to various other standards development organizations and industry fora, such as IEEE, O-RAN, WFA, WBA, ETSI, and 5G Americas.

Prior to working in the field of wireless (802.11/Wi-Fi and cellular) communications, Sasha was actively involved in the open source, primarily the Linux operating system. Having been an open source enthusiast since 1993, Sasha was one of the first to realize that the potential of Linux lies not so much in the desktop, but in embedded and real-time systems, which he worked to promote long before the first version of Android was conceived.

Sasha received an MSc in machine learning and BSc degrees in computer science and physics from Tel-Aviv University.

Sasha lives with his wife and children in Hod HaSharon, Israel. In his spare time, which his kids make sure he doesn't have too much of, he occasionally goes scuba diving and alpine skiing (usually not on the same day, even though that is sometimes technically possible in Israel), and practices Kyokushin Karate.

## Contributors

Reza Arefi, Intel Corporation – Washington DC, USA

Nicolas Chuberre, Thales Alenia Space – Pibrac, France

Stefano Cioni, European Space Agency – Noordwijk, the Netherlands

Alexei Davydov, Intel Corporation – Nizhny Novgorod, Russia

Thibault Deleu, Thales Alenia Space – Toulouse, France

Miltiadis Filippou, Intel Deutschland GmbH – Neubiberg, Germany

Yuri Gittik, RAD Data Communications, Ltd. – Tel Aviv, Israel

Georg Hampel, Qualcomm Incorporated – Hoboken, NJ, USA

Colby Harper, Pivotal Commware Inc. – Seattle, WA, USA

Thomas Heyn, Fraunhofer IIS – Erlangen, Germany

Ron Insler, RAD Data Communications, Ltd. – Petah Tikva, Israel

Sudeep Palat, Intel Corporation – Cheltenham, UK

Sergio Parolari, ZTE Corporation – Milan, Italy

Sridhar Rajagopal, Mavenir – Dallas, TX, USA

Leszek Raschkowski, Fraunhofer HHI – Berlin, Germany

Dario Sabella, Intel Corporation – Munich, Germany

Eiko Seidel, Nomor Research GmbH – Munich, Germany

Clare Somerville, Intel Corporation – Maidenhead, UK

Sebastian Speicher, Qualcomm Wireless LLC – Zürich, Switzerland

Yaakov (J.) Stein, RAD Data Communications, Ltd. – Jerusalem, Israel

Jianli Sun, Intel Corporation – Hillsboro, OR, USA

Feng Yang, Intel Corporation – Beijing, People's Republic of China

Vladimir Yanover, Cisco Systems, Inc. – Kfar-Saba, Israel

Andreas Neubacher, Deutsche Telekom – Korneuburg, Austria

Vishwanath Ramamurthi, Verizon Wireless – Walnut Creek, CA, USA

## Acronyms and Abbreviations

| | |
|---|---|
| 3GPP | 3rd Generation Partnership Project |
| 5G ACIA | 5G Alliance for Connected Industries and Automation |
| 5G AKA | 5G Authentication and Key Agreement |
| 5G MOCN | 5G Multi-Operator Core Network |
| 5G-PPP | 5G Infrastructure Public Private Partnership |
| 5GAA | 5G Automotive Association |
| 5GC | 5G Core |
| 5GS | 5G System |
| 5QI | 5G QoS Class Identifier |
| A/D | Analog to digital |
| AAS | Active Antenna System |
| ACK/NACK | acknowledgement/negative acknowledgement |
| ACM | Adaptive Coding and Modulation |
| ADSL | Asymmetric digital subscriber line |
| AECC | Automotive Edge Computing Consortium |
| AF | Application Function |
| AI | artificial intelligence |
| AISG | Antenna Interface Standards Group |
| AM | Acknowledged Mode |
| AMC | adaptive modulation and coding |
| AMF | Access and Mobility Management Function |
| AN | Access Network |
| AN | Access Node |
| ANDSP | Access Network Discovery and Selection Policy |
| ANR | Automatic Neighbor Relation |
| API | Application Programming Interface |
| APN | Access Point Name |
| APS | Automatic Protection Switching |
| AR | Augmented Reality |
| ARIB | Association of Radio Industries and Businesses |
| ARQ | Automatic Repeat Request |
| AS | Access stratum |
| ASF | Apache Software Foundation |

| | |
|---|---|
| ASG | aggregation site gateway |
| ASIC | application-specific integrated circuit |
| ATIS | Alliance for Telecommunications Industry Solutions |
| ATM | asynchronous transfer mode |
| AUSF | Authentication Server Function |
| B2B2C | business to business to consumer |
| BAP | Backhaul Adaptation Protocol |
| BBF | Broadband Forum |
| BBU | Baseband Unit |
| BC | Boundary Clock |
| BE | Best Effort |
| BFD | Bidirectional Forwarding Detection |
| BFRP | Beam Failure Recovery Response |
| BFRQ | Beam Failure Recovery Request |
| BGP | Border Gateway Protocol |
| BiDi | bidirectional traffic on a single fiber |
| BIOS | basic input/output system |
| BLER | Block Error Rate |
| BNetzA | Bundesnetzagentur |
| BSD | Berkeley Software Distribution |
| BSR | Buffer Status Report |
| BSS | Broadcast Satellite Services |
| BWP | Bandwidth Part |
| C-RNTI | Cell Radio Network Temporary Identifier |
| C-SON | centralized SON |
| CA | Carrier Aggregation |
| CAC | Connection Admission Control |
| CAG | Closed Access Group |
| CAPEX | Capital Expenditure |
| CB | code block |
| CBG | Code block group |
| CBRS | Citizens Broadband Radio System |
| CC | continuity check |
| CCCH | Common Control Channel |
| CCE | Control Channel Element |
| CCSA | China Communications Standards Association |
| CDM | code division multiplexing |
| CDR | Charging Data Record |
| CEPT | European Conference of Postal and Telecommunications Administrations |
| CGI | Cell Global Identifier |
| CGS | computer-generated Quadrature Phase Shift Keying (QPSK) sequence |
| CLI | Cross-Link Interference |
| CM | Configuration Management |
| CN | Core Network |
| CNF | container network function |

| | |
|---|---|
| CNI | container network interface |
| CoMP | Coordinated Multi-Point |
| CORESET | control resource set |
| COTS | Commercial Off-The-Shelf |
| CP | control plane |
| CP | Cyclic Prefix |
| CPA | Coverage Per area |
| CPP | Coverage Per Population |
| CPRI | Common Public Radio Interface |
| CPU | central processing units |
| CQI | channel quality indicator |
| CR | Change Request |
| cRAN | cloud RAN |
| CRC | Cyclic Redundancy Check |
| CRS | Cell-Specific Reference Signal |
| CSG | cell site gateway |
| CSG | Closed Subscriber Group |
| CSI | Channel State Information |
| CSR | cell site router |
| CTC | Convolution Turbo Codes |
| CU | central unit |
| CU | Centralized Unit |
| CU/DU | central unit/distributed unit |
| CU/DU | centralized unit/distributed unit |
| CUPS | Control- and user-plane separation |
| CUS | control, user, and synchronization |
| CV | connectivity verification |
| D-SON | distributed SON |
| D/A | Digital to analog |
| D/C | Data or Control |
| DA | destination address |
| DAG | Directed Acyclic Graph |
| DAS | Distributed Antenna Systems |
| DC | Dual Connectivity |
| DCCH | Dedicated Control Channel |
| DCI | Downlink Control Information |
| DCI/UCI | downlink and uplink control information |
| DCN | Dedicated Core Network |
| DDDS | Downlink Data Delivery Status |
| DDoS | Distributed Denial-of-Service |
| DECOR | dedicated core network |
| DEI | discard eligibility indicator |
| dEPC | distributed EPC |
| DetNet | deterministic networking |
| DGM | distributed GM |

| | |
|---|---|
| DL/UL | downlink/uplink |
| DM | domain manager |
| DM-RS | demodulation reference signals |
| DMRS | Demodulation Reference Symbols |
| DN | Data Network |
| DNCP | Dynamic Host Configuration Protocol |
| DNS | Domain Name System |
| DOCSIS | Data Over Cable Service Interface Specification |
| DoS | denial of service |
| DPDK | Data Plane Development Kit |
| DRB | Data Radio Bearers |
| DRX | Discontinuous Reception |
| DSCP | Differentiated Services Code Point |
| DSCP | DiffServ code point |
| DSL | digital subscriber line |
| DSP | digital signal processor |
| DTCH | Dedicated Traffic Channel |
| DU | Distributed Unit |
| DVFS | Dynamic Voltage and Frequency Scaling |
| DWDM | Dense Wavelength Division Multiplexing |
| E-RAB | E-UTRAN Radio Access Bearer |
| E-UTRA | Evolved Universal Mobile Telecommunications System Terrestrial Radio Access |
| E-UTRAN | Evolved Universal Terrestrial Radio Access Network |
| E2E | end-to-end |
| EAP | Extensible Authentication Protocol |
| EB | Exabytes |
| ECOMP | Enhanced Control, Orchestration, Management and Policy |
| EDR | Event Data Record |
| EIRP | Effective Isotropic Radiated Power |
| EM | Element Managers |
| eMBB | enhanced mobile broadband |
| EN-DC | E-UTRA-NR Dual Connectivity |
| ENG | Electronic New Gathering |
| EPC | Evolved Packet Core |
| EPL | Ethernet private line |
| EPON | Ethernet passive optical network |
| ePRC | enhanced PRC |
| EPS | Evolved Packet System |
| eRE | eCPRI Radio Equipment |
| eREC | eCPRI Radio Equipment Control |
| ESMC | Ethernet Synchronization Messaging Channel |
| ETSI | European Telecommunications Standards Institute |
| EVM | error vector magnitude |
| EVPL | Ethernet Virtual Private Line Service |

| | |
|---|---|
| F1-C | control-plane part of the F1 interface |
| F1-U | F1 User-Plane |
| F1AP | F1 Application Protocol |
| FCS | frame check sequence |
| FDD | Frequency Division Duplexing |
| FEC | Forward Error Correction |
| FFT | Fast Fourier Transform |
| FHBW | fronthaul bandwidth |
| FIB | Forwarding Information Base |
| FM | Fault Management |
| FOMA | Freedom of Mobile Multimedia Access |
| FOSS | free and open source software |
| FPGA | field programmable gate array |
| FRER | Frame Replication and Elimination for Reliability |
| FRR | fast reroute |
| FSF | Free Software Foundation |
| FSPF | free space propagation formula |
| FSS | Fixed Satellite Services |
| GAA | General Authorized Access |
| GEO | geostationary orbit |
| GGSN | Gateway GPRS Support Node |
| gNB-CU | gNB central unit |
| gNB-CU-UP | centralized user-plane node |
| gNB-DU | gNB distributed unit |
| GNSS | Global Navigation Satellite System |
| GoS | Grade of Service |
| GP | Guard Period |
| GPL | General Public License |
| GPL | GNU General Public License |
| GPON | gigabit passive optical network |
| GPP | general purpose compute |
| GPRS | General Packet Radio System |
| GPU | graphic processing unit |
| GSA | Global mobile Suppliers Association |
| GSMA | GSM Association |
| GTP | GPRS Tunneling Protocol |
| GTP-U | GPRS Tunneling Protocol User Plane |
| GUAMI | Globally Unique AMF ID |
| HAPS | High Altitude Platforms |
| HARQ | Hybrid ARQ |
| HEO | high elliptical orbit |
| HetNet | heterogeneous network |
| HFN | Hyper Frame Number |
| HPLMN | Home Public Land Mobile Network |
| HSS | Home Subscriber Server |

| | |
|---|---|
| I/Q | In-phase & Quadrature |
| IAB | Integrated Access-Backhaul |
| IE | Information Element |
| IEEE | Institute of Electrical and Electronics Engineers |
| IET | interspersing express traffic |
| IETF | Internet Engineering Task Force |
| iFFT | inverse FFT |
| IIOT | Industrial Internet of Things |
| IMS | IP multimedia subsystem |
| IMT-2020 | International Mobile Telecommunications-2020 |
| IOC | Information Object Class |
| IoT | Internet of Things |
| IPR | intellectual property rights |
| ISG | Industry Specification Group |
| ITS | Intelligent Transport Systems |
| ITU | International Telecommunication Union |
| ITU-R | ITU Radiocommunication Sector |
| ITU-T | International Telecommunication Union Telecommunication Standardization Sector |
| IWF | Interworking Function |
| JSON | JavaScript Object Notation |
| K8S | Kubernetes |
| KPI | Key Performance Indicators |
| KQI | key quality indicator |
| L1-RSRP | Layer 1 reference signal received power |
| L3VPN | Layer 3 VPN |
| LAA | licensed assisted access |
| LAG | link aggregation |
| LBT | Listen-Before-Talk |
| LCM | Life Cycle Management |
| LDPC | Low Density Parity Check |
| LEO | low -earth orbit |
| LFA | Loop Free Alternates |
| LLC | logical link control |
| LLS | Lower-Layer Split |
| LMLC | Low Mobility Large Cell |
| LPI | Low Power Idle |
| LPWA | low-power wide area |
| LSA | Licensed Shared Access |
| LSP | label switched path |
| LSR | label switch router |
| LTE | Long-Term Evolution |
| LWA | LTE-WLAN Aggregation |
| MAC | Medium Access Control |
| MANO | Management and Network Orchestration |

| | |
|---|---|
| MBB | Mobile Broadband |
| MCC | Mobile Country Code |
| MCG | Master Cell Group |
| MCL | maximum coupling loss |
| MCS | Modulation Coding Scheme |
| MCS/MPS | mission-critical and priority services |
| MDT | Minimization of Drive Tests |
| MEAO | MEC application orchestrator |
| MEC | Mobile Edge Compute |
| MEC | Multi-access edge computing |
| MEO | Mobile Edge Orchestrator |
| MEO | medium earth orbit |
| MEPM | Mobile Edge Platform Manager |
| MIB | Master Information Block |
| MIMO | Multiple-Input and Multiple-Output |
| MIT | Massachusetts Institute of Technology |
| ML | machine learning |
| MLB | mobility load balancing |
| MME | Mobility Management Element |
| MME | Mobility Management Entity |
| MN | Master Node |
| MNC | Mobile Network Code |
| MnF | management function |
| MNO | Mobile Network Operators |
| MnS | management service |
| MOI | Managed Object Instance |
| MPLS | multiprotocol label switching |
| MPLS-TP | MPLS Transport Profile |
| MR-DC | Multi-Radio Dual Connectivity |
| MRO | Mobility Robustness Optimization |
| MSI | Minimum System Information |
| MSS | Mobile Satellite Services |
| MT | mobility termination |
| MTC | Machine Type Communication |
| MU-MIMO | multi-user MIMO |
| N3IWF | Non-3GPP Interworking Function |
| NaaS | Network-as-a-Service |
| NAS | non-access stratum |
| NE | Network Elements |
| NE-DC | NR-E-UTRA dual connectivity |
| NEF | Network Exposure Function |
| NF | network function |
| nFAPI | Network FAPI |
| NFMF | Network Function Management Function |
| NFV | Network Function Virtualization |

| | |
|---|---|
| NFV/SDN | Network Function Virtualization and Software Defined Networks |
| NFVI | network function virtualization infrastructure |
| NFVO | network function virtualization orchestrator |
| NG-AP | NG Application Protocol |
| NG-C | NG control plane |
| NG-RAN | 5G Radio Access Network |
| NG-U | NG user plane |
| NGAP | NG Application Protocol |
| NGEN-DC | E-UTRA-NR dual connectivity |
| NGFI | Next Generation Fronthaul Interface |
| NGMN | Next Generation Mobile Networks |
| NHN | Neutral Host Network |
| NHOP | next hop |
| NIC | Network Interface Card |
| NID | network ID |
| nLOS | non-line-of-sight |
| NM | network manager |
| NMM | Network Monitor Mode |
| NMS | network management system |
| NNHOP | next next hop |
| NPN | Non-public networks |
| NR | New Radio |
| NR-DC | NR-NR dual connectivity |
| NR-U | NR user plane |
| NRF | Network Repository Function |
| NRM | Network Resource Model |
| NRPPa | NR Positioning Protocol A |
| NSA | Non-Standalone |
| NSI | Network Slice Instance |
| NSMF | Network Slice Management Function |
| NSSAI | Network Slice Selection Assistance Information |
| NSSF | Network Slice Selection Function |
| NSSI | Network Slice Subnet Instance |
| NSSMF | Network Slice Subnet Management Function |
| NSSP | network slice selection policies |
| NTN | Non-terrestrial network |
| NTP | network time protocol |
| NWDAF | network data analytics function |
| O-DU | O-RAN Distribution Unit |
| O-RAN | Open Radio Access Network |
| O-RU | O-RAN radio unit |
| OAI | Open Air Interface |
| OAM | Operation, Administration and Maintenance |
| OAM | operations, administration and management |
| OAM | Operations, Administration, and Maintenance |

| | |
|---|---|
| OBSAI | Open Base Station Architecture Initiative |
| OC | OpenCellular |
| OEM | original equipment manufacturer |
| OFDM | orthogonal frequency division multiplexing |
| OIF | Optical Internetworking Forum |
| ONAP | Open Networking Automation Platform |
| OPEN-O | OPEN-Orchestrator Project |
| OPEX | Operational Expenditure |
| ORAN FH | O-RAN Fronthaul |
| ORI | Open Radio equipment Interface |
| OSA | OpenAirInterface Software Alliance |
| OSI | Open Source Initiative |
| OSI | Other System Information |
| OSM | Open Source MANO |
| OSS | Operations Support System |
| OTA | over-the-air |
| OTN | Optical Transport Network |
| OVS | Open Virtual Switch |
| OWAMP | One-Way Active Measurement Protocol |
| P | polling bit |
| P-GW | Packet Data Network Gateway |
| PAL | Priority Access License |
| PAPR | peak to average power ratio |
| PBBN | Provider Backbone Bridge Network |
| PBCH | Physical Broadcast Channel |
| PBR | Prioritized Bit Rate |
| PCE | Path Computation Element |
| PCell | Primary Cell |
| PCF | Policy Control Function |
| PCI | Physical Cell Identity |
| PCP | priority code point |
| PCRF | Policy and Charging Rules Function |
| PDB | Packet Delay Budget |
| PDCCH | Physical Downlink Control Channel |
| PDCP | Packet Data Convergence Protocol |
| PDCP-RLC | Packet Data Convergence Protocol–Radio Link Control |
| PDH | plesiochronous digital hierarchy |
| PDN | Packet Data Network |
| PDP | packet data protocol |
| PDSCH | Physical Downlink Shared Channel |
| PDU | Protocol Data Unit |
| PDV | packet delay variation |
| PE | Provider Edge |
| PF | Paging Frame |
| PFD | power flux density |

| | |
|---|---|
| PGW | PDN Gateway |
| PGW-C | PGW control-plane function |
| PHY | Physical Layer |
| PLL | Phase Locked Loop |
| PLMN | Public Land Mobile Network |
| PLR | Packet Loss Ratio |
| PM | Performance Monitoring |
| PMI | precoding matrix indicator |
| PNF | physical network function |
| PNI-NPN | Public-network-integrated non-public network |
| PO | Paging Occasion |
| PON | Passive Optical Network |
| PoP | point of presence |
| PoPs | Points of Presence |
| PPI | Paging Policy Indicator |
| PRACH | Physical Random Access Channel |
| PRB | Physical Resource Block |
| PRC | primary (frequency) reference clock |
| PREOF | Packet Replication, Elimination, and Ordering Functions |
| PRG | Precoding Resource Group |
| PRTC | Primary Reference Time Clock |
| PSCell | Primary Secondary Cell Group Cell |
| PSS | Primary Synchronization Signal |
| PT-RS | phase tracking reference signals |
| PTP | Precision Time Protocol |
| PUCCH | Physical Uplink Control Channel |
| QFI | QoS Flow Identifier |
| QFI | QoS Flow Indicator |
| QoE | Quality of Experience |
| QoS | Quality of Service |
| QSFP | quad small form-factor pluggable |
| RACH | Random Access Channel |
| RAN | Radio Access Network |
| RAR | Random Access Response |
| RAT | Radio Access Technology |
| RATs | radio access technologies |
| RDI | reflective QoS flow to DRB mapping Indication |
| RE | Radio Equipment |
| REC | Radio Equipment Controller |
| REG | Resource Element Group |
| RIC | RAN intelligent controller |
| RIT | Radio Interface Technology |
| RLC | Radio Link Control |
| RLF | Radio Link Failure |
| RMSI | Remaining Minimum System Information |

| | |
|---|---|
| RNA | RAN Notification Area |
| RNI | radio network information |
| RNL | Radio Network Layer |
| RNTI | Radio Network Temporary Identifier |
| RoE | Radio over Ethernet |
| RoHC | Robust Header Compression |
| ROI | Return on Investment |
| RQI | Reflective QoS Indicator |
| RRC | RAN Control protocol |
| RRH | Remote Radio Head |
| RRM | Radio Resource Management |
| RSSI | Received Signal Strength Indicator |
| RSU | Road Side Unit |
| RSVP | Resource Reservation Protocol |
| RTT | Round Trip Time |
| RU | radio unit |
| RU | Remote Unit |
| RV | Redundancy Version |
| S-GW | Serving Gateway |
| S-NSSAI | Single Network Slice Selection Assistance Information |
| S1-AP | S1 Application Protocol |
| SA | source address |
| SAS | Spectrum Access System |
| SBA | Service-based architecture |
| SC | Software Community |
| SCEF | Service Capability and Exposure Function |
| SCell | Secondary Cell |
| SCG | Secondary Cell Group |
| SCS | subcarrier spacing |
| SCTP | Stream Control Transmission Protocol |
| SD | Slice Differentiator |
| SDAP | Service Data Adaptation Protocol |
| SDH | Synchronous Digital Hierarchy |
| SDN | Software Defined Networks |
| SDO | Standards Developing Organization |
| SDR | software-defined radio |
| SDU | Service Data Unit |
| SEQ | number of sequences |
| SFI | Slot Format Indicator |
| SFN | System Frame Number |
| SGSN | Serving GPRS Support Node |
| SGW | Serving Gateway |
| SGW-C | SGW control-plane function |
| SI | Segmentation Information |
| SI | System information |

| | |
|---|---|
| SIB | System Information Broadcast |
| SIB1 | System Information Block 1 |
| SLA | Service Level Agreement |
| SLO | service level objective |
| SmartNIC | smart network interface controller |
| SMF | Session Management Function |
| SN | Secondary Node |
| SN | Sequence Number |
| SNPN | Stand-alone non-public network |
| SO | Segment Offset |
| SoC | system on a chip |
| SON | self-organizing network |
| SOTA/FOTA | software over the air/firmware over the air |
| SpCell | Special Cell |
| SPS | Semi Persistent Scheduling |
| SR | Scheduling Request |
| SR-IOV | single root input–output virtualization |
| SRB | Signaling Radio Bearers |
| SRI | Satellite Radio Interface |
| SRIT | Set of Component RITs |
| SRP | Stream Reservation Protocol |
| SRS | Sounding Reference Signal |
| SSB | Synchronization Signal Block |
| SSC | Session and Service Continuity |
| SSCMSP | SSC mode selection policy |
| SSS | Secondary Synchronization Signal |
| SST | Slice/Service Type |
| SU-MIMO | single-user MIMO |
| SUL | Supplementary Uplink |
| SyncE | synchronous Ethernet |
| TA | Timing Advance |
| TA | Tracking Areas |
| TAC | Tracking Area Code |
| TB | Transport block |
| TBS | Transport Block Size |
| TC | Transparent Clock |
| TCO | Total Cost of Ownership |
| TDD | Time Division Duplex |
| TDD/TDD | time division duplex/time division duplex |
| TDM | time division multiplexed |
| TE | Traffic Engineering |
| TEID | Tunnel Endpoint Identifier |
| TI-LFA | topology independent LFA |
| TI-LFA | Topology Independent Loop Free Alternates |
| TIP | Telecom Infrastructure Project |

| | |
|---|---|
| TM | Transparent Mode |
| TNL | Transport Network Layer |
| TPR | Technical Performance Requirement |
| TSDSI | Telecommunications Standards Development Society |
| TSN | Time-Sensitive Networking |
| TTA | Telecommunications Technology Association |
| TTC | Telecommunication Technology Committee |
| TTI | Transmission Time Interval |
| TVWS | TV White Spaces |
| TWAMP | Two-Way Active Measurement Protocol |
| UAS | Unmanned Aircraft Systems |
| UCI | Uplink Control Information |
| UDM | Unified Data Management |
| UDM | unified date management |
| UDP | User Datagram Protocol |
| UE | User Equipment |
| UHD | Ultra High Definition |
| UL/DL | uplink/downlink |
| ULCL | Uplink Classifier |
| UM | Unacknowledged Mode |
| UMTS | Universal Mobile Telecommunications Service |
| UMTS | Universal Mobile Telecommunications System |
| UP | User Plane |
| UPF | User-Plane Function |
| URLLC | Ultra-Reliable Low-Latency Communication |
| URSP | UE Route Selection Policy |
| UTRAN | Universal Terrestrial Radio Access Network |
| V2X | Vehicle-to-Everything |
| vDU | virtualized gNB-DU |
| VID | VLAN identifier |
| VIM | Virtualized Infrastructure Manager |
| VM | Virtual Machine |
| VNF | virtual network function |
| VNI | Virtual Network Index |
| VR | Virtual Reality |
| VR/AR | Virtual Reality and Augmented Reality |
| vRAN | virtual RAN |
| VXLAN | Virtual Extensible LAN |
| W-AGF | Wireline Access Gateway Function |
| WAN | wide area network |
| WBA | Wireless Broadband Alliance |
| WDM | wavelength division multiplexing |
| WG7 | Working Group 7 |
| WiMAX | Worldwide Interoperability for Microwave Access |
| WLAN | wireless local area network |

| | |
|---|---|
| WRC | World Radiocommunication Conference |
| xDSL | digital subscriber line technologies |
| Xn-AP | Xn Application Protocol |
| Xn-C | Xn Control Plane |
| Xn-U | Xn User Plane |
| ZTP | Zero Touch Provisioning |

# 1

# Introduction

As a general rule of thumb, every 10 years the cellular industry introduces a new technology: 3G Universal Mobile Telecommunications Service (UMTS) circa 2000, 4G Long-Term Evolution (LTE) circa 2010, and now finally 5G in 2020. Within that evolution, every technology cycle comes with advancement in terms of performance and new services, which the technology makes possible. These are typically attributed (and justifiably so) to the air interface, including the physical layer and the protocol stack. What is often overlooked is the Radio Access Network (RAN), which is fundamental to the success of every technology and which also undergoes major changes when a new technology is released.

The RAN is arguably the most important component in a mobile network. At least in terms of deployment and operational complexity and cost it certainly is. The air interface, including the physical layer and the protocol stack, typically draw most of the attention at least in the research community as these determine to a very large extent the performance of any wireless technology. However, when it comes to deployments, RAN is what eventually makes it possible and economically feasible (or not).

RAN is typically defined as a collection of base stations, interconnected with each other and connected to the core network, providing coverage in a certain area through one or more radio access technologies. This is illustrated in the simplified Figure 1.1.

In Figure 1.1 the RAN is depicted as a collection of base stations (shown as a single network node) connected via network interfaces (shown as straight lines). The reality of RAN standards, implementations, and, even more so, practical deployments is significantly more complex:

- Not all base stations are equal in terms of the capacity, coverage, and throughputs they provide. These can range from macro base stations serving many hundreds of users and covering a few square kilometers to small cells serving just a handful of users in an office.
- Base stations often also differ in terms of the radio access technology they provide over the air interface. Some base stations only provide 5G radio, some may provide 4G and 5G, and in some cases base stations providing different radio access may work in conjunction with each other. In other words, base stations also differ in terms of how tightly they are coupled with base stations providing other radio access.
- While it is possible to implement a base station with all the components, from antennas, to radio, to baseband, to protocol stack, and finally applications and management

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

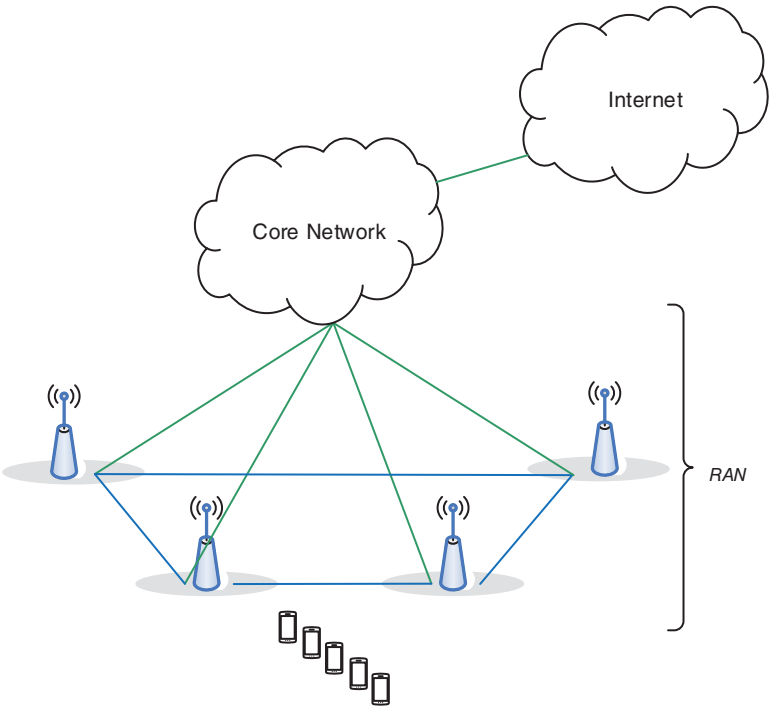



**Figure 1.1** Radio Access Network (RAN).

services in a single box (as shown in Figure 1.1), that is rarely the case. In practice, most base stations are split into multiple nodes in a variety of architectures, interconnected by sometimes standardized and sometimes proprietary network interfaces in a variety of architectures.

- Network interfaces themselves, illustrated as straight lines, in practice are anything but straight. What is often overlooked is that these interfaces run on a transport network, which often consists of various technologies – multiple transport network nodes interconnected in various network topologies.

This book is dedicated to the topic of RAN architectures and technologies. It is structured as follows:

- In Chapter 2 ("Market Drivers") we describe the technological, regulatory, and business driving forces behind 5G in general and how these diverse requirements, challenges, and marketing considerations affect the RAN.
- Before we dive into the details of RAN architectures, in Chapter 3 ("5G System Overview") we provide a high-level overview of all the components of a 5G system: the core network, the air interface protocol stack, and the air interface physical layer. These help put the RAN architectures discussed afterward into a proper context.
- Chapter 4 ("NG-RAN Architectures") is perhaps the main part of the book, where we describe in detail all the 5G RAN architectures defined in the 3rd Generation Partnership Project (3GPP), O-RAN Alliance, and Small Cell Forum, specifically: the high-level

gNB CU/DU (central unit–distributed unit) split, the multi-connectivity architectures, the gNB architecture with control/user separation, the low-level gNB intra-PHY split, and the small cell architectures.

- Chapter 5 ("NG-RAN Evolution") is dedicated to NG-RAN evolution beyond Release-15, describing technologies introduced in Release-16: e.g. relaying, also known as integrated access and backhaul (IAB, and satellite access, also known as non-terrestrial networks.
- Chapter 6 ("Enabling technologies") is dedicated to various technologies that are not always considered part of RAN architecture but are nevertheless fundamental to RAN deployments. These include implementation-related aspects, such as virtualization and open source, edge computing, Operations, Administration, and Maintenance (OAM), and last but not least the transport network technologies.
- We finish the book with Chapter 7 ("NG-RAN Deployment Considerations") by discussing the practical implications of selecting the right RAN architecture and deploying it to serve the practical needs of an operator.

A note on terminology: throughout this book, we generally try to use a consistent terminology. However, that is not always possible, or convenient – in particular, because similar technologies may sometimes be commonly referred to by different names in different standards, industries, or literature. As this book crosses multiple domains, it is challenging to use a uniform terminology, which is at the same time consistent with different terminologies used in their respective fields. One such example is the term "5G" itself – while it is used extensively in technical literature, marketing materials, product descriptions, etc. – many (but not all) 3GPP specifications intentionally avoid the term, using terminology such as New Radio (NR) when referring to the air interface and NG-RAN (which is not an acronym at all, but is considered a "monolithic term") when referring to the RAN. Another example is the network interface between the NG-RAN and the core network, which is referred to as the NG interface in RAN specifications and N2/N3 reference points in core network standards.

We therefore took the pragmatic approach of using common terminology where we felt it is appropriate, and otherwise using the terminology from the domain being described in the book, with appropriate definitions and explanations in each chapter.

# 2

# Market Drivers

*Reza Arefi*[1] *and Sasha Sirotkin*[2]

[1] *Intel Corporation, USA*
[2] *Intel Corporation, Israel*

## 2.1 Introduction

In this chapter we discuss various technological, regulatory, and market drivers that triggered the development of 5G and the problems 5G is expected to solve. We then attempt to derive how these affect the Radio Access Network (RAN) architecture and its evolution in order to support 5G, which is the primary focus of the book.

This is not an easy task, as there is no universally agreed definition of what constitutes 5G. To some, this is the technology that meets the International Telecommunications Union (ITU) IMT-2020[1] requirements and therefore will be able to make use of the newly identified spectrum for IMT. To others, this is an expansion of cellular technologies beyond their traditional mobile broadband (MBB) use cases and markets into Internet of Things (IoT), private networks (i.e. networks deployed by entities other than traditional cellular operators), and other markets where cellular technologies have not been commonly used before. Some others view 5G as simply an evolution of 4G (Long-Term Evolution [LTE]) to support higher throughputs, lower latencies, and better energy efficiency targeting primarily MBB; that is, the same use cases as 4G. Some point out that the primary technological advancement of 5G is the support of mmWave spectrum, while others believe that 5G is the turning point when cellular networks finally fully embrace virtualization (including RAN), driving down operational costs by opening up RAN to bigger competition.

Given such diverse views in the industry it is hard to pinpoint a single major market driver for 5G. Moreover, it is quite clear at the time of writing this book that, while at least some of the driving forces mentioned above (e.g. mmWave) do provide substantial technological improvements, these do not necessarily address an existing market need, but are rather

1 International Mobile Telecommunications-2020 (IMT-2020) is the codename used by International Telecommunications Union's Radiocommunication Sector (ITU-R) to describe the next generation of IMT technologies to be submitted to ITU-R and approved in a multi-year process of evaluations and consensus building scheduled to complete in the year 2020. The process, which started in 2015, aims at producing a new ITU-R Recommendation containing detailed specifications of IMT-2020 radio interfaces.

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

being developed in the hope that market need will "catch up" and eventually materialize to take advantage of these new technical advancements.

In our view, unlike previous generations of cellular technologies, it is better to view 5G not as a single technology, but rather as a flexible system designed to serve many use cases and many markets. Such extreme flexibility comes at a cost of increased network and device complexity and, perhaps even more importantly, greater uncertainty of which features of 5G will be deployed and when. It is quite possible that different market forces in different geographies will drive the deployment of different features. It appears that in Asia the major driving force is the increased throughput for the MBB, while European operators are exploring various options for breaking into new markets (e.g. IoT), whereas in North America one of the key driving forces (at least for the moment) is fixed wireless access to provide better internet service to suburban areas. In summary, 5G may not be a one-size-fits-all technology as it is often presented, but rather a toolbox of different technologies that different operators (and potentially new entities) will use for different purposes.

This is not new, as oftentimes this is historically how computing and networking technologies have been developed. A breakthrough in computing power and/or network throughput comes first; applications that make use of these new capabilities are developed later. The caveat is that it is unclear when exactly these new business cases and applications taking advantages of the progress in speed and power will emerge; it can take a while.

One good example of a similar case is 3G, which was initially deployed in the early 2000s,[2] but it was not until the late 2000s that 3G MBB market penetration became significant, in part thanks to the launch of the iPhone.

This is not to say that there is no need for better, faster, and more energy-efficient wireless networks supporting billions of devices. According to the Cisco Virtual Network Index (VNI) forecast, as shown in Figure 2.1, there will be 396 Exabytes (EB) per month overall IP traffic by 2022. Ericsson estimates in their Mobility Report that 80 EB of these will be consumed by mobile devices, as shown in Figure 2.2.

There are similar forecasts indicating growth of connected devices in general and IoT in particular, as well as other indicators pointing to the fact that it is reasonable to expect that

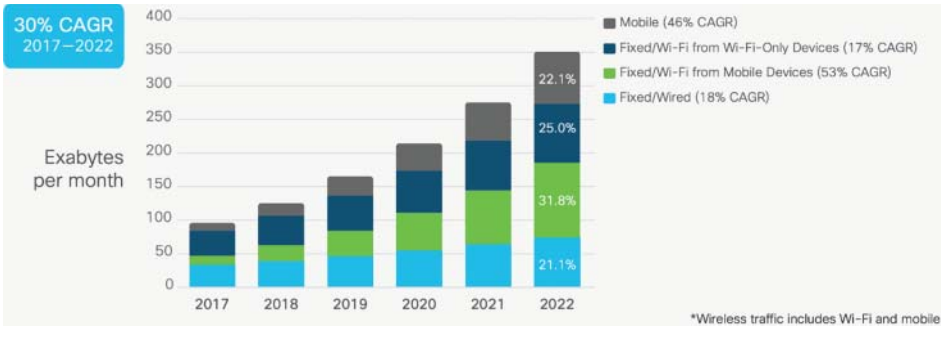



**Figure 2.1** Cisco VNI IP traffic forecast (Source: SISCO VNI Global IP traffic forecast 2017–22).

2 NTT DoCoMo's Freedom of Mobile Multimedia Access (FOMA) network is usually regarded as the first 3G deployment, even though initially it did not follow the Universal Mobile Telecommunications System (UMTS) standard.

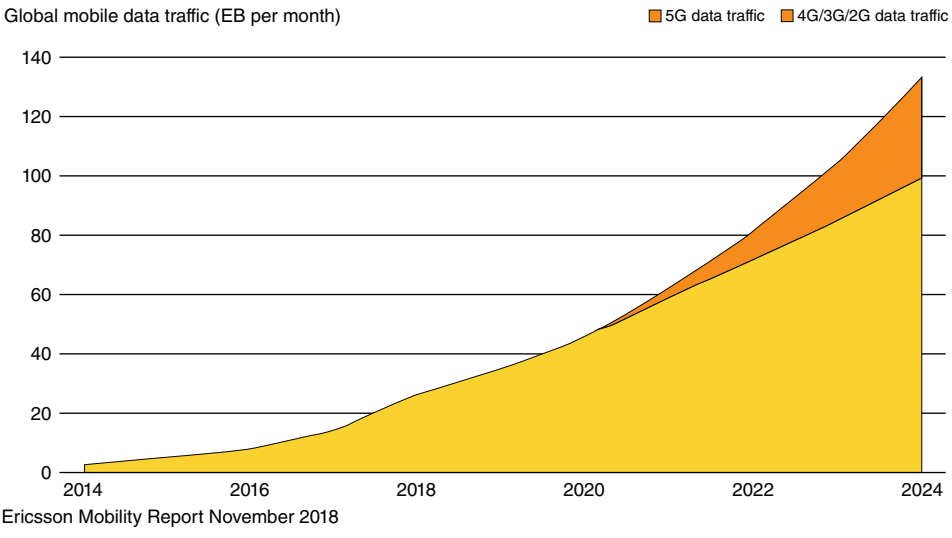



**Figure 2.2** Ericsson Mobility Report, global mobile data traffic (EB per month).

network traffic in general and mobile traffic in particular are likely to continue growing exponentially. Therefore, even though it may not be clear yet what applications will be served by 5G networks, the demand for 5G is there and mobile networks, RAN in particular, need to evolve to cope with such traffic in a cost- and energy-efficient manner.

Increased throughputs and new spectrum (e.g. mmWave) are not the only, and maybe not even the primary, 5G driving factors. Additional drivers are cost and energy efficiency considerations, competition (between operators, vendors, and even market sectors and technologies), and even politics, in what is sometimes referred to as the "race to 5G."

In this chapter we elaborate on the various forces driving 5G technology development and deployment with emphasis on how these impact RAN features, RAN-related technologies, and RAN architecture, which is the primary focus of the book.

## 2.2 Key Ideas

- Data traffic in general and mobile traffic in particular is expected to continue growing exponentially.
- In the past, spectrum needs forecasts significantly underestimated actual data usage. To alleviate this issue, the ITU Radiocommunication Sector (ITU-R) used a new approach that forecasts spectrum needs ranging from hundreds of MHz to tens of GHz. The 5G target spectrum consists of lower frequency ranges (below 1 GHz), middle frequency ranges (below 6 GHz), and higher frequency ranges (mmWave) to cater to different applications. As the 5G spectrum is expected to be an order of magnitude larger than 4G, this will have a direct impact on RAN.
- Spectrum-sharing models, such as Citizens Broadband Radio Service (CBRS) in the USA and Licensed Shared Access (LSA) in Europe, may further increase available spectrum.

Furthermore, they may trigger new RAN deployment options, such as the neutral host operator model. Even though CBRS and LSA are currently based on LTE, we expect that in the future spectrum-sharing models will become applicable to 5G as well.

- In order for a technology to qualify for IMT-2020, it must fulfill certain technical requirements broadly categorized as: enhanced mobile broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine-Type Communication (MTC). Of these URLLC in particular will have the biggest impact on RAN architecture and design, because most real-world applications are concerned with end-to-end latency, not just over the air, which is addressed by the New Radio (NR) design. URLLC scenarios and other latency-sensitive applications such as cloud gaming, require 5G networks to support significantly lower end-to-end latency, compared with 4G.
- 5G creates new business opportunities. It allows cellular operators to expand into new markets (which have been served by non-cellular technologies in the past or did not exist before), for example, by deploying IoT and Vehicle-to-Everything (V2X). Furthermore, it creates new business models with, for example, slicing, allowing mobile network operators (MNOs) to lease network capacity to other companies. On the other hand, 5G also helps new entities that have not used cellular technologies in the past to adopt 5G and in some cases compete with traditional cellular operators, with technologies such as private networks and the adoption of the 5G radio interface for satellite communications. Increased competition is likely to make standardized network interfaces more important and may eventually allow network multi-vendor interoperability in RAN (which is not quite the case in 4G).
- Standards will continue being important in 5G and it appears that the 3rd Generation Partnership Project (3GPP) will continue to have a central role in developing cellular standards. This has the positive effect of ensuring that there is only one major 5G standard, reducing market fragmentation. On the other hand, the increased interest in 3GPP triggers increased participation from many more companies and delegates, making a consensus harder to reach. The end result is that, unlike 4G, 3GPP 5G standard will have many options (sometimes presented as "flexibility"). This flexibility has a cost, as it is increasingly hard to predict which standard options will be deployed in the field. Furthermore, there are still many Standards Developing Organizations (SDOs) and industry fora working on technologies that may be considered competition (e.g. LoRa and the Institute of Electrical and Electronics Engineers [IEEE]), or may complement 3GPP standards (e.g. Broadband Forum [BBF], Open Radio Access Network [O-RAN], Small Cell Forum, etc.).
- Open source, which was extremely successful in the enterprise and data centers, is increasingly finding its way into telecom networks. There are number of open source LTE Evolved Packet Core (EPC) implementations available (e.g. Magma), open source Operations, Administration, and Maintenance (OAM) frameworks (e.g. Open Networking Automation Platform [ONAP] and Open Source Mano [OSM]), and finally even RAN implementations (e.g. OpenAirInterface). Open source may be considered an alternative to standardization, and while it is hard to see how it can replace standards for the radio interface, at least in the CN and even in RAN it may become a viable alternative.
- RAN sharing is likely to become more important in driving down costs; it may eventually evolve into a neutral host RAN sharing model.

- 5G will not only trigger a new round of competition among the usual cellular players, for example, operators, network and handset vendors, but also a competition between technologies and whole market segments. 5G is often touted as the next wireless revolution and the promises made are indeed somewhat grandiose. There is no doubt that the technology is capable of delivering these promises, provided there is a viable business model to support them.
- In this book, we illustrate how 5G market drivers affect RAN architecture and deployment considerations from the perspectives of increased throughputs, reduced latency, network densification, and competition within the traditional wireless ecosystem and between incumbents and new players.

## 2.3 Spectrum

### 2.3.1 Spectrum Needs

As with many previous generations of cellular technologies, availability of spectrum can be considered as one of the driving factors behind 5G. For illustrative purposes, Figure 2.3 shows spectrum allocations to various services in the US.

Over the past few decades, an exponential increase in data consumption has dominated the overall demand for 3G/4G services. This global data consumption of networks seems to undergo contiguous explosive growth. Figure 2.4 from an ITU-R Report in 2011 (ITU-R M.2243), compares the range of traffic growth estimates in 2005, the so-called baseball cap

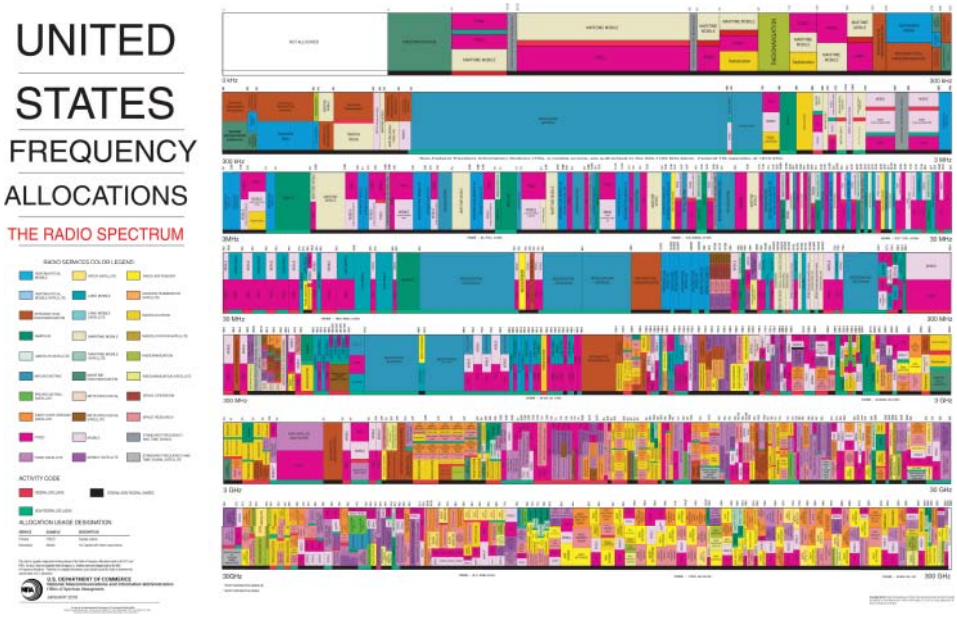



**Figure 2.3** United States Frequency Allocations Chart 2016.

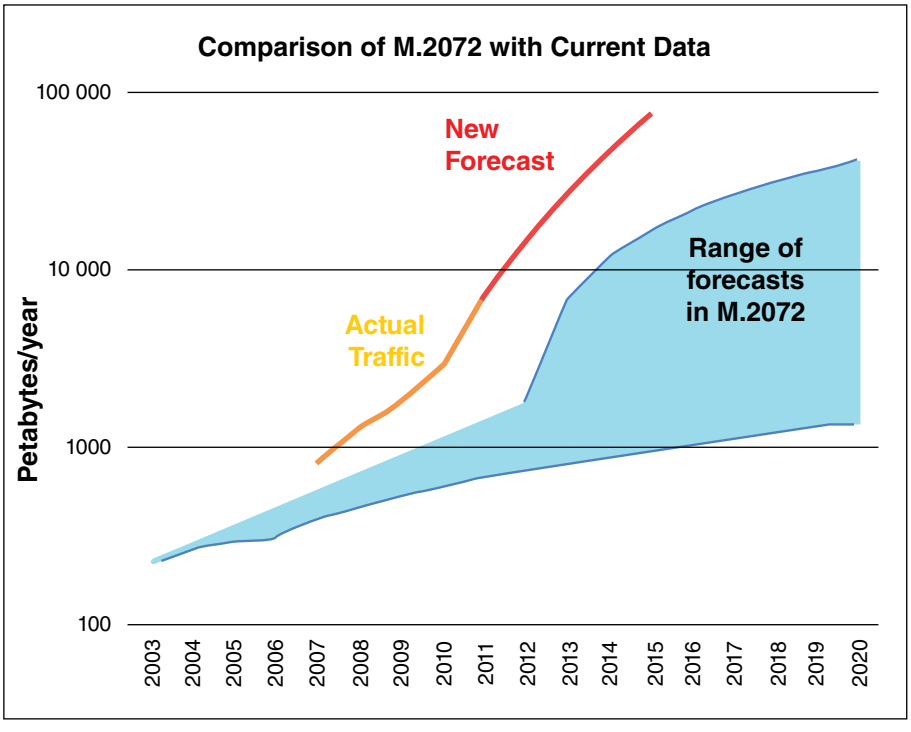



**Figure 2.4** Comparison of traffic estimates in 2005 with actual data. (Source: ITU-R).

figure, with actual traffic values observed. As can be seen from Figure 2.4, the actual traffic growth surpassed even the higher, more aggressive forecasts of 2005.

A similar attempt (ITU-R M.2290) to capture the traffic growth in the 4G era using similar methods is illustrated in Figure 2.5.

In this case, again a range of forecasts was used to estimate the amount of spectrum needed to support the growth in traffic. As a result, a range of total spectrum requirements between 1340 and 1960 MHz (including existing 3G/4G spectrum) was calculated to support mobile services up to the year 2020 (ITU-R M.2290). Considering how the mobile industry landscape has changed since 2013, it is evident that there are major shortcomings in the methodologies used to arrive at spectrum estimates.

This discrepancy between spectrum forecasts and actual data is at least partially due to the fact that increased data consumption of *individuals* (browsing, downloading, streaming, etc.) has been accompanied in the 5G era by addition of new and emerging *applications* requiring various types and amount of connectivity/data/resources dictating radio interface capabilities. As a result, new application-centric methodologies were needed to model this growth for the 5G era.

ITU-R, in a Recommendation (ITU-R M.2083) describing its vision for 5G framework and objectives, specifies several Key Performance Indicators (KPIs) as part of outlining future networks' Technical Performance Requirements (TPRs). These include, for instance, peak

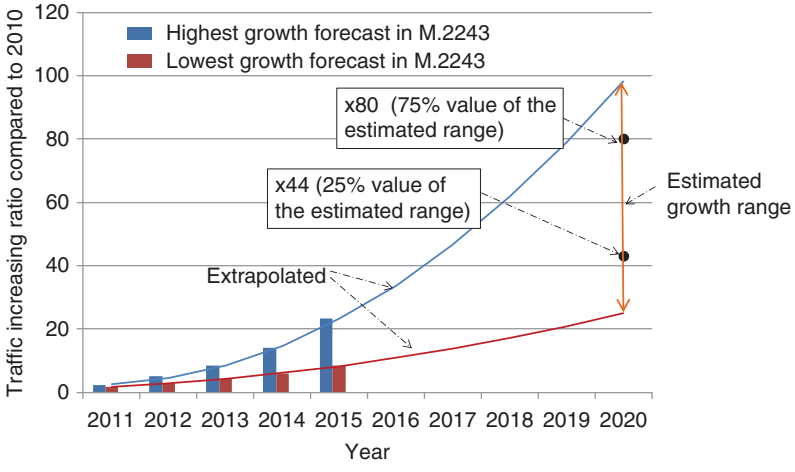



**Figure 2.5** IMT-Advanced spectrum estimation, 2013. (Source ITU-R).

and average data rates, latency, and spectral efficiency. These values, generally presenting aggressive leaps compared with previous generations, would then need to be supported by the radio interfaces of 5G systems.

A new approach based on requirements of various new applications and defined TPRs was chosen. This new approach is based on the simple principle that all other aspects held constant, a system targeting an application requiring a 100 Mbps user data rate would require 10 times more spectrum than a system targeting another application requiring only a 10 Mbps user rate. A simple equation was used (Eq. (2.1)), assuming full-buffer traffic, to calculate the bandwidth $B$ (in Hz), expressed as a product of the required user/device data rate $D$ (in bits/s) and the number of simultaneously served users/devices ($N$) in the cell, divided by the spectral efficiency $S$ (in bits/s/Hz).

$$B = (D \times N)/S \tag{2.1}$$

It was anticipated that different TPR values would result in different spectrum requirements. Table 2.1 (ITU-R WP5D) demonstrates the outcome of one such calculation for three different types of TPRs. Example 1 is based on cell-edge user data rate targets in M.2083 using Eq. (2.1). Example 2 additionally assumes sample deployments in two different environments. Example 3 considers the combined impact of latency and end-to-end data delivery.

The extent of spectrum requirements, many GHz in some cases, was at least partly the reason to consider higher spectrum ranges such as mmWave.

Other elements impacting the total amount of required spectrum for 5G also exist. One analysis (5G Americas) points to factors such as multi-operator deployments in the same area, the potential need for guardbands (e.g. in adjacent or same-area unsynchronized Time Division Duplex (TDD) networks), frequency re-use of greater than one in areas where additional carriers are needed for improving performance, advancements in spectral efficiency, and multi-antenna techniques.

**Table 2.1** IMT-2020 spectrum needs based on TPRs.

| Examples | Spectrum needs |
|---|---|
| 1 – Based on cell-edge user throughput and spectral efficiency targets in Recommendation ITU-R M.2083 with N simultaneously served users/devices at the cell edge | User-experienced data rate of 1 Gbps:<br>3.33 GHz (N = 1), 6.67 GHz (N = 2), 13.33 GHz (N = 4), e.g. indoor<br>User-experienced data rate of 100 Mbps:<br>0.67 GHz (N = 1), 1.32 GHz (N = 2), 2.64 GHz (N = 4), for wide area coverage |
| 2 – Based on cell-edge user spectral efficiency (obtained from 3GPP technical specifications) and data rate targets (from Recommendation ITU-R M.2083) in two given test environments | 0.83–4.17 GHz (for eMBB dense urban)<br>3–15 GHz (for eMBB indoor hotspot) |
| 3 – Impact of latency and spectral efficiency targets and a typical user throughout value on spectrum needs | With a file transfer of 10 Mb by a single user at cell edge in 1 ms: 33.33 GHz (one direction)<br>With a file transfer of 1 Mb by a single user at cell edge in 1 ms: 3.33 GHz (one direction)<br>With a file transfer of 0.1 Mb by a single user at cell edge in 1 ms: 333 MHz (one direction) |

ITU-R, ITU Radiocommunication Sector.
eMBB, enhanced mobile broadband.
Source: ITU-R.

### 2.3.2 Target Spectrum

From the early stages (NGMN5G) of envisioning 5G applications and requirements, various categories of use cases and their associated user experience targets were envisaged, each with specific requirements, which could impose potential conditions on radio interface design. User experience associated with various use case categories could have spectrum implications in order to optimize overall performance.

A methodical approach to categorization of future target applications was done in ITU-R, leading to the now-famous 5G usage scenario triangle as illustrated in Figure 2.9 (Section 2.5 ITU-R, M.2083).

It is generally expected that 5G will require substantially higher spectrum ranges compared with 4G in addition to lower and middle ranges. Certain applications require highly robust performance over long distances, which is a characteristic of lower frequencies. Other applications need very high throughput over shorter distances, which is a characteristic of higher frequencies.

These aspects could be optimally achieved through access to sufficient spectrum in a variety of bands to deliver full 5G service.

1. Lower frequency ranges, e.g. below 1 GHz, for wider reachability; examples include macro cells, robust obstacle penetration, sensor networks, and automotive.

2. Middle frequency ranges, e.g. below 6 GHz, for coverage/capacity trade-off; examples include small cells and capacity boost.
3. Higher frequency ranges, e.g. mmWave, for higher throughput; examples include hot spots, Ultra High Definition (UHD) video streaming, Virtual Reality (VR), and Augmented Reality (AR).

The wireless industry, therefore, has been encouraging regulators around the globe to designate sufficient amounts of spectrum in low, mid, and high ranges for 5G.

The World Radiocommunication Conference (WRC-19) was particularly important, with mmWave frequencies between 24.25 and 86 GHz being on its agenda. After several weeks of deliberations and intense negotiations, Administration members of ITU-R agreed on identification of more than 17 GHz of new spectrum for IMT. These bands are listed below[3]:

- 24.25–27.5 GHz – global identification
- 37–43.5 GHz – global identification
- 45.5–47 GHz – regional/country-specific identification
- 47.2–48.2 GHz – regional/country-specific identification
- 66–71 GHz – global identification.

In addition, WRC-19 agreed to further study identification for IMT of several bands below 10.5 GHz toward a decision at WRC-23.

IMT identification of frequency bands by a WRC has been historically followed, albeit in varying degrees, by national regulators' spectrum designations and availability for previous generations of cellular systems. To put the output of WRC-19 into perspective, it is worth comparing current 4G spectrum holdings with future 5G spectrum allocations. Currently, most operators hold spectrum of circa a few tens of MHz (typically ranging from 10 to 50 MHz), which is usually used for 4G and 3G.

It is generally understood that the majority of the 5G spectrum will be divided into two categories of mid-range (e.g. 3.5 GHz) and high range (e.g. 28 GHz). While prediction of the precise 5G spectrum availability in every country is difficult, it is reasonable to assume that a 5G network will need to make use of about 100 MHz of spectrum in the mid-ranges and about 1 GHz of spectrum in the millimeter wave or sub-millimeter wave bands. This is likely to be in addition to existing spectrum currently used by an operator in the sub-1 GHz spectrum range. While the additional mid-range spectrum is substantial, the addition in the high range is likely to be an order of magnitude larger than what operators have today and what their networks, RAN in particular, are designed to support.

### 2.3.3 Spectrum Implications

Such a radical increase in available spectrum will have a direct impact on RAN. 5G networks will have to support substantially higher throughputs compared with 4G, which inevitably will affect at least the transport fronthaul and backhaul network and also the RAN architecture. In particular, this affects base station architecture, making the usage of Common Public Radio Interface (CPRI) unfeasible and triggering different split gNB architecture designs, as explained in Chapter 4. Furthermore, high-band and even mid-band spectrum

3 For details, see WRC-19 Provisional Final Acts (https://www.itu.int/pub/R-ACT-WRC.13-2019).

require much higher network densification compared with what is used today (i.e. massive deployment of small cells), which is also likely to affect RAN architecture and deployment considerations.

There are also several challenges facing the implantation and development of 5G systems with direct or indirect impact on radio interface design. These challenges generally fall under three categories.

1. How to protect various incumbent systems in potential future 5G bands.
   Various incumbents have varying technical and/or regulatory requirements for their protection. Addressing these requirements needs technical as well as regulatory solutions.
2. How to overcome propagation impairments, especially in higher frequency ranges.
   In mmWave, atmospheric effects such as rain and gaseous losses limit propagation and cell range. In addition, obstacle penetration is another limiting factor due to the fact that propagation by reflection and scattering are dominant in mmWave as opposed to lower ranges where diffraction is the dominant propagation phenomenon.
3. How to develop required antenna and RFIC technology in a cost-effective manner.
   Use of high-gain arrays and beam-forming is encouraged by the relatively small sized antenna elements in order to compensate for excessive path loss of mmWave. However, designing cost-effective commercial components in such high frequencies is a new challenging area for the mobile industry. For instance, developing filter technology with sharp roll-off to curb unwanted emissions is more challenging in higher ranges than in traditional 3G/4G bands.

In addition, mmWave have other system-related impacts as well. Certain deployment scenarios such as dense urban could present excessive multipath through reflection and scattering. In many cases, however, especially with highly directional antennas, a single reflected path is observed, followed by many smaller components. Antenna characteristics (beamwidth, side-lobes) play a critical role in characterizing delay spread of these deployment scenarios. Similarly, Doppler spread due to multipath causes channel fluctuations, which in turn could be minimized by using highly directional antennas. Highly directional antennas, however, pose challenges in user tracking.

## 2.4 New Spectrum Models

Over the past few decades, some regulators, for example, in North America and Europe, have moved toward technology-neutral regulations, that is, providing maximum *flexibility* to licensees in deploying their technology of choice within the allocation of the service as long as a set of *least restrictive* technical conditions is met. The two important elements to technology neutrality, that is, flexibility and least restrictive conditions, are related to each other in a dialectic way. The flexibility element enables licensees to make decisions on their technology[4] of choice over the term of the license within the specified rules. The least

4 The duplexing mechanism is often considered a technology element. There are radio interfaces with more than one mode of duplexing where almost every other aspect of the interface remains the same. A full

restrictive element maintains that the flexibility given is within bounds, that is, only the minimum necessary for preventing harmful interference to other services in the band or in the adjacent bands. Technology neutrality has generally led to market-driven deployments and transition of 2G to 3G to 4G technologies in many countries without much need to repurpose, refarm, and reregulate cellular bands.

As mentioned in the previous section, significant spectrum is likely to be allocated for 5G in many parts of the world. However, in some cases even this may not resolve the spectrum shortage completely. One reason for this could be that the largest chunk of 5G spectrum is likely to be in the mmWave range, which may not be suitable for certain applications due to propagation characteristics. Even though availability of enough spectrum to allow implementation of ultra-wide channels in the mmWave spectrum may provide for extreme throughputs (by today's standards), it requires much denser network deployment, therefore substantially increasing operator's Capital Expenditure (CAPEX) and Operational Expenditure (OPEX). Moreover, extremely dense network deployment may lead to mobility issues and therefore may not be suited to address all the 5G use cases. To mitigate this issue, mobile industry and regulators are exploring new regulatory models to better use spectrum resources in the lower frequencies.

Spectrum has traditionally been made available for commercial use in two ways: exclusive license and license-exempt. The former is typically used by various radio services in lower and mid-range frequencies where exclusivity of a license is the main regulatory measure for protection of a licensee against interference from other services in the band (in adjacent areas) or in adjacent bands (in the same or adjacent areas). Each licensee would then have to comply with a certain emission level outside its spectrum block, outside its license area boundary, or both. In awarding terrestrial mobile licenses, a power flux density (PFD) value, or alternatively a field strength value, has been used in the past few decades to curb interference to other licensees at the boundary of a given license area, or at international borders.

While this method has many advantages and has led to proliferation of cellular technology all over the world, it could, in some cases, result in spectrum underutilization. The latter method is best suited for Wi-Fi access. However, user experience in license-exempt bands generally degrades with increased presence of other users. Therefore, license-exempt services cannot provide guarantees for any level of Quality of Service (QoS) to users and are limited to Best Effort (BE) methods.

### 2.4.1 New Ways of Sharing Spectrum

An approach, which may help resolve the spectrum crunch in the lower to mid-range frequencies, is through new methods in sharing of spectrum resources among multiple services in a way that certain QoS levels could be guaranteed. This scheme could potentially allow for availability of new spectrum licensed to incumbents (e.g. government agencies, satellite systems, Electronic New Gathering [ENG], amateur radio, and many others) for cellular communications without impacting the operation of the incumbent systems. It also

---

technology-neutral approach to paired bands, therefore, takes the form of allowing any duplexing scheme in the paired spectrum, e.g. uplink/downlink (UL/DL), downlink/uplink, (DL/UL), or time division duplex/time division duplex (TDD/TDD).

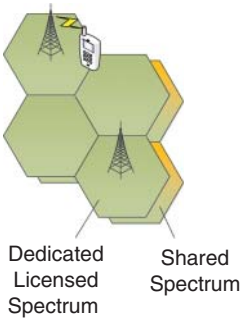



**Figure 2.6** Spectrum sharing.

improves overall spectrum utilization while providing some assurance to the licensees that they will get Return on Investment (ROI) on a network build-out using that spectrum. This is illustrated by Figure 2.6.

One of the first spectrum-sharing cases was the TV White Spaces (TVWS). TVWS allows unlicensed, secondary devices to access spectrum licensed to broadcasting at specific locations and time intervals, where they would not interfere with the operation of the incumbent systems. Unfortunately, TVWS technology adoption levels remain low.

Two new spectrum-sharing frameworks have been developed: LSA developed by the European Telecommunications Standards Institute (ETSI) and the European Conference of Postal and Telecommunications Administrations (CEPT) in Europe, and CBRS developed by CBRS Alliance in the US. These initiatives show great promise and at the time of writing of this book a number of companies had applied for the CBRS license and started testing CBRS networks. The Federal Communications Commission is expected to start auctioning the Priority Access License (PAL) part of the CBRS band by 2020.

Even though CBRS and LSA are conceptually similar, there are certain differences in how these technologies will be implemented, as summarized in Table 2.2.

In CBRS, the three tiers are defined as follows:

- First tier is an incumbent user, e.g. the federal government.

**Table 2.2** LSA vs. CBRS.

| | LSA | CBRS |
|---|---|---|
| Tier access | First tier: incumbent user<br>Second tier: (co-primary) licensee | First tier: incumbent user<br>Second tier: primary access license<br>Third tier: general authorized access |
| Operating band | 2.3–2.4 GHz (LTE band 40) | 3.55–3.7 GHz (LTE bands 42/43) |
| Incumbent protection | Using database | Using sensing and database |
| Licensing period | To be negotiated (target: >10 years) | 10 years (PAL) |

LTE, Long-Term Evolution.
PAL, Priority Access License.

- Second is PAL users – licensed users who acquire spectrum, e.g. through an auction. PAL users must protect incumbent Tier 1 users from harmful interference.
- Third tier is General Authorized Access (GAA) users, who may operate through registration. GAA users must protect both first tier incumbents and second PAL users from harmful interference.

LSA and CBRS may have RAN impact not only due to increased spectrum availability, but also because of special incumbent protection requirements, such as sensing and usage of a database for spectrum availability. Furthermore, while CBRS network architecture is essentially based on 3GPP, there are certain differences, such as for example the usage of Spectrum Access System (SAS), which controls access to CBRS spectrum.

A CBRS base station connects to a SAS (which is a network node unique to CBRS and not present in the 3GPP architecture) when it is powered on. The base station provides its coordinates and an identifier to the SAS. The SAS uses this information to provide to the base station the CBRS frequencies it can use, that is, those which are currently not in use by the first tier incumbent users.

Since as of now CBRS is not slated to use NR,[5] further details on SAS are beyond the scope of this book. Furthermore, shared spectrum as described here may give rise to new RAN deployment options such as neutral host. The neutral host concept extends the idea of RAN sharing and "tower business model." A Neutral Host Network (NHN) operator, which can be for example a venue owner, may build a network operating in a CBRS spectrum with relatively low investment and then lease the network capacity to other operators, such as conventional MNOs. This may be mutually beneficial to both, as it provides a new source of revenue for a venue owner and relieves MNOs from the burden of cell site acquisition, which can be substantial and is one of the primary reasons for the relatively insignificant small cell deployments so far.

### 2.4.2 Localized Licensing

Another important development has started in Europe in the form of making spectrum available for specific industry segments, such as Industrial Internet of Things (IIoT), Intelligent Transport Systems (ITS), and potentially others who would benefit from access to 5G but have strict performance requirements such as very low latency. For example, the German regulator, Bundesnetzagentur (BNetzA), has made spectrum in the 3.7–3.8 GHz band available for private networks[6] and this will be separate from the auction of spectrum for general 5G mobile broadband use. Other regulators, for instance some in Asia, have also started considering similar approaches to help facilitate introduction of 5G into many vertical industries. These efforts, however, can come to fruition using existing technology-neutral models.

Technology neutrality, as described earlier, does not mandate a specific network structure model. In other words, it does not require the licensee to provide wide-area service

5 At the time of writing of this book, CBRS and LSA systems are based on LTE, rather than 5G. However, we expect that over time the spectrum-sharing approach will be extended to 5G as well.
6 3GPP is addressing private network requirements in Release-16.

only. An example is use of Distributed Antenna Systems (DASs) to provide local coverage inside buildings using frequencies licensed for wide-area use. In the US, technology-neutral licenses have a wide range of variation in terms of license area.[7] All are exclusive type licenses, with no mandate on use of a specific wireless technology or a specific use case of a given technology. As a matter of fact, the same way certain KPIs of various 5G applications, such as peak and average throughput, dictate and drive the need for access to a variety of spectrum ranges from very low to very high, network architecture and certain other KPIs, such as end-to-end latency, require *very small license area sizes* in addition to the large license area size needed for wide-area eMBB applications.

There are no requirements for either license area size or network topology and structure for a technology-neutral license. The well-established technology-neutral framework could equally apply to all license area sizes and network topologies. Local-area, privately operated networks can also be regulated under a technology-neutral regulatory regime.

## 2.5 Regulations Facilitating 5G Applications

For future distributed communications and computing architectures to help vertical market segments maximize their benefit from 5G technologies a new look at regulations may be in order. It is becoming increasingly important that future regulatory regimes allow not only *exclusively licensed* (full wide-area QoS) and traditional *license-exempt* (BE) operation but also consider *locally licensed* and *semi-scheduled license-exempt*[8] operations. These two latter subcategories of licensing, which require further attention from regulators, could work in similar manners benefiting vertical markets but are different from the point of view of QoS and reliability, and hence cost and addressable market.

Regulators should consider creating favorable regulatory conditions and make spectrum available in a technology-neutral way that could facilitate industrial and enterprise applications (vertical industries) so they could benefit from upcoming availability of 5G radios and networks. It should be stressed that simple measures could be used to address the situation. As an example, the same metric/condition of compliance with a wide-area license at an international or a license area boundary, that is, compliance with a maximum electric field strength at a certain height above ground could also be used as the compliance condition for a local license at the boundary of the facility acquiring the license. There is also no reason to believe unwanted emission metrics should vary depending on the size of the license area.

Timely availability of spectrum under suitable licensing conditions could create a major growth area for the economy while enabling non-eMBB aspects of 5G toward implementation of Time-Sensitive Networking (TSN) and end-to-end integration of services, which

7 The license area size for terrestrial mobile systems vary from as large as nationwide to as small as counties.

8 Listen-Before-Talk (LBT) has been the foundation that enabled many consumer applications including Wi-Fi and Bluetooth, and will continue to benefit future applications including many in the 5G era. It is, however, anticipated that confined spaces of many industrial facilities, that is indoors, would reduce the inefficiencies of the "polite protocols" by moving toward more deterministic behavior for a specific set of use cases. Therefore, some level of time-synchronized scheduling implemented in the license-exempt protocols, for example IEEE 802.11ax, could increase reliability to levels required by some applications.

is needed if the full potential of 5G is to be reached. A few countries in Europe and Asia have already started, or plan to start soon, the process of allocation of spectrum for use by vertical industries in a locally licensed manner. It is also important for regulators to assign locally licensed spectrum within the ranges already defined in NR specifications in 3GPP to take advantage of economies of scale.

## 2.6 Network Deployment Models

Traditionally, each mobile operator deployed and operated its own network. However, in order to decrease CAPEX and OPEX of RAN, which are substantial, operators turned to RAN sharing models, which is also often encouraged by regulators. With RAN sharing, a portion of resources of a network deployed by operator A can be leased to operator B, as shown in Figure 2.7.

Different standardized and proprietary RAN sharing options exist, ranging from just sharing a cell site, to sharing a base station, to sharing a base station and the spectrum. Many operators deploy 4G using RAN sharing, and we expect this trend to continue with 5G.

The concept of a neutral host takes the RAN sharing idea one step further, allowing for cell site and RAN infrastructure sharing between operators, with the main difference between them being that the RAN is owned and operated by some other entity than a mobile operator.

Neutral host RAN infrastructure is a single, shared network solution provided to all MNOs by a third party (e.g. a venue owner) and can be used, for example, to resolve poor wireless coverage and capacity inside buildings or other busy locations, such as stadiums.

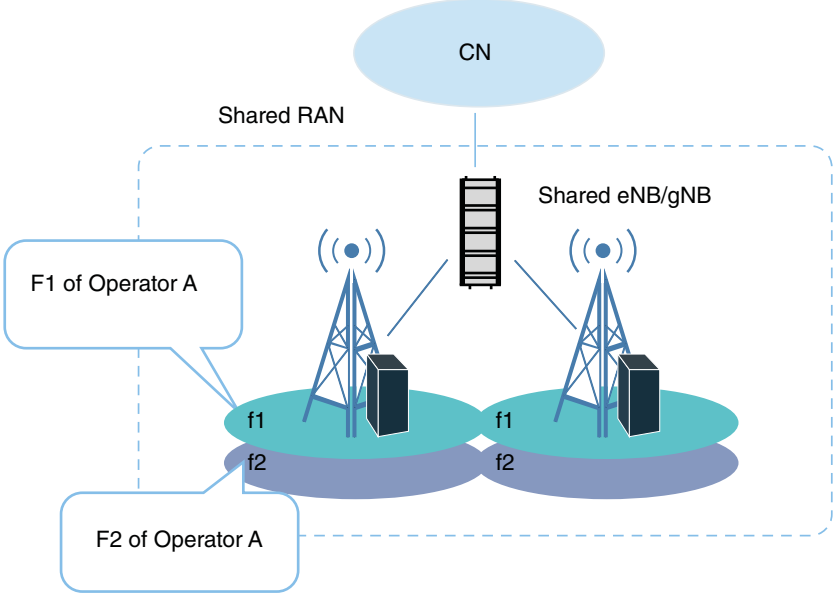



**Figure 2.7** RAN sharing.

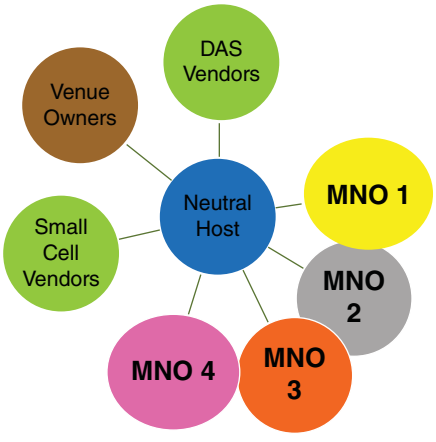



**Figure 2.8** Neutral host network deployment.

Various different neutral host approaches are used to provide wireless services in different environments, such as DASs and small cells, which are described in more detail in Sections 4.6 and 4.7. Figure 2.8 illustrates the neutral host ecosystem concept, in which small cell or DAS vendors provide the infrastructure, a venue owner builds and operates a network (perhaps using a third-party integrator), and MNOs lease capacity on that shared network to serve their users.

The term neutral host is often associated with small cells and, indeed, it is expected that most initial neutral host deployments will use small cells; however, there are no technical barriers to deploy the same concept in the macro network.

While the concept of neutral host is not new and not unique to 5G, as for example some CBRS networks are expected to operate in this mode, the integration of unlicensed spectrum into 5G (i.e. NR-U) will make it more easily available to cellular operators. In general, we expect more 5G networks to be rolled out in RAN sharing mode than in 4G, and more neutral host cellular networks to become available with 5G rollout.

## 2.7 Technical Requirements of 5G Radio Interfaces

In order for a proposed radio interface to qualify as an IMT-2020 radio interface (and to make use of IMT-2020 identified spectrum), it has to fulfill certain technical requirements specified in Recommendation ITU-R M.2083. According to M.2083, IMT-2020 applications would fall into the following three broad usage scenarios as illustrated in Figure 2.9:

- eMBB
- URLLC
- MTC.

The eMBB use case addresses human-centric scenarios. It is essentially the natural evolution of the 4G mobile broadband, which is meant to deliver the same services, but with higher throughputs, lower latencies, better power efficiency, and lower cost. It is expected

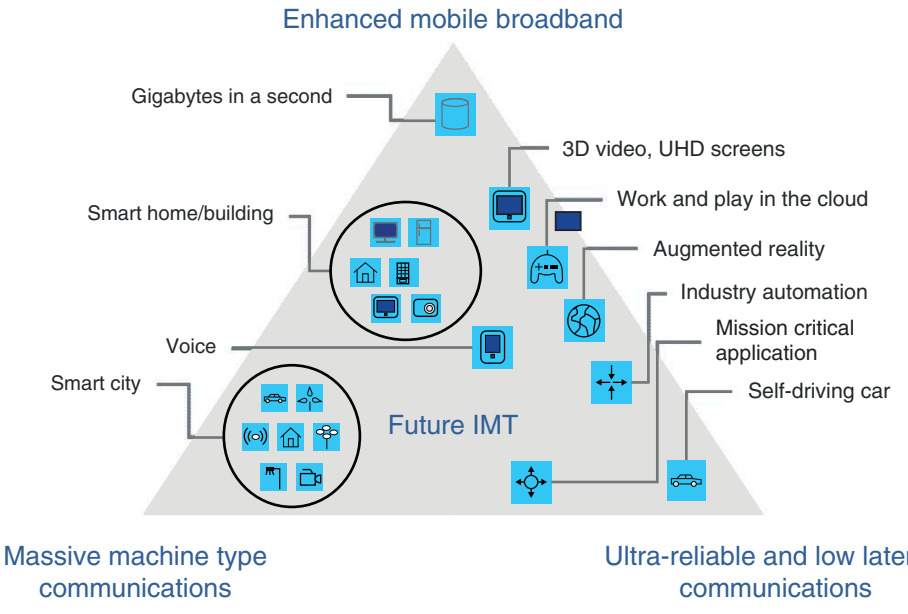



**Figure 2.9** IMT-2020 usage scenarios. (Source: ITU-R).

that these will spur innovation that will bring newer applications and services that are not available today.

The URLLC use case has stringent throughput, latency, and availability requirements. Its applications are under the early stages of development at the moment and it is perhaps the most ambitious goal of 5G. There is hardly anything equivalent as of now and it remains to be seen when future use cases, such as remote wireless control of industrial machinery, remote medical surgery, etc. actually emerge.

And finally, there is the MTC use case, which is characterized by a very large number of devices transmitting low volumes of data. Similar to eMBB, it is not a new use case – LTE NB-IoT, LTE MTC, and LoRaWAN, as well as many other technologies, have been developed in the past to address it.

The full list of IMT-2020 TPRs is beyond the scope of this book, but Figure 2.10 from Recommendation M.2083, and a short summary in Table 2.3, can be used to illustrate how these differ from 4G.

Compared with 4G, 5G technology addressing IMT-2020 requirements will need to support much higher peak and user-experienced data rates, and much lower latencies (albeit not necessarily simultaneously). Both increased throughput and reduced latency will have RAN as well as spectrum impacts. Specifically, regarding latency, while NR design addresses air interface latency, most real-world applications are concerned with latency end-to-end, not just over the air. Reducing end-to-end latency will require both core network (CN) and RAN changes, for example with the usage of multi-access edge computing[9] (MEC), as further explained in Section 6.4.

9 Formerly known as mobile edge computing.

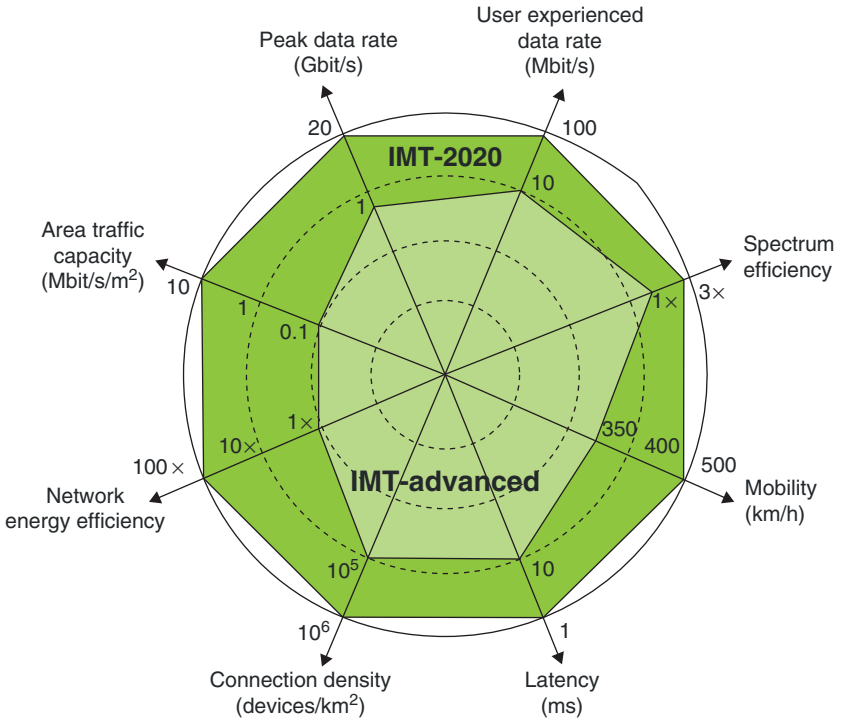



**Figure 2.10** IMT-2020 requirements. (Source: ITU-R).

**Table 2.3** Summary of IMT-2020 requirements.

| Capability | Description |
|---|---|
| Peak data rate | 10–20 Gbps |
| User-experienced data rate | 100 Mbps – 1 Gbps |
| Latency | 1 ms |
| Mobility | 500 km/h |
| Connection density | $10^6$/km$^2$ |

3GPP has decided to submit NR (which is described in this book) as a candidate Radio Interface Technology (RIT) to IMT-2020 as well as NR and LTE, as a set of component RITs (SRIT) (3GPP RWS-180004). In practice this means that while only NR will address all IMT-2020 requirements, LTE (as a component of an LTE plus NR submission), will be able to address some of the IMT-2020 requirements and use cases.

Note: while the primary focus of the book is 3GPP technologies, NR and NG-RAN in particular, it is worth pointing out that at the time of writing this book, there are also submissions toward IMT-2020 that are not based on 3GPP NR. The following RITs and

SRITs have been submitted to ITU-R WP5D for consideration toward becoming IMT-2020 technologies:

- 3GPP submission 1: SRIT – 5G, "Developed by 3GPP as 5G, Release 15 and beyond" (ITU-R WP5D 5D/1215 and 5D/1216)
  - Component RIT: NR
  - Component RIT: E-UTRA/LTE
- 3GPP submission 2: RIT – 5G, "Developed by 3GPP as 5G, Release 15 and beyond" (ITU-R WP5D 5D/1215 and 5D/1217)
  - NR
- China (People's Republic of) – "NR+NB-IoT" (ITU-R WP5D 5D/1268)
  - 'NR+NB-IoT' RIT, which is technically identical to NR RIT and NB-IoT part of 5G SRIT submitted from 3GPP
- South Korea (Republic of) – NR RIT (ITU-R WP5D 5D/1233)
  - 3GPP NR Technical Specifications (Release 15 and 16)
- ETSI (TC DECT) and DECT Forum: SRIT – (ITU-R WP5D 5D/1230 and 5D/1253)
  - Component DECT-2020 NR RIT
  - Component 3GPP 5G candidate for inclusion in IMT-2020: Submission 2 for IMT-2020 (RIT)
- TSDSI (India): TSDSI RIT – (ITU-R WP5D 5D/1231)
  - NB-IoT + NR, with Low Mobility Large Cell (LMLC) configuration as mandatory[10]
- Nufront – EUHT (ITU-R WP5D 5D/1238)
  - EUHT RIT.

One can note that most submissions are based on 3GPP technologies. However, the fact that other technologies, standardized (e.g. DECT) and proprietary (e.g. Nufront) are submitted (in some cases together with 3GPP NR as SRIT), and likely to be accepted, contributes to the confusion about what should be considered 5G. In this book we focus on the 3GPP submissions of NR and LTE, which are both supported by NG-RAN.

## 2.8 Business Drivers

New spectrum and new technical requirements in terms of throughput and latency are not the only forces driving the development of 5G. One additional obvious business driver for 5G is simply an upgrade cycle. The next generation of mobile networks is likely to trigger a network upgrade, which will benefit the network equipment vendors. This in its turn can trigger a handset upgrade, which will benefit the operators and handset vendors. It has been observed that historically every 10 years a new wireless generation is introduced into the market and 5G appears to follow that pattern.

However, 5G business drivers go beyond a mere network and handset upgrade cycle. From MNOs' point of view, while there is still growth opportunity in developing countries for the MBB use case, this is not the case anymore in most developed markets. There-

10 TSDSI submission is based on 3GPP NR RIT submission, with one feature being mandatory instead of optional.

fore, many North American and European operators see 5G as an opportunity to drive down operational costs and, more important, to expand into new markets. Massive IoT, industrial IoT, V2X communications and fixed access (i.e. providing internet connectivity to residential areas) are just some of many examples of use cases that have been served in the past by non-cellular technologies (or did not exist at all), but now are in the focus of 5G as a potential market a mobile operator can expand into.

As mobile phone penetration rates in developed countries are close to 100%, IoT is one of the few most promising growth opportunities for mobile operators (at least in terms of number of "subscribers"). Massive IoT refers to applications that are less throughput- and latency-sensitive but require wide coverage. IoT networks are expected to support large numbers of connections and low-cost, low-energy operation. Smart meters are one example of this use case; however, it is expected that many more use cases will emerge in the future. Mission-critical IoT is a new, somewhat futuristic use case, exemplified by remote machine driving and factory automation. V2X is yet another example of IoT application, which is poised to complement various camera and sensor technologies for assisted and autonomous driving. It is important to mention that in most of these areas 5G will face competition from other standards-based and proprietary technologies, for example, IEEE 802.11, DSRC, LoRa, Nufront, and others. Even though some of these technologies may not formally qualify for 5G (i.e. may not address all IMT-2020 requirements), they may nevertheless serve some of the same use cases (sometimes at lower cost). However, cellular technologies in general and 5G in particular will have an inherent edge over some of the competing wireless technologies due to operator backing and nation-wide coverage.

Therefore, in these new IoT markets, cellular operators will face competition from other players and other technologies. 5G is being designed with such competition in mind – with features such as slicing, which allows operators to rent out certain percentage of their network capacity to a third party, such as an enterprise, a factory, or a fleet operator. This represents yet another growth opportunity for operators.

While operators plan to use the 5G slicing feature to lease parts of their networks to non-operator entities, other 5G features (e.g. non-public network support) allow these new entities to build a 5G network themselves. This goes beyond the traditional model, in which a mobile network is built and operated by a cellular operator holding a spectrum licensed. With features such as unlicensed spectrum operation and non-public network support, 5G can be used by new entities, other than mobile operators – for example, enterprises, factories, etc. This once again shows that 5G is a toolbox of various technologies that can be used in very different cases with very different (sometimes competing) business models.

While cellular operators plan to use 5G to expand to new markets, new companies and whole ecosystems, which used to deploy proprietary technologies, plan to enjoy 5G market of scale and fit 5G into their needs. One such example is satellite communications, which in the past relied on proprietary technologies, but now there is an increased interest (at least from some satellite vendors and operators) in using a 5G technology. Even though non-terrestrial network support was not originally envisioned for 5G, the technology is being adopted for such use in later releases, which is explained in detail in Section 5.3.

Besides market expansion considerations, another important business factor is cost. In order to be successful, 5G deployment is likely to require much higher cell densification, at

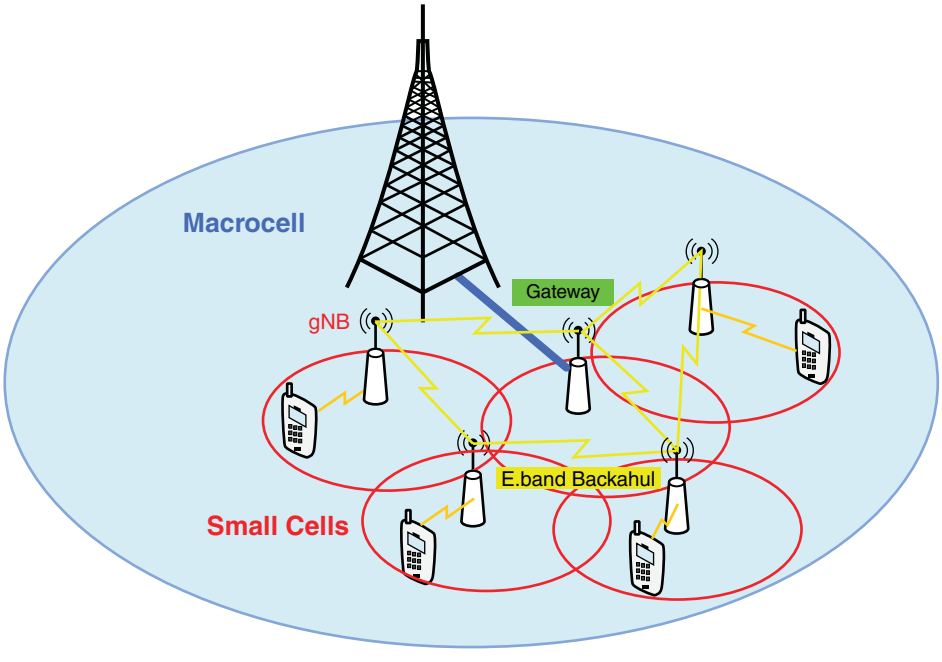



**Figure 2.11** 5G is likely to require massive small cell deployments.

least for the mmWave bands. Therefore, massive small deployments are crucial to realize the 5G potential, as illustrated in Figure 2.11.

Dense small cell deployment will come at a cost, and therefore operators are looking for ways to reduce the CAPEX required to build a 5G network. Since backhaul transport network (explained in Section 6.6) contributes substantially to both CAPEX and OPEX, especially for small cells, operators are considering more cost-efficient alternatives such as relays (explained in Section 5.2). Another important cost-related factor is the network equipment itself. In the past, most operators used to deploy equipment from a single vendor (at least in a given area) and therefore the importance of open network interfaces and multi-vendor interoperability was relatively low. This may change in 5G, considering that it is likely to require massive small cell deployments, which may not be economically viable without competition. Therefore, the importance of RAN architecture based on open interfaces may grow with 5G (this is covered in Chapter 4).

## 2.9 Role of Standards

Historically, standards in general and 3GPP in particular have been crucial for cellular technologies. The importance of standards is unlikely to change with 5G; however, certain market developments affect the way 5G standards are defined and deployed.

3GPP has been extremely successful so far, with LTE in particular. So much so, that it has put the organization in a very peculiar position. It is widely considered to be the

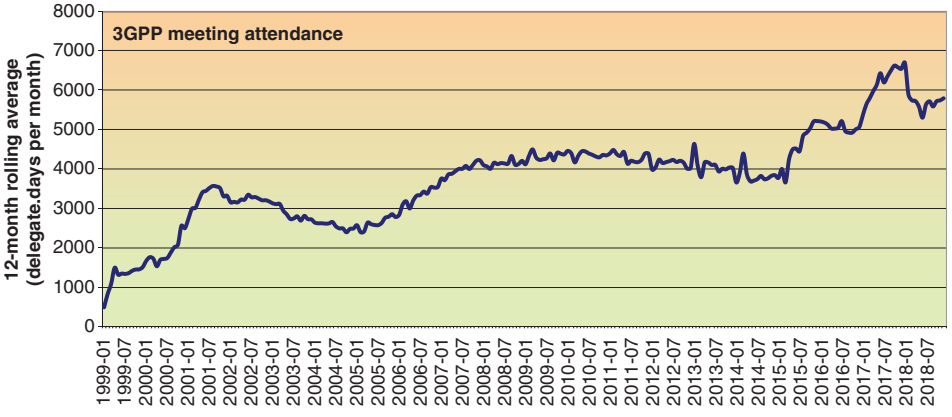



**Figure 2.12** 3GPP meeting attendance.

organization that should specify 5G[11] and therefore it is hard to overestimate its importance to the wireless industry. That importance now attracts many more companies and delegates to 3GPP, so that the numbers of delegates attending, member companies, and documents submitted to each meeting have grown dramatically when 3GPP has started working on 5G. Figure 2.12 illustrates that when 5G standardization activities started in 3GPP, the number of delegates attending increased by more than 50%.

Increased number of companies and delegates does not necessarily mean increased productivity; in fact oftentimes the opposite is true, as 3GPP works by consensus and with bigger numbers of participants consensus is harder to reach. Figure 2.13, for example, illustrates that the number of approved Change Requests (CRs) grew exponentially over the years, as 3GPP grew in popularity.

3GPP standards, of course, undergo an extreme amount of peer review and scrutiny, which helps to ensure high quality specifications. However, increased number of participants often means that more compromises have to be made and often the only way to reach consensus is by adopting multiple options into specifications. One consequence is that NR and NG-RAN specifications have many more options compared with LTE. On one hand, this makes the standards more flexible; on the other hand it makes it much harder to implement, as it is sometimes hard to tell which options are actually going to be deployed. One example of such "extreme flexibility" is the multitude of multiconnectivity options, described in Section 4.3.

While 3GPP is by no means the only SDO developing wireless technologies, with the only exception of IEEE, the situation with other SDOs is peculiar. Largely due to 3GPP success, other competing standards (i.e. standards defining mobile broadband technologies) went largely extinct. In the past, there were at least two more prominent SDOs:

- WiMAX forum
- 3GPP2: the 3rd Generation Partnership Project 2.

11 Indeed, most technologies being submitted to IMT-2020 are 3GPP-based.

| Release | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R99 | 1408 | 4398 | 2266 | 1004 | 581 | 512 | 111 | 42 | 23 | 5 | 5 |
| Rel-4 | | 376 | 2828 | 1900 | 690 | 257 | 122 | 63 | 48 | 22 | 20 |
| Rel-5 | | 27 | 644 | 3274 | 2842 | 2162 | 1357 | 509 | 94 | 25 | 22 |
| Rel-6 | | | | 172 | 1088 | 2458 | 3721 | 2074 | 1078 | 212 | 74 |
| Rel-7 | | | | | 1 | 20 | 663 | 2529 | 3132 | 1262 | 492 |
| Rel-8 | | | | | | | | 49 | 777 | 4609 | 7073 |
| Rel-9 | | | | | | | | | | 49 | 2918 |
| Rel-10 | | | | | | | | | | | 47 |
| Rel-11 | | | | | | | | | | | |
| Rel-12 | | | | | | | | | | | |
| Rel-13 | | | | | | | | | | | |
| Rel-14 | | | | | | | | | | | |
| Rel-15 | | | | | | | | | | | |
| Rel-16 | | | | | | | | | | | |
| Rel-17 | | | | | | | | | | | |
| Total | **1408** | **4801** | **5738** | **6350** | **5202** | **5409** | **5974** | **5266** | **5152** | **6184** | **10651** |

| 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | | | | | | | | **10356** |
| 13 | 11 | 8 | 7 | 2 | | | | | **6367** |
| 7 | 7 | 5 | 3 | 2 | | | | | **10980** |
| 19 | 17 | 10 | 4 | 4 | | | | | **10931** |
| 176 | 91 | 43 | 25 | 15 | | | | | **8449** |
| 2347 | 706 | 347 | 184 | 91 | 46 | 22 | 12 | 11 | **16274** |
| 4991 | 3252 | 1366 | 376 | 165 | 70 | 26 | 23 | 12 | **13248** |
| 1722 | 3800 | 3103 | 1636 | 584 | 267 | 50 | 21 | 22 | **11252** |
| 32 | 1152 | 4178 | 3525 | 2186 | 622 | 163 | 41 | 38 | **11937** |
| | 6 | 102 | 2254 | 5181 | 4287 | 1530 | 197 | 67 | **13624** |
| | | | 8 | 200 | 2648 | 5571 | 1681 | 309 | **10417** |
| | | | | | 24 | 2315 | 5865 | 1236 | **9440** |
| | | | | | | 6 | 1246 | 9728 | **10980** |
| | | | | | | | 35 | 464 | **499** |
| | | | | | | | | | **0** |
| **9308** | **9042** | **9162** | **8022** | **8430** | **7964** | **9683** | **9121** | **11887** | **144754** |

**Figure 2.13** Number of CRs per year.

However, these days 3GPP is pretty much the only organization developing mobile broadband standards. The positive effect of this is less market fragmentation, as the same technology is deployed all over the world, which leads to lower development costs and better user experience (e.g. when traveling). However, there is also a negative effect, since lack of competition puts less pressure on 3GPP in general and 3GPP members in particular, contributing to the difficulty of reaching consensus.

While 3GPP does the technical work of defining the 5G specifications, it becomes an official standard once it is formally adopted by a regional SDO (i.e. a 3GPP Organizational Partner). These are:

- ARIB: the Association of Radio Industries and Businesses, Japan
- ATIS: the Alliance for Telecommunications Industry Solutions, US
- CCSA: China Communications Standards Association
- ETSI: the European Telecommunications Standards Institute
- TSDSI: Telecommunications Standards Development Society, India
- TTA: Telecommunications Technology Association, Korea
- TTC: Telecommunication Technology Committee, Japan.

While 3GPP is arguable the most important group working on 5G specifications, there are a number of other relevant organizations and industry fora. Description of various SDOs is beyond the scope of this book, but to illustrate the abundance of organizations working on 5G and related technologies we provide the short list below:

- LoRa Alliance
- ETSI
- ITU-T: International Telecommunication Union Telecommunication Standardization Sector
- IEEE
- IETF: the Internet Engineering Task Force
- the O-RAN Alliance
- BBF
- Small Cell Forum
- SAE International.

While some of the organizations listed above develop standards that may in some cases be considered as competition (e.g. LoRa and 3GPP NB-IoT), some are actually complementary in the sense that they define standards for mobile networks aspects which are not handled by 3GPP for various reasons. For example, the BBF addresses issues related to the transport network that are often overlooked by 3GPP (see Section 6.6). Furthermore, O-RAN and Small Cell Forum define base station functional splits, which, even though they are conceptually in 3GPP scope, could not be defined there for largely political reasons (for details, see Sections 4.5 and 4.7).

In addition to SDOs and industry fora developing technical specifications, there are a number of associations and interest groups promoting certain 5G and related technologies and/or attempting to steer 3GPP standardization process toward certain areas of interest, primarily through publications of White Papers and by means of liaison exchange with SDOs, such as 3GPP. Some of these include:

- 5GAA: the 5G Automotive Association
- GSMA: the GSM Association
- GSA: Global mobile Suppliers Association
- 5G Americas
- NGMN: Next Generation Mobile Networks
- 5G ACIA: the 5G Alliance for Connected Industries and Automation
- AECC: Automotive Edge Computing Consortium
- 5G-PPP: the 5G Infrastructure Public Private Partnership
- WBA: Wireless Broadband Alliance.

While some of the above organizations are not new, 5G triggered an increased proliferation of such associations. This is perhaps due to 3GPP dominance, as now certain industries, instead of defining a standard to suit their needs, oftentimes attempt to influence 3GPP to adopt their use case into a global cellular standard (e.g. 5GAA).

And yet another category is community projects, which to some extent combine the roles of SDOs, special interest industry groups, and open source projects, such as the Telecom Infra Project (TIP). TIP largely relies on specifications developed by other organizations (such as 3GPP and O-RAN), or in some cases does not use standards at all, but instead allows different players (e.g. vendors and operators) to work jointly on projects of interest, developing solutions for these. Its primary importance lies in creating a platform allowing smaller vendors to work for large operators, which is something that is otherwise hard to achieve.

The above list is by no means exhaustive, but it shows the sheer amount of interest in 5G and the proliferation of different organizations working on 5G requirements, use cases, best practices, and standards. While some of them are undoubtedly important, such as 3GPP, IEEE, BBF, O-RAN, and IETF, the role of the others is hard to predict at the moment.

The end result of this situation is that, on one hand, there is likely to be only one major 5G standard to implement (based on 3GPP specifications), however that standard will have many more options compared with similar standards in the past. 5G standard has effectively become a "toolbox" rather than a single standard, out of which different vendors and different operators will select features that they find useful to implement. How the market will deal with that situation remains to be seen.

## 2.10 Role of Open Source

While standards remain undoubtedly important and arguably irreplaceable, at least on the air interface, an alternative to standards is starting to emerge in the form of open source. Ultimately, standards help to ensure multi-vendor interoperability, especially when there is a conformance certification process, which is often developed to accompany a standard.

The same result can be achieved through open source. With the open source paradigm, companies that would otherwise contribute to the development of a specification (to be used for development of interoperable products), contribute to an open source project that everybody can use – thus achieving interoperability by the virtue of the fact that all vendors would presumably have the same baseline implementation. This of course does not necessarily mean that open source makes all products identical, just as conformance to a common standard does not necessarily mean that all conforming products are the same. Just as a good standard leaves room for differentiation by specifying only what needs to be specified, an open source implementation may still allow extensions and enhancements (depending on the open source license used, as explained below).

Open source initiatives have been extremely disruptive in the software world, with the Linux operating system being the starkest example. Linux is extremely successful with most web servers in the world running on it and being at the core of the Android operating system. While Linux and most other open source projects in the past initially were driven by volunteers running code in their spare time, in the past few years the vast majority of open source codes are in fact contributed by corporations. The Tux penguin, the Linux mascot, has become well known and widely recognizable (see Figure 2.14).

Linux is also extensively being used by enterprises, however this does not necessarily mean that companies using Linux simply download the software (which is often, albeit not always, available free of charge), but rather most choose to rely on integrators (such as RedHat), which provide solutions based on open source, along with support and other services important for enterprises.

With open source being successful in enterprise and data centers, it inevitably drew attention from telecom operators, which at first started from the CN. At the time of writing this book, there are a number of open source LTE EPC implementations available; for example,

**Figure 2.14** Tux the penguin, mascot of Linux open source operating system.

Magma, which was developed by Facebook and distributed under the Berkeley Software Distribution (BSD) license.

Besides the EPC, there are several open source activities targeting orchestration, which is an important component required to run the network in a virtualized environment. The two best-known examples are:

- ONAP
- OSM.

ONAP was formed as a combination of the Linux Foundation's OPEN-Orchestrator Project (OPEN-O), and the AT&T ECOMP (Enhanced Control, Orchestration, Management and Policy). It is an open source software for design, creation, and orchestration of primarily CN services. OSM is an ETSI initiative for the development of open source NFV Management and Orchestration software, meant to achieve similar goals to those of ONAP.

So far, open source telecom projects have been primarily targeting the CN. While it is technically harder to apply the same concept to RAN, because RAN cannot be fully implemented in software, there are attempts to do so in the following organizations/projects:

- TIP
- O-RAN Alliance
- OSA: OpenAirInterface Software Alliance.

While OSA is primarily focused on open source LTE and 5G implementations, both TIP and O-RAN have much bigger scope, with open source being a part of it.

TIP OpenCellular (OC) is an ecosystem of open source projects focusing on hardware, software, testing automation, and manufacturing and building tools for ease of deployment and operation of a cellular network.

O-RAN Working Group 7 (WG7) has been established to promote 5G white box hardware, while O-RAN WG8 is developing software that conforms to O-RAN specifications.

It is important to point out that regardless of the maturity of the open source projects mentioned above (some of which are in rather initial stages of development at the time of writing this book), it is unlikely that an operator would be able to simply download and deploy such software. As has been the case in other industries using open source, oftentimes a third-party integrator is involved, whose role is to "assemble" and, more importantly, test and certify the final product based on open source components. While open source CNs and OAM are beyond the scope of this book, for more details about open source RAN, refer to Section 6.3.

Note: when dealing with open source, an important consideration is a license that an open source project is using. There are many open source licenses, which vary in particular in terms of the amount of freedom allowed and also in terms of restrictions imposed. A full overview of this subject is beyond the scope of this book; here we provide a few examples of the most popular open source licenses:

- A BSD license is considered "permissive", imposing very few restrictions on the use and distribution of the software, including re-use and extensions.
- An Apache license is essentially similar to BSD, allowing users to use the software for any purpose, to distribute it, and to modify it, without concern for royalties. The language of the Apache license is somewhat more elaborate compared with BSD, making it more appealing to enterprises and therefore more popular.
- The GNU General Public License (GPL) is a very popular open source license (GPLv2 is the license used by the Linux kernel, for example). It is considered a "copyleft" open source license, which guarantees end users the freedom to run, study, share, and modify the software. However, it imposes a number of important restrictions; for example, that derivative work must be open source and distributed under the same license terms.
- The OpenAirInterface Public License, even though nowhere near as popular as the ones mentioned above, is important to mention as it represents a different type of open source license, which is incidentally used for one of the most popular open source RAN implementations. It is a modified version of the Apache licenses, with one significant difference – it allows contributing parties to charge royalties based on patents for commercial exploitation of the software.

## 2.11 Competition

Competition certainly plays a big role in driving 5G. It is not confined, though, to the usual competition between mobile operators (in what is sometimes referred to the "race to 5G"), and network equipment and handset vendors. Additionally, with 5G we are likely to see a competition between industries, market segments, and technologies.

On one hand, 5G technology is aiming to expand beyond the traditional mobile markets, while on the other hand new ecosystems and market segments are looking to exploit the scale of the global 5G technology. Therefore, in addition to the traditional competition among:

- Mobile operators
- Network infrastructure vendors

- Handset vendors.

We are likely to see competition between:

- Wireless technologies (e.g. 3GPP NB-IoT vs. LoRa)
- Ecosystems or market segments (e.g. cellular operators vs. satellite operators).

Competition among mobile operators, their network equipment vendors, and handset manufacturers is not new. Neither is the competition between technologies, as was the case in the past, for example, between LTE and Worldwide Interoperability for Microwave Access (WiMAX). However, we expect that 5G will spur more aggressive competition over new markets and even spectrum, as we see in WRC-19.

For example, 5G is expected to open new markets – markets that already have their incumbent players. The satellite industry is looking into adopting 5G technologies to provide mobile broadband services, thus increasing competition with cellular operators. Cable companies are interested in that market too, through the use of 5G in unlicensed or shared spectrum. On the other hand, cellular operators consider using 5G in "fixed access" mode to provide internet service in, for example rural areas, thus competing with cable companies. There will also be a competition between private networks deployed by non-operator entities (e.g. enterprises) to serve their needs and cellular operators, who would like to serve them by leasing parts of their networks using, for example, slicing.

Some of this competition is already visible in discussions around 5G spectrum allocations. While some regulators consider putting aside certain chunks of spectrum for verticals, operators would prefer that spectrum awarded to them, so that they can serve the same industry using slicing. Additionally, there is an ongoing debate on spectrum allocations for mobile and satellite industries, for example, in WRC-19. Generally, competition is a positive force, driving innovation and reducing costs.

On the other hand, this kind of competition (between technologies and/or between industries) is likely to cause market fragmentation (something that we already observe in the V2X space and may actually drive costs up, not down).

All in all, 5G is likely to increase competition in the wireless space, which is probably a good thing. Increased competition will force providers to invest more and end users will hopefully see better prices with new and improved services. Whether and when these benefits materialize remains to be seen.

## 2.12 Challenges

5G is often touted as the next wireless revolution, promising faster speeds, lower costs, better energy efficiency, hyper connectivity, and new exciting use cases and applications. The 5G technology defined in 3GPP is certainly up to that task – it is capable of fulfilling the technical requirements of IMT-2020 (and in fact beating these) and can do much more than what has been envisioned for IMT-2020. The question, though, remains, when it will be deployed at large scale, which features out of many specified by 3GPP will be used, what applications will make use of the network capabilities, and, perhaps not the least important,

whether operators will be able to monetize their investments in 5G spectrum and network build-out.

Massive and mission-critical IoT are very promising technologies; however, as of now the main wireless network usage remains what has been the bread and butter for cellular operators – mobile broadband. It is obvious that mobile broadband will remain important, but the big unknown is how much growth potential there is in that market and whether users will be willing to pay premium for the higher speeds 5G can deliver. It is generally accepted that the smartphone market is saturated and therefore for 5G to live up to expectations it must be successful in new markets other than mobile broadband.

Massive IoT is widely believed to be the next big thing for the wireless industry. This may very well be the case, however the challenge here is that it requires a completely different business model than the cellular operators are used to. While the number of IoT devices will eventually be orders of magnitude bigger than the number of human cellular users, the revenue from each device will be much smaller. It is unclear how profitable massive IoT can be for traditional mobile operators and cellular equipment manufacturers.

Device cost is another big challenge. While 3GPP has made great efforts in reducing costs for IoT technologies such as NB-IoT, the cellular industry is known for producing technologies with high performance and high cost. Downsizing such technologies to low performance, low power, and, most importantly, low-cost use cases is a major challenge, especially taking into account the availability of cheaper alternatives designed specifically for massive IoT. Mobile operators will have a certain advantage compared with these alternatives in terms of coverage and reliability, as they can use licensed spectrum and in some cases would be able to rely on their existing networks to provide nation-wide coverage, but the business model issue still needs to be addressed.

While there are many advantages to licensed spectrum, which is often assumed to be used by 5G technologies, it comes with a cost. Mobile operators bearing the costs of 5G spectrum will look for new ways to monetize their investment by entering new markets and seeking new use cases. The issue is that while at least some of these use cases (e.g. V2X) can benefit from increased reliability of operation in licensed spectrum, the business model and operational complexity of (potentially multiple) mobile operators operating a V2X network in the same geographical area are significant. In such cases, a simpler technology, operating in a somewhat less reliable unlicensed band may prove to be easier to deploy and monetize.

Last but not least is the challenge of 5G network deployment and site acquisition in particular. In order to deliver Gbps speeds, dense network deployment of a large number of small cells is required. Even before 5G, network densification could have improved network speed dramatically; yet very few operators deployed small cells on a large scale. This is primarily due to difficulty and cost of massive cell site acquisition and maintenance, which will remain a problem for 5G as well. New business models centered around the neutral host idea may to some extent alleviate the problem; however, it remains a significant challenge that operators will have to find a way to overcome.

Despite these challenges, there is a significant momentum behind 5G and there is no doubt that it will be deployed. Timing remains the biggest unknown.

## 2.13 Summary

In this chapter we outlined the major market drivers behind 5G, ranging from technical requirements, to new spectrum, new deployment and business models, and deployment challenges. From these we attempted to derive how 5G technical requirements and business drivers affect RAN architecture design and deployment, and its evolution in 5G.

In our view, the main factors impacting RAN architecture redesign in 5G are:

- Increased throughputs
- Reduced latency
- Network densification
- Competition.

4G is typically associated with throughputs of few hundreds of Mbps, whereas 5G is expected to deliver (at least in the mmWave bands) throughputs of many Gbps. Such a drastic increase in throughput affects not just the air interface design, but also the fronthaul and backhaul transport network capabilities, and the RAN architecture. In particular, it is no longer reasonable to expect that even with fiber fronthaul transport, CPRI[12] architecture would be able to sustain such throughputs. This drives the desire to re-architect RAN in order to reduce fronthaul throughput requirements, for example, by moving more functionality closer to the edge. This is explained in detail in Chapter 4.

Significant efforts have been made to reduce latency in 5G, which mainly focused on the air interface. However, for most mission-critical low-latency applications what is important is the overall end-to-end latency, and therefore air interface improvements alone cannot fulfill that requirement. In order to achieve actual end-to-end latency reduction, RAN architecture changes are also required. For example, bringing the content closer to the edge (i.e. to the RAN) using MEC is one such option. MEC is explained in Section 6.4.

The usage mmWave in, for example, the 28 GHz frequency range can provide great throughput improvements; however, this comes with the limitation of reduced cell sizes. Realistically, mmWave can only be deployed with small cells, which so far have had little market traction. 5G is likely to see massive small cell deployments (in order to realize the mmWave potential); however, this brings certain challenges in terms of transport network and site acquisition, which need to be addressed. More importantly, though, potential massive small cell deployment creates an opportunity for new network equipment vendors. Section 4.6 covers small cells. If operators deploy network equipment from multiple vendors,[13] the importance of network interface standards and standardized RAN architecture will increase.

Slicing and private networks are important 5G features, allowing competition beyond traditional wireless models. With slicing, a mobile network operator can offer some of their

12 CPRI standard defines an interface between Radio Equipment Controllers (RECs) and Radio Equipment (RE). It is commonly used in 4G networks, however almost all vendors have implemented proprietary extensions on top of the standard interface, thus it cannot be considered multi-vendor interoperable. A CPRI link transports digitized RF signals and therefore has high transport network bandwidth requirements.

13 In 4G, most mobile operators deploy equipment from a single vendor in a given area.

network capacity to a third party. On the other hand, private network support and NR operation in unlicensed spectrum allow new market players, who may not hold licensed spectrum, to use 5G technologies. Similar, relay and non-terrestrial network support (explained in Sections 5.2 and 5.3, respectively), which have been added to NR in later releases, allow the usage of this technology in new markets, previously dominated by other (proprietary or standard) technologies. These features require proper RAN dimensioning and design, which is addressed in Chapter 7.

These and other factors are driving the RAN architecture evolution explained in this book.

## References

5GAmericas (2016). 5G Spectrum Recommendations. Bellevue, WA: 5GAmericas.

ITU-R M.2083-0 (2015). IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond. ITU.

ITU-R WP5D (2017). Spectrum needs for the terrestrial component of IMT in the frequency range between 24.25 GHz and 86 GHz, document TG5–1/36. ITU.

ITU-R WP5D 5D/1215 (2019). Alliance for Telecommunications Industry Solutions, 3GPP final technology submission – overview of 3GPP 5G solutions for IMT-2020. ITU. ITU-R WP5D 5D/1216, Alliance for Telecommunications Industry Solutions, "3GPP 5G candidate for inclusion in IMT-2020: Submission 1 (SRIT)."

ITU-R WP5D 5D/1216 (2019). Alliance for Telecommunications Industry Solutions, 3GPP 5G candidate for inclusion in IMT-2020: Submission 1 (SRIT). ITU.

ITU-R WP5D 5D/1216 (2019). Alliance for Telecommunications Industry Solutions, 3GPP 5G candidate for inclusion in IMT-2020: Submission 2 for IMT-2020 (RIT). ITU

ITU-R WP5D 5D/1230 (2019). European Telecommunications Standards Institute, Description template of SRIT candidate for inclusion in IMT-2020. ITU.

ITU-R WP5D 5D/1233 (2019). Korea (Republic of), Final submission of a candidate technology of IMT-2020. ITU.

ITU-R WP5D 5D/1238 (2019). Nufront (Beijing) Technology Co., Ltd, Submission of candidate IMT-2020 – radio interface technology (EUHT). ITU.

ITU-R WP5D 5D/1238 (2019). TSDS India (TSDSI), Updated submission of the candidate IMT-2020 Technology. ITU.

ITU-R WP5D 5D/1253 (2019). DECT Forum, Support of the IMT-2020 submission from ETSI. ITU.

ITU-R WP5D 5D/1268 (2019). China (People's Republic of), Final submission of candidate IMT-2020 radio interface technology. ITU.

Magma (n.d.). Magma https://github.com/facebookincubator/magma (accessed ).

NGMN (2015). 5G White Paper, 2015. NGMN.

Recommendation, ITU-R M.2083 (2017). IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond. ITU.

Report ITU-R M.2243 (2011). Assessment of the global mobile broadband deployments and forecasts for IMT. ITU.
Report ITU-R M.2290 (2013). Future spectrum requirements estimate for terrestrial IMT. ITU.
RP-190019 (2019). Support Team Report RAN#83, ETSI MCC. 3GPP.
RWS-180004 (2018). The 3GPP submission. 3GPP.

# 3

# 5G System Overview

## 3.1 Introduction

Before diving into the details of NG-RAN (5G Radio Access Network) architecture, it is important to have at least a high-level understanding of the whole 5G System (5GS). To provide such a high-level picture, in the present chapter we describe the functionalities of the physical layer, the protocol stack, the NG-RAN, and the 5G Core (5GC) network. This chapter is not meant to be a definitive guide to either, as each one deserves a separate book in order to describe all the details. Instead, we provide an overview of all the components of 5GS, with emphasis on what is new compared with 4G.

Readers who are sufficiently familiar with these can skip this chapter and go directly to the next one, for detailed discussion about NG-RAN architecture.

## 3.2 5G Core Network

*Sebastian Speicher*

*Qualcomm Wireless LLC, Switzerland*

### 3.2.1 Introduction

5GS consists of the 5G Radio Access Network (RAN), the 5GC, and the user equipment (UE).[1] In the present section we provide a 5GC overview.

Like earlier generations of the 3GPP system, the 5GC's tasks include:

- *Storing subscription information*, including identifiers, cryptographic information needed for authentication and to derive cipher keys, information about networks a UE may establish data sessions to, and restrictions to specific radio access technologies (RATs) or geographic areas, etc.;
- *Performing mutual authentication* between UE and network and subsequent authorization of the UE;
- *Registering UEs* and keeping track of the list of UEs that are registered with the system;

1 5GS may also include non-3GPP Access Networks (ANs).

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

- *Tracking the location of the UE* at different levels of granularity depending on whether a UE is involved in active communication or not;
- *Establishing data sessions* to different networks for different payloads types (e.g. IPv4, IPv6, Ethernet, etc.) as requested by a UE;
- *Traffic forwarding* in both uplink and downlink directions between RAN and the data networks a UE has established a session to;
- *Deriving Quality of Service (QoS) and charging rules* based on operator polices;
- *Enforcing charging and QoS rules* (the latter in tandem with the RAN);
- *Performing lawful interception*, i.e. providing meta data and access to payload of interception targets in line with legal obligations.

While fulfilling similar tasks as the core network of earlier generations, 5GC follows novel paradigms in various areas:

- *Service-based architecture (SBA)*: in contrast to the Evolved Packet Core (EPC), the core network of the Evolved Packet System (EPS), 5GC procedures are defined based on generic services that are exposed by 5GC network functions. The benefit of generic services is that they can be reused for different system procedures and can be accessed by different network functions (now or in a future release) or can also be leveraged for proprietary operator-specific services. This not only simplifies standardization of new system features but also reduces implementation and testing effort. Most importantly, the SBA approach, as well as the decision to define network function services as application programming interfaces (APIs) using well-established web technologies, addresses the market demand for programmability and extensibility of the 3GPP system by mobile network operators (MNOs) and third parties.
- *Control- and user-plane separation (CUPS)*: separating control and user plane allows for independent scaling and evolution of control-plane functions and user-plane functions. It also enables new deployment models where user-plane functionality is deployed closer to the access network while control-plane functionality remains centralized. As such CUPS is an important enabler for multi-access edge computing (MEC). While CUPS is already supported for EPS (3GPP TS 23.214, 2018), the user-plane deployment scenarios that can be supported using EPS CUPS are limited as CUPS was added on top of the already existing EPS architecture baseline that traces back to Release-8. In contrast to this, 5GC offers a higher degree of deployment flexibility for the user plane as separation of control and user plane was considered from the beginning during the 5GC architecture design phase.
- *Common access-agnostic core network*: 5GC has been defined as a converged core network capable of serving different types of access technologies. Unlike EPS, 5GS uses the same core network architecture and the same interface between access network (AN) and core network for both 3GPP accesses (e.g. NR or Evolved Universal Mobile Telecommunications System Terrestrial Radio Access [E-UTRA]) and non-3GPP accesses (e.g. wireless local area network [WLAN[2]] or wireline access technologies). Furthermore, in difference to EPS, a common non-access stratum (NAS) protocol is used regardless of the underlying access technology.
- *Concurrent and efficient access to localized and centralized services*: besides centralized services (e.g. internet-based services or operator services like IP multimedia subsystem

2 The term WLAN is often used in 3GPP specifications to refer to IEEE 802.11, i.e. Wi-Fi.

[IMS] voice), 5GC has been designed to also support localized services in an efficient manner. In this context localized refers to hosting services closer to the UE, e.g. colocated with a centralized unit (CU) serving a UE (see Section 4.2 for more details on centralized unit/distributed unit] CU/DU] split). Services that benefit from being deployed locally are, for example, Virtual Reality and Augmented Reality (VR/AR) applications since those require very low delays. This approach is also referred to as MEC. The key challenge for MEC is to establish an efficient data-forwarding path between the UE and the closest service instance and to adapt the data-forwarding path as the UE changes location. 5GC provides various enablers to achieve this.

- *Network slicing*: the cellular ecosystem has been expanding into new domains, including low-power wide area (LPWA) Internet of Things (IoT), mission-critical and priority services (MCS/MPS), and more recently industrial automation, also known as Industrial Internet of Things (IIoT). The requirements of those different verticals on the underlying 5G network are, however, very different. Therefore, 5GS deployments will need to support diverse core network functionalities and configurations in the same Public Land Mobile Network (PLMN). Earlier 3GPP system generations already provided an increasing level of support for selecting different core network functions and network configurations for different groups of UEs (e.g. based on the Release-13 dedicated core network[DECOR] feature for EPS). 5GS network slicing adds to this by increasing further the deployment flexibility and isolation of different sets of core network functions, enabling differentiation and isolation of groups of UEs within the RAN, and by specifying slice selection policies for the UE, which provides operators with more control over which application traffic will be handled by which set of network functions.
- *Private networks*: the promise of a radio interface, which supports very low latencies and is highly reliable at the same time, has attracted significant interest from various industries to replace existing wired networks, e.g. for motion control of robot arms or to address logistics use cases. However, traditional wide-area cellular network deployments, which are based on a very distributed radio network and a centralized core network, do not address the data privacy needs and reliability concerns of those industries. Therefore, one of the key themes of 5GS is to enable private network deployments where parts or even the entire network (RAN and core network) is deployed on site, e.g. in a factory.

The remainder of this section illustrates each of these paradigms in greater detail.

## 3.2.2 Service-Based Architecture

### 3.2.2.1 Fostering Functional Reuse

Comparing the EPS architecture (Figure 3.2.1) and the 5GS architecture (Figure 3.2.2) reveals a significant architectural difference.

As depicted in Figure 3.2.1, EPS is based on point-to-point interfaces between network functions[3] such as Mobility Management Entity (MME), Serving Gateway, and Packet Data Network (PDN) Gateway. Consequently, functionality supported on an interface between an EPS network function A and a network function B can only be used by those two network functions. If yet another network function (e.g. network function C) needs to use

3 Sometimes referred to as network nodes.

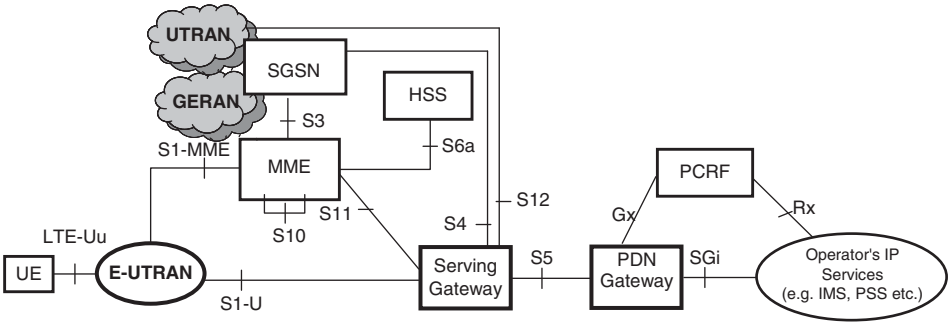



**Figure 3.2.1** EPS non-roaming architecture. (Source: reproduced with permission from © 3GPP).

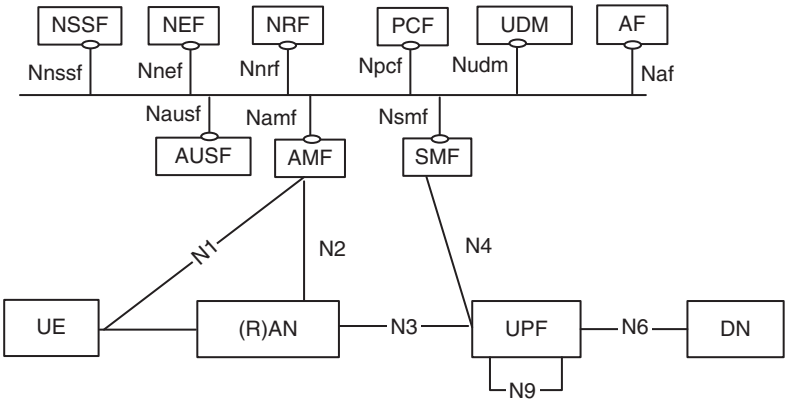



**Figure 3.2.2** 5GS non-roaming architecture. (Source: reproduced with permission from © 3GPP).

similar functionality from network function A, then a new interface needs to be introduced between A and C and the functionality in question needs to be replicated on this new interface. This not only slows down standardization projects in 3GPP, but more importantly poses an obstacle for operators to extend and customize the core network to their needs.

To illustrate this, it is worth looking at an example.

The S11 interface supports reporting of UE location changes (e.g. to report when the UE changes Tracking Areas or changes cells) from MME to Serving Gateway (SGW). As knowledge of user location is useful for many services, it would be interesting to make this information available for use by other network functions.

However, the S11 location reporting functionality cannot easily be used by network functions other than the SGW for the following reasons:

- S11 location reporting is piggybacked on top of session management-related signaling, e.g. Modify Bearer Request (3GPP TS 29.274, 2019). In other words, S11 location reporting assumes a common session context between MME and the other termination point of the S11 interface (typically the SGW).
- Only a single SGW (i.e. single S11 interface between MME and SGW) is allowed per registered UE.

For the same reasons, existing system functionality also cannot be extended easily, for example, to support location reporting for groups of UEs.

As a result, as part of the Release-13 work aiming at exposing network information to third parties via the Service Capability and Exposure Function (SCEF), location reporting from MME for individual UEs and groups of UEs was therefore specified via yet another interface (the T6a interface between MME and SCEF) instead of extending the existing location reporting mechanism.

To avoid such issues, 5GC follows software engineering paradigms such as modularity and self-containment for 5GC network functions to foster reuse and extensibility of system functionality (see the "Design guidelines for NF services" documented in Annex A.6 in 3GPP TS 23.502).

Practically speaking, this means that whenever possible 3GPP should define procedures (i.e. interactions between network functions) as services, so that they can be reused by other network functions and can also be extended more easily in the future.

To illustrate the difference to EPC it is worth looking at the same example for 5GC, that is, how UE location information is provided by the Access and Mobility Management Function (AMF), the 5GC equivalent of the MME. AMF supports the Namf_EventExposure service, which enables any other network function to subscribe for and subsequently get notified about various events including UE reachability status changes, time zone changes, and location changes (3GPP TS 23.502). For location changes additional filters can be defined to further narrow down the events that will be reported (e.g. report some or all tracking area changes only).

In conclusion, the modularized service approach fosters reuse since it makes it simple to access information like UE location for other standardized or proprietary 5GC network functions.

In a similar manner, the functionalities of most 5GC network functions have been specified as services (see clause 5.2 in 3GPP TS 23.502 for an overview of all network function services defined in 5GC).

It is worth mentioning that interfaces between the 5GC control plane and the UPF as well as the interfaces between 5GC and NG-RAN are still following the traditional point-to-point model, that is, are not based on network function services.

Another design decision that can be expected to simplify reuse, foster network programmability, and generally lower the entry barrier for software add-ons on top of the standardized 5GC network functions is the shift from telecom-specific protocols like Diameter (Fajardo et al. 2012) and the GPRS Tunneling Protocol (GTP) (3GPP TS 29.274) to web protocols such as HTTP/2 (Belshe et al. 2015) and JavaScript Object Notation (Bray 2017) as the basis for the 5GC control plane (further protocol details can be found in 3GPP TS 29.500).

#### 3.2.2.2 Overview of 5GC Control-Plane Functions

As illustrated in the previous section, 5GC is based on a new system architecture. Nevertheless, many concepts and system functionalities are similar to those in EPC.

This section presents a high-level overview of the 5GC control plane and the roles of the network functions it consists of (Figure 3.2.2) while also pointing out the key differences to their EPC counterparts.

The 5GC equivalent of EPC's MME is the AMF. AMF's key responsibilities include:

- *Registration management*: before a UE can use network services, the UE needs to perform an initial registration with the AMF. As part of this step, UE and AMF (supported by other functions such as the Authentication Server Function [AUSF] and Unified Data Management [see below]) perform mutual authentication. As an AMF typically serves part of a network only, UEs inform the network when they enter an area that may be served by a different AMF by performing a mobility registration. To enable the network to keep track of which UEs are still registered with the network, UEs perform periodic registrations.
- *Connection management*: to exchange signaling messages with the network (e.g. to request 5GC to establish a data session), UE and AMF need to first establish a secure signaling channel, also known as a NAS signaling connection. Establishing and releasing the NAS signaling connection is referred to as connection management.
- *Reachability management*: while UE and AMF have an active NAS signaling connection, i.e. while the UE is in CM-CONNECTED state, AMF is aware of the actual NG-RAN cell serving the UE and can directly exchange signaling messages with the UE. However, to save power and network resources, the NAS signaling connection is typically released when no further data or signaling is to be exchanged between UE and network. Without NAS signaling connection the UE is in CM-IDLE state. In this state the network is not aware of the exact location of the UE. To bring the UE back into CM-CONNECTED state, e.g. to deliver downlink data or to send signaling messages to the UE, the AMF triggers NG-RAN to broadcast paging messages for this UE.

To send and receive data, also referred to as protocol data units (PDUs), UEs need to first request establishment of a PDU session toward a data network. In 5GC, the Session Management Function (SMF) is in charge of establishing, modifying, and releasing PDU sessions. 5GS supports PDU sessions for IPv4, IPv6, Ethernet, and unstructured data.

The key difference between session management in 5GC and EPC is as follows: while in EPC session management functionality is split across two functions, the SGW and the PDN Gateway (PGW), 5GC session management functionality has been merged into a single entity, the SMF.

This decision is related to the user-plane design in 5GC: as illustrated in Section 3.2.3, 5GC does not have distinct user-plane entities like SGW-U and PGW-U but only a generalized UPF, which can take different roles.

In line with this, also the related control-plane functionality (SGW-C and PGW-C) has been generalized and merged into a single architectural entity (the SMF).

The Policy Control Function (PCF) is the equivalent of EPC's Policy and Charging Rules Function (PCRF). As in EPC, either based on interactions with applications or based on triggers received from the SMF, the PCF takes policy decisions and provides PCC rules to the SMF. In contrast to the PCRF, the PCF has the additional responsibility to provide policy information to the UE (via SMF and AMF).

Policy information that can be provided from PCF to UE includes the following:

- *Access Network Discovery and Selection Policy (ANDSP)*: this information is used by the UE to select and register to non-3GPP access networks.

- *UE Route Selection Policy (URSP)*: URSP enables UEs to determine how the traffic of a given application should be routed: via a 3GPP or non-3GPP access, using an already established PDU session or using an additional, yet-to-be established PDU session. URSP also conveys to the UE additional parameters to provide to the network in case a new PDU session needs to be established including type of PDU session (e.g. IPv4, IPv6, or Ethernet), network slicing-related information (see also Section 3.2.6), and identifiers for the data network the PDU session should be established to.

The *Network Exposure Function (NEF)* is the counterpart of EPC's SCEF as defined in 3GPP TS 23.682. NEF exposes the capabilities of network functions, that is, enables MNO or third-party applications to access 5GC network functions. External exposure supports:

- Monitoring capability, which allows applications to receive an indication when a UE becomes reachable;
- Provisioning capability to enable applications to provide information about the expected UE behavior (e.g. the UE's expected mobility pattern) to 5GC;
- Policy/charging capability to allow applications to request QoS and charging treatment for specific flows.

Additional auxiliary 5GC functions include:

- The *Unified Data Management (UDM)*, which is the equivalent of the Home Subscriber Server in EPC and performs subscriber data management.
- The *Network Repository Function (NRF)* allows for dynamic discovery of other network functions, e.g. enables AMF to discover an SMF when a UE requests to establish a PDU session. To support this, APIs have been specified (3GPP TS 29.510) for network functions to register their own profile (e.g. network type, supported network slices, etc.) with the NRF and APIs to enable network functions to query NRF. In the case of EPC, only query functionality was specified (network function registration was left to implementation) and was realized using the Domain Name System (DNS) (3GPP TS 29.303).
- The *Network Slice Selection Function (NSSF)* determines the set of network slice instances to serve a UE (see Section 3.2.6 for further details about network slicing).

### 3.2.3 Control-User Plane Separation (CUPS)

The motivation for separating control- from user-plane functionality is twofold:

- Independent scaling of control- and user-plane processing capacity;
- Increased flexibility when deploying control-plane functions and user plane functions, e.g. to enable operators to locate control-plane functions in a central location (e.g. in a data center) while placing UPFs close to the RAN.

It is worth noting that the ability to deploy user-plane functionality close to the RAN is a key enabler for efficient access to localized services (see Section 3.2.5 for further details).

CUPS has initially been studied and specified for EPS. As described in 3GPP TS 23.214 and as depicted in Figure 3.2.3, CUPS has been realized as an extension to the existing EPC architecture by splitting SGW and PGW into their control- and user-plane components. The SGW control-plane function (SGW-C) and PGW control-plane function (PGW-C) control

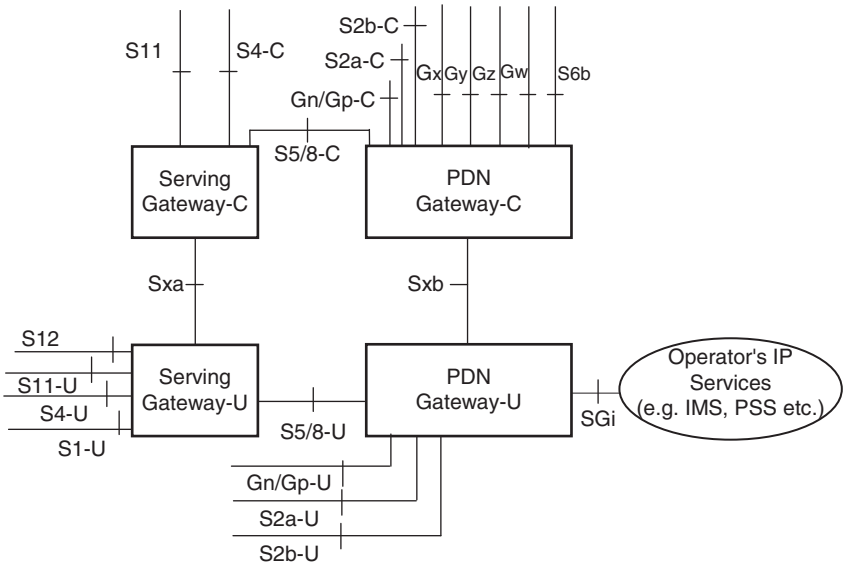



**Figure 3.2.3** Separation of user plane and control plane in EPS. (Source: reproduced with permission from © 3GPP).

the user-plane forwarding within their respective user-plane functions by providing rules to classify incoming packets, add/remove header information, and forward packets to next hop user-plane nodes.

While following a similar approach, 5GC offers a more flexible method for control and user-plane separation since CUPS was not superimposed onto an existing architecture as in EPC. Instead, separation of control and user plane was a design target from the beginning during the 5GS study phase. This resulted in the following key characteristics:

- Only a single UPF has been specified, which can be configured by the SMF to take different roles (e.g. the role of an intermediate user-plane entity like the SGW-U in EPS, which forwards packets between two tunnels, or the role of a session anchor like the PGW-U in EPS, which adds/removes tunnel headers, classifies packets for charging, and QoS rule enforcement, etc.).
- The number of UPFs that can be chained for a PDU session is not restricted by specifications, i.e. 5GC supports PDU sessions using only a single or multiple UPFs.

This flexibility enables the SMF to chain UPFs for different scenarios, for example, to enable access to both local and central data networks using a single PDU session (see Figure 3.2.4). Section 3.2.5 provides more details and examples of how the UPF can be configured for such scenarios.

### 3.2.4 Common Access-Agnostic Core Network

One of the goals during the design phase of the 5G system architecture was to define a common core network that could be used for 3GPP accesses (e.g. E-UTRA and NR) but also for non-3GPP access technologies (e.g. WLAN, digital subscriber line [DSL], or fiber). The

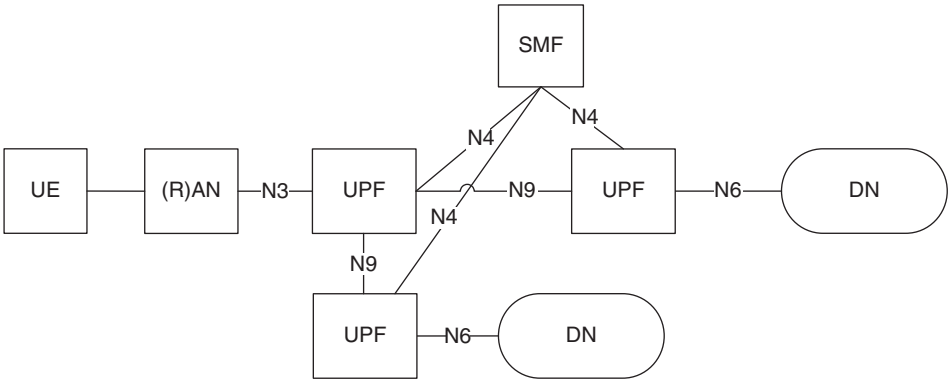



**Figure 3.2.4** 5GC user-plane configuration with concurrent access to a local and central data network using a single PDU session.

key motivation for this was to enable integrated operators who offer both fixed and mobile services to harmonize their core network infrastructure to reduce Capital Expenditure and Operational Expenditure.

When looking into the details of WLAN ANs or wirelines networks (e.g. as defined by Broadband Forum), it becomes clear that some system aspects will continue to be access specific, for example, security, mobility handling, or QoS enforcement.

Therefore, to achieve the goal of a common core network despite this, 3GPP decided to

- hide AN specifics from 5GC inside access network-specific adapter functions, e.g. the Non-3GPP Interworking Function (N3IWF) for WLAN or the Wireline Access Gateway Function (W-AGF) for wireline ANs as depicted in Figures 3.2.5 and 3.2.6, respectively;
- define common control- and user-plane interfaces between the core network and ANs that apply to both 3GPP accesses (E-UTRA and NR) as well as to the adapter functions for non-3GPP accesses, i.e. N3IWF and W-AGF.[4]

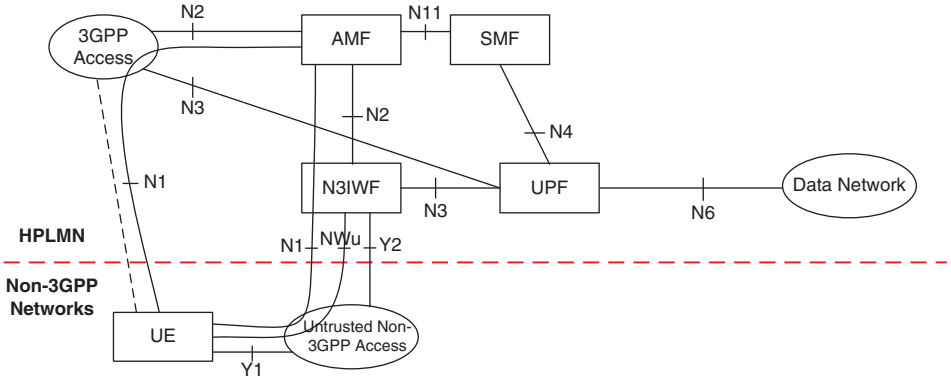



**Figure 3.2.5** N3IWF hides specifics of non-3GPP access networks (e.g. WLAN) from 5GC. (Source: reproduced with permission from © 3GPP).

4 In practice, even though the interfaces are largely common, there are still some access-specific messages and information elements.

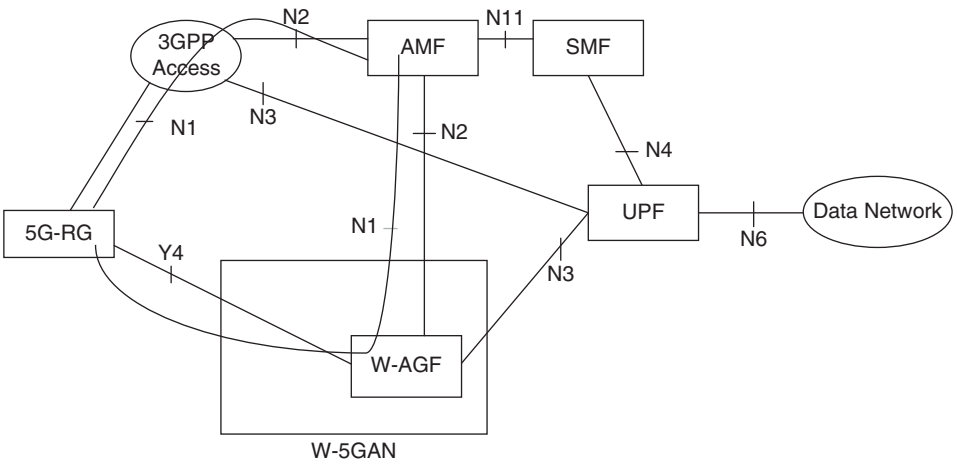



**Figure 3.2.6** W-AGF hides specifics of wireline access networks from 5GC. (Source: reproduced with permission from © 3GPP).

While this approach may come across as merely shifting complexity from one part of the system to another, it has two salient benefits:

- The same control- and user-plane architecture (AMF, SMF, UPF with common interfaces toward other functions like PCF, UDM, etc.) applies to all accesses. In contrast to this, EPS is based on different architectures for 3GPP and non-3GPP accesses (see 3GPP TS 23.401 (2019) and 3GPP TS 23.402 (2019) for details).
- A common NAS protocol for mobility and session management is used between the UE or 5G Residential Gateway and the core network for both 3GPP accesses and non-3GPP accesses (e.g. WLAN or wireline access technologies). This is a significant simplification compared with EPS where e.g. different session management protocols were used between UE and core network for 3GPP access and non-3GPP technologies.

### 3.2.5 Enablers for Concurrent and Efficient Access to Local and Centralized Services

#### 3.2.5.1 Overview

There has been an increasing demand to host services closer to the user (an approach also known as MEC) to reduce end-to-end latency, for example, for virtual and augmented reality applications. Another driver to host services at least partially closer to the user is to reduce traffic volume across backbone links by caching content within the RAN aggregation network. At the same time, many services will continue to be hosted in central locations, for example, IMS voice and many internet services. Therefore, it is important to support concurrent access to both local and centralized services.

This raises the following questions:

1. How can an efficient data-forwarding path be enabled between the UE and local services as well as centralized services?

2. How can the data-forwarding path be updated when the UE changes its location?
3. How can it be determined when to update the data-forwarding path and where to forward data to depending on the locations where a given service is hosted?
4. How can the operator or applications be enabled to influence which traffic should be routed locally and which traffic should be forwarded to a central location?

The 5GC mechanisms that have been specified (3GPP TS 23.501) to address these questions can be classified based on whether they operate within a single PDU session or across multiple PDU sessions and subsequently whether the UE is aware of and involved in the routing of traffic to local services or not.

#### 3.2.5.2 Single PDU Session-Based Access to Local Services

5GC supports two mechanisms operating within a single PDU session: Uplink Classifier (ULCL) and multi-homed PDU sessions.

As illustrated in Figure 3.2.7, the ULCL is a UPF functionality that can be activated by the SMF to divert some of the uplink traffic of a PDU session toward a local data network that hosts a local copy (or instance) of the service that the UE is trying to access. Similarly, in the downlink direction, ULCL injects traffic from the service instance hosted in the local data network into the PDU session and forwards the traffic to the UE. Traffic classification is done based on filtering rules (e.g. matching certain destination IP ranges, etc.) provided by the SMF based on operator configuration.

As the UE changes location, the SMF ensures an efficient routing path between the UE and the closest service instance by selecting a different UPF to terminate the N3 interface from NG-RAN and to act as a ULCL. The new UPF diverts the traffic to a local data network that is closer to the UE's new location and that also hosts an instance of the same service.

The UE is generally not aware of the ULCL. In other words, the UE is not aware that some of its traffic is routed to a local data network.

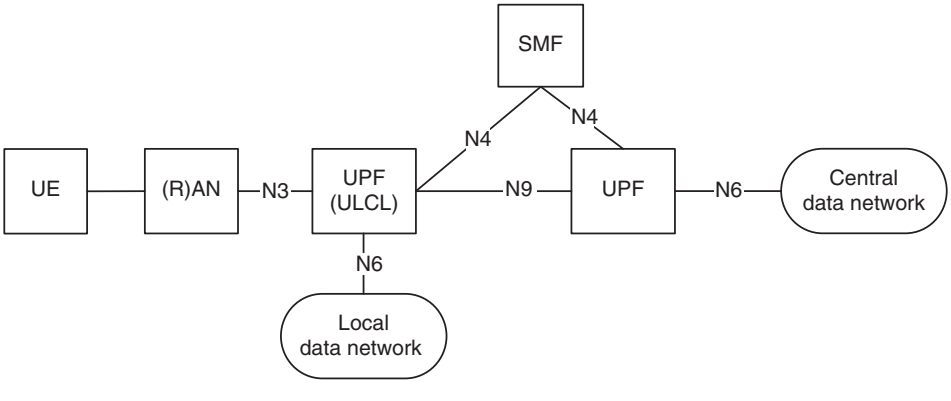



**Figure 3.2.7** Uplink classifier (ULCL) functionality in a UPF is used to divert some uplink traffic to a local data network and inject downlink traffic from the local data network to the UE.

#### 3.2.5.3 Multiple PDU Session-Based Access to Local Services

As depicted in Figure 3.2.8, a multi-homed PDU session refers to a configuration where the UE has been assigned different IPv6 prefixes by multiple PDU session anchors in the same PDU session. The SMF configures one UPF to act as a branching point, which forwards uplink traffic to the appropriate PDU session anchor (connected to the central or local data network) based on the source IPv6 prefix used by the UE. Furthermore, the branching point merges and forwards downlink traffic from the PDU session anchors to the UE.

As the UE changes location and the forwarding path to a previous PDU session anchor becomes inefficient, the SMF may add another UPF acting as a PDU session anchor and providing access to a closer local data network. The new PDU session anchor will then assign another IPv6 address to the UE. Previously used PDU session anchors for local data network instances that have become inefficient will be removed by the SMF.

In a multi-homed PDU session the IPv6 prefix that the UE uses when sending uplink traffic also determines the PDU session anchor that the uplink traffic is forwarded to by the branching point. Therefore, it is important for the network to be able to influence the UE's decision as to which source IPv6 prefix to use for which traffic.

The latter is achieved by enhancements to IPv6 Router Advertisements (Draves and Thaler 2005), which enable operators to provide routing information and preferences for source IPv6 selection to the UE.

Another approach to enable the UE to access both local and central services is to establish multiple PDU sessions: one PDU session terminating at a PDU session anchor located close to the RAN and one located deeper in the network for access to central services.

The challenge of this approach is how to ensure that the PDU session for local services is updated so that it always terminates at a PDU session anchor that is close to the UE's location. At the same time, the system needs to ensure that the PDU session for central services is not modified.

To address this challenge, three different Session and Service Continuity (SSC) modes have been defined, which determine whether a PDU session is to be relocated upon UE

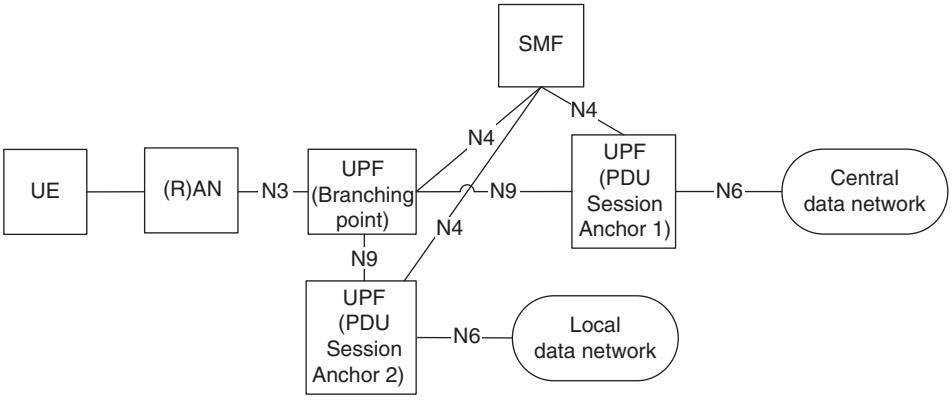



**Figure 3.2.8** A multi-homed IPv6 PDU session provides access to both a local and central data network in the same PDU session.

mobility and the level of session continuity that will be provided during PDU session relocation:

- *SSC mode 1*: a PDU session with SSC mode 1 will remain anchored on the same PDU session anchor throughout the lifetime of the PDU session. PDU sessions for central services would therefore typically use SSC mode 1.
- *SSC mode 2*: the goal of SSC mode 2 is to enable the network to re-anchor a PDU session, i.e. to re-establish a PDU session to a PDU session anchor that is located closer to the location of the UE. To achieve this, the network releases the PDU session and as part of this, requests the UE to establish a new PDU session to the same data network immediately. It is worth mentioning that this concept already existed in EPS where the MME can deactivate a PDN connection and request the UE to reactivate it again immediately to enable the MME to select a PGW that is located closer to the UE.
- *SSC mode 3*: this mode is like SSC mode 2 with the key difference that the PDU session to the new, more closely located PDU session anchor is established before the session to the previous PDU session anchor is released. The UE will receive a different IP address for the new PDU session. However, given that both PDU sessions remain active for some time, applications can be gracefully bound to the new IP address and, by this, get relocated from the old to the new PDU session. In summary, the benefit of SSC mode 3 (see Figure 3.2.9) is that the UE does not suffer a loss of connectivity while the user-plane path for access to local services is updated.

The SSC mode for a PDU session is determined by the SMF based on subscription information and based on the SSC mode requested by the UE (if provided). It is worth mentioning that the operator can influence the SSC mode to be requested by the UE for specific applications by providing SSC mode selection policy (SSCMSP) as part of the URSP (see Section 3.2.2.2).

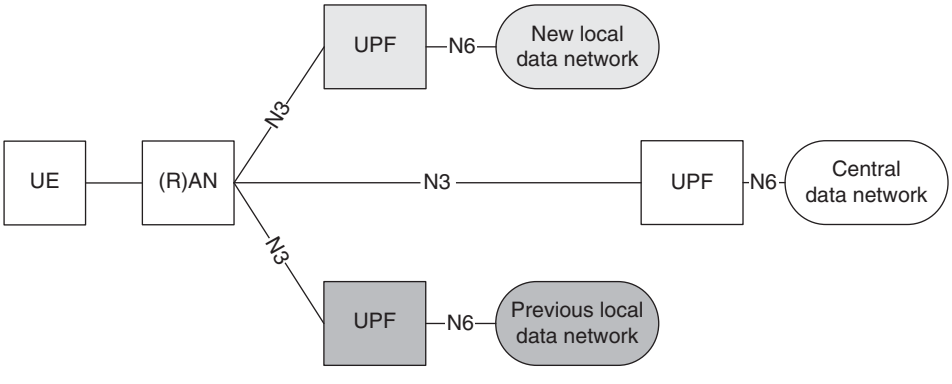



**Figure 3.2.9** SSC mode 3 – PDU Sessions to the previous and new local data network are temporarily active in parallel to allow for applications to bind to the new PDU session in order to avoid service interruption.

### 3.2.6 Network Slicing

The rise of vertical use-cases that go beyond traditional mobile broadband, i.e. the increasing interest in LPWA IoT, MCS/MPS as well as the emerging demand for ultra-reliable low latency communications (URLLC) services, emphasizes the need to support diverse core network functionalities and configurations in the same PLMN.

The key enabler for this is network slicing, which allows operators to

- deploy multiple independent and isolated sets of 5GC network functions in the same network;
- create multiple 5GC (and RAN) configurations, e.g. one configuration tailored to support many IoT devices such as sensors, which however only send small amounts of data, and yet another configuration for high-throughput mobile broadband customers;
- select the appropriate 5GC implementation and configuration dynamically for a given UE depending on subscription and applications running on the UE.

It is important to emphasize that the concept of network slicing as defined for 5GC is not fundamentally new. Earlier generations of 3GPP-based mobile networks supported a steadily increasing degree of flexibility for selecting core network functions and network configurations, for example, based on subscription, data network to establish a session to, etc.:

- *3GPP Release-97* introduced the General Packet Radio Service (GPRS), which enables transfer of packetized data via 2G and later 3G mobile radio (3GPP TS 03.60, 1997). GPRS allows for selection of the Gateway GPRS Support Node (GGSN), the equivalent of the PGW in EPS, based on Access Point Name (APN). The APN identifies the data network to connect to and typically gets signaled by the UE during packet data protocol (PDP) context establishment (session establishment). By configuring UEs with different APNs (e.g. "Internet" and "corporate"), this feature allowed for selection of different GGSNs for sessions toward different data networks.

  It is worth pointing out that the Serving GPRS Support Node (SGSN), the equivalent of the MME and SGW in EPS, is selected based on serving RAN node. In other words, GPRS does not support selection of different SGSNs for different groups of subscribers. The same limitation applies to EPS as defined in 3GPP Release-8, which selects MME and SGW only based on network topology and load.
- *3GPP Release-13* addressed this gap by adding the notion of *DECOR*. The key idea of DECOR (3GPP TS 23.401, 2019) was to select the serving MME not only based on serving RAN node and MME load but also based on a new subscription parameter, namely the subscribed UE usage type.

  MMEs that serve different UE usage types may also have different configurations for selection of other core network nodes (SGW and PGW), which enables operators for instance to use different SGWs and PGWs for the same APN depending on the subscribed UE usage type.

  In summary, DECOR enabled operators to deploy and select different Dedicated Core Networks (DCNs), consisting of all CN nodes (MME, SGW, PGW, PCRF, etc.), for different groups of subscribers, e.g. DCN 1 for massive IoT subscribers and DCN 2 for mobile broadband subscribers.

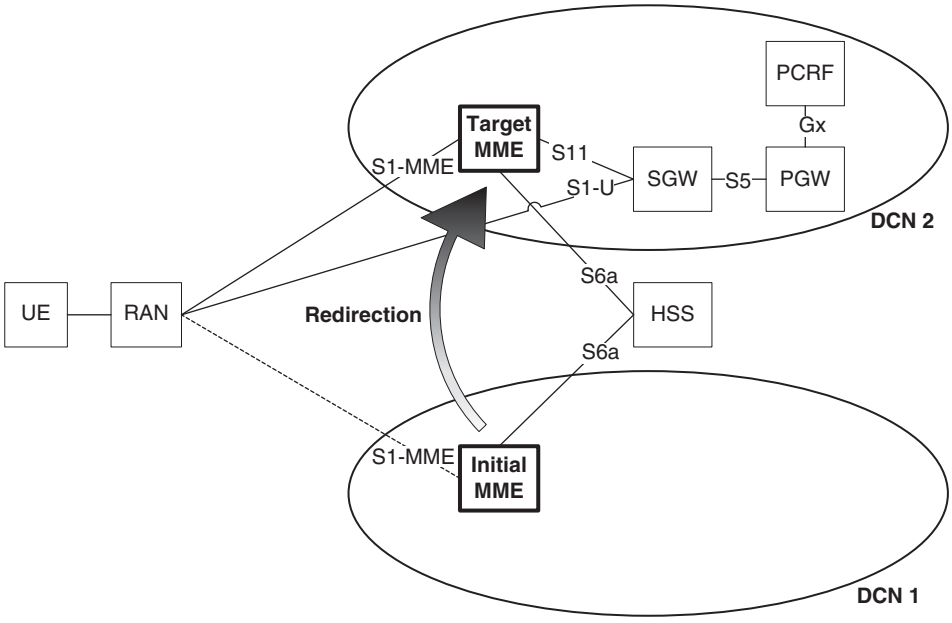



**Figure 3.2.10** Release-13 DECOR enables redirection to a target MME in the right DCN based on subscribed UE usage type received from the HSS.

The key drawback of Release-13 DECOR is that the RAN is not aware of the UE usage type.

When a UE attaches to EPS the RAN selects an initial MME, which then retrieves the UE's subscription. As depicted in Figure 3.2.10, if the initial MME selected by the RAN does not support the subscribed UE usage type, then the initial MME redirects the UE to a different MME that supports the subscribed UE usage type. In other words, the drawback of Release-13 DECOR is that the RAN cannot directly select the right MME; the right MME can only be selected by redirection.

Those redirections not only imply additional signaling load but also break isolation between different DCNs since all UEs initially attach to the same MMEs before they eventually get redirected to their correct DCN.

- To significantly reduce the need for redirections and to enhance the level of isolation, *Release-14* enhancements to DECOR (referred to as eDECOR and defined in 3GPP TS 23.401, 2019) enable the RAN to directly select the right MME (and thereby the right DCN). To achieve this, the UE provides a DCN-ID to the RAN via Radio Resource Control (RRC) protocol, which the RAN subsequently uses to select an MME that supports this DCN-ID (see Figure 3.2.11). The DCN-IDs to signal to the RAN in a given PLMN are provided to the UE by the Home Public Land Mobile Network (HPLMN).

As summarized above, various enablers to support multiple sets of core network functionalities and configurations in the same PLMN have already been specified across various 3GPP releases. Two aspects however have remained unchanged:

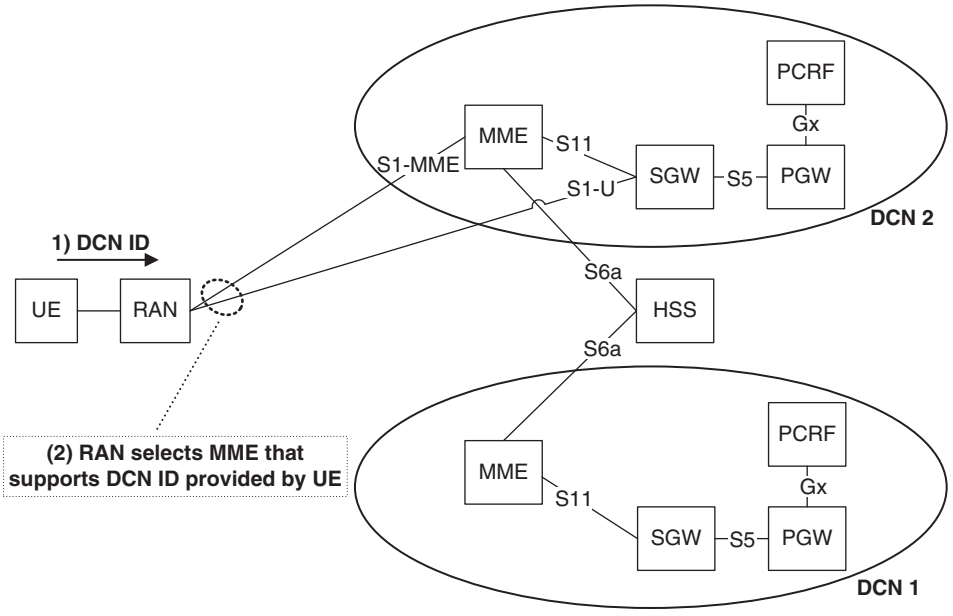



**Figure 3.2.11** Release-14 eDECOR reduces the need for redirections by enabling the RAN to directly select the right DCN based on information provided by the UE.

- *Single DCN per UE*: a UE can only connect to a single DCN at a time. In other words, even though it is possible to concurrently use different PGWs for different PDN connections of the same UE, all PDN connections should be part of the same DCN. In the EPS architecture this means practically that MME and SGW are shared across DCNs.
- *DCNs do not apply to RAN*: the RAN is not always aware of the DCN-ID so that RAN cannot differentiate UEs that belong to different DCNs.

5GS network slicing as defined in Release-15 (3GPP TS 23.501) addresses these shortcomings as follows:

- A *5GS network slice consists of both RAN and core network*, i.e. in contrast to DCNs in EPS 5GS network slices also cover the RAN.
- *Concurrent access to multiple network slices per UE*: a UE can be concurrently connected to multiple network slices with the only limitation that RAN and AMF are shared across those network slices.
  In contrast to EPS where the SGW was shared across all sessions of a UE, 5GC enables operators to deploy fully isolated sets of session-related network functions (SMF, UPF, PCF) for different PDU sessions of a UE as shown in Figure 3.2.12. To facilitate this, the UE provides assistance information (referred to as Single Network Slice Selection Assistance Information [S-NSSAI]) to the network during PDU session establishment. The network uses S-NSSAI as additional input for selection of session-related network functions (SMF, UPF, PCF) within the network slice as requested by the UE.
- *Support for RAN slicing*: the UE informs the RAN about the network slices the UE intends to access by sending Requested NSSAI. Like DCN-IDs in the Release-14 enhancements to

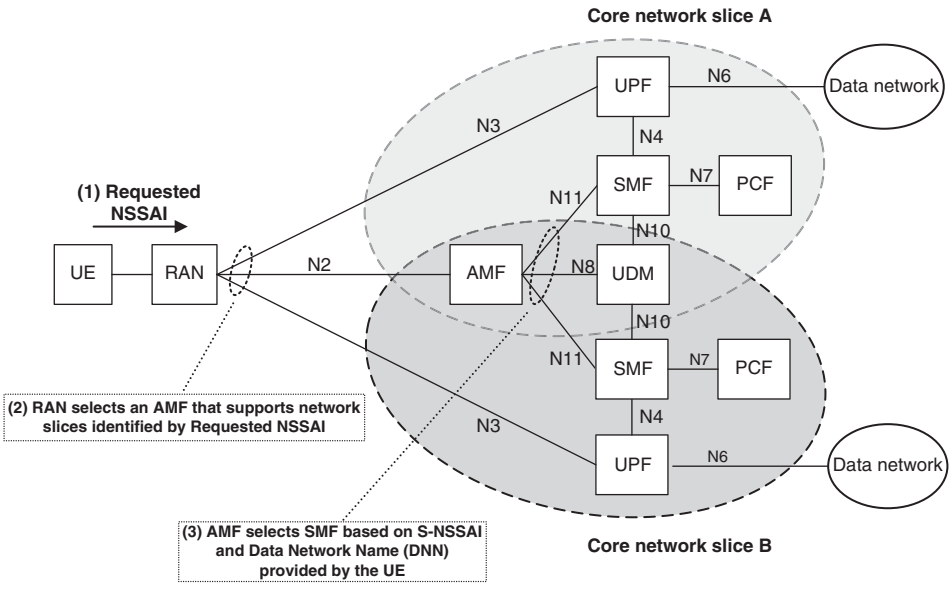



**Figure 3.2.12** Release-15 enables UEs to concurrently access multiple network slices with independent sets of session-related functions (SMF, UPF, PCF), AMF and UDM are shared across the network slices.

DECOR, Requested NSSAI enables RAN to select a serving AMF for the UE that supports the set of network slices requested by the UE, i.e. avoids unnecessary redirections. The core network also makes RAN aware of the network slice that a given PDU session of a UE belongs to by passing the S-NSSAI requested by the UE for a new PDU session to RAN. The latter enables RAN to perform resource isolation and differentiated handling of traffic of PDU sessions that belongs to different slices.

- *Network slice selection policies for the UE*: the fact that the UE can access multiple slices at the same time raises the question of how to determine which network slice to request for which PDU session/application. This is enabled by network slice selection policies (NSSPs) that the PCF provides to the UE based on operator policies. This mechanism provides operators with fine-grained control to ensure that specific applications running on the UE are served by the right network slice.

In summary, 5GS network slicing is an incremental next step that builds on top of concepts that were already introduced in earlier 3GPP system generation. 5GS network slicing increases deployment flexibility and isolation for session-related network functions, extends slicing into the RAN, and provides operators with new UE policies to influence the selection of network slices that applications running on the UE should be associated with.

## 3.2.7 Private Networks

### 3.2.7.1 Overview

Private networks address various use cases, for example, to enable 5G-based industrial automation for which hosting the entire network on site (e.g. in a factory) is key to

addressing data privacy needs and reliability concerns of many industries or to provide private cellular network coverage in rural or even offshore scenarios.

Tools to enable 5G-based private networks were added in different releases.

*Release-15 focused on authentication for private networks* and added support for the Extensible Authentication Protocol (EAP) (Aboba et al. 2004). EAP is a framework that enables private networks to use authentication methods other than the cellular network-specific 5G Authentication and Key Agreement (5G AKA) and SIM cards for storage of the related identifiers and credentials (3GPP TS 33.501, 2019). Based on the EAP framework, private networks can for instance use the EAP-TLS method to leverage certificates for mutual authentication (for an illustration of EAP-TLS for 5GS, see Annex B of 3GPP TS 33.501, 2019). Support of EAP-TLS is especially beneficial if 5G is added as another networking technology in scenarios where certificate-based authentication is already used, for example, EAP-TLS for device authentication in Ethernet networks deployed in offices or factories.

*Release-16 addresses unique network identification and access control for private networks*, also referred to as non-public networks (NPNs). Two NPN deployment models are supported:

- *Stand-alone non-public network (SNPN)*, i.e. scenarios where the entire network (RAN and core network) is deployed on-site (e.g. inside a factory).
- *Public-network-integrated non-public network (PNI-NPN)* refers to scenarios where the core network control-plane and typically subscription management (UDM) are shared between NPN and PLMN. RAN and core network user-plane may either be shared or are dedicated to the NPN.

#### 3.2.7.2 Stand-Alone Non-public Networks

The key challenge for SNPNs is that the existing identification scheme for cellular networks, i.e. the PLMN ID consisting of mobile country code (MCC) and mobile network code (MNC) has not been designed with a large number of non-public networks in mind. As a result, it is not feasible to assign a unique PLMN ID for each SNPN deployment.

Furthermore, while the International Telecommunication Union (ITU) has allocated MCC 999 for use by private networks (ITU 2018), MNCs under this MCC can be used by anyone without further consultation with ITU. In other words, PLMN IDs using MCC 999 also fall short of providing a unique network ID for SNPNs.

Non-unique network IDs can lead to SNPN UEs not receiving service, for example, if a UE serving a robot arm tries to connect to an SNPN B that uses the same network ID as its own SNPN A. SNPN B would reject the registration attempt, upon which the UE would blacklist the network ID and remain disconnected.

It is worth noting that connection attempts by unauthorized UEs (which are rejected by the network) consume radio resources and may therefore also degrade the performance of SNPNs. The latter may be an issue in the case of time-sensitive non-public network services, for example, if the SNPN is used for industrial automation.

To enable SNPN operators to configure a unique network ID, Release-16 introduced the network ID (NID), which is broadcast by NG-RAN cells in addition to the PLMN ID as depicted in Figure 3.2.13.

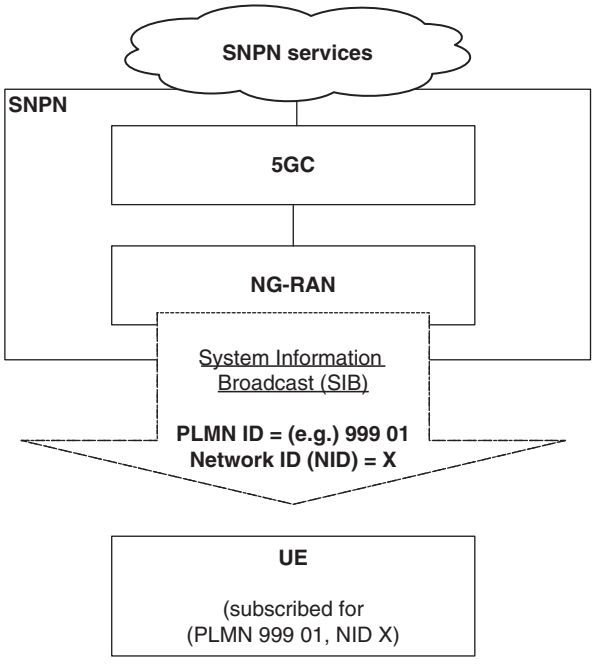



**Figure 3.2.13** Stand-alone non-public network (SNPN).

The NID supports two assignment models:

- Locally managed NIDs can be chosen by SNPNs at deployment time and may therefore not be unique.
- Universally managed NIDs refer to scenarios where the NID is assigned by a legal entity, e.g. by regulators.

The key idea is that Release-16 UEs in SNPN access mode only select networks broadcasting both PLMN ID and NID. In other words, Release-16 UEs in SNPN access mode ignore PLMNs and only register with SNPNs. Legacy and non-supporting UEs are prevented from accessing SNPN cells.

In contrast to PLMNs there is no support for emergency calls in limited service state in SNPNs. Furthermore, there is no notion of home SNPNs and therefore also no support for roaming between SNPNs.

#### 3.2.7.3 Public-Network-Integrated Non-public Network

The challenge for PNI-NPNs applies to deployment scenarios where dedicated NG-RAN cells serve the NPN, for example, dedicated small cells in a factory. Given that the PLMN and the NPN share the PLMN ID of the PLMN, PLMN UEs would also be able to access the small cells that are supposed to be dedicated to the factory. In other words, the question is how to ensure that only NPN UEs can access the dedicated NPN cells.

The key idea to limit cell access to a group of devices is the notion of Closed Access Groups (CAGs). In addition to the PLMN ID of the public network, the dedicated cells broadcast a CAG ID. The CAG IDs are managed by the PLMN operator as a result of which they can be assumed to be unique in combination with the PLMN ID.

Both the UE and the UE's subscription in the network are configured with the list of allowed CAG IDs that a UE is entitled to access. Based on this, UEs only select CAG cells that broadcast a CAG ID contained in the UE's allowed CAG list; the network double checks access attempts against the UE's subscription. Similarly RAN prevents connected mode mobility to non-allowed cells based on enhanced mobility restrictions from the core network.

As illustrated in Figure 3.2.14, three types of UE behavior can be distinguished for CAGs:

- Pre-Release-16 UE or UEs not supporting CAGs ignore CAG cells (UE 1).
- UEs can be configured to access both specific CAG cells and non-CAG cells, i.e. public cells (UE 2).
- UEs can be configured to only access specific CAG cells but no public cells (UE 3), e.g. to prevent machines in a factory from connecting to macro cells.

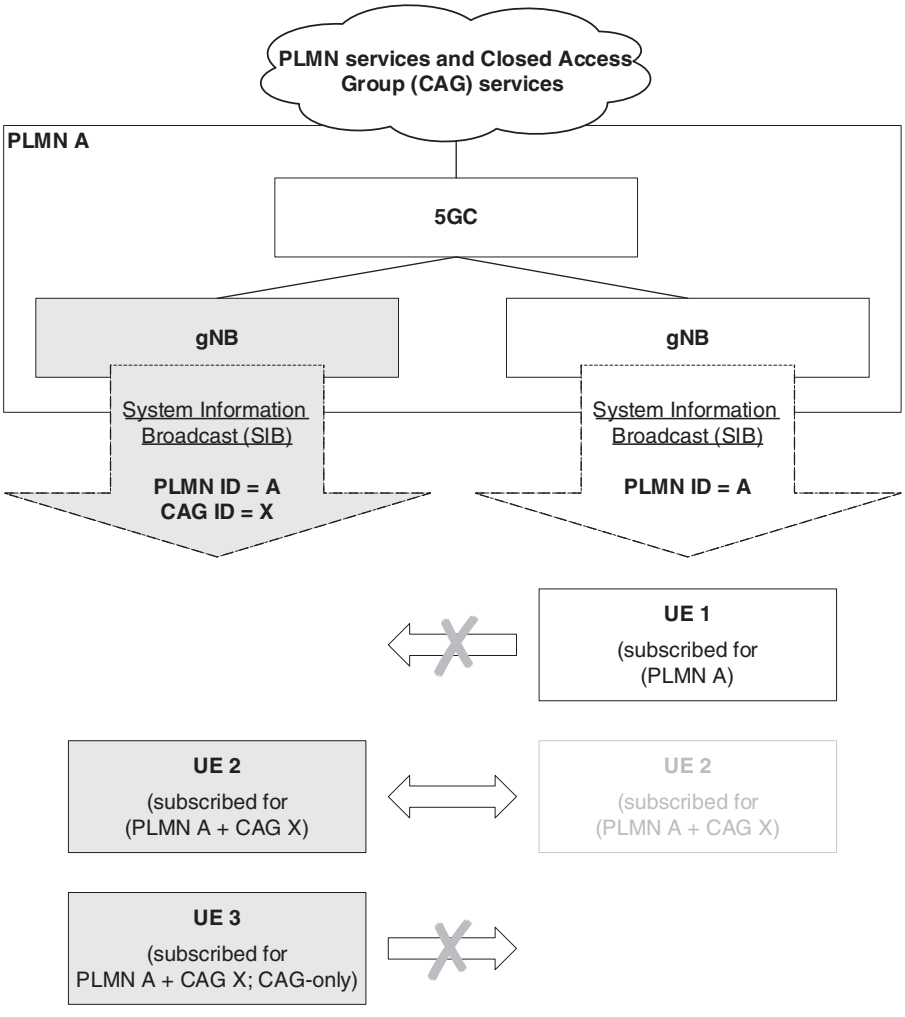



**Figure 3.2.14** Closed Access Group (CAG), an enabler for public-network-integrated non-public networks.

It is worth pointing out that CAGs are very similar to Closed Subscriber Groups (CSGs) in EPS (3GPP TS 23.401, 2019), which were introduced in Release-8 and Release-9 to enable femtocell deployments. The key difference is that the CAG concept additionally allows for restricting UEs to CAG cells only (UE3 in Figure 3.2.14).

## References

3GPP Technical Specification (TS) 03.60 v2.0.0 (1997). General Packet Radio Service (GPRS); Service description; Stage 2. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 23.214 v15.5.0(2018). Architecture enhancements for control and user plane separation of EPC nodes. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 23.401 v16.3.0 (2019).General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 23.402 v16.0.0 (2019). Architecture enhancements for non-3GPP accesses. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 23.501 v16.1.0 (2019).System architecture for the 5G System (5GS). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 23.502 v16.1.1 (2019).Procedures for the 5G System (5GS). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 23.682 v16.3.0 (2019).Architecture enhancements to facilitate communications with packet data networks and applications. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 29.274 v16.0.0. (2019).3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunneling Protocol for Control plane (GTPv2-C); Stage 3. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 29.303 v15.5.0. (2019).5G System; Network function repository services; Stage 3. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 29.500 v16.0.0 (2019).5G System; Technical Realization of Service Based Architecture; Stage 3. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 29.510 v16.0.0. (2019).5G System; Network function repository services; Stage 3. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 33.501 v15.5.0. (2019).Security architecture and procedures for 5G System. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and Levkowetz, H. (2004). Extensible Authentication Protocol (EAP), IETF RFC 3748. Available at: https://tools.ietf.org/html/rfc3748 (accessed August 28, 2019).

Belshe, M., Peon, R., and Peon, M. (2015). Hypertext Transfer Protocol Version 2 (HTTP/2), IETF RFC 7540. Available at: https://tools.ietf.org/html/rfc7540 (accessed August 28, 2019).

Bray, T. (2017). The JavaScript Object Notation (JSON) Data Interchange Format, IETF RFC 8259. Available at: https://tools.ietf.org/html/rfc8259 (accessed August 28, 2019).

Draves, R. and Thaler, D. (2005). Default router preferences and more-specific routes. IETF RFC 4191. Available at: https://tools.ietf.org/html/rfc4191 (accessed August 28, 2019).

Fajardo, V., Arkko, J., Loughney, J., and Zorn, G. (2012). Diameter Base Protocol. IETF RFC 6733. Available at: https://tools.ietf.org/html/rfc6733 (accessed August 28, 2019).

International Telecommunication Union (ITU), Standardization Bureau (2018). Operational Bulletin No. 1156. Available at: http://handle.itu.int/11.1002/pub/810cad63-en (accessed August 28, 2019).

## 3.3 NG Radio Access Network

*Sasha Sirotkin*
*Intel Corporation, Israel*

### 3.3.1 Introduction

In the present section we provide an overview of the RAN part of the 5G System.

In order to support 5GC and NR, 3GPP have developed a new RAN, referred to as NG-RAN. Conceptually, it resides between a UE and a 5GC, as shown in Figure 3.3.1.[5]

NG-RAN is a collection of base stations, or gNBs and ng-eNBs, interconnected by the Xn interface and connected to the 5GC via the NG[6] interface. NG-RAN terminates the Uu air interface toward a UE and therefore supports the NR physical layer (PHY) and protocol stacks, described in Sections 3.5 and 3.4, respectively.

Both ng-eNBs and gNBs are part of NG-RAN, as they terminate NG-RAN network interfaces (e.g. Xn and NG[7]), however they provide different air interface accesses – Long-Term Evolution (LTE) and NR, respectively. This is different compared to all other technologies previously defined by 3GPP (e.g. LTE), where a RAN only supports one air interface and interworking with other technologies is only supported using handovers via the core network or specific dual-connectivity technologies (e.g. LTE–Wi-Fi interworking using LTE-WLAN Radio Level Integration with IPsec tunnel [LWIP]). 5G supports not only

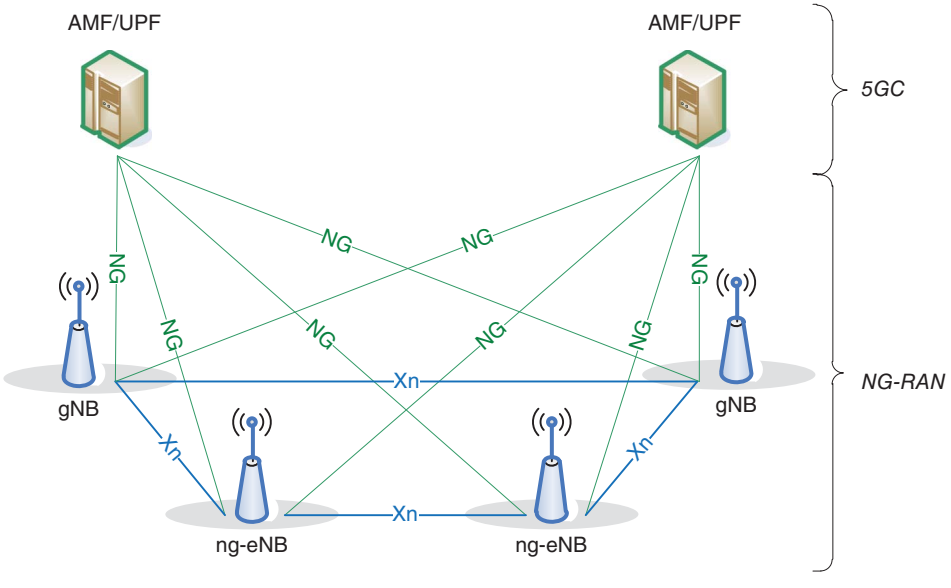



**Figure 3.3.1** Overall 5G System Architecture. (Source: Reproduced by permission of © 3GPP).

5 Only parts of 5GC and only monolithic (i.e. no-split) gNB are shown in the figure.
6 In some specifications, the NG interface is referred to as N2 and N3 (interfaces or reference points), corresponding to NG-C and NG-U, respectively.
7 Referred to as N2 and N3 in Chapter 3 and core network specifications.

handovers to LTE, but also tight NR-LTE interworking via dual connectivity (E-UTRA-NR Dual Connectivity [EN-DC]).

Compared with Evolved Universal Terrestrial Radio Access Network (E-UTRAN), NG-RAN is somewhat more complex not only because it supports both NR and LTE access technologies, but also because it can be deployed in various split and non-split architecture variants, as explained in Chapter 4.

As the NG-RAN and especially new architecture options introduced in 5G are the primary focus of this book, in the present section we only describe high-level NG-RAN functions and their differences compared with E-UTRAN. When discussing the NG-RAN interfaces, we make the (somewhat artificial) distinction between NG-RAN internal and external interfaces, with the focus of the present section being the external ones. For the purpose of this discussion we currently assume that a gNB (or en-gNB) is a single monolithic network node, terminating network interfaces toward the core network and other NG-RAN nodes and implementing all of the required functionality: from network interfaces, to air interface protocol stack, to physical layer, RF and antennas.

In 3GPP specifications, NG-RAN network nodes are defined in terms of radio and network interfaces they support, specifically:

- Functionality of NR (in the case of gNB) and LTE (in the case of ng-eNB) air interfaces toward a UE;
- Functionality of NG interface toward the core network (i.e. 5GC);
- Functionality of Xn interface toward other NG-RAN network nodes.

The functionality required to support the radio and network interfaces mentioned above is somewhat loosely defined to allow sufficient freedom of implementation and generally include:

- UE admission control over the radio interface;
- UE radio interface connection setup and release;
- Radio resource management, including UE radio bearer control, and uplink and downlink scheduling for a UE;
- UE mobility control in connected state (i.e. handover) and in inactive state;
- UE measurements, including measurement configuration and processing of UE measurement reports;
- Routing of user-plane and control-plane packets toward UPF and AMF, respectively;
- UE QoS flow management and mapping to radio bearers;
- Slicing;
- Tight interworking between NR and LTE, including multiple dual connectivity options (between these technologies);
- RAN sharing between multiple operators.

As mentioned above, some of the NG-RAN functions, for example the air interface or the NG interface protocol stacks, are well defined in the corresponding specifications and are described in the present book; some other functions (e.g. scheduling, resource isolation for slicing, radio resource management, and others) are intentionally left unspecified, in order to allow for implementation flexibility and differentiation.

The following key enhancements have been introduced in NG-RAN compared with E-UTRAN.

**Multiple air interfaces:** in contrast to E-UTRAN, NG-RAN supports two air interfaces; gNB supports the NR air interface and ng-eNB supports the LTE air interface. Both gNB and ng-eNB are considered NG-RAN network nodes.

**Split architectures:** in addition to the "flat" architecture with a monolithic gNB, which is similar to E-UTRAN, NG-RAN supports an architecture with a gNB split into two (high-level and low-level referred to CU and DU, respectively) logical nodes, with a standardized F1 interface between them. Furthermore, the central unit of a gNB, which implements the higher layers of the split, can be deployed as separate control-plane and user-plane logical network nodes with a standardized E1 interface between them. Moreover, in addition to the high-level split defined in 3GPP, the Open Radio Access Network (O-RAN) Alliance defined the low-level split, in which the gNB-DU, which implements the lower layers of the split, can be further split into two network nodes (this last split, however, is not reflected in the 3GPP architecture model) This is described in detail in Chapter 4.

**Multi-radio dual connectivity (MR-DC):** NG-RAN supports several options of multi-connectivity, in which a single UE may be connected to two different network nodes, one providing NR access and the other one providing either E-UTRA or NR access. Either node can act as the master node (MN), while the other one acts as the secondary node (SN). This is the generalization of the dual connectivity architecture introduced in E-UTRAN, with the main difference being that MN and SN can use different access technologies. This is described in detail in Section 4.3.

**Slicing:** slicing allows MNOs to categorize customers into different tenant types, each having different service requirements, as reflected in a Service Level Agreement (SLA). Slicing allows, for example, an MNO to lease parts of resources of their network to a "vertical" (e.g. a factory that may want to use that wireless technology without investing in their own deployment). A network slice always consists of a RAN part and a core network part. This is somewhat similar to DECOR and eDECOR available in E-UTRAN; however, it has better flexibility as both 5GC and NG-RAN have been designed from the beginning with slicing in mind.

**Integrated access and backhaul (IAB):** starting from Release-16, NG-RAN supports IAB functionality, which is roughly equivalent to the relaying support in E-UTRAN. Unlike E-UTRAN, IAB in NG-RAN supports multi-hop backhauling with topology adaptation and redundant links for better performance and resilience to links failures. IAB supports both in-band and out-of-band relaying, among other features. This is described in detail in Section 5.2.

**Non-terrestrial networks (NTNs):** starting from Release-17, NG-RAN will also support NTNs (i.e. satellite). This is described in detail in Section 5.3.

**Virtualization:** Even though 3GPP specifications do not explicitly define virtualized NG-RAN, nevertheless it has been designed from the beginning with virtualization in mind. Certain provisions have been made in the definition of all NG-RAN interfaces to allow deployments in virtualized environments. Additionally, for the scenario with split gNB, the CU hosting the high-level protocols can also be virtualized. Furthermore, the implementation for control–user-plane separation in both NG-RAN and 5GC makes it particularly well suited to be used with software defined networks (SDNs).

### 3.3.2 Network Protocol Stacks

Before discussing the functionalities of the NG-RAN interfaces, we first describe the protocol stacks used by these, specifically:

- Control-plane protocol stack (used on interfaces NG, Xn, F1, and E1);
- User-plane protocol stack (used on interfaces NG, Xn, and F1).

#### 3.3.2.1 Control-Plane Protocol Stack

All NG-RAN control-plane interfaces (NG, Xn, F1, and E1) use the same control-plane protocol stack (illustrated in Figure 3.3.2) with Stream Control Transmission Protocol (SCTP) (IETF RFC 4960) on top of IP.

SCTP is chosen for increased reliability of control-plane messages, which are generally considered more critical than user-plane messages. Even though network interfaces typically use wired transport network, which is assumed to be more reliable compared with the air interface, the transport network may still be prone to, for example, congestion resulting in packet loss, therefore the usage of a reliable transport is important. SCTP also supports in-sequence delivery of control-plane packets and multi-homing, among other features.

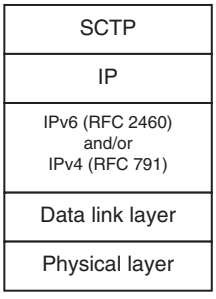



**Figure 3.3.2** Control-plane protocol stack.

It should be noted that the protocol stack in Figure 3.3.2 does not specify the transport network, other than saying that it should provide IP connectivity. This is the general approach taken in most 3GPP network interface specifications – they are abstract of the transport layer. On the one hand, this abstraction model simplifies standardization and development; however, on the other hand, it makes it easy to overlook transport-related issues when discussing NG-RAN network interfaces. This problem is discussed further in the book in Section 6.6, which is dedicated to the transport network.

In contrast to LTE, which also uses the same protocol stack for control-plane network interfaces, NG-RAN supports dynamic addition and removal of multiple SCTP endpoints, which is useful for deployments in virtualized environments.

Every control-plane network interface has its own Application Protocol on top of SCTP, which is described in the respective section for each interface further down in the book.

#### 3.3.2.2 User-Plane Protocol Stack

All NG-RAN network interfaces (with the exception of E1, described in Section 4.4) have a user plane, which uses the same protocol stack, shown in Figure 3.3.3.

The user-plane protocol stack used on NG, Xn, and F1 interfaces uses IP transport and GPRS Tunneling Protocol User Plane (GTP-U) on top of User Datagram Protocol (UDP)/IP to carry the user-plane PDUs between the NG-RAN node and the UPF, or between NG-RAN network nodes.

GTP-U uses the notion of bearer identified by source GTP-U Tunnel Endpoint Identifier (TEID), destination GTP-U TEID, source IP address, and destination IP address. GTP-U bearers can be mapped to NG-RAN bearers or to PDU sessions, depending on the interface it is used on.

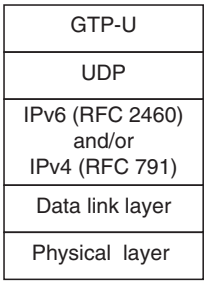



**Figure 3.3.3** NG-U protocol stack. (Source: Reproduced by permission of © 3GPP).

GTP-U is used extensively in the core network; however, its usage in RAN requires some additional functionality. For example, since RAN network interfaces may be considered somewhat less reliable (compared with the CN interfaces) and RAN network nodes may have smaller buffers, certain GTP-U enhancements specific to the use of GTP-U in RAN have been made, for example, flow control. These enhancements (specific to NG-RAN interfaces), use a GTP-U container, defined in 3GPP TS 29.281, whereas the content of the container is defined in a user-plane protocol specification of a respective NG-RAN interface. This is because even within NG-RAN different interfaces require somewhat different functionality.

#### 3.3.2.3 Standards

Traditionally, every NG-RAN (and E-UTRAN) network interface has been defined by a set of five to six specifications. This is, of course, somewhat redundant, as typically some of these specs overlap, partially or even fully. For the most part, what differs between NG-RAN interfaces is the Application Protocol, but for example not the transport.

Some effort has been made to improve this situation in 5G; however, it was limited to the definition of a single common user-plane protocol for Xn and F1 (3GPP TS 38.425), as it was considered that the functionality required is very similar. The NG interface still has a separate user-plane specification (3GPP TS 38.415).

### 3.3.3 NG Interface

The NG interface connects NG-RAN to 5GC. It is further divided into NG control plane (NG-C[8]), connecting NG-RAN to AMF and NG user plane (NG-U[9]), connecting NG-RAN to UPF.

Some key NG-C functions are:

- Paging, i.e. sending paging messages from 5GC to NG-RAN nodes in the UE's paging area;
- UE context management, to allow AMF to establish, modify, and release UE context in NG-RAN;
- UE mobility management, i.e. intra-system (i.e. within 5GS) and inter-system (i.e. between 5GS and EPS) handovers;
- PDU session management, to allow SMF to establish, modify, and release PDU sessions for a UE;
- tTrace, allowing AMF to control trace sessions;
- AMF load balancing, allowing AMF to indicate its capacity to potentially trigger AMF load balancing within a pool area, and overload control, allowing AMF to control the load NG-RAN generates;
- NR Positioning Protocol A (NRPPa) signaling transport, for positioning support.

NG-U connects an NG-RAN network node to a UPF. It supports non-guaranteed delivery of PDU session user-plane PDUs between the NG-RAN node and the UPF.

---

8 Note that the NG-C interface is referred to as N2 reference point in some stage-2 specifications (3GPP TS 23.501).

9 Note that the NG-U interface is referred to as N3 reference point in some stage-2 specifications (3GPP TS 23.501).

#### 3.3.3.1 NG-C Interface

As mentioned above, the NG-C interface (3GPP TS 38.413) connects a NG-RAN to an AMF. A gNB may be connected to multiple AMFs in the so called NG-Flex configuration, where a gNB is connected to all AMFs within an AMF region. A gNB selects an AMF for a UE based on configuration, service requirements, slicing information, and other parameters. NG-Flex allows AMF load balancing across an AMF pool.

NG Application Protocol (NG-AP) (3GPP TS 38.413) is used on the NG-C interface on top of SCTP, which is conceptually similar to the S1 Application Protocol (S1-AP) (3GPP TS 36.413) used in LTE.

One new distinct feature of NG-C, not present in LTE, is the capability of both NG-RAN and AMF to use multiple SCTP associations through multiple SCTP endpoints on both ends. This functionality facilitates the deployment of 5GC and NG-RAN in virtualized environments, where new computational resources can be added or removed "on the fly." As a computational (or network) resource may have a separate IP address, this new functionality allows adding these resources in a seamless manner, which has no impact on a UE.

Furthermore, in contrast to LTE, the NG-C interface also supports non-3GPP access technologies, e.g. WLAN (IEEE 802.11). When non-3GPP access is used, a new network node referred to as N3IWF terminates the NG interfaces toward 5GC, making the usage of non-3GPP access largely transparent for the 5GC. Of course, not all 3GPP functionalities are supported by all access technologies. For example, there is no notion of paging in WLAN. Therefore, 3GPP have specified (3GPP TS 29.413) which NG-AP procedures are supported for non-3GPP access and which information elements (IEs) of the procedures that are used are not applicable for non-3GPP access. The initial aspiration of the 3GPP work was to make 5GC "access-agnostic," which has been largely fulfilled; however, in some specific cases the 5GC has to be aware whether it is a 3GPP gNB or a non-3GPP N3IWF that terminates the NG interface and therefore it would probably be more correct to call this functionality a "common core," rather than "access-agnostic core."

NG-AP procedures can be categorized as follows:

- Interface management procedures
- UE context management procedures
- UE session management procedures
- UE mobility management procedures
- Paging procedures
- PDU session management procedures
- Transport of NAS messages procedures
- Others (trace, location reporting, UE radio capability management, etc.).

In the present chapter we describe only the subset of the most commonly used procedures. For a detailed description of all the procedures, please refer to 3GPP TS 38.413.

NG-AP protocol differentiates between UE-associated and non-UE-associated signaling. For the former, a logical association for that UE must be established between the NG-RAN node and the AMF, which generally involves the allocation of temporary UE identifiers at both sides, which are used to identify that UE association.

#### 3.3.3.1.1 NG Interface Management

Interface management procedures are used to establish the NG interface, to tear it down, to reset it (when and if needed), and to allow gNB and AMF to exchange and update configuration information. The relevant NG-AP procedures are:

- NG Setup
- RAN/AMF Configuration Update
- AMF Status Indication
- Overload Start/Stop
- Reset and Error Indication.

NG Setup is the first procedure initiated by a gNB after the Transport Network Layer (TNL) association with an AMF has become operational. It carries the list of Tracking Areas (TAs) supported by a gNB, a list of PLMNs within each tracking area, and some other information. If the procedure is successful, the AMF replies with NG Setup Response, carrying the list of Globally Unique AMF IDs (GUAMIs), the list of PLMNs supported by the AMF, and some additional information. If some of the information conveyed during the NG Setup procedure changes in either gNB or AMF, each node can notify the other about such changes using RAN Configuration Update and AMF Configuration Update procedures, respectively. Those procedures can also be used to add or remove additional SCTP associations.

If an AMF experiences overload, it can request the NG-RAN to reduce signaling traffic UEs may generate toward it. This is performed using the Overload Start procedure, which also carries the information about which traffic an AMF is requesting to reject. For example, an AMF may request to reject all RRC connection requests or allow only emergency services. The Overload Stop procedure is used to indicate that the AMF can now accept all signaling traffic.

Interface management messages use non-UE-associated signaling.

#### 3.3.3.1.2 NAS Transport and UE Context Management

When a UE connects to the network, for example, during the registration procedure, a gNB selects an AMF for the UE and establishes a UE association between the two network nodes. This is performed using the Initial UE Message (which is part of the NAS transport NG-AP messages category) and Initial UE Context Setup Request/Response (which are part of the UE context management NG-AP messages category).

NAS transport messages are:

- Initial UE message
- Downlink/Uplink NAS transport
- Others (NAS Non-Delivery Indication, Reroute NAS request).

UE context management messages are:

- Initial Context Setup Request/Response/Failure messages
- UE Context Release Request/Command/Complete messages
- UE Context Modification Request/Response/Failure messages
- RRC Inactive transition report message.

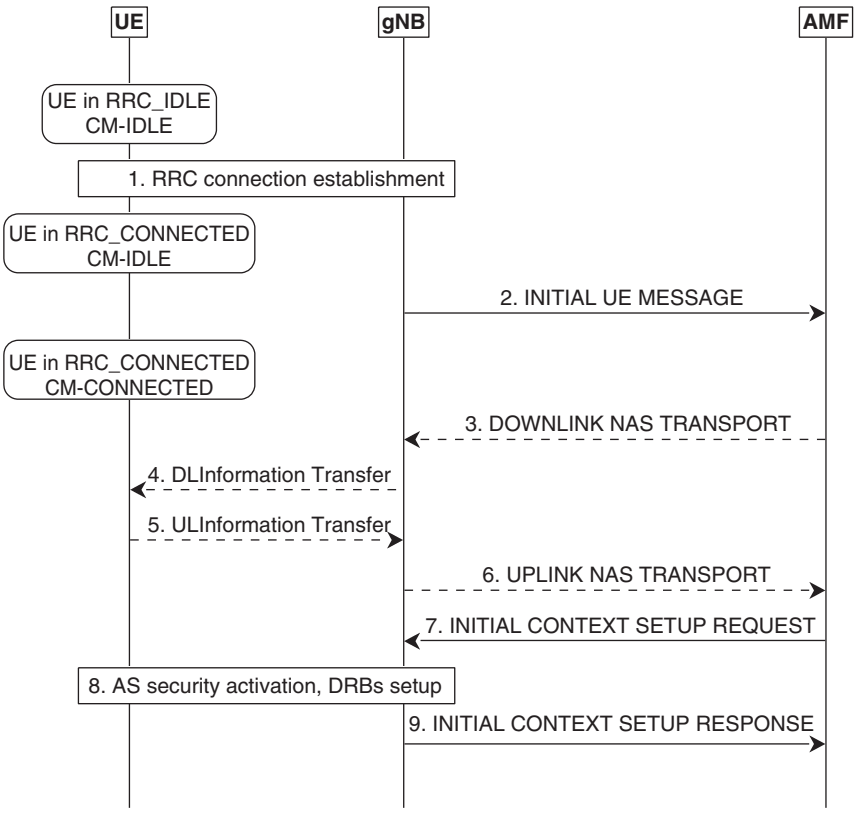



**Figure 3.3.4** IDLE to CONNECTED state transition. (Source: Reproduced by permission of © 3GPP).

Downlink and Uplink NAS transport procedures are rather self-explanatory – they are used to transfer NAS PDUs to/from a UE, respectively. Additionally, certain information can be piggybacked to them. For example, in the uplink, the NAS transport procedure can carry the UE location information. In the downlink, the message can also carry the mobility restriction list, UE aggregate maximum bit rate, and others, in case this information needs to be updated.

The usage of some of these procedures can be illustrated by the UE-triggered service request, which is the procedure performed when an idle mode UE has signaling or uplink data to send. The procedure is shown in Figure 3.3.4 in a simplified form, where interactions inside the 5GC are hidden. As a result of this procedure, a UE transitions from RRC_IDLE to RRC_CONNECTED RAN state (and from CM-IDLE to CM-CONNECTED 5GC state), establishes the signaling connection with an AMF, and may send messages.

0. A UE is in RRC_IDLE (from NG-RAN perspective) and in CM-IDLE (from 5GC perspective).
1. If the UE decides to send uplink data, it establishes an RRC connection. When an RRC connection is established, the UE moves from RRC_IDLE to RRC_CONNECTED.
2. In order to notify the core network and to allow e.g. establishment of PDU sessions for the UE, the gNB sends NG-AP Initial UE Message to the AMF, which carries the Service

Request NAS message received from the UE and other information, such as selected PLMN ID, user location information, etc./4./5./6. Additional NAS messages may be exchanged between UE and AMF (not shown for brevity).
7. The AMF prepares the UE context data and sends it to the gNB in the Initial Context Setup Request. The Initial Context Setup Request typically carries a request to establish one or more PDU sessions, security context, UE radio and security capabilities, mobility restriction list, allowed NSSAI, etc.
8. Access stratum (AS) security between UE and gNB is activated and Data Radio Bearers (DRBs) are set up (details are not shown for brevity).
9. The gNB informs the AMF that the procedure is completed using the Initial Context Setup Response. It typically carries the information (e.g. transport tunnels) of the established PDU sessions and the list of PDU sessions that failed to be established.

NAS transport and UE context management messages use UE-associated signaling, like most other NG-AP messages (except for those used for interface management). This means that there is a UE association in a gNB and an AMF, identified by a pair of temporary identifiers: AMF UE NG Application Protocol (NGAP) ID and RAN UE NGAP ID. These identifiers are present in every UE-associated message, with the exception of the Initial UE Message, which is used to initiate the UE association and carries only the RAN UE NGAP ID. When a new UE association needs to be created, a gNB allocates a new RAN UE NGAP ID and sends it to an AMF in the Initial UE Message. The AMF, in its turn, allocates a new AMF UE NGAP ID and sends it back (together with the received RAN UE NGAP ID) in the Initial Context Setup Request message. After that, the UE association in gNB and AMF is established.

#### *3.3.3.1.3 UE Mobility*

Mobility procedures are used primarily for connected mode UE mobility (i.e. handover). The messages are:

- Handover preparation (between source NG-RAN node and AMF)
  - Handover Required
  - Handover Command
  - Handover Preparation Failure
- Handover resource allocation (between target NG-RAN node and AMF)
  - Handover Request
  - Handover Request Acknowledge
  - Handover Failure
- Handover notification (from target NG-RAN node to AMF)
- Handover cancelation (between source NG-RAN node and AMF)
  - Handover Cancel
  - Handover Cancel Acknowledge
- Other
  - Path Switch Request
  - Uplink/Downlink RAN Status Transfer.

The usage of some of these procedures can be illustrated by the handover example. The procedure is shown in Figure 3.3.5 in a simplified form, where interactions inside the 5GC are hidden.

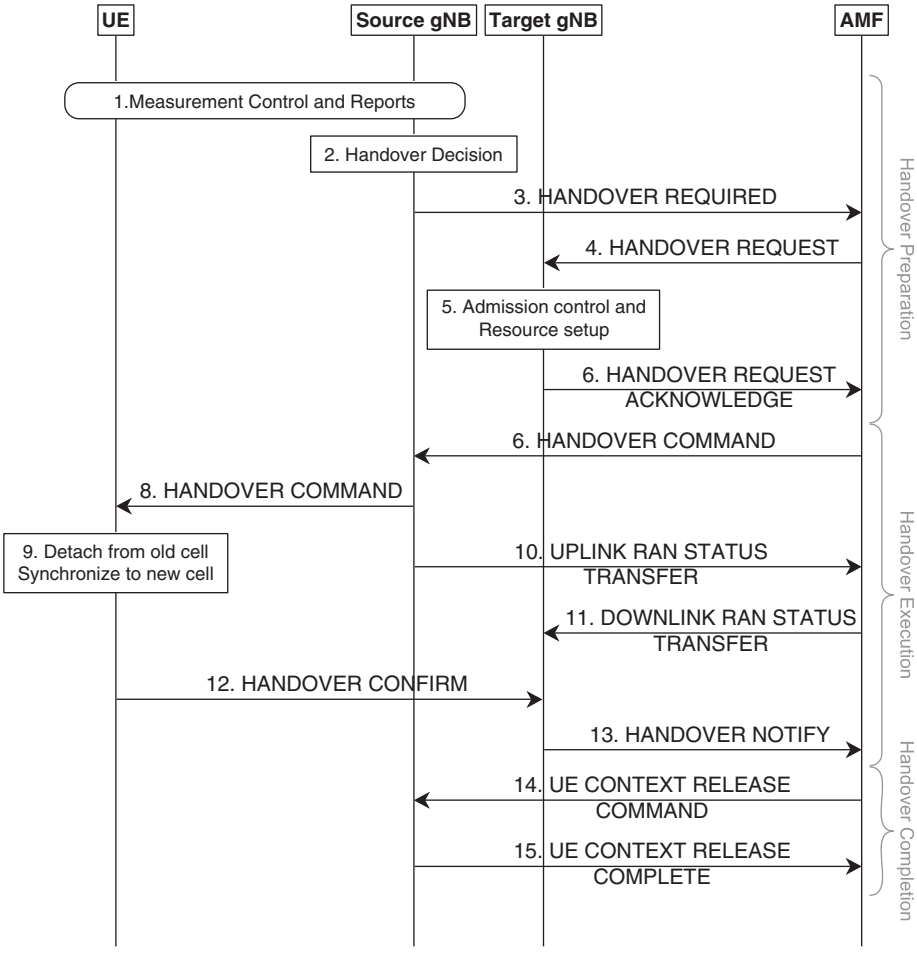



**Figure 3.3.5** NG handover. (Source: Reproduced by permission of © 3GPP).

Generally, a handover procedure is divided into two steps: handover preparation and handover execution.

1. The source gNB configures the UE measurement procedures and the UE reports the measurements according to the configuration.
2. Based on measurement report and potentially other information (e.g. load), the source gNB decides to hand over the UE.
3. The source gNB sends the Handover Required message to the AMF. The message carries a source-to-target transparent RRC container with necessary information to prepare the handover at the target side, the list of established PDU sessions with associated information, and optionally the direct forwarding path availability indication.
4. The AMF sends the Handover Request message to the target gNB. The message carries the source-to-target transparent RRC container and the list of PDU sessions, received from the source gNB, along with various other parameters, such as UE Aggregate Maximum Bitrate, UE Security Capabilities, Allowed NSSAI, and others.

5. The target gNB allocates the resources for the UE.
6. If the target gNB decides to admit the handover, it sends the Handover Request Acknowledge message to the AMF. The message carries the lists of admitted and not admitted PDU sessions, along with the target-to-source transparent RRC container to be sent to the UE as an RRC message to perform the handover. This completes the handover preparation stage.
7. The AMF sends the Handover Command to the source gNB. The message carries the list of PDU sessions to handover, the list of PDU sessions to release, the target-to-source RRC transparent container received in step 6, and some other information.
8. The source gNB triggers the handover by sending the Handover Command to the UE containing the information required to access the target cell.
9. The UE synchronizes to the target cell.
10. If Packet Data Convergence Protocol (PDCP) status preservation is required, the source gNB sends the Uplink RAN Status Transfer message to the AMF. The message carries the list of DRBs with their associated information about missing PDCP sequence numbers.
11. The AMF sends the information received in step 10 to the target gNB in the Downlink RAN Status Transfer message.
12. After the UE has successfully synchronized to the target cell, it sends the Handover Confirm message to the target gNB.
13. The target gNB sends the Handover Notify message to inform the AMF that the UE has been identified in the target cell and the handover has been completed.
14. The AMF sends the UE Context Release Command message to the source gNB to request the release of the UE-associated logical NG connection.
15. The source gNB sends the UE Context Release Complete message to the AMF to confirm the release of the UE-associated logical NG connection.

Note that in the present section we describe the handover via 5GC procedure, while the Xn handover is described in Section 3.3.3.

#### 3.3.3.2 NG-U Interface

The NG-U interface between an NG-RAN node and a UPF uses the same GTP-U-based protocol stack as the rest of the NG-RAN interfaces, with NG-specific user-plane protocol on top of GTP-U (i.e. in the NG-specific GTP-U container), as shown in Figure 3.3.6. This user-plane protocol is referred to as a PDU Session User-Plane protocol (3GPP TS 38.415).

Connectivity for a UE over the NG interface is provided over one or more PDU sessions, which are associations between the UE and a data network. When a UE registers with the network, it requests establishment of one or more PDU sessions. For every PDU session of a UE, there is one tunnel on the NG-U interface. However, each NG-U tunnel may have multiple QoS flows configured. This is one of the key differences compared with the S1-U interfaces used in LTE, where the EPC (and therefore S1) supported the E-UTRAN Radio Access Bearers (E-RAB bearers) and for every bearer there was only one tunnel on the S1-U interface. There was no further QoS differentiation within an E-RAB bearer.

In the downlink, a UPF maps a PDU to an NG-U tunnel and marks a PDU in accordance with its QoS flow. An NG-RAN node maps PDUs from QoS flows to radio access-specific

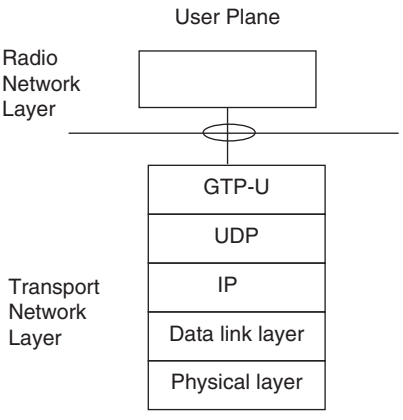



**Figure 3.3.6** NG user-plane protocol stack. (Source: Reproduced by permission of © 3GPP).

resources based on the QoS flow information, together with the information on NG-U tunnel.

A similar process happens in the uplink. An NG-RAN node selects an NG-U tunnel for an uplink PDU, and sends it to the UPF, together with the QoS flow information.

The QoS-related information is piggybacked on every GTP-U PDU in a GTP-U container, dedicated for the NG-U, and defined in 3GPP TS 29.281, whereas the content of the container is defined in 3GPP TS 38.415. The formats used in uplink (UL PDU SESSION INFORMATION) and downlink (UL PDU SESSION INFORMATION) are different. Both carry the QoS Flow Identifier (QFI), which indicates the QoS flow to which the transferred packet belongs. Additionally, the downlink packet may carry a Paging Policy Indicator (PPI) used for paging policy differentiation and a Reflective QoS Indicator (RQI), used for activation of the reflective QoS toward the UE.

### 3.3.4 Xn Interface

The Xn interface connects a gNB to another gNB or ng-eNB. It is primarily used for:

1. UE mobility control, i.e. handovers and UE mobility in inactive state;
2. UE data forwarding for lossless mobility;
3. Resource coordination, including coordination between NR and LTE;
4. Network energy saving, i.e. cell activation and deactivation;
5. Dual and multiconnectivity (described in Section 4.3).

#### 3.3.4.1 Xn Control Plane (Xn-C) Interface

The Xn-C protocol stack (3GPP TS 38.423) is similar to that of NG-C (3GPP TS 38.413). Therefore, in the present section we focus on the functions of the Xn Application Protocol (Xn-AP), used on the Xn-C interface.

Xn-C primary functions are:

- Xn interface management
- UE mobility in connected and inactive states
- Dual and multiconnectivity.

As the last one is described in Section 4.3, here we focus on the two first points.

Similar to the NG interface, Xn also supports multiple SCTP connection functionality to facilitate virtualized deployments.

In the present chapter we describe only the subset of the most commonly used procedures. For detailed description of all the procedures, refer to 3GPP TS 38.423.

#### *3.3.4.1.1 Interface Management*

The purpose and the general structure of the Xn interface management procedures are the same as those of the NG. The procedures are:

- Xn Setup Request/Response/Failure
- NG-RAN Node Configuration Update/Update Acknowledge/Update Failure
- Cell Activation Request/Response/Failure
- Reset Request/Response and Error Indication
- Xn Removal Request/Response/Failure.

One point worth mentioning regarding the interface management (and some other) Xn procedures is that because the interface supports two radio accesses (NR and LTE), depending on whether the interface is between two gNBs or between a gNB and ng-eNB, they support carrying the information about either NR or E-UTRA.

The Xn Setup Request/Response messages are used to establish the interface and to exchange the information about the served NR and E-UTRA cells, such as Physical Cell ID (PCI), Tracking Area Code (TAC), and frequency information.

The information about served cells exchanged during the setup can be updated using the NG-RAN Node Configuration Update procedure. Additionally, this procedure can be used to add, remove, and modify additional SCTP endpoints, which is used when new computational resources are added or removed in virtualized NG-RAN.

The Xn interface also supports switching off and on cells for energy-saving reasons. When a gNB decides to switch off a cell (e.g. because there are no UEs to serve), it indicates so using the Deactivation Indication IE in the NG-RAN Node Configuration Update messages. A neighbor node may request to turn that cell on using a different message – Cell Activation Request.

As for NG, Xn interface management messages use non-UE-associated signaling.

#### *3.3.4.1.2 Connected Mode Mobility*

We first describe the Xn-C support for connected mode mobility, which is performed using the following messages:

- Handover preparation
  - Handover Request
  - Handover Request Acknowledge
  - Handover Preparation Failure
- Handover execution
  - SN Status Transfer
- Handover completion
  - UE Context Release.

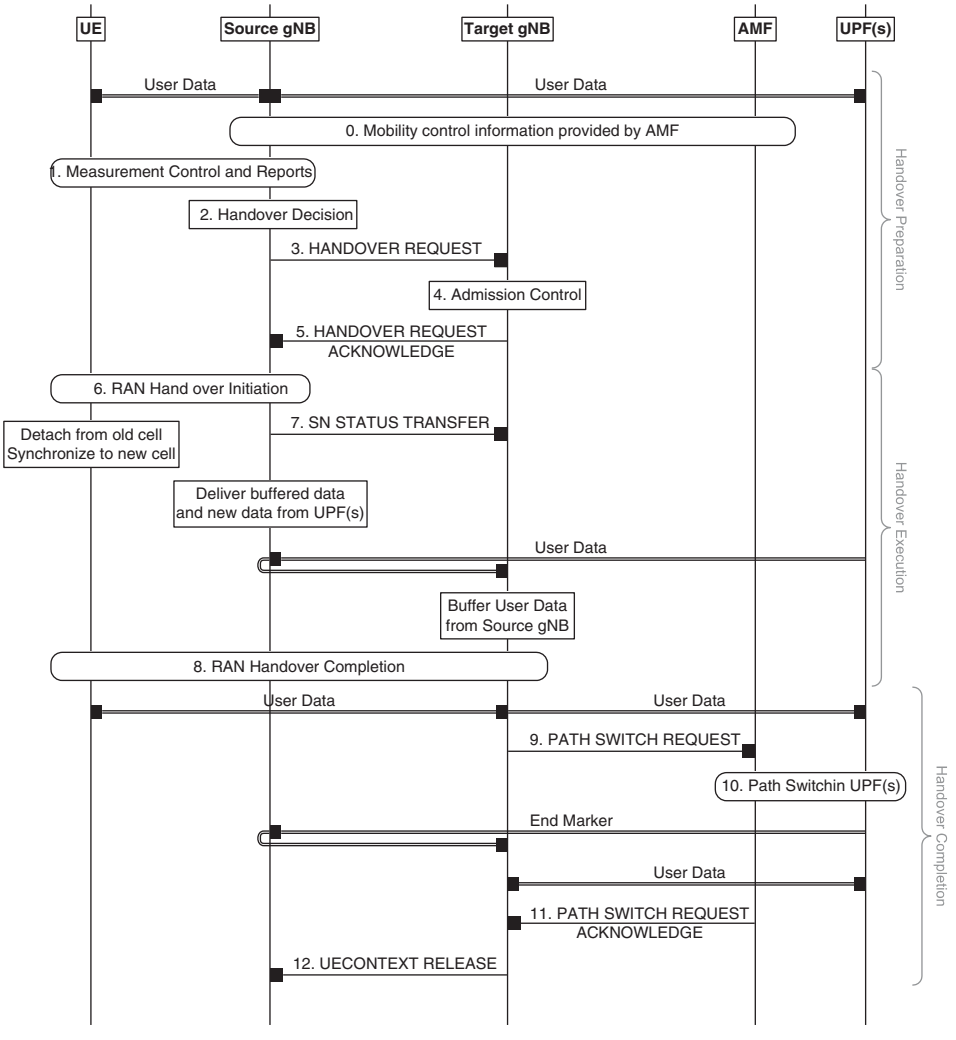



**Figure 3.3.7** Xn handover. (Source: Reproduced by permission of © 3GPP).

The usage of these procedures can be illustrated by the call flow example shown in Figure 3.3.7. Note that the procedure also uses some NG messages.

0. The UE is in RRC_CONNECTED, sending and receiving uplink and downlink data; the source gNB has the UE context, which contains information that may affect UE mobility decisions, such as mobility restrictions, radio capabilities, QoS, etc.
1. The UE has the measurement configuration provided by the source gNB, including frequencies to measure on, parameters to report, and thresholds to trigger the measurement reporting.
2. Based on the measurement report and potentially other information (e.g. cell load) and while taking into account UE mobility restrictions and radio capabilities, the source gNB decides to hand over the UE and selects the handover target.

3. The source gNB sends the Xn-AP Handover Request message to the target gNB over the Xn interface. The message caries the transparent RRC container with the Handover-PreparationInformation RRC message. Additionally, the message includes the target cell ID, list of PDU sessions, and other information.
4. The target gNB allocates the resources for the UE while taking into account the information received in the RRC container in step 3.
5. If the target gNB decides to admit the handover, it sends the Handover Request Acknowledge message to the source gNB, which includes a transparent container to be sent to the UE as an RRC message to perform the handover and the lists of admitted and not admitted PDU sessions. This completes the handover preparation stage.
6. The source gNB triggers the handover by sending the RRCReconfiguration message to the UE, containing the information required to access the target cell (received in step 5): at least the target cell ID, the new C-Radio Network Temporary Identifier (RNTI), and the target gNB security algorithm identifiers for the selected security algorithms.
7. The source gNB sends the SN Status Transfer message to the target gNB to transfer the uplink/downlink PDCP SN and Hyper Frame Number (HFN) status.
8. After the UE has successfully connected to the target cell, it completes the handover procedure by sending the RRCReconfigurationComplete message to target gNB.
9. The target gNB sends the NG-AP Path Switch Request message to AMF over the NG interface to trigger 5GC to switch the downlink data path toward the target gNB and to establish an NG-C interface instance toward the target gNB. The message also carries the list of PDU sessions to be switched and the list of PDU sessions which failed to set up at the target gNB.
10. 5GC switches the downlink data path toward the target gNB.
11. The AMF confirms the Path Switch Request message with the Path Switch Request Acknowledge NG-AP message. The message carries the list of PDU sessions that have been switched and the list of PDU sessions to be released.
12. Upon reception of the NG-AP Path Switch Request Acknowledge message from the AMF, the target gNB sends the UE Context Release Xn-AP message to the source gNB, which can then release resources associated with the UE.

#### *3.3.4.1.3 Mobility in Inactive State*

In the inactive state described in Section 3.4, a UE remains connected from the 5GC perspective and can move within a RAN-based Notification Area (RNA) without notifying the network. The last serving gNB (who sent the UE into the inactive state) keeps the UE context and the UE-associated NG connection with the serving AMF and UPF.

In order to send or receive data, a UE has to transition from inactive state to connected. This can be triggered by the network using paging or initiated by a UE. Two Xn procedures have been defined to support the inactive state:

- RAN paging
- Retrieve UE Context Request/Response/Failure.

The usage of UE context retrieval functionality can be illustrated by the UE initiated inactive to connected mode transition, shown in Figure 3.3.8.

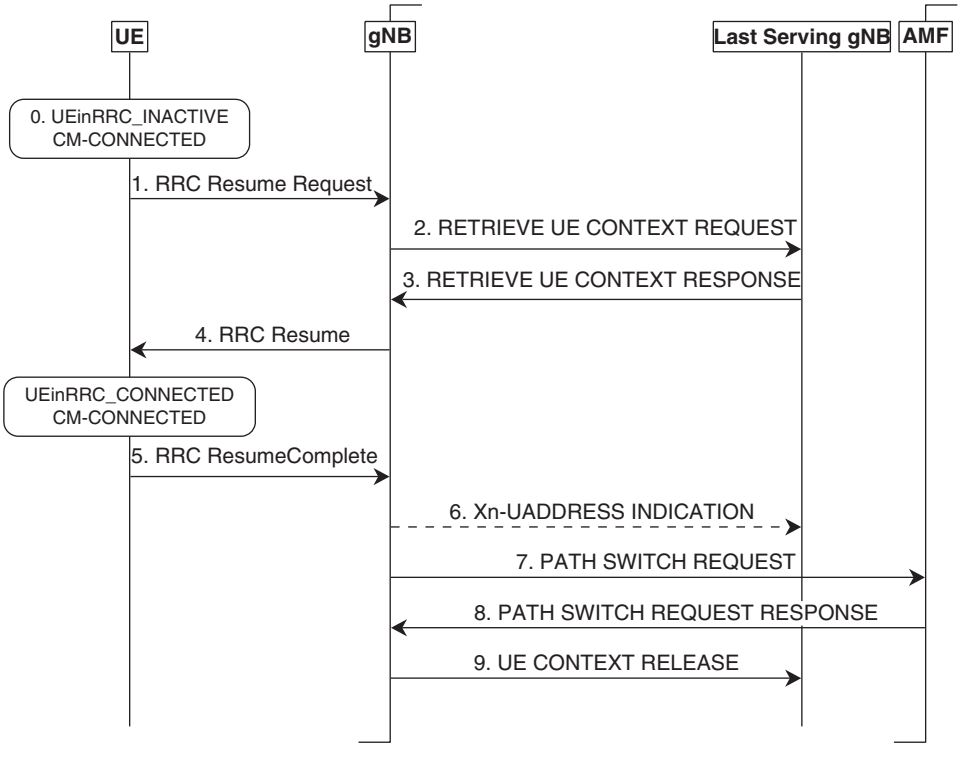



**Figure 3.3.8** UE-triggered transition from inactive to connected state. (Source: Reproduced by permission of © 3GPP).

0. A UE is in RRC_INACTIVE, it has the information about previously established connection with the network, however it is not actively sending or receiving data.
1. If the UE decides to resume from RRC_INACTIVE (e.g. to send uplink data), it sends the RRCResumeRequest message to the gNB controlling the cell it is camped on, providing the information with which the network can identify its context, i.e. I-RNTI (allocated by the last serving gNB).
2. As the gNB which the UE connects to may not be the one the UE has previously established the connection, the gNB sends the Xn-AP Retrieve UE Context Request message with the UE context ID (e.g. I-RNTI) to the last serving gNB over the Xn interface, in order to obtain the information about the UE context.
3. If the last serving gNB is able to identify the UE context by the ID received in step 2, it replies with the Retrieve UE context Response. This message carries UE security capabilities, PDU sessions information, RRC Context (HandoverPreparationInformation IE), and some other information.

4./5. The gNB and UE complete the resumption of the RRC connection. The UE uses the context information that it has stored and the gNB uses the context information that it has either stored (in the case it is the last serving gNB for that UE) or has received in step 3.

6. Optionally (if data forwarding is enabled), the gNB may send the Xn-U Address Indication message to the last serving gNB, which carries data forwarding address information for all the established PDU sessions and DRBs.

7./8. The gNB performs path switch toward the AMF as in the legacy Xn-based handover.

9. The gNB sends the Xn-AP UE Context Release message over the Xn interface to trigger the release of the UE resources at the last serving gNB. This step is also similar to the legacy Xn-based handover.Xn-AP RAN paging message is used, when the last serving gNB received downlink data for the UE. When a UE receives the RAN paging message, the UE initiated state transition procedure described above is performed.

#### 3.3.4.2 Xn User Plane (Xn-U) Interface

The Xn-U uses the same transport and protocol stack (3GPP TS 38.424) as NG-U, i.e. GTP-U/UDP/IP-based protocol stack as shown in Figure 3.3.6. The Xn-U supports three different types of payloads – PDCP Service Data Units (SDUs) (e.g. in case of DRB-level data forwarding upon handover), Service Data Adaptation Protocol (SDAP) SDUs (e.g. in the case of PDU session-level data forwarding upon handover), or PDCP PDUs (e.g. in the case of dual connectivity). For dual connectivity, the Xn-U uses an NR user plane (NR-U)[10] protocol (3GPP TS 38.425); that is, the content of the GTP-U container it uses and the functionality it supports are different to that of NG-U. As mentioned above, this user-plane protocol is common to Xn and F1 (described in Section 4.2).

Unlike the PDU session user plane protocol (3GPP TS 38.415) defined for NG-U, in the NR-U protocol GTP-U tunnels are mapped to NG-RAN bearers (corresponding to the DRBs), not PDU sessions. This is because even though the 5GC stopped using the notion of bearers, these are still present between UE and NG-RAN.

Furthermore, one additional reason for defining user plane enhancements for Xn and F1 network interfaces is that these interfaces are considered somewhat less reliable, compared with core network interfaces. Moreover, some NG-RAN network nodes may have smaller buffers. All this may lead to congestion on network interfaces and packet loss, which is why flow control and in-sequence delivery mechanisms have to be defined, which also helps when packet loss over the air interface occurs.

Primary NR-U functions are:

- *Data transfer*: transfer of data between NG-RAN nodes to support dual connectivity or CU/DU split deployment.
- *Flow control*: enabling a NG-RAN node to provide feedback information associated with the data flow received from a second NG-RAN node.
- *Retransmissions*: allowing retransmissions through the same node the data were forwarded to or through a different node.
- *Transfer of assistance information*: transfer of radio-related assistance information when the radio layer is not managed in the same node as the data control.

Data transfer is also used during the mobility operation, though in this case data are transferred without the NR-U extension, so all the NR-U functions are not available.

---

10 NR User Plane (NR-U) should not be confused with NR Unlicensed (NR-U), for which the same acronym is used in 3GPP specifications.

The NR-U functions are implemented as a set of additional messages carried inside the NR Container in a GTP-U extension header of a GTP-U PDU. The messages include:

- Downlink User Data
- Downlink Data Delivery Status (DDDS)
- Assistance Information Data.

The purpose of the Downlink User Data message is to carry the NR-U sequence numbers, to trigger the receiving node (e.g. gNB-DU in the case of downlink) to report various information (see below) and to request the receiving node to discard certain packets that are buffered. The NR-U sequence number is assigned consecutively for each and every Downlink User Data message transmitted to detect possible loss over the radio interface.

The receiving node may feed back the DDDS message carrying information on PDCP PDUs that have been successfully delivered to the UE, buffer status, lost NR-U sequence numbers, and some other information. The main purpose of this message is to prevent buffer underrun and overrun in the gNB-DU. Sending of the DDDS is up to the implementation, but can be requested by the sending node using a specific flag in the Downlink User Data message.

While the DDDS provides a coarse estimate of the UE downlink rate, the receiving node can also provide the Assistance Information Data message to report more precise radio-related information, such as: average channel quality indicator (CQI), average Hybrid ARQ (HARQ) retransmissions, etc. This information helps the sending node schedule downlink traffic for the UE.

The definitions of the NR-U procedures in the specification (3GPP TS 38.425) use somewhat obscure terms "the node hosting the NR PDCP entity" and "the corresponding node." These terms have been introduced to make it possible to generalize the protocol definitions so that they can be used on multiple network interfaces (e.g. F1 which is explained in Section 4.2) between different network nodes (e.g. between gNBs or between gNB-CU and gNB-DU), and for different purposes (e.g. dual connectivity, which is explained in Section 4.3). In the Xn-AP specification, the "corresponding node" is referred to as the "assisting node."

Xn-U PDUs may also be sent without payload, if delivery of the NR-U message is needed and there are no data to transfer.

### 3.3.5 Additional NG-RAN Features

In the present section we explain a few select NG-RAN features, which are either new in 5G or different compared with LTE.

#### 3.3.5.1 RAN Sharing

Similar to E-UTRAN in LTE, NG-RAN may be shared by multiple operators. In Release-15, only the 5G Multi-Operator Core Network (5G MOCN) network sharing architecture is supported, which is illustrated in Figure 3.3.9.

In MOCN, a NG-RAN can be shared between multiple operators each having their own 5GC. To support this functionality, NG-RAN can broadcast multiple PLMNs per cell.

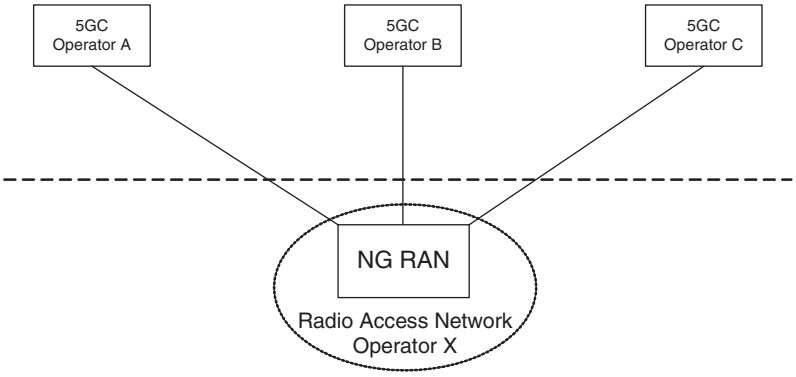



**Figure 3.3.9** RAN sharing. (Source: Reproduced by permission of © 3GPP).

In contrast to LTE, two slightly different architecture options for NG-RAN sharing have been defined. In the first option, which is similar to what is commonly used in LTE, all NG-RAN network interfaces (e.g. Xn and F1) are shared between all PLMNs (e.g. all core networks) that NG-RAN is connected to. In the second option, the notion of an interface instance is used on the NG-RAN network interfaces. The interface instance is identified by the Interface Instance Identification IE carried in all Xn-AP and F1-AP messages. For each interface instance a separate setup procedure is performed, however they all can use the same transport (i.e. the same SCTP association). The latter option is assumed to be beneficial for the case when RAN sharing with multiple cell IDs is deployed.

### 3.3.5.2 Slicing

Slicing is an important feature of 5G, which spans across an operator's network, including RAN, as shown in Figure 3.3.10.

It is generally assumed that a slice "tenant" will have a SLA with an operator to lease parts of an operator's network, including RAN and radio resources. It is then up to an

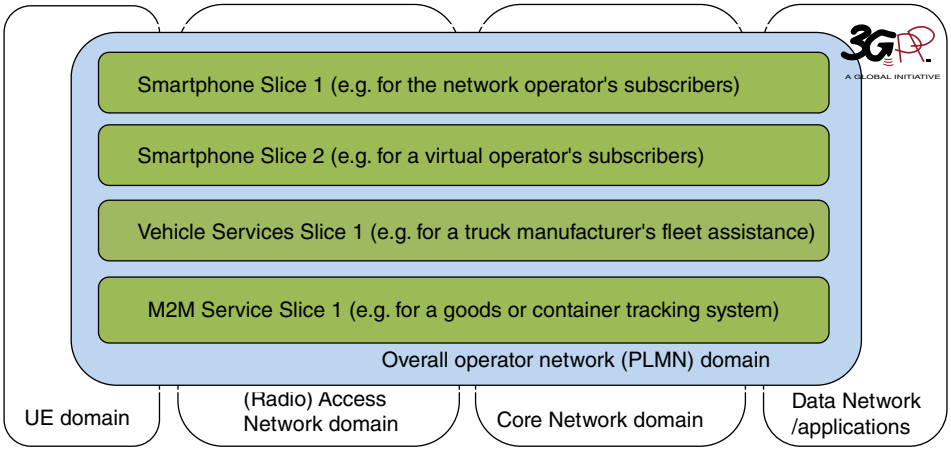



**Figure 3.3.10** Example of network slicing.

operator's policy and NG-RAN implementation to ensure that the SLA is fulfilled. For example, an NG-RAN implementation may need to provide resource isolation between different slices (if multiple slices are active at a time), which can be "soft isolation" or "hard isolation."

However, from the NG-RAN architecture, network interfaces, and protocols, very little is specified, leaving plenty of room for implementation. During network interface (e.g. NG and Xn) setup, both network nodes (gNB and AMF or gNBs, respectively) exchange information about slices they support. A UE may provide to the network the NSSAI it selected using the RRCSetupComplete message. Subsequently, when a UE context is established in NG-RAN, the AMF indicates the list of allowed slices (e.g. allowed NSSAI) for the UE. This information is then used by NG-RAN, among other things, for:

- Selecting an AMF for a UE and routing the Initial UE message over the NG interface to an appropriate AMF that supports the slices requested by the UE;
- UE access control;
- Resource isolation between slices;
- UE mobility decisions.

However, the actual logic to support the above is not specified. In the end, slicing functionality will depend a lot on implementation and may vary vendor by vendor.

As mentioned above, a network slice is identified by an S-NSSAI, which is a combination of:

- Mandatory SST (Slice/Service Type) field, which identifies the slice type and consists of 8 bits (with range is 0–255);
- Optional SD (Slice Differentiator) field, which differentiates among Slices with same SST field and consist of 24 bits.

The network may support a large number of NSSAIs, while a UE should not support more than eight slices simultaneously.

#### 3.3.5.3 Virtualization

Generally, all 3GPP network nodes can be virtualized, which has been the case in many EPC (i.e. core network) implementations. In 5G, the 5GC core network has been designed from the beginning with virtualization in mind (as explained in Section 3.2). In order to work efficiently with virtualized 5GC and also to support virtualization of NG-RAN network nodes, certain functionality has been added to NG-RAN interfaces.

In particular, and in contrast to LTE, all NG-RAN network interfaces (NG, Xn, F1, and E1) support dynamic addition and removal of multiple SCTP endpoints on each termination point of each interface. This facilitates certain types of virtualized deployments, in which new computational and transport network resources may be added or removed dynamically – since in some case some of these resources may use a different transport network (i.e. IP) address, NG-RAN and 5GC support that in a manner which does not disrupt UE operation.

More details about NG-RAN virtualization can be found in Section 6.2. Those details have, however, little protocol impact.

#### 3.3.5.4 Non-3GPP Access

Integration with non-3GPP access technologies, IEEE 802.11 (WLAN) in particular, is not new to 5G – various solutions have been defined by 3GPP in the past. For example, integration of WLAN with LTE was enabled at core network and RAN levels.

In 5G, 3GPP made an attempt to make the NG network interface "access-agnostic," so that from the point of view of 5GC a "common core" can be used with 3GPP NG-RAN or E-UTRAN, as well as with non-3GPP Access Node (AN). This is described in more detail in Section 3.2, but from the high-level point of view, the NG interface can terminate in either

- a gNB (or ng-eNB) providing NR (or LTE) radio access;
- or in the N3IWF network node, providing, e.g. WLAN access.

The following provisions have been made in the definition of the NG network interface:

- 3GPP TS 29.413 defines which NG-AP messages (e.g. paging) and which IEs (e.g. Trace Activation) are not applicable to non-3GPP access.
- Certain IEs in 3GPP TS 38.413 of the NG-AP are defined to facilitate non-3GPP access, e.g. the paging origin IE, which allows 5GC to page a UE via 3GPP access (e.g. NR) to establish a connection on a non-3GPP access.

Overall, the NG interface does not make 5GC truly access-agnostic, but rather allows an operator to use a common core with multiple access technologies. However, the 5GC should be aware which network node (gNB or N3IWF) an NG interface is terminated at.

## References

3GPP Technical Specification 29.281 (2019). General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 29.413 (2019). Application of the NG Application Protocol (NGAP) to non-3GPP access. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 23.501(2019). System architecture for the 5G System (5GS). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 36.413 (2019). Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.412 (2019). NG-RAN; NG signalling transport. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.413 (2019). NG-RAN; NG Application Protocol (NGAP). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.414 (2019). NG-RAN; NG data transport. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.415 (2019). NG-RAN; PDU Session User Plane protocol. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.423 (2019). NG-RAN; Xn Application Protocol (XnAP). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.424 (2019). NG-RAN; Xn data transport. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification 38.425 (2019). NG-RAN; NR user plane protocol. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

802.11-2016 IEEE Standard for Information technology – Telecommunications and information exchange between systems Local and metropolitan area networks – Specific requirements – Part 11 (2016). Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications IETF RFC 4960, Stream Control Transmission Protocol. Internet Engineering Task Force.

## 3.4 NR Protocol Stack

*Sudeep Palat*

*Intel Corporation, UK*

### 3.4.1 Introduction

As mentioned above, 5GS architecture is functionally split into 5GC and RAN (NR) functionalities similar to older cellular technologies. 5GC functionality is responsible for the overall user registrations and session management functions, while NR provides the radio access-related functionalities. 5GC functionality is discussed in more detail in Section 3.2. NG-RAN provides the connectivity over the radio interface between the network and the UE with the user-plane functions providing the required QoS for the data over the radio interface.

While the network aspects of NG-RAN are explained in Section 3.3, in the present section we focus on the air interface protocol stack, specifically:

- The NR protocol stack consists of user-plane and control-plane parts: the user plane handles the transfer of data across the radio interface with the required QoS, while the control plane handles the configuration of the radio interface connection.
- In the non-split NG-RAN architecture, the gNB hosts all of the protocol stack functionality described in this section; different protocol stack layers are hosted in different logical network nodes in the split NG-RAN architecture (see Section 4.2).
- The user-plane protocol stack is similar to LTE and consists of SDAP, PDCP, Radio Link Control (RLC), and Medium Access Control (MAC) layers. The SDAP layer (not present in LTE) maps a packet to a DRB based on QFI. The PDCP layer provides encryption, integrity protection, and IP header compression functionalities. Additionally, PDCP is responsible for reordering. The RLC layer provides segmentation and reliability with the Automatic Repeat Request (ARQ) functionality. Finally, the MAC layer is multiplexing data from different logical channels into a transport block and scheduling.
- Even though 5GC no longer uses the notion of bearers, the concept of radio bearers is kept in NR, which maps 5GC flows to DRBs based on QFI.
- RRC, which is the control-plane protocol of NR, is also similar to LTE. The protocol has a number of "transparent containers" defined, which can be used to convey e.g. UE configuration messages to be delivered to the UE by an intermediate network node without interpreting the content, e.g. in EN-DC operation (see Section 4.3). Other main differences are related to beam-based measurements and on-demand System Information Broadcast.
- Unlike LTE, NR RRC supports three different UE states – IDLE, CONNECTED, and INACTIVE. The INACTIVE state is similar to IDLE in terms of UE actions, with the main difference being that the NG-RAN maintains the UE context and the connection to 5GC.

These are discussed in more detail in the following sections.

### 3.4.2 NG-RAN Architecture

The RAN functions in NR reside in the logical node, gNB. The gNBs are connected by means of the NG interfaces to the 5GC, more specifically to the AMF by means of the NG-C interface and to the UPF by means of the NG-U interface. The gNBs are interconnected with each other by means of the Xn interface. The simplified NG-RAN architecture is shown in Figure 3.4.1. A gNB can be further split into different logical nodes as discussed in Chapter 4.

### 3.4.3 NR User Plane

The primary role of the user plane is to transfer data across the radio interface efficiently and with the required QoS.

Figure 3.4.2 shows the protocols across the different interfaces for the transfer of a user IP packet between the UPF and the UE. The radio protocols, also called AS, are shown in gray. The higher layers (also called NAS) indicate the QoS handling required for a packet.

Each RB has four L2 protocol layers – the upper layers of SDAP (used only for DRBs) and PDCP, along with the lower layers of RLC and MAC (called RLC bearer, which corresponds roughly to a logical channel). Such a split between the upper and lower layers of the DRB allows easier logical separation of the RLC bearer into a different logical node from the upper protocol layers. This split is applied for the CU–DU architecture as discussed in Section 4.2 and dual connectivity architectures as discussed in Section 4.3. Each RB can be configured with more than one of the RLC bearers that may reside in the same cell group for Carrier Aggregation (CA) or a different cell group in case of dual connectivity. If the

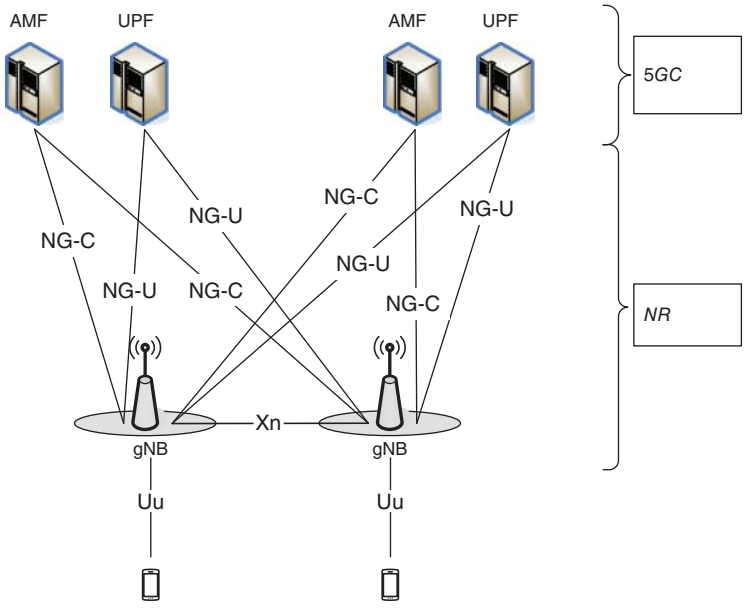



**Figure 3.4.1** NG-RAN architecture.

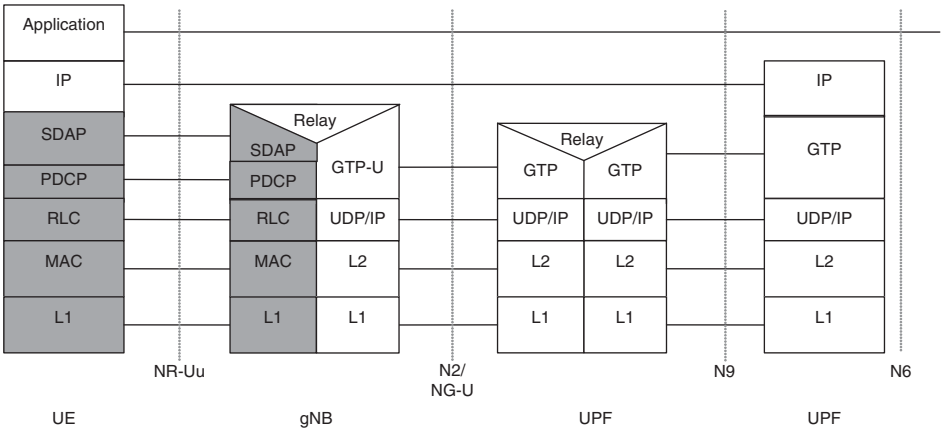



**Figure 3.4.2** User-plane protocol stack.

RLC bearers are in different nodes as in dual connectivity, it can provide higher throughput for the RB. Alternatively, data packets can be duplicated and sent over the different RLC bearers for higher reliability, for example with CA duplication.

The SDAP layer is common across all the RBs of a PDU session and its role is to split the data into the different RBs of the PDU session in accordance with the QoS information.

The overall user-plane functionality can be summarized in Figure 3.4.3.

The protocol layers are discussed in more detail in the following sections.

Primary functionality of the SDAP (specified in 3GPP TS 37.324) is to map a packet to a DRB based on its QFI. The mapping table is configured either using RRC signaling (often called explicit) or in-band using a SDAP header (called AS reflective) and is discussed further below. The SDAP layer has an optional configurable header and carries different information in the uplink and downlink as shown in Figure 3.4.4.

In the downlink, the SDAP header includes reflective QoS flow to DRB mapping Indication (RDI) and RQI bits along with the QFI. The RDI bit is the AS reflective mapping bit. On receipt of a packet with this bit set, the UE updates the mapping table to map uplink packets of that QFI to the DRB the packet was received in. The RQI bit is the NAS reflective mapping bit and is provided to the upper layers of the UE along with QFI of the data packet. In the uplink, the header carries only the QFI and is used by the network for handling the packet over the CN network nodes and interfaces.

There is one SDAP entity in the UE per PDU session. On the network side, there may be up to two SDAP entities per PDU session – one in the MN and the other in the SN in case of dual connectivity (as described in Section 4.3).

The PDCP layer (specified in 3GPP TS 38.323) provides the security functions of encryption and integrity protection for the data packets. It also offers Robust Header Compression (RoHC) header compression for IP traffic. Both these functions are similar to LTE with the main exception that user data packets can also be configured with integrity protection for better security. PDCP data PDU consists of a header containing the Sequence Number, the data payload, and an optional integrity protection checksum.

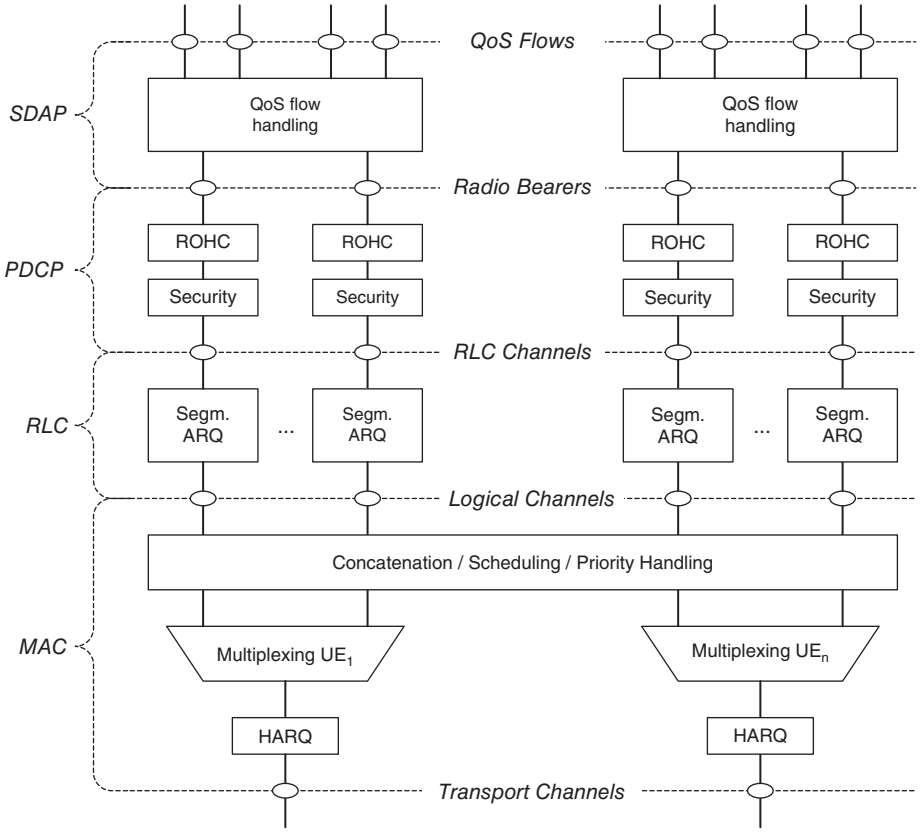



**Figure 3.4.3** Downlink Layer 2 Structure Protocol. (Source: Reproduced by permission of © 3GPP).

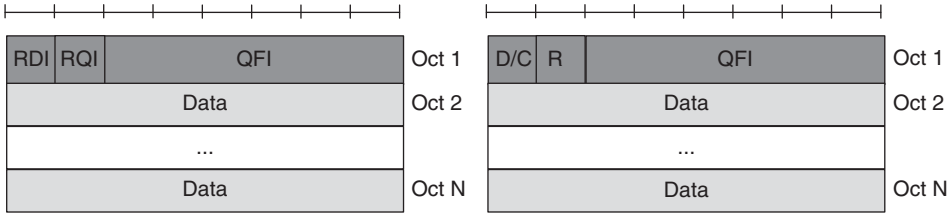



**Figure 3.4.4** Downlink (left) and uplink (right) SDAP Data PDU format with SDAP header. (Source: Reproduced by permission of © 3GPP).

PDCP is also responsible for retransmission of packets based on PDCP status reports to ensure lossless delivery of packets for the cases where RLC cannot ensure lossless delivery. This could be when RLC bearers are released or moved to another network node, such as during handover, or with dual connectivity, or CA. PDCP can also be used to duplicate packets across multiple RLC bearers of the RB as a mechanism to improve reliability with lower latency compared with RLC-based retransmission.

Unlike LTE, PDCP is responsible for reordering and duplicate detection continuously rather than just during a handover. Duplicate detection and reordering based on PDCP SN is built into PDCP protocol itself at the receiving end and operates continuously without the need for additional configuration.

Each DRB has exactly one PDCP entity and one or more RLC bearers. The interface between PDCP and RLC is well specified to allow the PDCP and RLC to be in different nodes such as in the case of dual connectivity or split CU–DU architecture.

The RLC layer (specified in 3GPP TS 38.322) provides the required reliability for the data transmission. Similar to LTE, it supports three transmission modes – Transparent Mode (TM) used for data that are broadcast, Unacknowledged Mode (UM) used for services that can tolerate data loss such as voice, and Acknowledged Mode (AM) with retransmission mechanism (ARQ) for services such as TCP/IP and RRC signaling (Signaling Radio Bearers [SRBs]) that require reliable delivery of data. Both RLC UM and AM modes also support segmentation and resegmentation of data at the transmitter to fit into the transport block size and reassembly at the receiver. RLC status PDUs provide positive and/or negative acknowledgments of RLC SDUs (or portions of them) for ARQ functionality. RLC AM also provides duplicate detection.

Figure 3.4.5 shows one of the RLC PDU structures. SI indicates the Segmentation Information on whether the data are segmented and whether it is the first, last, or an intermediate segment. SN is the RLC Sequence Number, and the Segment Offset (SO) field indicates the position of the RLC SDU segment in bytes within the original RLC SDU. P is the polling bit, which can be used to trigger status reporting from the peer AM RLC entity. D/C indicates whether the PDU is a Data or Control PDU.

Unlike LTE, RLC in NR does not do reordering of data as PDCP offers continuous (i.e. not just during handovers) reordering. RLC also does not perform concatenation of multiple RLC SDUs and this function is now part of the MAC multiplexing. This allows the UE to pregenerate uplink RLC PDUs from RLC SDUs before receipt of an uplink grant and therefore speeds up the generation of MAC PDU after the uplink grant is received. This in turn allows low uplink grant to transmission delay and reduction in latency. The last RLC segment of the MAC PDU still need to be generated after reception of an uplink grant as it requires information about the number of bits available. This is considered acceptable as

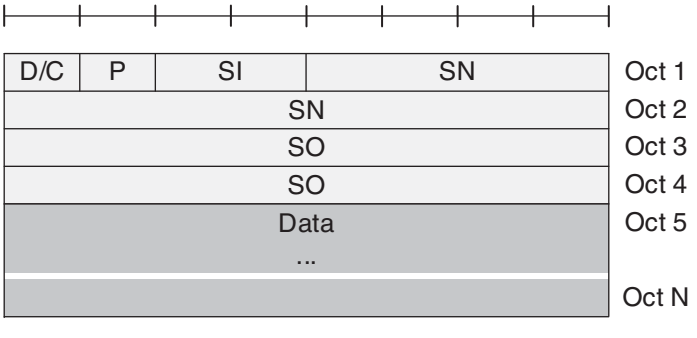



**Figure 3.4.5** RLC AMD PDU with 12 bit Sequence Number with Segment Offset. (Source: Reproduced by permission of © 3GPP).

**Table 3.4.1** Downlink channel mapping.

| Transport channel<br>Logical channel | BCH<br>(broadcast channel) | PCH<br>(paging channel) | DL-SCH (downlink<br>Shared channel) |
|---|---|---|---|
| BCCH (Broadcast Control Channel) | X | | X |
| PCCH (Paging Control Channel) | | X | |
| CCCH (Common Control Channel) | | | X |
| DCCH (Dedicated Control Channel) | | | X |
| DTCH (Dedicated Traffic Channel) | | | X |

the UE can perform this while it is putting together and possibly already providing to the physical layer the initial bits of the MAC PDU for transmission.

The MAC layer (specified in 3GPP TS 38.321) multiplexes the data (RLC PDUs) from different logical channels into a transport block for transmission over the radio interface. Table 3.4.1 shows the mapping between logical channels and transport channels in the downlink.

The Common Control Channel (CCCH) is used during the initial access before dedicated channels are established, such as for RRC Connection Request. The Dedicated Control Channel (DCCH) carries UE-dedicated RRC signaling messages, while the Dedicated Traffic Channel (DTCH) carries user-plane traffic.

As discussed above, this mapping function multiplexes RLC PDUs from different logical channels and also does the concatenation function of including different RLC PDUs from the same logical channel.

Each logical channel is assigned a scheduling priority and Prioritized Bit Rate (PBR) by the network. On receipt of an uplink grant, the UE MAC fills each MAC PDU with data from the highest priority logical channel up to the PBR and then cycles through the lower priority logical channels in sequence. This functionality is similar to LTE. NR MAC also includes a new functionality to restrict data from certain logical channels to be sent only on certain numerologies/cells. This is used for example to send the duplicate PDCP PDUs data of the RB on different logical channels of different cells in the case of CA.

Semi Persistent Scheduling (SPS) functionality similar to LTE is also used in NR with some enhancements. Two types of Configured Grants are used – Type 1 where the periodicity and resources for the uplink grant is provided to a UE by RRC, and Type 2 where the resource is configured by a Layer 1 control channel, Physical Downlink Control Channel (PDCCH), and the periodicity by RRC.

MAC is also responsible for handling Hybrid ARQ (HARQ) processing and retransmissions as in LTE.

Apart from data transfer, the MAC protocol also supports several control elements that are used for signaling and configuration of the user plane. The Scheduling Request (SR) is used for requesting UL-SCH resources for new transmission. SR can be transmitted on Physical Uplink Control Channel (PUCCH) resources if available or using the Random Access (RACH) procedure. Buffer Status reports are generated by MAC to inform the network

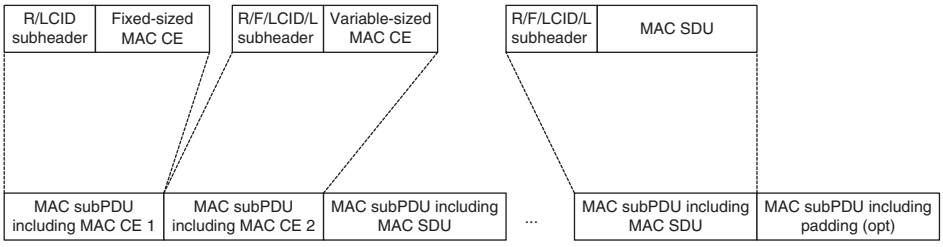



**Figure 3.4.6** Example of a downlink MAC PDU structure. (Source: Reproduced by permission of © 3GPP).

about pending uplink data in the buffer. PDCP duplication can also be enabled/disabled using a MAC control element.

Furthermore, MAC includes several power saving functionalities as in LTE. Discontinuous reception (DRX), when configured, controls the UE's PDCCH monitoring activity and is used, as in LTE, to save UE power consumption. When CA is configured, Secondary Cells (SCells) can be activated or deactivated using MAC control elements to save power.

Figure 3.4.6 shows a MAC PDU structure consisting of MAC control elements and MAC sub-PDUs. Each MAC sub-PDU includes the MAC subheader and MAC SDU. This is unlike LTE where all the MAC subheaders are grouped and placed at the head of a MAC PDU. Such distribution of a MAC subheader to each MAC SDU in NR is again for faster generation of MAC PDU after receipt of an uplink grant by maximizing preprocessing.

RACH procedure is handled by MAC. NR follows the contention-based and contentionless RACH procedures of LTE where successful completion of contention resolution in message 4 or successful Random Access Response (RAR) for contentionless access represents successful completion of the RACH procedure. In addition to the RACH triggers as in LTE, such as for connection establishment, handover access, Timing Advance (TA) update in CONNECTED state, and SR, RACH in NR is also used for beam failure recovery and SI request.

The RACH beam failure recovery procedure is used to indicate to the serving gNB when beam failure is detected on the serving Synchronization Signal Blocks (SSBs)/Channel State Information Reference Signals (CSI-RSs). A UE indicates beam failure to the network when the number of beam failure instance indications from the lower layers to the MAC entity exceeds a certain configured threshold. Successful completion of the RACH procedure represents successful beam failure recovery. RACH can also be used to request on-demand System Information as discussed in Section 3.4.5.2.7.

A single MAC entity in the UE can support multiple numerologies, transmission timings, and cells, such as in the CA.

### 3.4.4 Supporting QoS with 5GC

5GC uses a different QoS concept from 4G EPC. Each downlink data packet is marked by the 5GC with a QFI indicating the QoS requirement for the packet. In the uplink such marking of an IP packet to the QFI is performed by the NAS layer, based on its configuration.

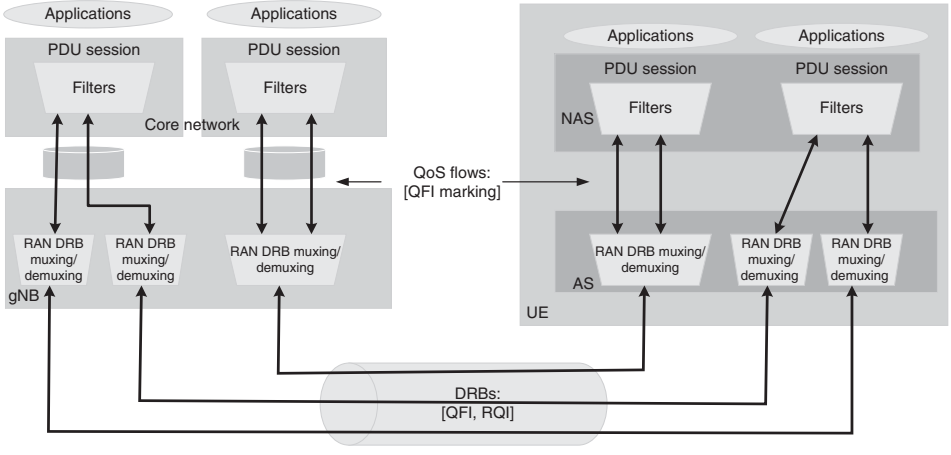



**Figure 3.4.7** QoS flow mapping in the CN, RAN, and UE.

The QFI is carried in the GTP header (3GPP TS 29.281) and provided to the gNB with each packet. Similar to LTE, Radio Bearers are used for transferring data over the radio interface, that is, RRC messages are carried by SRBs and user data is transferred over DRB. Each radio bearer provides a certain QoS for the data exchanged over it and multiple RBs can be set up per UE as needed to provide different QoS requirements for different services.

RAN is responsible for setting up DRBs supporting different QoS and also for data to be sent on the DRB that meets its QoS requirement. RAN node has the flexibility to decide when and how many DRBs are to be set up for the UE. For example, data with different QFIs can be mapped into one DRB, provided that the QoS of the DRB meets the QoS requirements of all the QFIs mapped to it. The mapping of QoS flows to DRBs is shown in Figure 3.4.7.

In the downlink, this mapping can be decided by the gNB without any preconfiguration to the UE by simply sending data for the QFI on a DRB. For the uplink, the mapping information from QFI to DRB has to be provided to the UE by gNB. This could be by explicit signaling using RRC Reconfiguration message or using the new concept of AS reflective mapping. With AS reflective mapping, the mapping table for a QFI can be updated to use the DRB on which the UE received a downlink packet for the QFI with a RQI bit set in the SDAP header. This allows RAN to change the uplink mapping dynamically using user-plane signaling without using RRC.

The signaling for allocation of the QoS flow to a DRB is shown in Figure 3.4.8.

0. UE is registered with the network, and PDU sessions and DRBs are established to send data between UE and the network.
1. gNB receives a new data packet with a new QFI marking.
2. gNB decides to send this QoS flow on an existing DRB that meets the QoS requirement and to use AS reflective mapping.
3. gNG sends the data packet on the chosen DRB with the QFI and RQI bit set in the SDAP header.

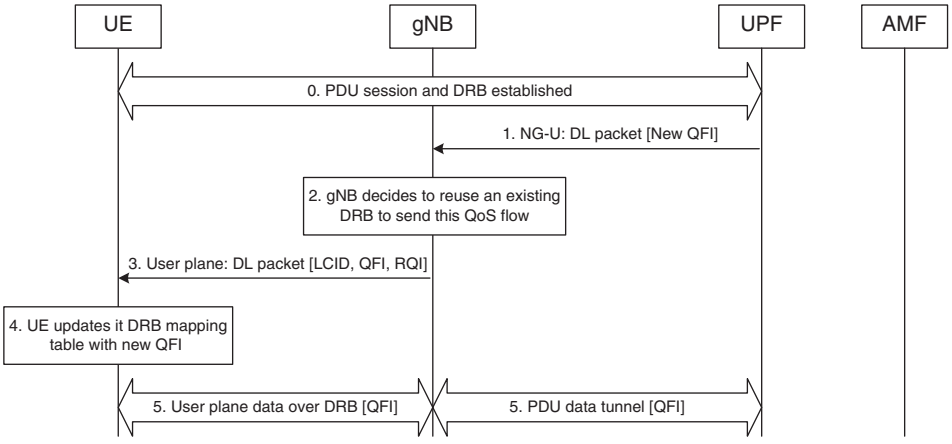



**Figure 3.4.8** An example signaling for a UE configuration of the DRB mapping for a new QoS flow using AS reflective mapping.

4. UE, upon receipt of the packet with the RQI bit set, updates the mapping table to send UL data with this QFI on this DRB.
5. Data for this QoS flow are exchanged over the DRB.

### 3.4.5 NR Control Plane

RRC (specified in 3GPP TS 38.331) is control-plane protocol used to configure and manage the radio interface connection. RRC provides a set of procedures that are used for connection establishment, security configuration, mobility, physical layer, and user-plane configurations.

#### 3.4.5.1 RRC States

RRC supports three different UE states:

- IDLE
- CONNECTED
- INACTIVE.

The IDLE and CONNECTED states are similar to LTE. A UE in active communication exchanging data with the network will be in the CONNECTED state. In this state, physical channels and user-plane protocols are configured and set up. UE mobility is controlled by the network using RRC messages. A UE context is created in the RAN when the UE moves to the RRC CONNECTED state. The UE context includes the information about the connection to the CN, the user plane, and physical layer configurations. The UE context is cleared when the UE leaves RRC CONNECTED and goes back to RRC IDLE.

A UE is in IDLE or INACTIVE state (see below) when not in active communication. In these states, a UE is just performing cell reselection to camp on the best cell, staying up to date with System Information Broadcast information, and monitoring Paging for downlink data and Public warning messages.

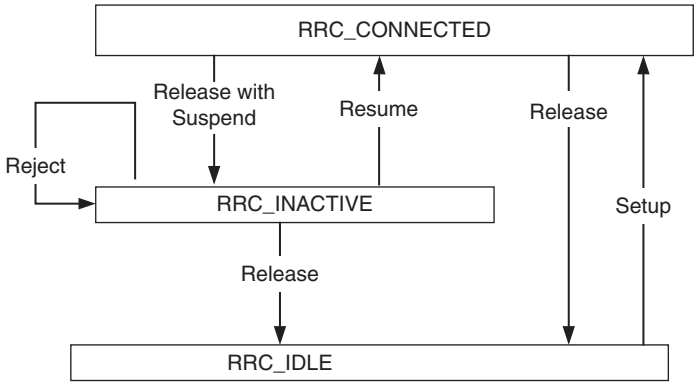



**Figure 3.4.9** RRC states and state transitions.

NR introduces a new RRC state called INACTIVE, which is similar to the IDLE state in terms of the UE actions. A UE in this state keeps a copy of the previous RRC configuration, including the security configuration before it went into the INACTIVE state. The UE configuration, context, and the connection to the core network are also maintained in the gNB. These configurations are kept suspended.

Figure 3.4.9 shows the transitions between different RRC states. The procedures to trigger state transitions are discussed in the following subsections.

#### 3.4.5.2 RRC Procedures and Functions

NR RRC defines a set of procedures that are functionally and procedurally similar to LTE. RRC connection control is a set of procedures that support the establishment and release of the RRC connection. It covers the procedures responsible for UE initial access to the network and the set up of security in the AS.

##### *3.4.5.2.1 RRC Connection Establishment*

The RRC Connection establishment procedure is used to move the UE from RRC state IDLE to CONNECTED in order to communicate with the network. From the core network perspective, this procedure may also involve registration with the 5GC to obtain 5G services. An example AS message flow for registration is shown in Figure 3.4.10.

1. A UE in IDLE that intends to perform a NAS registration procedure to attach to the network initiates a RACH procedure with a RACH preamble.
2. The gNB responds with a RAR message providing the timing advance and resources to use for the subsequent uplink message. These messages are defined in MAC specification (3GPP TS 38.321).
3. The UE then sends the RRC Setup Request message as a CCCH message, also referred to as RACH message 3. It contains the NAS UE identifier if available, or a random number if not.
4. The UE ID contained in the RACH message 3 is then echoed back to the UE in the UE Contention Resolution Identity MAC CE (3GPP TS 38.321) for contention resolution in RACH message 4. The RRC Setup message, also a CCCH message, could be included in

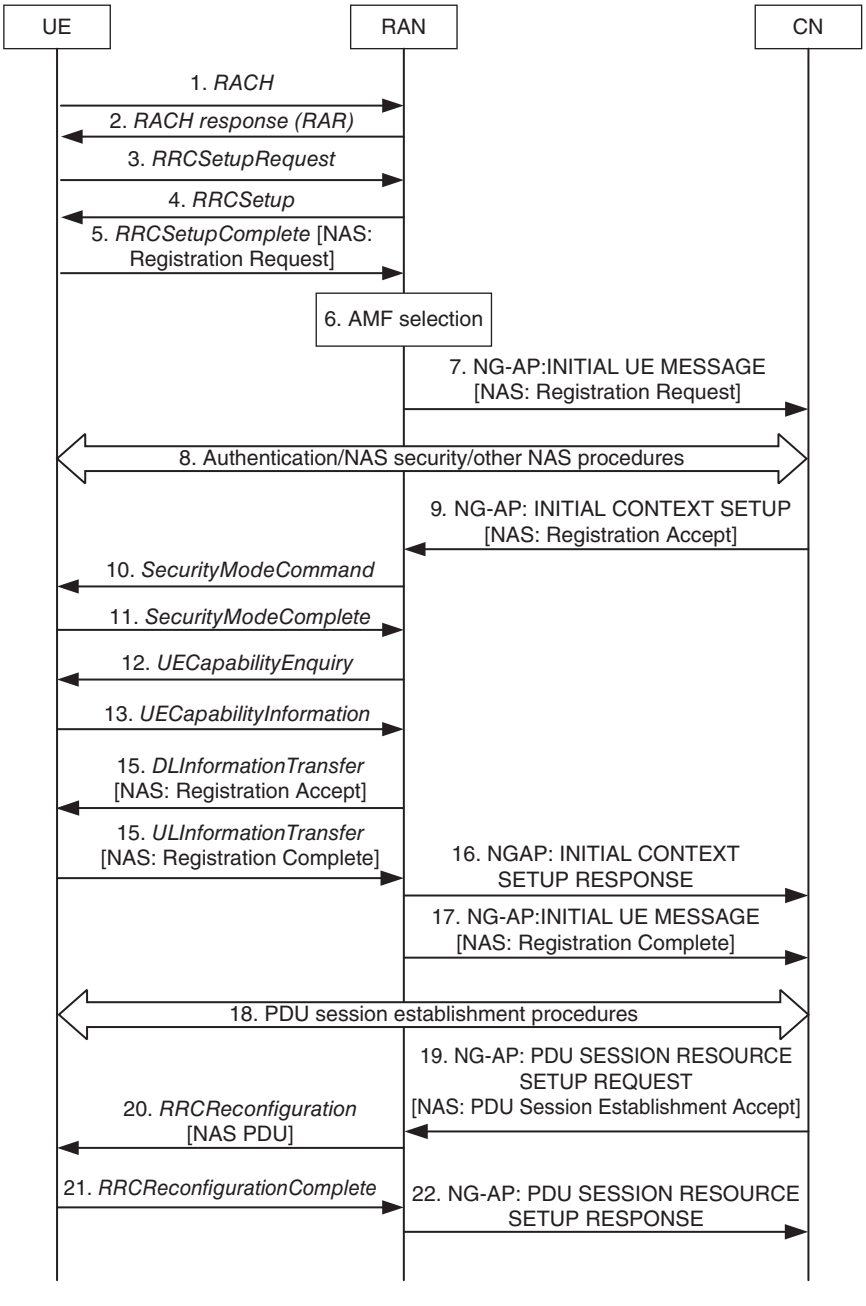



**Figure 3.4.10** An example message flow for initial NAS registration.

the RACH message 4 or sent separately to the UE. It contains the initial configuration to set up the DCCH logical channel and SRB1. Subsequent RRC messages are sent over DCCH on SRB1.

5. After the set up of SRB1, the UE can send NAS messages to the core network that are encapsulated in RRC messages and subsequently sent to the core network by the gNB. The NAS registration request is encapsulated in the RRC Setup Complete message. Additional NAS messages can be exchanged from the UE to the core network encapsulated in the RRC ULInformationTransfer message over SRB1 and SRB2 after SRB2 is established.
6. The gNB selects the AMF based on the UE ID and other NAS selection information such as NSSAI and the registered AMF included in the RRC Setup Complete message.
7. The gNB forwards the NAS registration request to the AMF in the NG-AP INITIAL UE MESSAGE (specified in 3GPP TS 38.413).
8. The AMF may initiate additional procedures with the UE NAS that are encapsulated and transferred using RRC DLInformationTransfer and ULInformationTransfer messages.
9. The CN provides the NG-AP INITIAL CONTEXT SETUP REQUEST message (specified in 3GPP TS 38.413, also explained in Section 3.3) to the gNB, which contains information to establish the AS security and optionally PDU session information to set up DRBs.
10. RRC SecurityModeCommand configures the security algorithms and establishes AS security.
11. The UE acknowledges successful completion of the security configuration with RRC SecurityModeComplete.
12. RRC UECapabilityEnquiry is used to retrieve the UE AS capability. The UE capability in NR is subdivided into NR Standalone and capability related to dual connectivity in the MR-DC capability containers. The gNB may request one or more of these capabilities. The UE capability enquiry procedure can be executed at any time though this is now recommended to be after security activation.
13. The UE responds with the requested capabilities in RRC UECapabilityInformation.
14. RRC DLInformationTransfer encapsulates the NAS Registration Accept message to be sent to the UE.
15. RRC ULInformationTransfer encapsulates the NAS Registration Complete message to be sent to the network.
16. The gNB sends the NG-AP INITIAL CONTEXT SETUP RESPONSE message to the AMF indicating the successful completion of the UE context set up procedure.
17. The gNB forwards the NAS Registration Complete message received from the UE to the AMF.
18. The UE initiates the PDU Session Establishment procedure with the CN. These are carried transparently over RAN using RRC uplink and downlink InformationTransfer messages.
19. The core network initiates the bearer set up procedures for the PDU session using the NG-AP PDU SESSION RESOURCE SETUP REQUEST message. It encapsulates the NAS PDU to be sent to the UE.

20. The gNB configures the DRB(s) for the PDU session using an RRC Reconfiguration message, which also encapsulates the NAS PDU to be sent to the UE. It can also provide other radio configuration such as physical layer, measurements, etc.
21. The UE replies with an RRC Reconfiguration Complete message upon successful completion of the configuration.
22. The gNB indicates the successful completion using the NG-AP PDU SESSION RESOURCE SETUP RESPONSE message.

An RRC Reconfiguration message is used to provide most of the UE configurations while the UE is in the CONNECTED state, such as RB configurations, measurement configuration, physical layer configuration, etc.

Additionally, it can be used to configure dual connectivity. Some part of the configurations may be generated by another node, either by the DU in the case of CU/DU split of a gNB or another gNB in the case of dual connectivity. The configuration generated by another node is carried within containers in the RRC Reconfiguration message and can be forwarded transparently by the serving gNB or gNB-CU putting together the final RRC Reconfiguration message sent to the UE. This is to allow the other nodes to support configurations or even to be of different 3GPP releases that may not be supported or comprehended by the serving gNB-CU. Some notable examples are discussed further below.

The RLC bearer and physical layer configuration may be put together by the gNB-DU and is provided in the container, *CellGroupConfig*. The *RBConfig* carries the SDAP and PDCP configurations and could be provided by the master or secondary gNB-CU in the case of dual connectivity. There are two containers defined in the RRC Reconfiguration message so that the UE behavior is the same irrespective of whether the *RBConfig* is originated from the master or secondary cell group. The snippet of the ASN.1 below shows the usage of the containers in the RRC Reconfiguration message.

```
masterCellGroup OCTET STRING (CONTAINING CellGroupConfig) OPTIONAL, – Need M

radioBearerConfig 2 OCTET STRING (CONTAINING RadioBearerConfig) OPTIONAL, – Need M
```

NR Standalone (i.e. NR connected to 5GC) supports two SRBs as in LTE. An additional SRB, SRB3, is introduced for direct RRC signaling to and from the SN when a UE is configured with dual connectivity (see Section 4.3). Up to 16 DRBs can be supported in NR.

#### *3.4.5.2.2 Security*

Security procedures for NR are similar to LTE; that is, all RBs are encrypted and SRBs are also integrity protected. Additionally, integrity protection is introduced for DRBs and both encryption and integrity protection are configurable per DRB. Another difference to LTE is that not all handovers involve a change of security key. The reason for this is that inter-DU handover within a gNB-CU does not change the security location and hence such a handover does not require a key change. To cater for handovers without key change, the security

configuration is not carried within the *reconfigwithSync* IE that triggers the handover procedure.

The *reconfigwithSync* is carried within the *CellGroupConfig* put together by the gNB-DU while the security configuration is provided by the gNB-CU directly in the RRC Reconfiguration message along with other possible configurations.

#### *3.4.5.2.3 Mobility*

Apart from the abovementioned differences, the mobility procedures in NR are similar to other cellular technologies. Measurement configuration provided by a gNB to a UE indicates which objects (frequencies) to measure, what quantities should be measured, and the trigger events for measurement reporting. When a measurement report is triggered, the UE sends the report with the measured results to the gNB. If the source gNB decides to hand over the UE, it indicates so to the target gNB via the Xn handover preparation procedure. The target then gNB reserves resources and provides the target cell configuration to source gNB, to be delivered to the UE over the source cell using the RRC Reconfiguration message including the *reconfigwithSync*.

The main differences between LTE and NR regarding mobility come from the beam-based measurement configurations. NR allows the reference signal to be SSB for the IDLE state, and SSB and/or CSI-RS for the CONNECTED state. The UE measures multiple beams of a cell and derives the cell quality from multiple beams. Measurement reports may contain beam results (beam identifier and optionally its measurement result) in addition to cell quantities.

#### *3.4.5.2.4 Radio Link Failure Recovery*

Reestablishment procedures similar to LTE are reused in NR for recovery after radio link failure (RLF). Unlike LTE, the RRC Reestablishment message itself does not contain any configuration information other than security parameters. This reestablishes security between the UE and the network. A subsequent RRC Reconfiguration message is then sent with security to provide all the configuration information. Another difference compared with LTE is that if the target cannot continue with the reestablishment, for example if it does not have the UE context, then it can use a fallback procedure and convert the reestablishment to a new RRC connection set up procedure without the UE having to initiate a RACH attempt again. Figure 3.4.11 shows the successful RRC connection reestablishment and Figure 3.4.12 illustrates the fallback message flow.

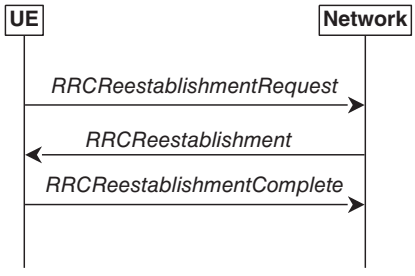



**Figure 3.4.11** Successful RRC connection reestablishment. (Source: Reproduced by permission of © 3GPP).

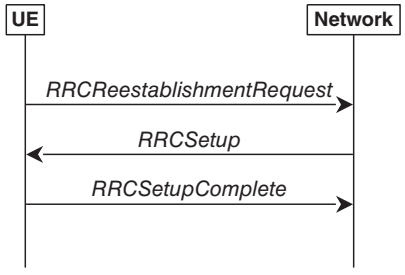



**Figure 3.4.12** RRC connection reestablishment with fallback. (Source: Reproduced by permission of © 3GPP).

#### *3.4.5.2.5 UE AS Capability Retrieval*

The UE capability enquiry procedure is used to retrieve and store the UE AS capability in the core network as part of the UE connection establishment procedure. This procedure is normally only run once per UE registration. For subsequent connection establishments, the core network provides the stored UE capability to gNB. During handover and dual connectivity establishment, the source node provides the UE capability to the target node during the preparation phase.

The stored UE capability can be released and updated with a new NAS registration procedure.

Figure 3.4.13 shows parts of an example message flow for UE capability handling and storage in the CN.

0. A UE is in RRC IDLE state.
1. The UE goes into CONNECTED state using the Connection Establishment procedure, for example at NAS registration.

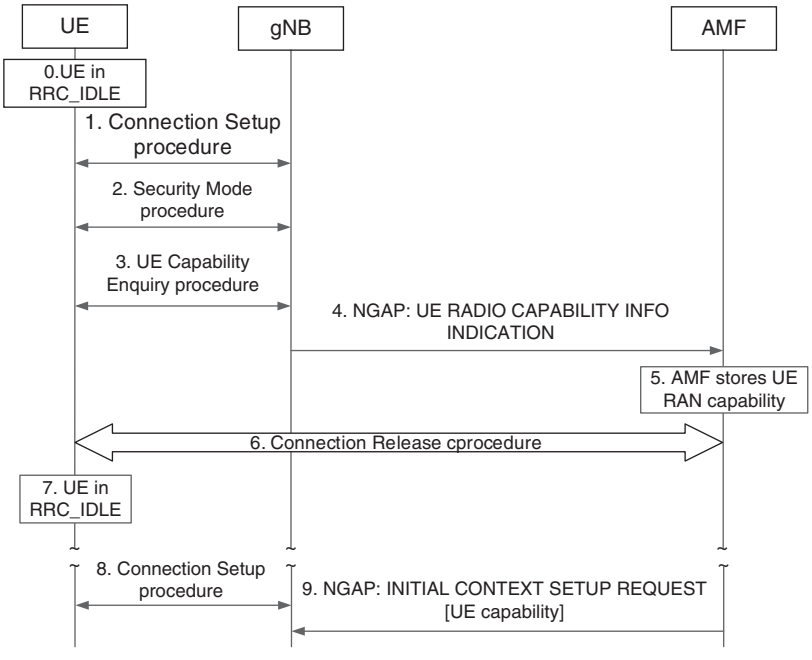



**Figure 3.4.13** UE AS capability enquiry, network storage, and retrieval.

2. AS security is configured.
3. The gNB retrieves the UE RAN capability from the UE using the UE Capability Enquiry procedure.
4. The gNB forwards the retrieved UE RAN capability to the AMF using the NGAP UE RADIO CAPABILITY INFO INDICATION message.
5. The AMF stores the UE RAN capability as part of the UE context stored in the AMF as a transparent container.
6. The UE RRC connection is subsequently released. The UE context in the gNB is released.
7. The UE goes to RRC IDLE state.
8. The UE starts a new RRC Connection Establishment procedure.
9. The AMF includes the stored UE RAN capability in the NGAP INITIAL CONTEXT SETUP REQUEST message sent to the gNB. The gNB can use this for RRC connection without having to retrieve it from the UE again.

#### *3.4.5.2.6 RRC INACTIVE State*

As discussed above, when a UE is in the INACTIVE state, the network and the UE store the previously used configurations. Additionally, the network provides to the UE a context RAN ID called I-RNTI.

Figure 3.4.14 shows the message flow for a UE moving to CONNECTED from INACTIVE that also involves a change of gNB.

0. A UE is in the CONNECTED state.
1. The network releases the connection and moves the UE to the INACTIVE state by providing an RRC Release message with suspend configuration. The suspend

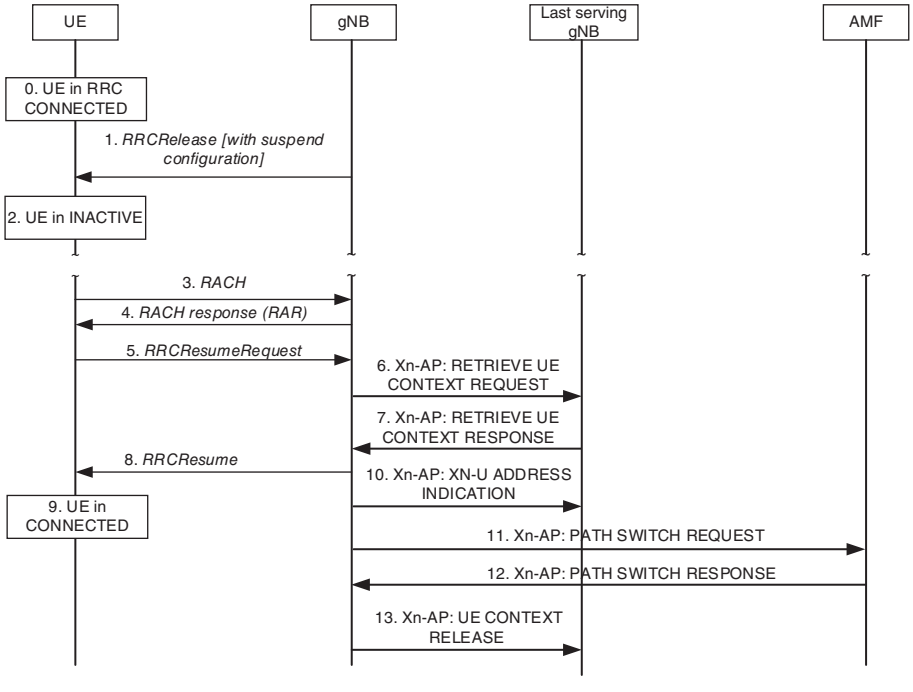



**Figure 3.4.14** Example message flow for state transition from INACTIVE to CONNECTED.

configuration includes a UE RAN ID called I-RNTI. It also includes security configuration to use when resuming the connection.
2. The UE moves to the INACTIVE state.
3. When the UE wants to move to RRC CONNECTED, either to initiate communication with the network or in response to RAN or CN paging message, the UE initiates the connection using the RACH procedure. This is the same as when the UE is moving to CONNECTED from IDLE.
4. The network responds with a RACH response, the same as for an IDLE to CONNECTED connection establishment procedure.
5. The UE requests the transition to CONNECTED using an RRC Resume Request message in RACH message 3 that contains the I-RNTI, authentication information ResumeMAC-I, and cause value.
6. If the resume request is received in a new gNB, the UE context has to be relocated from the previous gNB (i.e. the last serving gNB) to the current gNB. The new gNB requests the UE context using the NG-AP RETRIEVE UE CONTEXT REQUEST message.
7. After successful authentication of the UE based on the ResumeMAC-I, the previous gNB provides the UE context using the NG-AP RETRIEVE UE CONTEXT RESPONSE message.
8. The gNB indicates the successful resumption to the UE using the RRC Resume message. It is sent with security over DCCH as the security context in the UE was already provided when the UE went to INACTIVE. Hence the Resume message can provide additional UE configuration.
9. Upon successfully processing the RRC Resume message, the UE enters RRC CONNECTED.
10. The Xn-AP XN-U ADDRESS INDICATION message is used to provide the previous gNB with the address of the new gNB to forward data to.
11. The AMF is updated with the new gNB address with the NG-AP PATH SWITCH REQUEST message.
12. The AMF responds with the NG-AP PATH SWITCH RESPONSE message.
13. Upon successful completion of the resumption, the UE context in the previous gNB is released using the Xn-AP UE Context Release message.

If the resume request cannot be successfully processed by the gNB, for example if the UE context cannot be retrieved, a fallback procedure to a connection setup similar to the one discussed above for reestablishment is used.

Similar to the tracking area concept in the CN, an RNA is used in INACTIVE to keep track of the UE's location. Consequently, the UE performs an RNA update when crossing an RNA boundary. When downlink data arrives at the gNB for a UE in INACTIVE state, the gNB will trigger a RAN originated Paging message to the cells of the RNA. The UE, upon receipt of the Paging message, will initiate a transition to CONNECTED state using the RRC Resume Request message.

The UE also continues to perform CN tracking area updates while in INACTIVE.

#### *3.4.5.2.7 Broadcast Information*

System Information Broadcast (SIB) is primarily used to provide configuration information required for IDLE UEs to make an initial access to the network. The main components of the broadcast information are:

- Master information block (MIB), also called Minimum System Information (MSI), which provides information about the cell and how to acquire SIB1.
- SIB1, also called Remaining Minimum System Information (RMSI), which contains information necessary for UE to decide whether it can camp on and to initiate access in the cell. It also contains information about how to acquire the rest of the SIBs.
- Other SIBs (from SIB2 to SIB9 at the time of writing this book) contain function-specific information for cell reselection functions, Public Warning, etc.

The System Information update procedure is similar to LTE in that an indication about System Information change is provided to the UE, which triggers the System Information acquisition procedure. The main difference compared wiht LTE is that the System Information change indication and presence of Public Warning System message is provided directly in PDCCH itself, called Short Message (specified in 3GPP TS 38.331), rather than by the RRC Paging message.

In NR, the network may decide not to continuously broadcast the other SIBs (i.e. SIB2 to SIB9). This concept is called On Demand System Information. When an IDLE UE requires those SIBs that are not broadcast, it can make a request to the network using an RRC System Info Request message for the SIB that it is interested in. This can trigger the broadcast of the SIB in the cell.

Another piece of information that is broadcast by the network is UE Paging, which is used to inform an IDLE or INACTIVE UE that there is pending downlink data to the UE and to request the UE to initiate an RRC connection. A Paging message is sent on a specific paging occasion that the UE monitors and includes the UE identity. Paging could be triggered by the core network for UEs in IDLE or by RAN for UEs in INACTIVE.

#### *3.4.5.2.8 Slicing*

Slicing is a new functionality supported in 5GS to allow partitioning of the RAN and core network resources across the different slices. A slice could be used to provide a particular service or belong to an administrative domain. During a connection establishment, a UE indicates the slice it wants to connect to, which the network uses to select the appropriate core network node that supports the requested slice. Beyond this, most of the slicing functionality is internal to the gNB implementation, and the gNB scheduler should ensure that data for a slice uses only its allocated partition and that the slice does not exceed its allocated resources.

### 3.4.6 Summary

This section provided a brief overview of the air interface protocols of the NG-RAN and some of the basic procedures. NR draws most of the protocol aspects from LTE with some enhancements to support new 5G services and functions. For example, a new protocol layer SDAP was introduced to support the 5G QoS concept. The PDCP, RLC, and MAC protocol layers were enhanced (compared with LTE) to support lower latency and higher throughput more efficiently. The NR MAC also supports beam failure recovery procedure. Finally, the NR RRC supports a new state, INACTIVE, to provide for quick and efficient transition to CONNECTED state.

## References

3GPP Technical Specification 29.281 (2019). General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).
3GPP Technical Specification 37.324 (2020). E-UTRA and NR; Service Data Protocol (SDAP) specification. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).
3GPP Technical Specification 38.321 (2020). NR; Medium Access Control (MAC) protocol specification. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).
3GPP Technical Specification 38.322 (2020). NR; Radio Link Control (RLC) protocol specification. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).
3GPP Technical Specification 38.323 (2020). NR; Packet Data Convergence Protocol (PDCP) specification. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).
3GPP Technical Specification 38.331 (2020). NR; Radio Resource Control (RRC); Protocol specification. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).
3GPP Technical Specification 38.413 (2020). NG-RAN; NG Application Protocol (NGAP). 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

## 3.5 NR Physical Layer

*Alexei Davydov*

*Intel Corporation, Russia*

### 3.5.1 Introduction

In this section we focus on the NR physical layer, which is in large part designed to meet IMT-2020 requirements for 5G systems. In addition to improving the performance over LTE/LTE-A, NR offers fundamentally new technology components at the physical layer compared with LTE cellular systems. In particular, the following key physical layer enhancements are introduced in NR:

- **New services:** In addition to the traditional enhanced mobile broadband (eMBB) use case addressed by LTE systems, NR wireless access technology supports new services including IIoT, URLLC, and cellular V2X.
- **Scalable numerology:** In contrast to LTE, NR wireless access technology supports a variety of deployment scenarios (from macro cells to indoor hotspots), different bandwidth sizes (from several MHz up to several GHz), and carrier frequencies (ranging from sub-GHz to millimeter wave bands). For efficient operation of NR systems in such diverse scenarios, scalable numerology with variable subcarrier spacing and cyclic prefix duration is supported. The numerology of NR system can be selected to optimize the system performance to the actual deployment scenario determined, for example, by the cell radius, mobility, etc., and at the same time address some of the practical impairments, for example, radio frequency impairments like phase noise, which are more detrimental at millimeter wave bands.
- **Low latency and high reliability:** Unlike LTE, NR at the physical layer supports transmission providing low latency (with sub-millisecond delays) and high reliability (with error rates below $10^{-5}$). To facilitate low latency processing at the receiver, the physical channels of NR systems are accordingly designed. In particular, in addition to conventional subframe level scheduling (Type A mapping) supported by LTE-A, NR system allows special "mini-slot"-based transmission (Type B mapping) with more flexible scheduling in time. To facilitate "early" decoding at the receiver, mini-slot-based transmission also supports a special frontloaded position of the control channel and demodulation reference signals (DM-RS).
- **Waveforms:** Unlike LTE-A, which only supports the DFT-s-OFDM (Discrete Fourier Transform-Spread Orthogonal Frequency-Division Multiplexing) waveform in the uplink, the NR system adopts both DFT-s-OFDM and OFDM waveforms. The use of the OFDM waveform for the NR uplink provides better spectral efficiency, while the DFT-s-OFDM waveform addresses power efficiency issues in the uplink for coverage-limited scenarios.
- **Wider bandwidth:** The maximum bandwidth of NR system is extended beyond the maximum bandwidth supported by LTE-A. More specifically, the NR system in Release-15 supports a maximum of 400 MHz bandwidth to allow for efficient operation of the system in mmWave bands. The maximum bandwidth of NR is planned to be further increased in future releases, for example, to accommodate NR operation at carrier frequencies above 52.6 GHz.

- **Massive multiple-input and multiple output (MIMO):** mmWave frequencies supported by NR offer significantly more spectrum than the existing spectrum utilized by LTE-A in bands below 6 GHz. However, at mmWave frequencies the transmitted signal is subject to severe attenuation. To compensate for the encountered propagation losses, multi-antenna transmission techniques with beamforming at both gNB and UE are supported in the NR system. To support efficient beamforming at both gNB and UE, beam management procedures are introduced. The purpose of beam management schemes is to establish a highly directional communication link by using high-dimensional antenna arrays at both ends of the link.
- **Dynamic time division duplexing (TDD):** In TDD systems, time resources can be optimally allocated for downlink and uplink depending on traffic conditions. In LTE-A system, time resource adaptation is performed at the subframe level, comprising 14 OFDM symbols. In NR, downlink and uplink adaptation can be performed with the finer granularity of one OFDM symbol, which further improves the TDD system performance compared with LTE-A.
- **Forward compatibility support:** Similar to LTE, future releases of the NR system will support backward compatibility. However, unlike LTE, the NR system is also prepared for future technology evolution. In particular, co-existence of NR system transmissions to UEs from different releases can be performed through the use of a feature termed special "blank resources" (rate-matching resources). Blank resources can be flexibly configured for UEs of a previous 5G NR release (not supporting the new functionality of the latest NR release) in time and frequency, thus avoiding overlap between the transmissions to UEs from different releases.
- **Power efficiency:** To reduce power consumption, NR system avoids "always on" transmissions, for example, reference signals such as the cell-specific reference signal (CRS) used in LTE. Instead, demodulation of the physical channels in NR mainly relies on user-specific DM-RS. Moreover, for better power consumption efficiency at the UE, the bandwidth of the physical channels used by a UE can be adapted through the configuration of a new NR feature called bandwidth part (BWP). BWP allows the allocation of smaller bandwidths to a UE to reduce its power consumption or wide bandwidths to support very high data rates.
- **New Channel Coding:** NR supports new types of channel coding based on Low-Density Parity Check (LDPC) and Polar coding, which are used for the data and control channels, respectively. The key advantages of LDPC (compared with convolution turbo coding used in LTE-A) are improved performance with very low error floors, reduced decoding complexity and latency, better power and area efficiency, and support of multi-Gbps data rates. Polar coding yields better performance compared with the convolutional codes supported in LTE for control channel transmission.

In the present chapter we describe the NR physical layer, focusing on differences compared with LTE.

### 3.5.2 Waveform and Numerology

Unlike LTE, which supports different waveforms for downlink and uplink, Release-15 NR adopts the cyclic prefix (CP)-OFDM waveform for both downlink and uplink (3GPP TS

**Table 3.5.1** Scalable numerology supported by NR.

| μ | $2^{\mu} \cdot 15$ kHz | Frequency range | CP duration | Physical channels and reference signals |
|---|---|---|---|---|
| 0 | 15 | 1 | Normal | All |
| 1 | 30 | 1 | Normal | All |
| 2 | 60 | 1, 2 | Normal, extended | All |
| 3 | 120 | 2 | Normal | All |
| 4 | 240 | 2 | Normal | Only SS/PBCH |

38.211). The common waveform based on CP-OFDM simplifies system design and provides better co-existence for downlink and uplink transmissions in dynamic TDD systems; see Section 3.5.3 for a description of dynamic TDD for NR. Additionally, for coverage-limited scenarios in the uplink, Release-15 NR specifies DFT-s-OFDM with transmission of a single MIMO layer. The DFT-s-OFDM waveform has lower peak to average power ratio (PAPR) properties compared with CP-OFDM and, therefore, provides better UE power efficiency for uplink transmission.

Compared with LTE, NR is expected to offer more flexibility in supporting different deployment scenarios and carrier frequencies. In particular for Release-15, two frequency ranges denoted as FR1 (corresponding to carrier frequencies up to 7.125 GHz) (3GPP TS 38.101-1) and FR2 (corresponding to millimeter wave carrier frequencies up to 52.6 GHz) (3GPP TS 38.101-2) are introduced for NR operation. To support such a wide spectrum range, multiple numerologies – defined conveniently by their subcarrier spacing (SCS) – are defined. The actual SCS for NR is determined as $2^{\mu} \cdot 15$ kHz, $\mu = 0,1,\ldots,4$, where the scaling factor $2^{\mu}$ ensures alignment of slots and symbols in the time domain, which is important to efficiently enable TDD networks. In order to maintain similar overhead, the CP duration in NR is scaled down by a factor of $2^{\mu}$. The choice of the scaling parameter $\mu$ depends on several factors such as the deployment scenario, frequency range, radio frequency impairments, the type of service, etc. In particular, the narrower subcarrier spacing of 15 and 30 kHz, corresponding to scaling parameter $\mu = 0$ and $\mu = 1$, respectively, can be used in FR1 to accommodate larger delay spread channels inherent to the deployments with larger cell sizes. Numerologies with the wider subcarrier spacing of 60 and 120 kHz, corresponding to scaling parameter $\mu = 2$ and $\mu = 3$, respectively, can be utilized in FR2 to provide additional robustness of the transmitted signal to radio frequency impairments such as phase noise. The additional parameters related to the support of different numerologies in NR are summarized in Table 3.5.1.

### 3.5.3 Frame Structure

Similar to LTE, NR also supports the concept of frame in time (3GPP TS 38.211). A frame has a duration of 10 ms irrespective of the SCS used and is divided into two half frames, each with five subframes. Each subframe always has a duration of 1 ms and is further subdivided into one or multiple slots depending on the scaling parameter $\mu$ (see Figure 3.5.1). The number of OFDM symbols within a slot is always 14 and does not change with SCS.

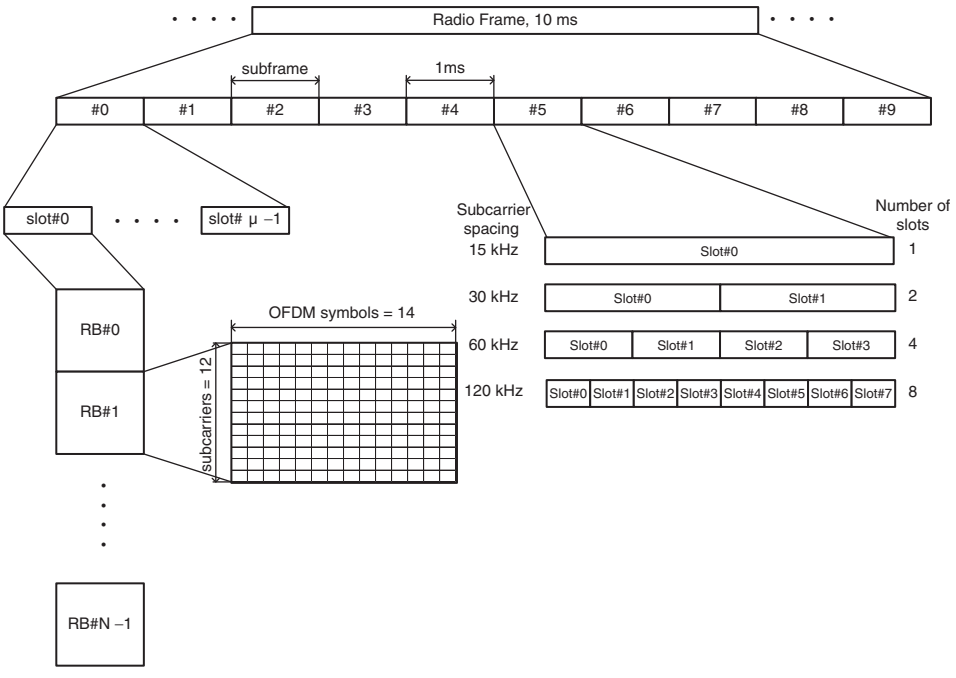



**Figure 3.5.1** Frame structure supported by NR.

Similar to LTE-A, a set of 12 subcarriers in the frequency domain across 14 OFDM symbols in the time domain is grouped into a physical resource block (PRB). The total number of the available PRBs depends on the SCS and the frequency range. The maximum supported bandwidth is 100 MHz in FR1 and 400 MHz in FR2. The supported bandwidth sizes and the maximum number of PRBs in Release-15 NR are summarized in Table 3.5.2.

In NR, PRB boundaries corresponding to different SCS values are frequency aligned relative to a common reference point termed "Point A" as illustrated in Figure 3.5.2. Such frequency alignment is supported for efficient multiplexing of transmissions corresponding to different numerologies in the same cell or for the same UE.

NR supports conventional slot-level based scheduling denoted in NR as Type A mapping. Slot-level transmission can only start at specific OFDM symbols, but has flexible duration

**Table 3.5.2** Supported bandwidth sizes in Release-15 NR for different subcarrier spacing.

| Subcarrier spacing | Minimum number of PRBs | Maximum number of PRBs | Minimum bandwidth, MHz | Maximum bandwidth MHz |
|---|---|---|---|---|
| 15 | 24 | 275 | 4.32 | 49.5 |
| 30 | 24 | 275 | 8.64 | 99 |
| 60 | 24 | 275 | 17.28 | 198 |
| 120 | 24 | 275 | 34.56 | 396 |
| 240 | 24 | 138 | 69.12 | 397.44 |

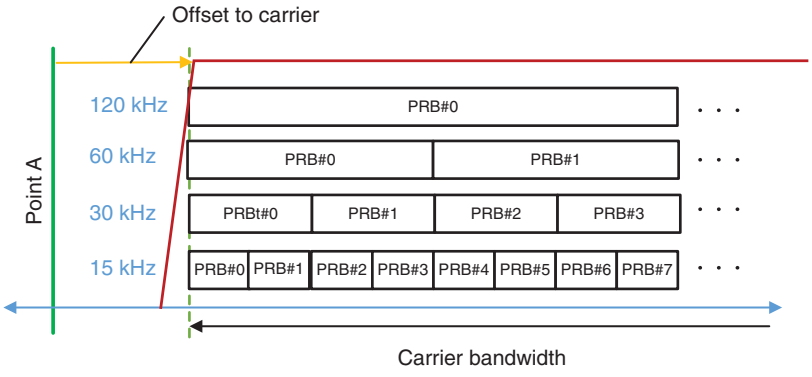



**Figure 3.5.2** PRB alignment for different numerologies.

up to 14 OFDM symbols within a slot. Type A mapping typically has a relatively long transmission time interval, which helps to reduce the overhead from reference signals and the control channel as well as to increase coverage. Conventional slot-level-based scheduling, however, is not efficient for all deployment scenarios. For instance, for 5G NR operation in unlicensed spectrum (NR-U), it is necessary to start transmission as early as possible after Listen-Before-Talk. In the case of mmWave, high payload transmission can be realized within just a few OFDM symbols due to the use of large bandwidth sizes. Finally, in the case of low latency transmission required for time-critical data applications, it is beneficial to start the transmission at any OFDM symbol without constraints. To optimize the system performance for such deployment scenarios, 5G NR also supports mini-slot-based transmission, denoted as Type B mapping, in addition to the slot-based scheduling. Mini-slot-based scheduling enables physical shared channel transmission to start at any OFDM symbol within a slot and to have flexible duration. To facilitate early decoding in mini-slot-based scheduling, the control channel and reference signals are located at the beginning of the transmission.

The frame structure in NR supports both TDD and Frequency Division Duplexing (FDD) operations. However, in contrast to LTE where downlink or uplink assignment is performed at the subframe level, downlink or uplink assignment is performed with the finer granularity of one OFDM symbol in NR (3GPP TS 38.213). In particular, each OFDM symbol in a slot is classified as "downlink," "flexible," or "uplink." In a downlink slot, the UE assumes that downlink transmissions can be performed only in "downlink" or "flexible" symbols. In an uplink slot, the UE only transmits in "uplink" or "flexible" symbols. The actual characteristics of the OFDM symbols are dynamically determined from the slot format indicator (SFI) field defining the link direction for one or more slots based on preconfigured values. By applying different slot format configurations, various types of the scheduling can be implemented for the TDD system, which is in contrast to LTE supporting only a limited set of predetermined uplink/downlink configurations. The FDD system can be also realized under the same framework by configuring symbols as all downlink or all uplink. Some examples of slot configuration for downlink and uplink transmission are shown in Figure 3.5.3.

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Slot with DL and UL | | | | | | | | | | | | | |
| D | D | D | D | D | D | D | D | D | F | F | F | U | U |
| Slot with DL only | | | | | | | | | | | | | |
| D | D | D | D | D | D | D | D | D | D | D | D | D | D |
| Slot with UL only | | | | | | | | | | | | | |
| U | U | U | U | U | U | U | U | U | U | U | U | U | U |

**Figure 3.5.3** Examples of slot configurations.

## 3.5.4 Synchronization and Initial Access

### 3.5.4.1 Downlink Synchronization Signals

The first step UE performs during the connection to the cell is synchronization. NR defines a special type of reference signal denoted as the SS/Physical Broadcast Channel (PBCH) block for that purpose (3GPP TS 38.211). The SS/PBCH block includes synchronization signals such as the Primary Synchronization Signal (PSS), Secondary Synchronization Signal (SSS), and PBCH. In contrast to LTE, each serving cell of the NR system can transmit more than one SS/PBCH block with different periodicities within the set of {5, 10, 20, 40, 80, 160} ms. However, irrespective of the used SS/PBCH block periodicity, each of the multiple SS/PBCH block transmissions in the cell is always confined to a duration of 5 ms. Multiple SS/PBCH blocks are required to support downlink beamforming on the synchronization signals in order to achieve better coverage, for example, in mmWave spectrum. The maximum number of SS/PBCH block transmissions depends on the configuration but cannot exceed a maximum of 8 SS/PBCH blocks in FR1 and 64 SS/PBCH blocks in FR2. A larger maximum number of SS/PBCH blocks is required in FR2 to support a larger number of downlink beams with narrower beamwidth compared with FR1. An SS/PBCH block can be transmitted using different numerologies depending on the FR: 15 or 30 kHz SCS can be used for FR1, and 120 or 240 kHz can be used for FR2.

The SS/PBCH position in the time domain is determined from the numerology, which in most cases can be uniquely derived from the frequency band. In the frequency domain, the position of the SS/PBCH block is configured by higher layers and, unlike LTE, does not necessarily coincide with the center of the system bandwidth (carrier raster). Moreover, the SS/PBCH block position is not necessarily aligned with the PRB grid. To reduce the UE complexity and power consumption associated with cell search in NR, the synchronization raster used for SS/PBCH block transmission (and therefore for NR cell search at the UE) is sparser than the carrier raster and also made dependent on the considered band. The position of the SS/PBCH block in the frequency domain is indicated relative to "Point A" (see Figure 3.5.4), which is the common reference point used for the alignment of PRB grids, reference signals, etc. (see also Figure 3.5.2).

Each SS/PBCH block in the time domain occupies four adjacent OFDM symbols and is transmitted over a maximum of 240 subcarriers, which corresponds to 20 PRBs. A detailed SS/PBCH structure is illustrated in Figure 3.5.5.

In NR three sequences are used for PSS modulation, where each PSS conveys partial information on the physical cell identity. For PSS modulation, a maximum length sequence – often denoted as M-sequence – with three cyclic shifts is used instead of the Zadoff–Chu sequences in LTE in order to provide better autocorrelation properties under

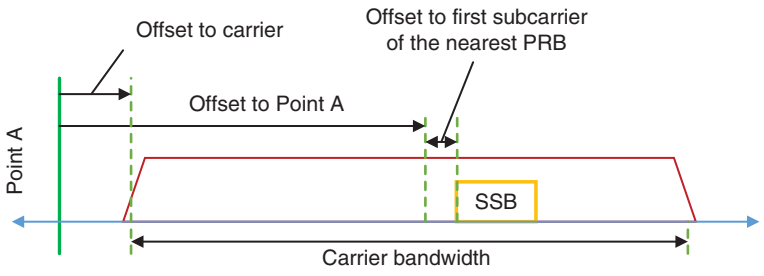
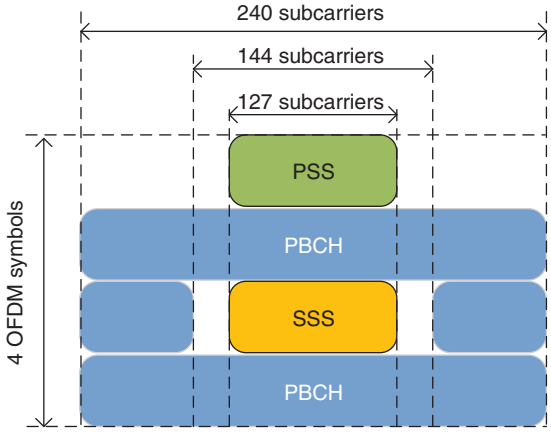



**Figure 3.5.4** SSB position indication in frequency domain.

**Figure 3.5.5** NR SS/PBCH block structure.




non-ideal synchronization assumptions. The SSS sequences in NR are also constructed based on M-sequences and convey information on the physical cell identity. More specifically, each SSS sequence is obtained by using a cyclic shift operation applied to the XOR function of two M-sequences. To support more dense deployments with larger number of cells, NR defines 1008 SSS sequences, which doubles the number of supported physical cell identities compared with LTE.

During the initial phase of the UE connection to the cell, after detecting the PSS/SSS synchronization signals the UE also decodes the PBCH. The PBCH of NR carries the remaining critical information such as SS/PBCH block index, OFDM index, slot index, system frame number (SFN) index, etc., and has a payload of 56 bits including cyclic redundancy check (CRC). The transmission time interval for PBCH is 80 ms, allowing for soft combining of PBCH transmissions over multiple PBCH occasions. To assist the demodulation of PBCH, DM-RS are transmitted together with the PBCH. The DM-RS for PBCH has a regular comb structure and its density is three resource elements per PRB. The actual position of the DM-RS resource elements in the frequency domain vary depending on the subcarrier shift, which is determined by the physical cell identity known to the UE after the PSS/SSS demodulation. The DM-RS of PBCH is modulated by a pseudo-noise random binary sequence and carries the index of the half radio frame as well as partial information on the index of the associated SS/PBCH block.

### 3.5.4.2 Random Access Channel

Similar to LTE, Release-15 NR enables a UE to access a cell by using a four-step RACH procedure, that is, synchronization signal detection, broadcast information acquisition, connection establishment using random access, and contention resolution. However, configuration of the signals and transmissions is different from LTE. In particular, for msg2 (Message 2) transmission from a UE to a gNB, the Physical Random Access Channel (PRACH) supports long and short sequences (preambles) (3GPP TS 38.211). For the long sequence four preamble formats are supported, mainly targeting FR1 deployment scenarios with large cells. These formats inherit the LTE PRACH design principles and support SCS of 1.25 or 5 kHz. For the short sequences nine preamble formats are supported in NR, targeting small/regular cells and indoor deployment scenarios. Short PRACH formats are based on a new structure and can be used for both FR1 and FR2. Short preamble formats are constructed by the composition of multiple shorter OFDM symbols per PRACH preamble and are transmitted using the same SCS as the shared channel, i.e. $2^{\mu} \cdot 15$ kHz, $\mu = 0,1,2,3$. This structure offers several benefits including possible support of Rx beam sweeping at gNB, robustness in the scenarios with time varying channels, single Fast Fourier Transform operation with uplink data, etc. The additional details of the supported PRACH formats are summarized in Table 3.5.3.

The generic structure of the PRACH preamble in NR is shown in Figure 3.5.6, where the number of sequences (SEQ), CP duration, and guard period (GP) duration depend on the

**Table 3.5.3** Supported PRACH preambles.

| PRACH formats | | | | | | | |
|---|---|---|---|---|---|---|---|
| Long sequence, 839 samples | | | | Short sequence, 139 samples | | | |
| Format | SCS, kHz | Duration, ms | Use case | Format | SCS, kHz | Duration, ms | Use case |
| 0 | 1.25 | 1 | LTE | A1 | $2^{\mu} \cdot 15$ | 0.0094 | Small cell |
| 1 | 1.25 | 3 | Large cell, >100 km | A2 | $2^{\mu} \cdot 15$ | 0.0188 | Normal cell |
| 2 | 1.25 | 4.3 | Coverage enhancement | A3 | $2^{\mu} \cdot 15$ | 0.0281 | Normal cell |
| 3 | 5 | 1 | High speed | B1 | $2^{\mu} \cdot 15$ | 0.0070 | Small cell |
| | | | | B2 | $2^{\mu} \cdot 15$ | 0.0117 | Normal cell |
| | | | | B3 | $2^{\mu} \cdot 15$ | 0.0164 | Normal cell |
| | | | | B4 | $2^{\mu} \cdot 15$ | 0.0305 | Normal cell |
| | | | | C0 | $2^{\mu} \cdot 15$ | 0.0404 | Normal cell |
| | | | | C2 | $2^{\mu} \cdot 15$ | 0.0667 | Normal cell |


**Figure 3.5.6** General structure of the RACH preamble.

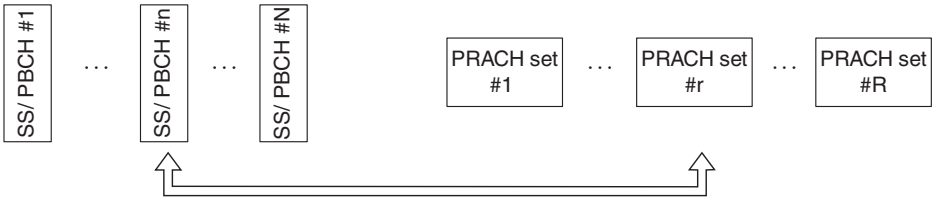



**Figure 3.5.7** Association of SS/PBCH blocks with PRACH.

PRACH format. In contrast to LTE, the length of an OFDM symbol in the PRACH preamble equals the length of a data OFDM symbol. Moreover, due to the OFDM symbol repetition in the new preamble formats, the last part of each OFDM symbol acts as a CP for the next OFDM symbol, thus providing several advantages as discussed above. Similar to LTE, each sequence of the PRACH preamble is based on a Zadoff–Chu sequence, which offers good autocorrelation and power efficiency properties.

For scenarios with multi-beam transmission of SS/PBCH blocks, it is necessary to apply the same beamforming for PRACH reception at the gNB as was used for the corresponding SS/PBCH transmission to achieve better coverage. As a result of this requirement, NR defines an association between one or multiple SS/PBCH blocks to one PRACH transmission resource occasion, so the gNB can use for PRACH reception the same beamforming as for SS/PBCH transmission (see Figure 3.5.7). A UE selects a PRACH resource for subsequent PRACH preamble transmission based on the detected SS/PBCH block by using a predefined association rule (3GPP TS 38.213). The PRACH resource configuration information is signaled in SIB1 (System Information Block 1).

In addition to synchronization and initial access, PRACH can be also used for other purposes, including new features not supported in LTE such as transition from RRC_INACTIVE state, establishment of time alignment at secondary cell addition, request for other system information (OSI), and beam failure recovery request, where some of the usages are supported using contention-free PRACH resources.

### 3.5.5 Downlink Control Channel

Similar to LTE, NR uses the PDCCH to transmit control information in the downlink (3GPP TS 38.211). For PDCCH reception, a UE can be configured with a control resource set (CORESET), which is similar to the control channel region in LTE, but has more flexible structure and position in time and frequency (see Figure 3.5.8). Such flexibility of the CORESET configuration allows addressing a wide range of new use cases and applications supported by NR systems. In particular, unlike LTE PDCCH, which always spans the entire system bandwidth, a CORESET can occupy certain subcarriers and OFDM symbols in a slot depending on higher layer configuration (e.g. RRC). To enable more efficient signaling of the CORESET configuration, the frequency domain resources for CORESET can be allocated with granularity of six REGs (Resource Element Groups) – each REG consists of 12 subcarriers over one OFDM symbol and a bundle of six REGs is termed a CCE (Control Channel Element). The CORESET position in time domain can be also flexible within a slot depending on the PDCCH search space set configuration. This is in contrast to LTE, where

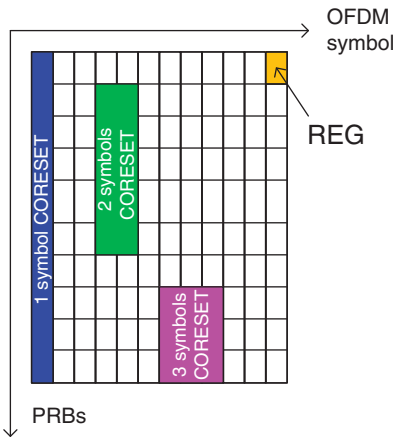



**Figure 3.5.8** CORESET transmission within a slot.

PDCCH can be only transmitted at the beginning of the downlink subframe. CORESET in NR can span one, two, or a maximum of three contiguous OFDM symbols within a slot to trade-off between low latency processing at the UE and PDCCH coverage, and to control the CORESET-induced overhead depending on the traffic load. When the CORESET spans multiple OFDM symbols, each control channel candidate is mapped to all symbols of the CORESET following time-first mapping.

To improve the channel estimation performance for PDCCH, multiple adjacent REGs are grouped into a REG bundle in the frequency domain and transmitted using the same precoding vector. To facilitate the use of efficient channel estimation techniques, 5G NR also supports common precoding for all contiguous REGs of the CORESET. To enable coherent processing of the received PDCCH, each REG bundle is transmitted with DM-RS enabling UE-specific beamforming of the control channel. PDCCH supports single-port DM-RS with its resource elements uniformly distributed in the frequency domain with a density of 3 REs per REG. To better control the trade-off between channel estimation performance and frequency diversity, the size of a REG bundle in NR can also be configured by the higher layer and can take the values of two, three, or six REGs.

Similar to LTE, PDCCH in NR can be transmitted by using 1, 2, 4, 8, or 16 CCEs depending on the aggregation level. A different number of CCEs is required to adapt the coding rate to the desired coverage of PDCCH and to support different payload sizes of the downlink and uplink control information (DCI/UCI). In NR, a UE can be configured to blindly monitor a number of PDCCH candidates of different DCI formats and different aggregation levels. Depending on the higher layer configuration, CCE-to-REG mapping for a CORESET can be interleaved to provide frequency diversity and interference averaging, or non-interleaved to support localized PDCCH transmission with a UE-specific precoding.

There are two types of CORESETs supported by NR: Common CORESET and UE-Specific CORESET (3GPP TS 38.213). The configuration of the common CORESET (denoted as CORESET 0) is indicated by the MIB and is used to transmit control information before RRC connection establishment. In the connected mode, a UE can be configured with one or multiple additional UE-specific CORESETs for further optimized control channel transmission.

**Table 3.5.4** DCI formats supported in NR (Release-15).

| Format Type | Direction | Description |
|---|---|---|
| 0_0 | Uplink | Grant for PUSCH |
| 0_1 | Uplink | Grant for PUSCH |
| 1_0 | Downlink | Scheduling of PDSCH |
| 1_1 | Downlink | Scheduling of PDSCH |
| 2_0 | N/A | Slot format indicator |
| 2_1 | N/A | No transmission indication/preemption indicator |
| 2_2 | Uplink | Power control commands for PUCCH and PUSCH |
| 2_3 | Uplink | Group power control for SRS |

Similar to LTE, NR supports two types of search spaces for PDCCH: common and UE-specific. A UE performs search of the possible DCI transmissions only in the specific time and frequency resources determined by the corresponding search spaces. The UE-specific search space is indicated to the UE after RRC connection establishment. However, the common search space, for example, for SIB1 transmission, is obtained in accordance with a predefined PDCCH candidates search scheme.

PDCCH is used to transmit control information included in the DCI, indicates scheduling information for the transmitted downlink data, and provides grants for uplink transmission. DCI can be also used to convey other information to the UE such as activation or deactivation of the semi-persistent physical downlink shared channel (PDSCH) transmission, transmission of power control commands for PUSCH, PUCCH and Sounding Reference Signals (SRS), switching of the BWP, initiating a random access procedure, and notifying one or more UEs of the slot format. Similar to LTE, the type of information transmitted in the DCI is determined from the RNTI parameter of the DCI. Table 3.5.4 summarizes the DCI formats supported in NR (Release-15) (3GPP TS 38.212).

### 3.5.6 Uplink Control Channel

NR uses the PUCCH to transmit uplink control information (UCI) from a UE to a gNB (3GPP TS 38.213). UCI consists of the following information:

- HARQ feedback, which is used to convey ACK/NACK information in response to PDSCH transmissions to the UE;
- CSI and beam reporting indicating the quality of the downlink channel, preferred digital precoding as well as information related to analog beamforming;
- SR.

In contrast to LTE, which typically supports PUCCH transmission at the edges of the carrier bandwidth with a time span of one uplink subframe, the location and structure of the PUCCH in 5G NR is more flexible in time and frequency. Moreover, up to two PUCCH transmissions from a UE are supported in one slot. Such configuration flexibility for PUCCH has a number of advantages including support of UEs with smaller bandwidth capabilities,

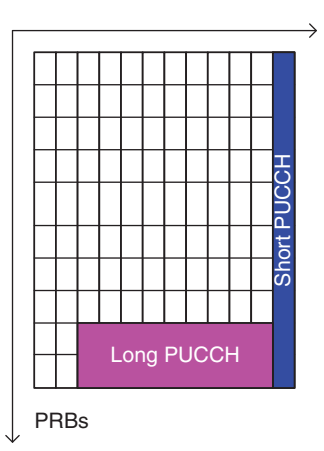



**Figure 3.5.9** Short and long PUCCH structures.

more efficient usage of the available physical resources depending on the target coverage, and capacity and/or support of low latency transmissions.

There are two PUCCH structures supported by NR: one with short and one with long duration; see Figure 3.5.9 for an illustrative example. More specifically, NR PUCCH with short duration spans one or two symbols in a slot, and can be multiplexed in the time domain with downlink or uplink transmissions. This PUCCH format can be used to support low latency use cases such as URLLC. In such use cases, PUCCH can be transmitted in the last OFDM symbol of a slot to enable fast HARQ-ACK feedback.

Similar to LTE, to provide robust transmission and high capacity of UCI, NR also supports PUCCH transmission with long duration, where multiple OFDM symbols are allocated for PUCCH to ensure the desired coverage and size of the UCI payload. In total, NR defines five different formats for PUCCH. The number of bits, PUCCH duration, and the number of PRBs that can be used for PUCCH transmission in NR are summarized in Table 3.5.5.

**Table 3.5.5** Supported PUCCH formats in 5G NR.

| Format | Type | Duration | Number of bits in UCI | Number of PRBs | Types of UCI | Summary of key elements |
|---|---|---|---|---|---|---|
| 0 | Short | 1–2 | 1, 2 | 1 | HARQ, SR | Cyclic shift (CS) per UCI<br>UE multiplexing using CS |
| 1 | Long | 4–14 | 1, 2 | 1 | HARQ, SR, CSI | TDM of UCI and DM-RS<br>UE multiplexing using CS and OCC |
| 2 | Short | 1–2 | >2 | 1–16 | HARQ, SR, CSI | FDM of UCI and DM-RS |
| 3 | Long | 4–14 | >2 | 1–6, 8–10, 12, 15, 16 | HARQ, SR, CSI | TDM of DM-RS and UCI<br>UCI using DFT-s-OFDM |
| 4 | Long | 4–14 | >2 | 1 | HARQ, SR, CSI | TDM of DM-RS and UCI<br>Pre-DFT OCC multiplexing of UEs |

The long PUCCH has variable duration in time domain, that is, the actual number of OFDM symbols for long PUCCH in a slot is configurable and can range from 4 to 14. For short PUCCH formats, the transmission of the UCI is always confined within one or two OFDM symbols. Short- and long-duration PUCCHs formats can support different payload sizes to carry the UCI. A particular PUCCH format can be selected depending on the use case and target deployment scenario.

In PUCCH formats 1, 3, and 4, to maintain low PAPR, DM-RS symbols and symbols used for UCI transmission are time-division multiplexed (TDM). Depending on the number of used PRBs, the DM-RS of PUCCH is modulated using either a low PAPR Zadoff–Chu sequence or low PAPR computer-generated Quadrature Phase Shift Keying (QPSK) sequence (CGS). In PUCCH format 2, the DM-RS is frequency domain multiplexed with subcarriers carrying the UCI. For short PUCCH format 0, the transmission of the UCI bits is performed via sequence selection without DM-RS transmission. In particular for HARQ-ACK feedback in PUCCH Format 0, for the configured base sequence, a specific cyclic shift is determined by the HARQ-ACK. For SR the configured base sequence is transmitted using the preconfigured cyclic shift on the preconfigured time and frequency resource. Cyclic shifts or orthogonal cover codes can be also used for PUCCH formats 0, 1, and 4 to support multi-user multiplexing on the same time and frequency resources.

A UE can be configured with a PUCCH resource set comprising multiple PUCCH resources of the same or different PUCCH formats, which can be used to carry HARQ-ACK in response to dynamically scheduled PDSCH transmissions (see Figure 3.5.10 for an illustrative example). UE determines a PUCCH resource for UCI transmission based on the UCI size and a 3-bit field in DCI (PUCCH resource indicator – PRI field). For example, for multiplexing of dynamic HARQ-ACK/SR and CSI with overlapping PUCCH resources if the UE is configured with more than one PUCCH resource set, one PUCCH resource set is determined based on the total UCI payload size, while the PUCCH resource within the selected PUCCH resource set is determined based on the PRI in the scheduling DCI.

Unlike LTE, CSI reporting on PUCCH in NR is accomplished in one slot. To support efficient usage of allocated PUCCH resources, CSI content is divided into two parts. The first part of CSI (i.e. rank indicator) has low and fixed payload size to facilitate UCI decoding at gNB without blind decoding. The payload size for the second CSI (i.e. precoding matrix indicator [PMI]) is variable and depends on the CSI content transmitted in the first part. Similar to LTE, HARQ-ACK and the first part of the CSI are transmitted on OFDM symbols around DM-RS symbols to provide more reliable transmission, while

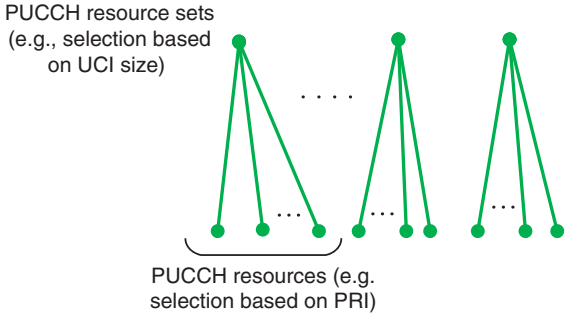



**Figure 3.5.10** PUCCH resource sets.

the remaining resource elements are allocated for the transmission of the second CSI part. To facilitate early decoding of the UCI, mapping of UCI to PUCCH resources is performed in a frequency-first and time-second manner.

### 3.5.7 Reference Signals

In LTE, most of the downlink transmission schemes are based on the always available downlink CRS, which are transmitted by evolved Node B (eNB) irrespective of the actual data traffic. Such an approach for the reference signal transmission has several drawbacks including high overheads, low energy efficiency at the base station, and excessive inter-cell interference.

Unlike LTE, NR was designed to support more efficient resource usage and therefore the reference signals are transmitted only when needed. This design enables lower overheads, greater base station power savings, and reduced levels of inter-cell interference. The time and frequency positions of the reference signals in NR are also configurable, which is in contrast to the CRS design in LTE, which only allows for wideband CRS transmission at predetermined positions.

The following types of reference signals are supported in NR (3GPP TS 38.211):

- CSI-RS: for CSI acquisition, beam management, and reporting.
- DM-RS: reference signals that are UE-specific and designed in the same way for data and control channels; DM-RS transmission is confined to a set of PRBs where the corresponding physical channel is transmitted.
- SRS: signals to assist reciprocity-based precoding in the downlink as well as to acquire CSI for the uplink.

NR also supports new types of reference signals not supported by LTE (3GPP TS 38.211):

- CSI-RS for tracking: tracking reference signals for time and frequency offset tracking and estimation of the channel delay spread and Doppler spread.
- PT-RS: phase tracking reference signals for fine time domain granularity phase estimation that may occur, e.g. due to phase noise impairments in mmWave bands.

To facilitate orthogonal multiplexing and sharing of reference signals among UEs with different bandwidth capabilities, the modulation sequence for reference signals in NR is typically performed starting from common point A (see Figure 3.5.10). The actual sequence for the UE is determined based on the offset parameter $k_0$ signaled to the UE as part of the reference signal configuration. For forward compatibility, the range of the offset parameter $k_0$ is also selected to be a relatively large value to accommodate the introduction of UEs with larger bandwidth capabilities.

In the following text, each NR reference signal is discussed in more detail.

#### 3.5.7.1 CSI-RS

Similar to LTE-A, CSI-RS in NR can be used for downlink CSI acquisition (3GPP TS 38.211). Based on channel measurements from CSI-RS, UE reports to gNB the preferred parameters for downlink transmission, for example, number of MIMO layers, precoding, and modulation and coding scheme. When CSI-RS is used for CSI feedback purposes, the time and

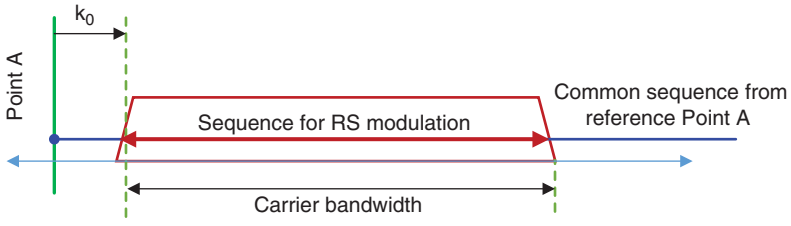



**Figure 3.5.11** Modulation sequence for reference signal.

frequency density of CSI-RS is low incurring small overheads. More specifically, CSI-RS in NR has a density of one resource element per PRB per antenna port. Additional overhead reduction for CSI-RS is achieved by transmitting CSI-RS every other PRB with total overheads of 0.5 resource elements per PRB per antenna port.

Unlike LTE-A, NR offers high degree of flexibility when configuring CSI-RS. Depending on the number of antenna ports, CSI-RS may be constructed by aggregation of one or multiple basic units transmitted within a slot (see Figure 3.5.12). Multiple antenna ports within each basic unit of the CSI-RS can be supported by using different orthogonal cover codes.

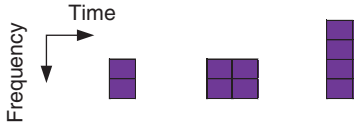
In the time domain, CSI-RS may start at any OFDM symbol of a slot as long as there is no collision with other reference signals. Depending on the number of configured antenna ports, CSI-RS spans 1, 2, or 4 OFDM symbols. CSI-RS can be periodic, semi-persistent, or aperiodic. Periodic and semi-persistent CSI-RS can be transmitted with periodicity of {4, 5, 8, 10, 16, 20, 32, 40, 64, 80, 160, 320, 640} slots depending on the configuration. Semi-persistent CSI-RS is transmitted based on the activation/de-activation command from the MAC control element, while aperiodic CSI-RS transmission is based on the DCI and is limited to a single occasion unlike periodic and semi-persistent CSI-RS, which are typically transmitted over multiple instances.

Unlike LTE, which only supports wideband CSI-RS, the CSI-RS bandwidth can be configurable and indicated to the UE in contiguous units of four PRBs. CSI-RS transmission can be UE-specific or cell-specific and supported by the unified UE-specific configuration procedure of CSI-RS. CSI-RS is modulated by QPSK using a pseudo-random sequence that depends on the configuration parameter and the OFDM symbol index. For cell-specific CSI-RS transmission that enables CSI-RS sharing among UEs possibly supporting different bandwidth, the reference for the CSI-RS sequence mapping is point A (cf. Figure 3.5.11).

Unlike CSI-RS for larger number of antennas, CSI-RS with one antenna port has a uniform pattern and has a density of three resource elements per PRB. The increased density of CSI-RS is required to support Layer 1 reference signal received power (L1-RSRP) measurements used for the acquisition of the DL transmission beams from gNB during the beam management procedure.

Unlike LTE, CSI-RS in NR can be also used for time-frequency tracking with periodic (with periodicity of 10, 20, 40, or 80 ms) and aperiodic CSI-RS configurations (3GPP TS

**Figure 3.5.12** Basic units for CSI-RS.



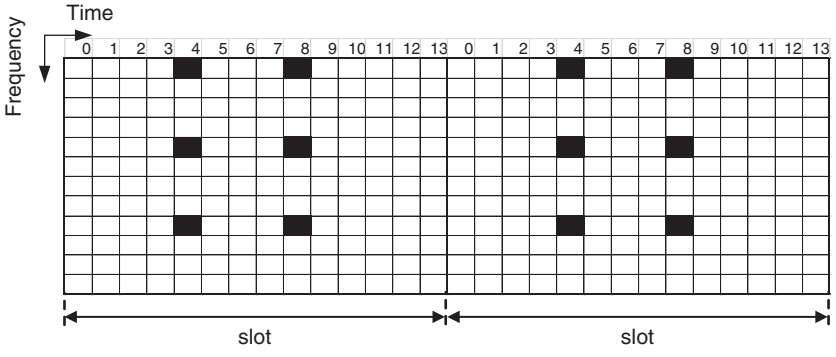



**Figure 3.5.13** CSI-RS for time-frequency tracking.

38.214). In this use case, CSI-RS can be only transmitted with a single antenna port over two OFDM symbols within a slot using the same set of subcarriers. Depending on the FR, one or two slots (see Figure 3.5.13), can be used for CSI-RS transmission to provide a better trade-off between overhead and accuracy for synchronization as well as for channel delay spread and Doppler spread estimation.

### 3.5.7.2 DM-RS

The demodulation of the physical channels in NR is performed based on the channel estimated from DM-RS (3GPP TS 38.211). DM-RS are precoded in the same way as a physical channel and confined within the scheduled resource elements of the corresponding channel. Unlike LTE-A, for the physical downlink shared channels (PDSCH or PUSCH), NR offers more flexible DM-RS structure supporting a wide range of use cases. In particular, two DM-RS types denoted Type I and Type II can be used to provide trade-off between frequency domain density and DM-RS overhead. In particular, DM-RS Type I supports two code division multiplexing (CDM) groups, while DM-RS Type II supports three CDM groups per OFDM symbol as shown in Figure 3.5.14. For both DM-RS types, up to two DM-RS ports can be multiplexed in one CDM group by using orthogonal cover codes in the frequency

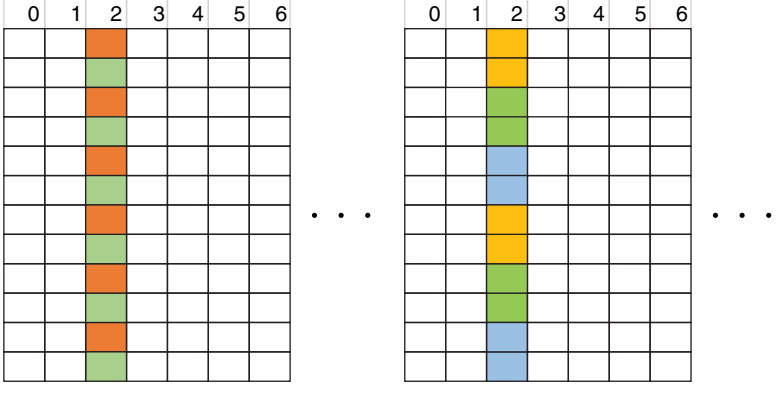



**Figure 3.5.14** DM-RS Type I and Type II.

domain, thus providing a total DM-RS multiplexing capacity of four and six DM-RS ports per OFDM symbol, respectively. For scenarios with massive multi-user MIMO (MU-MIMO) or single-user MIMO (SU-MIMO) transmission with a large number of MIMO layers, the number of DM-RS ports can be doubled in NR by using the DM-RS configuration with two adjacent symbols and orthogonal cover codes in the time domain. DM-RS sequences corresponding to different ports are orthogonal due to the use of different subcarriers (e.g. following a comb pattern) or by using different orthogonal cover codes. The actual number of orthogonal antenna ports depends on the type of DM-RS with a maximum of 12 DM-RS antenna ports, allowing efficient MU-MIMO transmission schemes for up to 12 UEs.

In the time domain, the DM-RS structure is also configurable and may include one or multiple DM-RS occasions with the first DM-RS symbol always located at the beginning of the transmission to enable early channel estimation processing. For efficient PDSCH or PUSCH transmission in scenarios with medium and high mobility, NR supports a maximum of three additional DM-RS occasions within a slot, thus allowing more frequent channel estimation updates. The channel estimation using DM-RS in downlink can be performed based on the unit of precoding resource group, which may include two, four, or all contiguous scheduled PRBs. The actual physical resource block group (PRG) size in NR can be configured for the UE by higher layers, for example, RRC, or dynamically indicated by the DCI. For uplink transmission, the PRG always corresponds to all scheduled PRGs.

The DM-RS sequence for the CP-OFDM waveform is modulated by QPSK using Gold sequences and for the DFT-s-OFDM waveform by low PAPR constant amplitude zero autocorrelation sequences. Depending on the number of allocated PRBs for uplink transmission, DM-RS is modulated by QPSK using either a low PAPR Zadoff–Chu sequence or a low PAPR CGS similar to LTE. The constant modulus property of the sequences in frequency domain guarantees perfect autocorrelation in the time domain for DM-RS and therefore optimal channel estimation performance.

#### 3.5.7.3 PT-RS

PT-RS is used for fine granularity tracking of the phase in time, which helps to suppress RF impairments such as phase noise (3GPP TS 38.211). Since the phase noise increases with carrier frequency, the use of PT-RS is more beneficial for high carrier frequencies, such as mmWave. PT-RS in NR has low density in the frequency domain but high density in the time domain. It is supported for both downlink (associated with PDSCH) and uplink (associated with PUSCH). To facilitate phase tracking at the receiver, the antenna port of PT-RS is always associated with one antenna port of DM-RS. For PUSCH transmission the association with the DM-RS port is indicated by the DCI, while for PDSCH transmission the association is fixed in the NR specification.

For the CP-OFDM waveform, PT-RS has uniform structure in both frequency and time. In the frequency domain, PT-RS resource elements can be present every second or every fourth PRB depending on the number of allocated PRBs. In the time domain, PT-RS can be transmitted every OFDM symbol, every second OFDM symbol (see Figure 3.5.15), or every fourth OFDM symbol depending on the modulation and coding scheme (MCS). The PT-RS sequence for CP-OFDM is derived from the DM-RS symbol of the associated DM-RS port.

For DFT-s-OFDM, PT-RS is transmitted using multiple PT-RS groups, which are multiplexed with the PUSCH symbols prior to DFT spreading. The number of PT-RS groups as

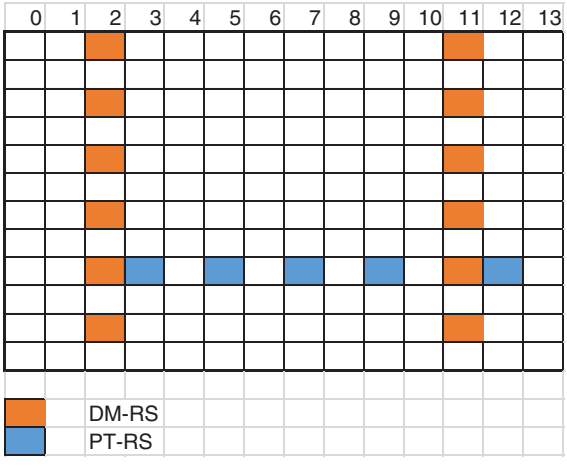



**Figure 3.5.15** PT-RS within PRB for CP-OFDM.

well as the number of PT-RS symbols per group can be adapted based on the number of allocated PRBs for PUSCH transmission. In DFT-s-OFDM, the PT-RS resource elements are modulated by $\pi/2$ BPSK to guarantee power-efficient transmission with low PAPR.

#### 3.5.7.4 SRS

Similar to LTE, NR supports SRS, which are used to assist link adaptation and precoding matrix selection for uplink transmission as well as to provide downlink CSI for reciprocity-based precoding in TDD systems (3GPP TS 38.211). SRS for DL CSI acquisition supports antenna switching when UE has fewer transmit chains than receiving antennas (3GPP TS 38.214). Compared with LTE, SRS in NR also offers new functionality such as non-codebook-based precoding for uplink and uplink beam management. For non-codebook-based precoding, multiple SRS signals (SRS resources) are transmitted by the UE using different precoding. gNB indicates the actual precoding for PUSCH transmission based on SRS measurements through signaling of the indices of the selected SRS resources (SRS resource index – SRI). For SRS used for beam management, the actual beam indication for PUSCH and PUCCH transmission is also accomplished by similar principles using SRI.

Unlike LTE, the physical resource used for SRS in NR is configured in a UE-specific manner. In the time domain, SRS can be transmitted using one, two, or four OFDM symbols, which can be located in the last six symbols of a slot. Multiple SRS symbols can be utilized at the gNB to support uplink beam management or to provide coverage extension by coherent processing of multiple SRS symbol transmission.

Similar to other reference signals in NR, SRS support periodic, semi-persistent, and aperiodic transmission, where semi-persistent SRSs are transmitted based on MAC activation command.

### 3.5.8 Beam Management

mmWave bands offer significantly more spectrum than the conventional sub-6 GHz bands currently used by LTE/LTE-A cellular systems and therefore the mmWave spectrum is considered as a key enabler of multi-Gbps data rates for the NR system in certain deployment

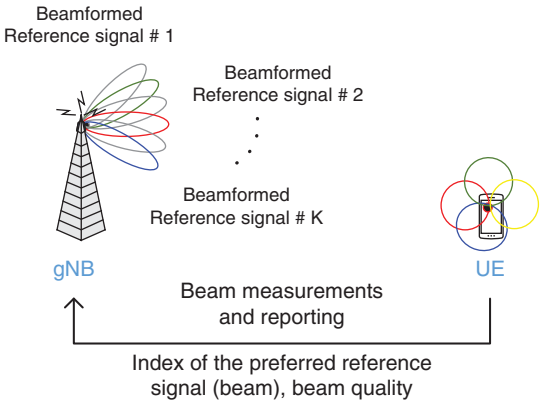



**Figure 3.5.16** Beam measurement and reporting.

scenarios. Such a spectrum, however, makes propagation conditions more severe than at the conventional lower frequencies. In particular, signals in the mmWave spectrum are subject to higher attenuation due to outdoor-to-indoor penetration losses, blockage, oxygen absorption, etc. To compensate for these impairments, highly directional multi-antenna transmission techniques with beamforming at both gNB and UE can be used. To establish such highly directional transmission links, fine alignment of the transmitter and receiver beam pairs is required. Such beam alignment in NR is achieved through a set of operations denoted beam management procedures, which include Tx/Rx beam pair acquisition, beam measurement and reporting, and beam indication for the transmission. In beam acquisition, a UE finds one or more Tx/Rx pairs of beams that can be used for subsequent communication, based on the transmitted beamformed reference signals (e.g. SS/PBCH, CSI-RS) from gNB (see Figure 3.5.16). The UE measures the link quality for the corresponding beam pairs using L1-RSRP (3GPP TS 38.215) and reports the measurement results along with the selected index of the beamformed reference signal back to the gNB.

Based on the reported information, the gNB assigns a transmission beam to a physical channel (PDSCH or PDCCH) or other reference signal by using the beam indication procedure. The indication of the used beam is supported through quasi colocation signaling (3GPP TS 38.214), which establishes the connection between antenna ports with respect to the spatial channel properties. In particular, the antenna port of the reference signal used for beam management, for example, SS/PBCH or CSI-RS, can be quasi colocated with the antenna port of the corresponding physical channel, for example, PDCCH, PDSCH, or other reference signal (see Figure 3.5.17). For example, the antenna port of PDCCH

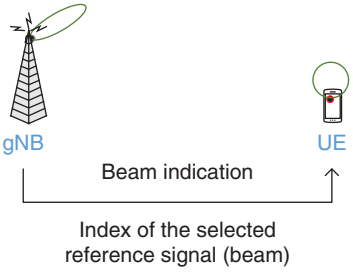



**Figure 3.5.17** Beam indication for the physical channel.

can be associated with the antenna port of one of the beamformed reference signals (SS/PBCH or CSI-RS) by using the index of the corresponding reference signal. Based on the corresponding quasi colocation association, the beam obtained by the UE during beam pair acquisition can be used for PDCCH reception. A similar principle is also used for PDSCH transmission, where the actual beam can be dynamically indicated to the UE by DCI signaling. To provide a sufficient time to switch the receive beam at the UE in accordance with the gNB indication, a time gap of several OFDM symbols between the last PDCCH symbol and the first PDSCH symbol is applied.

NR also supports beam indication for uplink transmission. In particular, for the UE supporting beam correspondence, that is UE implementation allowing the reuse of the beam acquired by the UE in downlink for uplink transmission, beam indication for uplink transmission can be performed through the index of the downlink reference signal. The beam indication for uplink can be also performed independently from downlink by indicating the index of the SRS transmitted by the UE for beam acquisition purposes.

In certain scenarios, the highly directional link between gNB and UE may fail, for example due to channel blockage. This event in NR is denoted as beam failure. To avoid time-consuming cell reselection procedures involving higher layers, NR specifies a beam failure recovery procedure at the physical layer, where a UE in the event of beam failure acquires an alternative beam pair without invoking RLF (3GPP TS 38.213). Beam failure recovery in NR includes the following steps:

- Beam failure detection
- New beam identification
- Beam failure recovery request
- Beam failure recovery response.

The detection of beam failure in 5G NR is performed based on the measurements of the periodic reference signals, that is, SS/PBCH or CSI-RS, transmitted with the same beamforming as for the downlink control channel. Similar to RLF in LTE, beam failure detection in 5G NR is declared by the UE based on block error rate calculation (BLER) for the PDCCH. If the measured quality becomes low, the UE declares beam link failure and proceeds to the acquisition of an alternative candidate beam pair denoted as new beam identification. After finding a new beam pair based on L1-RSRP, the UE transmits a beam failure recovery request (BFRQ) message using a preconfigured PRACH resource to the gNB. The gNB then transmits a beam failure recovery response (BFRP) to the UE.

### 3.5.9 Channel Coding and Modulation

In NR, due to the demand for Gbps-level user throughput and lower implementation complexity, the conventional Convolution Turbo Codes (CTCs) supported in LTE-A for data channel transmission were replaced by LDPC codes (3GPP TS 38.212). The NR LDPC codes are represented by a special parity check matrix that has a quasi-cyclic structure. The parity check matrix can be also represented via a bipartite graph with variable nodes corresponding to the columns and check nodes corresponding to the rows of the matrix. If the entry in the parity check matrix has a non-zero element, the corresponding variable and parity nodes in the graph are connected with the edge.

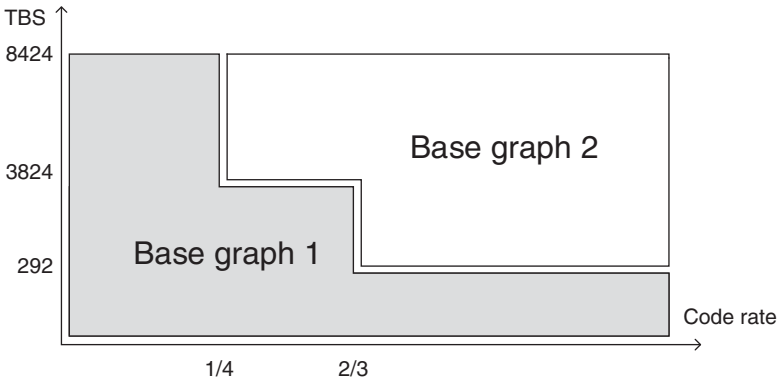


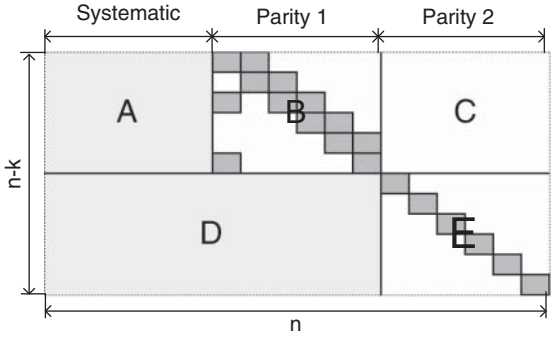

**Figure 3.5.18** Code rate and transport block size scenario for LDPC base graphs.

LDPC codes are decoded by exchanging information between variables and parity checks inside a graph in an iterative manner, where the messages which are exchanged between nodes represent probability distributions for the associated bits. At each iteration, the messages are further processed and updated at the nodes of the graph. Unlike CTCs in LTE-A, which always decode the received code block under the assumption of the low code rate of 1/3, the design of the NR LDPC codes allows for the decoding of high rate code blocks. In this case, the bits that are not transmitted are not accounted for in the decoding process, thus reducing the complexity of LDPC decoding as the code rate increases. Such LDPC code design makes it easier to support higher data rates with reasonable implementation complexity.

The parity check matrix of the NR LDPC codes is defined by a smaller base matrix and each entry of the base matrix represents either a $Z \times Z$ zero matrix or a shifted $Z \times Z$ identity matrix. For better performance optimization and improved decoding latency for the supported range of block lengths and code rates, LDPC in NR system is defined through two base graphs, with the first base graph designed to support larger code blocks and higher rates and the second base graph targeting smaller code blocks and lower rates as shown in Figure 3.5.18.

The general structure of the LDPC parity check matrix supported in NR is illustrated in Figure 3.5.19.

**Figure 3.5.19** Illustration of the parity check matrix for LDPC codes.



In Figure 3.5.19 the parity check matrix consists of five submatrices which are denoted as A, B, C, D, and E:

- Matrix **A** corresponds to the systematic bits.
- Matrix **B** with dual diagonal structure corresponds to the first set of parity bits denoted as parity 1, where the first column in matrix B has weight of 3 and the other columns have weight of 2.
- Matrix **C** corresponds to the all-zeros matrix.
- Matrix **D** corresponds to the incremental redundancy part.
- Matrix **E** is the identity matrix corresponding to the second set of parity bits.

NR base graph 1 has 317 edges with the base matrix size of $46 \times 68$ corresponding to 22 systematic columns. NR base graph 2 has 197 edges with matrix dimensions $42 \times 52$ corresponding to 10 systematic columns. For both base graphs from row 20 (counting from 0), consecutive rows do not overlap to ensure the row orthogonality required for more efficient decoding. The maximum supported effective code rate for LDPC code is 0.95, which is in contrast to LTE-A CTC supporting a maximum coding rate of 0.931, that is, LDPC codes can offer higher coding gains than CTC without error floors.

For the control channels (DCI, PBCH, UCI) in NR, Polar coding is adopted, which has better performance compared with the convolution codes used in LTE-A at the cost of increased decoding complexity (3GPP TS 38.212). Polar codes are used for coding the downlink and uplink control information (DCI/UCI) as well as the broadcast channel MIB. Depending on the payload size of the control channel, different types of Polar codes with different number of CRC bits can be used (see Figure 3.5.20).

The key idea of Polar codes relies on the "polarization" of binary input channels into "low"- and "high"-quality channels after linear transformation, where the number of "high"-quality channels will be determined by the channel capacity. In practice, such "polarization" of the channels is exploited by transmitting information bits using "high"-quality channels, while predetermined "frozen" bits (e.g. fixed to 0) are transmitted for the "low"-quality channels. The knowledge of the "frozen" bits is used at the receiver to correct the errors in the received signal. The linear transformation for Polar codes is determined by a special square matrix (often denoted as Kernel), which typically has a dimension of $2^N$, $N = 5, 6, \ldots, 10$. In NR the sequence of indices determining the set of "frozen" bits is common for all supported lengths of $2^N$.

The structure of Polar codes used in NR is shown in Figure 3.5.21. First, depending on the length of the information sequence the input information sequence can be segmented into multiple blocks. Code block segmentation is followed by CRC attachment, with CRC in Polar coding being used to detect false decoding (as in conventional procedures) as well as to improve the error rate performance through CRC-aided list decoding algorithms. Unlike

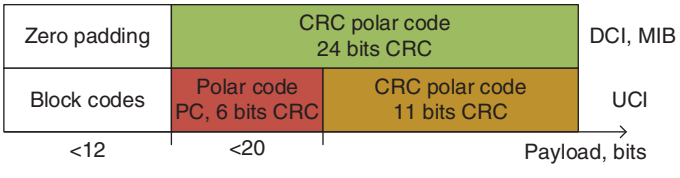



**Figure 3.5.20** Polar coding control information in NR.

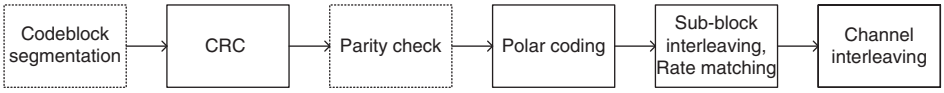



**Figure 3.5.21** Polar coding chain supported in NR.

LTE where the CRC bits are always contiguous and transmitted as the last bits of the block, in Polar codes the CRC bits can be distributed across the input sequence of DCI bits to facilitate early detection of false decoding.

Depending on the length of the information sequence, parity check outer coding can be used to obtain some of the "frozen" bits based on the input sequence. Subblock interleaving and rate matching are used to adjust the number of coded bits to the actual number of the available resources using puncturing, shortening, and repetition. Finally, interleaving is applied in order to improve the performance of Polar codes in fading channels.

Depending on the waveform, Release-15 NR supports the following modulation schemes:

- CP-OFDM: QPSK, 16QAM, 64QAM, 256QAM
- DFT-s-OFDM: $\pi/2$-BPSK, QPSK, 16QAM, 64QAM, 256QAM.

The constellation mapping used in 5G NR is the same as in LTE and is based on Gray coding. $\pi/2$-BPSK modulation was added to support efficient uplink transmission in coverage-limited scenarios. Due its low PAPR, the use of $\pi/2$-BPSK modulation in conjunction with pulse-shaping filters can achieve better efficiency for power amplifiers, thus improving the link budget of NR system in the uplink.

### 3.5.10 Co-Existence with LTE, Forward Compatibility and Uplink Coverage Enhancement

NR supports several techniques that ensure efficient co-existence of the system with previous technologies such as LTE/LTE-A. In particular, to ensure smooth migration from LTE to NR, simultaneous operation of NR and LTE on the same downlink carrier is supported. For that purpose, the physical channels and reference signals of the NR system are designed to avoid collision with the physical channels necessary for operation of the LTE system. For example, to avoid collision with LTE CRS signals, special patterns for DM-RS, SS/PBCH, and CSI-RS for tracking are supported. Moreover, PDSCH in NR can be also configured with mapping patterns to avoid transmission on the resource elements occupied by LTE CRS signals. The NR signals and physical channels are also designed to avoid overlapping transmission with other LTE signals such as PSS, SSS, and PBCH. In particular, configurable RB-level rate-matching resources (or "blank resources") are defined to avoid PDSCH transmissions in the PRBs and OFDM symbols used by the corresponding LTE reference signals and channels (3GPP TS 38.214). Due to flexibility of the configuration, the RB-level rate-matching resources can be also used for forward compatibility of NR to avoid collision with transmissions of future yet-to-be-defined NR releases. This allows unrestricted development of new physical layer signals and channels for currently unknown use cases.

One of the common deployment options for NR system implies high carrier frequencies (e.g. 3.5GHz). Due to harsh propagation conditions at high frequency bands, the coverage of NR especially in uplink may be reduced compared with the coverage provided by existing LTE systems (see Figure 3.5.22).

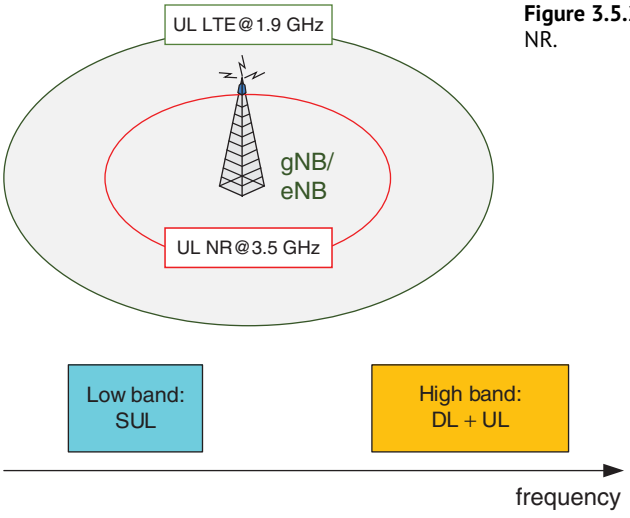



**Figure 3.5.22** Deployment option for 5G NR.

**Figure 3.5.23** Frequency allocation for supplemental uplink.

To allow NR deployments with the same coverage as LTE (and therefore site density, so that the same cell sites can be used for NR), 5G NR defined a new band operation mode denoted as Supplementary Uplink. With Supplementary Uplink (SUL), a UE can use a low-frequency band (e.g. 1.9 GHz) for uplink transmission as supplementary to higher frequency downlink and uplink operation (see Figure 3.5.23). Due to the use of lower carrier frequencies in Supplementary Uplink, an uplink coverage can be significantly improved.

## References

3GPP Technical Specification (TS) 38.101-1 (2020).NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 38.101-2 (2020). NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 38.211 (2020). NR; Physical channels and modulation. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 38.212 (2020). NR; Multiplexing and channel coding. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 38.213 (2020). NR; Physical layer procedures for control. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 38.214 (2020). NR; Physical layer procedures for data. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

3GPP Technical Specification (TS) 38.215 (2020). NR; Physical layer measurements. 3GPP. Available at: www.3gpp.org (accessed May 22, 2020).

# 4

# NG-RAN Architecture

*Colby Harper*[1] *and Sasha Sirotkin*[2]

[1] *Pivotal Commware Inc., USA*
[2] *Intel Corporation, Israel*

## 4.1 Introduction

With a high-level understanding of the whole 5G System (described in the previous chapter), we now dive into the details of the NG-RAN architectures. We use the plural term intentionally – not so much because there are multiple NG-RAN architecture options that vendors and operators need to choose from (which is indeed the case), but because these options are defined in different standards development organizations and industry fora (not just 3GPP), for different use cases and development scenarios and even sometimes with different business objectives in mind.

In this chapter we focus on standards-based NG-RAN architectures (even though some proprietary options are also covered), but one must understand that unlike, for example, the standards-based NR Uu (air) interface (described in Chapter 3), NG-RAN network standards are not as rigorous. In practice, while most implementations do try to follow them, the result is not always multi-vendor interoperable. When equipment from different vendors is used in the same network, it usually requires a fair amount of integration and interoperability testing – in particular, because there are no standards-based conformance testing and certification programs for network architectures and interfaces (unlike the air interface). Oftentimes operators working with multiple RAN vendors chose to deploy equipment from only one vendor in a given geographical area in order to reduce the integration effort and potential interoperability issues.

Therefore, NG-RAN architectures described in this chapter should be viewed primarily as a model that implementations follow with varying degrees of rigor and precision. That being said, most implementations do follow these architectures. Furthermore, unlike 4G, in 5G we may see larger numbers of multi-vendor network deployments (which is already happening with at least one green field operator), in which case the standards-based network interfaces and architectures described in the present chapter will be of great importance.

It is perhaps worth noting that the term "NG-RAN architecture" as used in this chapter (and often in the industry) is somewhat confusing, as sometimes it refers to a collection

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

of base stations and their interactions (as is the case of multiconnectivity described in Section 4.3) and sometimes it refers to an architecture of a "split" base station, in which a "monolithic" base station is further split into functional blocks with standards-based interfaces between them.

With this in mind, we start by describing what we consider as a "monolithic" gNB architecture, followed by high-level considerations about studies undertaken in 3GPP and other organizations on potential split of gNB functionalities into separate logical or physical network nodes. In the following sections of this chapter we finally describe various split architectures defined in relevant standards development organizations and industry fora.

### 4.1.1 Monolithic gNB Architecture

In the "monolithic" architecture, a gNB is depicted as a single logical network node implementing all the required functionality, which include among other things:

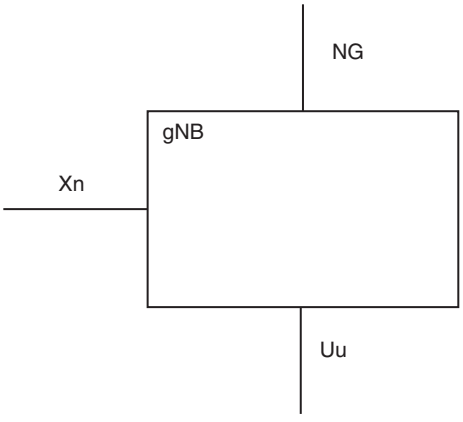
- User equipment (UE) admission control over the radio interface;
- UE radio interface connection setup and release;
- Radio resource management, including UE radio bearer control, and uplink and downlink scheduling for a UE;
- UE mobility control in connected state (i.e. handover) and in inactive state;
- UE measurements, including measurement configuration and processing of UE measurement reports;
- Routing of user-plane and control-plane packets toward user-plane function (UPF) and Access and Mobility Management Function (AMF), respectively;
- UE Quality of Service (QoS) flow management and mapping to radio bearers;
- Slicing;
- Tight interworking between NR and Long-Term Evolution (LTE), including multiple dual connectivity (between these technologies) variants;
- Radio access network sharing between multiple operators.

In standards (e.g. 3GPP), a gNB is typically described in terms of the network and air interfaces it supports, protocols it implements to run on these interfaces, and the functionality it provides. While the interfaces and protocols are typically specified in a sufficient level of detail to allow interoperability, the rest of the functionality (in particular, significant parts of the functionalities listed above) is intentionally described at the high level to allow different implementations and differentiation between vendors. As this book is primarily concerned with architectural aspects and because gNB implementation details are proprietary and differ between different vendors, we therefore proceed with the same approach as the one taken by standards to describe the "monolithic" gNB architecture in terms of the interfaces and protocols it supports.

Figure 4.1.1 illustrates the gNB architecture at the most abstract level.

The gNB terminates the Xn (which is used to communicate with other gNBs) and NG (which is used to communicate with the 5GC) network interfaces and the Uu air interface (which is used to communicate with a UE). Detailed description of the Xn and NG network interfaces can be found in Section 3.3, while the physical layer and the protocol stack of the Uu interface are described in Sections 3.5 and 3.4, respectively.

**Figure 4.1.1** Monolithic gNB architecture.




Note that this gNB representation omits certain important aspects of the physical connections implementing the logical network interfaces shown. For the network interfaces, these details are explained in Section 6.6. The Uu interface is described in Chapter 3. The description of the network interfaces provided in these sections omits the details of the transport network (3GPP specifications generally assume IP transport and do not define it in further detail). To fill this gap, we describe transport network aspects in Section 6.6.

With regard to the Uu interface, while the specifications do provide a rigorous description of all the functionality, one aspect is sometimes overlooked – the connection to the physical antenna (3GPP specifications operate at the logical abstraction of an antenna port). To fill this gap, we describe the antenna interface (which is used to configure the antenna) and the Common Public Radio Interface (CPRI) interface (which is commonly used to transfer data between a gNB and a Remote Radio Head [RRH]) in the following subsections.

### 4.1.2 Common Public Radio Interface (CPRI)

CPRI is the interface between a Radio Equipment (RE) (sometimes referred to as an RRH, or remote unit [RU]), and the rest of the base station (i.e. gNB), referred to as Radio Equipment Control (REC). CPRI is primarily used to transfer user data (along with relevant control information) between REC and RE.

Formally, CPRI specification is not a standard as it is produced by an industrial cooperation by a few network equipment vendors. Nevertheless, CPRI has historically had large market adoption, and some portions of CPRI and its evolutions have offered a minimal level of openness.

That is, in the CPRI architecture the "monolithic" gNB shown in Figure 4.1.1 is split into RE and REC, as shown in Figure 4.1.2.

The CPRI physical layer can support an electrical interface (primarily for internal connections of an integrated base station) and optical interface (primarily for remote installations). CPRI was originally defined for 3G, later extended to support 4G, and the new eCPRI specification supports 5G.

In terms of functionality split, the RE mostly implements radio frequency functions and A/D–D/A conversion, while all the other gNB functionality is implemented in the REC. This is summarized in Table 4.1.1.

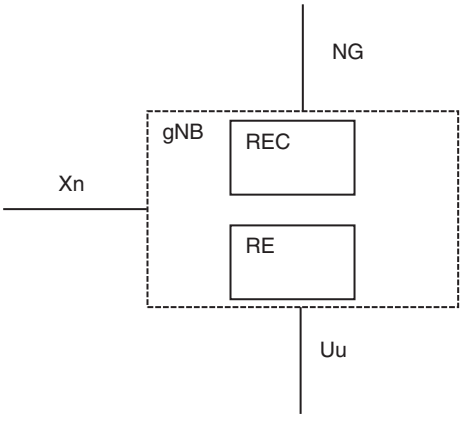



**Figure 4.1.2** gNB architecture with CPRI REC and RE split.

**Table 4.1.1** Common Public Radio Interface (CPRI) Radio Equipment Control (REC) and Radio Equipment (RE) functions.

| Functions of REC (combined central unit/distributed unit) | | Functions of RE (remote unit) | |
|---|---|---|---|
| **Downlink** | **Uplink** | **Downlink** | **Uplink** |
| Network interfaces termination, protocol stack, radio resource management (RRM), scheduling, etc. | | Cyclic prefix (CP) addition (optional) | |
| | | Channel filtering | |
| | | Digital to analog (D/A) conversion | Analog to digital (A/D) conversion |
| Channel coding, interleaving, modulation | Channel decoding, deinterleaving, demodulation | Up conversion | Down conversion |
| Inverse Fast Fourier Transform (iFFT) | Fast Fourier Transform (FFT) | ON/OFF control of each carrier | Automatic gain control |
| CP addition (optional) | CP removal | Carrier multiplexing | Carrier demultiplexing |
| Multiple-input and multiple-output (MIMO) processing | | Power amplification and limiting | Low noise amplification |
| Signal aggregation from signal processing units | Signal distribution to signal processing units | Antenna supervision | |
| Transmit power control of each physical channel | Transmit power control and feedback information detection | RF filtering | RF filtering |
| Frame and slot signal generation (including clock stabilization) | | Time Division Duplexing (TDD) switching in the case of TDD mode | |

CPRI protocol supports:

- Transfer of user-plane data (IQ samples)
- Transfer of control and management messages
- Transfer of synchronization signals
- Transfer of vendor-specific information.

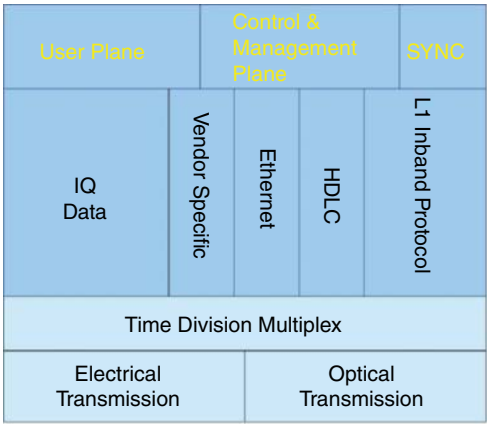



**Figure 4.1.3** CPRI protocol stack.

This is illustrated in Figure 4.1.3.

In terms of CPRI network topology, multiple CPRI links can be used between a REC and RE, multiple REs can be connected to one REC, and multiple RECs can be connected to one RE. Three types of topologies are supported:

- Chain topology (Figure 4.1.4)
- Tree topology (Figure 4.1.5)
- Ring topologies (Figure 4.1.6).

In the physical (L1) layer, CPRI can use electric and optical transports and supports bitrates that range from 614.4 Mbit/s to 24 330.24 Mbit/s.

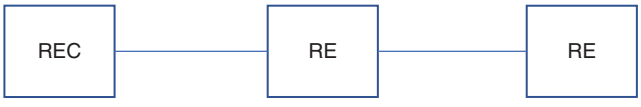



**Figure 4.1.4** CPRI chain topology.

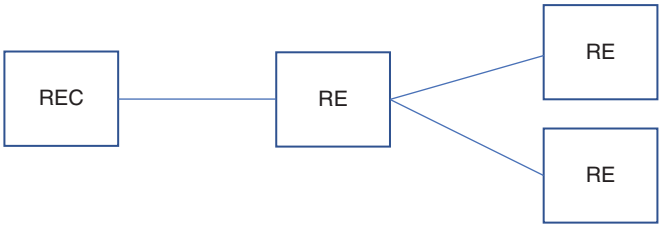



**Figure 4.1.5** CPRI tree topology.

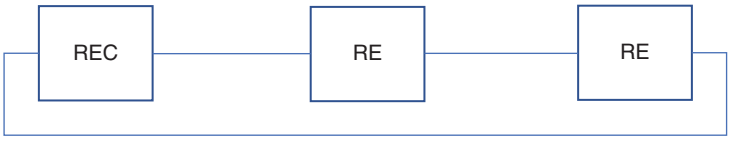



**Figure 4.1.6** CPRI ring topology.

While CPRI defines In-phase & Quadrature (I/Q) data transfer, synchronization, and a few other mechanisms in rigorous detail, many important pieces (such as management and control, for example) are intentionally left unspecified. These vary greatly between different vendors and therefore CPRI is not a fully multi-vendor interoperable standard. This gap is addressed by O-RAN specifications, described in Section 4.5.

The latest CPRI version at the time of writing this book is CPRI v7.0, which supports GSM, the Universal Mobile Telecommunications System (UMTS), LTE (including LTE-Advanced), and Worldwide Interoperability for Microwave Access (WiMAX). In order to support NR, CPRI released a new specification referred to as eCPRI.

While eCPRI (eCPRI v2.0) has many similarities to CPRI, it is a complete redesign compared with the previous generation, with the following main objectives:

- Reduction of the required bandwidth of the transport network;
- Usage of Ethernet-based transport.

It supports NR and is capable of other gNB functional split options other than PHY/RF split. The key word is "capable," as eCPRI specification does not actually define messages and procedures needed to support these functional splits.

In terms of architecture, eCPRI defines eCPRI Radio Equipment Control (eREC) and eCPRI Radio Equipment (eRE). Further, there is the notion of eCPRI/CPRI Interworking Function (IWF), providing a bridge between eCPRI and CPRI nodes. This is illustrated in Figure 4.1.7.

As one can see from the architecture Figure 4.1.7, eCPRI is designed to co-exist with legacy CPRI devices, through the usage of type 0, 1, and 2 of IWF. In particular, this allows upgrading a transport network from CPRI to eCPRI, while retaining the legacy REC and RE devices.

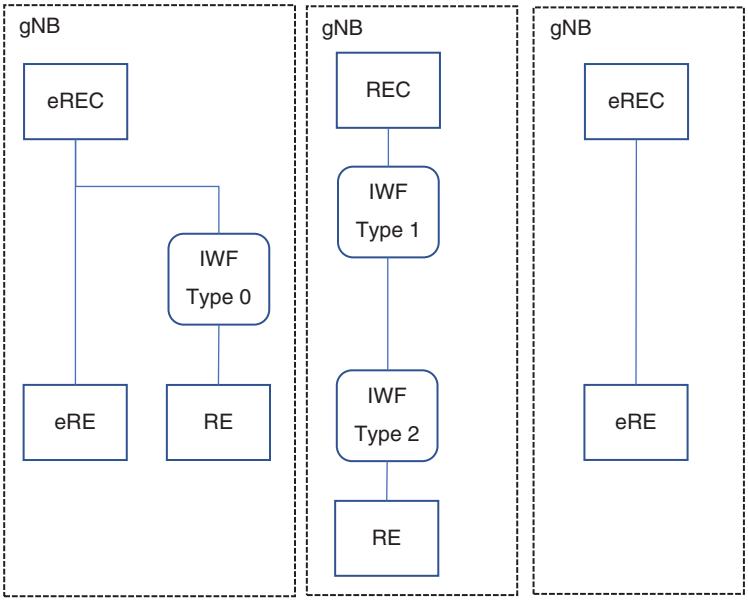



**Figure 4.1.7** eCPRI architecture.

In contrast to CPRI, eCPRI user data payload can be either I/Q samples (as in CPRI), or a bit sequence. In the latter case, the information carried over eCPRI depends on the functional split chosen and is vendor-specific. Therefore, eCPRI does not support different functional splits per se, but rather allows (in an undefined fashion) all possible splits. O-RAN low-level split specification, described in Section 4.5, uses CPRI along with Next Generation Fronthaul Interface (NGFI) as a transport and defines all the remaining details needed to support the 7.2 (see below) functional split.

For reference, it is worth mentioning that besides CPRI, the Open Base Station Architecture Initiative (OBSAI) and the ETSI Open Radio equipment Interface (ORI) also produced related specifications, which however had very limited market adoption.

### 4.1.3 Antenna Interface

So far, we have ignored the fact that in order to communicate with a UE over the air, the gNB (in the monolithic architecture) or the RE (in the CPRI architecture) must be connected to one or multiple antennas. The gNB architecture that illustrates this can be depicted as shown in Figure 4.1.8.

In this subsection we consider antenna interface specifications and what they enable over time. Here we choose to limit the focus of the antenna interface to the antenna and directional beam-centric control/monitoring/management interface functions, rather than one used to convey user traffic itself (the latter is described in the CPRI subsection). We briefly look at where we've been with antenna interface standards before 5G, where we are in the new 5G era, and where we're likely headed in 3GPP Release-17 and beyond.

Antenna interfaces are closely associated with and in various cases may be included as part of the RAN "fronthaul" interface between the baseband processing function/controller (e.g. REC in Figure 4.1.8) and one or more RRH units (or RE in Figure 4.1.8, also referred to as radio unit [RU]), to which antenna subsystems are attached. This is also the case in, for example, the O-RAN low-level split architecture, described in Section 4.5 (in which case the fronthaul interface is between O-DU and O-RU).

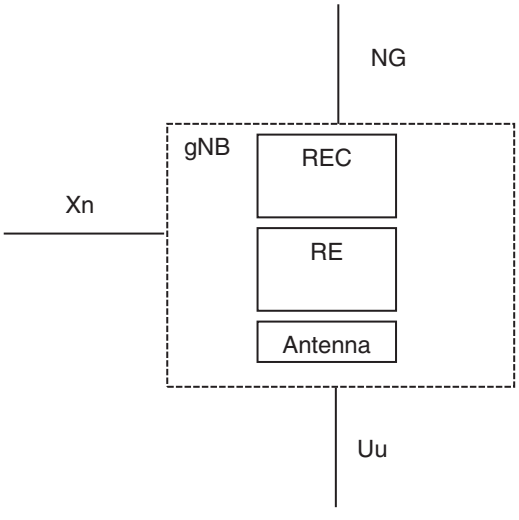



**Figure 4.1.8** gNB architecture with antenna.

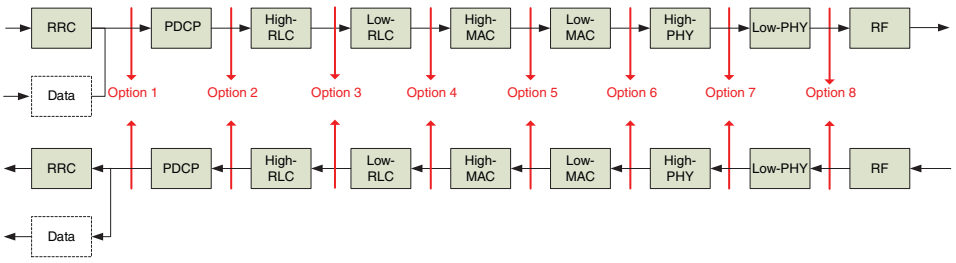



**Figure 4.1.9** gNB spit architectures considered in 3GPP study. (*Source:* Reproduced by permission of © 3GPP).

Generally, fronthaul carries control/monitoring/management messages as well as user-plane traffic. The control/monitoring messages may alternately include explicit antenna interface messages as a subset. User-plane traffic (including multi-stream multiple-input and multiple-output [MIMO]) is generally transported in the form of I/Q samples to RU, and that user-plane traffic is ultimately conveyed to/from an antenna subsystem and transmitted as energy over the air by the antenna subsystem.

#### 4.1.3.1 Before 5G: Where We Have Been

Since the 2G cellular days and on into 4G, base stations have offered sectorization rather than simple omnidirectional antenna radiation patterns, where each sector has an antenna providing a static directional beam of quite broad horizontal (azimuthal) beam width of, say, 120 or 60°. While planning, design, and operations remained mostly in the two-dimensional horizontal domain, a vertical (elevation) adjustment or "tilt" of the antenna pointing direction would be chosen at upfront installation.

Over time operation teams would more frequently adjust the horizontal and vertical antenna pointing direction to fine-tune coverage. Initially and still in many cases this required rolling a truck to the base station site for fully manual mechanical adjustments. Increasingly, antenna subsystems offered electro-mechanical and then fully electronic adjustment, which allowed remote control of these adjustment mechanisms.

This remote control was utilized in a fairly slow, static, and semi-static human-scale workflow control loop latency mode, though it did provide the initial transition from a horizontal 2D-centric to a more 3D-centric (azimuth + elevation) way of thinking about and operating antenna boresight directional tuning.

While antenna and base station vendors had their own ways of effecting these adjustments, a specification and standard was made available from the Antenna Interface Standards Group (AISG) that enjoyed and still enjoys significant market uptake. The AISG standard includes, among other features, support for antennas with remote electrical tilt for vertical/elevation adjustment. AISG has become and remains part of the 3GPP specification by way of reference within 3GPP TS 25.802 "Remote control of electrical tilting antennas," 3GPP TS 25.460 "UTRAN luant interface: General aspects and principles," 3GPP TS 25.461 "UTRAN luant interface: Layer 1," 3GPP TS 25.462 "UTRAN luant interface: Signaling transport," and 3GPP 25.463 "Tilting (RET) antennas Application Part (RETAP) signaling." These have further been transferred to the 37-series TS (3GPP TS 37.460, 37.461, 37.462, 37.466) to reflect the fact that the Iuant interface is now applicable to 5G NG-RAN,

not just Universal Terrestrial Radio Access Network (UTRAN). Furthermore, AISG v3.0 itself was released in late 2018.

#### 4.1.3.2 New 5G Era: Where We Are

While AISG specs will continue to be referenced, provide value, and undergo evolution, newer antenna interface specifications have come to the fore due much to needing improved support for new capabilities associated with historical transitions nascent in the pre-5G, LTE era including:

- From 2D to 3D tunable directional radiation patterns;
- From a "static directional antenna" concept to "dynamic antenna beamforming system concept;"
- From a human-workflow-scale slower semi-static antenna control approach to a lower latency dynamic machine-oriented antenna control approach.

In contrast to 4G, 5G on the other hand, is (if anything) beam-centric. That is, 5G has a conceptual model of time-frequency resource management that inherently contemplates an explicitly integrated spatial dimension, rather than 3G/4G's mostly implicit management of spatial domain management via power control, static cell/sector laydown, and low-rank MIMO, or 4G's more explicitly, though arguably (from a standards perspective rather than vendor-specific implementation perspective) "tacked on" "3D/FD-MIMO" beamforming nature minimally integrated into the standardized 4G RAN system architecture.

These recent dynamic beam-centric antenna control/monitoring and management interfaces currently consist of complementary specifications from the O-RAN Alliance and 3GPP, and are part of the accelerating move from closed/proprietary antenna and fronthaul interfaces toward open and standardized interfaces.

As of March 2019, the O-RAN Alliance O-RAN Fronthaul CUS Plane spec and follow-on July 2019 Version 2.0 specifies a functional split that is within the PHY layer of the 5G protocol stack (see Section 4.5 for details). Sending of antenna beamforming subsystem control messages is one function of this open fronthaul interface specification. This new open interface enables new very low latency machine-scale control loops – switching beams on even a symbol-by-symbol basis if needed.

Within the O-RAN Fronthaul CUS Plane specification, there is an abstract (high-level) open beamforming interface that provides a standard abstract interface for operators and vendors to specify just what directional beams they desire to be formed as energy over the air. As an abstract interface across different beamforming antenna subsystem technologies and products, it improves interoperability, technology-independence, and forward and backward compatibility, as well as reducing the amount of space beam control traffic takes up on bandwidth-limited fronthaul transport links. This abstract interface allows operators to describe the "what" of the desired directional beam by: beam attributes of boresight Azimuth (phi $\varphi$); elevation/zenith (theta $\theta$); beam width; and sidelobe suppression values rather than the low-level (more concrete) technology-specific "how" of generating those beams. These beam attributes provide the most compact and abstract representation of the operator-desired directional energy over the air, allowing improved real-time controllability and global reasoning by higher-level software controllers, such as those, say, utilizing AI and machine-learning optimization routines. The O-RAN Fronthaul CUS

Plane spec also provides a lower-level antenna interface that allows sending of either a batch or stream of low-level (and more verbose) beam weights that do specify exactly how each different antenna type will form the beams, rather than the "what" of the higher-level abstract interface.

To support antenna beam monitoring and control capabilities at the somewhat slower latency control loops of Operation Support System (OSS) operation, administration, and maintenance (OAM) (described in Section 6.5) network management tools, the 3GPP's Network Resource Model (NRM) as specified in 3GPP TS 28.540 "Management and orchestration; 5G Network Resource Model (NRM); Stage 1" and the associated 3GPP TS 28.541 "5G; Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and Stage 3″ has been expanded to include a beam concept. The attributes of the beam concept in the NRM also correspond to the beam attributes within the O-RAN Alliance low-latency O-RAN Fronthaul CUS Plane spec in 3GPP 5G New Radio Release-15, the PHY, Medium Access Control (MAC), and radio resource management (RRM) layers (see Chapter 3) work together to provide beam performance measurements that now have a specified place in the NRM to provide fault, performance, and configuration management OSS tools information they need to better analyze and manage ongoing performance and fault events related to directional antenna beams operation. It is expected this specification will be further enriched over the course of Release-17 efforts, including further beam operation-related enhancements.

#### 4.1.3.3 Release-17 and Beyond: Where We Are Going

While the conceptual and operational model of beamforming antenna subsystem beams in Release-15 is predominantly single TxRx point-based, in Release-16 (and more so on into upcoming Release-17 and beyond) there is a shift to support a multi-point, networked system-oriented conceptual model of multiple coordinated TxRx points.

To a degree, the Release-15 beam concept is a useful transition and stopgap: the approach is somewhat to just treat beams like "narrower and more dynamic cell sectors" for the sake of RRM link monitoring. This initial similarity allowed Release-15 to be standardized and fielded faster as well as to help operators and network equipment vendors to minimize immediate redesign of existing 4G-based management systems and technical operations processes/approaches.

That said, supporting the emerging multi-point TxRx points capability will greatly benefit from, if not demand, better and more explicit spatial/beam resource management. For instance, gNBs and other complimentary controllers such as those from the O-RAN Alliance will need to have a more explicit, precise, and low-latency knowledge of multi-beam multi-point antenna TxRx point relationships. In keeping with both O-RAN and 3GPP advanced beamforming antenna system modeling, this will include increased usage of a global coordinate system in which to orient the antennas' individual local coordinate systems.

To deliver these performance attributes, beamforming antenna subsystem interface control/monitoring will need to be distributed over higher performance fronthaul to enable finer multi-point coordination capabilities in time and space. Furthermore, we expect that over time antenna interfaces will evolve beyond mere multi-point coordination to full multi-point cooperation. For example, future antenna interfaces may support distributed

coherent beamforming, which demands that fronthaul and its antenna interface messages be delivered with higher precision time and phase sync capabilities provided by that fronthaul transport. Open, distributed, low-latency antenna control is already contemplated in, for instance, O-RAN's split-PHY fronthaul specification, and the performance enabled by this type of split is essential to go-forward antenna interfaces for many emerging advanced multi-point TxRx capabilities.

An additional emerging need is near real-time support. Between the high-latency OSS 3GPP beamforming antenna interfaces at the top of the 3GPP and O-RAN stack and the extremely low-latency antenna interfaces at the bottom, there is an emerging need to enable value-added intermediate-latency near real-time control/monitoring interfaces in support of various O-RAN Alliance and 3GPP value-added AI-enhanced controllers and distributed-self-organizing network (SON) and centralized SON efforts.

Finally, it is also worth noting that those emerging beamforming antenna subsystem interface improvements and capabilities, which are both feasible and make business sense for LTE, will likely be backported for certain market segments.

### 4.1.4 gNB Functional Split(s)

As we have already seen in the previous subsection, the monolithic base station architecture proved to be too limiting and there has always been a desire to standardize a more flexible split base station architecture. Such split architectures are more flexible in adapting to various deployment scenarios, in which it may not be possible to deploy baseband processing hardware close to an antenna or there may be a need to control multiple cells from a centralized remote location. Even in cases when RF and baseband processing units are installed in close proximity to each other, oftentimes these are developed by different vendors and therefore there is a benefit in having a standardized interface between these.

To cater to these deployment needs, many vendors have implemented various base station split architectures, in which an eNB or a gNB is further split into a number of logical or physical nodes. Many of these implementations are based on the CPRI specification, which however does not provide true multi-vendor interoperability as it severely underspecifies the functionality required to implement a fronthaul interface. While in 4G all these architectures have been proprietary, in 5G 3GPP (and in fact other Standards Developing Organizations [SDOs] and industry fora) considered and eventually standardized a number of split gNB architectures.

During the 5G study, split options illustrated in Figure 4.1.9 have been considered by 3GPP.

A brief description of each functional split option is provided below. In this preliminary study, an architecture in which a gNB is split into two logical nodes, referred to as central unit (CU) and distributed unit (DU), has been considered, as illustrated in Figure 4.1.10.

Note that here and in the rest of the book, we omit the antenna interface for brevity.

- **Option 1**

In this option, the radio resource control (RRC) layer resides in the CU. The rest of the protocol stack layers, that is, Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), MAC, PHY, and RF are in the DU. In other words, this option separates the control

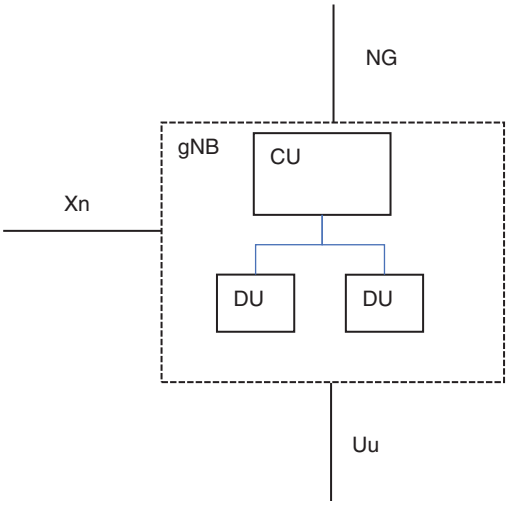



**Figure 4.1.10** gNB architecture with centralized unit and multiple distributed units.

plane (i.e. RRC) from the user plane, allowing centralized control-plane and distributed user-plane deployments.

The main benefit of this option is the centralized RRM, allowing a certain level of optimization mainly in terms of UE mobility, as these decisions are taken by a CU (potentially connected to a large number of DUs) having a higher level of visibility into potentially many cells it can control. Furthermore, this option is the least stringent in terms of throughput and latency requirements on the fronthaul transport network. The downside is that it offers little to no gains in terms of resource pooling and scheduling, as user plane functions are not centralized.

- **Option 2**

In this option, also referred to as *PDCP-RLC split*, the RRC and the PDCP protocol stack layers reside in the CU. The rest, that is, RLC, MAC, physical layer, and RF, reside in the DU.

It has similar advantages (e.g. centralized RRM) and disadvantages (e.g. little to no performance and resource pooling gains) as option 2, with two additional benefits:

1. It becomes possible to realize some resource pooling gains, as PDCP processing is done in CU.
2. The PDCP-RLC functional split is very similar to the functional split supported in dual connectivity (see Section 4.3 for details), which allows a fair amount of reuse both in terms of standardization and implementation of both options.

Note: this is the option that eventually was standardized in 3GPP, see Section 4.2.

- **Option 3**

In this option, also referred to as intra-RLC split, low-RLC (certain low-level/real-time functions of RLC), MAC, PHY, and RF reside in the DU. PDCP and high-RLC (the remaining parts of RLC) are in the CU.

Two flavors of option 3 have been considered, specifically:

  - **Option 3-1**

In this option, low-RLC implements the segmentation function, while high-RLC implements the Automatic Repeat Request (ARQ) and other RLC functions.

In terms of benefits, it allows further resource pooling gains compared with option 2, as more functionality is centralized in the CU. Additionally, it offers better flow control and can tolerate failures in the fronthaul transport network, as ARQ (and retransmission) functionality is implemented in the CU. This comes at a cost of somewhat more stringent latency requirements for the fronthaul transport network.

- **Option 3-2**

In this option, the split is based on transmit and receive RLC entities. That is, low-RLC implements transmitting Transparent Mode (TM) RLC entity, transmitting Unacknowledged Mode (UM) RLC entity, or a transmitting side of Acknowledged Mode (AM) and the routing function of a receiving side of AM. Consequently, high-RLC implements receiving TM RLC entity, receiving UM RLC entity, or a receiving side of AM except the routing function and reception of RLC status report.

The main advantage of this option (compared with option 3-1) is that it does not introduce additional latency requirements on the fronthaul.

- **Option 4**

In this option, also referred to as *RLC-MAC split*, MAC, PHY, and RF reside in the DU. PDCP and RLC are in the CU.

This option was considered in the study primarily for completeness, as no advantages have been identified.

- **Option 5**

In this option, also referred to as *intra-MAC split*, RF, PHY, and some part of the MAC layer (e.g. Hybrid ARQ [HARQ], which has more stringent timing requirements) are in the DU. The rest, that is, upper-MAC, RLC, PDCP, and RRC, are in the CU.

The main benefits of this option are:

1. Even further resource pooling gains, as more functionality (compared with lower-numbered options) is centralized in the CU.
2. Efficient interference management across multiple cells and enhanced scheduling technologies such as Coordinated Multi-point (CoMP) and Carrier Aggregation (CA), with multi-cell view, as the scheduler (which is normally part of the MAC protocol layer) is centralized in the CU.

The main disadvantage is the extra complexity of the interface between the CU and the DU.

- **Option 6**

In this option, also referred to as *MAC-PHY split*, physical layer and RF are in the DU. All the protocol stack layers are in the CU.

This option has all the benefits of option 5 (resource pooling, efficient scheduling, etc.). Additionally, resource polling gains are somewhat bigger (as even more functionality is centralized in the CU) and more advanced scheduling optimizations become possible, for example joint transmission.

This comes at the cost of subframe-level timing requirements on the fronthaul interface. Furthermore, round trip fronthaul delay may affect HARQ timing and scheduling.

Note: this is the option that is standardized in the Small Cell Forum, see Section 4.7 for details.

- **Option 7**

  In this option, also referred to as *intra-PHY split*, part of physical layer function and RF are in the DU. The remaining parts of the PHY and all the protocol stack are in the CU. Multiple PHY partitioning options have been considered (and even more options are technically possible), in particular:

  - **Option 7-1**

  In this option, in the uplink, Fast Fourier Transform (FFT), cyclic prefix (CP) removal, and possibly Physical Random Access Channel (PRACH) filtering functions reside in the DU, the rest of the PHY functions reside in the CU. In the downlink, iFFT and CP addition functions reside in the DU, the rest of PHY functions reside in the CU.

  The main benefit is that this option allows the implementation of advanced receivers.

  - **Option 7-2**

  In this option, in the uplink, FFT, CP removal, resource de-mapping, and possibly prefiltering functions reside in the DU, the rest of the PHY functions reside in the CU. In the downlink, inverse FFT (iFFT), CP addition, resource mapping, and precoding functions reside in the DU; the rest of the PHY functions reside in the CU.

  Note: this is the option selected for standardization by the O-RAN Alliance for the low-level split. More details can be found in Section 4.5.

  - **Option 7-3 (Only for DL)**

  In this option, only the encoder resides in the CU, and the rest of the PHY functions reside in the DU.

  The main benefit is that it reduces the fronthaul requirements in terms of throughput to the baseband bitrates as the payload for Option 7-3 is encoded data.

- **Option 8** In this option, also referred to as *PHY-RF split*, RF functionality is in the DU and all the upper layers are in the CU.

  The main benefit is that the interface between CU and DU is relatively simple (as the interface needs to convey mostly I/Q data samples, together with some control and synchronization information); however, it has the most stringent requirements on the fronthaul interface in terms of latency and especially throughput.

  This is the option supported by CPRI and eCPRI.

Generally, one can observe that lower splits offer higher gains in terms of performance (RRM, scheduling, interference mitigation, CoMP) and bigger resource utilization gains, as more functionality is centralized in the CU. Furthermore, a CU may be connected to a large number of DUs controlling multiple cells in a relatively large area – this allows implementation of more advanced RRM and scheduling algorithms, which take into account the radio conditions of multiple cells and large numbers of UEs. This, however, comes at a cost of increased complexity (in terms of standardization, implementation, and interoperability testing) of the fronthaul interface. Furthermore, lower splits can substantially increase fronthaul requirements in terms of latency and throughput, to the point that certain splits (e.g. split option 8) may not be feasible at all. For example, in the mmWave frequency range,

certain functional splits' fronthaul bandwidth requirements may be hard to satisfy even if fiber backhaul is available.

In the course of the study on the gNB split, 3GPP came to realize that there is interest and market demand for two somewhat different split options:

- Low-level split (options 6, 7, and 8), which is appealing to operators that have sufficient fiber fronthaul transport;
- High-level split (options 1, 2, 3, and 5), which can be deployed by operators that do not have fiber fronthaul transport or consider investment in fiber transport not justifiable.

Therefore, one can generalize a gNB functional split architecture as shown in Figure 4.1.11, in which a gNB is split into three logical nodes:

- CU, providing high-layer protocol stack functionality (e.g. PDCP);
- DU, providing low-layer protocol stack functionality (e.g. RLC and MAC);
- RU, providing lower-layer functionality (e.g. parts of PHY).

Notes on Figure 4.1.11:

- As mentioned above, the antenna interface is not shown for brevity.
- Eventually, this architecture was generalized even further, when control- and user-plane functions have been separated into their own logical network nodes.

Eventually, 3GPP have standardized the option 2 (PDCP-RLC split) in the CU/DU split architecture with fronthaul interface between a CU and a DU called F1 – this is described in detail in Section 4.2. Even though this option provides little to no gain in terms of scheduling and resource pooling, it was considered a relatively "low-hanging fruit" as it uses similar functional split to the multi-connectivity architecture (described in Section 4.1), which has also been standardized in 3GPP. Additionally, this option becomes more appealing and can provide significant resource pooling gains if an operator deploys a CU in a virtualized environment (described in Section 6.2), potentially together with the Mobile Edge Compute system (described in Section 6.4) running on the same hardware platform. Furthermore, a variant of split option 1, in which the control plane (e.g. RRC) is separated from the user plane, has also been standardized in 3GPP and is supported by the E1 interface – this is described in detail in Section 4.4.

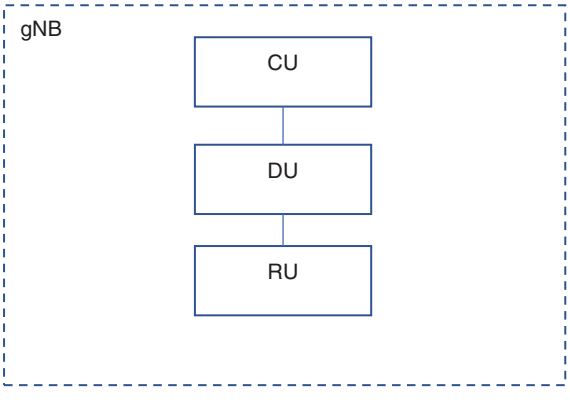



**Figure 4.1.11** Generalized gNB functional split architecture with CU, DU, and RU.

Despite significant interest in a low-level functional split option, 3GPP could not reach the consensus required to standardize it. While nobody challenged the technical benefits of the low-level split, some companies argued that providing a truly multi-vendor interoperable standardized interface to support it is not feasible. As it was not possible to conduct this work in 3GPP, some companies turned to O-RAN Alliance, which produced the low-level functional split specification (a variant of option 7-2). For more details see Section 4.5.

### 4.1.5 Conclusions

In this introductory section we defined the "monolithic" gNB architecture, in which all (but the RF and the antenna) parts are implemented in a single network node. We then described the CPRI and eCPRI interfaces,which are commonly used to convey user data to the RF/antenna component (often referred to as RRH or RRU) and the Iuant, AISG, and other interfaces that are used to control the antenna.

We then discussed that the "monolithic" architecture proved to be too limiting to cater to all relevant deployment scenarios. To this end, 3GPP (and other SDOs and industry fora) studied various functional splits, some of which have been standardized and are described in the following sections of this chapter.

### 4.1.6 Further Reading

Most of the topics mentioned in this section are described in more detail in the book. However, there are some detailed descriptions of what (beyond what appears in the present section) was considered out of scope, specifically:

- CPRI and eCPRI
- Antenna interfaces.

Readers interested in more detailed description of the former should consider reading the relevant specifications provided by CPRI – CPRI v7.0 and eCPRI v2.0.

Readers interested in historical perspective about standardization activities related to fronthaul interfaces may consider reading ETSI ORI and OBSAI specifications.

For details about antenna interfaces, refer to Iuant 3GPP specifications (3GPP TS 37.460 and TS 37.466) and AISG specifications (AISG v3.0, AISG ST-RET and AISG ST-TMA).

## References

3GPP Technical Report 25.802 (2004). Remote control of electrical tilting antennas. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Report 28.540 (2019). Management and orchestration; 5G Network Resource Model (NRM); Stage 1. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Report 28.541 (2019). Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3. Availake at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 23.501 (2019). System architecture for the 5G System (5GS). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 23.502 (2019).Procedures for the 5G System (5GS). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 37.460 (2019). Iuant interface: General aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 37.461 (2019). Iuant interface: Layer 1. Available at: www.3gpp .org (accessed May 29, 2020).
3GPP Technical Specification 37.462 (2019). Iuant interface: Signalling transport. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 37.466 (2019). Iuant interface: Application part. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.413 (2019). NG-RAN; NG Application Protocol (NGAP). Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.414 (2019). NG-RAN; NG data transport. Available at: www .3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.415 (2019). NG-RAN; PDU Session User Plane protocol. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.424 (2019). NG-RAN; Xn data transport. Available at: www .3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.425 (2019). NG-RAN; NR user plane protocol. Available at: www.3gpp.org (accessed May 29, 2020).
Common Public Radio Interface (2015). CPRI Specification V7.0. Available at: www.cpri.info (accessed May 29, 2020).
Common Public Radio Interface (2019). eCPRI Specification V.2. Available at: www.cpri.info (accessed May 29, 2020).
ETSI(2014). Open Radio equipment Interface (ORI); Requirements for Open Radio equipment Interface (ORI) (Release 4). Available at: www.etsi.org (accessed May 29, 2020).
Open Base Station Architecture Initiative (2018). OBSAI System Spec V2.0. Available at: www .obsai.com (accessed May 29, 2020).
The Antenna Interface Standards Group (2018). AISG v3.0 base standard. Available at: http:// aisg.org.uk (accessed May 29, 2020).
The Antenna Interface Standards Group (2018). Remote electrical tilt (ST-RET). Available at: http://aisg.org.uk (accessed May 29, 2020).
The Antenna Interface Standards Group (2018). Tower mounted low noise amplifier (ST-TMA). Available at: http://aisg.org.uk(accessed May 29, 2020).

## 4.2 High-Level gNB-CU/DU Split

*Sasha Sirotkin*
*Intel Corporation, Israel*

In this section we discuss the gNB-CU/DU split architecture, which is equivalent to option 2 from the 5G study, which we mentioned in Section 4.1.

The RAN architecture has been evolving over the years and 3GPP Releases in somewhat circular fashion. In UMTS, the RAN consists of two network nodes: Node B and RNC. The LTE RAN architecture, on the other hand, is flat – there is only one eNB node, which combines the roles previously assigned to Node B and the RNC. In 5G, 3GPP enabled the hierarchical architecture again. However, even though the high-level architecture may look similar to UMTS, there are important differences, as we will show.

Generally, 3GPP defines a logical network architecture, meaning that logical network nodes defined in the specifications can be mapped in different ways into physical network nodes in actual implementations. Multiple logical nodes can be combined into a single physical node, in which case network interfaces "collapse" into internal Application Programming Interfaces (APIs), which may or may not follow the standard. The same does not necessarily apply to the case of "splitting" a single logical node into multiple physical ones – this can, of course, be done in a proprietary manner, but if multi-vendor interoperability of the split nodes is desired, such architecture must be standardized. On the other hand, having both split and non-split architecture in the specification is not required. The fact that the NR standard, contrary to the UMTS specifications, supports both architectures (single-node gNB and a gNB split into two logical nodes) illustrates the internal 3GPP struggles to define open and multi-vendor interoperable network interfaces and the reluctance of some parties to standardize these.

In this section we describe the NG-RAN architecture in which a gNB is split into two (high-level and low-level) logical nodes, with a standardized F1 interface between them. This is just one deployment option – as mentioned above, a gNB can also be implemented as a single physical network box. Moreover, multiple gNB split architectures have been defined, as we explain further in the book. The current section only describes the RAN aspects of these procedures supported by the F1 interface, which assumes some knowledge of the air interface and core network functionality. For an overview of these, please refer to Chapter 3.

### 4.2.1 Key Ideas

- LTE eNB is defined as a single logical network node, implementing RRC, PDCP, RLC, MAC, and PHY layers. Many implementations follow 3GPP LTE logical architecture; however, some vendors chose to split an eNB into two (or even more) physical nodes. The interface between these nodes is proprietary and therefore not interoperable between multiple vendors.
- In 5G, 3GPP have defined both a single-node "monolithic" NG-RAN architecture and a split architecture. In the latter, a gNB is split into two logical nodes: a centralized gNB central unit (gNB-CU) node and one or many distributed gNB distributed unit (gNB-DU) nodes, connected via the standardized F1 interface.

- The underlying assumption is that a single gNB-CU controls a large number of cells (much larger than a typical gNB would), thus allowing centralized RRM across many cells for improved performance, resource pooling across many gNBs (at least for PDCP functionality) for lower cost and energy, and virtualization support.
- In the split architecture, a gNB-CU hosts PDCP, Service Data Adaptation Protocol (SDAP), and RRC layers, and a gNB-DU node hosts RLC, MAC, and PHY layers. This functional split does not allow the realization of all the benefits of centralization (discussed above), as most of the functionality resides in a gNB-DU and there is relatively little to be centralized in a gNB-CU.
- 3GPP has decided to define a high-level split as a compromise to support at least deployments where a high-speed (e.g. fiber) fronthaul transport network is not available. The PDCP-RLC split, as opposed to an intra-RLC split and a few other options considered, was chosen for alignment with the dual connectivity architecture, which follows the same protocol split.
- The F1 standardized interface between a gNB-CU and gNB-DU(s) includes the F1 User-Plane (F1-U) protocol and control plane F1 Application Protocol (F1AP), which generally follow the same design principles as other RAN interfaces (e.g. Xn and X2): F1-U uses GPRS Tunneling Protocol user plane (GTP-U) as the transport and F1AP uses Stream Control Transmission Protocol (SCTP).
- F1AP supports: interface management procedures, UE context management procedures, RRC message transfer procedures, system information procedures, and paging procedures.
- F1-U uses the common NR user-plane protocol, which is also used on other NG-RAN interfaces (e.g. Xn). It extends GTP-U with functionalities such as flow control, delivery status feedback, and others.
- gNB-DU and gNB-CU are managed separately via OAM. Certain information has to be preconfigured in both nodes by OAM when new network nodes are deployed. For example, a gNB-DU must be preconfigured with the transport network address of a gNB-CU, a list of cells it supports, etc.

### 4.2.2 Market Drivers

Most LTE base stations are deployed today as a single physical network node, often connected to remote antennas controlling what is sometimes referred to as a "three-sector cell," that is, a single base station often controls three cells covering three sectors. Nevertheless, the idea of splitting a base station into separate nodes and centralizing at least parts of the base station functionality (similar to UMTS) has been considered for LTE as well. In such an architecture, a base station is "split" into two network nodes, one "high-level" node deployed in a centralized location serving potentially many "low-level" nodes distributed close to antennas.

Such centralization has numerous benefits in terms of cost and performance. A centralized node can serve more cells and as the load in these cells can vary, it can reduce hardware costs by resource pooling and sharing. Moreover, as the centralized node has larger visibility into more cells and UEs in these cells, it provides improved RRM and better scheduling (to an extent, as the scheduler itself is not centralized). However, these gains often come

at a cost of potentially increased latency and more stringent requirements on the transport network.

Such architecture was used in 3G, where one centralized RNC controls multiple distributed Node Bs. In LTE, 3GPP moved away from this architecture and designed E-UTRAN as a "flat" network with a single network node – the eNB. With the advent of 5G, history repeats itself and the RAN architecture evolves from hierarchical 3G architecture, through flat LTE architecture, to hierarchical architecture once again in 5G.

It is important to point out that even though the standardized E-UTRAN architecture in LTE is flat, some vendors have implemented split RAN architectures. Obviously, in such implementations network nodes are connected via proprietary interfaces, which makes it impossible to deploy equipment from different vendors in the same network. One of the intentions of the 3GPP standardization process is to allow a split NG-RAN implementation with standardized interfaces, to allow multi-vendor deployments.

Such an interface has been defined; however, it remains to be seen whether actual multi-vendor NG-RAN deployments materialize. In the past, operators had the tendency of deploying equipment from a single vendor in a given area, in order to minimize interoperability issues. This may or may not change in 5G, as operators are increasingly interested in using equipment from multiple vendors to drive down costs. Moreover, there are new companies trying to exploit that desire to enter the marker. On the other hand, "mixing and matching" equipment from different vendors shifts the burden of interoperability testing and issue resolution from a vendor to an operator, who may or may not be willing or capable of carrying this burden.

As was concluded during the 5G study, the lower (in the protocol stack) the base station is split, the more its functionality is centralized, resulting in bigger performance gains (due to e.g. centralized RRM) and bigger resource sharing gains. Moreover, only sufficiently "low" splits (e.g. below MAC or intra-PHY) allow for centralized scheduling, which provides for substantial performance gains. However, lower split (especial intra-PHY) introduces much more stringent requirements to the fronthaul transport network, in terms of throughput and latency. In practice, "low" (i.e. intra-PHY) splits can only be deployed by operators which have a sufficiently large fiber transport network. Deploying a new fiber network just for the sake of using the low-level split architecture may not be economically feasible.

Therefore, the 3GPP has decided to standardize the so-called "high-level split," in contrast to the "low-level split" described in Section 4.5. It is worth mentioning that this architecture, initially defined for 5G gNB, was later "backported" to 4G ng-eNB as well. The decision to focus on this particular flavor of the high-level split in 3GPP was driven by the following somewhat contradictory factors:

1. Maximization of centralization gains.
2. Relaxed requirements on the transport network.
3. Limited impact on latency (of UE procedures involving a UE).
4. Consistency (and potential reuse of) with other protocols and architectures already defined in 3GPP.

For non-technical reasons, 3GPP could not reach consensus on standardizing the low-level split, even though numerous advantages of such split have been identified during the study. Eventually, the work on the low-level split was pushed to the O-RAN Alliance, which

produced the low-level split specification, covering control, user, and management planes (see Section 4.5). 3GPP has only standardized the high-level split, which is described in the present section.

Additional information about the 3GPP study of the gNB split architecture can be found in 3GPP TR 38.801,[1] section 11.1.

### 4.2.3 Functional Description

As mentioned above, 3GPP defined two logical gNB architectures: a monolithic architecture with a gNB defined as a single logical network node encompassing all the functionality and a split architecture with a gNB-CU connected to one or multiple gNB-DUs. Explicit standardization of two such options is somewhat unusual (and strictly speaking unnecessary), as normally 3GPP aims at defining only a single solution. As already explained above, normally any modifications to the defined single solution are up to the implementation; in the case of network architecture, this means that each standardized logical network node can be implemented in a different number of physical network nodes.

The split and the no-split architectures are virtually indistinguishable from the point of view of a UE, a 5GC, or another gNB. Split deployment may affect latency or performance, but functionality-wise it is equivalent to the monolithic deployment. In the split case, all the network interfaces (i.e. NG and Xn) are terminated at the gNB-CU, the air interface to a UE is terminated in the gNB-DU, and a gNB-CU and gNB-DU(s) are connected via the F1 interface, as illustrated in Figure 4.2.1. In this case, all the "external" interfaces, that is NG, Xn and Uu, are exactly the same as in the case where they terminate at a monolithic gNB. For brevity, Figure 4.2.1 does not show the OAM system, which manages a gNB-CU and a gNB-DU separately. OAM aspects are explained in more detail in Section 6.5.

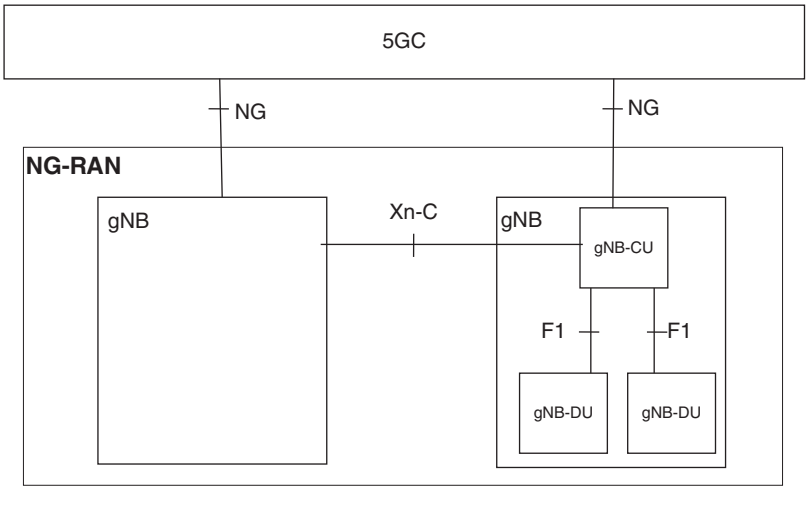



**Figure 4.2.1** Overall NG-RAN architecture. (*Source:* Reproduced by permission of © 3GPP).

1 3GPP Technical Report TR) 38.801, as all 3GPP TRs, is not normative and should only be used as a source of information about 3GPP study progress.

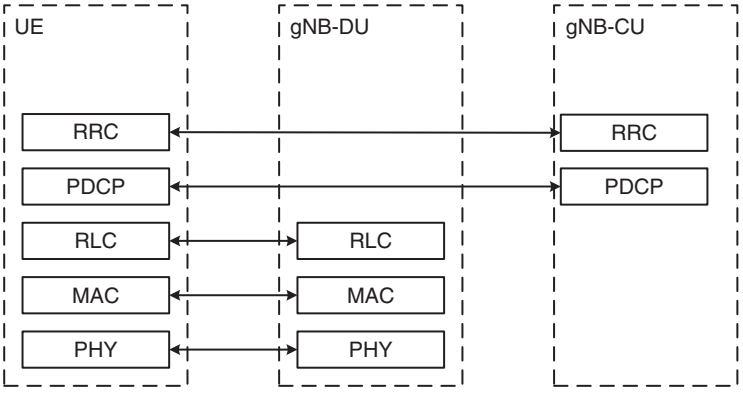



**Figure 4.2.2** gNB-CU/gNB-DU protocol stack split.

A gNB-CU hosts all the upper-layer protocols and functionality: RRC, SDAP, and PDCP. A gNB-DU hosts all the lower-level protocols and functionality: RLC, MAC, and PHY. This is illustrated in Figure 4.2.2.

Note that for brevity Figure 4.2.2 does not illustrate the F1 protocol stack, which is used to carry PDCP Protocol Data Units (PDUs) between gNB-CU and gNB-DU. Another point worth mentioning is that even though according to the gNB split architecture the RRC layer resides in the gNB-CU, certain functions with tight timing requirements, such as System Information Broadcast, which otherwise could be considered as part of the RRC layer, are actually implemented in the gNB-DU. In other words, the layer separation is not perfect.

A gNB-CU is connected to potentially multiple gNB-DUs via the standardized F1 interface; a gNB-DU is connected to only one gNB-CU. A gNB-DU (and by extension a gNB-CU) may control multiple cells; however, a single cell is always controlled by a single gNB-DU. 3GPP decided not to allow a single cell to span multiple gNB-DUs for simplicity of standardization and deployment.

The F1 standardized interface between a gNB-CU and gNB-DU(s) supports the user-plane protocol (F1-U) and control-plane protocol (F1AP). The protocol design of both the user and the control plane generally follow the same principles as other RAN interfaces, as we show in more detail below.

More details about the overall NG-RAN architecture can be found in 3GPP TS 38.401.

#### 4.2.3.1 F1 Control-Plane Protocol

The control-plane part of the F1 interface (F1-C) uses the F1AP. The F1AP protocol design is similar to other application protocols defined for RAN in 5G and LTE (e.g. X2-AP and Xn-AP). As with most other RAN control protocols (e.g. Xn-AP described later in the chapter), it relies on SCTP, which provides reliable transport for in-sequence delivery of the control-plane messages. The F1-C protocol stack is illustrated in Figure 4.2.3.

F1AP procedures[2] can be categorized as follows:

- Interface management procedures
- UE context management procedures

2 In the interest of clarity, this chapter only describes the most important procedures and functionalities. Some procedures, such as Warning Message Transmission procedures, are left out.

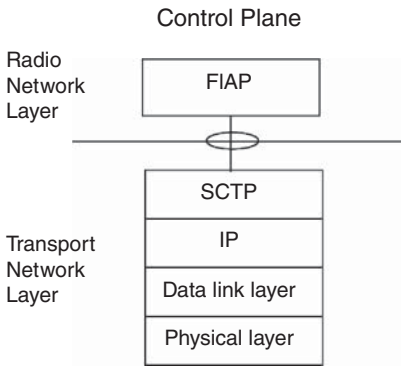



**Figure 4.2.3** F1-C protocol stack. (*Source:* Reproduced by permission of © 3GPP).

- RRC message transfer procedures, system information procedures, and paging procedures.

As for all other RAN interfaces, F1AP procedures can be class 1 and class 2. The class 1 procedures are the most common and they have a Request followed by a Response or an Error message call flow. Class 2 procedures do not require a response.

The F1 interface has been specifically designed to allow gNB-CU deployments in the virtualized environment. This is different from LTE in which all the control-plane interfaces always use a single transport network layer (TNL) association (i.e. a single SCTP connection for a single interface instance). F1, on the other hand, can use multiple TNL associations on a single interface instance. This is helpful when gNB-CU is deployed as a virtual machine in, for example, a cloud, where computational resources can be added "on the fly" without disturbing the network operation, even if new resources need additional transport network capacity (e.g. additional transport network interfaces). Even though it is somewhat less likely that a gNB-DU would be virtualized, it can also use multiple SCTP endpoints for a single instance of an F1-C interface. This is not unique to the F1 interface, as a similar mechanism is also defined for NG and Xn interfaces.

Full details about the F1AP control-plane protocol can be found in 3GPP TS 38.473.

#### *4.2.3.1.1 Interface Management Procedures*

Interface management procedures are used to establish an F1 interface, to tear it down, to reset it (when and if needed), and to allow gNB-CU and gNB-DU to exchange and update configuration information (such as the list of supported cells, etc.). These procedures are:

- F1 Setup
- gNB-DU/gNB-CU Configuration Update
- gNB-DU Resource Coordination
- gNB-DU Status Indication
- Reset and Error Indication.

Once a TNL association between a gNB-DU and gNB-CU becomes operational, the F1 Setup procedure must be initiated by the gNB-DU. As it is expected that new gNB-DUs can be added (e.g. to extend coverage and/or to add capacity) when the network is operational, it is the gNB-DU that initiates the interface setup by sending the F1 Setup Request message, rather than a gNB-CU.

The F1 Setup message contains information about cells supported by the gNB-DU and its RRC version. In the normal F1 Setup scenario, when the gNB-CU accepts the request,

it responds with an F1 Setup Response message carrying the indication of which cells the gNB-CU requests the gNB-DU to activate, list of Public Land Mobile Networks (PLMNs) supported (which is important for RAN sharing between operator deployments), and gNB-CU's RRC version. The RRC version information is exchanged to ensure that both network nodes support the same version. This is important because, as mentioned above, the layer separation model in which RRC resides in the gNB-CU is not perfect. One such example is the case when a gNB-DU may trigger the RRC Reconfiguration procedure for a UE to be performed by the gNB-CU.

If a gNB-CU cannot accept the setup request, for example due to RRC version mismatch or overload, it may also respond with a F1 Setup Failure message. If the issue is expected to be resolved automatically within a certain period of time, a gNB-CU may include the "time to wait" indication in the F1 Setup Failure message, during which a gNB-DU is not allowed to reattempt the setup. After the F1 Setup procedure is completed successfully, the gNB-DU is considered operational and can serve UEs. It is implicitly assumed in the setup procedure described here that certain information, such as a list of supported cells and PLMNs, is preconfigured in a gNB-DU and a gNB-CU by OAM before the F1 Setup procedure takes place.

During the F1 interface operation, a gNB-DU can inform the gNB-CU about configuration changes using the gNB-DU Configuration Update procedure. Similar to gNB-DU, the gNB-CU can use the gNB-CU Configuration Update procedure to inform any of the gNB-DUs about changes on its side. For example, a gNB-DU may indicate to a gNB-CU when new cells are configured, previously configured cells are removed, or existing cell configuration (e.g. Cell Global Identifier [CGI] or System Information [SI]) is changed. Note that the configuration update procedure names are somewhat misleading, as these are not used to update the configuration (this is done via OAM), but to notify the other network node about the configuration update that has already taken place.

The gNB-CU Configuration Update procedure is used for:

- Activation and de-activation of cells in a gNB-DU;
- Addition, removal, and update of TNL associations;
- Indication of cells to be barred;
- Information transfer on LTE/NR co-existence.

A gNB-CU can request a gNB-DU to activate cells (among those that can be served by the gNB-DU, as configured by OAM and indicated to the gNB-CU during F1 Setup or gNB-DU Configuration Update). Along with the activation indication, a gNB-CU sends information such as the list of available PLMNs and a container with part of the SI, which is controlled by a gNB-CU. When certain cells are no longer needed, for example, owing to energy saving or other reasons, a gNB-CU can request to deactivate them.

As mentioned above, a single F1 interface can use multiple TNL associations to support RAN deployment in virtualized environments. In such cases, a gNB-CU may be deployed, for example, in a server farm with multiple computational resources and multiple network interfaces. One of the key features of virtualized deployments is the capability to add computational resources "on the fly" and to this end the F1 interface allows addition and removal of TNL associations. This is performed using the gNB-CU Configuration Update procedure, which allows a gNB-CU to add, remove, and update TNL association information. To this

end, *gNB-CU TNL Association To Add List* IE, *gNB-CU TNL Association To Remove List* IE, and *gNB-CU TNL Association To Update List* IE are used. The gNB-DU can also add and remove SCTP endpoints on its side; however, the procedure used is slightly different – a gNB-DU can add endpoints "implicitly" by simply initiating a SCTP connection from a new endpoint, but it must "explicitly" remove an endpoint using *gNB-DU TNL Association To Remove List* IE. There is no good technical reason for using two different procedures to achieve a very similar goal.

Since LTE and 5G can share spectrum, a coordination mechanism (between gNB-CUs) has been defined to prevent co-existence issues. Information about which E-UTRA resources needs to be protected is exchanged between gNB-CUs via the Xn interface and propagated to gNB-DUs via the F1 interface, using the gNB-CU Configuration Update procedure and the *Protected E-UTRA Resources List* IE.

In the case where a gNB-DU is experiencing an overload, it can indicate so to the gNB-CU using the gNB-DU Status Indication class 2 procedure. This indication is rather coarse and is limited to "overloaded"/"not-overloaded" states. It can be used by the gNB-CU for load-balancing purposes, for example. More details about load in the DU can be deduced based on the per-bearer assistance information that is provided within the NR user plane (e.g. Average HARQ Failure or Power Headroom Report). This information may not always be available: first, it is part of the user plane and thus its usage for control-plane decisions requires implementation of cross-domain triggers; and second, in the case of dual connectivity, the NR user plane may connect the hosting node and the DU directly, so the content is not available for the CU (dual connectivity is explained in Section 4.3).

We further illustrate the concepts described above by one typical F1 interface management call flow involving the F1 Setup and the Configuration Update procedures, shown in Figure 4.2.4. We then focus on the detailed description of the F1 Setup message and we

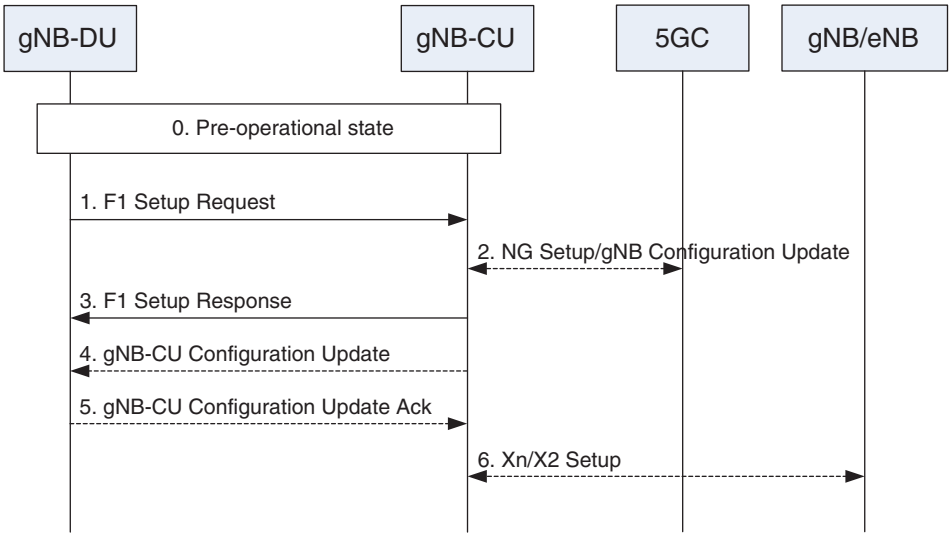



**Figure 4.2.4** F1 startup and cell activation. (*Source:* Reproduced by permission of © 3GPP).

encourage the reader to study further details about F1 maintenance procedures in 3GPP TS 38.473.

0. At the beginning, basic information is assumed to be preconfigured by the OAM in a gNB-DU (e.g. a gNB-CU TNL address, list of cells, etc.) and in a gNB-CU (e.g. an AMF TNL address, list of PLMNs, etc.).
1. The gNB-DU initiates the setup procedure by sending the F1 Setup Request message to the gNB-CU, carrying the following information: gNB-DU ID, Name, RRC version, and a list of cells (with relevant information, e.g. frequency, bandwidth, etc.).
2. Optionally, a gNB-CU performs ether an NG Setup or gNB Configuration Update procedure with the AMF (in most cases, this should not be needed, as the NG interface would be operational already).
3. If the gNB-CU accepts the request, it replies with the F1 Setup Response message, carrying the following information: gNB-CU Name, RRC version, and a list of cells to activate.
4. During the F1 interface lifetime, both the gNB-CU and the gNB-DU may notify each other about the change in their configuration or operational state, using gNB-CU/gNB-DU Configuration Update procedures.
5. If the receiving node accepts the new configuration, it replies with gNB-CU/gNB-DU Configuration Update Acknowledge (or Configuration Update Failure otherwise).
6. Optionally, the gNB-CU performs the Xn Setup procedure (in most cases, this should not be needed, as the Xn interface would be operational already).

The F1 Setup Request message, which is the first message sent by a gNB-DU on a newly established F1 interface, carries the following information:

- gNB-DU ID
- gNB-DU Name
- A list of gNB-DU served cells, with their information
- gNB-DU RRC version.

The gNB-DU ID and gNB-DU Name are used by the receiving gNB-CU to identify the sending NG-RAN network node (the Name IE is needed mainly for the purpose of human identification). The gNB-DU RRC version is used to prevent interoperability issues in case a gNB-DU and gNB-CU use different versions of the RRC specification. Every entry in the list of served cells contains the identifiers of a cell, for example, NR Physical Cell Identity (PCI), a list of served PLMNs (to support RAN sharing between multiple operators), and cell frequency and bandwidth information. This information can be used by a gNB-CU, for example, in decisions related to UE mobility (e.g. handover). Additionally, the message carries the gNB-DU System Information, that is Master Information Block (MIB) and System Information Block 1 (SIB1). Even though the gNB-DU is responsible for generating MIB and SIB1 parts of the overall system information, this information has to be available in the gNB-CU as well, in case the system information is provided to a UE in dedicated RRC signaling (as opposed to broadcast).

The F1 Setup Response message, which is sent back to a gNB-DU in the case of successful F1 interface establishment, carries the following information:

- gNB-CU Name
- A list of cells to be activated
- gNB-CU RRC version.

The gNB-CU Name and the RRC version IEs serve the same purpose as in the initiating message (see above). The list of cells to be activated has multiple purposes:

- To enable a gNB-CU to activate only some cells that a gNB-DU supports (e.g. for energy saving reasons);
- To indicate a list of PLMNs, which is needed for network sharing deployments;
- To carry the SI per cell that is generated by a CU (for the case when SI is delivered to a UE using broadcast signaling).

All the F1AP interface maintenance-related messages are specified in 3GPP TS 38.473, clause 9.2.1, and the detailed definition of the F1 Setup Request message can be found in 3GPP TS 38.473, clause 9.2.1.4.

#### *4.2.3.1.2 UE Context Management Procedures*

UE context management procedures are arguably the most important aspect of the F1 interface. With these procedures, a gNB-CU can establish, modify, or release a UE context in the gNB-DU. These procedures are also used to admit a new UE in a cell, for example, during an initial access, Secondary Node Addition in DC or EN-DC or a handover. These procedures are:

- UE Context Setup/Modification/Release
- UE Inactivity Notification
- Notify procedure.

Generally, a UE context management is controlled by a gNB-CU, which is the node that initiates most of the UE context management procedures. In certain cases, when a procedure needs to be triggered by a gNB-DU, the gNB-DU sends an indication to the gNB-CU, which triggers the relevant procedure.

The UE Context Setup Request procedure is used by a gNB-CU to establish a UE context in the gNB-DU. This procedure can only be fully understood in the context of general initial access, handover, or secondary cell addition procedures (involving air interface and also core network), which are described in Chapter 3. Here we only describe the F1 aspects of these procedures, assuming that the reader is familiar with the overall concept.

We further illustrate the F1 UE context management described above by the UE mobility procedure, which involves the UE Context Setup, Modification, and Release messages, shown in Figure 4.2.5. We then focus on the detailed description of the message establishing a UE Context in the gNB-DU and we encourage the reader to study further details about F1 maintenance procedures in 3GPP TS 38.473.

0. At the beginning, it is assumed that a UE is connected to the network and sends/receives uplink/downlink traffic.
1. Furthermore, the UE is assumed to be configured to report measurements to the network; when a measurement report is triggered, the UE sends a *Measurement Report* message to the source gNB-DU.
2. Since the mobility decisions are taken by a gNB-CU, the source gNB-DU sends an Uplink RRC Transfer message to the gNB-CU to convey the received *Measurement Report*.
3. If the gNB-CU decides to handover the UE, it sends an UE Context Setup Request message to the target gNB-DU to create an UE context and setup one or more bearers. This step assumes intra-gNB-CU mobility.

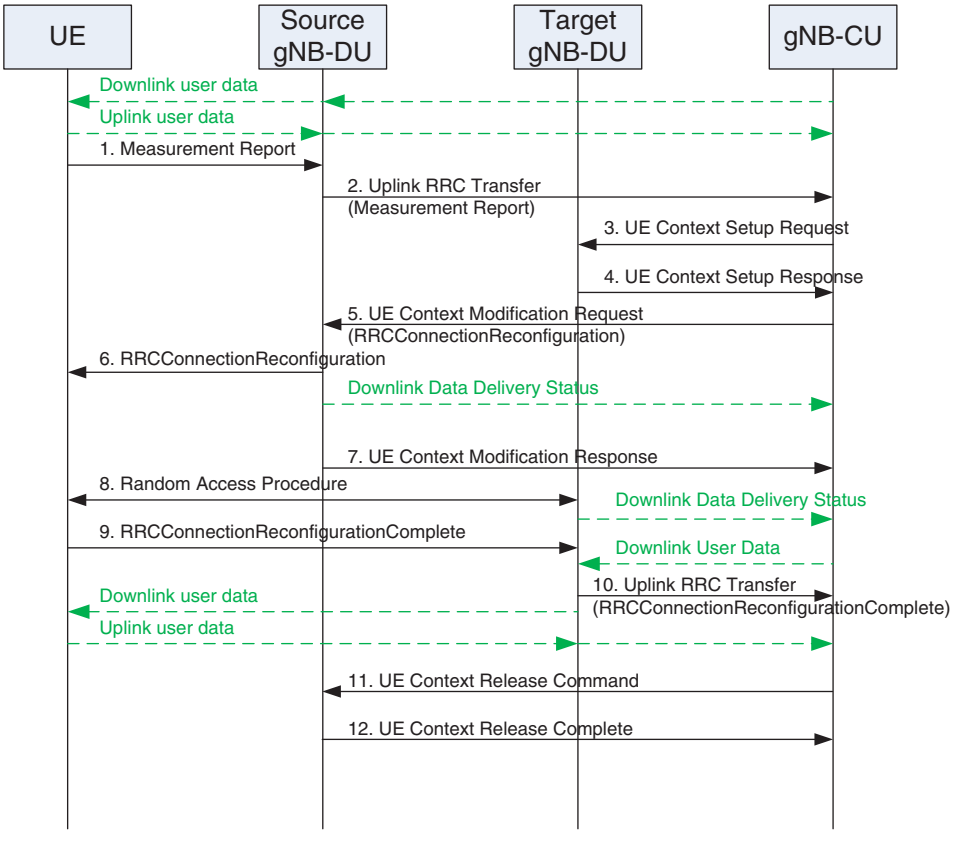



**Figure 4.2.5** Inter-gNB-DU mobility for intra-NR. (*Source:* Reproduced by permission of © 3GPP).

4. If the target gNB-DU accepts the request, it responds to the gNB-CU with an UE Context Setup Response message.
5. The gNB-CU sends an UE Context Modification Request message to the source gNB-DU, which includes a generated *RRCConnectionReconfiguration* message and indicates to stop the data transmission for the UE.
6. The source gNB-DU forwards the received *RRCConnectionReconfiguration* message to the UE. The source gNB-DU also sends a Downlink Data Delivery Status frame to inform the gNB-CU about the unsuccessfully transmitted downlink data to the UE.
7. The source gNB-DU responds to the gNB-CU with the UE Context Modification Response message to complete the context modification procedure.
8. The UE performs a Random Access procedure toward the target. The target gNB-DU sends a Downlink Data Delivery Status frame to inform the gNB-CU. Downlink packets, which may include PDCP PDUs not successfully delivered through the source gNB-DU, are sent from the gNB-CU to the target gNB-DU.[3]

3 It is up to gNB-CU implementation whether to start sending downlink user data to gNB-DU before or after reception of the Downlink Data Delivery Status.

9. The UE completes the Random Access procedure with an *RRCConnectionReconfigurationComplete* message.
10. The target gNB-DU sends an Uplink RRC Transfer message to the gNB-CU to convey the received *RRCConnectionReconfigurationComplete* message. Downlink packets are sent to the UE. Also, uplink packets are sent from the UE, which are forwarded to the gNB-CU through the target gNB-DU.
11. The gNB-CU sends an UE Context Release Command message to the source gNB-DU.
12. The source gNB-DU releases the UE context and responds to the gNB-CU with an UE Context Release Complete message.

The UE Context Setup Request message, used by a gNB-CU to establish a context for a new UE in the gNB-DU, carries UE identifiers on the F1 interface, information about cells for the UE, Signaling Radio Bearers (SRBs), and Data Radio Bearers (DRBs) for the UE, and others. While the explanation of all the IEs in that message is beyond the scope of this chapter, we focus on a few select IEs that are important to illustrate the concept of UE context establishment. Specifically:

- gNB-CU UE F1AP ID and gNB-DU UE F1AP ID are the identifiers of the UE on the F1 interface.
- Special Cell (SpCell) ID and a list of candidate SpCells (in non-multiconnectivity operation, SpCell refers to Primary Cell [PCell], which operates in the primary frequency and in which the UE performs initial connection establishment).
- CU to DU RRC Information, which is a collection of RRC containers carrying information defined in 3GPP TS 38.331, such as measurement configuration and handover preparation information.
- Discontinuous Reception (DRX) Cycle.
- Secondary Cell (SCell) to be setup list (in the case of Carrier Aggregation [CA]).
- A list of SRBs to be set up.
- A list of DRB to be set up.
- Inactivity monitoring request to trigger (optional) UE inactivity monitoring in the gNB-DU.
- RRC-Container with the DL-DCCH-Message IE, defined in (3GPP TS 38.331).
- Serving PLMN.
- gNB-DU UE Aggregate Maximum Bit Rate Uplink.

All the F1AP UE context management-related messages are specified in 3GPP TS 38.473, clause 9.2.2, and the detailed definition of the F1 UE Context Setup Request message can be found in 3GPP TS 48.473, clause 9.2.2.1. Multiple RRC containers referenced in 3GPP TS 38.473 are defined in 3GPP TS 38.331.

#### *4.2.3.1.3 RRC Message Transfer, SI and Paging Procedures*

Since the gNB-CU hosts the RRC protocol functionality, RRC and SI transfer procedures need to be defined, to carry the information to the gNB-DU, which then delivers it to a UE. An example of this has been already shown as part of the mobility procedure above (see Figure 4.2.5), where the procedure was used to transfer the measurement results. In theory, most RRC messages carried on SRBs could be transferred over the F1 interface in

exactly the same manner as the user-plane messages on DRBs. However, since the F1 user plane uses unreliable transport (i.e. GTP-U), 3GPP decided to specify RRC message transfer over the F1 control plane, which uses SCTP. This was done for increased reliability of control-plane information which is considered more important than the user plane. To this end, the following F1AP messages have been defined:

- Initial UL RRC Message Transfer
- DL RRC Message Transfer
- UL RRC Message Transfer.

They carry the RRC-Container IE with an actual RRC message, defined in the RRC specification (3GPP TS 38.331). These are largely transparent to gNB-DU, which simply transmits over the air to the UE upon reception from gNB-CU; however, they are typically carried over the F1 interface along with some supplementary information, for example, SRB ID in the case of DL RRC Message Transfer to indicate to a gNB-DU which SRB shall be used for that RRC message.

Despite the fact that according to the functional split architecture RRC resides in the gNB-CU, certain functions that otherwise could be considered as part of RRC are actually implemented in the gNB-DU. These are generally time-critical functions, the implementing of which in a gNB-CU would have affected the timing of sending them over the air because of the F1 interface latency. For example, MIB and SIB1, which need to be transmitted periodically with precise timing, are generated in gNB-DU – sending these from gNB-CU would be suboptimal. Therefore, even though RRC is considered a part of a gNB-CU, quite a few RRC procedures involve both gNB-DU and gNB-CU, and thus require F1AP signaling support, as these are not completely transparent to gNB-DU.

We further illustrate the concepts described above by the RRC inactive to RRC connected states transition procedure (for details about RRC states, see Chapter 3), which involves most of the RRC-related F1 messages: F1 Paging, Initial UL RRC Message Transfer, DL RRC Message Transfer, and UL RRC Message Transfer. This is illustrated in Figure 4.2.6. We then focus on the detailed description of the F1 Paging message and we encourage the reader to study further details about F1 RRC-related procedures in 3GPP TS 38.473.

1. When downlink data are received from the 5GC, the gNB-CU sends the F1AP Paging message to the gNB-DU with the paging-related information (most of which it receives from the 5GC), such as paging DRX, paging priority, and list of cells in which to page a UE.
2. Using the information received from the gNB-CU, the gNB-DU constructs the RRC paging message and sends it to a UE over the air interface.
3. If a UE receives the RRC paging message triggering the UE to resume the RRC connection, it sends the RRC Resume request to the gNB-DU.
4. As the RRC functionality is implemented in the gNB-CU, the gNB-DU includes the RRC Resume request received from the UE in a non-UE-associated F1AP Initial UL RRC Message Transfer message and transfers it to the gNB-CU.
5. To establish an association for that UE over the F1 interface, the gNB-CU allocates gNB-CU UE F1AP ID and sends it in the F1AP UE Context Setup Request message to the gNB-DU. The message may also include SRB ID(s) and DRB ID(s) to be set up, and

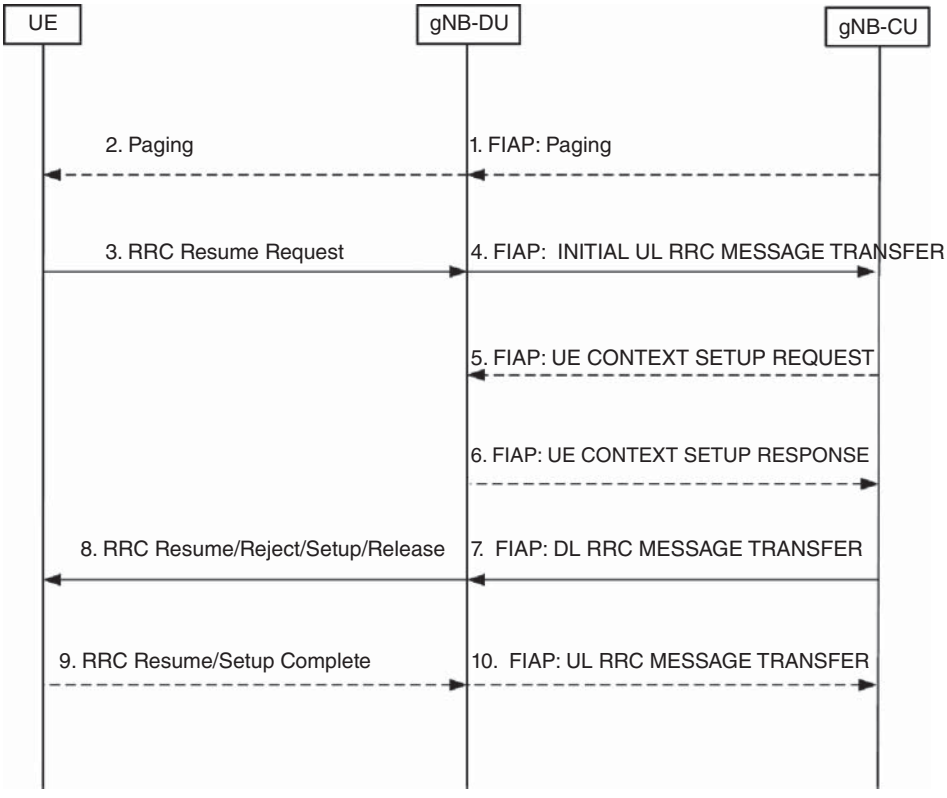



**Figure 4.2.6** RRC inactive to other RRC states transition procedure. (*Source:* Reproduced by permission of © 3GPP).

CellGroupConfig stored in gNB-CU or retrieved from the old NG-RAN node, which are used by the gNB-DU to communicate with the UE.

6. If the procedure succeeds, the gNB-DU responds with the F1AP UE Context Setup Response message carrying the gNB-DU UE F1AP ID that it has allocated – this completes the establishment of the F1 connection for that UE. The messages also contains a list of successfully established SRBs and DRBs, along with their RLC/MAC/PHY configuration, the list of bearers the gNB-DU failed to establish, and other UE-related information, which is handled by the gNB-DU but must be known to the gNB-CU as well.
7. Using the information received in step 6 from the gNB-DU, the gNB-CU generates the RRC Resume message for the UE. The RRC message is encapsulated in the F1AP DL RRC Message Transfer message together with the corresponding SRB ID and sent to the gNB-DU.
8. The gNB-DU forwards the RRC message received in step 7 to the UE either over SRB0 or SRB1 as indicated by the SRB ID.
9. The UE sends the RRC Resume Complete message to the gNB-DU.
10. The gNB-DU encapsulates the received RRC message in the F1AP UL RRC Message Transfer message and sends it to the gNB-CU, which completes the RRC state transition procedure.

The paging procedure explained above illustrates why some RRC functions involve both gNB-DU and gNB-CU. For example, a paging message can arrive at a gNB-CU either from core network (i.e. AMF) or another NG-RAN node (i.e. gNB). It is then processed by gNB-CU and gNB-DU as follows: gNB-CU determines what is sent in a paging message and gNB-DU determines how the message is sent over the air interface. This is because the information about which UEs to page in which cells is received at a gNB-CU (either from core network or from another NG-RAN node); however, gNB-CU cannot determine by itself scheduling-related parameters – only gNB-DU is aware of these. Therefore, the gNB-CU sends to the gNB-DU either a core network or RAN UE paging identity, paging priority, and a list of cells to page the UE using the F1AP Paging message. The gNB-DU determines a Paging Occasion (PO) and a Paging Frame (PF), and eventually sends the paging message to the UE over the air.

All the F1AP messages for RRC message transfer are specified in 3GPP TS 38.473, clause 9.2.3, and Paging messages in 3GPP TS 38.473, clause 9.2.6.

#### 4.2.3.2 User-Plane Protocol

The F1-U user-plane protocol design also follows the same principles as many other user-plane protocols used on various RAN interfaces. It is based on GTP-U, with the protocol stack described in Section 3.3. A GTP-U tunnel is mapped to an NG-RAN bearer one-to-one and a single F1-U GTP-U PDU carries a single PDCP PDU. Furthermore, the NR user-plane protocol is applicable for the Xn interface (3GPP TS 38.425) as well, hence sequence numbering, retransmissions (if needed), and flow control are supported for PDCP PDUs carried over the NG-RAN user-plane interfaces.

More details about the F1-U protocol are can be found in 3GPP TS 38.425.

#### 4.2.3.3 OAM Aspects

As mentioned above, both gNB-DU and gNB-CU need to have certain information pre-configured before they establish the F1 interface between them – this is performed via OAM. It is generally assumed that a gNB-CU and a gNB-DU have independent management interfaces and are managed by the OAM separately. In particular, this means that certain information needed for NG-RAN to operate is preconfigured via OAM in a gNB-CU and a gNB-DU and the peer nodes are not necessarily aware of each other's configuration. Therefore, F1AP signaling (e.g. configuration update) is used by gNB-DU and gNB-CU to exchange configuration information, which has been configured in each node by OAM.

One example that illustrates the principle described above is the SI. Generally, gNB-DU is responsible for generating MIB and SIB1 system information, while the rest of the system information is controlled by gNB-CU. The information for the MIB and SIB1 is configured by OAM. However, since in certain cases (e.g. SI delivery using dedicated RRC signaling) a gNB-CU needs to know that information, whenever it is changed a gNB-DU propagates that information to the gNB-CU using F1AP signaling.

For additional information about OAM support for CU/DU split, see Section 6.5. Additional details related to OAM are specified in 3GPP TS 28.541.

### 4.2.4 Further Reading

Full information about the CU/DU split NG-RAN architecture and related protocols described in this chapter can be found in the 3GPP technical specifications provided below.

3GPP TS 38.401 is the general stage-2 specification covering all the NG-RAN aspects, including the CU/DU split, and it is a good starting point for an interested reader to understand the details beyond what is described in this chapter. Once a reader has familiarized himself with the high-level aspects, we suggest learning the details of the control-plane signaling, defined in 3GPP TS 38.473. To understand the user-plane aspect, a reader must proceed to read both 3GPP TS 38.425 and 3GPP TS 29.281.[4] Finally, 3GPP TS 28.541 can be used to gain an understanding of OAM aspects of the CU/DU split, noting that this specification covers other nodes and architectures as well.

## References

3GPP Technical Report 38.801 (2017). Study on new radio access technology: Radio access architecture and interfaces. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.541 (2019).Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 29.281 (2019). General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.401 (2019). NG-RAN; Architecture description. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.425 (2019). NG-RAN; NR user plane protocol. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.470 (2019). NG-RAN; F1 general aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.471 (2019). NG-RAN; F1 layer 1. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.472 (2019). NG-RAN; F1 signalling transport. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.473 (2019). NG-RAN; F1 Application Protocol (F1AP). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.474 (2019). NG-RAN; F1 data transport', www.3gpp.org (accessed May 29, 2020).

Request for Comments 4960 (2007). The Internet Engineering Task Force (IETF), Network Working Group, Stream Control Transmission Protocol. Available at: www.ietf.org (accessed May 29, 2020).

4 Those specifications define the user-plane protocol used on many NG-RAN and 5GC interfaces and not everything is relevant to the F1 interface.

## 4.3 Multi-Radio Dual Connectivity

*Sergio Parolari*

*ZTE Corporation, Italy*

Most of the NG-RAN architectures described elsewhere in this chapter (e.g. in the previous section) are concerned with various options for splitting the gNB functionality into multiple logical network nodes. Multi-radio dual connectivity (MR-DC) described in the present section is different as, in this case, gNB functionality is split across two fully functional base stations, which are otherwise capable of serving UEs by themselves, that is, also without the cooperation of another base station.

From the UE point of view, MR-DC refers to the technology where a UE is simultaneously connected to two different radio access network nodes, one providing NR access and the other one providing either LTE or NR access. This allows a UE with multiple RX/TX capabilities to increase the data transfer rate thanks to simultaneous reception and transmission over the two different radio links. Unlike some other RAN architectures described in the book, MR-DC is not just a choice of RAN architecture, as it has substantial UE impacts. Therefore, even though the primary focus of the book is RAN architecture, in the present chapter we also describe certain UE-related aspects, which are important to fully understand how MR-DC works.

From the network architecture point of view, one of the two network nodes serving the UE acts as the Master Node (MN), that is, the access node that terminates the control-plane connection with the core network and provides primary radio resources via one or more cells – the Master Cell Group (MCG) – under its control. The other node acts as a Secondary Node (SN), that is, an access node with no control-plane connection to the core network (and with or without a user-plane connection to the core network) providing additional radio resources to the UE via one or more additional cells, that is, the Secondary Cell Group (SCG).

5G MR-DC is the evolution of LTE Dual Connectivity (DC) defined as early as Release-12. Even though LTE DC has not been widely deployed, it heavily influenced a number of technologies defined in 3GPP, ranging from LTE-WLAN Aggregation (LWA) to 5G MR-DC.

The MN and SN are connected together via a network interface that allows the exchange of control-plane and user-plane information. The MN and the SN are logical nodes, which may be deployed in a single physical node or separate ones, as is the case for all other network architecture options described in the book.

It is generally assumed that the MN node is a macro base station providing coverage, whereas the SN node is a small cell providing additional capacity; however, the technology is designed in such a way that other deployments are also possible.

One specific flavor of MR-DC, referred to as E-UTRA-NR dual connectivity (EN-DC), is particularly important as it is expected to be used in the first phase of 5G deployments. In EN-DC, a 5G "capacity layer" is added to an existing 4G network in the form of 5G SN nodes, while using the 4G Evolved Packet Core (EPC) network, as opposed to 5GC. This allows 5G deployment with relatively low capital expenditure (CAPEX) (as opposed to upgrading the whole network) and provides an evolution path toward "full" 5G deployment.

### 4.3.1 Key Ideas

- From the UE point of view, MR-DC allows simultaneous connection to two different radio access nodes, one providing NR access and the other one providing either LTE or NR access. From the network point of view, one network node (typically deployed for coverage) serves as the MN and the other network node (typically deployed for capacity) serves as the SN.
- Different MR-DC network architecture options are supported in 5G (EN-DC, NGEN-DC, NE-DC, and NR-DC), which differ in connectivity to the core network (EPC or 5GC) and in which nodes serve as MN or SN.
- One particular MR-DC flavor (EN-DC) is especially important for early 5G deployment, as it allows gradual 5G roll out while still using the legacy EPC CN. EN-DC has been selected as the first 5G deployment option by many operators.
- EN-DC uses the legacy network interfaces X2 and S1, while MR-DC connected to 5GC relies on new interfaces called Xn and NG corresponding to X2 and S1, respectively.
- In order to support spectrum sharing between LTE and NR, network interfaces used for MR-DC allow for resource coordination between MN and SN network nodes.
- Control-plane MR-DC aspects covered in this section include: dual RRC architecture, split SRB, SRB3, UE capability coordination, radio resource coordination, measurement framework, and security.
- User-plane aspects of MR-DC operation covered in this section include: different bearer types (MN and SN terminated, MGG/SCG/split bearer), QoS aspects, and bearer type selection.

### 4.3.2 MR-DC Options

Several different MR-DC variants have been specified by 3GPP.

The first option, providing connectivity to the EPC, is EN-DC, in which a UE is connected to one eNB (providing E-UTRA access) that acts as an MN and one gNB (providing NR access) that acts as an SN.

In this architecture, the gNB may also be referred to as en-gNB, to discriminate this from a gNB that connects to the 5GC. Similarly, in some architectures, when an eNB is connected to a 5GC it is referred to as ng-eNB. That is, the prefix "en-" indicates connectivity to EPC and the usage of X2 and S1 network interfaces, whereas the prefix "ng-" indicates connectivity to 5GC and the usage of Xn and NG network interfaces. From a 3GPP specifications point of view, strictly speaking EN-DC is not an NG-RAN architecture, as it uses "old" network interfaces (X2 and S1) and EPC. Nevertheless, it is described in the book as this is an important 5G deployment option.

Figure 4.3.1 illustrates the EN-DC architecture, in which the eNB is connected to the EPC via the S1 interface and to the gNB via the X2 interface. Optionally, the gNB might also be connected to the EPC via the S1-U interface.

All the other MR-DC options provide connectivity to the 5GC.

In NG-RAN, E-UTRA-NR dual connectivity (NGEN-DC) illustrated in Figure 4.3.2, a UE is connected to one ng-eNB that acts as an MN (connecting the UE to the 5GC) and one gNB that acts as an SN connected to the ng-eNB via the Xn interface. In NR-E-UTRA dual

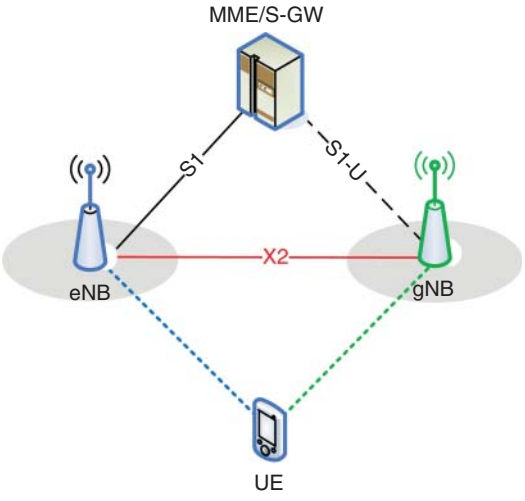



**Figure 4.3.1** EN-DC architecture. (*Source:* Reproduced by permission of © 3GPP).

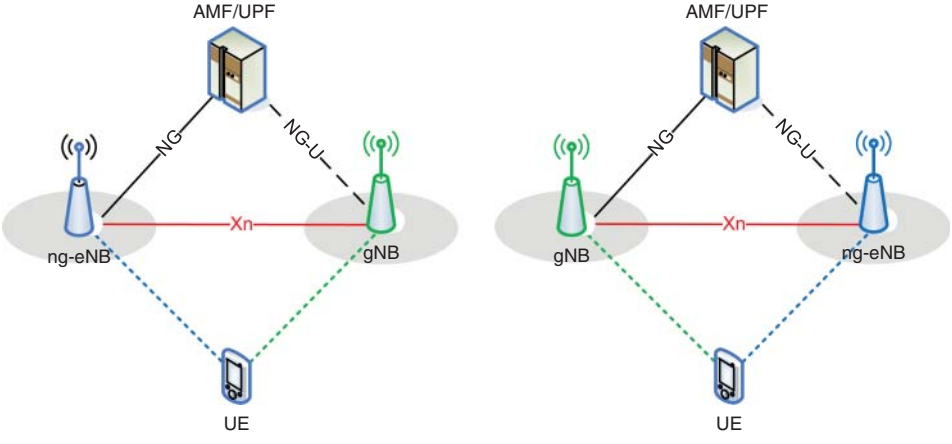



**Figure 4.3.2** NGEN-DC (left) and NE-DC (right) architectures. (*Source:* Reproduced by permission of © 3GPP).

connectivity (NE-DC) also shown in Figure 4.3.2, a UE is connected to one gNB that acts as an MN (again connecting the UE to the 5GC) and one ng-eNB that acts as an SN.

Finally, in NR-NR dual connectivity (NR-DC), a UE is connected to one gNB that acts as an MN and another gNB that acts as an SN (see Figure 4.3.3). The master gNB is connected to the 5GC via the NG interface and to the secondary gNB via the Xn interface. The secondary gNB might also be connected to the 5GC via the NG-U interface. NR-DC can also be used to provide a UE with NR access via two separate gNB-DUs, serving different cells, connected to the same gNB-CU, acting both as an MN and as an SN.

### 4.3.3 Market Drivers

The full benefits of 5G can only be exploited by a 5G "Standalone" architecture, where gNBs providing NR access are directly connected to the 5G core network. This architecture is the

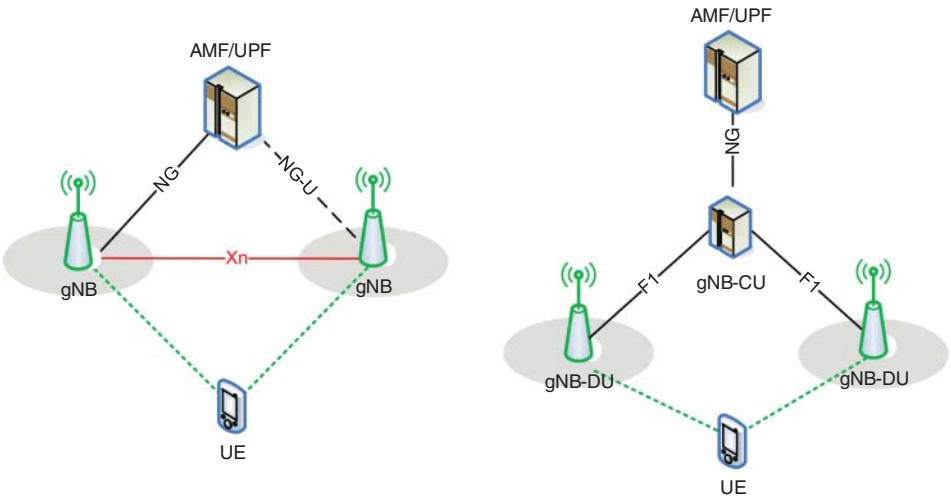



**Figure 4.3.3** NR-DC inter-gNB (left) and intra-gNB (right) architectures. (*Source:* Reproduced by permission of © 3GPP).

final destination of the technological migration in many operators' networks, but it is also the 5G architecture used from day one in several deployments in different parts of the world. Whenever the Standalone architecture is not adopted from the very beginning, EN-DC is useful to realize early 5G deployments because it allows the benefits of NR technology to be experienced, with its increased data rates and the possibility to exploit new frequency bands, while fully reusing the existing 4G infrastructure. In fact, EN-DC is the architectural option that allows the deployment of NR, that is a 5G radio access technology, without the need to deploy a 5G core network at the same time, but simply reusing the EPC, that is the 4G core network. It also does not require ubiquitous NR coverage from the very beginning, as it is always possible to fall back to the connectivity provided by the LTE network and maintain service without interruptions due to NR coverage holes. Thanks to this attribute of allowing NR access while relying on an existing LTE infrastructure to be able to connect to the operator's network, EN-DC is also known as the 5G "Non-Standalone" (NSA) option and it was also the first 5G solution specified in the first version (known as the "early drop") of 3GPP Release-15.

The other MR-DC options, NGEN-DC, NE-DC, and NR-DC, were introduced in the last version (i.e. the "late drop") of 3GPP Release-15.[5]

NGEN-DC shares the same benefit as EN-DC of allowing the reuse of an existing LTE radio infrastructure to provide basic service, which can be enhanced using the dual connectivity functionality when NR coverage is available. On top of this, NGEN-DC allows connection to the 5GC and the related services and enhancements, for example, including the support of the 5G QoS framework.

5 3GPP Release-15 is different compared with other releases not only because this is the first release introducing 5G, but also because it effectively contains three "sub-releases," also referred to as "drops": EN-DC/NSA in the "first drop," "standalone"/full 5G released at the originally planned scheduled date, and NGEN-DC, NE-DC and NR-DC in the "late drop."

NE-DC also allows the reuse of an existing LTE radio infrastructure. The difference in this case is that the anchor is the NR access network and the LTE access is used to increase the performance. For instance, this architectural option is useful when the NR network operates in a lower band than the LTE one; then providing an even better coverage than the existing LTE network.

While EN-DC is arguably the most important MR-DC option, at least in the first phase of 5G deployment, the benefits of NGEN-DC and NE-DC are somewhat less evident. While these options do have certain benefits, as described in this chapter, it is not clear whether these benefits justify the cost of their deployments. At the time of writing this book, the majority of 5G deployment announcements and plans are centered around EN-DC and "Standalone" 5G.

The benefit of NR-DC is different: it is not related to the reuse of legacy core and radio access networks, as it actually applies to a 5G Standalone architecture and it can be compared with the use of the NR CA feature. CA already allows the combination of different NR carriers (also belonging to different bands) to increase the maximum data rate for UEs supporting such features. However, CA requires that the different carriers are tightly synchronized and controlled by the same MAC entity. In practice this means that, on the network side, the different carriers need to be controlled by the same gNB, or possibly by different access nodes connected with an ideal backhaul (e.g. in a proprietary manner, as opposed to using a standardized Xn network interface). Also, in cases where a split architecture is used, with a single gNB-CU connected to multiple gNB-DUs (e.g. each one operating in a different frequency band) inter-band CA is not possible. NR-DC can then be used to combine the radio resources controlled by different gNBs connected via a non-ideal backhaul with standardized Xn network interface (inter-gNB NR-DC) or by different gNB-DUs connected to the same gNB-CU via a F1 network interface (intra-gNB NR-DC).

### 4.3.4 Functional Description

#### 4.3.4.1 Control Plane

From the UE point of view, each UE configured with MR-DC has one single control-plane connection to the corresponding core network entity (Mobility Management Element [MME] or AMF), through the MN.

From the network point of view, there is an interface (X2 or Xn, depending on the MR-DC option) between the MN and the SN for coordination messages between them and for exchanging RRC control messages intended for the UE. The interface is used for example for SN node addition, modification, and release messages used to establish an MR-DC operation for a UE, to modify it, or to release it, respectively. The details of these message exchanges are described in more detail below.[6]

In particular, in EN-DC, the MN and the SN are interconnected via X2-C, and the involved core network entity is the MME, to which MN is connected via S1-MME. In the MR-DC options with connection to the 5GC (NGEN-DC, NE-DC, and NR-DC), the involved core

6 General principles of the Xn interface operation are described in Section 3.3. In the present section we focus on Xn functions defined for MR-DC operation. The X2 interface in general and MR-DC messages and procedures in particular are very similar to those defined for Xn.

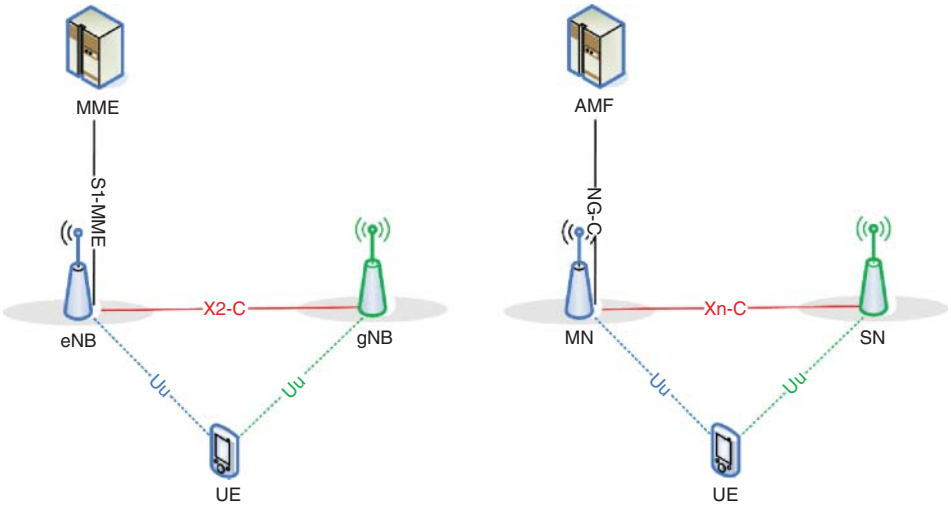



**Figure 4.3.4** Control-plane connectivity for EN-DC (left) and MR-DC with 5GC (right). (*Source:* Reproduced by permission of © 3GPP).

network entity is the AMF, NG-C interface is used between the AMF and the MN, and the MN and the SN are interconnected via Xn-C.

The control-plane connectivity for the different MR-DC options is shown in Figure 4.3.4.

Both radio nodes have their own RRC entity that can generate RRC messages to be sent to the UE. Each node, that is the MN or the SN, is responsible for the handling of its group of cells (called MCG and SCG, respectively) and then for the generation of the corresponding MCG/SCG RRC Reconfiguration messages. RRC messages generated by the SN can be conveyed from the SN to the MN via the X2/Xn interface and then encapsulated in an RRC message generated by the MN that may also carry an MCG reconfiguration decided by the MN. The combined configuration can be jointly processed by the UE and the UE uses a joint success/failure procedure for RRC messages transmitted by the MN and encapsulating an RRC message generated by the SN. Each RRC Reconfiguration message has its own RRC response message even when the RRC message is encapsulated in another RRC message. The RRC response message for the SN can then be encapsulated in the RRC response message for the MN and then forwarded over X2/Xn to the SN.

If the SN is a gNB (i.e. for EN-DC, NGEN-DC, and NR-DC), the UE can also be configured to establish an SRB with the SN, called SRB3. SRB3 is different from the other SRBs (SRB0, SRB1, and SRB2) established between the UE and the network, as it is defined exclusively for MR-DC options where the SN is a gNB, to enable SN RRC messages to be sent directly between the UE and the SN. RRC Reconfiguration messages generated by the SN can only be transported directly to the UE if the reconfiguration does not require any coordination with the MN. Additionally, measurement-reporting messages for mobility within the SN can be transmitted directly from the UE to the SN, if SRB3 is configured.

The dual connectivity functionality provided by MR-DC can also be applied to the SRBs. The split SRB option, supported for all MR-DC options, allows the duplication of RRC messages generated by the MN, via the direct path and via the SN. However, duplication of RRC

messages generated by the SN is not possible. In other words, split SRB is supported for both SRB1 and SRB2, but not for SRB3. For downlink, duplication of RRC messages over split SRB is up to network implementation. For uplink, the UE uses the MCG path or the SCG path depending on configuration from the MN.

For full details of network functions defined for MR-DC refer to 3GPP TS 36.423, sub-clause 9.1.4, for X2-based MR-DC options, and 3GPP TS 38.423, sub-clause 9.1.2, for Xn-based MR-DC options.

#### *4.3.4.1.1 UE Capability Coordination*

A UE configured with MR-DC needs to share its baseband and RF capabilities between the two connections to the network. As a consequence, the network needs to know and then coordinate the capabilities that can be used over the two links. More specifically, the capabilities of a UE supporting MR-DC are included in different containers. The capabilities that need to be visible to both MN and SN are carried in a specific MR-DC container, including for instance the supported MR-DC band combinations. Other capabilities that only need to be visible to the node of the concerned RAT are then contained in two separate E-UTRA and NR capability containers.

When the MN needs to retrieve the MR-DC related capabilities, the MN transmits an RRC UE Capability Enquiry message to the UE, providing an MR-DC filter to retrieve the corresponding MR-DC-related capabilities in MR-DC, E-UTRA, and NR capability containers. The MN then stores the retrieved capabilities and the corresponding filter in the core network for later use.

A number of UE capabilities require coordination between E-UTRA and NR: band combinations, baseband processing capabilities, and the maximum power for FR1 the UE can use in SCG. For these capabilities the MN decides how to resolve the dependency between MN and SN configurations. The MN then provides the resulting UE capabilities usable for SCG configuration to the SN. The SN may also indicate the desired UE capabilities to be used for SCG configuration. In this case it is up to the MN to accept or reject the request from the SN.

#### *4.3.4.1.2 Radio Resource Coordination*

The MN and SN may also coordinate the use of their radio resources, for instance when a UE configured with MR-DC cannot simultaneously receive or transmit over both the links at the same time. Another important case is the one of spectrum sharing between LTE and NR. The MN and SN can exchange messages over the X2 and Xn interfaces to coordinate the use of the radio resources in a time-division multiplexed (TDM) manner. For a given UE, this is done via UE-specific signaling, including the Resource Coordination Information IEs in the relevant messages during the SN Addition/Modification procedures. On the other hand, the spectrum sharing is done via non-UE-associated signaling, using the E-UTRA–NR Cell Resource Coordination Request/Response messages. To this end, either the MN or the SN can include the Data Traffic Resource Indication IE, indicating in a semi-static manner the resource block allocation to LTE or NR.

#### *4.3.4.1.3 Measurement Configuration and Coordination*

In MR-DC, UE measurements on neighbor cells can be configured independently by the MN and by the SN (only for intra-Radio Access Technology [RAT] measurements), with

some coordination to ensure that UE capabilities in terms of measurement configuration are not exceeded.

Measurements configured to the UE in preparation for the procedure that establishes MR-DC, that is the SN Addition procedure described in Section 4.3.4.3.1, are configured by the node serving the UE, and becoming the MN when the procedure successfully completes. But in the case of intra-secondary node mobility, as described in Section 4.3.4.3.2, UE measurements are configured by the SN, if required in coordination with the MN. The SN Change procedure, described in Section 4.3.4.3.3, can be triggered by both the MN (only for inter-frequency SN Change) and the SN. In the first case, UE measurements are configured by the MN but in the latter they are configured by SN, which also processes the measurement reports, with no need for providing the measurement results to the MN.

Measurement reports for measurements configured by the SN are sent by the UE directly to the SN on SRB3, if SRB3 is configured. Otherwise, reports for measurements configured by the SN are sent on SRB1 to the MN first, and then forwarded via the X2 or Xn network interface to the SN. Measurement results can be exchanged among the involved network nodes during the different mobility procedures. In an SN Change procedure initiated by the MN, measurement results related to the target SN can be provided by the MN to the target SN. In an SN Change procedure initiated by the SN, measurement results of the target SN can be forwarded from the source SN to the target SN via the MN. In an inter-MN handover procedure, measurement results related to the SN can be provided by the source MN to the target MN.

To be able to perform measurements on the different serving or not serving frequencies, a UE might need to be configured with gaps during which measurements can be performed. A UE can be configured with a single measurement gap configuration ("per-UE" measurement gap) or with two measurement gap configurations ("per-FR" measurement gaps): one for the lower frequency range (FR1, below 7 GHz) and one for the upper frequency range (FR2, above 7 GHz), depending on the UE capability to support independent measurements for different frequency ranges. If per-UE gap is used, the decision on the gap configuration is taken by the MN. If per-FR gap is used, in EN-DC and NGEN-DC, the MN decides the gap configuration for FR1, while the SN decides the gap configuration for FR2; in NE-DC and NR-DC, the MN decides both the FR1 and FR2 gap configurations.

#### *4.3.4.1.4 Security-Related Aspects*

A UE can be configured with MR-DC only after security activation in the MN. The security key used by the UE depends on whether the bearers are terminated in the MN or in the SN. For bearers terminated in the MN, the UE is configured to use the master security key. For bearers terminated in the SN, the UE is configured to use the secondary security key, which is derived from the master security key being used with the MN.

In 5G, integrity protection, which was only applied to SRBs in 4G, can be also used for user-plane data on a PDU session basis. This is also applicable to MR-DC, where all DRBs belonging to the same PDU session have the same user-plane integrity protection configuration (i.e. on or off), no matter whether a PDU session is served by both MN and SN in MR-DC with 5GC.

Considering that in NR an intra-gNB handover does not necessarily require a security key change (i.e. in cases where gNB-CU is not changed), whenever the MN is a gNB (i.e. in

NE-DC and NR-DC), a PCell change without a security key change does not require a secondary security key change. Similarly, whenever the SN is a gNB (i.e. in EN-DC, NGEN-DC, and NR-DC), for a Primary Secondary Cell Group Cell (PSCell) change that does not require a corresponding master security key change (e.g. when there is no simultaneous PCell handover in EN-DC and NGEN-DC), a secondary security key refresh is not required if the PDCP termination point of the SN is not changed. On the other hand, in NE-DC, a PSCell change always requires a secondary security key change.

#### 4.3.4.2 User Plane

Both MN and SN network nodes have user-plane interfaces to the CN (S-GW in the case of EPC and UPF in the case of 5GC). A UE configured with MR-DC has a user-plane connection to the corresponding core network entity, through the MN, the SN, or both. Furthermore, there is a network interface (X2 or Xn) between the MN and the SN for transferring data between the two nodes.

In particular, in EN-DC, the involved core network entity is the S-GW – S1-U is terminated in the MN and/or the SN, and the MN and the SN are interconnected via X2-U. In the MR-DC options with connection to the 5GC (NGEN-DC, NE-DC, and NR-DC), the involved core network entity is the UPF – NG-U is terminated in the MN and/or the SN, and the MN and the SN are interconnected via Xn-U.

All four network interfaces (S1-U, X2-U, NG-U, and Xn-U) involved in MR-DC operation use the same protocol stack, with GTP-U protocol used on top of UDP. Additionally, X2-U and Xn-U also provide flow control functionality and enhancements for lossless packet delivery (in case the transport network is not reliable), similar to the ones provided by the F1-U interface (described in Section 4.2.4).

An overview of the user-plane connectivity for the different MR-DC options is shown in Figure 4.3.5.

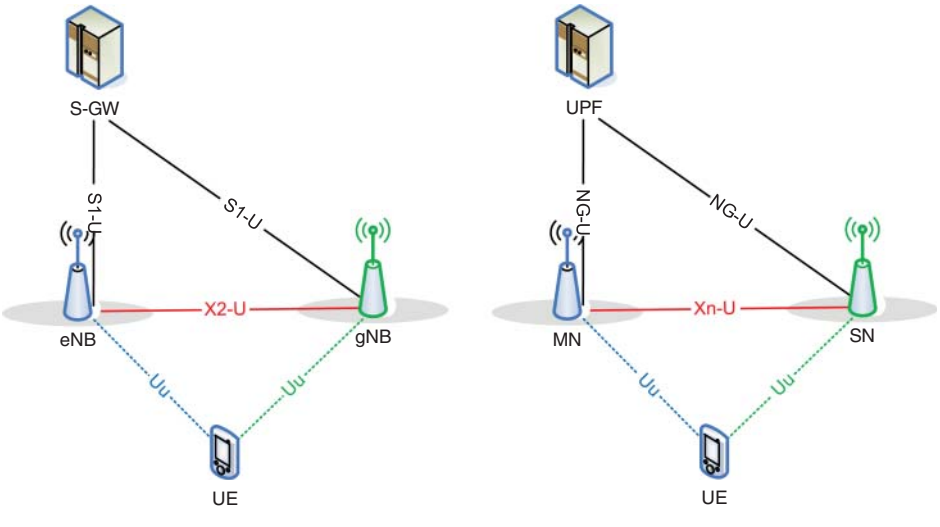



**Figure 4.3.5** User-plane connectivity for EN-DC (left) and MR-DC with 5GC (right). (*Source:* Reproduced by permission of © 3GPP).

For further details on network user-plane protocols used on network interfaces S1-U, X2-U, NG-U, and Xn-U refer to 3GPP TS 36.414 for S1-U, 3GPP TS 36.424 and TS 36.425 for X2-U, 3GPP TS 38.414 and TS 38.415 for NG-U, and 3GPP TS 38.424 and TS 38.425 for Xn-U.

#### *4.3.4.2.1 Bearer Types*

The user-plane connection to the core network is performed via the MN or the SN depending on the bearers that are configured for a UE. In fact, a UE can be configured with multiple bearer types in MR-DC.

From a network perspective, a UE can be configured with *MN terminated bearers*, where the user-plane connection to the core network entity is terminated in the MN, and *SN terminated bearers*, where the user-plane connection to the core network entity is terminated in the SN. Regardless of the network termination point, bearers can be further categorized into *MCG bearers*, *SCG bearers*, or *split bearers*, depending on whether the transport of user-plane data over the Uu interface involves MCG, SCG radio resources, or both. For *MCG bearers*, only MCG radio resources are used; for *SCG bearers*, only SCG radio resources are used; while for *split bearers*, both MCG and SCG radio resources are used. For *split bearers*, *MN terminated SCG bearers*, and *SN terminated MCG bearers*, user-plane data are transferred between the MN and the SN via the X2-U/Xn-U interface.

A summary of all the possible MR-DC bearer types, from a network perspective, for the EN-DC case is shown in Figures 4.3.6 and 4.3.7.

It is important to note that even if only SCG bearers are configured for a UE (i.e. no MCG resources are used to exchange user-plane data with the network), SRB1 and SRB2 always use at least MCG resources. In other words this is still an MR-DC configuration. Also, if only MCG bearers are configured for a UE (i.e. no SCG resources are configured), this is still considered as MR-DC as long as at least one of the bearers is terminated in the SN.

In terms of protocol stack, at the network side, for each radio bearer the PDCP entity (for EN-DC) or the SDAP/PDCP entities (for NGEN-DC, NE-DC, and NR-DC) is always hosted by the node that terminates the radio bearer: in the MN for MN terminated bearers and in the SN for SN terminated bearers. Furthermore, a radio bearer may have an RLC and MAC

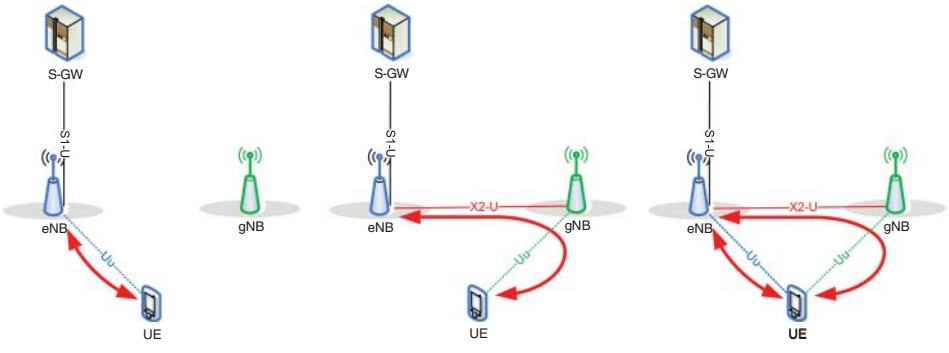



**Figure 4.3.6** MN terminated bearers: MCG bearer (left), SCG bearer (center), split bearer (right).

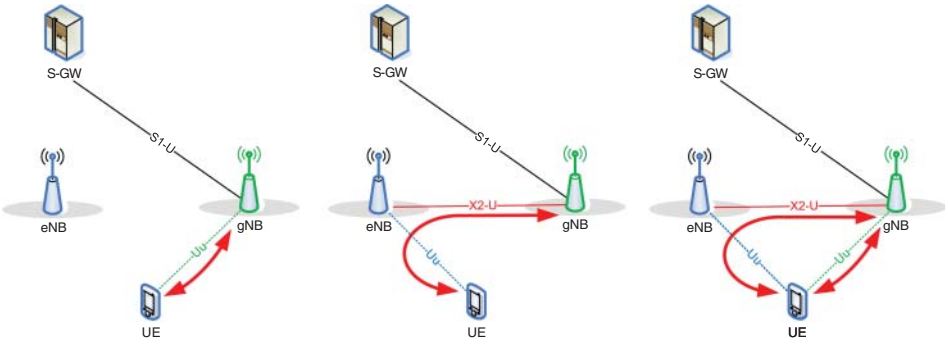



**Figure 4.3.7** SN terminated bearers: SCG bearer (left), MCG bearer (center), split bearer (right).

logical channel configuration (i.e. an RLC bearer) in one cell group (for MCG bearers and SCG bearers) or two cell groups (for split bearers).

For EN-DC, the network can configure either E-UTRA PDCP or NR PDCP for MN terminated MCG bearers, while NR PDCP is always used for all other bearers. However, in MR-DC with 5GC, NR PDCP is always used for all bearer types. In (NG)EN-DC, E-UTRA RLC/MAC is used in the MN while NR RLC/MAC is used in the SN. In NE-DC, NR RLC/MAC is used in the MN while E-UTRA RLC/MAC is used in the SN. In NR-DC, NR RLC/MAC is used in both MN and SN.

The user-plane radio protocol architecture at the network side for the different bearer types is shown in Figure 4.3.8 for EN-DC, and Figure 4.3.9 for MR-DC with 5GC (for NGEN-DC, NE-DC, and NR-DC).

The user-plane handling at the UE side is independent of where the bearer is terminated on the network side. So from a UE perspective only three bearer types exist: MCG bearer, SCG bearer, and split bearer. The user-plane radio protocol architecture for the three bearer types from the UE point of view is shown in Figure 4.3.10 for EN-DC, and Figure 4.3.11 for MR-DC with 5GC (NGEN-DC, NE-DC, and NR-DC).

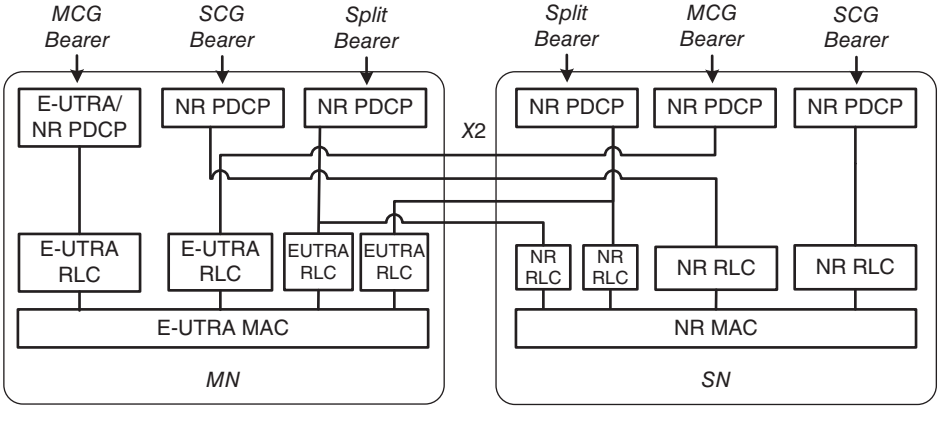



**Figure 4.3.8** Radio protocol architecture at the network side for MCG, SCG, and split bearers in EN-DC. (*Source:* Reproduced by permission of © 3GPP).

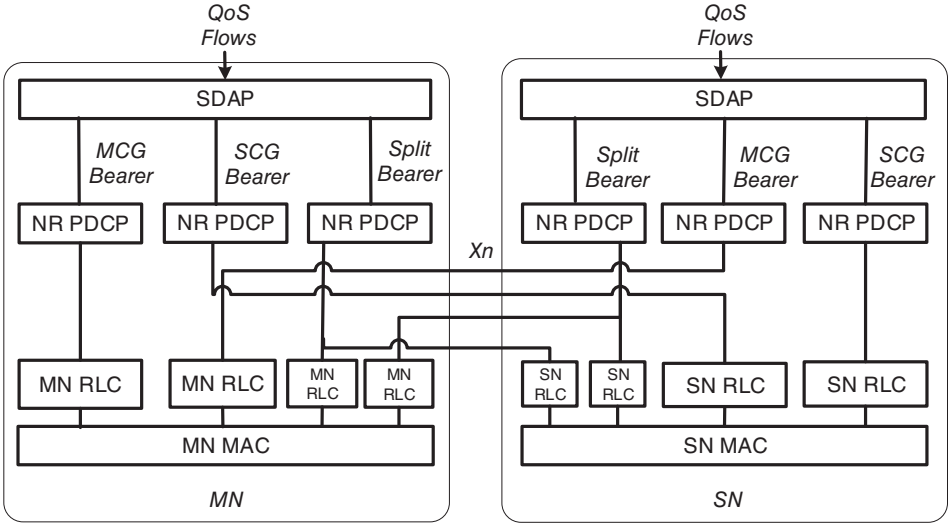



**Figure 4.3.9** Radio protocol architecture at the network side for MCG, SCG, and split bearers in MR-DC with 5GC (NGEN-DC, NE-DC, and NR-DC). (*Source:* Reproduced by permission of © 3GPP).

**Figure 4.3.10** Radio protocol architecture for MCG, SCG, and split bearers from a UE perspective in EN-DC. (*Source:* Reproduced by permission of © 3GPP).

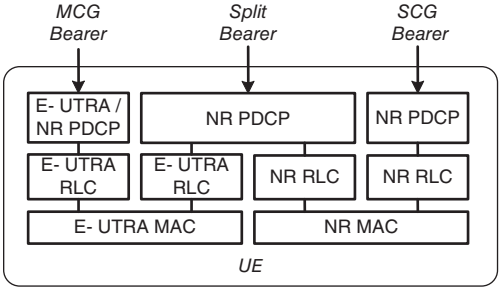
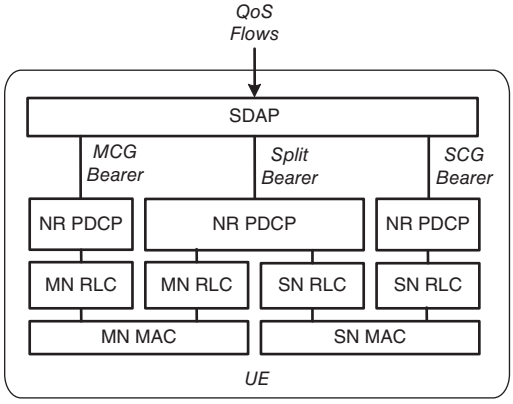



**Figure 4.3.11** Radio protocol architecture for MCG, SCG, and split bearers from a UE perspective in MR-DC with 5GC (NGEN-DC, NE-DC, and NR-DC). (*Source:* Reproduced by permission of © 3GPP).



#### 4.3.4.2.2 QoS Aspects and Bearer Type Selection

**EN-DC** The QoS aspects for EN-DC follow the QoS framework that applies in E-UTRAN and defined in 3GPP TS 36.300. The MN decides whether a given bearer should be terminated at the MN or at the SN. For an MN terminated bearer, the corresponding S1-U bearer is established between the EPC and the MN. For an SN terminated bearer, the corresponding S1-U bearer is established between the EPC and the SN.

Besides deciding which PDCP entity (in the MN or in the SN) terminates each radio bearer, the MN also decides in which cell group(s) radio resources are to be configured, that is whether a bearer is an MCG, an SCG or a split bearer. However, once an SN terminated split bearer is established, through a Secondary Node Addition procedure (as described in Section 4.3.4.3.1) or a Secondary Node Modification procedure (as described in Section 4.3.4.3.2), the SN may later on remove SCG resources, as long as QoS for the respective E-UTRAN Radio Access Bearer (E-RAB) is guaranteed.

In EN-DC, all the possible bearer type change options are supported: MCG bearer to/from split bearer, MCG bearer to/from SCG bearer, and SCG bearer to/from split bearer. Also bearer termination point change (MN terminated bearer to/from SN terminated bearer) is supported for all bearer types, and can be performed with or without a simultaneous bearer type change. For E-RABs for which a bearer termination point changes from/to MN terminated bearer to/from SN terminated bearer, user data forwarding may be performed. In this case, the behavior of data forwarding follows that of handover, where the node from which data are forwarded behaves as a "source eNB" and the node to which data are forwarded behaves as a "target eNB" during handover.

**MR-DC with 5GC** Similarly, the QoS aspects for MR-DC with 5GC follow the same principles that are applied in NG-RAN, which are described in Section 3.4 and defined in 3GPP TS 38.300. It is always up to the MN to decide whether a QoS flow for a given PDU session is handled by the MN or the SN. As a result, QoS flows belonging to the same PDU session may be handled by different nodes. This means that it may happen that, at the network side, a given PDU session may use two different SDAP entities: one at the MN and another at the SN, which is referred to as a split PDU session.

For split PDU sessions, the MN requests the 5GC to establish two NG-U tunnel terminations at the NG-RAN side: one is to exchange user-plane traffic for a subset of the QoS flows of the PDU session handled by the SDAP entity at the MN and the other is to exchange user-plane traffic for the rest of the QoS flows of the PDU session handled by the SDAP entity at the SN. For QoS flows assigned to the SDAP entity in the SN, if the SN realizes that it cannot host a given QoS flow any longer, the SN can request the MN to remove the QoS flow from its SDAP entity. Also, if the MN realizes that it can host a given QoS flow previously assigned to the SDAP entity in the SN, the MN may inform the SN and remove the QoS flow from the SN's SDAP entity.

Once a QoS flow is assigned to a SDAP entity, the node hosting SDAP entity can then decide how to map the QoS flow to an actual radio bearer, as well as its bearer type (MCG bearer, SCG bearer, or split bearer). In particular, if the SDAP entity for a given QoS flow is hosted by the MN and the MN decides that an SCG or a split bearer should be configured (in other words, SCG resources should be configured), the MN provides the relevant information to the SN: the DRB-level QoS parameters, the QoS flow to DRB mapping information,

and the respective per-QoS flow information, based on which the SN configures appropriate SCG resources. On the other hand, if the MN decides the SDAP entity for a given QoS flow is hosted by the SN, the MN first provides sufficient QoS-related information to enable the SN to configure appropriate SCG resources and to request the configuration of appropriate MCG resources. For instance, the MN may offer MCG resources to the SN (for Guaranteed Bit Rate QoS flows, may indicate the amount offered to the SN on a per-QoS flow level), then the SN decides whether to map the QoS flow to an SCG bearer, an MCG bearer, or a split bearer. If the SN decides that an MCG or a split bearer should be configured (i.e. MCG resource is required), the SN needs to provide the relevant information back to the MN: the DRB-level QoS parameters and the QoS flow to DRB mapping information.

Since both the MN and the SN are allowed to establish/modify/release DRBs on their own for QoS flows terminated at their SDAP entities, coordination of DRB IDs between the MN and the SN is needed in MR-DC with 5GC to ensure unique allocation of DRB IDs for a UE. This is unlike EN-DC where the SN is not allowed to establish/modify/release a DRB on its own (it has to request the MN to or be requested by the MN) – and thus only the MN assigns a DRB ID (5 bit space) for a UE regardless of whether it is served by the MN or the SN.[7] For unique allocation of DRB IDs in MR-DC with 5GC, the SN assigns DRB IDs for the bearers it terminates, based on the IDs available for use offered from the MN. If a termination point is changed for a DRB (e.g. from the MN to the SN or vice versa), the node initiating the procedure indicates the corresponding DRB ID and is informed (by the node who accepted DRB offloading) whether the indicated DRB ID is available.

For QoS flows assigned to the SN SDAP entity, the SN may remove or add SCG resources for the corresponding bearers, as long as the QoS for the respective QoS flow is guaranteed. For each PDU session, including split PDU sessions, at most one default DRB may be configured at the MN or the SN.

As in EN-DC, also in all the other MR-DC options all the possible bearer type change options are supported, as well as bearer termination point changes for all bearer types. Whenever a bearer termination point is changed, user data forwarding may be performed between NG-RAN nodes following the behavior of handover, that is the node from which data are forwarded behaves as a “source NG-RAN node” and the node to which data are forwarded behaves as a “target NG-RAN node” during handover.

#### 4.3.4.3 Procedures

Different procedures are used to set up a MR-DC configuration, to modify the characteristics of an MR-DC configuration between an MN and an SN, and to change the SN involved in the MR-DC configuration.

The procedures described in the present section are defined in generic terms and are applicable to all MR-DC architectures. However, in practice, different (albeit quite similar) protocols are used on the network interface between MN and SN, that is X2-AP for EN-DC and Xn-AP for MR-DC with 5GC. The former is defined in 3GPP TS 36.423 and the latter in 3GPP TS 38.423. Note that in the procedure diagrams below EPC entities (S-GW and MME) and therefore X2-AP protocol on the X2 interface are used. The procedures for MR-DC with 5GC are very similar.

7 This is also because the DRB ID management by a single node makes security updates due to wraparound less frequent (DRB ID is used as an input to the PDCP encryption algorithm of a DRB).

#### 4.3.4.3.1 Secondary Node Addition

The Secondary Node Addition procedure is initiated by the node serving the UE (that plays the role of the MN) and is used to add an SN providing additional resources to the UE. For bearers requiring SCG radio resources, this procedure is used to add at least the first cell of the SCG. But this procedure can also be used to configure an SN terminated MCG bearer (where no SCG configuration is needed). Figure 4.3.12 shows the message flow for the Secondary Node Addition procedure in the EN-DC case.

1. The procedure is triggered by the MN based on the measurement reports provided by the UE and also based on QoS and traffic load considerations. For instance, a UE configured to perform NR measurements may eventually report a B1 event (triggered when a neighboring inter-system cell becomes better than a threshold) to the serving eNB (MN).
2. If the MN decides to initiate the Secondary Node Addition procedure, it sends the X2-AP SgNB Addition Request message requesting the SN to allocate the needed resources, for the DRBs and possibly also for the split SRB operation, also providing information about the usable UE capabilities. The MN also provides the latest measurement results for the SN to choose and configure the SCG cell(s). For a specific E-RAB, the MN may request the direct establishment of an SCG or a split bearer, without first having to establish an MCG bearer. It is also possible that all E-RABs can be configured as SN terminated bearers. The SN performs admission control and allocates the necessary radio resources and, depending on the bearer option, the respective transport network resources. For bearers requiring SCG radio resources, the SN decides the PSCell and other SCG SCells and triggers Random Access so that synchronization of the SN radio resource configuration can be performed.
3. If the SN decides to admit the request, it responds with the X2-AP SgNB Addition Request Acknowledge message providing the resource configuration to the MN in an NR RRC configuration container.
4. The MN sends the RRCConnectionReconfiguration message to the UE with the NR RRC configuration received from the SN.

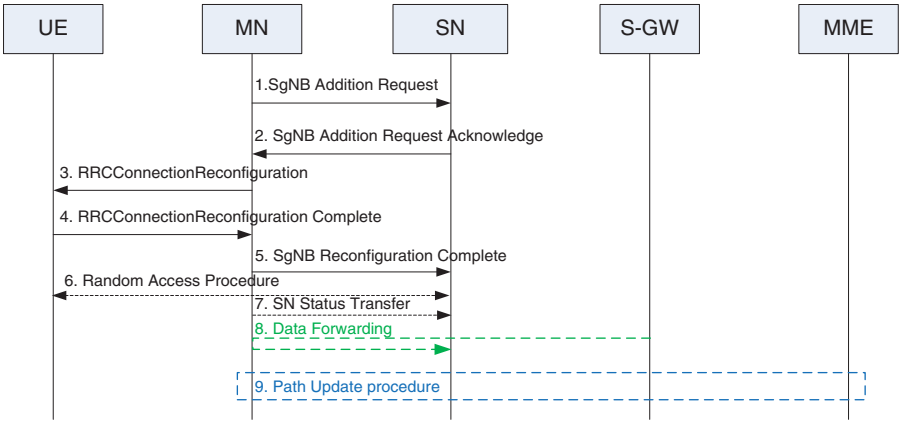



**Figure 4.3.12** Secondary Node Addition procedure. (*Source:* Reproduced by permission of © 3GPP).

5. The UE replies with an RRCConnectionReconfigurationComplete message, which may also contain a NR RRC response message.
6. The MN sends the X2-AP SgNB Reconfiguration Complete message to the SN, which may also contain the encoded NR RRC response message, if received in step 4.
7. The UE may perform the random access procedure with the SN to synchronize to the PSCell, if SCG radio resources are configured.
8. The MN may also send the SN Status Transfer message with the PDCP SN and HFN status, if the PDCP termination point is changed to the SN for bearers using RLC AM.
9. Depending on the bearer characteristics of the respective E-RAB, the MN may initiate data forwarding to minimize the service interruption due to the EN-DC activation.
10. For SN terminated bearers, the update of the user-plane path toward the EPC is performed using the S1-AP Path Switch procedure.

#### *4.3.4.3.2 Secondary Node Modification*

The Secondary Node Modification procedure may be initiated either by the MN or by the SN and be used for several purposes: to set up, release, or modify bearer contexts, to transfer bearer contexts to and from the SN, or to modify other properties of the UE context within the same SN. This procedure does not necessarily need to involve signaling toward the UE.

The MN may initiate the SN Modification procedure to change the SCG configuration within the same SN, for example, addition, modification, or release of SCG bearer(s) and the SCG RLC bearer of split bearer(s), as well as configuration changes for SN terminated MCG bearers. The MN may also use this procedure to perform handover within the same MN while keeping the SN.

The SN may initiate an SN Modification procedure involving the MN to perform SCG configuration changes within the same SN, for example, to trigger the release of SCG bearer(s) and the SCG RLC bearer of split bearer(s), and to trigger a PSCell change with a new security key is required. For instance, a UE configured with EN-DC may eventually report an A2 event (triggered when the serving cell becomes worse than a threshold) to the gNB (SN). In the case of an SN terminated bearer, the SN may then decide to trigger the SN-initiated SN Modification to change the bearer type by removing the SCG RLC bearer while keeping the bearer termination point in the SN. Figure 4.3.13 shows the message flow for an SN-initiated SN Modification procedure, with MN involvement, in the EN-DC case.

1. If the SN decides to initiate the SN Modification procedure, it sends the X2-AP SgNB Modification Required message, which may contain: PDCP change indication, E-RABs to be modified list, E-RABs to be released list, and the NR RRC configuration message with UE context-related information and the new SCG radio resource configuration.
2. The reception of the X2-AP SgNB Modification Required message may also trigger a nested MN-initiated SN Modification procedure, in which case the MN sends the X2-AP SgNB Modification Request message.
3. If the nested MN-initiated SN Modification procedure was triggered in step 2, the SN responds with the X2-AP SgNB Modification Request Acknowledge. The nested procedure is used, for example, to provide information such as data forwarding addresses, new SN security key, measurement gap configuration, etc.
4. The MN sends the RRCConnectionReconfiguration message to the UE with an NR RRC configuration message carrying the new SCG radio resource configuration.

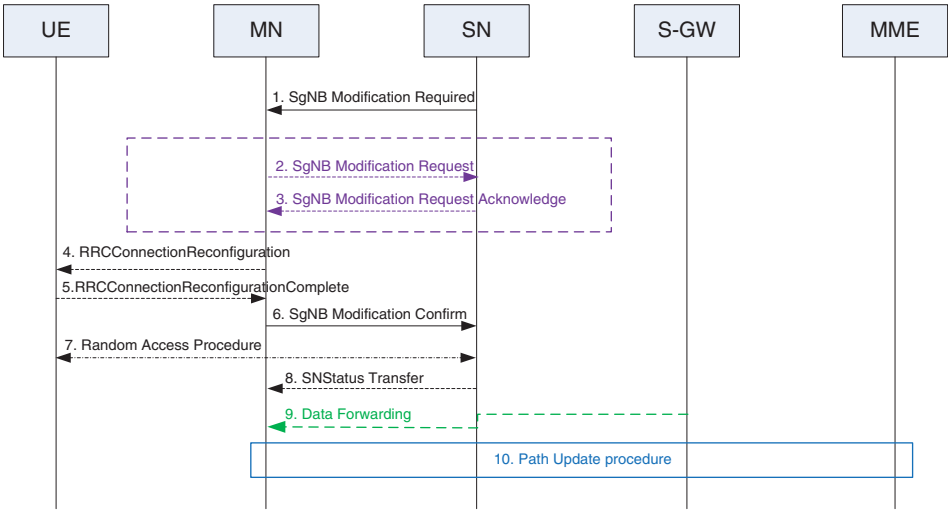



**Figure 4.3.13** SN Modification procedure – SN-initiated with MN involvement. (*Source:* Reproduced by permission of © 3GPP).

5. The UE replies with an RRCConnectionReconfigurationComplete message, which may also carry an NR RRC response message.
6. If the procedure is successful, the MN sends the X2-AP SgNB Modification Confirm message to the SN, which may carry the encoded NR RRC response message via the MeNB to SgNB Container IE, if the NR RRC message has been received from the UE (step 5).
7. The UE performs synchronization toward the PSCell if instructed by the network. Alternatively, the UE may immediately start uplink transmission after having applied the new configuration.
8. The SN may also send the X2-AP SN Status Transfer message with the PDCP SN and HFN status, if the PDCP termination point is changed for bearers using RLC AM.
9. Data forwarding between MN and the SN may be performed, if configured.
10. The update of the user -plane path toward the EPC is performed using the S1-AP Path Switch procedure.

In some cases the SN Modification procedure can be performed without involving the MN, if SRB3 is established. This may happen if no coordination with the MN is required, for example, in the case of addition, modification, or release of SCG SCells. The most typical case is a PSCell change where the security key does not need to be changed. For instance, a PSCell change may be triggered when a UE configured with EN-DC eventually reports an A3 event (triggered when a neighboring cell becomes better than the serving cell by an offset). An example of the message flow for an SN-initiated SN Modification procedure without MN involvement, in the EN-DC case, is shown in Figure 4.3.14.

1. If If the SN node decides to perform SN medication, for example, after receiving a measurement report from the UE), it sends the RRCConnectionReconfiguration message to the UE through SRB3.

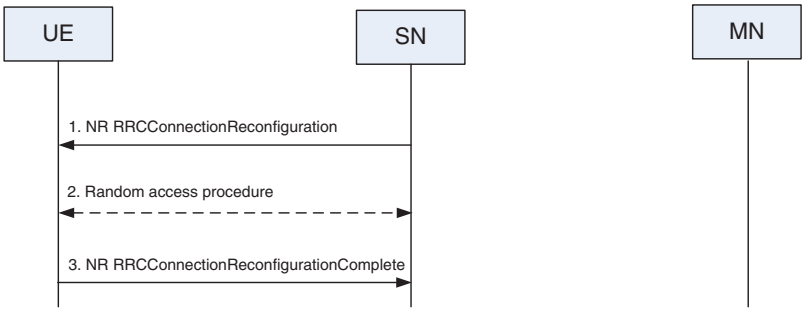



**Figure 4.3.14** SN Modification – SN-initiated without MN involvement. (*Source:* Reproduced by permission of © 3GPP).

2. If instructed by the network, the UE may (optionally) perform synchronization toward the (new) PSCell.
3. When the procedure is finished, the UE replies with the RRCConnectionReconfigurationComplete message.

#### *4.3.4.3.3 SN Change*

The SN Change procedure may be initiated either by the MN or the SN and used to transfer a UE context from a source SN to a target SN and at the same time change the SCG configuration in UE from one SN to another. This procedure always involves signaling over MCG SRB toward the UE.

For instance, an SN may decide to trigger an SN-initiated SN Change procedure when the UE reports an A3 event to the SN, if the neighboring cell becoming better than the serving cell is controlled by another SN. Figure 4.3.15 shows the message flow for an SN-initiated SN Change procedure, in the EN-DC case.

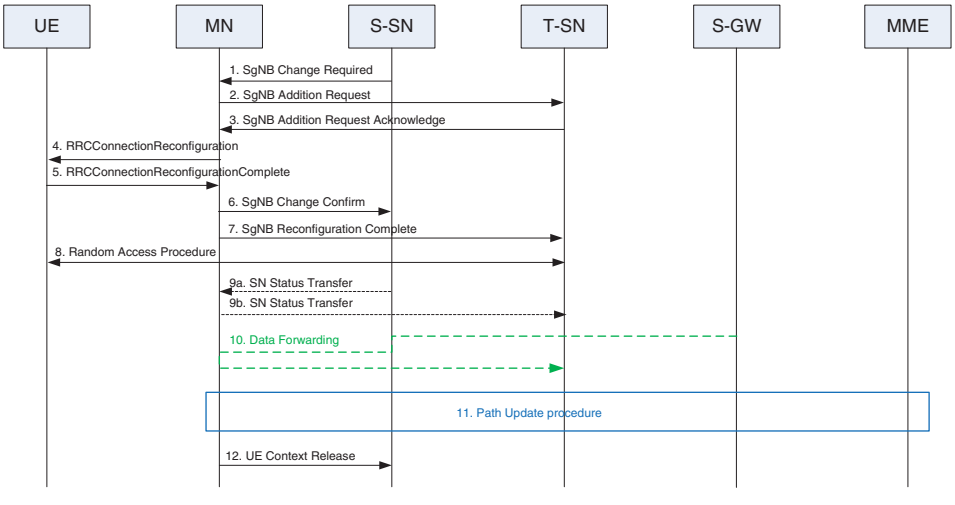



**Figure 4.3.15** SN Change – SN-initiated. (*Source:* Reproduced by permission of © 3GPP).

1. If the source SN decides to initiate the SN Change procedure, it sends the X2-AP SgNB Change Required message to the MN, carrying the target SN identifier information and optionally the SCG configuration via the SgNB to MeNB Container IE.
2. If the MN accepts the request, it sends the X2-AP SgNB Addition message to the target SN requesting it to allocate resources; the message may also carry the measurement results related to the target SN received from the source SN.
3. If the target SN accepts the request, it responds with the X2-AP SgNB Addition Request Acknowledge message.
4. The MN then provides the new configuration to the UE in the RRCConnectionReconfiguration message, including the NR RRC Reconfiguration message generated by the target SN.
5. The UE applies the new configuration and sends the RRCConnectionReconfigurationComplete message, including the encoded NR RRC response message for the target SN, if needed.
6. The MN then confirms the release of the source SN resources by sending the X2-AP SgNB Change Confirm message, which may carry the downlink (and uplink) Forwarding GTP Tunnel Endpoint IE if data forwarding is configured. When the source SN receives the SgNB Change Confirm message it stops providing user data to the UE and, if applicable, starts data forwarding.
7. If the RRC connection reconfiguration procedure was successful, the MN then sends the X2-AP SgNB Reconfiguration Complete message to the target SN, which may contain the MeNB to SgNB Container IE carrying the encoded NR RRC response message for the target SN, if received from the UE.
8. The UE performs the random access procedure to synchronize to the new PSCell in the target SN.
9. For SN terminated bearers using RLC AM, the source SN sends the X2-AP SN Status Transfer message with the PDCP SN and HFN status, which the MN sends then to the target SN in a separate X2-AP message.
10. Data forwarding from the source SN takes place either at this step or as early as step 6, if configured.
11. For source SN terminated bearers, the MN then triggers the update of the user-plane path toward the EPC using the S1-AP Path Switch procedure.
12. The MN also sends the X2-AP UE Context Release message to the source SN, upon which the SN releases all the resource associated to the UE context.

### 4.3.5 Further Reading

Readers interested in further details about various options of MR-DC, should start from the general stage 2 description in 3GPP TS 37.340. Note that while the overall stage 2 RAN specification is normally provided in the 300 series, that is, 3GPP TS 36.300 for LTE and 3GPP TS 38.300 for NR, MR-DC is described separately.

Once a reader has familiarized himself with the high-level aspects, readers interested in the RAN aspects of MR-DC operation should familiarize themselves with the X2-AP control-plane protocol (3GPP TS 36.423) and the similar but somewhat different Xn-AP control-plane protocol (3GPP TS 38.423).

Network interface user-plane details can be found in 3GPP TS 36.424 and TS 36.425 for EN-DC, and 3GPP TS 38.424 and TS 38.425 for MR-DC with 5GC.

While the present chapter and the specifications mentioned above focus on the network aspects, the understanding of MR-DC operations would not be complete without the radio interface details, which can be found in 3GPP TS 36.331 TS 38.331.

## References

3GPP Technical Specification 36.300 (2019). Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 36.331 (2019). Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 36.413 (2019). Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 36.423 (2019). Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 36.424 (2019). Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 data transport. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 36.425 (2019). Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 interface user plane protocol. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 37.340 (2019). Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.300 (2019). NR; Overall description; Stage-2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.331 (2019). NR; Radio Resource Control (RRC); Protocol specification. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.423 (2019). NG-RAN; Xn Application Protocol (XnAP). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.424 (2019). NG-RAN; Xn data transport. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.425 (2019). NG-RAN; NR user plane protocol. Available at: www.3gpp.org (accessed May 29, 2020).

## 4.4 Control–User Plane Separation

*Feng Yang*

*Intel Corporation, China*

In the Section 4.2, we described the high-level gNB functional split architecture (using the terminology introduced in Section 4.1), consisting of a gNB-CU and gNB-DU, wherein the gNB-CU consists of both the control-plane functions and user plane functions of the upper-layer protocols, including RRC and RRM (for the control plane) and PDCP and SDAP (for the user plane).

While the previously described gNB-CU/gNB-DU split architecture is a viable and beneficial deployment option, 3GPP have additionally specified a more fine-grained split architecture, in which the gNB-CU is further split into control-plane (gNB-CU-CP) and user-plane (gNB-CU-UP) logical network nodes. This architecture option is inspired by the popular software-defined networking (SDN) concept, which is in large part based on the idea of control- and user-plane separation. 3GPP applied the separation principle to EPC in 4G for the first time,[8] which was later extended to NG-RAN in 5G, specifically to the gNB-CU. Standardized control/user-plane separation can bring numerous benefits, including independent scaling of control and user planes, enabling multi-vendor interoperability, and centralized RRM.

The gNB-CU described in Section 4.2 can be deployed as separate control-plane and user-plane logical network nodes with a standardized E1 interface between them, which is the topic of the present chapter. With this enhancement, a gNB can be deployed as a single network node, or split into gNB-CU and gNB-DU nodes, or further split into gNB-CU-CP (centralized control-plane node), gNB-CU-UP (centralized user-plane node), and gNB-DU. These are deployment options an operator can choose from to suit their needs best, depending on, for example, network capacity requirements, transport network capabilities, etc.

### 4.4.1 Key Ideas

- 3GPP has defined a number of NG-RAN architecture options: a monolithic (i.e. single logical network node) gNB, split gNB-CU and gNB-DU, and an option allowing control- and user-plane separation. In the latter case, a gNB-CU is separated into two logical nodes: a control-plane node (gNB-CU-CP) and one or multiple user-plane nodes (gNB-CU-UP), connected via a standardized E1 interface.
- Key reasons for control/user plane separation are: independent scaling of control and user planes, centralized RRM, alignment with the SDN concept, and better support of network slicing.
- The gNB-CU-UP logical network node hosts SDAP and PDCP (for user plane) layers, and gNB-CU-CP hosts RRC and PDCP (for control plane) layers.
- 3GPP have standardized E1 interface between a gNB-CU-CP and gNB-CU-UP, which supports control-plane protocol only. The E1 interface and the E1-AP protocol

8 3GPP work item on control- and user-plane separation of EPC nodes (CUPS), (3GPP TS 23.214).

design generally follow the same design principles as other RAN control-plane interfaces.[9]

- The standardized E1 interface may allow multi-vendor deployments of gNB-CU-CP and gNB-CU-UP network nodes.
- The E1AP supports interface management procedures and bearer context management procedures.
- gNB-CU-CP and gNB-CU-UP are managed separately via OAM. Certain information has to be preconfigured in both nodes by OAM. For example, a gNB-CU-UP must be preconfigured with the transport network address of a gNB-CU-CP, etc.
- gNB-CU-CP and gNB-CU-UP control/user-plane separation with the E1 interface by itself is not equivalent to SDN, but rather follows a similar design concept. SDN is normally characterized by additional features, such as flow-based forwarding and routing, use of OpenFlow, and others.

### 4.4.2 Market Drivers

The idea of decoupling control- and user-plane protocols and deploying them in separate network nodes has been around for a while. It inspired the development of SDN, which become very successful in data centers and transport networks, such as those described in Jain et al. (2013). In 3GPP, the idea was first discussed in the context of 4G EPC, where S-GW and P-GW core network nodes were separated into S-GW-C/P-GW-C and S-GW-U/P-GW-U (3GPP TS 23.214) control- and user-plane nodes, respectively. In 4G, the concept remained confined to the core network.

In 5G, the core network (5GC) was designed according to the principle of control- and user-plane separation from the beginning. This is also the case for NG-RAN, which since Release-15 supports both deployment options of co-located and separated control- and user-plane NG-RAN network nodes. The fact that both options have been specified is a bit unusual, as normally there is no need to do so – it is sufficient to specify the separated option only and the implementation is free to "collapse" two network nodes in one, as this has no standards impact.[10] This reflects the fact that in the initial 5G study, control- and user-plane separation of RAN has attracted broad interest but also faced concerns from some companies (the concerns have been mostly not with the technology as such, but with standardizing an open multi-vendor interoperable network interface to support it, which has commercial implications).

During the study of control/user-plane separation in NG-RAN (3GPP TR 38.806), three deployment options have been identified, as illustrated in Figure 4.4.1. These have the following benefits for each deployment option:

1. Allow to take maximum advantage of cloud technologies.
2. Ensure low latency for critical control-plane procedures.

---

9 E1 is one of the few examples of control-plane only network interfaces; all other NG-RAN interfaces are with both control- and user-plane protocols.

10 In EPC, for example, there are implementation that contain both S-GW and P-GW in a single network node, in which case the interface between them is internal. Some implementations go as far as containing the whole EPC in a single network node. Even though such implementation options are not explicitly specified, they are allowed by the standard.

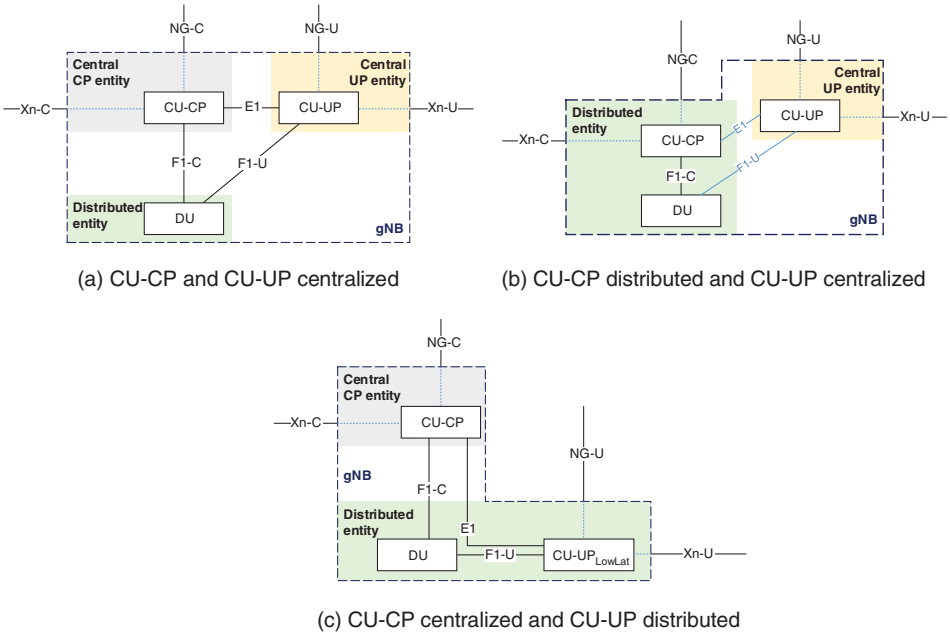



**Figure 4.4.1** Deployment scenarios for CU/UP separation. (*Source:* Reproduced by permission of © 3GPP).

3. Ensure low latency for user-plane traffic, which is important for some applications, e.g. critical MTC.

These deployment options trade off applicability to cloud deployments, control-plane latency, and user-plane latency, and thus give operators sufficient flexibility to meet diverse requirements of different applications of 5G.

The study report (3GPP TR 38.806) also summarizes the following advantages that control–user plane separation could bring:

- Flexibility to operate and manage complex networks, efficiently supporting different transport network topologies.
- Ability to tailor NG-RAN deployment to address various service requirements.
- Alignment with the SDN concept of a functional decomposition of the RAN into user- and control-plane entities.
- Independent scaling of control and user plane, allowing, e.g. addition of new UP hardware resources when required.
- Support of multi-vendor interoperability between NG-RAN control- and user-plane network nodes provided by different vendors.
- Capability to deploy separate control- and user-plane network nodes while optimizing for desired scenarios and performance in terms of throughput and latency. For example, a gNB-CU-CP can be deployed in close proximity to a gNB-DU to optimize for critical control-plane latency. Additionally, a gNB-CU-UP can be deployed in a centralized manner, e.g. in a regional or national data center, allowing resource sharing and supporting

cloud implementation. Furthermore, a gNB-CU-UP can be deployed close to a gNB-DU to provide a local termination point for ultra-reliable low-latency communication (URLLC) traffic, thus significantly improving latency.

- Support of centralized RRM within a single gNB-CU-CP controlling a large number of gNB-CU-UPs covering a large geographical area for improved radio performance and resource utilization.
- Better support for network slicing, as centralized RRM makes it easier to provide slice-level isolation as well as improved resource utilization.

Moreover, with the emergence of big data analytics and artificial intelligence, a centralized control plane opens the door to applying these cutting-edge technologies for RAN optimization. One example is the work conducted in the O-RAN Alliance to leverage emerging deep learning techniques to empower an intelligent RAN controller (O-RAN).

Despite numerous benefits of control- and user-plane separation, there are also challenges. For example, network maintenance complexity and cost may increase due to the introduction of new logical nodes. While some deployment options improve latency, generally, the implementation of control- and user-plane functionalities in separate network nodes with a network interface between them carrying signaling messages (as opposed to an internal interface in a co-located scenario) introduces additional delay. These issues should be carefully addressed during deployment planning.

### 4.4.3 Functional Description

In the NG-RAN architecture with control- and user-plane separation, the gNB-CU is further split into control plane (gNB-CU-CP) and user plane (gNB-CU-UP) logical network nodes. It generally follows the same design principle as the gNB split into gNB-CU and gNB-DU, that is – deployments with and without control- and user-plane NG-RAN separation shall be indistinguishable from the point of view of a UE, a 5GC, or another gNB. Furthermore, a monolithic gNB-CU and split gNB-CU-CP/gNB-CU-UP shall be indistinguishable from the perspective of a gNB-DU.

A gNB-CU-CP hosts the RRC and PDCP (for control-plane) protocols and a gNB-CU-UP hosts SDAP and PDCP (for user-plane) protocols. A gNB-CU-CP is connected to potentially multiple gNB-CU-UPs via the standardized control-plane E1 interface, as illustrated by Figure 4.4.2.

Figure 4.4.3 shows the protocol structure for E1. As with any other NG-RAN control-plane interface, it uses IP and SCTP. However, unlike most other NG-RAN

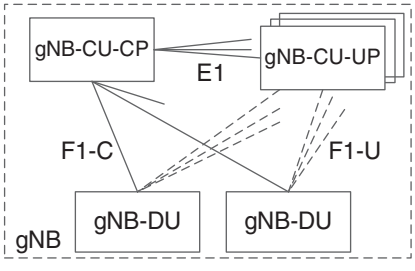



**Figure 4.4.2** : Overall architecture for separation of gNB-CU-CP and gNB-CU-UP. (*Source:* Reproduced by permission of © 3GPP).

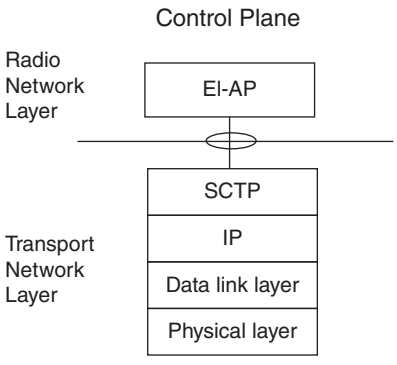



**Figure 4.4.3** Interface protocol structure for E1. (*Source:* Reproduced by permission of © 3GPP).

network interfaces, there is no user plane for E1 as it is purely a control-plane interface. All user-plane interfaces (e.g. NG-U and Xn-U) are terminated in the gNB-CU-UP – E1 is used to control these, but data-plane packets do not go via gNB-CU-CP.

gNB-DU is connected to only one gNB-CU-CP, since in the monolithic architecture a gNB-DU is connected to only one gNB-CU. Similarly, a gNB-CU-UP is connected to only one gNB-CU-CP. Such architecture with only one control-plane network node simplifies both design and implementation. The number of connections between a gNB-DU and a gNB-CU-UP can be arbitrary, that is, one gNB-DU can be connected to multiple gNB-CU-UPs and vice versa, as long as they all are under control of the same gNB-CU-CP. In Figure 4.4.2, (3GPP TS 38.401) solid lines represent interface instances whereas dashed lines represent F1-U tunnels, for example, carrying data for multiple UEs and/or multiple bearers. A single E1 (and F1-C) interface instance between a pair of gNB-CU-CP and gNB-CU-UP (and a pair of gNB-CU-CP and gNB-DU) can support multiple SCTP associations, for example, for NG-RAN deployment in virtualized environments (where, e.g., a single instance of a gNB-CU-UP or a gNB-CU-CP network node can be deployed in a virtualized platform with multiple hardware instances having multiple transport network addresses).

The E1 standardized interface between gNB-CU-CP and gNB-CU-UP uses the E1AP, which is a control-plane protocol. It generally follows the same design principles as other RAN control-plane application protocols. Figure 4.4.1 zooms in on NG-RAN and does not show the 5GC and other NG-RAN network nodes; however, from the overall 5GS perspective the control-plane interfaces NG-C and Xn-C are terminated in the gNB-CU-CP, while user-plane interfaces NG-U[11] and Xn-U are terminated in the gNB-CU-UP (or directly in the gNB-DU). As mentioned above, this is defined in such a manner that the 5GC or the other NG-RAN network nodes are not exposed to the details of the NG-RAN architecture.

More details about the overall NG-RAN architecture can be found in the 3GPP specification (3GPP TS 38.401).

#### 4.4.3.1 Control Plane

Unlike most other NG-RAN interfaces (e.g. F1 and Xn), E1AP is a pure control-plane protocol, as there is no need to carry user data between the control- and user-plane network

11 Referred to as N2 in 3GPP specifications describing 5GC.

nodes. Similarly to F1-C and other NG-RAN control-plane interfaces, E1 relies on SCTP, which provides reliable transport for in-sequence delivery of E1AP messages. E1AP functionalities can be categorized as follows:

- Interface management procedures
- Bearer context management procedures.

Note that, unlike many other NG-RAN interfaces, formally there is no "UE management" procedure – the high-level gNB-CU-CP design principle is that it operates on bearers, rather than UEs. Therefore, unlike, for example, an Xn interface, there is no notion of "UE context establishment" on E1. However, a UE context is implicitly established using the Bearer Context Setup procedure and is identified by a pair of gNB-CU-CP and gNB-CU-UP E1AP identifiers. Moreover, some bearer management procedures (e.g. the Bearer Context Setup Request) carry some UE-specific information (e.g. UE inactivity timer).

As for all other RAN interfaces, E1AP procedures can be class 1 (request and response) and class 2 (single message).

In E1AP design, provisions have been made to facilitate the deployment of a gNB-CU-UP node in a virtualized environment. For example, it is the gNB-CU-UP rather than the gNB-CU-CP that allocates TNL addresses for a UE during initial access or handover procedures, which allows an arbitrary part of the gNB-CU-UP to serve the particular gNB-DU that the UE was attached to – a feature essential for resource virtualization and pooling. More details and examples are provided in the subsequent section.

Full details about the E1AP control-plane protocol can be found in 3GPP TS 38.463.

#### *4.4.3.1.1 Interface Management Procedures*

Interface management procedures are used to establish, to release, and to reset (when and if needed) the E1 interface, and also to allow a gNB-CU-CP and a gNB-CU-UP to update configuration information. These procedures are:

- gNB-CU-CP E1 Setup and gNB-CU-UP E1 Setup
- gNB-CU-CP Configuration Update and gNB-CU-UP Configuration Update
- Reset, Error Indication, and Release
- gNB-CU-UP Status Indication.

One noticeable difference compared with the F1 Setup discussed previously (which can be only initiated by the gNB-DU), is that both the gNB-CU-CP and the gNB-CU-UP are allowed to initiate the E1 Setup procedure once a TNL association between the gNB-CU-CP and gNB-CU-UP becomes operational. This generally follows the design principle of other NG-RAN horizontal interfaces (e.g. Xn) and is meant to facilitate the case in which more control- and user-plane capacity is added by an operator when the network evolves. Consequently, a race condition can occur if both a gNB-CU-CP and a gNB-CU-UP simultaneously perform the E1 Setup procedure. To resolve the issue, the specification clarifies that the network node that initiates the TNL association shall also initiate the E1 Setup procedure. If the setup procedure fails, the other network node may respond with the E1 Setup Failure message, carrying the "time to wait" indication during which it is not allowed to attempt reinitiation of the E1 Setup procedure. This is normally used during load spikes or temporary issues, which are expected to pass within a reasonable time and prevents the initiating node from generating additional load during that time.

Since NG-RAN supports connectivity to 5GC, legacy EPC, and both (e.g. in EN-DC) as described in more detail in Chapter 3, the node initiating the E1 Setup procedure signals the type of core network connectivity it supports. Additionally, certain core network network-related information is signaled by the node initiating the E1 Setup procedure, such as list of supported PLMNs and a list of supported slices and cells within that PLMN. This helps resolving potential misconfiguration issues, especially in the early days of 5G deployment and during migration, for example, from NG-RAN deployed with EPC to NG-RAN deployed with 5GC.

During the E1 interface operation, both the gNB-CU-CP and the gNB-CU-UP can inform each other about their configuration changes, using the gNB-CU-CP Configuration Update and the gNB-CU-UP Configuration Update procedures, respectively. For example, a gNB-CU-UP may initiate the gNB-CU-UP Configuration Update procedure if the list of supported NR CGIs has changed. The lists of supported PLMNs, cells, and slices in gNB-CU-CP and gNB-CU-UP connected via E1 does not have to match. If it does not, the setup procedure can still succeed, and the gNB-CU-CP will take this information into account for decisions on bearer establishment; for example, by allowing bearer establishment only for UEs requesting access to the PLMN supported by both nodes. These details are, however, not specified in the standard and are left for implementation. Therefore, different implementations may handle these cases differently.

#### *4.4.3.1.2 Bearer Context Management Procedures*

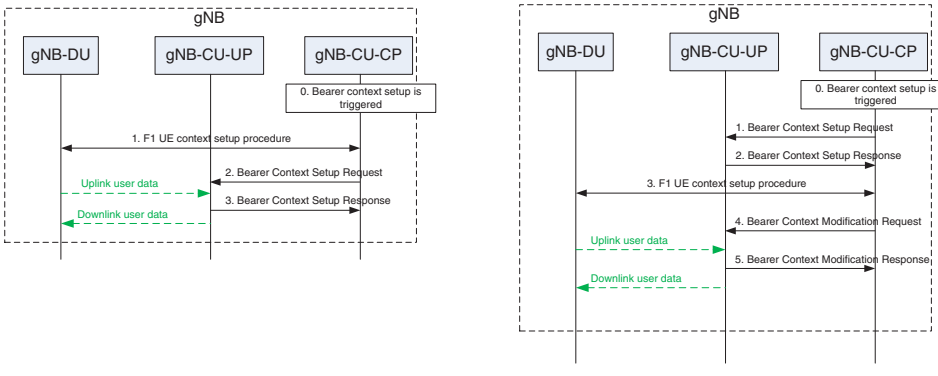
Bearer context management procedures are arguably the most important aspect of the E1 interface. Using these procedures, a gNB-CU-CP can establish, modify, and release a bearer context in the gNB-CU-UP.

There procedures are:

- Bearer Context Setup
- Bearer Context Release
- Bearer Context Modification (gNB-CU-CP initiated) and Bearer Context Modification Required (gNB-CU-UP initiated).

**Bearer Context Setup Procedure** The purpose of the Bearer Context Setup procedure is to allow the gNB-CU-CP to establish a bearer context in the gNB-CU-UP. The bearer can either connect the gNB-DU and the gNB-CU-UP, or the UPF/S-GW and the gNB-CU-UP. Messages include Bearer Context Setup Request, Bearer Context Setup Response, and Bearer Context Setup Failure. Among the numerous IEs defined for these messages, uplink or downlink TNL address information of F1-U, S1-U/NG-U play a vital role, as they represent either end of a bearer.

From that perspective, there are two somewhat different procedures defined that are used to admit a new UE in a gNB-CU-UP, for example, during an initial access or a handover. The key difference between these two procedures is which node (gNB-CU-CP or gNB-CU-UP) allocates the TNL address (which can be a F1-U UL TNL address or NG-U/S1-U DL TNL address for a particular bearer) at the gNB-CU-UP side for a UE. These procedures are illustrated in Figure 4.4.4. In the former case, shown in Figure 4.4.4a, in a "one-shot" procedure the gNB-CU-CP selects TNL addresses.




(a) "One-shot" procedure (b) Two-step procedure

**Figure 4.4.4** Two options to admit a new UE in a gNB-CU-UP.

For the sake of brevity we only explain the two-step procedure shown in Figure 4.4.4b, which is performed as follows:

0. Bearer context setup (e.g. following the Initial UE Context Setup Request from the core network) is triggered.
1. The gNB-CU-CP sends a Bearer Context Setup Request message containing uplink TNL address information for S1-U or NG-U (for comparison, in the one-step procedure the Bearer Context Setup Request message contains downlink TNL address information for F1-U, as well as uplink TNL address information for S1-U or NG-U).
2. The gNB-CU-UP responds with a Bearer Context Setup Response message, containing uplink TNL address information for the F1-U and downlink TNL address information for the S1-U or NG-U it has assigned. The message indicates that one or more bearers are established in the gNB-CU-UP.
3. The F1 UE Context Setup procedure is performed to set up one or more bearers in the gNB-DU.
4. The gNB-CU-CP sends a Bearer Context Modification Request message, containing the downlink TNL address information for F1-U.
5. The gNB-CU-UP responds with a Bearer Context Modification Response message, indicating that one or more bearers, which are terminated at the gNB-DU, are established in the gNB-CU-UP.

**Bearer Context Modification Required Procedure** The purpose of the Bearer Context Modification Required procedure is to allow the gNB-CU-UP to inform the gNB-CU-CP that the Bearer Context Modification procedure has been completed or that a modification is required. Relevant E1AP messages include Bearer Context Modification Required and Bearer Context Modification Confirm.

Similar to Bearer Context Setup procedure, TNL address information is still one of the most important IEs defined for these messages.

The procedure is arguably another design relevant to virtualization, as it can be executed when the gNB-CU-UP needs to modify a bearer context and to inform the gNB-CU-CP, due to, for example local problems or virtual machine (VM) migration. It is well established that

cellular networks have so-called tidal effects where the traffic periodically and significantly changes over time. When the network is underutilized, performing VM migration to consolidate computing resources and shut down unused VMs could save energy. The details of live VM migration including transferring of context, installation, and teardown of VM, etc. are out of the scope of the book, however general aspects related to NG-RAN virtualizing are explained in Section 6.2.

The E1 (and F1) procedures required to support VM migration are illustrated in Figure 4.4.5.

1. The procedure is triggered by a gNB-CU-UP sending the E1-AP Bearer Context Modification Required message when VM migration requires changes in the transport network addresses used. The message may carry new uplink TNL address information of F1-U, and downlink TNL address information of S1-U or NG-U.
2. The TNL information received by the gNB-CU-CP in step 1 needs to be forwarded to the the gNB-DU, which is done using the F1AP UE Context Modification Request message.
3. If the gNB-DU accepts the request, it replies with the UE Context Modification Response message.
4. If this is successful, the procedure is completed with the E1AP Bearer Context Modification Confirm message sent by the gNB-CU-CP to the gNB-CU-UP.

**Notes on Select IEs** In this section we described in detail certain important Information Elements present in E1AP messages mentioned above.

- CHOICE system {E-UTRAN, NG-RAN}.

In almost every single bearer context management message defined in (3GPP TS 38.463), there is a "system" IE of type choice, which can be set either to E-UTRAN or NG-RAN. The CHOICE type indicates that a gNB can either interface an EPC or a 5GC. The latter case is the so-called "Standalone" NR deployment, with a gNB connected directly to the 5GC. In the former case, gNB may indirectly connect to EPC by EN-DC (refer to Section 4.3)

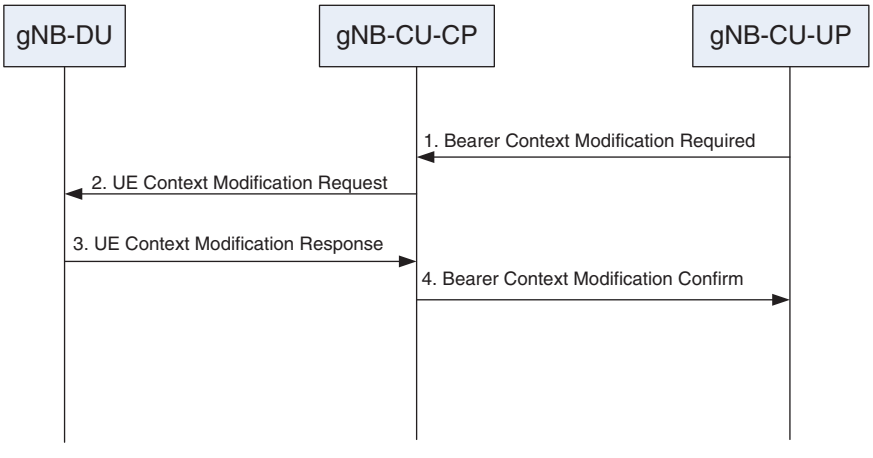



**Figure 4.4.5** NG-RAN procedure to support VM migration.

operation. The EN-DC is one of the MR-DC options, in which a UE may be connected to one eNB that acts as an MN and one en-gNB that as acts as an SN, both connecting to the EPC via a S1-U interface.

- IE E-UTRAN.

In the Evolved Packet System (EPS), the elementary connectivity between a UE and an S-GW is an E-RAB, which uniquely identifies traffic flows that receives a common data forwarding treatment. An E-RAB is the concatenation of a radio bearer represented by the DRB ID and a S1 bearer represented by the S1-Tunnel Endpoint Identifier (TEID), the one-to-one mapping of which is managed by the eNodeB or en-gNB. Therefore, under the IE E-UTRAN, both DRB ID and S1-U UL/DL Transport Layer Information consisting of GTP-TEID are present, indicating the binding of the two identifiers. For the internal interface F1-U, the gNB-CU-UP provides uplink UP Parameters in the message Bearer Context Setup Response and the gNB-CU-CP provides downlink UP Parameters in the message Bearer Context Modification Request, both of which contain UP Transport Layer Information. A bearer can be an MCG bearer (with an RLC bearer only in the MCG), an SCG bearer (with an RLC bearer only in the SCG) or a split bearer (with RLC bearers both in the MCG and the SCG) and that's why under UP Parameters there is a Cell Group ID associated with each UP Transport Layer Information, specifying which cell group (MCG or SCG) the bearer belongs to. A split bearer has two UP Parameters items, one for MCG and the other for SCG.[12]

- IE NG-RAN.

Compared with EPS, 5GS employs a different means to enforce QoS. First, the UPF classifies the user-plane traffic and performs user-plane marking using a QoS Flow Indicator (QFI), which takes over the role of S1-TEID in EPS, leaving GTP-TEID less important in 5GS. Second, Unlike in EPS where the eNB performs one-to-one mapping of S1 bearers to DRBs, the gNB decides how to bind QoS flows to DRBs and what's more important there is no strict 1:1 relation between them. It is up to the gNB to establish the necessary DRB that QoS flows can be mapped to, and to release them when not needed (3GPP TS 23.501). That's why under each PDU Session Resource To Setup Item[13] (uniquely represented by PDU Session ID), there are one or more DRB To Setup Items, each of which contains a list of QFIs to indicate the mapping between QoS flows and DRBs. The flow mapping information shall be included only in the messages sent by the gNB-CU-CP, because the gNB-CU-CP is in charge of mapping.

#### *4.4.3.1.3 Full UE Initial Access Procedure (with Focus on E1)*

A comprehensive call flow that incorporates E1 Bearer Context Management and F1 UE Context Management during UE initial access is illustrated in Figure 4.4.6. In this call flow we focus on E1 procedures, which are explained in detail, while other messages exchanged (e.g. over the F1 interface) are provided for context.

1. When the UE needs to establish a connection with the network (e.g. to send uplink data), it sends an RRC Connection Request message to the gNB-DU.

12 In other words, a split bearer is terminated in two gNB-DUs.

13 The elementary connectivity between a UE and a UPF is a PDU session, which associates the UE and the data network to provide a PDU connectivity service.

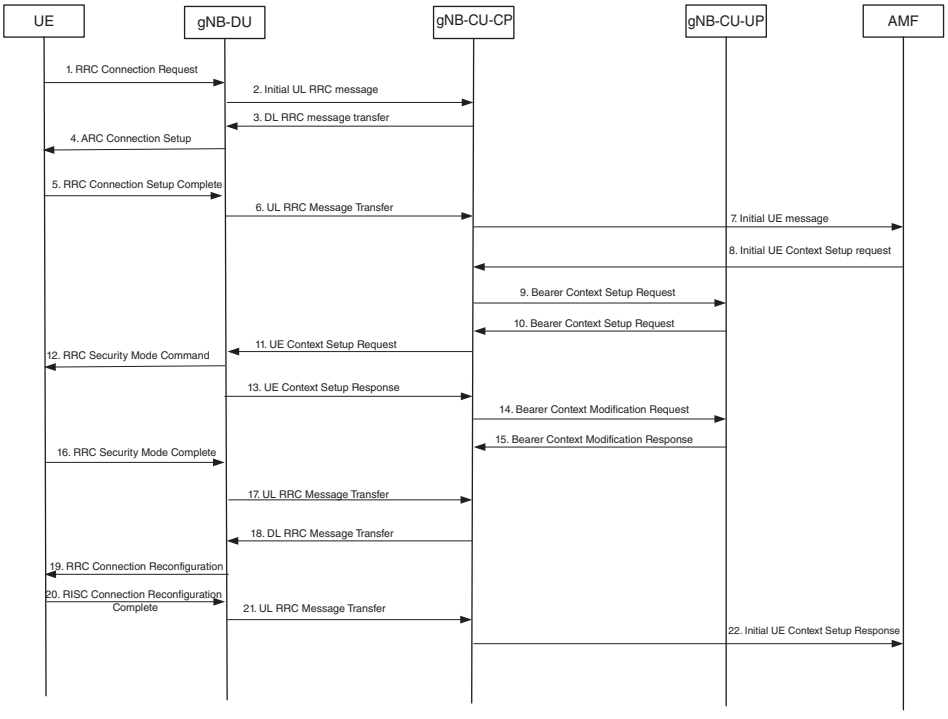



**Figure 4.4.6** UE initial access procedure involving E1 and F1. (*Source:* Reproduced by permission of © 3GPP).

2. If the gNB-DU decides to admit the UE, it forwards the RRC message received from the UE to the gNB-CU in the F1AP Initial UL RRC Message Transfer message, together with the corresponding low-level configuration (which the gNB-DU is in charge of). The Initial UL RRC Message Transfer message also includes the Cell Radio Network Temporary Identifier (C-RNTI) allocated by the gNB-DU.
3. If the gNB-CU-CP accepts the request, it allocates a gNB-CU UE F1AP ID for the UE and generates the RRC Connection Setup message toward the UE. The RRC message is encapsulated in the F1AP DL RRC Message Transfer message and sent to the gNB-DU.
4. The gNB-DU sends the RRC Connection Setup message (received from the gNB-CU) to the UE.
5. The UE sends the RRC Connection Setup Complete message to the gNB-DU.
6. Similar to step 2, the gNB-DU encapsulates the RRC message in the F1AP UL RRC Message Transfer message and sends it to the gNB-CU-CP.
7. The gNB-CU-CP sends the Initial UE Message to the AMF, thus triggering the connection establishment with the core network.
8. If the AMF accepts the request, it sends the Initial UE Context Setup Request message to the gNB-CU-CP.
9. The gNB-CU-CP sends the E1AP Bearer Context Setup Request message to establish the bearer context in the gNB-CU-UP. In the case of NG-RAN, the message carries the list of

PDU sessions to set up, which in turns contains a list of DRBs to set up up for each PDU session. The PDU sessions and DRB IEs contain all the information required by the gNB-CU-UP, such as QoS parameters, inactivity timers, slicing information, transport layer information, etc.

10. If the procedure is successful, the gNB-CU-UP sends the E1AP Bearer Context Setup Response message to the gNB-CU-CP. In the case of NG-RAN, the message contains the list of successfully established PDU sessions and the list of PDU sessions that could not be established. For each successfully established PDU session a list of successfully established DRBs and QoS flows is included. Furthermore, the response message also carries F1-U UL TEID and a transport layer address allocated by the gNB-CU-UP.
11. The gNB-CU-CP sends the UE Context Setup Request message to establish the UE context in the gNB-DU. In this message, it may also encapsulate the RRC Security Mode Command message.
12. The gNB-DU sends the RRC Security Mode Command message to the UE.
13. The gNB-DU sends the UE Context Setup Response message to the gNB-CU-CP.
14. The gNB-CU-CP sends the E1AP Bearer Context Modification Request message to the gNB-CU-UP, with a list of PDU sessions to modify, including F1-U DL TEID and the transport layer address received from the gNB-DU. This step is necessary since it is the gNB-DU that allocates the F1-U downlink transport layer addresses.
15. Similar to step 10, if the procedure is successful, the gNB-CU-UP sends the E1AP Bearer Context Modification Response message to the gNB-CU-CP. The message carries the list of successfully modified PDU sessions, together with the list of DRBs for each PDU session and their parameters.
16. The UE responds with the RRC Security Mode Complete message.
17. The gNB-DU encapsulates the RRC message in the F1AP UL RRC Message Transfer message and sends it to the gNB-CU-CP.
18. The gNB-CU-CP generates the RRC Connection Reconfiguration message and encapsulates it in the F1AP DL RRC Message Transfer message.
19. The gNB-DU sends the RRC Connection Reconfiguration message to the UE.
20. The UE sends the RRC Connection Reconfiguration Complete message to the gNB-DU.
21. The gNB-DU encapsulates the RRC message in the F1AP UL RRC Message Transfer message and sends it to the gNB-CU-CP.
22. The gNB-CU-CP sends the Initial UE Context Setup Response message to the AMF.

#### 4.4.3.2 OAM Aspects

It is generally assumed that a gNB-CU-CP and a gNB-CU-UP have independent management interfaces and are managed by the OAM somewhat separately. Therefore, in the beginning, the peer nodes are not aware of each other's configuration and certain E1AP procedures have been defined to allow both nodes to exchange relevant configuration information. For example, once the gNB-CU-UP is reconfigured, parameters such as the NR CGI Support List, Slice Support List, or QoS Parameters Support List will be updated and communicated to the gNB-CU-CP via the gNB-CU-UP Configuration Update procedure.

For additional information about OAM support for control- and user-plane separation, see Section 6.5. Full details of OAM are specified in 3GPP TS 28.541.

#### 4.4.3.3 Relation to SDN

SDN technology is an approach to networking architecture in which control logic is implemented in a network entity separate from, for example, routers and switches. This represents a paradigm shift compared with previous networking architectures and has been successfully applied to many networking technologies, for example, transport network and cloud computing, to name a few. The SDN concept makes it easier to meet new requirements from, for example, data centers that cannot be easily satisfied by traditional IP networks. Examples of such new requirements are: enforcement of dynamic routing policies, simplifying network configuration, accelerating the deployment of new networking features, and support of network customization.

A typical SDN network generally consists of three layers:

- Data plane, which perform packets forwarding.
- Control plane (or SDN controller), which installs forwarding rules in data forwarding devices via the south-bound interface, e.g. OpenFlow.
- Management plane (or network applications), which leverage the functions offered by the north-bound interface to implement network control and operation logic.

Even though 3GPP has not explicitly defined SDN, NG-RAN with separated control/user-plane architecture makes it easy to map 3GPP network entities to the SDN architecture:

- In the data plane, a gNB-CU-UP becomes a simple device to process and forward GTP or IP packets.
- In the control plane, a gNB-CU-CP role is similar to that of an SDN controller with the E1 interface acting as the south-bound interface, which installs processing rules for inbound packets of the data plane.
- The management plane is not explicitly standardized in 3GPP; however, RRM ranging from radio access control to mobility management can be viewed as embedded network applications. Applications offered to verticals, e.g. to customize the data plane via the north-bound interface are not there yet but worth further investigation.

Having said that, despite the fact that E1 makes it possible to apply the SDN concept to NG-RAN, there are still certain obstacles that need to be overcome by implementations. The TCP/IP protocol suite is relatively simple compared with 3GPP protocols, and therefore the design of a south-bound interface such as OpenFlow is simpler compared with the E1 interface specified by 3GPP. For example, the E1-AP Bearer Context Modification Request message can consist of a large number of IEs, while an OpenFlow message typically has fewer fields. OpenFlow essentially creates a two-phase pipeline for each input packet, that is, match and action. Match specifies which fields of the header, for example, Ethernet, VLAN, or IP shall be examined, and action specifies what needs to be done, for example, output, push/pop a VLAN tag, or increase/decrease TTL once a rule or multiple rules have been matched. As of OpenFlow 1.5, roughly 40 fields and 30 actions are defined. In conclusion, E1 does not directly map to OpenFlow, and therefore 3GPP control–user plane separation is not an implementation of SDN as such, but more of a concept alignment, which is illustrated by Figure 4.4.7.

#### 4.4.3.4 Relation to 5GC

As mentioned above, the principle of control- and user-plane separation in 3GPP originated in the core network domain. In the initial 4G EPC design, some control functions, including

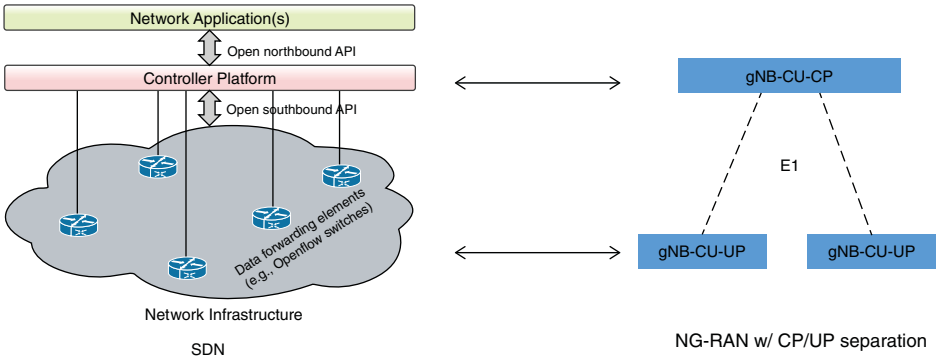



**Figure 4.4.7** Mapping of gNB-CU-CP and gNB-CU-UP to the elements of SDN.

those related to connection management and bearer management, have been separated to form a control node, that is, MME. However, some other EPC user-plane network nodes, specifically Serving Gateway (S-GW) and PDN Gateway (P-GW), as originally defined in 3GPP, also contain control-plane functions, for example, UE IP address management. In Release-14, the principle was applied to S-GW and P-GW, creating two pairs of logical nodes, that is, S-GW-C and S-GW-U, and P-GW-C and P-GW-U. As a result, most control-plane functions including UE IP address management are hosted in S-GW-C or P-GW-C, leaving S-GW-U or P-GW-U as a user-plane node to simply forward and route user's packets.

5G development takes this principle one step further. To allow independent scalability, evolution, and flexible deployments, 3GPP adopts control- and user-plane separation as one of the key principles to design the 5G core network (3GPP TS 23.501). A single logical node UPF, rather than two as in 4G is defined to perform user plane functions, such as packet routing and forwarding, packet inspection, anchor point for intra-/inter-RAT mobility, etc. Unlike the P-GW, which should be deployed centrally to support functions like UE IP address management and lawful intercept, the UPF is either responsible for part of some functions, for example, user-plane collection of lawful intercept, or does not have to support functions like UE IP address management. Therefore, UPF is appropriate for distributed deployment and offers native support of edge computing.

The role of UPF in 5G core network is similar to that of gNB-CU-UP in 5G RAN as:

- UPF routes user's packets between data network and gNB under the control of AMF, while gNB-CU-UP processes and forwards user's packets between UPF and gNB-DU under the control of gNB-CU-CP.
- Neither processes control-plane messages, but only routes them to the control-plane functions. For example, UPF routes the Dynamic Host Configuration Protocol (DNCP) request to SMF, which allocates an IP address for the UE while the user plane of gNB (if gNB-DU is regarded as part of the user-plane function) routes RRC to gNB-CU-CP.

The fact that both 5GC and NG-RAN support control- and user-plane separation makes it easy to deploy (and scale) both in a coherent manner.

### 4.4.4 Further Reading

A comprehensive survey on SDN can be found in 'Software-defined networking: a comprehensive survey'.

Full information about the CP/UP split NG-RAN architecture and related protocols described in this chapter can be found in the 3GPP technical specifications listed below.

3GPP TS 38.401 is the general stage 2 specification covering all NG-RAN aspects, including the CP/UP separation, and it is a good starting point for an interested reader to understand the details beyond what is described in this chapter. Once the reader has familiarized himself with the high-level aspects, we suggest learning the details of the E1AP, defined in 3GPP TS 38.463. Finally, 3GPP TS 28.541 can be used to gain an understanding of OAM aspects of the CP/UP separation, noting that the specification covers other nodes and architectures as well.

3GPP TR 38.806 is a good source of background information about the study on control- and user-plane separation, their benefits, and deployment options.

## References

3GPP Technical Specification 23.214 (2019). Architecture enhancements for control and user plane separation of EPC nodes. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Report 38.806 (2019). Study of separation of NR Control Plane (CP) and User Plane (UP) for split option 2. Available at: www.3gpp.org (accessed May 29, 2020).

O-RAN Alliance (2018). O-RAN: Towards an open and smart RAN. Available at: www.o-ran.org (accessed May 29, 2020).

3GPP Technical Specification 38.401 (2019). NG-RAN; Architecture description. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.460 (2019). NG-RAN; E1 general aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.461. (2019). NG-RAN; E1 layer 1. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.462 (2019). NG-RAN; E1 signalling transport. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.463 (2019). NG-RAN; E1 Application Protocol (E1AP). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.541 (2019). Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification TS23.501. (2019). System Architecture for the 5G System. Available at: www.3gpp.org (accessed May 29, 2020).

Jain. S., Kumar, A., Mandal, S. et al. (2013). B4: Experience with a globally-deployed software defined wan. Proceedings of the ACM SIGCOMM Conference, 2013, Hong Kong, pp. 3–14.

## 4.5 Lower-Layer Split

*Jianli Sun*
*Intel Corporation, USA*

As we mentioned previously, operators have many choices of different NG-RAN architectures to deploy; in particular, choices of mapping logical NG-RAN functions into different physical network nodes and further partition of RAN functions. A study conducted by 3GPP in Release-14 discussed some of the possible RAN split options, which are elaborated on in Section 4.1. In Section 4.2, we discussed the high-level split option specified in 3GPP, which divides the RAN into CU and DU. In that split, the CU hosting PDCP and RRC functions can be implemented in a cloud to realize the benefits of centralized RRM and resource pooling. The advantage of the high-level CU/DU split is that it can be deployed using a non-ideal fronthaul transport network, that is, it is particularly well-suited for the case when fiber transport is not available. This, however, comes at a cost as the high-level split does not provide significant gains compared with the monolithic gNB. In the present section we discuss the characteristics of physical layer functions, and the challenges separating these functions into separate network nodes.

Below we explain the details of the lower-layer split (LLS) of an NG-RAN and the reasons why that particular choice of functional split has been adopted by xRAN (which later became O-RAN Alliance). The fronthaul traffic flow procedures and the requirements on the fronthaul transport network to satisfy timing and latency constraints of such deployment are also discussed in this section.

### 4.5.1 Key Ideas

- The benefits of putting more RAN processing resources at a central location were established long time ago. The centralized RAN has the benefits of performances gains due to centralized scheduling and RRM, and resource pooling. Furthermore, it allows the leveraging of commercial off-the-shelf (CoTS) hardware and virtualization. However, these benefits are only limited to functions centralized, hence the desire to centralize as much functionality as possible.
- The desire to move more RAN functions into a centralized processing unit requires a gNB functional split point further down the physical layer, resulting in a physical layer implementation which processes the radio signals that is split across two network nodes. The network nodes in the LLS gNB architecture are connected by the fronthaul interface, which has to transmit a large quantity of the radio data in a short period of time, imposing strict requirements on the transport network bandwidth and latency. Furthermore, physical layer processing requires stringent timing in order to guarantee timely radio signal transmission timing over the air. All of these factors need to be considered when designing the LLS NG-RAN.
- 3GPP studied (3GPP TR 38.801) several LLS options. In option 8, which separates the RF function from the DU, the fronthaul transport carries the time-domain I/Q data streams. In option 7, which puts some physical layer functions in the RU and leaves the rest of the

physical layer functions in the DU, the fronthaul carries the frequency-domain I/Q data streams. Option 7 can be further divided into several suboptions, based on the level of physical layer functions residing in the RU.

- LLS can only be deployed if the fronthaul transport network satisfies its bandwidth, latency, and timing synchronization requirements. If the existing transport cannot sustain these requirements, the transport network needs to be upgraded.
- Despite considering the LLS in the 5G study, 3GPP decided not to standardize it due to concerns with the complexity of the fronthaul network interface. As the 5G deployment schedule approached, several industry players decided to carry on that work in xRAN (later to become O-RAN Alliance). Eventually, O-RAN produced the LLS specification (option 7-2), which is described in this section. Furthermore, CPRI and IEEE1914.3 offer technical solutions for split option 8. The pros and cons of these options are discussed below.
- The fronthaul network interface supporting the LLS is somewhat more complex, compared with the other interfaces used in NG-RAN (e.g. Xn and F1). Therefore, while the standard makes it possible to deploy NG-RAN network nodes (e.g. an RU and a DU) from different vendors, this comes at a cost of more complex system integration and interoperability testing.

### 4.5.2 Market Drivers

NR introduced a plethora of new technologies, such as massive MIMO, flexible numerology, high bandwidth, and support for the mmWave spectrum frequency range. These technologies, among other things, substantially increased the 5G network speed, but they also created new technical challenges compared to 4G. For example, NR baseband processing requires much higher computational capability. In addition, NR serves more diversified user groups such as URLLC, eMBB, and massive Internet of Things (IoT). Deploying a separate mobile network for each user group is prohibitively expensive and may not be feasible due to logistic complexity. Therefore, the traditional integrated physical NG-RAN architecture with monolithic gNB may not be suitable to serve those diverse 5G use cases due to lack of flexibility, performance, and cost. A new NG-RAN physical node architecture is required, and cloud-based NG-RAN is widely considered as a viable solution for addressing the challenges mentioned above.

Another important consideration is the network management, as NR will likely rely on massive small cell deployment alongside macro cells, which creates more challenges in terms of network operation and maintenance. Hence the desire to simplify network management, which becomes significantly simpler with NG-RAN centralization.

The usage of CoTS hardware and virtualization in the cloud are not new concepts, as these have been successfully used in the IT industry for over a decade (see Section 6.2). Most of the cloud-building blocks are off-the-shelf standardized equipment, which greatly reduces cost and increases deployment flexibility. On the transport side, the network building blocks including the switches and routers are readily available (see Section 6.6). Moreover, Ethernet has been widely used for long haul and local networks – 100 Gbps Ethernet has been successfully deployed for IT data center networks, which is capable of satisfying the fronthaul bandwidth requirement. Furthermore, the experiences that the IT industry gained in

cloud network implementations could help with cloud-based NG-RAN. All these factors made RAN ripe for centralization and virtualization, and indeed during LTE, the service operators and equipment vendors made efforts to find technical solutions for cloud RAN implementations. However, these implementations have been mostly based on proprietary technologies.

With 4G, different scale levels were considered for virtualization, ranging from virtual RAN, which uses the virtualization technology on radio network, to distributed RAN, which separates the RAN processing unit into multiple nodes, to cloud RAN, which uses the cloud native technology for RAN. However, these implementations were mostly confined to trials and there have not been massive cloud RAN deployments in LTE so far.

Even though 3GPP considered the physical layer NG-RAN split options during the Release-14 study on 5G (3GPP TR 38.801), this technology was not selected for standardization. That was primarily due to concerns with the complexity of the intra-PHY low-level split, its sensitivity to network interface latency, and stringent requirements on timing synchronization, compared with, for example, the high-level CU/DU split (described in Section 4.1). These factors are closely tied to RAN implementation and pose additional challenges in the interoperability testing, which can be somewhat challenging to be agreed in a large SDO such as 3GPP. Nevertheless, this idea has been considered in several smaller industry consortiums, where reaching consensus and defining a standardized split RAN interface are relatively easy to achieve.

CPRI industry cooperation has specified the split option 8 (described in Section 4.1), which has been used by multiple vendors during LTE. That work continued for NR, leading to the eCPRI specification.

eCPRI is a complete redesign of CPRI, which uses Ethernet-based transport. It supports NR and is capable of supporting several split options besides option 8 supported by CPRI. The eCPRI has specified message types for control plane, user plane, and synchronization plane. Furthermore, it is able to support both eCPRI and CPRI nodes in the same network. However, CPRI specifications are rather abstract and do not provide sufficient details for true multi-vendor interoperability. Later we discuss how the xRAN/O-RAN addresses these issues.

At the same time, another standards organization – IEEE – formed a working group 1914 to explore the fronthaul protocol solutions. IEEE1914.3 specification, also referred to as NGFI, defines the encapsulation and mapping of radio protocols for transport over Ethernet frames. The IEEE1914.3 Ethernet-based transport can be used for example with the 3GPP split option 8 and option 7.1. It offers both structure-agnostic and structure-aware (CPRI frame structure) frames for I/Q data transmission. However, NGFI only defined the transport network aspects, not a fully functional fronthaul interface. For details about Radio over Ethernet (RoE) protocol, refer to the IEEE1914.3 specification.

In addition to CPRI and IEEE, related works were conducted in xRAN Forum and C-RAN Alliance. xRAN Forum is an operator-driven industry consortium focusing on developing standardized RAN network interfaces (including fronthaul interface). Based on the technical feasibility study of the low-level split in 3GPP, xRAN decided to concentrate on a single-split option, leveraging the eCPRI-defined transport protocol and message framework. The goal was to fill in all the missing details, so that the interface between DU and RU would be multi-vendor interoperable. The C-RAN alliance charter was to

promote cloud-based technologies in RAN and white box hardware. After xRAN merged with C-RAN to form O-RAN Alliance, the work on the low-level split started by xRAN continued and produced fronthaul network interface specifications.

### 4.5.3 Functional Split

During the initial 5G study on NG-RAN architecture, 3GPP outlined eight split options, which are explained in Section 4.1. These options cover all possible functional split points within a gNB. These functional splits differ in functionality implemented in the RU and the transport network bandwidth, latency, and synchronization requirements.

3GPP have standardized the high-level CU/DU split (see Section 4.2), while O-RAN defined the architecture in which gNB-DU is further split into DU and RU network nodes. Figure 4.5.1 shows the RAN functional partition with both high-layer and low-layer splits.

In the present section we focus on the low-layer functional split in which the physical layer functionality is distributed across two network nodes (RU and DU), that is, split option 7 (as defined in Section 4.1 and 3GPP TS 38.801).

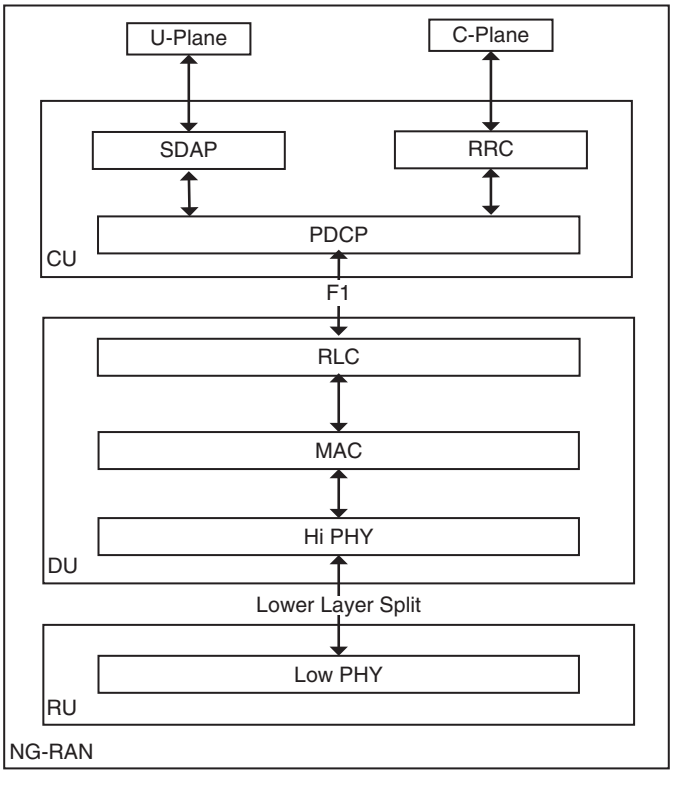



**Figure 4.5.1** RAN architecture with CU, DU, and RU.

#### 4.5.3.1 Fronthaul Bandwidth Requirements

The fronthaul bandwidth requirement is one of the critical factors to be considered when choosing the gNB split option. For some operators, higher fronthaul bandwidth requires deploying (or leasing) new fiber, thus substantially increasing capital spending. Of course, the actual bandwidth requirement for each split option varies depending on configuration and air interface features deployed; however, in order to illustrate the impact on transport network bandwidth of each split option we use one typical network configuration example. In this exercise we assume a 128 TxRx Massive MIMO system with 100 MHz system bandwidth, 16 + 16 I/Q samples, 256QAM, 30 KHz SCS in a sub-6 GHz spectrum. Table 4.5.1 provides a rough estimation of downlink bandwidth for each split option.

This analysis shows that lower split options impose higher requirements on the fronthaul transport network bandwidth. The calculation only takes into account user data transmission; however, in reality there will be also control messages and transport protocol overhead. One obvious way to reduce the transport network bandwidth demand is to select a higher split option (at a cost of reduced performance); however, other solutions are possible (some of which are elaborated upon below).

Data compression is one such technique that may help alleviate the bandwidth demand to a certain extent. There are some popular compression algorithms that are suitable for I/Q data compression, which have been analyzed based on the compression ratio and the error vector magnitude (EVM). The current version of the O-RAN specification defines three compression methods:

- Block floating point
- Modulation compression
- μ-law.

**Table 4.5.1** Front haul transport downlink bandwidth comparison.

| Split option | Bandwidth (Gbps) | Description |
|---|---|---|
| 6 | 5.8 | The Medium Access Control (MAC) and upper layers are in the central unit<br>PHY and RF are in the distributed unit |
| 7-1 | ~376 | Inverse/Fast Fourier Transform (FFT), cyclic prefix (CP) addition/removal and Physical Random Access Channel (PRACH) filtering are in the remote unit<br>All other PHY functions are in the distributed unit |
| 7-2 | ~23 | i/FFT, CP addition/removal, PRACH filtering and precoding are in the remote unit<br>All other PHY functions are in the distributed unit |
| 7-3 | ~5.8 | In downlink only, the encoder resides in the distributed unit<br>All other PHY functions are in the remote unit |
| 8 | ~503 | RF chains are in the remote unit<br>All other PHY functions are in the distributed unit |

The compression is not a mandatory feature – it is up to the vendor whether or not to implement it. Data compression can be very helpful in reducing fronthaul bandwidth requirements; however, there is a cost as data compression may cause signal precision loss, and compression/decompression may increase RU complexity.

#### 4.5.3.2 Low-Level Functional Split Details

Generally, the lower the RAN is split, the more functions can be centralized, and the higher gains due to centralized scheduling and resource sharing can be achieved. On the other hand, lower splits increase fronthaul bandwidth and transport latency requirements, and the RU complexity. Therefore, a good NG-RAN split design needs to balance between these considerations. Those contradicting objectives are hard to address with a single solution.

To better understand the issue we first consider the physical layer processing flows using the Physical Downlink Shared Channel (PDSCH) as an example. Generally, PHY Tx processing includes two stages – bit block processing and complex value symbol processing. The bit stream codeword is scrambled with a 3GPP-defined scrambling bit sequence (3GPP TS 38.211) before the modulation. Scrambling randomizes transmitted bits and makes them less prone to interference. The modulation function converts the bit block into a complex value modulation symbol using 3GPP-specified modulation schemes (3GPP TS 38.211), for example, QPSK, 16QAM, 64QAM, and 256QAM. The next processing function is the layer mapping, which maps the symbols into one or multiple layers. The layer mapping prepares the symbols to be transmitted using spatial multiplexing. The precoding function and beamforming function are used to maximize the radio signal power at the receiving antennas. For the precoding, the weighting parameters are predefined by 3GPP. The last processing block is the iFFT function, which transfers the frequency-domain I/Q samples into time-domain I/Q samples to be transmitted over the air by Tx antenna.

Out of many possible split options, O-RAN has chosen the option 7-2x, which provides a good trade-off, balancing the requirements of reducing fronthaul transport bandwidth and the desire to simplify the radio unit implementation. In particular, it was deemed important to make the RU independent of future 3GPP specification updates, so that it would be possible to implement it entirely in hardware. Therefore, in the split option 7-2x (ORAN-WG4.CUS.0) most of the 3GPP-specific functions of PHY processing reside in the DU, whereas the O-RAN radio unit (O-RU) hosts the following PHY functions:

- Beamforming
- CP addition/removal
- FFT/iFFT functions.

The O-RAN Distribution Unit (O-DU) hosts RLC, MAC and the rest of the physical layer functions (3GPP TS 38.211, 3GPP TS 38.212):

- Scrambling/descrambling
- Modulation/demodulation
- RE mapping/demapping.

For the 7-2x split, the lower physical functions within O-RU are generic and not affected by the future 3GPP spec changes, which will reduce O-RU OPEX and maintenance costs.

O-RAN fronthaul transmits the spatial I/Q data streams instead of antenna I/Q data streams (as in, e.g. split 8). Typically, the number of spatial streams is less than the number of antennas, significantly reducing the transport bandwidth requirement. It is especially beneficial for the base station using a massive MIMO antenna system, which has much larger number of antennas than the spatial streams. The downside of this design choice is that many physical layer functions reside in the O-RU, which requires more processing capabilities and makes the O-RU design somewhat more complex.

The single-split option 7-2x chosen by O-RAN allows two types of O-RU radio categories. The first O-RU type is called Category "A," where the physical layer functions below precoding are located in the O-RU; the second type is the Category "B," which additionally hosts the precoding in the O-RU. Standardization of two O-RU categories provides flexibility for a system designer to choose the best option depending on the target use case. Further details can be found in the O-RAN fronthaul specification (ORAN-WG4.CUS.0).

Figures 4.5.2 and 4.5.3 show the O-RAN downlink functional split for Category "A" and Category "B" radio unit, respectively.

The uplink functional split is similar, as illustrated in Figure 4.5.4, which lists functional blocks residing in O-DU and O-RU. Similar to the downlink, the uplink processing functions perform operation to recover the bit stream transmitted by the UE. In the uplink, after the time-domain I/Q data received from the antenna, I/Q data samples go through the FFT and CP removal to be converted into frequency-domain I/Q samples. Through the beamforming, the radio data are combined into the reduced number of I/Q sample streams. The RE demapping separates UE data and reference signals. The reference signals are sent to the channel estimation function to produce the channel information. The equalization function uses the channel information and received data samples to recover the data sent by UE. The demodulation processing converts complex value I/Q samples into bit blocks. The descrambling function performs the descrambling operation using the same scrambling sequence as the one applied by the UE. The optional compression/decompression function can be used to reduce the fronthaul bandwidth.

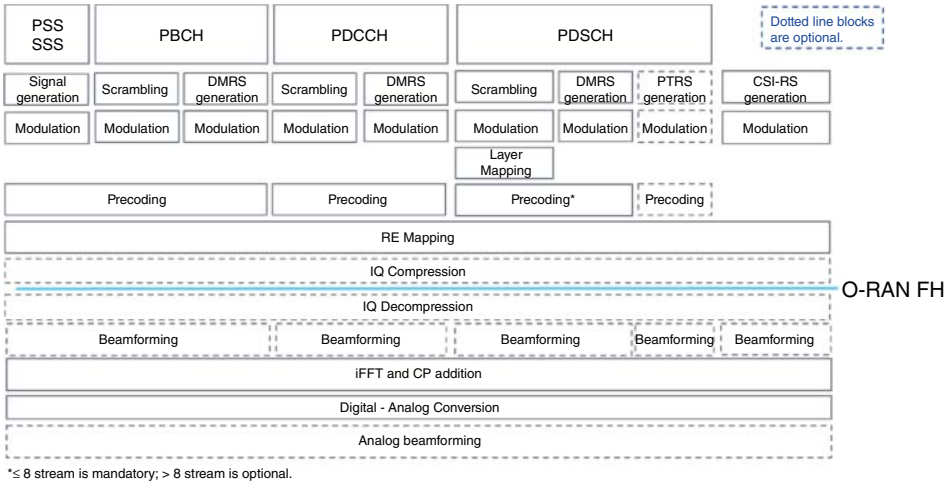



**Figure 4.5.2** Downlink split description, NR, Category "A" Radio. (*Source:* Reproduced by permission of © O-RAN).

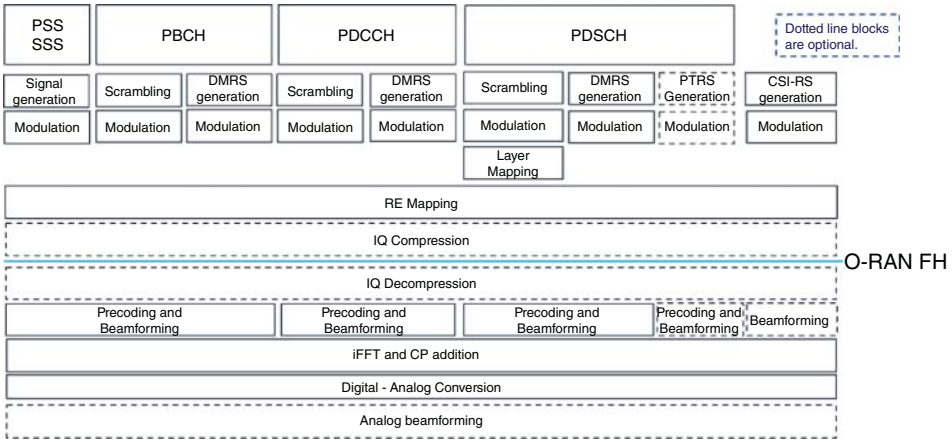



**Figure 4.5.3** Downlink split description, NR, Category "B" Radio. (*Source:* Reproduced by permission of © O-RAN).

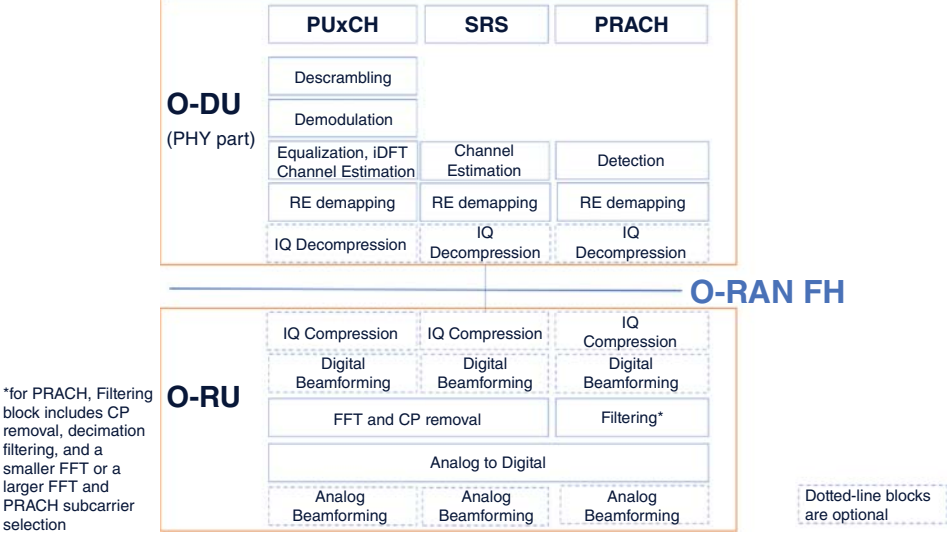



**Figure 4.5.4** Uplink split block diagram. (*Source:* Reproduced by permission of © O-RAN).

#### 4.5.3.3 Latency Management

The move to Ethernet-based fronthaul simplifies implementation and is cost-effective; however, it also brings additional challenges. One of the issues is latency, which is caused by packetizing the radio data in Ethernet transport network (for additional details, see Section 6.6). For RAN, the end-to-end latency includes the delay of radio data during Ethernet encapsulation, data transmission between O-RU and O-DU, and the O-RU processing. The total latency budget also restrains the physical distance between the O-DU and O-RU. Therefore, the fronthaul, like any other time-critical transport network, has a stringent end-to-end latency requirement. It is determined by the radio signal frame timing requirements of the air interface.

Another issue is jitter, which is caused by the network traffic load variation and Ethernet switch buffering. The latency and jitter of the fronthaul interface impact O-RU and O-DU processing budget and timing accuracy. While it is impossible to completely eliminate them, mitigation of these negative factors with a reasonable cost is an important design consideration. The final target is to guarantee that the RAN meets overall system timing and accuracy requirements.

To build a fronthaul network that meets the latency requirements of NR, the deployment use case, network topology, traffic loading, and switching performance should be taken into account. For better understanding of the fronthaul latency issue, an end-to-end latency model is required. We use the latency model defined in eCPRI (which can be used together with O-RAN fronthaul) to analyze different transport network and use case scenarios. The eCPRI model defines timing reference points that are used for delay management, which breaks up the latency into multiple time periods to analyze the latency for downlink and uplink transmission. The reference points defined by eCPRI are reflected in Figure 4.5.5 (eCPRI). These reference points are:

- R1/ R4 – Transmit/Receive interface at O-DU
- R2/R3 – Receive/Transmit interface at O-RU
- Ra - Antenna interface at O-RU.

The latency measurement framework shown above helps to budget (in terms of time) each processing step in both downlink and uplink: from the radio data generation until its transmission over the air, and from radio data reception at the antenna to radio data decoding.

When data are sent across the fronthaul interface between O-DU and O-RU, many uncertain factors such as buffer size, processing time at O-RU, and data transmission over the fiber will affect the timing of the air interface.

In the downlink data transmission, the latency includes: transmission time from O-DU to O-RU, queueing time at O-RU, O-RU data processing time, and the time until the data are transmitted over the air. O-DU must make sure that the data reach the antenna at the right time. Even though the O-RU has buffer to store the received data, the O-DU cannot send the data too early, otherwise they may overflow the buffer.

Using the time reference points, we define the downlink delay as T1a = T12 + T2a. The transmit window is then defined as the time period that allows the transmitter to send all the data to the receiver. It is not an exact transmission time, but rather a specified boundary

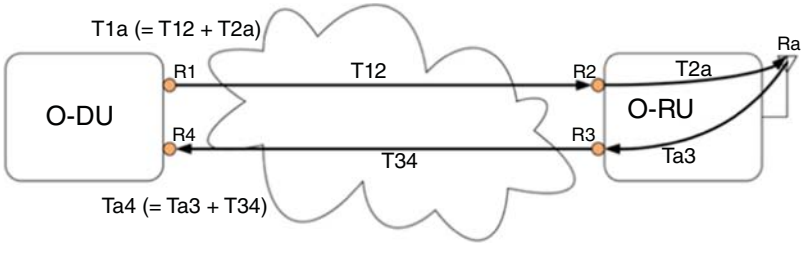



**Figure 4.5.5** Definition of reference points for delay management. (*Source:* Reproduced by permission of © O-RAN).

**Table 4.2.1** Uplink transmit and receive window.

| | Downlink | Uplink |
|---|---|---|
| Transmit window | T1amax – T1amin | Ta3max – Ta3min |
| Receive window | T2amax – T2amin | Ta4max – Ta4min |

that the transmitter has to observe. Similarly, the receive window is the time period that the receiver is able to accept data, with some buffer time to process the data.

In the uplink direction, when O-RU receives a piece of sampled data from the antenna, it starts processing when radio samples of one symbol are collected. After the processing is completed, data are packed and sent through the fronthaul interface toward the O-DU for further processing. The time for the radio data to reach the O-DU should be sufficiently short to leave enough time for the O-DU to process the data within the HARQ timing requirement.

Similar to downlink, we define the uplink delay as Ta4 = Ta3 + T34. The uplink transmit and receive windows are expressed in Table 4.5.2.

The transmit and receive windows defined above are key parameters in managing the fronthaul transport network latency and jitter. Once the transmit and receive windows are calculated based on the dynamic or static latency measurement, the O-DU and O-RU can ensure that the fronthaul I/Q data streams are sent or received within the corresponding window.

### 4.5.4 Fronthaul Interface

In this section we describe the O-RAN fronthaul interface, which has been defined to support the functional split 7-2 described above.

O-RAN specifications allow several options for the transport protocol. The choice of the transport protocol is up to a vendor and a service provider. Regardless of the protocol used, the transport network should fulfill the QoS requirements defined in the standard. For example, even though the underlying transport media can be shared (e.g. with other radio access technology), the latency requirements of the 5G radio data need to be protected. One way to achieve this is to assign the highest priority to the fronthaul traffic flow, compared with other types of traffic carried by the same transport network. This is further elaborated on in Section 6.6, while some high-level considerations on the transport network usage with O-RAN fronthaul are provided below.

Ethernet is the most commonly used fronthaul transport, even though other protocols such as IP/UDP can also be used. The fronthaul traffic identifier can be used for accelerating data package processing. When Ethernet is used to carry both user-plane and control-plane traffic, the fronthaul packets can be identified by a specially allocated Ethertype field. Either eCPRI Ethertype or IEEE1914.3 Ethertype can be used. When IP/UDP is used as fronthaul transport protocol, the fronthaul messages may be identified by the designated UDP port configured via the management plane during system startup.

The fronthaul transport is considered secured, and therefore no additional security measures are specified in O-RAN to protect the control-plane and user-plane traffic. That is, it is assumed that the security is provided by the transport network itself.

#### 4.5.4.1 Messages

The fronthaul traffic falls into three categories in term of message types:

- Control plane
- User data
- Synchronization.

The control messages are used to control user data scheduling, beamforming weight selection, and numerology selection, etc. User data messages carry radio data between O-DU and O-RU. Synchronization messages are used for timing synchronization between O-DUs and O-RUs via Ethernet, which rely on Precision Time Protocol (PTP) (IEEE 1588-2008).

All three types of message are illustrated in Figure 4.5.6 (ORAN-WG4.CUS.0).

O-RAN control- and user-plane protocols stacks are the same: the Ethernet physical layer as the first layer, then Ethernet MAC. If Ethernet is chosen as the fronthaul transport, the eCPRI or RoE protocol layer becomes the Ethernet payload to carry the control or user data message. For the fronthaul using IP/UDP, the IP and UDP layers are added after Ethernet MAC, and eCPRU/RoE is embedded as UDP payload. The fronthaul control message payload within the eCPRI/RoE is the same regardless of using either Ethernet or IP/UDP transport.

Synchronization messages can only use the Ethernet protocol, but not IP/UDP.

Figure 4.5.7 further illustrates the O-RAN protocol stack message frame in greater detail for both Ethernet and IP/UDP transports. Besides the transport protocol layer, the core of the O-RAN package is the payload.

The O-RAN message payload consists of three building blocks: the transport header, the payload common header, and the payload section fields. In the interest of brevity, we use one control message and one data message to explain some key fields of these building blocks. Exhaustive description of all messages is beyond the scope of this section and can be found in the O-RAN specification (ORAN-WG4.CUS.0).

As mentioned earlier, both eCPRI and IEEE1914.3 headers can to be used in the transport frame differentiated by their corresponding Ethertype. Both headers have the same length of 8 bytes. Below we describe how eCPRI and IEEE1914.3 can be used with O-RAN fronthaul.

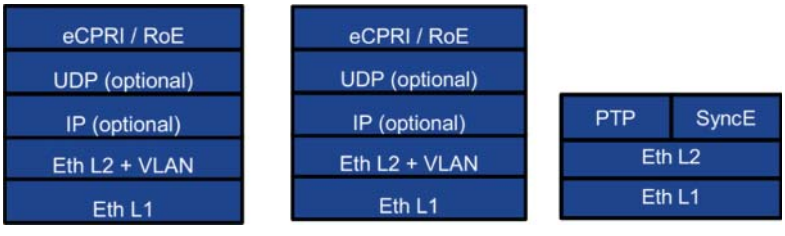



**Figure 4.5.6** Control (left), user (center), and synchronization (right) message protocol stack. (*Source:* Reproduced by permission of © O-RAN).

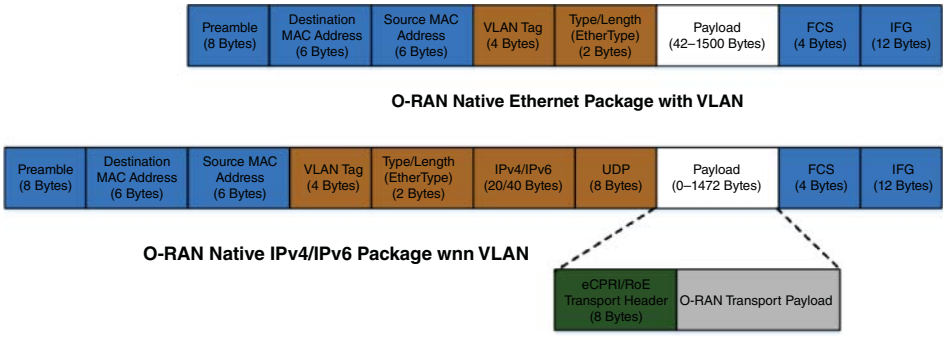



**Figure 4.5.7** O-RAN transport protocol stack diagram.

The eCPRI Transport Header, shown in Figure 4.5.8, carries the following information:

- ecpriVersion – The eCPRI protocol version. Default Value: 0001b (eCPRI version 1.0).
- ecpriReserved – The reserved field, not used. Default Value: 000b.
- ecpriConcatenation – The indicator of multiple eCPRI messages concatenation within single payload. Default value: 0 b (no concatenation).
- ecpriMessage – The message type. Valid Values: 0×0 (user-plane data) or 0×2 (control-plane data) or 0×5 (network delay measurement messages).
- ecpriPayload – The eCPRI message payload size in bytes.
- ecpriRtcid/ecpriPcid – The component_eAxC identifier (c_eAxC ID) which is used to identify the specific data flow associated with each control-plane (ecpriRtcid) or user-plane (ecpriPcid) message.
- ecpriSeqid – The message sequence ID to identify the message ordering within an eAxC message stream.

The IEEE1914.3 header, shown in Figure 4.5.9, carries the following information:

- RoEsubType – The payload type of the IEEE 1914.3. Value range 128–131, which maps to the eCPRI header field combination of ecpriMessage and cpriConcatenation.
- RoEflowID – The flow ID between endpoints. It is not used in O-RAN.
- RoElength – The message payload size in bytes.
- RoEorderInfo – The field is split into seven subfields:
  - DU_Port_ID – Identifier of processing units at O-DU.
  - BandSector_ID – The band and sector ID supported by O-RU.
  - CC_ID – The carrier components ID supported by the O-RU.
  - RU_Port_ID – The spatial streams or beams ID used by the O-RU.
  - Sequence_ID – The message sequence ID.
  - E_Bit – The last message indicator pertaining to the section.
  - Subsequence_ID – The subsequence ID.

Figure 4.5.10 shows the O-RAN message payload common header and section fields. We describe the two blocks separately below.

Payload common header fields:

- dataDirection – The indicator of data direction (gNB Tx/Rx).

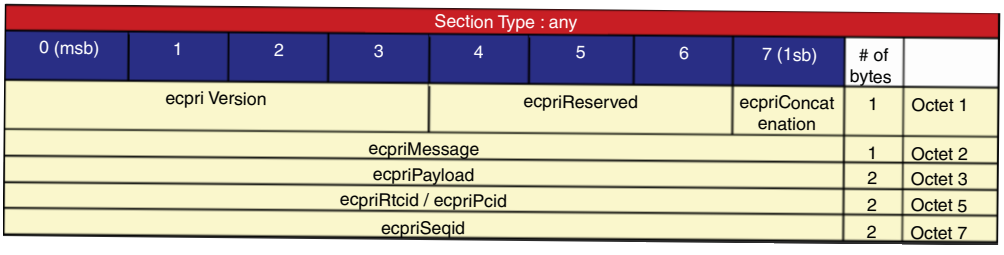



**Figure 4.5.8** eCPRI header table. (*Source:* Reproduced by permission of © O-RAN).

| Section Type : any | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 (msb) | 1 | 2 | 3 | 4 | 5 | 6 | 7 (1sb) | # of bytes | |
| RoEsubType | | | | | | | | 1 | Octet 1 |
| RoEflowId | | | | | | | | 1 | Octet 2 |
| RoElength | | | | | | | | 2 | Octet 3 |
| RoEorderInfo | | | | | | | | 4 | Octet 5 |

**Figure 4.5.9** IEEE1914.3 header table. (*Source:* Reproduced by permission of © O-RAN).

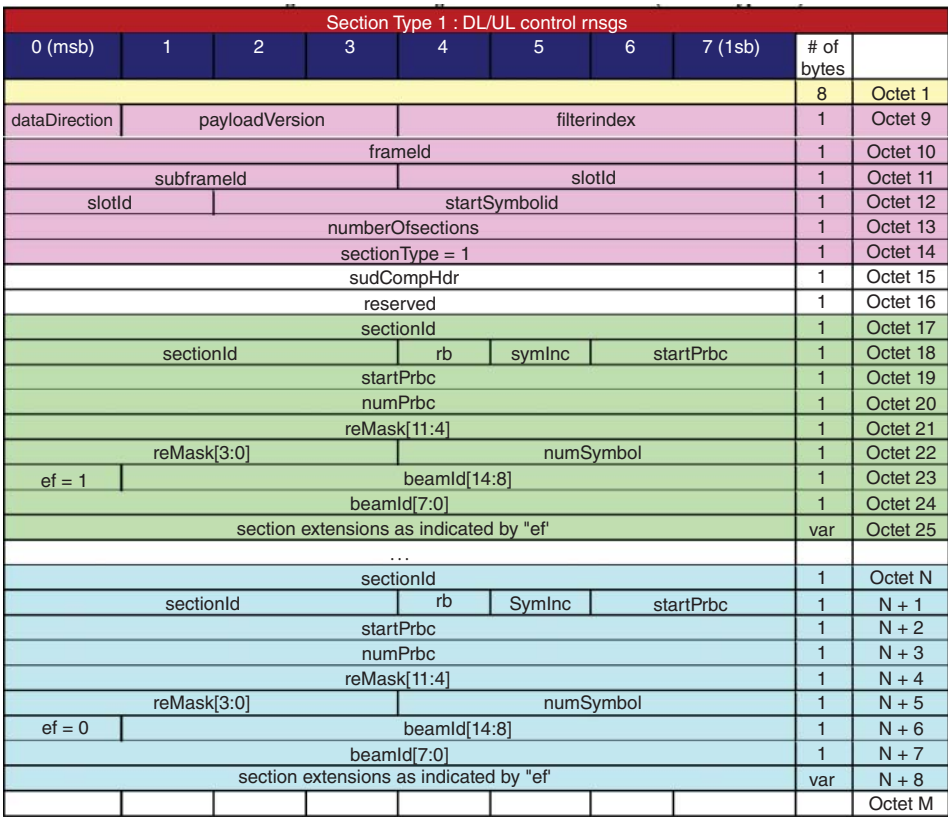



**Figure 4.5.10** Control plane section type 1 message format. (*Source:* Reproduced by permission of © O-RAN).

- payloadVersion – The payload version field. Value = "1" (first version for payload format).
- filterIndex – The filter index.
- frameId – The frame identifier.
- subframeId – The subframe identifier.
- slotID – Slot identifier.
- startSymbolid – The start symbol ID number.
- numberOfsections – The number of sections.
- sectionType – The section type. Value = "1" for section type 1.
- udCompHdr – The user data compression header containing compression method and IQ bit width information.
- reserved – The reserved field.

Payload section fields:

- sectionID – The section ID used to identify the section.
- rb – The resource block indicator.
- symInc – The symbol number increment indicator.
- startPrbc – The starting physical resource block (PRB) of data section description.

| Section Type 1,3 : DL/UL IQ data msgs | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 (msb) | 1 | 2 | 3 | 4 | 5 | 6 | 7 (1sb) | # of bytes | |
| transport header | | | | | | | | 8 | Octet 1 |
| dataDirection | payload Version | | | filterIndex | | | | 1 | Octet 9 |
| frameId | | | | | | | | 1 | Octet 10 |
| subframeId | | | | slotId | | | | 1 | Octet 11 |
| slotId | | symbolid | | | | | | 1 | Octet 12 |
| sectionId | | | | | | | | 1 | Octet 13 |
| sectionId | | | | rb | symInc | startPrbu | | 1 | Octet 14 |
| startPrbu | | | | | | | | 1 | Octet 15 |
| numPrbu | | | | | | | | 1 | Octet 16 |
| udCompHdr (not always present) | | | | | | | | 1 | Octet 17 |
| reserved (not always present) | | | | | | | | 1 | Octet 18 |
| udCompParam (not always present) | | | | | | | | 1 | Octet 17/19 |
| iSample (1st RE in the PRB) | | | | | | | | 1* | Octet 18/20 |
| qSample (1st RE in the PRB) | | | | | | | | 1* | Octet 19/21* |
| . . . | | | | | | | | | |
| iSample (12th RE in the PRB) | | | | | | | | 1* | Octet 40/42* |
| qSample (12st RE in the PRB) | | | | | | | | 1* | Octet 41/43* |
| udCompParam (not always present) | | | | | | | | 1* | Octet 42/44* |
| iSample (1st RE in the PRB) | | | | | | | | 1* | Octet 43/45* |
| qSample (1st RE in the PRB) | | | | | | | | 1* | Octet 44/46* |
| . . . | | | | | | | | | |
| iSample (12th RE in the PRB) | | | | | | | | 1* | Octet 65/67* |
| qSample (12st RE in the PRB) | | | | | | | | 1* | Octet 66/68* |
| . . . | | | | | | | | | |
| sectionId | | | | | | | | 1 | Octet M |
| sectionId | | | | rb | symInc | startPrbu | | 1 | M + 1 |
| startPrbu | | | | | | | | 1 | M + 2 |
| numPrbu | | | | | | | | 1 | M + 3 |
| udCompHdr (not always present) | | | | | | | | 1 | M + 4 |
| reserved (not always present) | | | | | | | | 1 | M + 5 |
| udCompParam (not always present) | | | | | | | | 1 | M + 4/5 |
| iSample (1st RE in the PRB) | | | | | | | | 1* | M + 5/7 |
| qSample (1st RE in the PRB) | | | | | | | | 1* | M + 6/8* |
| . . . | | | | | | | | | |
| iSample (12th RE in the PRB) | | | | | | | | 1* | M + 27/29* |
| qSample (12st RE in the PRB) | | | | | | | | 1* | M + 28/30* |
| udCompParam (not always present) | | | | | | | | 1* | M + 29/31* |
| iSample (1st RE in the PRB) | | | | | | | | 1* | M + 30/32* |
| qSample (1st RE in the PRB) | | | | | | | | 1* | M + 31/33* |
| . . . | | | | | | | | | |
| iSample (12th RE in the PRB) | | | | | | | | 1* | M + 52/54* |
| qSample (12st RE in the PRB) | | | | | | | | 1* | M + 53/55* |

**Figure 4.5.11** Data plane message format. (*Source:* Reproduced by permission of © O-RAN).

- numPrbc – The number of contiguous PRBs per data section description) field: 8 bits.
- reMask – The resource element mask.
- numSymbol – The number of symbols.
- ef – The section extension flag.
- beamId – The beam ID.

Figure 4.5.11 shows the I/Q data message. The payload common header fields and section fields are described below.

Payload common header fields:

- dataDirection – The data direction indicator (gNB Tx/Rx).
- payloadVersion – The payload version number. Value = "1" (first version for payload format).
- filterIndex – The filter index.
- frameId – The frame ID number.
- subframeId – The subframe ID number.
- slotID – The slot ID number.
- symbolId – The symbol ID number.

Payload section header fields:

- sectionID – The section ID number.
- rb – Indicates whether every resource block is used or every other resource block is used.
- symInc – The symbol number increment indicator.
- startPrbu – The starting PRB ID number of the user-plane section.
- numPrbu – The number of contiguous PRBs per data section.
- udCompHdr – The user data compression header containing compression method and IQ bit width information.
- reserved – The reserved field.
- udCompParam – The compression method specific parameter.
- iSample – The in-phase data sample.
- qSample – The quadrature data sample.

The rest of the definitions of O-RAN messages can be found in the O-RAN spec (ORAN-WG4.CUS.0).

#### 4.5.4.2 Scheduling Procedure

In this subsection we illustrate the LLS split operation by one typical call flow of the scheduling procedure, which is controlled by the O-DU. When the MAC scheduler allocates the UE on a specific subframe, a series of control- and user-plane messages are exchanged between O-DU and RU. It is possible to bundle multiple control messages together in a single message or send each one separately. The O-DU is also able to send the UE scheduling Information as well as the PRB usage information to the O-RU. When the system is not fully utilized (some of the PRBs are not used), the I/Q data of unused PRBs are not sent to the O-DU, which may lead to power saving.

Figure 4.5.12 shows the steps of the scheduling procedure (ORAN-WG4.CUS.0).

The downlink transmission procedure (Figure 4.5.12) is performed as follows:

- For given time slot "n," one or more control messages are sent by the O-DU to the O-RU.
- The O-RU makes preparation for receiving IQ data based on received control message(s).
- The O-DU sends the IQ data to the O-RU in multiple messages.
- The RU receives the IQ data, assembles the data in the right sequence, and puts I/Q data at allocated buffers.
- The above steps are repeated for the next subframe.

The uplink transmitting procedure is similar and is therefore not shown here.

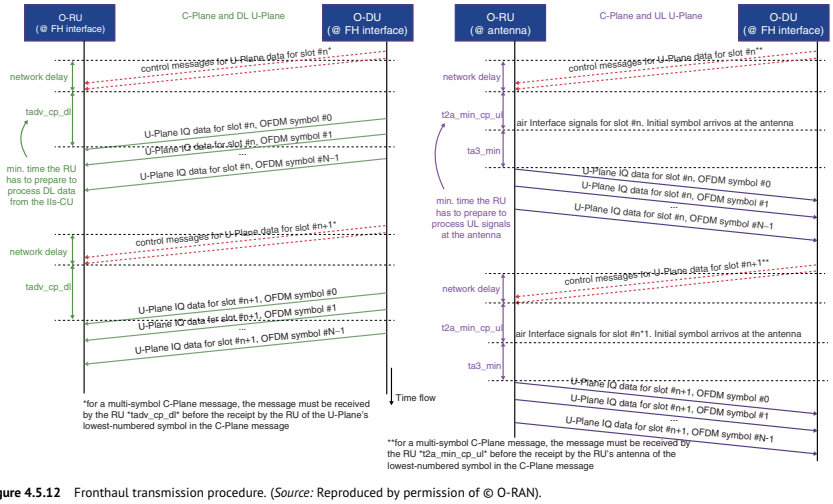



**Figure 4.5.12** Fronthaul transmission procedure. (*Source:* Reproduced by permission of © O-RAN).

#### 4.5.4.3 Beamforming Methods

Beamforming is an optional feature in 5G, which is supported by the O-RAN fronthaul via the control-plane procedures described below. All three types of beamforming – digital, analog, or hyper beamforming are supported.

When selecting a beamforming method, one needs to consider both fronthaul transport network impact and O-RU complexity. Some methods described below require significant amounts of beamforming-related information to be transferred over the fronthaul interface. Furthermore, the frequency of beamforming information update is also different for different methods. Finally, some methods require more processing capabilities in an O-RU, which may increase the hardware complexity and cost.

##### *4.5.4.3.1 Beamforming Indexing Method*

To use this method, the O-RU must be preconfigured with a table of beamforming weights and their assigned indexes. The beam index can then be used in either digital or analog beamforming, and one additional method described below.

The management-plan specification defines the file download method that can be used to provision the beamforming table. Alternatively, the OAM can be used or the beamforming table could be preconfigured in the persistent memory of the O-RU by the vendor.

##### *4.5.4.3.2 Real-Time Weights Method*

In this method, the O-DU sends the beamforming weights generated in real time to the O-RU to be associated with a specific user's data. A beam index can be assigned (and used in subsequent messages) if the beamforming weights are stable for a period of time.

##### *4.5.4.3.3 Beam Attributes Method*

In contrast to the previous method in which the O-DU provides the weights, this method relies on the O-RU to generate the beamforming weights using the beam attribute, which requires additional processing capability in the O-RU.

##### *4.5.4.3.4 Channel Information Method*

In this method the RU generates beam weights based on the channel information.

### 4.5.5 Fronthaul Timing Synchronization

IEEE1588 (IEEE1588) is used as the timing synchronization protocol for the O-DU and O-RU local clock synchronization. There are several ways to synchronize O-DU and O-RU, depending on the location of the master clock. Below we explain differences between those various configurations referred to as C1, C2, and so on. In configurations C1 and C2, the O-DU holds the timing source, which can be either a grandmaster or boundary clock. Alternatively, in C4, Global Navigation Satellite System (GNSS)-based synchronization is used by the O-RU. These configurations are illustrated by Figure 4.5.13.

- In configuration C1, the O-DU is directly connected to the O-RU. The O-DU acts as a grandmaster clock or boundary clock, so O-RU synchronizes with O-DU via PTP.
- With configuration C2, the O-DU is connected with the Primary Reference Time Clock (PRTC). The clock source acts as the master clock and distributes timing to the O-RU. The fronthaul allows more than one Ethernet switch hop. The total allowed network noise limits the number of switches.

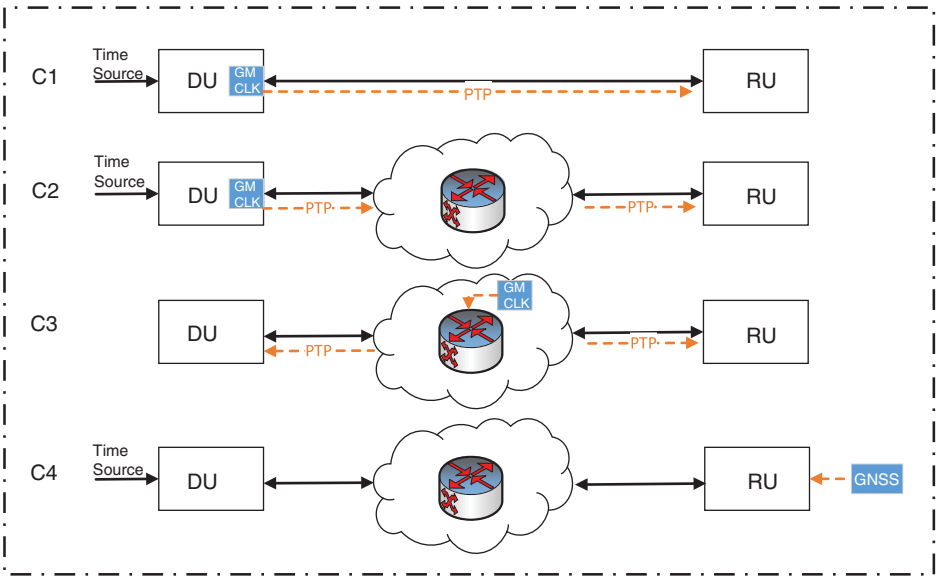



**Figure 4.5.13** Fronthaul timing synchronization configurations.

- In configuration C3, the PRTC grandmaster is provided by the network switch within the fronthaul network, the timing synchronization is distributed to the O-DU and the O-RU via the PTP network. The number of hops between switches is limited by frequency and time error in the distribution chain.
- Configuration C4 is a special case, which does not distribute the timing via fronthaul networks. The O-RU has a local PRTC-traceable time source such as GNSS.

The choice of the timing synchronization configuration depends on the deployment use case of the NG-RAN, transport network capability, and availability of the timing source.

### 4.5.6 Operation, Administration and Maintenance (OAM)

There are two architecture options for the fronthaul OAM: hierarchical and hybrid, as shown in Figure 4.5.14.

In the hierarchical model (ORAN-WG4.MP.0), the network management system (NMS) communicates with the O-RU via the O-DU. All the messages from the NMS are passed to the O-RU by the O-DU. On the other hand, in the hybrid architecture (ORAN-WG4.MP.0) the NMS directly connects both the O-DU and the O-RU. The configuration information, the radio operation status, and error conditions can be directly sent and received between the network management and the O-RU.

The OAM can share the transport network with the control and user data traffic. Alternatively, a dedicated transport network can be used. In the case of shared transport network, the OAM traffic shall be labeled as lower priority to minimize impact on the regular fronthaul traffic (control, user, and synchronization [CUS] plane messages) latency.

The OAM protocol stack is shown in Figure 4.5.15.

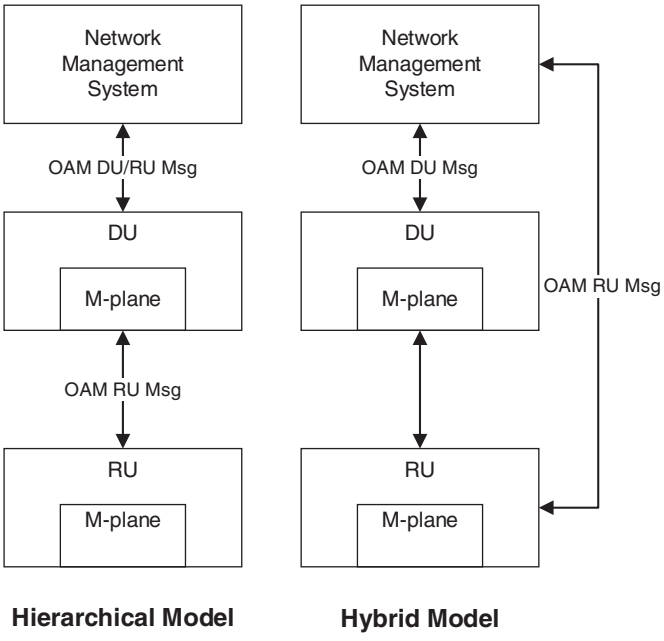
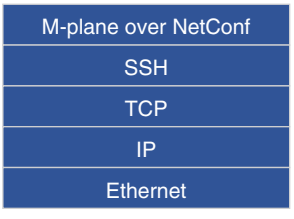



**Figure 4.5.14** Management plane architecture options.

**Figure 4.5.15** Management plane protocol stack.




The management plane protocol, which is based on NETCONF/Yang data models, supports the following features:

- Initial installation
- Fronthaul interface management
- O-RU and O-DU software management
- Configuration management
- Performance management
- Fault management
- Timing synchronization.

For further details refer to the O-RAN OAM specification (ORAN-WG4.MP.0).

### 4.5.7 Further Reading

This section outlines the basic principles of the O-RAN functional split and the fronthaul interface defined to support it, covering control, data, and management protocols between O-DU and O-RU.

For further details, refer to the O-RAN Fronthaul CUS Plane specification and management plane specification provided below.

## References

3GPP Technical Report 38.801 (2017). Study on new radio access technology: Radio access architecture and interfaces. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.211 (2019). NR; Physical channels and modulation. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.212 (2019). NR; Multiplexing and channel coding. Available at: www.3gpp.org (accessed May 29, 2020).

eCPRI Specification V1.0 (2017). Common Public Radio Interface: eCPRI Interface Specification. Available at:www.cpri.info (accessed May 29, 2020).

IEEE1588 (2002). IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. Available at: www.ieee.org (accessed May 29, 2020).

IEEE1914.3 (2019). IEEE Standard for Radio over Ethernet Encapsulations and Mappings. Available at: www.ieee.org (accessed May 29, 2020).

IEEE802-1CM (2018). Time-Sensitive Networking for Front Haul. Available at: www.ieee.org (accessed May 29, 2020).

ORAN-WG4.CUS.0-v01.00 Technical Specification (2019).O-RAN Fronthaul Working Group Control, User and Synchronization Plane Specification. Available at: www.organ.org (accessed May 29, 2020).

ORAN-WG4.MP.0-v01.00 Technical Specification (2019).O-RAN Alliance Working Group 4 Management Plane Specification. Available at: www.organ.org (accessed May 29, 2020).

ORAN-WG4.MP.0-v01.00 Technical Specification (2019). O-RAN Fronthaul Working Group Management Plane Specification. Available at: www.organ.org (accessed May 29, 2020).

## 4.6 Small Cells

*Clare Somerville*
*Intel Corporation, UK*

The concept of small cells first became popular in the 3G era, when they were used to solve the problem of coverage blackspots at home. In this first use case they were called femtocells and provided to a subscriber, by a network operator, who plugged the femtocell into their home broadband to gain a 3G service. With the advent of 4G the scope for small cell deployments grew as they were utilized as a way of providing improved indoor coverage in offices and shopping malls, but also, as way to offload data from a macro cell in traffic hotspots. However, although there is this history of small cells stretching back over a decade, it is 5G that is the first generation which is likely to truly embrace them.

The prominence of small cells in 5G comes from the ever-increasing data consumption by users, as data demand continues to grow eventually the only way to support this growth will be by building a denser network with more cells, so there will be more base stations but each transmitting at a lower power and over a shorter distance, creating an ultra-dense network of small cells.

In addition, most data are generated indoors where coverage is often poor, and small cells are the best way to improve indoor mobile coverage.

Low transmission power is the key constraint of a small cell, with 3GPP specifying a local area base station class with maximum transmission power of 24 dBm, whereas for their wide area base station (macro cell) there is no upper power limit (3GPP TS 38.104, Section 6.2.1). Beyond this 3GPP power definition there are no other specific constraints on what a small cell is, resulting in small cells being used in many variants across a large range of use cases. This section highlights popular use cases, different variants of small cells, the NG-RAN architecture used for them, and interfaces defined to enable the small cell ecosystem.

### 4.6.1 Key Ideas

- In 5G small cells are mainstream, and the different 5G disaggregation options and frequency choices apply equally to both small cell and macro cells.
- Barriers to small cell deployment are predominately business related, not technical.
- The trend to virtualization and centralization in 5G also applies to small cells.
- Small cells will carry a significant portion of the data in a mobile network.
- The largest market for small cells is indoor, while the biggest growth area for small cells is outdoor.
- Private networks and unlicensed networks are more specialized network options uniquely suited to small cells.
- mmWave deployments will use small cells, but there will also be many small cells operating in the sub-6 GHz frequency range.
- Private enterprise networks bring together technologies of small cells, edge services, and distributed EPC.

- The small cell industry has developed common APIs to allow an open RAN small cell ecosystem.
- One of these APIs, called FAPI, is internal to a small cell and is widely adopted by the small cell industry. FAPI is defined between one instance of MAC and one instance of the PHY layer.
- nFAPI is a wrapper around FAPI that supports 3GPP split option 6 with the MAC located in a CU, and the PHY located in the DU.
- nFAPI is designed to operate over non-ideal connections such as Ethernet found in indoor enterprise small cell networks.

### 4.6.2 Market Drivers

Interest in small cells and the continued drive to increase their deployment can be broadly split into two main market drivers: first, the desire to improve indoor coverage for mobile users; and second, the need to densify networks to support increasing data throughput demands.

Indoor coverage for mobile users has always been problematic due to the challenges of radio waves propagating into buildings. This challenge is becoming progressively harder as the insulation characteristics of buildings improve, which makes it even harder for radio waves to enter, and the increase in carrier frequencies used for cell transmission resulting in lower inherent abilities for radio wave penetration. With 5G networks deploying higher frequencies than any previous generations of networks (2G, 3G, and 4G all focused on sub-6 GHz, whereas 5G additionally introduces mmWave), this makes indoor coverage an increasingly pressing deployment problem to solve with small cells.

The market driver for indoor coverage is also related to the outdoor small cell market driver, in that the majority of mobile users generate most of their data while stationary and indoors. Today, this is frequently offloaded to a Wi-Fi network, but as mobile data packages become more generous there is less and less incentive for mobile users to seek an alternate Wi-Fi network and instead they use operator network for all their data use. This makes indoor small cell networks a key part of both providing ubiquitous 5G coverage and the network densification required to support the increasing data demand from smartphones.

In an outdoor environment small cells are critical components for improving data capacity of a network in an urban area as part of a heterogeneous network (HetNet). HetNets are a well-known concept and are networks with an over-arching macro cell with multiple smaller cells providing coverage; the small cells are placed either where extra capacity is needed or where there is a coverage deadspot. These small cells can be fully independent from the macro cell, meaning that a UE must hand over into the small cell, or they can be deployed as a carrier of the macro cell via CA. All of these outdoor scenarios are already used in today's LTE networks, but are expected to expand in usage and importance for 5G as the smartphone data demands continue to increase.

Finally, in addition to the market drivers for smartphones and today's mobile networks, small cells are also expected to play an important role as 5G becomes adopted in vertical segments. Many of these vertical segments generating initial interest are private enterprise networks, or indoor networks, for example, smart factories embracing industry 4.0. Both enterprise and indoor networks are typically created from small cells.

### 4.6.3 Barriers and Solutions

With the many small cell market drivers, which already existed in LTE networks, why aren't small cells already ubiquitous? The answer is that small cells have deployment challenges, and these deployment challenges are a mix of both technical and business oriented.

As the small cell industry matures these challenges are being overcome, and this section highlights these barriers and the current status of the solutions to them. Frequently, the barriers and solutions are different for indoor and outdoor scenarios.

#### 4.6.3.1 Site Locations

For outdoor small cell deployments sourcing sufficient site locations is a significant barrier. With small cell inter-site distances of less than 100 m, compared with 0.5–>1 km for a suburban macro cell, significantly more sites need to be found. A city center location will have street furniture, including, for example, streetlights and bus/tram stops. But equally a city center is likely to have three or four mobile operators looking for site locations. These site locations also need access to both power and backhaul connections.

For indoor small cell deployments, the initial barrier is that the cooperation of the building owner and possibly tenant are required. This is easier if the building owner is looking to improve their indoor coverage, but will still require an agreement between the building owner and mobile operator over the installation costs and ownership of the installed equipment. These indoor locations will also typically require a server room, for hosting the connection back to the operator's network, and a good-quality connection from the server room to the indoor cell site locations.

There are new initiatives designed to help overcome these barriers. The first is in the creation of relevant documentation: recommendations for footprints of small cells (Global 5G study) and what is required for new buildings to future-proof indoor installations (SCF 214). The second is the introduction of new small cell business and deployment models, specifically, neutral host small cells, which allow multiple operators to share the same site.

#### 4.6.3.2 Scaling Up Deployment

A challenge to the 5G vision for an ultra-dense network is the requirement to successfully deploy, maintain, and manage the network, and it is worth noting that one of the most expensive tasks in network management is to send an engineer to a cell site.

To aid in the initial deployment task SONs are a key aspect and reduce the level of cell planning required. At the beginning of the small cell industry, with residential femtocells, a consumer would buy, or be given, a femtocell, which self-installed. With today's indoor enterprise and outdoor small cell networks, this has changed as an operator or system integrator will install each cell; however, due to limited site choice, self-optimization is still popular as it makes it easier for a small cell network to adjust parameters to ensure it operates efficiently.

To reduce the complexity and cost of both maintenance and management of small cell networks, Network Function Virtualization and Software Defined Networks (NFV/SDN) can be used. This technique has been most extensively adopted for core networks but can also extend to the RAN. An NFV/SDN network separates control and data planes and abstracts network functions from underlying hardware. This makes it easier to scale up

and down processing as required and easier to automate control of the network functions. The improved manageability of virtualized networks is a big driver for their adoption.

#### 4.6.3.3 Backhaul

Providing a suitable backhaul connection to a small cell site is a major challenge, and barrier, for the expansion of small cell networks. The backhaul requirements of a small cell are less than a macro cell; however, there are significantly more cell sites and, as they are typically in urban locations, gaining a physical connection to a small cell site is a challenge.

As a consequence of this difficulty multiple types of backhaul have been proposed to help solve the backhaul challenge for small cells (SCF 049):

- Fiber is the traditional backhaul for macro cells and where feasible is a good choice for small cells.
- Wireless, where both sub-6 GHz and mmWave solutions have been implemented.
- Widely available wired networks such as broadband (digital subscriber line technologies [xDSL]) and cable networks.
- Satellite is an option useful for remote or rural small cell networks.

Furthermore, technologies such as Integrated Access-Backhaul (IAB) (see Section 5.2), which are being developed in 3GPP Release-16, may help alleviate the backhaul issue.

#### 4.6.3.4 Edge Compute

Finally, in this section it is worth mentioning multi-access edge compute (MEC). It is not a barrier to small cell deployments, rather an enabler, as small cells combined with MEC are a valuable solution for many low-latency services promised with 5G, and in particular this combination is expected to be the preferred solution for 5G deployments in vertical segments, such as industrial control and autonomous driving (ETSI-MEC).

The applicability of MEC solutions to 5G small cell networks will require a small cell network to be dimensioned to provide compute capacity for MEC services. In indoor networks this MEC compute is likely to be located in a server room, while for outdoor networks this compute will be at an aggregation point.

For further details about MEC, see Section 6.4.

### 4.6.4 Small Cell Variants

With small cells covering a plethora of use cases there are multiple types of small cells and these variations can be broadly split into different groups, each covered in this section:

- Disaggregation architectures
- Hardware/platform architectures
- Operating frequencies
- Operational modes.

#### 4.6.4.1 Disaggregation Architectures

An introduction to and the motivations for improved support of disaggregated networks have already been described earlier in Section 4.1. It is important to note that disaggregation is equally beneficial to small cell networks; in fact LTE small cells were some of the first

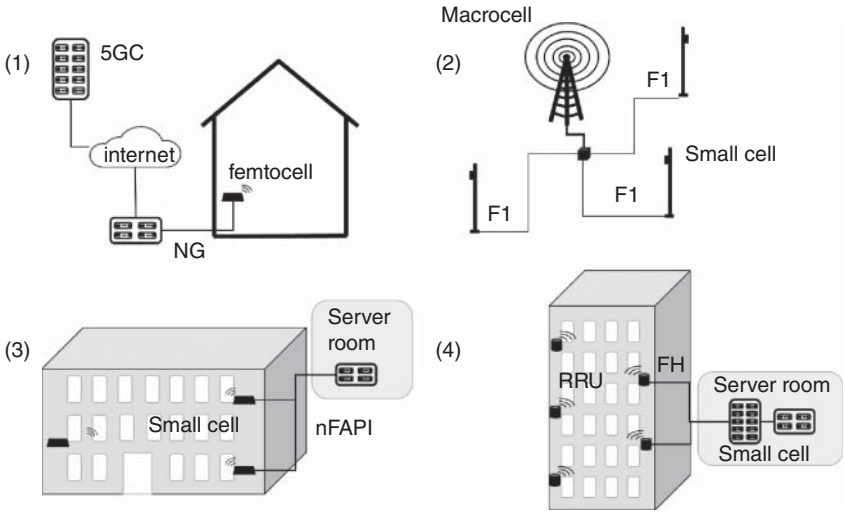



**Figure 4.6.1** Small cell disaggregation architectures.

architectures to adopt disaggregation, in particular with small cell stadium deployments. Figure 4.6.1 shows the four most common disaggregation points within the context of small cell networks. While examples of likely deployments are given for each network split, for a specific deployment there are often several disaggregation possibilities, and frequently the availability and characteristics of the backhaul are the decider in the adopted split.

The first disaggregation point is the traditional wireless network architecture where the NG interface (3GPP TS 38.41x series) is used as the connection over the transport network back to the 5GC. The example given for this split is an indoor residential small cell with xDSL or Data Over Cable Service Interface Specification (DOCSIS) backhaul, which was the first widely adopted use case for small cells. Asymmetric digital subscriber line (ADSL) or cable is used to route the backhaul to the telephone or cable TV provider's core network, from where it is subsequently routed to the wireless network operator's core network. This routing method results in a connection that is called a non-ideal backhaul (3GPP TR 36.932) and has latency and bandwidth characteristics only suitable for the NG interface. Another notable advantage of the traditional NG split is that integration (interoperability) with surrounding base station equipment is optional; small cells may still utilize the Xn interface (3GPP TS 38.42×) to coordinate with neighboring cells, but this is not a requirement.

The second disaggregation point is the F1 interface (3GPP TS38.47x), which separates the base station functionality into non-real-time components at the CU and real-time components at the DU. This option is described in more detail in Section 4.1. In the context of a 5G network this means the latency between the CU and DU can be of the order of 10 ms (SCF 049). The example given for this split is an outdoor small cell network with the small cells used to add extra capacity in a location where the network was previously congested. The small cells are shown with fiber connections back to a central 4G or 5G macro cell, although this could be a standalone CU for a small cell network. The attractiveness of the F1 interface is that it enables joint SON and resource management across multiple small cells, reducing interference and improving system throughput, while its bandwidth

requirements are similar to that of the NG interface. F1 does require interoperability between the CU and DU, or macro and small cell; however, this barrier is lowered in 5G by 3GPP standardizing this interface.

The third disaggregation point is termed split option 6 in 3GPP terminology (3GPP TR 38.801).[14] For LTE this interface was specified by the Small Cell Forum as nFAPI (SCF 082); for 5G the SCF has announced an intention to specify this interface as 5G nFAPI and providing an overview of this activity is a key focus for this chapter. The example given for the nFAPI split is an indoor small cell network within a hotel, the PHY and RF components are located in the DU, with the remainder of the base station at the central location. The attractiveness of the nFAPI split is that it introduces the benefits of centralized scheduling to deployments that have low-latency backhaul, while not increasing the backhaul bandwidth requirements compared with NG and F1. This type of backhaul situation is relatively common in enterprise deployments.

The fourth disaggregation point is termed split option 7 in 3GPP terminology (3GPP TR 38.801) and has been further refined and specified by the O-RAN Alliance as O-RAN Fronthaul (ORAN FH). This option is described in more detail in Section 4.5. The example given for the fronthaul split is an indoor small cell network within an office. The DU contains the lower PHY and RF components, with the remainder (including the upper PHY) location in the CU situated within a server room. This split is possible due to an assumption that a high bandwidth fronthaul connection is available between the CU and DU. This split opens up the possibility of joint transmission and reception between small cells, which, in particular, can help improve reliability for URLLC scenarios.

#### 4.6.4.2 Platform Architectures

With small cells having multiple disaggregation architectures they also have several different platform architectures. This section describes two types of small cell platform architectures, their pros and cons, and their suitability for different types of disaggregation architectures.

The traditional, and most common, hardware architecture for small cells is a system on a chip (SoC), where a SoC consists of several subsystem components specifically designed for the small cell. The other hardware architecture considered is a virtualized architecture, which is an extension of the 5G trend for virtualization to the gNB. These two architectures are shown in Figure 4.6.2.

The SoC architecture typically consists of a mix of different compute resources, hardware accelerators, internal memory, and I/O, where these components are sized for a specific set of small cell use cases. This means that a small cell designed for a residential deployment will consist of a different configuration than an outdoor small cell. It also means that the small cell is optimized for a specific disaggregation architecture, and probably won't support all split options.

While the dimensioning of the small cell components will vary between SoCs, there is much commonality in what these components are. Typically, there will be two different types of compute architecture: the first is general purpose compute (GPP) where the L3

14 Note that this option was discussed in 3GPP during the 5G study, but has not been standardized in 3GPP.

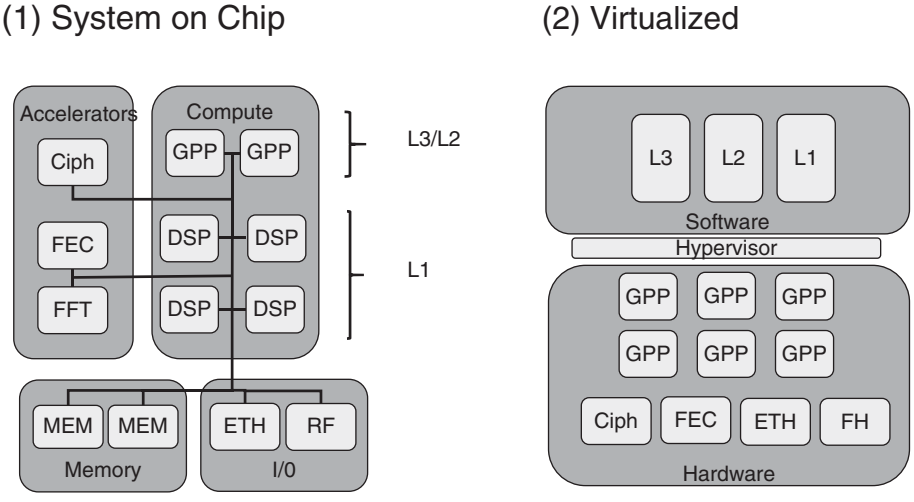



**Figure 4.6.2** Small cell platform architectures.

(RRC) and L2 (PDCP, RLC, and MAC) functionality resides; and the second is digital signal processor (DSP) where the L1 (PHY) functions are performed. The GPP and DSP within small cell SoCs are chosen to be low-power options with just enough processing power for the task. There are also tasks within a small cell that are typically implemented with hardware accelerators as they are too compute intensive for either the GPP or DSP. This group of tasks is decreasing over time, with the most likely candidates for acceleration being the ciphering in the PDCP layer and the channel coding, also called forward error correction coding (FEC) in the PHY. Finally, the SoC will include I/O capability with Ethernet for backhaul and a different bus protocol for the fronthaul. It should be noted that this is only a high-level description of the key components – a SoC may contain additional components and I/O interfaces.

A virtualized architecture consists of a software layer for the gNB RRC/PDCP/RLC/MAC and PHY functionalities, a set of hardware resources including GPP, hardware accelerators and I/O, and a hypervisor to manage the interface between software and hardware. Virtualization is being adopted in 5G networks to improve manageability and automation of the network, and for network deployments with some centralization, to reduce hardware requirements by pooling multiple cells onto the same hardware.

A virtualized small cell typically doesn't include DSPs, with these PHY functions instead using GPP compute. However, the rest of the hardware components, such as GPP, hardware accelerators, and I/O interfaces are similar to a SoC. What is different between the hardware components is the processing capabilities of each component. Virtualized small cells are most suitable where at least some functions are centralized so each platform is running several small cells. This means the choice of hardware components focuses on flexibility and scalability.

The pros and cons of SoC and virtualized platform architectures are shown in Table 4.6.1 for each disaggregation option. When considering the pros and cons it is assumed that the CU is managing multiple cells and the DU hosts a single small cell.

**Table 4.6.1** Pros and cons of disaggregation points.

| Disaggregation | System on a chip (SoC) small cell | Virtualized small cell |
|---|---|---|
| NG | Pros<br>– Can be efficiently designed to operate as one cell<br>Cons<br>– Can lack flexibility if multi-access edge compute (MEC) is present in small cell | Pros<br>– Virtualization can help if small cell also includes MEC<br>Cons<br>– Difficult to compete with cost and power against specially designed SoC |
| F1 | Pros<br>– Distributed unit can be efficiently designed to operate as one cell<br>Cons<br>– Central unit would be virtualized system | Pros<br>– Scalable at central unit to support multiple centralized cells and exploit pooling gains<br>Cons<br>– At distributed unit difficult to compete with cost and power against specially designed SoC |
| Network FAPI (nFAPI) | Pros<br>– Distributed unit can be efficiently designed to operate as one cell<br>Cons<br>– Central unit not able to scale or benefit from pooling gains | Pros<br>– Central unit scalable to support multiple centralized cells and exploit pooling gains<br>Cons<br>– At distributed unit difficult to compete with cost and power against specially designed SoC |
| O-RAN Fronthaul (ORAN FH) | Cons<br>– Central unit not able to scale or benefit from pooling gains | Pros<br>– Central unit scalable to support multiple centralized cells and exploit pooling gains |

#### 4.6.4.3 Operating Frequency Impacts on Architecture

Previous mobile technologies – 2G, 3G, and 4G – have focused on similar frequency ranges typically referred to as the sub-6 GHz spectrum, whereas 5G additionally extends into new frequency ranges for mobile networks, referred to as mmWave. For 2G to 5G this spectrum is predominately licensed, meaning an operator has exclusive access, but small cells are sometimes also deployed using the same unlicensed frequencies as Wi-Fi. Each of these frequency options results in different specifications and architectures for a small cell.

Below 6 GHz a 5G small cell can have the same functions and features as a macro cell, with the only difference being a reduced transmit power. However, in real deployments small cells typically have a reduced set of features compared with a macro cell, where this reduction is due to a several factors:

- The reduced cell coverage area results in a smaller number of users. The number of users per subframe or slot that a cell supports is frequently reduced. This reduces the small cell processing requirements.
- A small cell's enclosure has a smaller footprint, and it is installed at a lower height, for example, on a lamppost. Therefore, a large number of antennas is not desirable. This reduces the maximum throughput and small cell processing requirements.

**Table 4.6.1** Typical cell dimensioning.

| Parameter | Macro cell | Small cell <6 GHz | Small cell mmWave |
|---|---|---|---|
| Channel bandwidth | 3 sectors of 100 MHz | 100 MHz | N × 100 MHz |
| Users/transmission time interval (TTI) | 16 | 8 | 8 |
| Number layers | 8 downlink, 4 uplink | 4 downlink, 2 uplink | 4 downlink, 4 uplink |
| Number antennas | 32 | 4 | 64 |

- Small cells often use a smaller channel bandwidth than macro cells. For example, an operator may allocate a specific portion of their frequency spectrum to indoor coverage. This also reduces the maximum throughput and small cell processing requirements.

While at sub-6 GHz frequencies a mobile operator can choose the size of the cell they wish to deploy, at the new mmWave frequencies the cell will always be small. The propagation characteristics of mmWave frequencies limit the size of the cell up to 100 m. In order to maximize the cell size multiple antennas, also called massive MIMO, are used and since antenna size is directly related to transmission frequency these antennas are small enough to be deployed with a small cell sited at street level.

The final dimension for frequency spectrum is whether the spectrum is licensed or unlicensed. Licensed is the standard operating model with the spectrum available for the exclusive use of the mobile operator. This means the operator chooses where specific frequencies are used and the interference levels within the cell will be dependent on the cell planning performed by the network operator. The alternate is unlicensed spectrum, typically at either 2.4 GHz or 5 GHz, where the spectrum can be used by multiple technologies (e.g. Wi-Fi and Bluetooth) or multiple instances of the same technology. Unlicensed access was first used in 4G with two different modes – unlicensed (LTE-U) where a cell operates entirely in unlicensed spectrum and licensed assisted access (LAA) where CA is used to combined licensed and unlicensed frequencies. It is also important to mention shared spectrum, for example Citizens Broadband Radio Service (CBRS), which was first introduced in 4G, where a central database is accessed to determine if a cell can use a specific frequency based on its location. For 5G the same options of unlicensed, LAA, and shared spectrum will be available, and their use is expected to grow. These unlicensed methods are all used predominately with small cells.

Table 4.6.2 gives some typical examples of cell dimensions for a sub-6 GHz macro cell, a sub-6 GHz small cell, and an mmWave small cell. The use of licensed or unlicensed spectrum typically doesn't impact the cell dimensioning so is not included.

#### 4.6.4.4 Operational Models

The final variant for small cells is the relationship between a mobile network operator (MNO) and the small cell. For macro cells this relationship is clear – a macro cell is owned and operated by an MNO. However, for small cells with their wide variety of deployment scenarios this relationship is not so clear cut.

The most common form of small cell is the same as macro cells, with the MNO both owning and operating the small cell. Alternate options are a private network and neutral host architectures.

Private networks are where the small cell is owned and operated by someone other than a mobile operator. Typical scenarios might be an indoor network inside a factory, or an outdoor setup to provide a service at a remote outdoor location, such as a mining facility. The purpose of this type of network is to provide a service to a specific set of UEs for operational reasons. Therefore, in this type of network only specific permitted UEs will be able to connect, with no roaming from other networks. It is worth noting that the network slicing mechanism introduced in 5G is another method of operating a private network, where a mobile operator is still involved.

Neutral host is a more common and growing sector for small cell deployments and is where one small cell is used to provide a service for multiple mobile networks. This segment has developed from both the challenges of providing indoor 5G coverage and lack of suitable site locations from small cells. The neutral host small cell may operate one cell providing a service for all mobile networks or operate multiple cells, one for each mobile operator. Neutral host is popular in scenarios where a business benefits from ensuring its customers have a good mobile signal; for example, a hotel or shopping center.

### 4.6.5 Key Interfaces for Small Cells

The diversity and variety of small cell architectures and deployments have resulted in an ecosystem dominated by relatively small companies. The size of these companies reflects the size of small cell deployments, in that there are many small cell network deployment opportunities (in particular, indoor opportunities), but that each of these opportunities is relatively small; for example, a hotel chain, a single factory, office locations of one company.

These small cell vendors will procure the constituent parts of the small cell from several companies – in particular, there is a small cell ecosystem of protocol stack vendors (L3/L2), an ecosystem of PHY vendors (L1), and an ecosystem of RF vendors. This has resulted in the small cell industry adopting APIs to supporting integration of these different components.

This chapter focuses on one of these APIs, nFAPI, which defines the MAC/PHY split (split option 6) identified by 3GPP (3GPP TR38.801) but not standardized. It is important to note that nFAPI is a wrapper around the FAPI interface described next.

#### 4.6.5.1 FAPI

The FAPI interface, which originally meant Femtocell Application Programming Interface, is now simply called FAPI and defines the interface between MAC (L2) and PHY (L1), as the term femtocell was replaced by small cell.

FAPI is specified by the Small Cell Forum and the original interface defined the interaction between a 3G MAC and PHY (SCF 048). Subsequently a version was defined between a 4G MAC and PHY (SCF 082), and recently the 5G version has been completed (SCF 222). For each wireless technology FAPI defines a set of control messages for configuring a PHY, a set of messages for exchanging data between MAC and PHY, and indicates the timing constraints of these messages.

The architecture of FAPI is shown in Figure 4.6.3 showing the high-level FAPI principles:

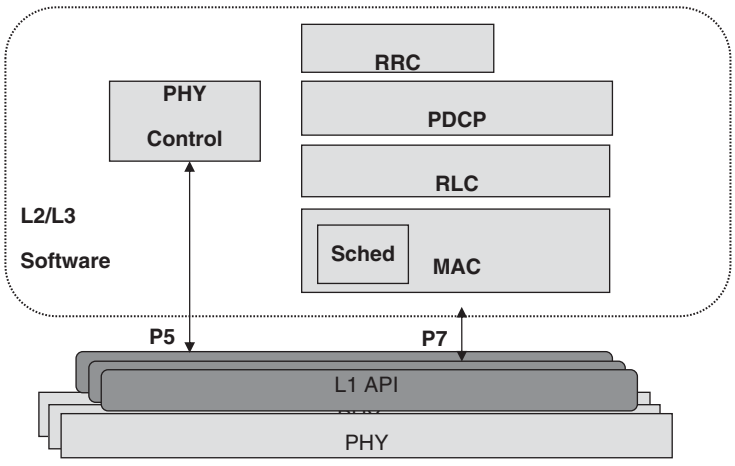



**Figure 4.6.3** FAPI architecture. (*Source:* Reproduced by permission of © Small Cell Forum).

- FAPI is defined between one instance of MAC and one instance of PHY.
- In scenarios such as CA there will be multiple instances of FAPI, one per carrier.
- The PHY configuration messages are defined by the logical interface P5. This configuration is semi-static and is being generated by a PHY control entity.
- The PHY data-plane messages are defined by the logical interface P7. This configuration is generated once per slot and consists of:
  - Slot configuration messages, which include information required by the PHY to encode and decode data.
  - Downlink data messages to transfer MAC PDUs to the PHY.
  - Uplink data messages to transfer MAC PDUs, Uplink Control Information (UCI), Sounding Reference Signal (SRS), and Random Access Channel (RACH) PDUs from the PHY.

To support consistent PHY behavior, FAPI defines a state machine for the PHY and this is shown in Figure 4.6.4:

- A PHY starts in IDLE mode, where the MAC can query the PHY's capabilities using the PARAM message sequence.
- A PHY can be moved from IDLE to CONFIGURED state by the MAC initiating the CONFIG message sequence.
- A PHY in CONFIGURED state can be configured further, or have its PHY capabilities queried.
- A PHY can be moved from CONFIGURED to RUNNING state by the MAC initiating the START message sequence.
- A PHY in RUNNING state is transmitting over the air and provides a small cell service.
- To stop the small cell the MAC must initiate the STOP message sequence.

Once a PHY is operational the interaction between the MAC and PHY occurs at the periodicity of a slot. Since 5G supports several different subcarrier spacing values these per-slot messages can be exchanged every 125 μs, 250 μs, 500 μs, or 1 ms. In fact, if a carrier

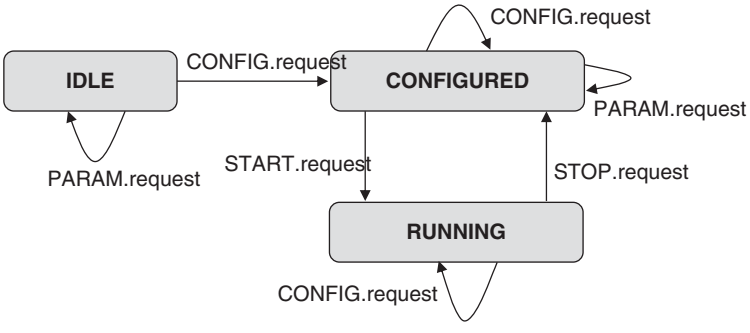



**Figure 4.6.4** FAPI state machine. (*Source:* Reproduced by permission of © Small Cell Forum).

is supporting several different bandwidth parts (BWP), with different subcarrier spacing, then the MAC may be exchanging FAPI messages with the PHY at different periodicities for different BWP.

FAPI reduces this complexity by providing all per-slot control messages in a single downlink control message, DL_TTI.request, and all downlink MAC PDUs in a single downlink data message, TX_Data.request. Similarly, for the uplink there is a single uplink control message, UL_TTI.request, and all similar uplink is grouped together into one message, RX_Data.indication, UCI.indication, SRS.indication, and RACH.indication.

The FAPI procedure to transmit data on the DL-SCH is shown in Figure 4.6.5:

- The MAC sends a DL_TTI.request message to the PHY. In this message there will be one PDSCH PDU for each UE with data in slot N, where the PDSCH PDU provides configuration information instructing the PHY how the MAC PDU, received in TX_Data.request, will be encoded. For two codeword transmissions then only one PDSCH PDU is included, containing the modulation and coding scheme of both codewords. The DL_TTI.request also includes PDCCH PDUs, which specify Downlink Control Information (DCI) sent to the UE to indicate data are scheduled. It should be noted that not every DL-SCH transmission requires a DCI; for example, data using configured grant scheduling.
- The MAC sends a TX_Data.request message to the PHY. In this message there will be one MAC PDU for PDSCH PDU with one codeword, and two MAC PDUs for PDSCH PDU with two codewords. The PDUs in TX_Data.request and DL_TTI.request contain an identifier, included in both messages, used to link the MAC PDU and PDSCH PDU together. The structure of TX_Data.request is flexible to support different underlying small cell architectures by allowing the MAC PDU data to be referenced by a pointer to memory, a list of pointers to memory locations, or the actual data.
- The PHY encodes the MAC PDUs received in TX_Data.request using the control information provided in DL_TTI.request and transmits the resultant data over the air to the UE.
- The UE will return an ACK/NACK message indicating whether it correctly received the data in slot N + K, where for 5G the value K is flexible and can be different for each UE. The MAC schedules the ACK/NACK reception by sending an UL_TTI.request message to the PHY. If the UE is transmitting uplink data then the ACK/NACK is indicated in a PUSCH PDU, and if there are no uplink data a PUCCH PDU is used.

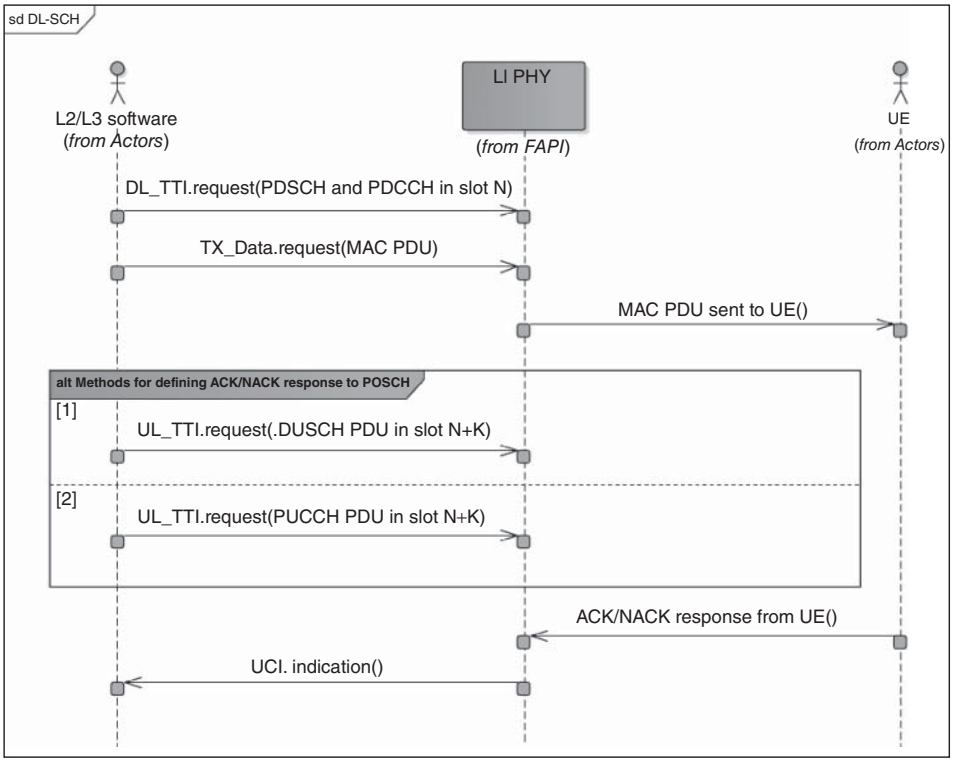



**Figure 4.6.5** FAPI downlink data procedure. (*Source:* Reproduced by permission of © Small Cell Forum).

- The PHY uses this information to locate and decode the ACK/NACK received from the UE. The received ACK/NACK is sent to the MAC in the UCI.indication message.

The FAPI procedure to receive data on the UL-SCH is shown in Figure 4.6.6. Assuming the uplink transmission needs to be scheduled the procedure starts with downlink messages.

- The MAC sends a UL_DCI.request message to the PHY. In this message there will be PDCCH PDUs that specify DCIs sent to the UE to indicate data are scheduled in the uplink. It should be noted that not every UL-SCH transmission requires a DCI; for example, configured grant.
- The UE will transmit the uplink data in slot N + K and when that slot arrives the MAC schedules the UL-SCH reception by sending an UL_TTI.request message to the PHY including a PUSCH PDU for each UE scheduled to transmit.
- The PHY uses this information to locate and decode the MAC PDU received from the UE. The received data are sent to the MAC in the RX_Data.indication message and the success or failure of the decoding is sent to the MAC in the CRC.indication message. This CRC information is separated from the MAC PDU since it gives the opportunity for the PHY to send the CRC.indication before the RX_Data.indication, which can help the MAC when performing uplink scheduling.

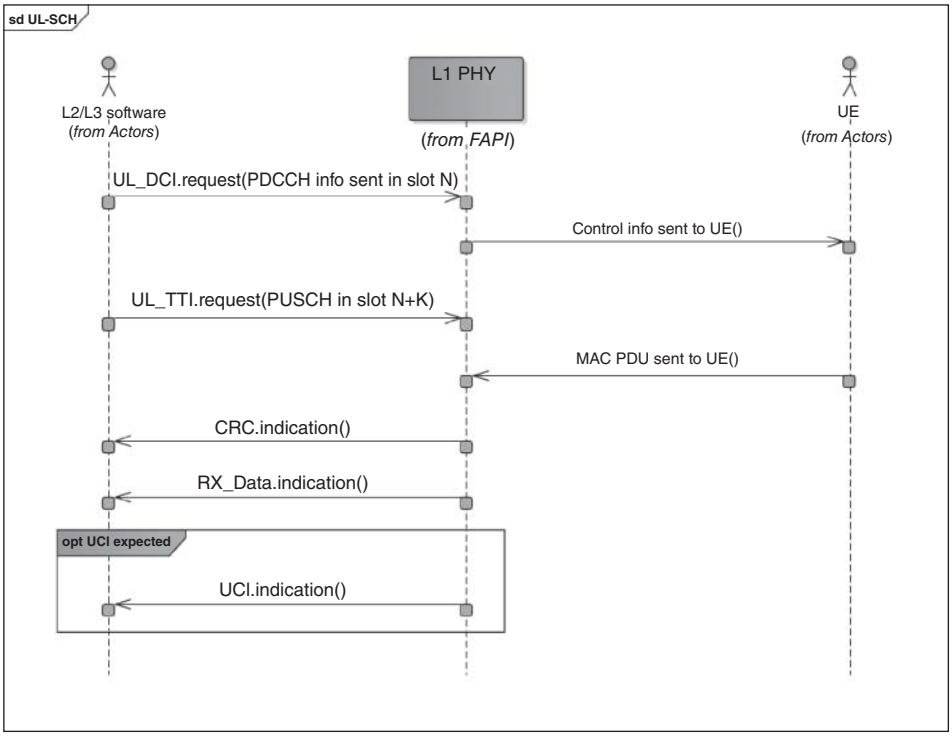



**Figure 4.6.6** FAPI uplink data procedure. (*Source:* Reproduced by permission of © Small Cell Forum).

- If the UE was also due to transmit uplink control information then the PHY will also locate and decode this information (based on control received in the UL_TTI.request PUSCH PDU). This uplink control information is sent to the MAC by an UCI.indication message.

FAPI has additional procedures to transmit and receive all logical channels; BCH, PCH, and RACH. These are primarily variations of the downlink and uplink data procedures, so not detailed here.

#### 4.6.5.2 nFAPI

Network FAPI (nFAPI) (SCF 082) is an extension introduced by the Small Cell Forum to support a disaggregated small cell eNB for 4G. The MAC and above protocol layers are located in a CU, while the PHY and RF are situated in a DU.[15] nFAPI assumes the connection between the CU and DU is non-ideal, where non-ideal means that there is a limitation in either latency, or throughput capabilities of the link between CU and DU. An example of a non-ideal connection that is particularly relevant to the small cell industry is Ethernet, which is a prime candidate for providing connectivity within indoor enterprise small cell

15 Note that CU and DU network nodes described in the present chapter are different from gNB-CU and gNB-DU nodes defined by 3GPP to support split option 7.

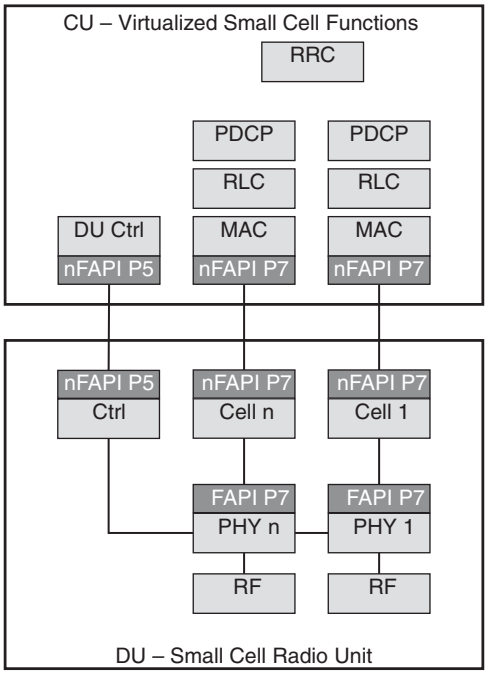



**Figure 4.6.7** nFAPI architecture. (*Source:* Reproduced by permission of © Small Cell Forum).

networks. The Small Cell Forum is in the process of building on 5G FAPI to develop 5G nFAPI supporting the 3GPP split option 6 interface (3GPP TR 38.801); this section highlights concepts from 4G nFAPI to capture the latest views on 5G nFAPI.

The relationship between nFAPI and FAPI is shown in Figure 4.6.7 and can be described such that an nFAPI interface provides a wrapper around FAPI to enable support for a non-ideal backhaul and multiple carriers (PHY) in one DU. This extension defines one instance of nFAPI per DU and supports scenarios where a DU may implement CA or is a neutral host small cell.

The top-level architecture principles of nFAPI can be summarized as:

- nFAPI is a wrapper around FAPI.
- nFAPI is defined between one instance of CU and one instance of DU.
  - nFAPI will have one instance of P5 (control).
  - nFAPI can have more than one instance of P7 – one for each PHY in the DU.
- Additional nFAPI P5 procedures are defined to support configuration and control of the DU.
- Additional nFAPI P7 procedures are defined to support a non-ideal connection between CU and DU.

Following the FAPI philosophy nFAPI defines a state machine for the DU to ensure consistent behavior and this is shown in Figure 4.6.8:

- A DU starts in IDLE mode, where the CU can query the DU's capabilities using the PARAM message sequence.
- A DU can be moved from IDLE to CONFIGURED state by the CU initiating the CONFIG message sequence.

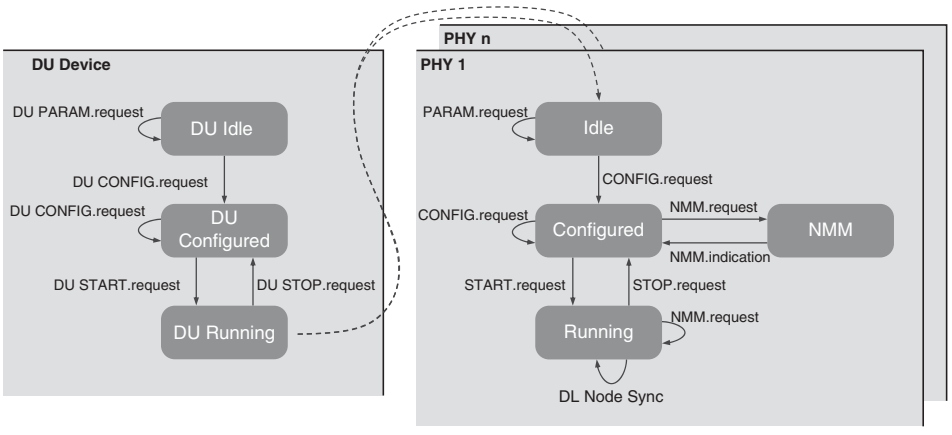



**Figure 4.6.8** nFAPI state machine. (*Source:* Reproduced by permission of © Small Cell Forum).

- A DU in CONFIGURED state can be configured further, or have its capabilities queried.
- A DU can be moved from CONFIGURED to RUNNING state by the DU initiating the START message sequence.
- A DU in running state instantiates a PHY in IDLE mode. FAPI P5 messages are used to query, configure, and start the PHY.
- Stopping a DU deletes any PHY which have been instantiated.

The nFAPI state machine in Figure 4.6.8 additionally includes a Network Monitor Mode (NMM) state, where NMM is unique to small cells and is designed to support self-configuration of the small cell. A small cell has an NMM for each technology it wants to monitor, so previously, 4G small cells had a 2G, 3G, and 4G NMM, while going forward 5G small cells will have 2G, 3G, 4G, and 5G NMM. The procedures in NMM can be summarized as:

- Measure the Received Signal Strength Indicator (RSSI) over a range of frequencies and channel bandwidths to identify potential neighbor cells.
- Search for neighbor cells by identifying their synchronization signals.
- Collect MIB information from a neighbor cell.
- Collect SIB information from a neighbor cell.

The hierarchal nature of the four procedures ensures an efficient way to build up detailed knowledge of its surrounding environment.

Finally, it is worth reiterating that nFAPI is a wrapper for FAPI, so in order to transmit downlink and uplink data, the procedures in Figures 4.6.5 and 4.6.6 are still followed.

#### 4.6.5.3 Management Plane

Within a mobile network every cell requires a management plane to ensure that the operator can configure a cell and monitor relevant key performance indicators. Even though 3GPP have defined OAM specifications, historically these management planes have been proprietary. Small cell networks still require a management plane but there have been efforts to provide a standardized interface used across small cell vendors.

In 4G this management plane for small cells was defined by the Broadband Forum (BBF TR-196). The motivation of involvement from the Broadband Forum was that the first small cells were residential cells, allowing the industry to develop a new data model to support the 4G-specific parameters, but also leverage the Broadband Forum data models for home broadband. The reused models focus on functions that support transmission across the non-ideal backhaul where available uplink backhaul bandwidth was often the system bottleneck. One aspect of 5G is fixed-mobile convergence, which has led the Broadband Forum to continue to look at developing specifications in this area.

For 5G management planes another configuration option is also being pursued, namely NETCONF/YANG (IETF RFC6241) (IETF RFC7950). 3GPP SA5 specifications for 5G rely on it and it is being promoted by the O-RAN Alliance as an open interface for the management plane for all 5G cells (ORAN-MP) (ORAN Yang). The O-RAN Alliance (see Section 4.5) has created standard interfaces for any 5G system, not just small cells, so the adoption of NETCONF/Yang for small cells would aid the integration of small cell and macro cell networks.

### 4.6.6 Worked Examples

So far, we have shown that small cells will be a critical network element to deliver 5G, with many of the 5G expansion use cases, such as new frequency ranges and new vertical industries, particularly suited to small cells. This results in a somewhat confusing selection of disaggregation, hardware architectures, and frequency options. We now take several examples of specific small cell deployments and identify how a small cell could be deployed. Other sources of information to locate specific small cell deployments are "NGMN small-cells") and "SCF deployment stories."

#### 4.6.6.1 Indoor Enterprise Example

The first example small cell deployment scenario is indoor enterprise, which is considered a key market for small cells. It is indeed viewed as the biggest opportunity due to the importance of improving indoor coverage (SCF 050).

An example of this scenario is shown in Figure 4.6.9 and is based around an indoor network within a corporate office.

Figure 4.6.9 highlights location, disaggregation, and platform choices suitable for this scenario and, while other combinations are possible, to provide an easy-to-understand example only one is selected.

*Location* – In this example the indoor network requires space in a server room and locations around the office to install DUs with the connection between these locations provided by a dedicated Ethernet connection. The option of space in a server room, and the desire to keep the DUs small help influence the architectural options for this network.

*Disaggregation* – With a central location available many of the small cell functions can be centralized. This has the advantage of enabling centralized SON and scheduling functions and potentially CoMP where transmissions are received from several DUs and joined in the upper PHY. The split between the CU and DU will follow the ORAN FH.

*Platform* – With much of the small cell centralized this scenario is ideal for a pool of virtualized small cells running on off-the-shelf server hardware. The RRU will be a low-cost SoC design.

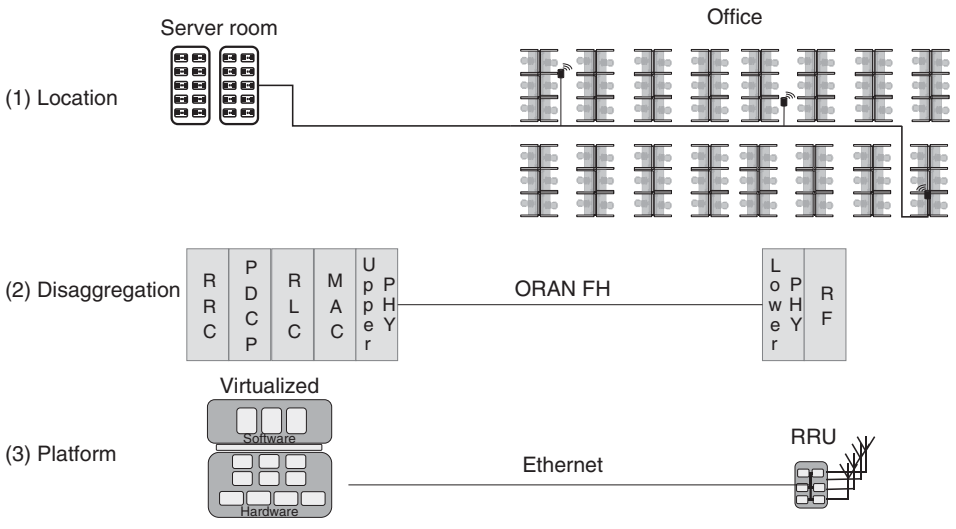



**Figure 4.6.9** Indoor enterprise scenario.

*Frequency* – The small cell is operating in the sub-6 GHz range on licensed frequencies.

*Operational mode* – Although the small cell network was implemented to benefit employees it is still likely to be a public network, meaning that visitors will also be able to attach and gain a service. If the corporate company already has a business relationship with one mobile operator, then this may be a small cell network serving just one operator. Alternatively, it might be neutral host to serve both the work and private smartphones of all employees regardless of their network.

#### 4.6.6.2 Outdoor Urban Example

The second example of a small cell deployment scenario is outdoor urban and is included since this is also a key market for small cells. Specifically, it is viewed as the biggest growth area for small cells (SCF 050) due to the expected 5G trend for ultra-dense networks.

An example of this scenario is shown in Figure 4.6.10 and is based around an outdoor network within a city.

Figure 4.6.10 highlights location, disaggregation, and platform choices suitable for this scenario and, while other combinations are possible, to provide an easy-to-understand example only one is selected.

*Location* – In this example the outdoor network requires space in a point of presence (PoP), where a PoP is a location within the mobile network where an operator installs equipment. For example, this could be an aggregation point bringing multiple base stations' data together, and in an urban area should be no more than 5–20 km from a base station. This outdoor network will require sites for each small cell, for example on street-lighting, and the connection between the CU and DU is likely to be fiber.

*Disaggregation* – The CU is up to 20 km from the DU, which means that due to latency the real-time portion of the small cell is likely to reside in the DU. By centralizing the RRC and PDCP, this will still improve handover and interference management in the small cell network. The split between the CU and DU will follow the 3GPP F1 interface.

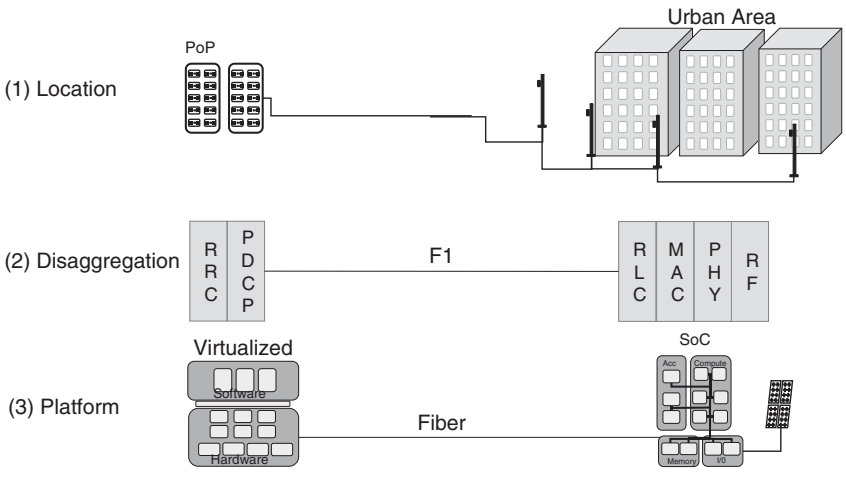



**Figure 4.6.10** Outdoor urban scenario.

*Platform* – With much of the small cell in the DU this scenario is ideal for a SoC at the small cell site. The small cell functions at the PoP will be virtualized.

*Frequency* – This example shows mmWave and the small cell operating on licensed frequencies. The choice of mmWave can allow the network to reserve its sub-6 GHz frequencies for the overlaying macro cell network.

*Operational mode* – This type of outdoor network is expected to be public and part of a HetNet solution for one mobile operator. Each mobile operator will have their own outdoor urban small cell network.

#### 4.6.6.3 Private Network Example

The third example of a small cell deployment scenario is a private network and is included since this is a new vertical market for small cells.

An example of this scenario is shown in Figure 4.6.11 and is based around a private indoor enterprise network within a factory, which is often referred to as a smart factory.

Figure 4.6.11 highlights location, disaggregation, and platform choices suitable for this scenario and, while other combinations are possible, to provide an easy-to-understand example only one is selected.

*Location* – In this example the private network requires space in a server room and locations around the factory to install DUs with the connection between these locations provided by a dedicated Ethernet network. This is a similar requirement to the indoor enterprise example and is typical of many indoor scenarios where there is a server/equipment room available and a desire to keep the footprint at the cell site small.

*Disaggregation* – With a central location available many of the small cell functions can be centralized. However, for a private network the centralized location contains some distributed EPC (dEPC) functionality and for a factory MEC is also likely. The combination of dEPC and MEC permits the factory to keep some, or all, of its data within the factory, which is beneficial from both a latency and privacy perspective.

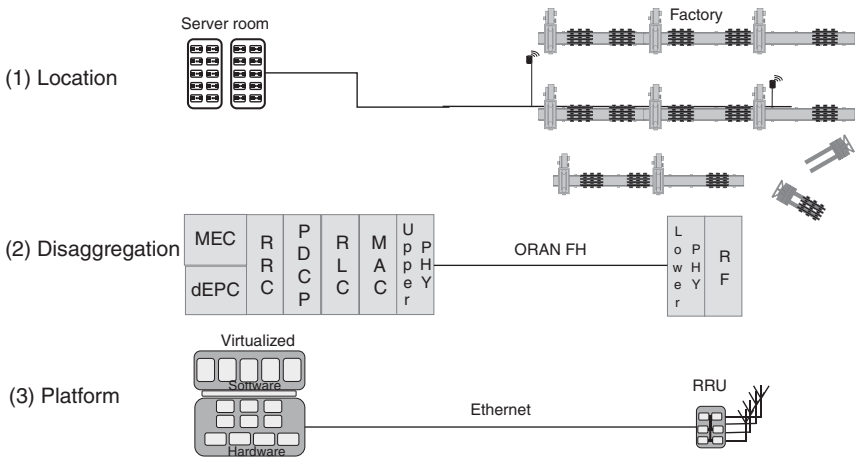



**Figure 4.6.11** Private enterprise scenario.

*Platform* – With much of the small cell centralized this scenario is ideal for a pool of virtualized small cells running on off-the-shelf server hardware, and the dEPC and MEC can also run on the same server hardware. The RRU will be a low-cost SoC design.

*Frequency* – In this example the small cell is operating in the sub-6 GHz range, but it could be licensed or unlicensed frequencies. The factory may have a relationship with an operator allowing it to use licensed spectrum, could be in a region where spectrum is being reserved for these types of vertical industries, or could use unlicensed frequencies.

*Operational mode* – This example is for a private network. There may still be a relationship with a network operator for some EPC functionality, or the network could be self-contained within the factory.

### 4.6.7 Further Reading

The most abundant source of information on small cells is the Small Cell Forum, which has produced numerous documents covering all types of small cell deployments and architectures as part of its release program. To date it has published over 125 documents on small cells, including its release 10 set of documents dedicated to 5G. The SCF references used in this chapter can be broadly split into two groups: technical specs relating to FAPI interfaces; and SCF 048, SCF 082, SCF 222, SCF 223, and reports designed to share knowledge within the small cell ecosystem and remove deployment barriers. From this second set of documents the most notable is SCF 050, which is a periodically updated report based on interviews with network operators and service providers, which identifies the top motivations to adopt small cells and the key barriers.

## References

3GPP Technical Specification 36.932 (2018). Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.401 (2019). NG-RAN; Architecture description. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.410 (2019).NG-RAN; NG general aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020)
3GPP Technical Specification 38.470 (2019).NG-RAN; F1 general aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020)
3GPP Technical Specification 38.801 (2019).Study on new radio access technology: Radio access architecture and interfaces. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.420 (2019).Xn general aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020).
BBF TR-196 (November 2011). Femto Access Point Service Data Model. TR-196, 2.
ETSI-MEC (June 2018). MEC in 5G networks. ETSI White Paper, 28.
Global 5G Study (December 2017). Study on small cells and dense cellular networks regulatory issues. Available at: https://www.global5g.org/small-cells (accessed May 29, 2020).
IETF-6241 (June 2011). Network Configuration Protocol (NETCONF). RFC 6241.
IETF-7950 (August 2016). The YANG 1.1 Data Modeling Language. RFC7950.
NGMN-Small-cells (September 2015). Recommendations for small cell development and deployment. NGMN Alliance.
ORAN FH (March2019). control, user and synchronization plane specification. O-RAN Fronthaul Working Group.
ORAN-M (March 2019). Management plane specification. O-RAN Fronthaul Working Group.
ORAN Yang (March 2019). Yang models. O-RAN Fronthaul Working Group.
SCF 048 (December 2013). 3G SCAPI. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF 049 (February 2013). Backhaul technologies for small cells Use cases, requirements and solutions. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF 050(December2018). Small cells market status report. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF 082 (February 2017). FAPI and nFAPI specifications. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF 214 (February 2018). Making buildings small cell ready. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF 222 (June 2019). 5G FAPI: PHY API Specification. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF 223 (2019). 5G FAPI: Frontend API Specification. Available at: http://scf.io/en/index.php (accessed May 29, 2020).
SCF (2018). Deployment stories. Available at: https://www.smallcellforum.org/small-cell/deployment-stories (accessed May 29, 2020).

## 4.7 Summary

In this chapter we discuss various NG-RAN architecture defined in 3GPP, O-RAN, and Small Cell Forum. Most of the architectures deal with various options of splitting gNB functionality into a number of separate logical network nodes, as opposed to a monolithic

architecture with a single network node hosting all gNB functionalities. One important exception is the MR-DC architecture, which splits the functionality across two base stations (which can be 5G gNBs or LTE eNBs) and which can also function independently (e.g. for UEs that do not support MR-DC).

These split architectures offer numerous benefits in terms of deployment flexibility, independent upgrades, and scaling of network components (only where scaling is needed). Furthermore, split deployments are typically easier to virtualize.

An important non-technical benefit of split architectures is that (if defined, implemented, and tested properly) they may allow network operators to source different network nodes from different vendors, thus increasing competition and potentially driving cost reductions. One must remember though that mixing and matching equipment from different vendors shifts the burden of interoperability testing and troubleshooting from a vendor to an operator, or an integrator employed by an operator for that task.

Another important point to keep in mind is that, while most network vendors do follow architecture standards to some extent, many chose to add proprietary features and some even develop non-standardized architectures. One such example (not covered in this book, as we focus on standardized architectures) is the Distributed Antenna System (DAS).

Another interesting trend we are observing is that some NG-RAN architectures that have been initially defined for 5G (e.g. gNB-CU/DU split and gNB control–user plane separation) are being "backported" to LTE.

This chapter is dedicated to NG-RAN architectures defined by 3GPP in Release-15 and by O-RAN and Small Cell Forum around the same time frame. In the next chapter we discuss how NG-RAN evolves in Release-16 and how it is likely to evolve further in Release-17 and beyond.

# 5

# NG-RAN Evolution

## 5.1 Introduction

In the previous chapter we discussed various NG-RAN architectures, defined in 3GPP, other standards bodies, and industry fora. The architectures described so far are either part of 3GPP Release-15 specifications or targeting 3GPP Release-15 networks.

Release-15, being the first 5G release by 3GPP, introduced the biggest changes into RAN architecture (compared with 4G); however, NG-RAN continues to evolve in Release-16 and beyond.

In the present chapter we describe some (but definitely not all) NG-RAN evolution paths, which will help expanding 5G networks into new market segments and new deployment options. In particular, we define two important NG-RAN architecture enhancements:

- Relays, also referred to as Integrated Access-Backhaul (IAB), which will help deploying 5G in areas where operators do not have sufficient backhaul capacity. This is especially beneficial for dense small-cell deployments and mmWave.
- Satellite (access and backhaul), also referred to as non-terrestrial networks, will mark an important milestone with yet another industry embracing 3GPP technologies as opposed to the proprietary technologies used in the past.

As mentioned above, relays and satellites are by no means the only NG-RAN evolution and expansion areas; however, these appear to be among the most important ones defined in Release-16.

## 5.2 Wireless Relaying in 5G

*Georg Hampel*

*Qualcomm Incorporated, US*

IAB introduces wireless relaying to 5G. While 3GPP has already specified relay solutions for 4G none of them enjoyed major commercial success. This is expected to change for 5G with the introduction of mmWave access, which depends on relaying for economic network deployment.

The wide bandwidth in the mmWave range provides abundant capacity, but the high propagation loss at these frequencies limits the practically achievable cell size requiring

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

densified network deployments to obtain sufficient area coverage. IAB aims to reduce the need for backhaul in such densified networks by allowing cells without fiber connection that wirelessly "self-backhaul" via neighbor nodes to the next fiber point. While wireless backhauling uses part of the mmWave capacity, the integration with access enables flexible and spectrally efficient resource utilization. IAB can also be applied for or in combination with sub-6 GHz frequencies. It is possible, for instance, to support sub-6 GHz access networks with mmWave-based backhauling.

The standardization of IAB provides interoperability and therefore allows integration of relays from different vendors into the same network. This lowers cost through commoditization, which is an important factor for a technology that relies on rollout in large quantities. Standardization further establishes test scenarios to ensure that performance guarantees can be met. However, one must remember that, even though IAB network interfaces are fully defined in 3GPP, it is unlikely that it would be possible to deploy IAB network nodes in a "plug-and-play," manner and a certain amount of integration and interoperability testing will still need to be performed by operators deploying an IAB network.

In this chapter, we describe the IAB architecture and elaborate on the design decisions made to accommodate multi-hop transport, Quality of Service (QoS), and spectrally efficient resource sharing between backhauling and access. We further highlight critical factors to be considered such as flow and congestion control, duplexing constraints, cross-link interference, and inter-node time synchronization. Finally, autonomous network integration of IAB nodes and dynamic topology changes are discussed.

### 5.2.1 Key Ideas

- IAB enables economic rollout of 5G mmWave access network. Due to flexible and spectrally efficient resource sharing between access and backhaul, it can substantially reduce the number of fiber connections without need for additional spectrum. Access and backhaul can also be operated in different bands. IAB is not limited to mmWave and can also be applied for or in combination with sub-6 GHz frequencies.
- IAB supports multi-hop backhauling to provide sufficient range extension in urban infrastructure with pronounced shadowing. It further supports topologically redundant backhauling for robustness and enables load balancing.
- Relaying is conducted on the radio link layer, which enables efficient processing and confines signaling to the RAN with minimum impact on the core network.
- IAB is compliant with all the relevant 5G deployment options. It leverages the existing NR air interface procedures (with appropriate extensions) to connect to 5GC in standalone mode and to connect to the Evolved Packet Core (EPC) via E-UTRA-NR dual connectivity (EN-DC). Release-15 user equipment (UE) can connect to the 3GPP network via IAB.
- Access and backhaul interfaces are synchronized at frame level to enable efficient time division multiplexed (TDM) resource sharing while minimizing the latency of multi-hop backhauling. IAB supports inter-node time synchronization via Global Navigation Satellite System (GNSS) and cellular over-the-air signaling.
- IAB support plug-and-play integration of network nodes (assuming sufficient interoperability testing has been performed), dynamic topology changes during operation, and autonomous recovery in response to link obstruction or node failure.

- IAB builds on the gNB central unit/distributed unit (CU/DU) split architecture (described in Section 4.2) with centralized topology, route, and resource management following software-defined networking (SDN) principles.
- IAB has been standardized by 3GPP in Release-16 and is being enhanced in Release-17.

### 5.2.2 Market Drivers

The mmWave bands provide abundant capacity, but the high propagation loss in this frequency range confines cells to small size so that densified deployments are necessary to achieve market-wide coverage. While small cells have been available for 3G and 4G radio access technologies, the densified small-cell network has only seen limited commercial success, in part due to the high backhauling cost. Moreover, 3G and 4G technologies operated at sub-6 GHz frequencies, which have more favorable propagation characteristics and allow deployment of heterogenous networks (HetNets). These HetNets combine macro cells for area coverage with small cells to meet capacity demand at hot spots (Figure 5.2.1). By adapting wireless capacity to the spatial traffic distribution, adequate service can be provided in an economic manner. It was possible to grow the network incrementally by restricting the initial rollout to a macro-cellular layer while adding capacity through cell split or small-cell fill-ins with increasing traffic demand. Incremental deployment therefore permitted market-wide service from day one while stretching deployment expenses over years.

For mmWave access, a densified small-cell network is necessary to provide coverage even where demand is low. Network densification implies high cost for backhaul connectivity. To match the large wireless access capacity, backhaul connections require high-capacity fiber (in some cases microwave backhaul can be used, but for mmWave access microwave backhaul capacity is unlikely to be sufficient), which exacerbates the economic burden. There is obviously no path to incrementally evolve the network in compliance with traffic growth.

Apart from high propagation loss, mmWaves also exhibit more prominent shadowing in the presence of obstructions such as cars and buildings due to diminished diffraction at these short wavelengths. The prominent shadowing can be addressed with deployment of more cells, which further increases network density and backhaul cost.

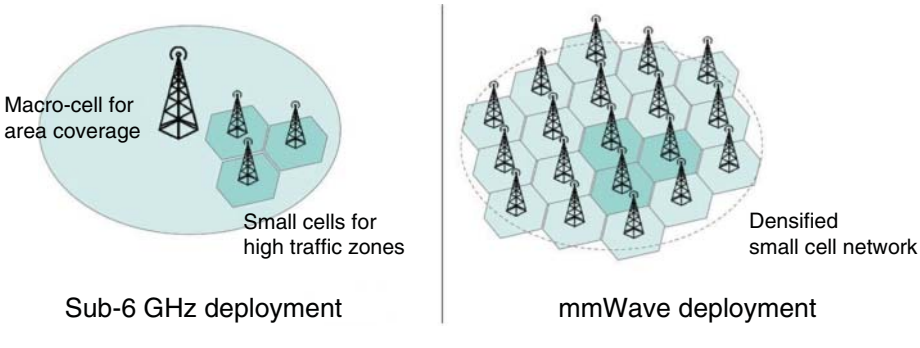



**Figure 5.2.1** Sub-6 GHz access can be deployed as HetNets while mmWave access requires a densified small-cell network.

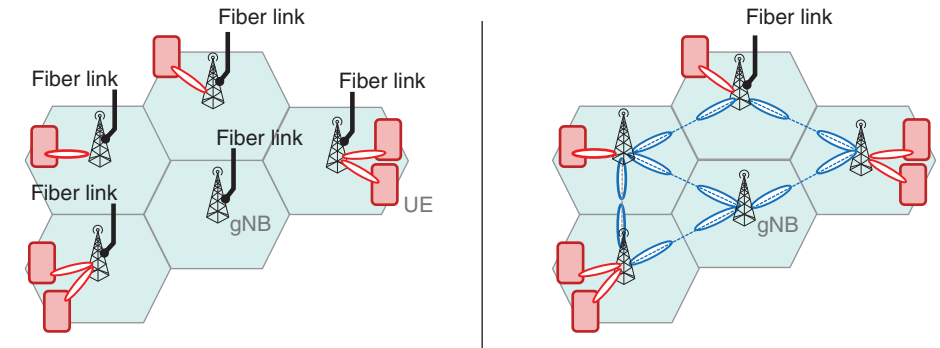



**Figure 5.2.2** Small-cell deployment with and without IAB: without IAB, a separate fiber link is needed for every cell. IAB allows reducing the number of fiber links through self-backhauling.

The economic constraints of rolling out 5G in the mmWave frequency range can be considerably lowered by also using the wireless access spectrum for backhauling and leveraging the mmWave-inherent advantages (Figure 5.2.2):

- *The abundant capacity* available at mmWave frequencies represents a significant benefit that can be leveraged to allocate capacity for backhauling.
- *Resource sharing between access and backhaul* adds deployment flexibility and provides an incremental deployment strategy. To achieve area coverage at small traffic load, as will apply in most cases during an initial rollout, only a few fiber points are necessary to support a large quantity of mmWave cells, and most of the mmWave spectrum will be used for backhauling. As traffic grows, additional fiber points can be added, which locally shifts the resource usage from backhauling toward access.
- *High gain beamforming* can be applied on both endpoints of the backhaul link, which suppresses inter-link interference creating "wires through the air." This allows extension of backhaul to multiple hops with efficient resource reuse.
- *Beam steering* – another unique feature developed for mmWave access – is leveraged on the backhaul links for autonomous neighbor detection and connection establishment. In this manner, plug-and-play deployment of mmWave small cells can be achieved.

Wireless backhauling can also be used for sub-6 GHz technologies. However, the benefits are less prominent due to smaller beamforming gain, which increases inter-link interference and therefore lowers the achievable spectral efficiency. Moreover, the need for multi-hop relaying is less critical than in the mmWave frequency range. Finally, mmWave backhauling can be paired with sub-6 GHz access; for example, to enable densified network deployments in the sub-6 GHz range while taking advantage of flexible and low-cost backhauling.

The economic necessity of wireless backhauling using 5G mmWaves was instrumental when 3GPP defined the scope of the initial IAB feature set in Release-16:

- Focus was set on stationary infrastructure relays. Support for relay mobility was considered less important and therefore left for later releases.

- Multi-hop backhauling and topological redundancy were considered essential for deployments in convoluted urban environments.
- Plug-and-play network integration and dynamic topology adaptation were included to lower deployment cost of a densified network and to ensure robustness to changes in propagation environment and load distribution.
- Compliance with the relevant 5G architecture deployment scenarios, in particular operation with EPC and 5GC, was considered critical. It was also desirable to support network connectivity via IAB for Release-15 UEs.
- Finally, efforts were made to retain high spectral efficiency through integration of access and backhaul links at the radio frame and enabling TDM operation among access and backhaul links.

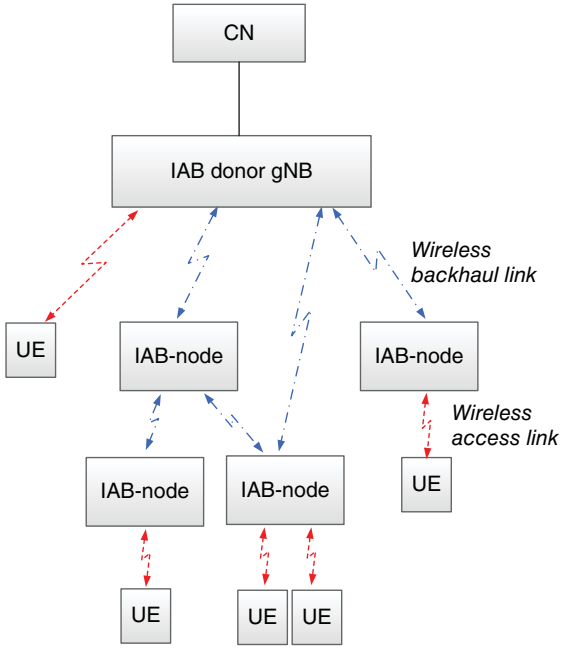
At the time of writing, the normative specification of Release-16 IAB had not been finalized. The following description may therefore lack some details or differ in some minor aspects with respect to the final Release-16 specification.

## 5.2.3 Functional Description

### 5.2.3.1 IAB Architecture

The IAB architecture introduces the *IAB node* as the 5G relay and the *IAB donor gNB* as the anchor point to the wireline network (Figure 5.2.3). The IAB node provides access to UEs, and it wirelessly backhauls the access traffic to/from the IAB donor gNB. It also forwards backhaul traffic via peer nodes. While typical deployments are expected to require only a few backhaul hops the architecture does not pose any limitation to the achievable hop count. The IAB donor holds gNB functionality and includes additional features for the

**Figure 5.2.3** IAB topology.



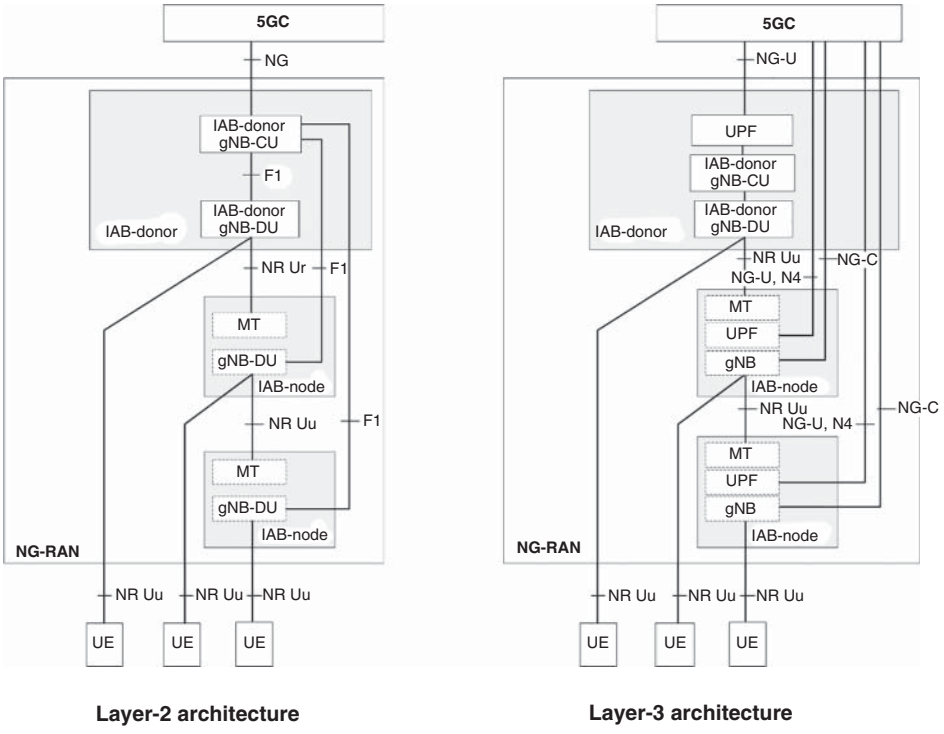



**Figure 5.2.4** Examples for layer 2 and layer 3 IAB architectures. 3GPP decided to use the layer 2 architecture (left).

support of IAB. Alternatively, the IAB donor gNB can be split into gNB-DU and gNB-CU, and the gNB-CU can be further split into gNB-CU-CP and gNB-CU-UP in compliance with Release-15 CU/DU split architecture. These split architectures are described in more detail in Sections 4.2 and 4.4. The IAB donor therefore has the same deployment flexibility as a conventional gNB.

3GPP considered multiple layer-2 and layer-3 relaying architectures and decided in favor of the layer-2 shown in Figure 5.2.4. Another architecture, which uses layer-3 relaying, is also shown in the figure and discussed further below.

In the layer-2 relaying architecture, the IAB node holds a gNB-DU and a *mobility termination* (MT) function. The IAB node gNB-DU interfaces to the gNB-CU on the IAB donor via the F1 interface (with relevant extensions), which was originally defined for the gNB CU/DU split architecture. The MT replicates UE functionality to establish connections to a gNB and the core network. In this manner, IAB node MTs can connect to gNB-DUs on other IAB nodes or on the IAB donor. The gNB-CU on the IAB donor further becomes the central control function for all IAB- node gNB-DUs and MTs defining an interconnected IAB topology.

The advantages of this architecture are:

- The IAB node is lightweight and low in complexity since it only supports a gNB-DU function.

- The IAB donor gNB-CU offers centralization of control and management tasks allowing performance optimization and reducing need for specification of distributed functionality.
- The RAN to core network interface is terminated at the IAB donor gNB and therefore shielded from the IAB nodes while control and management are retained to the RAN.
- IAB is compliant with conventional wireline CU/DU split architecture, where the gNB-CUs can be virtualized and moved into the cloud. It is therefore relatively easy to upgrade an IAB node to a regular gNB-DU by furnishing it with a fiber connection.
- The backhaul link reuses the existing NR-Uu interface (with relevant extensions) reducing specification effort and allowing easy integration of backhaul links with access links.

The dual personality of the IAB node, represented by gNB-DU and MT functionality, imposes a hierarchical structure onto the IAB topology, where the MT connects to an upstream *parent* node (which may be another IAB node or the IAB donor) while the gNB-DU connects to a downstream *child* IAB node. A spanning tree topology is formed when multiple IAB node MTs connect to the same parent node gNB-DU on multiple hops. The topology becomes a directed acyclic graph, when one or more IAB nodes also connect to multiple parent nodes. The restriction to hierarchical topologies was motivated by the nature of the traffic flow in a backhaul network, which moves between access nodes and IAB donor, and it makes loop-free signaling and transport easier to enforce.

IAB node DU and IAB donor CU together appear as a conventional gNB to the UE. In this manner, UEs can use standalone (5GC) mode or EN-DC (for details on EN-DC and more generalized MR-DC variant, see Chapter 3) to connect to the network. Since the IAB node MT uses Uu procedures, it can also use standalone mode and integrate with an 5GC, or EN-DC mode to integrate with an EPC. IAB nodes using EN-DC have to support an additional Long-Term Evolution (LTE) link to an eNB for control-plane signaling while keeping backhauling confined to NR (for details about multi-connectivity architectures, refer to Section 4.3). The operation mode of a UE and an IAB node MT are independent from each other. It is therefore possible to combine an EPC for UEs with a dedicated NGC for IAB nodes (Figure 5.2.5).

The layer-3 relaying architecture shown in Figure 5.2.4 is in line with one of the 4G relaying solutions (3GPP TS 36.300, clause 4.7). In this architecture, each backhaul link supports an IP connection via a self-contained Protocol Data Unit (PDU) session. For this purpose, the IAB donor and each IAB node have to hold a full gNB function and a user-plane function (UPF) in addition to the MT functionality. While this solution can reuse existing procedures for PDU session management, it requires core network signaling for every reconfiguration of the backhaul link. The amount of core network signaling was one of the main concerns. The layer-3 relaying architecture further lacks compatibility with wireline CU/DU split deployments. For these reasons, the layer-2 architecture was chosen.

One of the objectives during the IAB standardization introduced in Release-16 was to allow Release-15 UEs to connect to the network via IAB. For this reason, all access-related protocol layers were retained in the IAB protocol stack (i.e. protocol layers color-coded in white in Figure 5.2.6). This includes Radio Link Control (RLC) channels including PHY, Medium Access Control (MAC) and RLC sub-layers between a UE and an IAB node DU, which are extended via F1 connections between the IAB node DU and the IAB donor

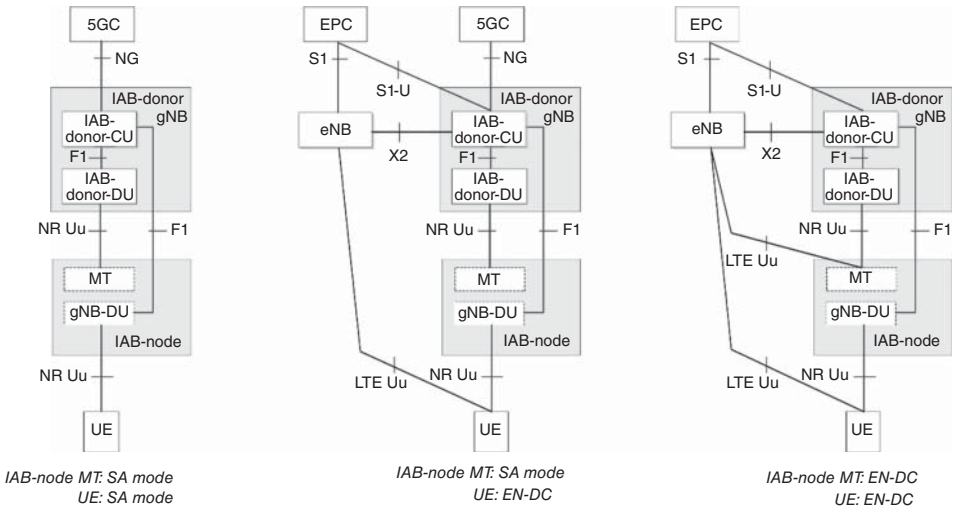



**Figure 5.2.5** Various options to integrate IAB with EPC and/or 5GC.

CU across the wireless multi-hop backhaul plane, and Packet Data Convergence Protocol (PDCP), Service Data Adaptation Protocol (SDAP), and radio resource control (RRC) layers between the UE and the IAB donor CU. For details about RRC, SDAP, PDCP, RLC, and MAC layers, please refer to Section 3.4.

Since an IAB node MT operates like a UE, it supports the same access radio channels to its parent node DU and the IAB donor CU. This includes the RRC connection with the IAB donor CU, and it may include one or more data radio bearers, for example, to carry PDU sessions for its own traffic (see the OAM discussion below). The protocol layers for the IAB node MT's access traffic are therefore the same as those for a UE, shown in Figure 5.2.6.

In addition to the RLC channels for access, the IAB node MT also supports RLC channels for backhauling. While these backhaul RLC channels provide full RLC functionality on each hop, for example, such as Automatic Repeat Request (ARQ) for RLC Acknowledged Mode (AM) operation mode, they are not mapped to a specific PDCP entity. Instead, they carry the Backhaul Adaptation Protocol (BAP) as an upper protocol layer to enable forwarding across the backhaul topology (the backhaul-related layers are shown as "BH RLC" in Figure 5.2.6). The BAP carries the native F1 protocol stack including IP (shown as "F1-C" in Figure 5.2.6), which is used for transport across wireless and wireline networks. The F1 interface can be security protected via a Network Domain Security framework (3GPP TS 33.210) in compliance with 5G security architecture (3GPP TS 33.501), which inserts an IPsec layer on top of IP.

#### 5.2.3.2 Backhaul Transport and QoS

Support for fine granular QoS and fairness on the wireless backhaul is critical since traffic aggregation may lead to congestion high up in the topology close to the IAB donor DUs. For QoS differentiation, multiple RLC channels can be established on each backhaul link.

IAB supports 1 : 1 mapping between UE bearers and backhaul RLC channels, where each UE bearer is carried on a separate backhaul RLC channel (Figure 5.2.7, top). This enables

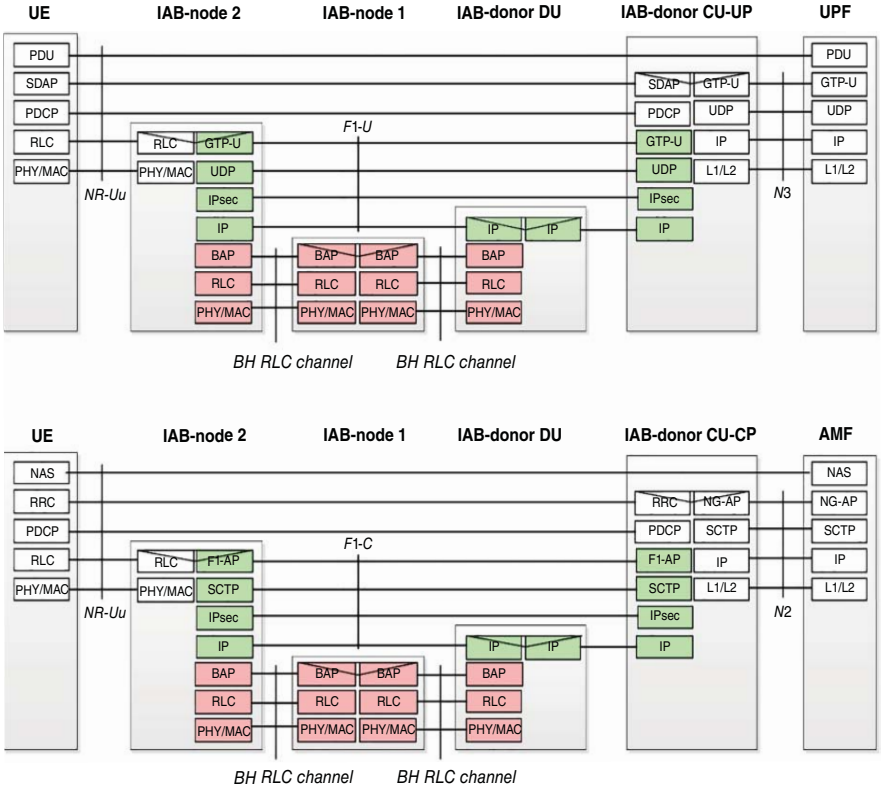
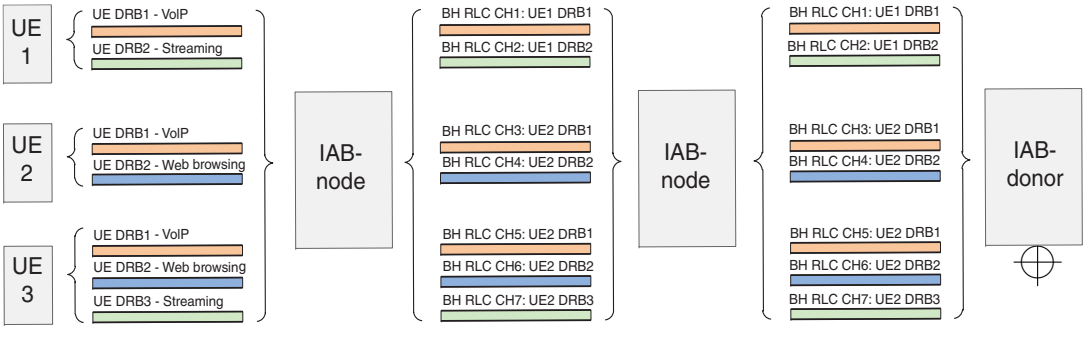
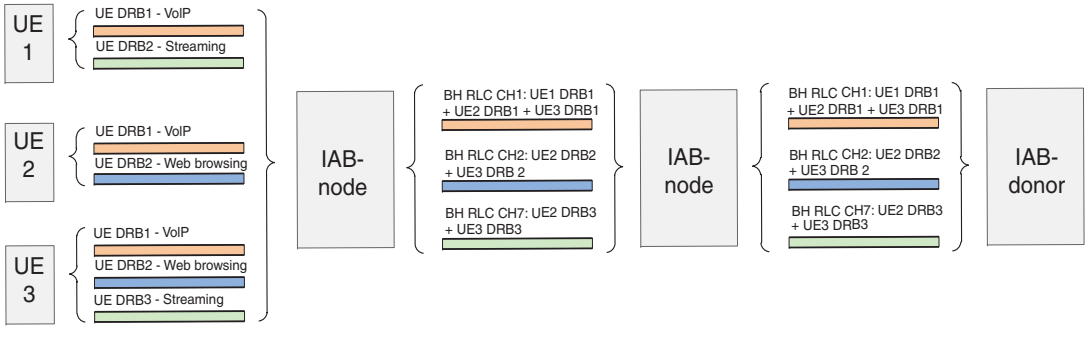



**Figure 5.2.6** Protocol stacks for UE-access with two-hop backhaul. Top: User-plane stack. Bottom: Control-plane stack.




1:1 mapping between UE bearers and backhaul RLC channels




N:1 mapping between UE bearers and backhaul RLC channels

**Figure 5.2.7** 1 : 1 and N : 1 mapping between UE bearers and backhaul RLC channels for QoS support.

bearer-specific QoS enforcement, such as guaranteed bitrate for streaming voice or video services. However, 1 : 1 bearer mapping requires update of RLC channels whenever a UE bearer is established or released. Therefore, IAB further supports N : 1 aggregation of multiple UE bearers onto a common backhaul RLC channel, which still allows traffic prioritization based, for example, on the traffic's QoS profile (Figure 5.2.7, bottom). Since such aggregated N : 1 mapping can be configured semi-statically, it reduces the signaling overhead compared with 1 : 1 bearer mapping.

The support of both features, 1 : 1 and N : 1 bearer mappings, was a compromise reached after long discussions in 3GPP. Some companies argued that fine granular QoS support on the backhaul is necessary considering congestion due to traffic aggregation at the IAB donor. Other companies were concerned such complexity would jeopardize timely availability of IAB without providing any significant benefit during an initial rollout. The final agreement was made to include the fine granular QoS support (i.e. 1 : 1 bearer mapping) but using a rather low complexity transport design to ensure timely completion of specification work as well as low-effort configuration (i.e. N : 1 bearer mapping) during an initial rollout.

Each backhaul IAB -node DU and IAB donor DU can further be configured with information on the number of descendant IAB nodes or UE bearers to permit the scheduler to apply appropriate weights to backhaul links, which ensures fairness across the IAB topology. The IAB node 1 in Figure 5.2.8, for instance, serves seven UEs via descendant nodes 2, 4, and 5, but only two UEs via descendant nodes 3 and 6. The DU scheduler on the IAB node 1 therefore needs to provide more weight to the first than to the second backhaul link. Configuration of RLC channels and scheduler-relevant information is conducted by the IAB donor CU using RRC and F1-AP signalings.

The *BAP routing ID* was introduced in the BAP layer to enable forwarding across the wireless backhaul topology. The BAP routing ID consists of *BAP address* and *BAP path ID*, and it is carried via a BAP header. The BAP address uniquely identifiers the destination node of a packet within the IAB topology while the BAP path ID allows differentiating redundant

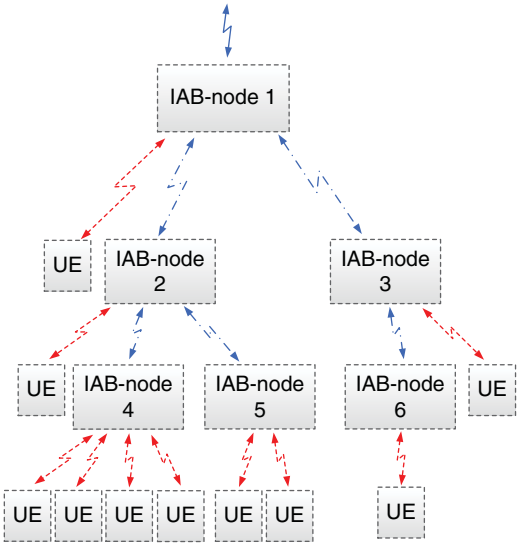



**Figure 5.2.8** Motivation for scheduler weighting on backhaul links to provide fairness.

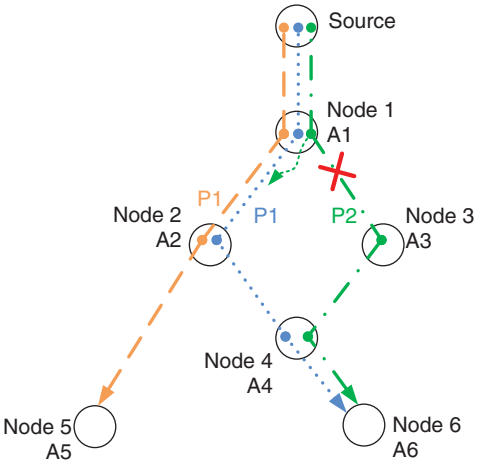



**Figure 5.2.9** BAP routing with address and path identifier.

routes with the same BAP address. The usage of both BAP address and BAP path ID is somewhat redundant and is indicative of the difficulties with the 3GPP decision making process. Nevertheless, it has the benefit of flexibility, which is elaborated upon below.

The IAB node holds a routing table with the next-hop link for each BAP routing ID. When a BAP packet has reached its final destination, it is passed to the upper layers (i.e. IP layer and above).

The combination of BAP address and BAP path ID enables source-based as well as local route selection. One example is shown in Figure 5.2.9, where the source of the packet (e.g. the IAB donor) can select between two routes, P1 and P2, that lead to the same destination (node 6 with BAP address = A6). The source node can select one of the routes by enclosing the corresponding BAP path ID together with the BAP address into the packet header. In case node 1, where both routes bifurcate, observes radio link failure on the source-selected route (path P2), it can overwrite the source's decision and send the packet on the alternative route (path P1), which it finds based on the BAP address (address A6).

The BAP header information is configured by the IAB donor CU to ensure topology-wide uniqueness of BAP routing IDs and BAP addresses. The routing table entries on the IAB nodes could either be configured via a routing protocol or by a central controller, such as the IAB donor CU, following SDN principles. Routing protocols can be designed to scale to large topologies, and they do not require a central control mechanism. The route selection rules applied by these protocols, however, need to be standardized so that interoperability is guaranteed. SDN-based route configuration has limited scalability and requires a controller function. For IAB, 3GPP decided in favor of an SDN-based mechanism since the IAB donor CU was available for central control. Also, scalability was not considered critical as topologies underneath the IAB donor were expected to have at most a few tens of IAB nodes. The configurations of BAP routing IDs and routing table entries is performed using a new class 1 F1-AP procedure defined for that purpose, called BH ROUTING CONFIGURATION.

The IP layer on top of the BAP layer enables IP routability between IAB node and IAB donor CU across wireless and wireline networks. Since the IP layer is primarily used to carry F1 protocols, it has been logically associated with the IAB donor DU. This sets it apart from the IAB node MT's IP connectivity to a data network, which is provided via a PDU

session (for IAB nodes using standalone 5GC mode) or PDN connection (for IAB nodes using EN-DC). The IAB node DU's IP address is different from that allocated to the collocated IAB node MT since they connect to different IP networks. While the IAB node MT's IP address is configured by the core network using Uu procedures, the IAB node DU's IP address is allocated within RAN by the IAB donor.

In case the IAB donor is split, the IAB donor DU holds an IP router function, which interconnects wireless and wireline IP networks. For IP packets arriving from the wireline network (downstream direction), the IAB donor DU inserts the BAP header and selects the backhaul link as well as RLC channel. For this purpose, it needs to hold a mapping between IP packet header information and L2 parameters (i.e. BAP routing ID and RLC channel ID). It is not possible to use inner packet headers for this mapping since they are protected via a security layer such as IPsec. For this mapping, a combination of destination IP address, Differentiated Services Code Point (DSCP) field, and Flow Label field for IPv6 are used to indicate BAP routing ID and RLC channel ID. The mapping rules are configured by the IAB donor CU on the IAB donor DU via F1-C. In the case where the IAB donor CU is split into control-plane and user-plane parts, the CU-CP has to configure the selection of the corresponding IP header field values (e.g. such as Flow Label values) on the CU-UP via the E1 interface.

Backhaul transport is subject to packet loss on the wireless links as well as congestion-related packet drop. Packet loss on the wireless access and backhaul links is mitigated via Hybrid ARQ (HARQ) on the MAC layer and ARQ on the RLC layer. Packet drop due to congestion may occur if the capacity on the ingress link is much larger than that of the egress link. This phenomenon is also encountered on wireline networks and generally handled by flow and congestion control mechanisms on higher protocol layers, such as TCP. On the F1 interface, congestion-related packet drops have an aggravating impact since the UE's PDCP layer on top of F1 applies packet reordering and stops packet delivery to upper layers if a packet is missing. Congestion is likely to occur at the wireline-to-wireless interface (i.e. the IAB donor DU for IAB) since the wireline capacity is likely substantially higher than that of the wireless interface. For that reason, a flow and congestion control mechanism referred to as the NR user-plane protocol (3GPP TS 38.425) is used on F1, which embeds feedback on downlink congestion-related packet loss. In the upstream direction, congestion is likely to occur on the wireless part of the backhaul close to the IAB donor DU due to traffic aggregation. Such congestion, however, can be throttled by the MAC layer scheduler on the upstream side of the wireless links by limiting uplink grants to its child nodes and UEs in case it experiences high buffer load.

#### 5.2.3.3 Resource Coordination

For access systems, duplexing schemes are typically handled by the scheduler. For IAB, the gNB-DU scheduler needs to consider half-duplexing among access links and backhaul links to child nodes. Additional duplexing constraints need to be considered between links to parent and child nodes and among multiple parent nodes. Due to the interconnectedness of backhaul links throughout the IAB topology, it is necessary to enforce topology-wide resource coordination, which translates into a graph edge coloring problem.

Figure 5.2.10 shows a few examples that illustrate topology-wide resource coordination in a spanning tree topology. In one scheme, orthogonal resources are assigned in an

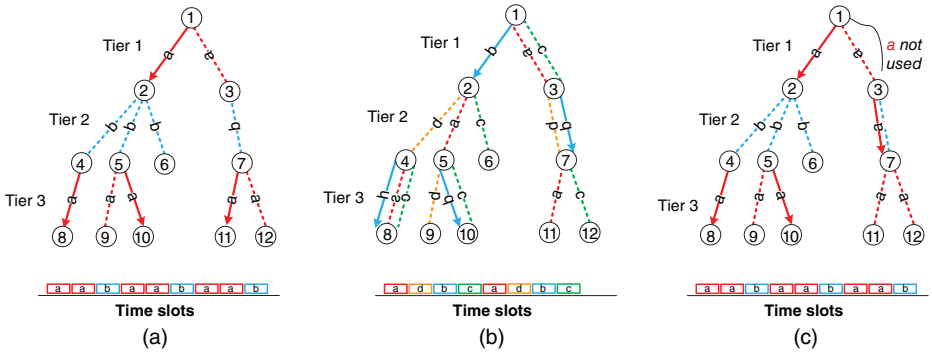



**Figure 5.2.10** Resource allocation across spanning tree: (a) resources configured per DU, (b) configured per link, and (c) configured per DU with local optimization.

alternating pattern across tree levels (Figure 5.2.10a). Within each tier, the gNB-DU scheduler can dynamically multiplex the tier's resources among its downstream links (the link selected by the scheduler is indicated by a solid line with an arrow). While this scheme meets the duplexing constraint with efficient resource reuse, it may lead to suboptimal resource utilization: if node 1 assigns its resource (red) to node 2, the other link to node 3 as well as its child links (links controlled by node 3) remain idle. It is possible to define a topology-wide resource allocation scheme, where each backhaul link obtains its own subset of resources (Figure 5.2.10b). While this multi-edge coloring problem allows optimal resource allocation for a given traffic distribution, it confines the local flexibility of the scheduler and cannot respond to traffic load fluctuations.

IAB allows semi-static resource allocation per DU and per link. In addition, a local signaling mechanism is supported via the physical downlink control channel, where the DU can dynamically lease the resource it owns to a child node in case it is not needed. Figure 5.2.10c shows an example of DU-wide resource allocation where the scheduler in tier 1 leases its "a" resource to child node 2. In this manner, scheduling flexibility can be retained while local resource leasing can optimize inefficiencies.

Another factor to be considered is cross-link interference among non-adjacent access and backhaul links. For mmWaves, it is expected that the use of narrow beams highly reduces cross-link interference. This especially applies to backhaul links where both link endpoints can support narrow beams via extended antenna arrays. On the access link, the UE usually has limited beamforming capabilities due to its small size, and its beams are therefore rather broad. For IAB deployments using sub-6 GHz, cross-link interference (CLI) becomes a more significant factor that needs to be considered during operation.

Since the problem of CLI also applies to access links, 3GPP has developed a CLI measurement framework in Release-16. While this framework aims to enable dynamic Time Division Duplexing (TDD) on access links, it can also be applied to CLI issues in IAB.

Resource coordination relies on time synchronization among IAB nodes and IAB donor. For this reason, IAB supports time synchronization via GNSS and via hop-by-hop over-the-air (OTA) signaling across the IAB topology. The OTA solution is beneficial if GNSS is not available such as for indoor deployments. It leverages existing access link

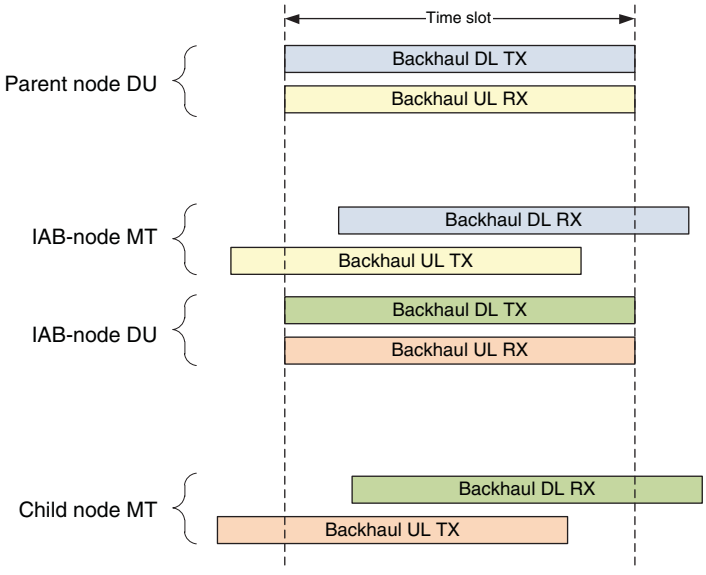



**Figure 5.2.11** Time alignment of IAB node transmission and reception.

synchronization methods between a gNB-DU and a UE (and therefore also MT), which achieves sufficient accuracy within the cyclic prefix defined by the orthogonal frequency division multiplexing (OFDM) numerology.

While time synchronization achieves clock alignment among IAB nodes, it is still necessary to determine the relative offset of the system frame boundaries within the topology. Due to the finite propagation delay of the air interface, data transmission and reception cannot be time-aligned. NR compensates for this delay by moving the UE's uplink transmissions up in time so that the DU's uplink reception and downlink transmission are sufficiently time-aligned. In a multi-hop topology, there are various options for relative time alignment of adjacent IAB nodes with different trade-offs.

In one scheme, which is used for the first IAB release (Release-16), the frame boundaries are aligned for downlink transmissions of all IAB nodes. This makes the entire IAB topology appear to the UE like a time-aligned small-cell system (Figure 5.2.11). Since the uplink transmission by an MT occurs earlier to account for the propagation delay, it cannot be time-aligned with the downlink transmission of the collocated DU. This mismatch leads to slight, but tolerable inefficiency for TDM-based resource partitioning between upstream and downstream links. SDM, however, cannot take advantage of common Fast-Fourier-Transform processing for upstream and downstream transmissions or receptions under such conditions.

Other time-alignment schemes allow for an efficient SDM operation, but they lead to misalignment of downlink transmissions across the IAB topology, which may limit the tolerable hop count. These schemes may be considered for later releases of IAB.

#### 5.2.3.4 Plug-and-Play Network Integration

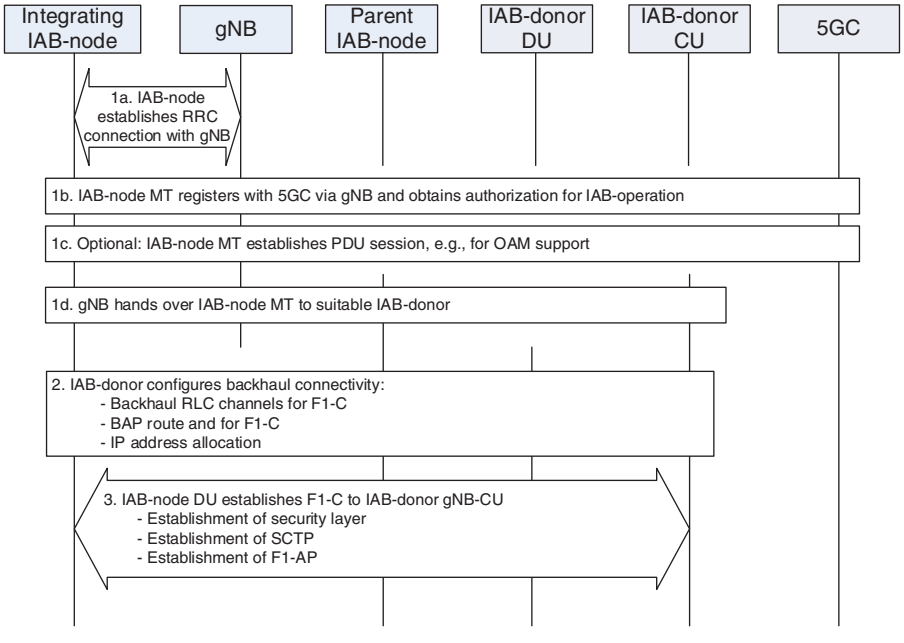
To enable lower-cost rollout of IAB nodes, the integration of IAB nodes has to follow a fully automated procedure. This procedure should ensure interoperability, where IAB nodes of one vendor connect with IAB donors and neighbor IAB nodes of other vendors (assuming proper interoperability testing has been performed). The integrating IAB node should further enjoy OAM connectivity to allow for vendor-specific configuration and optimization. Network integration needs to include the following tasks:

- Bootstrapping of IAB node MT and DU functions;
- Discovery of network, including discovery of suitable parent nodes and IAB donor CU;
- Authentication to the network and authorization of IAB operation by network;
- Support of OAM connectivity;
- Establishment of signaling and data connectivity for backhauling.

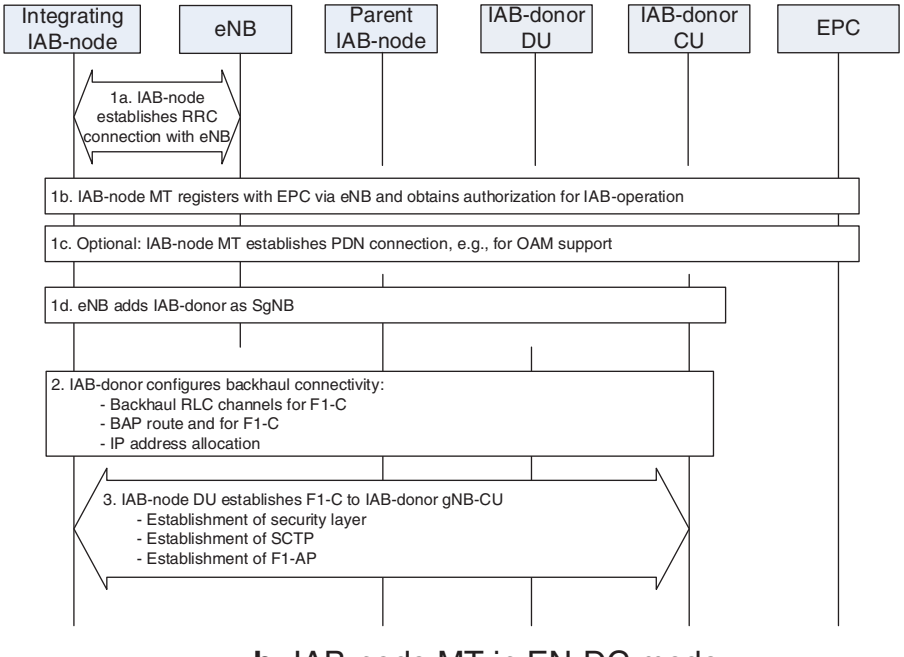
To reduce specification efforts, IAB node integration leverages the existing procedures. It includes the following sequence of steps (Figure 5.2.12):

1. The IAB node launches the MT function and connects to the network using the existing Uu procedures.
   a. It selects a cell and establishes an RRC connection. The RRC connection uses NR access for IAB nodes operating in standalone mode or LTE when operating in EN-DC mode.
   b. The IAB node MT registers to the core network (5GC or EPC, depending on standalone or EN-DC mode of CN connectivity, respectively) and is authenticated for IAB operation.
   c. The IAB node MT function may establish IP connectivity for OAM support. For IAB nodes operating in standalone mode, the MT establishes a PDU session to a data network via the 5GC. For IAB node operation in EN-DC, the MT establishes a PDN connection to an Access Point Name (APN) via the EPC. These procedures are the same as for a UE.
   d. When operating in standalone mode, the IAB node's MT is handed over to a suitable IAB donor CU unless it is already connected to such a suitable IAB donor. If operating in EN-DC, the eNB uses the dual connectivity procedure to add a suitable IAB donor CU as SgNB. In either case, establishment of the connection to the IAB donor CU occurs via a gNB-DU, which may reside on an IAB node or the IAB donor.
2. The IAB donor configures backhaul connectivity.
   a. The IAB donor gNB configures backhaul RLC channels to carry the IAB node DU's F1-C. Potentially, additional backhaul RLC channels are pre-emptively configured to also carry F1-U. This configuration is performed via the IAB node MT's RRC connection.
   b. The IAB donor gNB configures the BAP routing identifiers for the IAB node and populates routing entries on IAB nodes residing on the route between the new IAB node and the IAB donor gNB.

An IP address is allocated for the IAB node DU to support IP connectivity for F1. This IP address is either allocated by the IAB donor CU or the IAB donor DU. In both cases, it is configured via RRC.




**a:** IAB-node MT in standalone mode




**b:** IAB-node MT in EN-DC mode

**Figure 5.2.12** IAB node integration into network (a) IAB node MT operating in SA mode with 5GC, (b) IAB node MT operating in EN-DC with EPC.

In the topology change scenario, the IAB node MT leverages inter-gNB handover to migrate from the source to the target IAB donor CU (3GPP TS 38.300, clause 9.2.3). The configuration of new RLC channels, BAP route, and IP addresses is the same as in step 2 for intra-IAB donor parent node change described in Figure 5.2.13. The IAB node DU needs to establish a new F1-C association with the target IAB donor CU, where it adopts the target IAB donor CU's cell identifier (5CGI) and potentially also different Physical Cell IDs (PCIs). The UEs served by the migrating IAB node will experience this reconfiguration as a radio link failure and attempt the RRC Configuration Reestablishment procedure to the reconfigured cell. This procedure is certainly more disruptive than the parent node change underneath the same IAB donor CU. Enhancements to mitigate this issue can be introduced for UEs of later releases.

IAB supports topological redundancy when an IAB node has multiple routes to an IAB donor. In Release-16, the IAB network is restricted to directed acyclic graph (DAG) topology, which imposes hierarchical structure and prohibits an IAB node from becoming a child and parent node of a peer at the same time.

The IAB topology can be extended from spanning tree to DAG by allowing an IAB node to concurrently connect to multiple parent nodes. For this purpose, an IAB node can leverage the dual connectivity procedures (3GPP TS 38.300 clause 6.8), which allows an IAB node MT to establish independent links to an IAB donor CU via two parent node DUs. This solution can only be applied for IAB nodes operating in standalone mode. IAB nodes operating in EN-DC already use one of the two links for LTE.

Figure 5.2.13 shows the procedure for the establishment of a redundant route using dual connectivity. This procedure is shown for the same topology as in Figure 5.2.12a. The following steps are included in this procedure:

1. The IAB node MT establishes a link to a second parent node.
   a. Via NR DC procedures (3GPP TS 38.300 clause 6.8), a secondary cell group is established on the second parent IAB node DU. As part of these procedures, a second RRC connection is established between the IAB node MT and the IAB donor CU via the second parent IAB node.
2. The IAB donor CU configures backhaul connectivity via the second parent node.
3. F1 connectivity is adapted to the additional backhaul route.
   a. For F1-C, the new IP address can be added as an alternative SCTP connection.
   b. Each F1-U can only use one IP address at a time. It is possible, however, to load-balance traffic by distributing F1-U tunnels to different routes, or to establish two F1-Us for each bearer.

Route redundancy can be extended to descendant IAB nodes by adding BAP route and IP connectivity to these nodes via the second parent node. The degree of redundancy can be increased by adding a second parent link to multiple IAB nodes.

The overall support of route redundancy remains limited since each IAB node can at most have two parent nodes. This restriction was made for Release-16 since it allowed the use of NR dual connectivity to reduce specification efforts. Another option proposed was to permit multiple MTs in each IAB node, which would remove this limitation. This feature may be added in later releases.

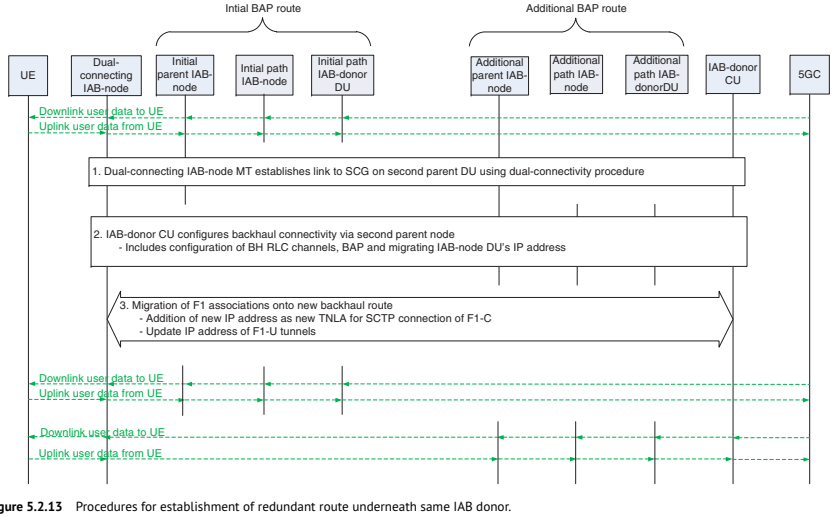



**Figure 5.2.13** Procedures for establishment of redundant route underneath same IAB donor.

Since IAB nodes may be deployed in urban areas with rather low antenna height the wireless backhaul may be vulnerable to changes in the propagation environment, for example, due to moving obstructions, changes in foliage, etc. IAB therefore supports a recovery mechanism in case backhaul radio link failure (RLF) occurs. For this purpose, the IAB node MT monitors the backhaul link.

If the IAB node MT observes an RLF and has a redundant backhaul link, it informs the IAB donor CU about the event via a measurement report, which can switch all traffic to the alternative route. This applies to NR dual-connected IAB nodes. IAB nodes operating in EN-DC can use the LTE link to forward such measurement reports.

If the IAB node MT is only single-connected, it conducts RLF recovery using the RRC Connection Reestablishment procedure (3GPP TS 38.300, clause 9.2.3). In this procedure, the IAB node MT selects a new parent node cell, where it reestablishes RRC connectivity to this or another IAB donor CU. The detailed recovery after reestablishment of RRC connectivity essentially follows steps 2 and 3 in the parent change procedures described above.

If the backhaul RLF recovery procedure fails, the IAB node MT enters RRC IDLE state. At this point, it may reestablish network connectivity through the network integration procedure.

When backhaul RLF recovery fails and the IAB node has descendant nodes, it can inform the child nodes about backhaul RLF via a notification message (Figure 5.2.14). The child nodes may pass this notification on its downstream direction. This procedure permits the descendant nodes to proactively engage in RLF recovery and reestablish their own backhaul connectivity. In this process, topology flips may occur, where an IAB node recovers backhaul connectivity via a former child or descendant node.

Backhaul RLF recovery may introduce service interruption and packet loss. However, it allows the IAB network to autonomously retain connectivity. After backhaul RLF recovery, the parent change and topological redundancy procedures can be invoked by the IAB donor to reoptimize the network.

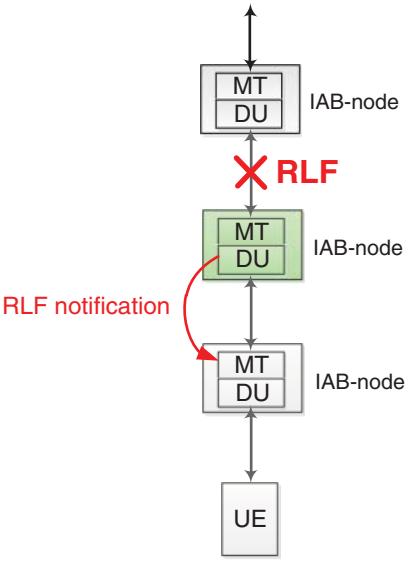



**Figure 5.2.14** Notification of RLF to downstream IAB nodes.

### 5.2.4 Outlook

IAB has been designed for the support of highly densified 5G deployments. By leveraging the benefits of mmWaves for backhauling, IAB can significantly reduce deployment costs for the rollout of mmWave access over an extended area. Integration of access and backhaul is crucial to achieve spectral efficiency. IAB supports a variety of features that are novel for cellular networks such as multi-hop backhauling, dynamic topology changes and support, topological redundancy, and autonomous failure recovery procedures.

The initial IAB specification in Release-16 primarily focused on stationary small cells. Further work on IAB in Release-17 is expected to enhance the feature set and improve on spectral efficiency and robustness. Also, performance enhancements are expected, for example, to support low-latency services over multiple hops. Specifically, the following IAB enhancements are expected to be defined in Release-17:

- Enhancements for resource multiplexing between child and parent links of an IAB node;
- CLI and interference mitigation improvements;
- Topology adaptation improvements;
- Routing improvements.

It has also been proposed to support IAB node mobility, which enables IAB node deployments on trains, buses, cabs, low and high-altitude drones, satellites, etc. At the time of writing this book, mobile IAB is not part of the 3GPP work plan; however, it may be added in the future.

## References

3GPP Technical Report 22.862 (2016). Service requirements for the 5G system; Stage1. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Report 38.874 (2019). Study on Integrated Access and Backhaul for NR. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 23.401 (2019). GPRS enhancements for E-UTRAN access. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 23.501 (2019). System Architecture for the 5G System; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 23.502 (2019). Procedures for the 5G system; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 33.210 (2019).Network Domain Security (NDS), IP network layer security. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 33.501 (2019).Security architecture and procedures for 5G system. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 36.300 (2019). E-UTRA and EUTRAN; Overall description; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.300 (2019). NG-RAN; Architecture description. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.401 (2019). NG-RAN; Architecture description. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.425 (2019). NG-RAN; NR user plane protocol. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.470 (2019). NG-RAN; F1 general aspects and principles. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.471 (2019). NG-RAN; F1 layer 1. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.472 (2019). NG-RAN; F1 signalling transport. Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.473 (2019). NG-RAN; F1 Application Protocol (F1AP). Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.474 (2019). NG-RAN; F1 data transport. Available at: www.3gpp.org (accessed May 29, 2020).
Request for Comments 4301 (2005). The Internet Engineering Task Force (IETF), Network Working Group, Security Architecture for the Internet Protocol. Available at: www.ietf.org (accessed May 29, 2020).
Request for Comments 7296 (2014). The Internet Engineering Task Force (IETF), Network Working Group, Internet Key Exchange Protocol Version 2 (IKEv2). Available at: www.ietf.org (accessed May 29, 2020).

## 5.3 Non-terrestrial Networks

*Leszek Raschkowski[1], Eiko Seidel[2], Nicolas Chuberre[3], Stefano Cioni[4], Thibault Deleu[5], and Thomas Heyn[6]*

[1] *Fraunhofer HHI*

[2] *Nomor Research GmbH*

[3] *Thales Alenia Space*

[4] *European Space Agency*

[5] *Thales Alenia Space*

[6] *Fraunhofer IIS*

Prior to 5G, 3GPP networks were traditionally designed to support terrestrial mobile networks. However, with the advent of 5G, there was on one hand a desire from mobile vendors to expand the ecosystem into new verticals and, on the other hand, a desire from the satellite ecosystem to adopt 3GPP technologies to benefit from their economies of scale, which is expected to drive down costs of satellite connectivity. To this end, 3GPP conducted a number of studies in Release-15 and Release-16 for the support of non-terrestrial network (NTN) connectivity and committed to specify enhancements needed for NTN in Release-17.

At the time of writing this book, the work on Release-17 has just started and therefore the present section describes various options that have been considered and provides a high-level outlook into what is going to be specified in Release-17 and beyond.

An NTN refers to a wide range of systems operating in various frequency bands allocated by the International Telecommunication Union (ITU) to Broadcast Satellite Services (BSS), Fixed Satellite Services (FSS), or Mobile Satellite Services (MSS). It is possible to distinguish between two types of NTNs according to the type of terminals targeted:

- *Terminals with omni-directional antennas*: The targeted terminals are handheld or Internet of Things (IoT) devices. The service links typically operate in frequency bands below 6 GHz and provide narrow to wide band services directly to end-user devices.
- *Terminals with directional antennas*: The targeted terminals are mounted on a fixed building or on a moving platform (e.g. bus, train, ship, aircraft, etc.). The service links typically operate in frequency bands above 6 GHz. These systems typically provide direct-to-home/office or backhaul services.

The first version of 5G specifications (i.e. Release-15) covers the most urgent use cases and scenarios and does not include satellite support. However, the flexibility of NR design and NG-RAN architecture allows it to be extended to new use cases, such as NTNs. Both geostationary orbit (GEO) and low -earth orbit (LEO) based NTN access is considered in the 3GPP study item described in 3GPP RP-190710 as reference deployment scenarios for NTN. While GEO satellites are located at an altitude of about 35 786 km at a fixed position above the equator, LEO satellites are circling the earth at altitudes between 500 and 2000 km at velocities of about 7 km/s. As exemplary frequency bands, both S-band (∼2 GHz for downlink and uplink) and Ka-band (downlink: 20 GHz; uplink: 30 GHz) are considered. While for GEO based NTN access only fixed beams relative to the ground are considered, for LEO based access moving beams are possible as well. Note that Unmanned Aircraft Systems (UAS), including High Altitude Platforms (HAPS) based access, has not been studied since these

**Table 5.3.1** Typical performances of NTN for considered usage scenarios.

| Usage scenarios | Experience data rate | | Max UE speed | Environment | UE categories |
|---|---|---|---|---|---|
| | Downlink | Uplink | | | |
| Pedestrian[a)] | 2 Mbps | 60 kbps | 3 km/h | Extreme coverage | Handheld |
| Vehicular connectivity | 50 Mbps | 25 Mbps | 250 km/h | Along roads in low population density areas | Vehicular mounted |
| Stationary | 50 Mbps | 25 Mbps | 0 km/h | Extreme coverage | Building mounted |
| Airplane connectivity | 360 Mbps | 180 Mbps | 1000 km/h | Open area | Airplane mounted |
| Internet of Things (IoT) connectivity[b)] | 2 kbps | 10 kbps | 0 km/h | Extreme coverage | IoT |

[a)] Better performances may be achieved.
[b)] Considering low-power wide area service capability.

could be considered as a special case of non-terrestrial access with lower delay/Doppler value and variation rate. HAPS are airborne vehicles, that is, planes or balloons, deployed in the stratosphere. HAPS operate like satellites, although being closer to earth typically at 20 km altitude, they float above conventional aircraft.

Typical throughputs that can be provided by NTNs for the usage scenarios considered in the study on extending 5G specifications to support non-terrestrial systems conducted by 3GPP in Release-16 are presented in Table 5.3.1.

The definition of a global standard based on the 5G technology for all NTN platforms in different orbits, frequency bands, and for different devices will create new service capabilities critical for consistent service continuity, reliability, and availability. Furthermore, it will decrease the cost of the network infrastructure and devices due to the economies of scale of the 5G ecosystem.

There are several effects that have an impact on 5G standards that need to be handled in order to support NTN, as shown in Table 5.3.2 for HAPS, LEO, medium earth orbit (MEO), GEO, and high elliptical orbit (HEO) satellites. These depend on the considered NTN reference scenarios (3GPP TR 38.811).

In the present section we discuss these aspects in detail.

### 5.3.1 Key Ideas

- The definition of a global standard based on the 5G technology framework for all NTN platforms, whatever orbit, frequency band, or device, will enable a smooth integration of non-terrestrial networks into the 5G system. This will contribute to create new service capabilities critical for consistent service continuity, reliability, and availability, opening up access to various verticals and a decrease in the cost of network infrastructure and devices due to the economies of scale of the 5G ecosystem.

**Table 5.3.2** NR impacts to support the NTN reference scenarios.

| Effects | | High Altitude Platforms (HAPS) | Low earth orbit (LEO) | Medium earth orbit (MEO) | Geo-stationary orbit (GEO) | High elliptical orbit (HEO) |
|---|---|---|---|---|---|---|
| Motion of the space/aerial vehicles | Moving cell pattern | Yes if beams are moving on earth | Yes if beams are moving on earth (hence high speed)[a] | Yes if beams are moving on earth (hence high speed) | No | Yes if beams are moving on earth (hence high speed) |
| | | No if beams are fixed on earth | No if beams are fixed on earth | No if beams are fixed on earth | | No if beams are fixed on earth |
| | Delay variation | No | High[b] | Medium[b] | No | Low[b] |
| | Doppler | To be determined | High[b] | Medium[b] | Negligible | Low[b] |
| Altitude | Latency | Negligible | Low | Medium | High | High |
| Cell size | Differential delay | Small | Typically relatively medium | Typically relatively medium | Possibly relatively high | Possibly relatively high |
| Propagation channel | Frequency selectiveness impairments | [c] | [c] | [c] | No | No |
| | Delay spread impairments | [c] | [c] | [c] | No | No |
| Duplex scheme | Regulatory constraints | Frequency Division Duplexing (FDD) and possibly Time Division Duplexing (TDD) | FDD and Possibly TDD | Only FDD | Only FDD | Only FDD |

[a] Assuming a fixed relation between beams and cells.
[b] Doppler and delay variation can be pre-compensated at beam center. In such cases residual Doppler and delay variation can be accommodated by the UE.
[c] Some delay spread and frequency selective effect can be experienced in the case of an omni-directional antenna device especially at a low elevation angle.

- The following NR and NG-RAN design aspects may need to be reconsidered for NTN: maximum cell size (especially for LEO and GEO based access), transparent or regenerative payload options, earth fixed or mobile beams (especially for HAPS and LEO based access scenarios), UE with and without location determination capability, e.g. GNSS (especially for LEO and GEO based access scenarios), targeted usage scenarios as in table B.2.1 in 3GPP TR 38.821, and UE type (3GPP class 3 or other).

- For direct communications to mass market devices (3GPP defined class 3 UE), the following should be considered: operation of the satellite service link in the FR1 frequency range to allow maximum commonality in the RF front end of the devices, and Frequency Division Duplexing (FDD) mode with CP-OFDM on the downlink and DFT-S-OFDM access scheme on the uplink.
- From the standards perspective, the following physical layer, protocol stack, NG-RAN, and performance adaptations need to be considered:
  - The NR user-plane protocol stack may need to be modified to accommodate larger propagation delays for the satellite access, which primarily affects the range of various protocol stack timers and RLC and PDCP sequence numbers.
  - Longer latency may also affect RRC procedures and state transitions. One promising enhancement, which is being studied in 3GPP in other contexts as well, is the reduction of the Random Access Channel (RACH) procedure from four steps to two steps.
  - Both idle and connected mode mobility have to be revisited in the context of satellite access, to accommodate the issue of UE measurement reports not necessarily being up to date due to longer propagation delays, and also because with NTN not only UE mobility but also network mobility has to be accounted for.
  - The higher Doppler shifts in the case of LEO based NTN pose another issue to be addressed in the context of timing and frequency acquisition and tracking by a UE.
  - HARQ timers and buffer sizes may need to be extended to support NTN propagation delays; alternatively, NR may need to be modified to support limited HARQ functionality or disabling HARQ altogether for NTN access.
  - For terrestrial links, the timing advance (TA) typically does not require very fast updates, which may not be the case with NTN. While TA variations are bigger with NTN, they are also more predictable, which opens up possibilities for optimizations of TA adjustments.

### 5.3.2 Market Drivers

For the cellular ecosystem, NTN can be a complementary solution addressing new market segments that are currently hard to serve using terrestrial networks. Thanks to the wide service coverage capabilities and reduced vulnerability of space/airborne vehicles to physical attacks and natural disasters, NTNs are expected to:

- Foster the rollout of 5G services in unserved areas that cannot be covered by terrestrial 5G networks (isolated/remote areas, on board aircraft, or ships) and underserved areas (e.g. suburban/rural areas) to improve the performance of limited terrestrial networks in a cost effective manner;
- Reinforce the 5G service reliability by providing service continuity for M2M/IoT devices or for passengers on board moving platforms (e.g. passengers on air vehicles, ships, high-speed trains, and buses) or by ensuring service availability anywhere, especially for critical communications and future railway/maritime/aeronautical communications;

- Enable 5G network scalability by providing efficient multicast/broadcast resources for data delivery toward the network edges or even the user terminal.

These benefits relate to either NTNs operating alone or to an integrated solution encompassing terrestrial and NTNs. The design choice will have an impact on coverage, user bandwidth, system capacity, service reliability/availability, energy consumption, and connection density.

A role for NTNs in the 5G system is expected for the following verticals: transport, public safety, media and entertainment, e-health, energy, agriculture, finance, and automotive.

For the satellite ecosystem, one of the key drivers in adopting 3GPP standardized 5G technology (as opposed to the proprietary technologies used today) is the desire to reduce both the network equipment and especially the terminal cost, by reusing a mass market technology.

### 5.3.3 NTN Based NG-RAN Architecture

The outcome of the study conducted by 3GPP in Release-16, devoted to the analysis of the key aspects and potential issues involving the integration of NTNs in 3GPP cellular systems, is reported in 3GPP TR 38.821. Hereafter, some possible reference architectures considered in the study for such satellite exploitation in terrestrial networks have been depicted. These configurations are instrumental for the next sections to highlight and review the impacts of the specific peculiarities of satellite communication links. It shall be noted that the following architectures are applicable to all possible constellations (e.g. GEO or LEO) and all available radio frequencies (e.g. S-band or Ka-band).

Multiple NG-RAN architecture options to support NTN are being considered in the study and are described in detail below. However, it is expected that only a limited subset of these options will be standardized in Release-17.

#### 5.3.3.1 Access Network with Transparent NTN Payload

The first and easiest configuration for connecting user terminals to the NG-RAN is shown in Figure 5.3.1, where a transparent satellite payload is assumed. The satellite payload implements only frequency conversion and power amplification functions in both uplink and downlink directions. In practice, the NR-Uu interface is terminated in the gNB, which is on the ground, and it is transparent for the satellite. As far as the user terminal is concerned, both handhelds and dedicated satellite equipment are envisaged.

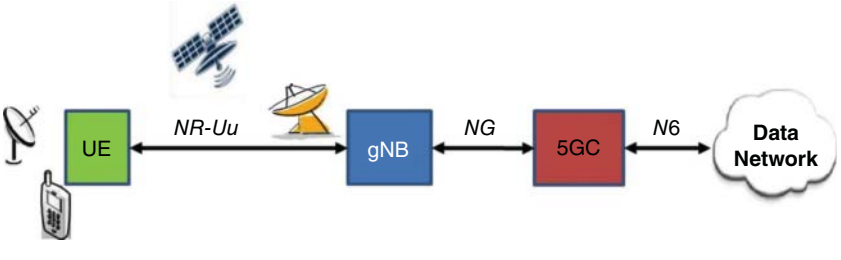



**Figure 5.3.1** Access network based on NTN platform with transparent payload.

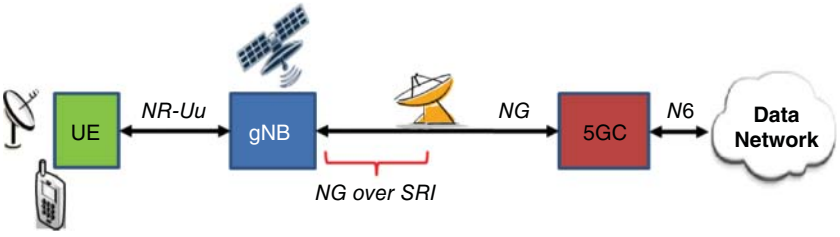



**Figure 5.3.2** Access network based on NTN platform with regenerative payload.

#### 5.3.3.2 Access Network with Regenerative NTN Payload

The architecture in Figure 5.3.2 differs from the previous one owing to the choice to embark the gNB directly on board the satellite. This requires more computational power in the satellite payload, but it has numerous benefits; for instance, it halves the propagation delay between the UE and the gNB with respect to the transparent solution. It is important to note (see Figure 5.3.2) that the Satellite Radio Interface (SRI) between the gNB and the satellite gateway on-ground is used as a transport network for the NG interface.

#### 5.3.3.3 Transport network based on NTN

For the sake of completeness, a third NTN based architecture is reported in Figure 5.3.3. In this configuration, a satellite backhaul is used between the core and terrestrial access network providing transport capabilities for the NG interface (e.g. N1/N2/N3 reference points). In this case the satellite system transparently conveys all needed 3GPP protocols on all relevant interfaces, effectively being just part of the transport network. Furthermore, the radio link can be either based on 3GPP RAT or not. This scenario is not considered further in the chapter, since it does not impact the 3GPP defined radio access network.

### 5.3.4 NTN radio protocol

A key issue to be resolved within the scope of the 3GPP work on non-terrestrial networks (NTN) is to enhance the 5G New Radio (NR) radio access protocol to support the long latencies for satellite communication. While one of the design goals of the NR air interface is extremely fast processing to enable Ultra Reliable Low Latency Communications (URLLC) services with a one-way delay of 1 ms, the same protocol needs to work over non-terrestrial satellite networks with several 100 ms propagation delay.

The NR protocol stack consists of the RRC, SDAP, PDCP, RLC, and MAC layers, which are explained in detail in Section 3.4. Depending on the architecture options described above, the NTN radio access protocol may be distributed differently across the network nodes.

In case of regenerative architecture the radio protocol is fully hosted by the satellite (Figure 5.3.4), while for transparent architecture it is located at the ground station (Figure 5.3.5).

The NG-RAN supports the new split architecture comprising a gNB-CU and a gNB-DU, which is described in Section 4.2. This architecture was also investigated for NTNs and is depicted in Figure 5.3.6. In this case the gNB-DU, which hosts the RLC, MAC, and PHY layers is in the satellite, while the gNB-CU, which hosts the RRC, SDAP, and PDCP layers is hosted by the ground station.

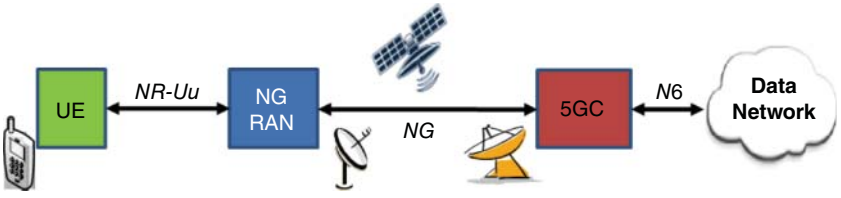
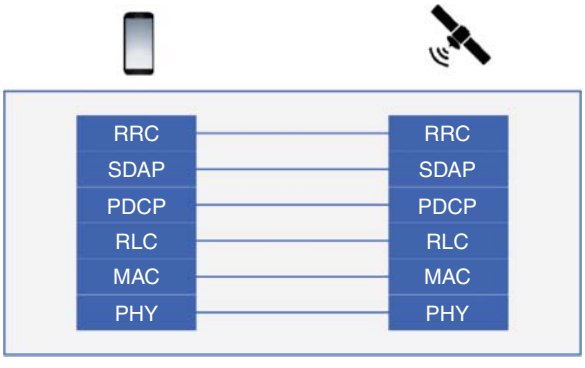



**Figure 5.3.3** Satellite backhauling configuration.

**Figure 5.3.4** Protocol stack for regenerative architecture, all of gNB functionality is in the satellite.



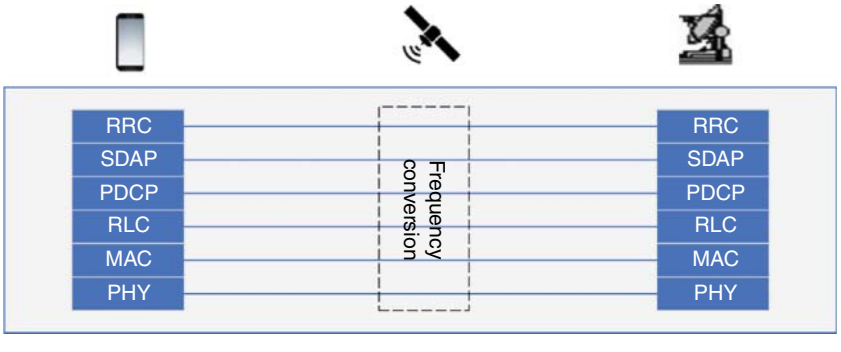




**Figure 5.3.5** Protocol stack for transparent architecture, all of gNB functionality is in the ground station.

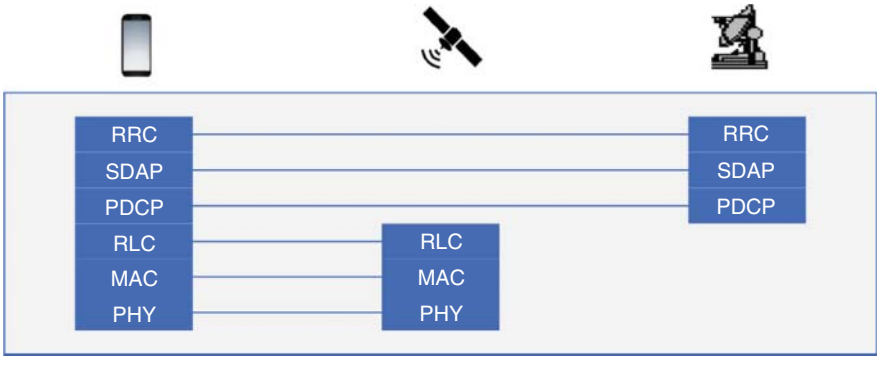



**Figure 5.3.6** Protocol stack for split architecture, parts of gNB functionality are in the satellite and parts in the ground station.

In principle, this architecture could improve radio performance toward the UE compared with the transparent architecture, as the propagation delays are reduced at least for the time-critical functions (e.g. buffer status information, scheduling, TA, and retransmission, which are part of the MAC layer), while the RRC procedures still experience a larger two-hop delay, which is somewhat less critical.

#### 5.3.4.1 Scheduling and Link Adaptation

Scheduling and link adaptation algorithms can be affected significantly by the increased propagation delay of NTNs. GEO satellites are more challenging in terms of propagation delay, while fast-moving LEO satellites are prone to faster channel variations and can cause frequent handovers. The scheduler in the MAC layer must consider the propagation delay, for instance, for the uplink/downlink resource allocations. Different to terrestrial networks, there is a large delay between measurement and reporting of the downlink channel quality, which is reported via Channel State Information feedback on the Physical Uplink Control Channel (PUCCH), and the actual downlink scheduling decision is indicated to the UE via Downlink Control Information on the Physical Downlink Control Channel (PDCCH). In the uplink, this effect might even be more severe, because of the additional delay produced by uplink Scheduling Requests (SRs) and Buffer Status Reports (BSRs) that have to be received by the gNB before a scheduling decision can be made. One possible way to mitigate this issue is an uplink preconfiguration of resources, which can reduce the uplink scheduling delay in NTN.

Furthermore, link adaptation algorithms in the MAC layer can also be affected by the large round-trip time. Examples of algorithms affected are: adaptive coding and modulation, uplink power control, TA, adaptive multiple-input multiple-output (MIMO), and outer loop link adaptation. While a scheduler in a terrestrial network might track the fast fading, it might not be feasible in NTN.

The importance of an accurate link adaptation will be even bigger if the usage of a HARQ retransmission protocol is not feasible. At least for the GEO satellites, the propagation delay will prevent the use of HARQ for high-rate eMBB applications. The round-trip time required for the retransmissions to arrive at the receiver will become too large to store all the quantized receive symbols of the erroneous packets in the HARQ protocol. Therefore, an NTN system without HARQ may need to operate at a significantly more conservative block error rate (BLER) to avoid a large number of retransmissions at higher layer, for example, 1% BLER in a system without HARQ versus 10–20% BLER in a system with HARQ.

#### 5.3.4.2 NR Layer 2 Enhancements for NTN

While the previously described link adaptation and scheduling algorithms are mostly vendor specific and are typically not standardized, several other enhancements of the NR protocol specification are required. Specification enhancements to support NTN have been studied and discussed on a per-layer basis in 3GPP: for MAC enhancements see 3GPP R2-1818511, for RLC see 3GPP R2-1818512, for PDCP see 3GPP R2-1818513, and for SDAP see 3GPP R2-1818514. Most of the enhancements of the NR radio protocol stack to support NTNs are related to timer values (3GPP R2-1900119), as the currently defined value ranges of many timers are not sufficiently large to cope with the propagation delays of NTNs. Two approaches are considered: adding an offset to the timer value equal to

the increase in propagation delay or increasing the value range in the RRC configuration protocol. Furthermore, the size of sequence numbers at RLC and PDCP layers as well as the size of the layer 2 buffer will have to be increased to ensure sufficient eMBB throughput over satellite links. Of course, some of the QoS classes and respective radio bearer configurations defined for terrestrial networks (e.g. URLLC) may not be supported due to the large propagation delay.

#### 5.3.4.3 NR Control-Plane Procedure Adaptations for NTN

The latency of RRC control-plane procedures is another concern for NTN that may benefit from enhancements (3GPP R2-1901493). The time it takes to perform RRC procedures such as connection setup and RRC state transitions may have an impact on the actual user experience. On the one hand, NTN cell sizes can be of several 100 km and not all the UEs in that area can be kept active. On the other hand, RRC state transitions toward an active state will take quite some time. Therefore, means must be found to reduce the duration between the arrival of a packet in idle state and the transmission of the first packet in connected state in order to maintain control-plane latency at reasonable levels. The use of the new RRC Inactive state (see Section 3.4) is one of the protocol enhancements defined in Release-15 5G NR, which could be useful in addressing this issue.

Another promising method to reduce the overall delay of most of the procedures is the reduction of the RACH procedure from four steps to two steps. The random access procedure specified in LTE and NR has four messages and is illustrated on the left side of Figure 5.3.7:

1. The UE starts the procedure by transmitting the Random Access Preamble (Msg1).
2. The gNB responds with a Random Access Response (RAR) providing a temporary UE identity, a TA, and an uplink grant (Msg2).
3. The following Layer 3 Message, originated at the UE, depending on the RRC procedure may include for instance an RRC Connection Request message (Msg3).

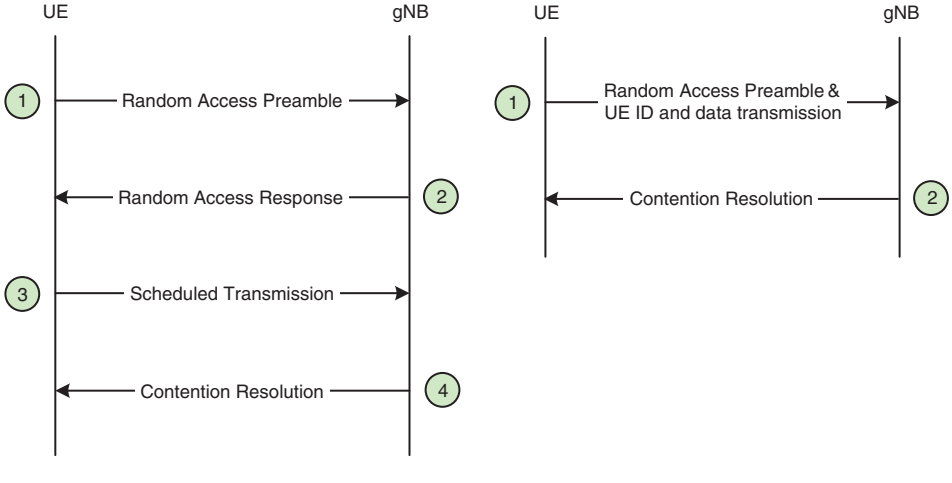



**Figure 5.3.7** Four-step RACH versus two-step RACH procedure. (Source: Reproduced by permission of © 3GPP).

4. The NR gNB resolves any possible collision by including a UE contention resolution identity (Msg4).

A two-step RACH procedure is currently being studied in 3GPP in Release-16. In this two-step RACH procedure, a new combined MsgA consists of both a preamble as well as uplink data, while a new combined MsgB includes the RAR as well as some downlink data (including contention resolution). While this new procedure can help enable low-rate, low-latency uplink services, it also provides significant benefits for the RRC procedures in NTNs (3GPP R2-1818510). The two-step RACH procedure (when standardized) will essentially halve the required latency for most of the RRC procedures, if the first RRC message can already be sent together with the Random Access Preamble.

#### 5.3.4.4 NR Mobility within NTN

Connected mode mobility is another critical area due to the additional delay in the availability of UE measurement reports in uplink and the reception of handover commands in downlink. Several enhancements are under consideration in 3GPP.

Mobility issues are quite different for stationary GEO and fast-moving LEO satellites. While handover for GEO satellites may still be triggered by signal strength-based measurements as in terrestrial networks, LEO satellites may benefit from additional handover decision criteria. There are different realizations to provide coverage for LEO satellites: earth moving beams and earth fixed beams. While earth fixed beams could reduce the number of handovers and mobility issues, earth moving beams as shown in Figure 5.3.8 are prevalent in satellite networks nowadays.

Unlike terrestrial networks, where cells are stationary, in LEO systems, cells and beams move quickly over the coverage area on the earth causing a high frequency of handovers. In order to support LEO, 3GPP will have to define mobility enhancements to minimize the handover failure rate and to reduce interruptions of the user-plane data transmission. Fortunately, LEO satellite constellations are known and LEO satellite movements are deterministic. If the UE location is known, handover triggers based on satellite ephemeris could

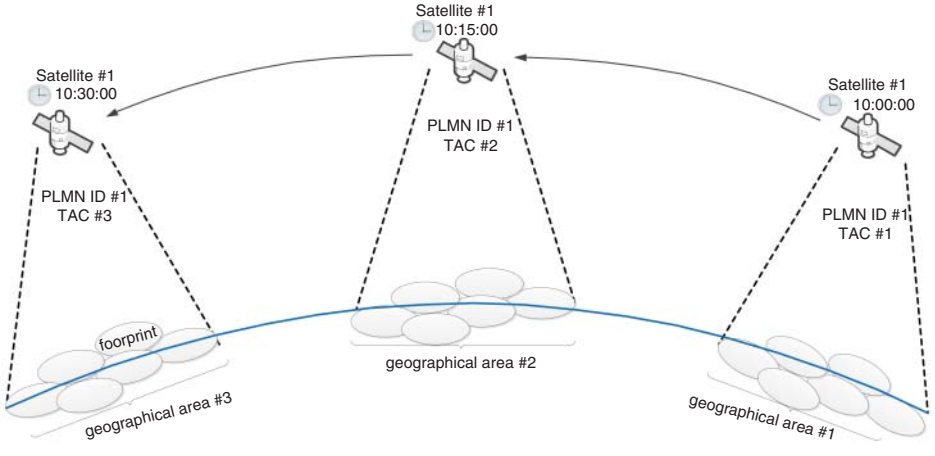



**Figure 5.3.8** Moving satellites with moving beams with earth fixed tracking area codes. (Source: Reproduced by permission of © 3GPP).

be used (3GPP R2-1818050). This can be used for UEs that support GNSS and therefore know their location; however, for other types of UEs a different method must be considered.

Similar approaches considering UE positioning information can be considered for idle mode mobility, that is,tracking area updates. Large Tracking Areas and Tracking Areas linked to a geographical area (see Figure 5.3.8) are preferred to avoid regular updates for stationary UEs. However, the size will still be limited by the overall paging capacity of the NR standard, which should not be exceeded.

Other areas of study include seamless mobility between terrestrial and NTNs as well as dual connectivity (see Section 4.3). In dual connectivity, a UE may be connected simultaneously with a terrestrial (which can host, e.g. a MgNB) and an NTN (which can host, e.g. a SgNB); this is one possible realization of seamless mobility.

### 5.3.5 NR Physical Layer Adaptations for NTN

The differences in NTN propagation channel characteristics compared with terrestrial systems, which are mentioned above (e.g. long delay, larger cell sizes, fast movements of non-GEO satellites resulting in a high and variable Doppler shift as well as a variable distance between gNB and UE), affect not just the NG-RAN architecture and protocol stack, but also the physical layer. Some of the required PHY enhancements considered in the 3GPP study are described in the present section.

For other aspects of the NR physical layer in satellite applications like the cyclic prefix dimensioning and demodulation reference symbols (DMRS) density in time and frequency, no changes are expected in NR (3GPP TR 38.811). Limited impact on the NR PTRS design is expected for scenarios in FR2 (i.e. frequencies between 24.25 and 52.6 GHz), where the maximum received SNR is limited by the phase noise of satellite on-board payloads.

#### 5.3.5.1 Timing and Frequency Acquisition and Tracking

A fundamental difference between terrestrial networks and LEO satellite or other non-GEO satellite networks is the relative mobility of the transmission infrastructure (gNB or gNB-DU), causing frequency and timing synchronization and tracking issues in a UE due to high Doppler offsets and drifts (3GPP TR 38.811). In cellular networks, the transmission infrastructure is usually fixed, except when on board a moving platform such as a train. In contrast, the transmission equipment is not static in most cases of NTN:

- For GEO systems, the transmission equipment is quasi-static with respect to the UE with only a small Doppler shift.
- For LEO, the satellites move relative to the earth and create higher Doppler shift than for GEO systems.
- For HAPS, the transmission equipment is moving around or across a theoretical central point but creates a small Doppler shift.

In order to synchronize to the 5G network, the UE has to detect the Primary Synchronization Signal (PSS) and the Secondary Synchronization Signal (SSS). Those synchronization signals allow time and frequency correction, as well as Cell ID detection. The requirement for a UE in a cellular network is to get a one-time detection probability at an SNR of the FDD downlink baseband signal of −6.4 dB with less than 1% false alarm rate (3GPP TS 38.101-4),

with robustness against initial frequency offset up to 5 ppm (3GPP TR 38.802). It is assumed here that these requirements apply to an NTN UE as well (in practice, it may be possible to relax these requirements to some extent).

The Doppler shift depends on the relative satellite or HAPS velocity with respect to the UE, and on the frequency band. In terms of Doppler shift, the worst-case example that was considered in 3GPP TR 38.811 is a non-geostationary satellite at an altitude of 600 km with a satellite speed of 7.5 km/s. The resulting Doppler shift in the downlink signal yields up to ±48 kHz at 2 GHz (equal to 24 ppm) and ±480 kHz at 20 GHz center frequency. These Doppler offsets exceed the maximum initial frequency offset of 5 ppm for the synchronization requirement. To meet the 5G requirement of 5 ppm, the satellite altitude needs to be at least 13 000 km.

As a solution, satellites can pre-compensate the Doppler offset for each of its spot beams, for example, for the beam center. The remaining Doppler offset within a beam footprint/satellite cell is significantly less than without pre-compensation.

The movement of non-GEO satellites result in a drift of the Doppler offset of up to 544 Hz/s at 2 GHz and 5.44 kHz/s at 20 GHz according to 3GPP TR 38.811. However, the remaining Doppler variation after pre-compensation for the individual satellite spot beams can be tracked by a UE in a way similar to terrestrial networks, when the DMRS have sufficient density in time.

#### 5.3.5.2 HARQ

The impacts on NR HARQ operation due to the long RTT delay of an NTN were also studied in 3GPP TR 38.811. The impacts are considered for the NTN UEs as well as the serving gNBs, when the number of HARQ processes is either extended to satisfy high reliability scenarios or limited/disabled for longer NTN delays.

Two options are envisaged for using HARQ in satellite systems:

1. Enhancing the existing HARQ operation to extend the HARQ processing accommodating low to moderate NTN RTT delays like in scenarios C1, C2, D1, and D2 as defined in 3GPP TR 38.821. Here it is important to study the impact on the possible HARQ process numbers, HARQ timers, HARQ process IDs, number of Redundancy Versions (RVs), and the possibility for bundling. These will have a direct impact on the UE, for example in terms of memory and power limitations, as well as the collected gNB signaling and feedbacks for a large number of UEs.
2. Limiting HARQ capabilities and/or disabling HARQ for long RTT delays. For that, it is important to study the direct impact on NR due to disabling the HARQ feature, including the required signaling for on/off switching of HARQ and possible metrics like the target BLER/QoS, the Modulation Coding Scheme (MCS)/Transport Block Size (TBS) selection, etc. It is also important to consider a backup mechanism to support high QoS requirements in an attainable RTT delay. This should include, e.g. HARQ-less repetitions including its impact on the data rate and utilizing terrestrial NTN dual connectivity including its impact on the used timers.

These two aspects are described in more detail below.

#### 5.3.5.2.1 Enhancing the Existing HARQ Operation

Potential solutions to extend the HARQ operation in satellite systems are proposed in 3GPP RP-180664, while a comprehensive summary is given in 3GPP R2-1817757:

- Extending the minimum number of HARQ processes.
  It was agreed in NR to support a flexible number of HARQ processes to support different use cases with moderate RTT delays, e.g. satellites with MEO and LEO constellations as well as extreme coverage. Currently NR supports 8 and 16 HARQ processes as a baseline. However, for NTN integration the number of HARQ processes needs to be increased. Therefore, it is necessary to further study the optimal extension of the minimum number of HARQ processes.
- Flexible HARQ timing.
  The HARQ processing time in the Downlink Control Information (DCI) (see Section 3.5 for details) includes at least the time between downlink data reception and the corresponding HARQ-ACK transmission in uplink. It also includes the time between uplink grant reception and the corresponding uplink data transmission, which is affected by the satellite delays. A modification of the corresponding DCI control field/format should be further investigated.
- Adaptive HARQ process ID.
  Asynchronous HARQ can be scheduled at any time slot after $N$ sub-frames, with $N$ being the number of HARQ processes configured. Therefore, a HARQ process ID is used to indicate which HARQ process refers to an existing transmission. Hence, it is important to study how to introduce more HARQ processes with less impact on the number of HARQ process IDs.
- Code block group (CBG) aggregation for reduced HARQ acknowledgements.
  NR supports CBG aggregation, which schedules data transmissions to span over one or multiple code blocks (CBs). For each CBG, transmission of one HARQ feedback is sufficient for all aggregated CBs, reducing the number of required HARQ processes and HARQ feedbacks. Therefore, an impact from longer delays in NTN needs to be considered for CBG aggregation, and whether or not the aggregation among Transport Blocks (TBs) (rather than CBs) can enhance the HARQ procedure.

#### 5.3.5.2.2 Limiting the Number of HARQ Processes or Disabling HARQ

For some satellite constellations, the number of HARQ processes is too high, for example, GEO can reach 500 HARQ processes for a 1-ms sub-frame length. In this case, mechanisms are required to reduce the number of HARQ processes. In some cases, for example GEO satellite systems, disabling HARQ completely is beneficial and allows only the initial transmission RV0.[1] The HARQ deactivation/(re)activation can be initiated either by a gNB or by a UE. This can be done dynamically (e.g. per transmission of one or more transport blocks) or semi-statically (e.g. by a deactivation for some time or over a geographical region). Different mechanisms and metrics are required to decide whether to enable or disable HARQ, for example, the UE QoS, memory size, round-trip time, network type, or transport block size,

1 Refers to HARQ RV configuration, as defined in 3GPP TS 38.312.

etc. Hence, it is important to study the impact of disabling HARQ on NR (e.g. impacts on the TBS, MCS, etc.). It is also important to consider the following HARQ reliability mechanisms to support HARQ disabling:

- HARQ-less operation and redundancy transmission.
  Transmission Time Interval (TTI) bundling and HARQ-less repetition as specified in LTE Release-15 can be used to completely avoid retransmissions, thus avoiding a large number of required HARQ processes and excessive HARQ feedbacks. In NR Release-15 a similar mechanism was also agreed for uplink, the k-repetition. Those k repetitions can be similar to the initial transmission or following a certain RV sequence. Currently, NR supports only bundling and repetition in the time domain. Therefore, the impact on repetition value k, rate reduction (i.e. 1/k), and the TBS/MCS tables needs to be studied.
  In order to mitigate longer delays between scheduling the grant and uplink transmissions, grant-free retransmissions can be used and should be studied in the case of varying channel conditions.
- Utilizing terrestrial dual connectivity.
  In the case of deactivated HARQ, when it is required to guarantee a certain QoS, the HARQ operation can still be resumed with reduced latency, utilizing a dual connectivity with a ground station (gNB) in a terrestrial network-offloading scenario (3GPP R1-1904650). The ground gNB might be backhauled using a satellite reliable link or lossless connection to the network. Once a UE is in the coverage of a terrestrial and a non-terrestrial node, the initial transmission can be sent over the NTN link, while HARQ retransmissions (other RVs) and HARQ acknowledgement/negative acknowledgement (ACK/NACK) feedbacks can simply flow over the terrestrial link (i.e. rather than the longer delay NTN link) as illustrated in Figure 5.3.9. Here, the gNB transmits the RVs to the UE upon a NACK.

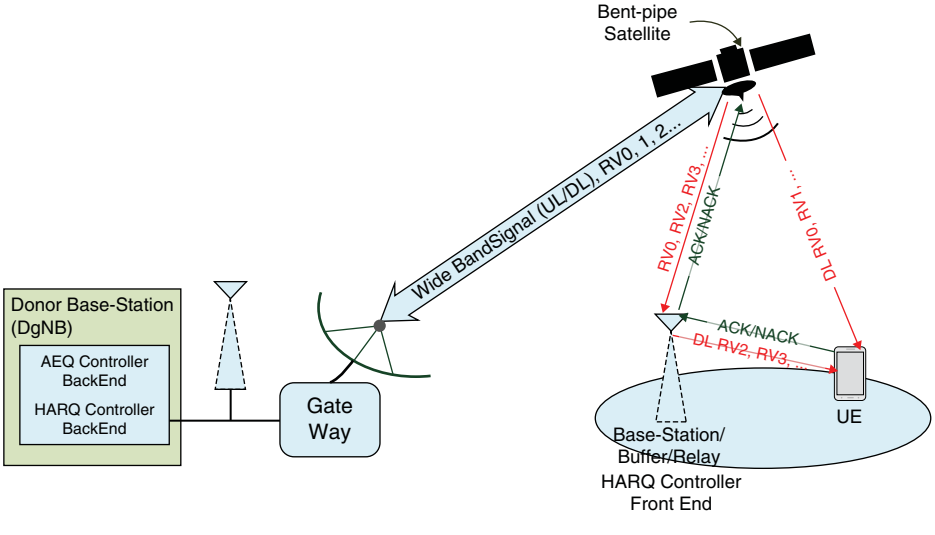



**Figure 5.3.9** Transmission of HARQ RVs to the gNB via satellite backhauling or direct satellite access.

In other scenarios, dual connectivity can gain performance by utilizing L2 aggregations, for example, PDCP aggregation. Hence, it is important to further study the impact on the NR HARQ operation in the case of dual connectivity with an existing terrestrial gNB to include HARQ redundancy decoding at the physical layer and/or L2 aggregation if possible. The impact on HARQ timing, the number of HARQ processes, and the HARQ process ID/synchronization needs to be identified in this case.

#### 5.3.5.3 Timing Advance (TA)

According to 3GPP TR 38.811, fast-moving satellites (e.g. in LEO, MEO orbits) cause fast change in the overall distance of the propagation from the UE over satellite to the gNB, and consequently a strong delay drift. In the case of a satellite, the delay drift is quite predictable because the motion of the satellite follows known paths.

The technical challenges for NR adaptations are as follows:

1. A strong delay variation is caused by fast-moving satellites (e.g. in LEO up to 7.5 km/s) generating a fast change in the overall distance of the propagation from a UE over satellite to a gNB, requiring fast adaptations for the uplink receiver synchronization and uplink transmission timing.
2. The delay is much longer over a satellite link than one TTI.

For terrestrial links, the TA typically does not require very fast updates, because the distance of the terminal to the base station only varies slowly due to the terminal mobility. In the case of GEO satellite links, the terminal mobility also dominates the TA requirements. However, this may not be the case, for example, for LEO satellites.

Another technical issue that arises is the delay variation over the satellite link being much larger than a TTI. For example, if the subcarrier spacing (SCS) is increased from 15 kHz (NR numerology 0) to 60 kHz (NR numerology 2), the TTI length decreases from 1 ms to 250 μs.

Therefore, the transmission timing of the UE has to be adjusted over the borders of individual TTIs in NTN applications, causing a high number of NR TA update commands (3GPP R1-1904225, 3GPP TS 38.213). To mitigate this effect, in LEO scenarios a UE should be enabled to perform the compensation of the predictable delay variation by itself.

The delay variation for satellite networks is quite predictable. Depending on the available information, the prediction of the delay variation can be done in the terminal for different time durations, resulting in a significant reduction of the required control plane overhead for TA adjustments. Two options were considered:

- The longest prediction of the fast timing drift is possible, where the UE knows its position (e.g. by GNSS) and the satellite ephemeris. Therefore, the UE is able to calculate the exact TA and apply it for uplink transmissions.
- For UEs with fewer capabilities (e.g. for MTC use cases), not knowing the position and ephemeris, the delay drift prediction can still be done for a smaller time window. Different solutions are possible, when the UE has been enabled for timing prediction, e.g. based on the last TA commands and their updated values, or based on the observed downlink timing drift via SSB tracking. Supporting information by the gNB to the UE is also possible, e.g. provisioning of timing drift rates.

#### 5.3.5.4 Physical Layer Control Loops

A UE operating in GEO satellite access networks can experience a one-way propagation time of up to 270 ms (3GPP TR 38.811). Using a LEO satellite access network at 600 km altitude, the one-way propagation delay can range between 2 and 7 ms. The slow reaction time caused by this delay has a performance impact on physical layer procedures, particularly on those with closed loops such as power control and adaptive modulation and coding (AMC)[2].

While slower reaction affects the performance of all control loops between a UE and a gNB, most of them require some adjustments in the implementation, not a different design. Satellite power is typically limited and an optimum selection of operating point concerning transmission power, modulation, and coding is of utmost interest. Due to the large free-space path loss, and the limited Effective Isotropic Radiated Power (EIRP) and battery power available at the UE, the power margin is also limited for mobile terminals. With the long delay over satellites in the loop, the power control is not able to track fast-fading channel conditions, only slower power variations.

For Ka-band satellites in static scenarios, Adaptive Coding and Modulation (ACM) used in the digital video broadcasting satellite specification DVB-S2 is an essential mechanism that maintains connection through rain fades, which typically change somewhat slower than the half-second round-trip delay. It works well in some cases helping to avoid excessive oscillations between two modulation and coding modes. Having said that, the reaction time of ACM is too slow to adapt for changes of signal strength for mobile terminals, especially when the line of sight is interrupted by shadowing events.

For S-band GEO systems in mobile scenarios, the main issue is multi-path fading, which can be much faster than a half-second round-trip delay. AMC applied to 5G NR in this scenario will not be able to follow it. The AMC algorithm typically attempts to settle on a modulation and coding mode that closes the link if possible, by giving up some power to maintain a margin.

For LEO satellites, AMC can be used in NR to adapt for the free-space path loss variation. This variation is sufficiently slow compared with the 20-ms worst case round-trip delay. AMC should be able to react to shadowing fades to a large extent as well, although it is still unable to follow fast fading. Therefore, instead of using closed-loop power control, open-loop power control as defined in 3GPP TS 38.213 seems beneficial for NTN scenarios and should be used in mobility scenarios to compensate for the path loss and adjust the signal power to an optimum target level.

### 5.3.6 NTN Channel Model

The 3GPP NR channel model is defined in 3GPP TR 38.901 for terrestrial links. It was necessary to define a specific NTN channel model as it may differ from terrestrial channels for several reasons:

- The elevation angle between the satellite/HAPS and the UE can be much higher resulting in different scattering statistics.

2 AMC and ACM are used interchangeably in different specifications.

**Table 5.3.3** Objectives of non-terrestrial network (NTN) channel modeling.

| | Satellite | High Altitude Platforms (HAPS) |
|---|---|---|
| Outdoor/indoor | Only outdoor conditions | Both |
| Support frequency range | From 0.5 GHz up to 100 GHz | |
| UE mobility | Up to 1000 km/h | Up to 500 km/h |
| User environment | Open, rural, suburban, urban, and dense urban | |

- The long distance between the satellite and the UE leads to almost no angular spread from the satellite.
- Atmospheric effects may attenuate the transmitted signal.

An NTN channel model has therefore been defined in clause 6 of (3GPP TR 38.811) with modeling objectives summarized in Table 5.3.3.

The methodology that was adopted is to consider a new satellite link model between the satellite/HAPS and the terrestrial model, as depicted in Figure 5.3.10. The satellite link consists of the dynamic delay and Doppler shift (for LEO satellites only) and the dynamic attenuation due to rain, clouds, and scintillation; the latter being described in ITU recommendations ITU-R P.681-10 and ITU-R P.618-13. The terrestrial part is similar to the 3GPP TR 38.901 channel model with a specific elevation-dependent parameterization obtained by ray-tracing simulations for all of the considered frequency bands and scenarios in line of sight and non-line of sight.

As a simplified alternative to the fast-fading model defined in 3GPP TR 38.901, the flat-fading model defined in ITU-R P.681-10 may also be considered if the following conditions are met:

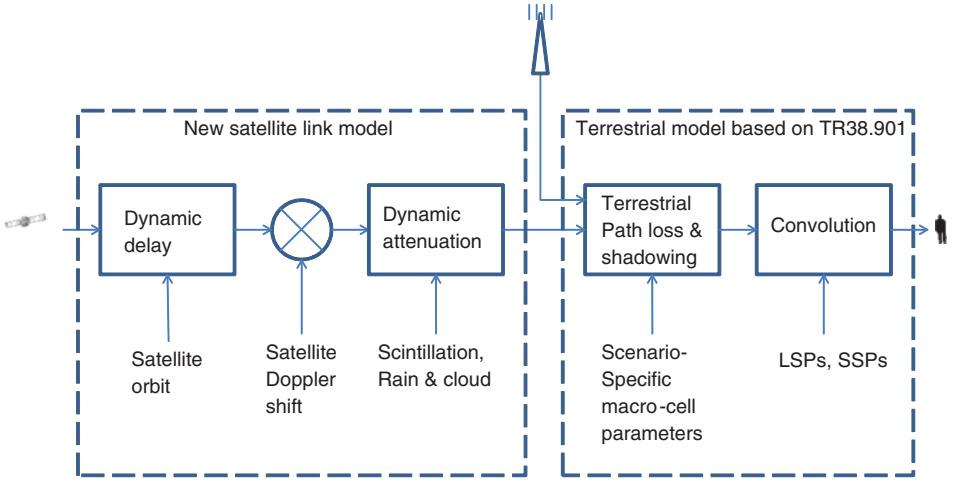



**Figure 5.3.10** Combined satellite and terrestrial channel model methodology. (Source: Reproduced by permission of © 3GPP).

- S-band with channel bandwidth up to 5 MHz;
- Minimum of 20° elevation;
- Rural, suburban, or urban environment;
- Quasi-line-of-sight conditions (fading margin of approx. 5 dB).

### 5.3.7 Outlook

3GPP has conducted several studies in Release-15 and Release-16 on enhancements required for NG-RAN architecture, as well as the NR physical layer and protocol stack to be deployed with NTN. At least some players in the satellite community envision deploying NR-based NTN by 2025, encompassing all deployment options, that is, GEO, MEO, LEO, and HAPS.

The vision is to deploy NTNs as part of 5G by 2025 in order to meet the challenges of mobile network operators and verticals in terms of reachability, reliability, and resiliency. This encompasses all deployment options like GEO, MEO, LEO, as well as HAPS.

It has been identified that NTN solutions introduce potentially new constraints compared with typical cellular deployments due to moving cell patterns, larger Doppler shifts and variation, larger and varying propagation delays, larger cell sizes, the highly frequency selective propagation channel, the power limited link budget, and feeder link handover.

Depending on the considered NTN scenario (orbit, device, frequency band), the following NG-RAN features may need to be enhanced: random access, uplink/downlink synchronization, HARQ, user-plane timers, idle/connected mode mobility, feeder link handover procedure, radio resource management, and multi-connectivity/mobility management across cellular/NTN access.

At the time of writing this book, 3GPP has approved a work item (3GPP RP-193234) for normative specification work to support NTN, based on the outcome of the studies described in this section. This initial normative work to be done in Release-17 should cover some of the scenarios outlined above, specifically:

- GEO, LEO, and HAPS;
- Transparent payload only (formally this implies that a gNB is deployed on a ground station, however the same system most likely will be able to support gNBs deployed on a satellite);
- UEs with GNSS;
- Satellite-specific RRM requirements.

## References

3GPP R1-1904225 (2019). Timing Advance Adjustments for Satellite Communications (NTN), Fraunhofer IIS, Fraunhofer HHI, 3GPP TSG-RAN WG1 Meeting #96bis, Xi'an, April 2019.

3GPP R1-1904650 (2019). Doppler Compensation, Uplink Timing Advance, Random Access and UE Location in NTN, Nokia, Nokia Shanghai Bell, 3GPP TSG RAN WG1 Meeting #96bis, Xi'an, China, April 2019.

3GPP R2-1817757 (2018). NR-NTN: HARQ in Satellite Systems, Fraunhofer IIS, Fraunhofer HHI, 3GPP TSG-RAN WG2 Meeting #104, Spokane, November 2018.

3GPP R2-1818050 (2018). Satellite Ephemeris Data and their Use for Handover Decisions between Satellites, Hughes, 3GPP TSG-RAN WG2 Meeting # 104, Spokane, USA, November 2018.
3GPP R2-1818510 (2018). Initial Random Access Procedure in Non-Terrestrial Networks (NTN), Nomor Research GmbH, 3GPP TSG-RAN WG2 Meeting # 104, Spokane, USA, November 2018.
3GPP R2-1818511 (2018). Considerations on MAC Timers and on RTD Compensation Offset in Non-Terrestrial Networks (NTN), Nomor Research GmbH, 3GPP TSG-RAN WG2 Meeting # 104, Spokane, USA, November 2018.
3GPP R2-1818512 (2018). Considerations on RLC Control Loops and Timings in Non-Terrestrial Networks (NTN), Nomor Research GmbH, 3GPP TSG-RAN WG2 Meeting # 104, Spokane, USA, November 2018.
3GPP R2-1818513 (2018). Considerations on PDCP Control Loops and Timings in Non-Terrestrial Networks (NTN), Nomor Research GmbH, 3GPP TSG-RAN WG2 Meeting # 104, Spokane, USA, November 2018.
3GPP R2-1818514 (2018). Considerations on SDAP in Non-Terrestrial Networks (NTN), Nomor Research GmbH, 3GPP TSG-RAN WG2 Meeting # 104, Spokane, USA, November 2018.
3GPP R2-1900119 (2019). Report of email discussion [104#51] [NR - NTN] – Impacts on user plane timers, Nomor Research GmbH (email discussion rapporteur), 3GPP TSG RAN WG2 Meeting #105, Athens, February 2019.
3GPP R2-1901493 (2019). RRC Connection Control and State Management for NTN, Nomor Research GmbH, 3GPP TSG RAN WG2 Meeting #105, Athens, February 2019.
3GPP RP-180664 (2018). NR-NTN: Preliminary solutions for NR to support non-terrestrial networks, Thales, Nokia, Nokia Shanghai Bell, HNS, June 2018.
3GPP RP-190710 (2019). SID for study on solutions for NR to support non-terrestrial netsworks (NTN), March 2019.
3GPP Technical Report 22.822 (2018). Study on using satellite access in 5G. Available at: www .3gpp.org (accessed May 29, 2020).
3GPP Technical Report 23.737 (2019). Study on architecture aspects for using satellite access in 5G Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Report 38.802 (2017). Study on New Radio Access Technology Physical Layer Aspects Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Report 38.811 (2019). Study on New Radio (NR) to support non-terrestrial networks Available at:www.3gpp.org (accessed May 29, 2020).
3GPP Technical Report 38.821 (2020). Solutions for NR to support non-terrestrial networks Available at:www.3gpp.org (accessed May 29, 2020).
3GPP Technical Report 38.901 (2020). Study on channel model for frequencies from 0.5 to 100 GHz Available at:www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 22.261 (2020). Service requirements for next generation new services and markets Available at:www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 36.300 (2020). E-UTRAN; Overall description; Stage 2 Available at: www.3gpp.org (accessed May 29, 2020).
3GPP Technical Specification 38.101-4 (2020). User Equipment (UE) radio transmission and reception; Part 4: Performance requirements Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.213 (2020). NR; Physical layer procedures for control Available at: www.3gpp.org (accessed May 29, 2020).
ITU-R M.2047-0 (2013). Detailed specifications of the satellite radio interfaces of International Mobile Telecommunications-Advanced (IMT-Advanced), December 2013.
ITU-R P.618-13 (2017). Propagation data and prediction methods required for the design of earth-space telecommunication systems, December 2017, December 2017.
ITU-R P.681-10 (2017). Propagation data required for the design of earth-space land mobile telecommunication systems, December 2017.
RP-193234 (2019). Solutions for NR to support non-terrestrial networks (NTN).

# 6

# Enabling Technologies

## 6.1 Introduction

In the previous chapters we discussed the NG-RAN architectures defined in Release-15 and their evolution in Release-16 and -17, covering both architectures defined in 3GPP and other standards bodies and industry fora, for example, the O-RAN Alliance and Small Cell Forum.

Architectures defined in these standards developing organizations (SDOs) are typically limited to what is essential to define a RAN and intentionally use an abstraction model that separates other technologies, which are required to build and operate a wireless network such as the transport network, virtualization technologies, and applications (e.g. multi-access edge computing [MEC]).

In the present chapter we describe various technologies required to deploy NG-RAN, which are often either overlooked completely by 3GPP specifications or are addressed only partially.

## 6.2 Virtualization

*Sridhar Rajagopal*

*Mavenir, USA*

In computing, virtualization typically refers to the idea of decoupling software from "real" hardware and thus allowing one to run said software on any platform, which provides a virtualized environment.

The term was originally applied to hardware virtualization, which allows the running of multiple instances of (often different) operating systems on the same hardware, through the usage of virtual machines that appear to the operating system as a real hardware. Subsequently, the concept evolved to support more advanced management features, such as taking snapshots of a virtual machine state, failover, and migration.

Apart from the economic benefits of decoupling software from hardware (and thus allowing one to source software and hardware components from different vendors), virtualized environments are typically more robust in the case of failures and provide better hardware utilization.

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

After virtualization had been extremely successful in the cloud and enterprise, there has been an increased interest in applying the same principles in networking. In fact, in many cases even 4G Evolved Packet Core (EPC) runs in virtualized environments and this trend is expected to continue with 5GC, which was designed with virtualization in mind.

Once the core network is virtualized, the next logical step is to consider virtualizing (at least parts of) NG-RAN. While virtualized NG-RAN deployments are not explicitly defined in the standards, certain provisions have been made to make such implementations possible (see Chapter 4). In the present chapter we describe how NG-RAN virtualization may be realized.

Note that the term "virtualization" is somewhat loosely defined and is often used in many different contexts to carry different meanings, especially in the area of networking. At least in some cases it means moving away from proprietary to a general purpose hardware, which may or may not require actual virtualization. Additionally, it may mean "softwarization" – that is, moving parts of functionality from custom hardware to software. In the context of this section we describe various technologies that allow virtualization in its broad meaning, covering most of the cases mentioned above.

### 6.2.1 Key Ideas

- RAN virtualization refers to an ability to run the RAN protocol stack (and other RAN functions) largely in software in an off-the-shelf hardware platform consisting of servers utilizing commercially available general purpose processors and accelerators such as x86 central processing units (CPUs), field programmable gate arrays (FPGAs), and smart network interface controllers (SmartNICs), where the stack implementation is not tied to a particular hardware (i.e. is virtualized) and the solution can be scaled based on the workload and resources available.
- With more powerful processors coming every generation and cloud-based data centers, virtualization enables operators to reduce capital expenditure (CAPEX) and operational expenditure (OPEX) by not requiring proprietary hardware, resource pooling across sites, and enabling solutions to be upgraded and operated in a centralized and scalable manner.
- The two most common virtualized platforms are hypervisors (virtual machines [VMs]) and containers. While the hypervisor is a more mature technology, which has been deployed for many applications (e.g. data centers), the growing trend in virtualization is moving toward a container-based cloud native platform, which enables lower memory footprints, lower processing overheads, and better support for real-time requirements.
- The 3GPP and O-RAN standards have evolved to allow functional splits of the RAN that enable most of the RAN to be virtualized (and deployed in a cloud or a data center) while keeping radio and a small portion of the physical layer processing (e.g. time-domain processing) to be implemented in hardware and remain deployed on the site. That being said, the relevant standards typically do not define virtualized RAN explicitly, but rather make it possible in implementation.
- There are several challenges to RAN virtualization, which is getting more complex as we move toward high data rate and low latency requirements for 5G. These include support for hardware accelerators, timing/synchronization, scaling with load, high availability, and managing power consumption.

### 6.2.2 Market Drivers

5G has been designed to support various use cases including high data rate and low latency services. In order to provide such services, the network has to be densified (e.g. to support mmWave) and in many cases MEC needs to be deployed (e.g. to support extreme low latencies), as described in Chapter 4. Both network densification and MEC will require additional investments. However, based on the latest GSMA reports, the operator mobile revenues have been increasing by only 2% year-on-year with 1% growth all the way to 2025. For example, in India, the cost of 1 GB of data to a consumer in 2019 was $0.26 (source: http://cable.co.uk) and it has reduced by 48% year-on-year on average since 2014, leading to a nine-fold increase in data consumption in the country. Furthermore, the cost of the spectrum has been around $0.02/MHz/pop/year in the developing countries, according to GSMA. For example, the 5G spectrum auction in India in June 2019 was priced at $7 billion/100 MHz. Relatively high spectrum fees encourage operators to use the spectrum more efficiently in order to control the deployment costs. Both factors increase pressure on operators to provide cost-efficient deployments for 5G. With RAN being the single largest contributor to mobile network CAPEX and OPEX, operators are looking for ways to reduce these costs and RAN virtualization is believed to be one of the key means to achieve that.

In one total cost of ownership (TCO) model analysis (MobileWorldLive 2018), the TCO shows a weighted total reduction of 37% deployment and operational costs over five years, derived from a 49% saving in CAPEX in the first year, and an annual 31% saving in OPEX over the full period. The CAPEX savings are mainly due to the virtualization of the Baseband Units (BBUs),[1] which enables the use of lower-cost commercial off-the-shelf (CoTS) hardware and pooled resources. OPEX savings are from reduced maintenance routines and power and operations savings resulting from the usage of centralized data centers that are cheaper and easier to access and operate. With less equipment needed to be located at cell sites, lease costs are reduced, while operators also benefit from faster upgrades and more flexible deployments.

Another key cost factor in a virtualized RAN deployment is the fronthaul capacity and latency requirements. Enabling multiple functional splits (described in detail in Chapter 4) gives the operator choices to virtualize portions of the RAN based on the transport network bandwidth and latency availability. In addition, compression schemes have been identified in open RAN standards (see Section 4.5 for details), which allow scaling the bandwidth with traffic. This enables operators to scale capacity dynamically where it is needed in the network during peak busy times or scale down during quiet times. This capability also enables operators to move capacity to where users are as they move around the network, thereby improving the network performance and reducing costs. Thus, RAN virtualization significantly reduces TCO compared with traditional implementations and brings the flexibility, scalability, and cost savings of network virtualization to the mobile network edge.

Even though RAN virtualization provides significant benefits, there is also a cost associated with it. General purpose hardware used in virtualized platforms often tends to consume more power, compared with a customized hardware specifically designed for the task. It is expected, though, that silicon process advancements and processor optimizations will be

1 BBUs may contain, for example, the functionality of gNB-CU and parts of the functionality of gNB-DU.

able to mitigate such disadvantages and eventually bring the gap in power consumption to acceptable levels.

### 6.2.3 Architecture Evolution Toward Virtualization

Traditionally, all networking components were deployed as standalone network nodes with dedicated hardware. Many of these had proprietary hardware components with limited interoperability. The hardware-based approach also made it difficult to scale up and down for different configurations and requirements, including future upgrades and optimizations. To address these issues operators have been pushing to allow for NG-RAN functional splits that enable most of the RAN functionality (e.g. gNB-CU and O-DU, as defined in Chapter 4) to be run in a virtualized manner in data centers, using an open hardware. This allows an open and software-based model where the system can scale with the workload leaving only the radios and a small amount of physical layer processing (e.g. O-DU, as defined in Chapter 4) to be implemented in hardware and deployed on site (see Section 4.5). Figure 6.2.1 shows the evolution of the RAN from dedicated hardware and fixed dimensioning to a scalable, flexible, and software-based model running in the cloud where the telco applications are now applied to the innovations happening in the IT industry.

### 6.2.4 Containers and Microservices

VMs and containers are two methods to realize virtualization on a compute platform. VMs, which were historically the first widely deployed virtualization method, allow an infrastructure platform to be abstracted from the associated hardware (i.e. virtualized). This is typically implemented using a hypervisor running on the host operating system (or sometimes on the real hardware, i.e. "bare metal"), which is a software component that manages

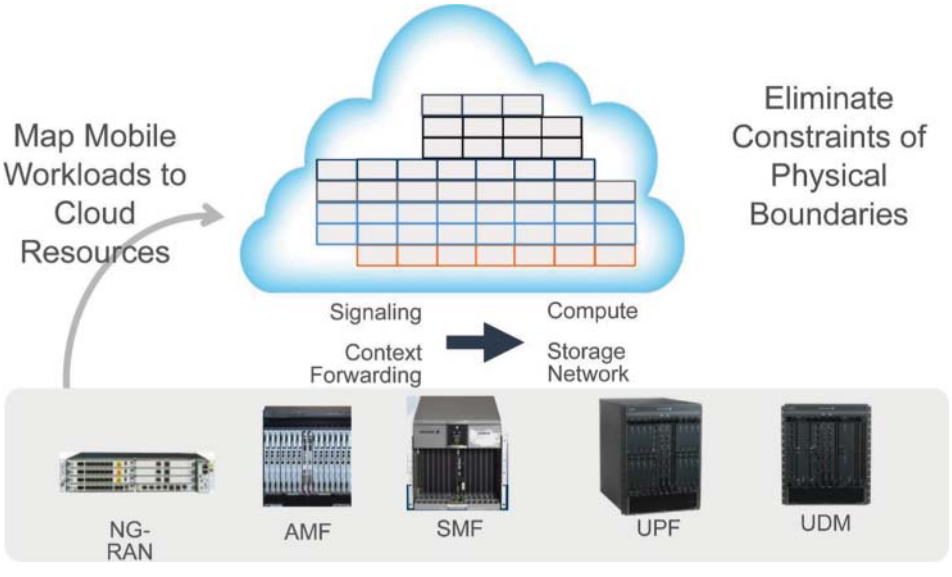



**Figure 6.2.1** Network evolution toward virtualized RAN.

VMs and presents them with a virtualized hardware, which is decoupled from the actual hardware. Thus, each application can run independently (i.e. in a separate operating system) from other applications on the server platform without impacting (or even the knowledge of) other applications running on the same hardware. In particular, this enables cloud providers to support many applications without constraining them to run on a particular set of resources.

The VM approach, which has been used by most virtualized platforms initially, is advantageous in that it does not require any changes in the operating system. Furthermore, it allows the running of different operating systems (e.g. Linux and Windows) on the same hardware. The disadvantage of this approach is that each VM runs both a full copy of an operating system (which can be different than that of the host) and a virtual copy of all the hardware that the operating system expects. This leads to significant overheads in terms of RAM and CPU, but provides extreme flexibility to the cloud provider.

As virtualization has been maturing and performance has become more of a priority, the container approach has gained popularity. A container is a self-contained piece of software that packages everything required to run the application. A container image includes the application code, libraries and other dependencies, additional processes (if needed), configuration, etc. In contrast to a VM, a container does not include a full copy of an operating system. Therefore, containers are more suitable to RAN applications that have strict timing requirements (especially in the case of gNB-distributed unit [DU]) and very conscious of memory, CPU, and storage load in order to optimize the cost of ownership and power consumption.

This is different compared with the VM approach of using a hierarchy of multiple levels of operating systems, which can cause significant variation in the real-time processing at the gNB-DU and increased use of resources. In contrast, with the container approach there is a single operating system that is enhanced to support containers and provides essentially the same level of resource isolation to an application as in the VM approach, but without the overhead of running a whole operating system (for each application) to achieve that. This is particularly suitable for RAN applications in order to enable better real-time support as well as reduced memory and CPU footprint.

With the advent of 5G, telecom architectures are evolving from virtual network functions (VNFs) into container network functions (CNFs), realizing the gains mentioned above. Figure 6.2.2 highlights the difference between two approaches, VMs and containers, using a Docker[2] framework as an example. As can be observed, the hypervisor and the guest operating system components (used in the VM architecture) are replaced by a thin Docker engine, eliminating the concept of the guest operating system entirely. This reduces the memory and the CPU footprint, making it lightweight and more efficient.

In order to make full use of the container platform, a good software design practice is to break the application into microservices. The key motivation of a microservices-based architecture is to partition a large and complex monolithic software application into multiple smaller self-contained modules (referred to as microservices), where each microservice can be designed, built, and tested individually. Thus, when an application has to be scaled

2 Docker is a popular container implementation. The software that hosts the containers is called Docker Engine. For more details refer to (Docker).

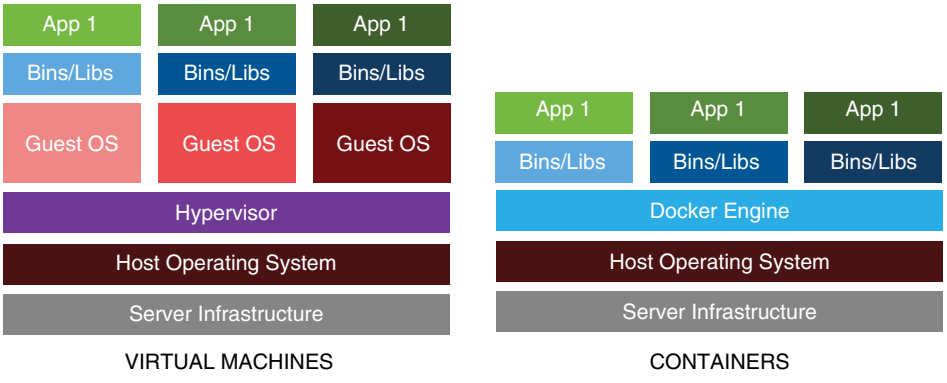



**Figure 6.2.2** Virtual machines versus containers.

based on workload, only the particular microservices that are impacted by the workload are scaled instead of the entire application. As shown in Figure 6.2.3, scaling a VM typically involves putting all the application functionalities into one server and replicating them across multiple servers. In contrast, in a microservice architecture, each functional component of an application is implemented as a service. Scaling is achieved by distributing these services across servers as needed.

Furthermore, in the microservice approach, each service works independently, that is, can function and fail (in the case of an error) without impacting the other services. Therefore, microservices allow faster recovery (often in hundreds of milliseconds) since only the failed services needs to be recovered, as opposed to the entire process. This is particularly useful in RAN applications to provide high availability (e.g. at the gNB-central unit [CU]) in the case of failures.

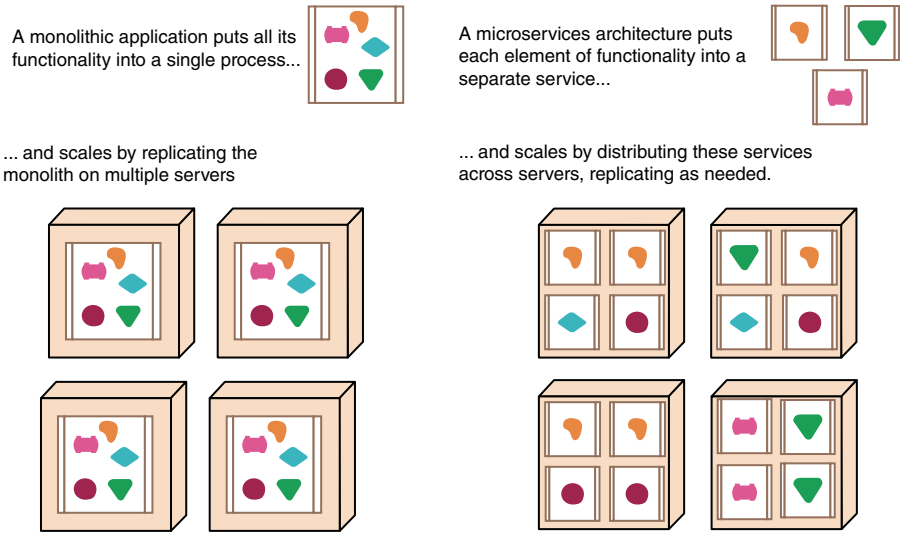



**Figure 6.2.3** Migrating to microservice-based architectures.

While breaking the NG-RAN functionality into microservices and running in multiple container instances help with scalability, an orchestration layer to manage these containers is needed in order to realize their full potential. The term "orchestration" refers to the automated configuration, coordination, and management of software. While the Docker platform provides an open standard for packaging and distributing containerized applications, it does not provide methods to scale, run, and monitor these applications. An alternative container platform is Kubernetes (K8S), which is a popular open source system for automating deployment, scaling, and management of containerized applications. In particular, this provides an abstraction layer for a cluster of machines to work together and behave as a single virtual machine, which is vital for a large-scale deployment of a virtualized environment.

As mentioned above, neither 3GPP nor O-RAN have explicitly standardized NG-RAN virtualization at the time of writing this book. However, the orchestration and automation framework being discussed in O-RAN (see Figure 6.2.4) does define how the design, inventory, policy, and configuration from the operators are pushed into the RAN configuration. The model also includes the RAN intelligent controller (RIC),[3] which can use artificial intelligence (AI)/machine learning (ML) methods to optimize RAN performance. When finalized, the platform will feature a set of functions and interfaces that enable optimization through policy-driven and closed loop automation. The RIC is also poised to create faster and more flexible service deployments and programmability within the RAN.

Open Networking Automation Platform (ONAP) is another open source effort to define a comprehensive management and network orchestration (MANO) framework for delivering software-defined networking (SDN) services that support both VNFs as well as CNFs. Both K8S and ONAP are supported by the Linux Foundation and can co-exist to support container modules on top of existing platforms in the operator network. ONAP (see Figure 6.2.5) provides a platform for real-time and policy-driven orchestration and automation of physical and virtual network functions (NFs) that will enable software, network, IT and cloud providers, and developers to rapidly automate new services and support complete lifecycle management ONAP interfaces with three major external subsystems:

- Operations Support System (OSS), Broadcast Satellite Services (BSS), big data analytics, and e-services applications through the northbound interface (the interface with higher-level layers of software);
- Virtualized infrastructure manager (VIM), the network function virtualization infrastructure (NFVI), and the SDN controller jointly constituting the network function virtualization (NFV) cloud through the southbound interface (the interface with lower-level layers of software);

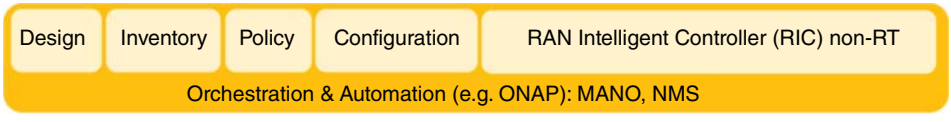



**Figure 6.2.4** Orchestration layer (Source: Reproduced by permission of © O-RAN).

3 RIC is currently being defined by O-RAN and is not described in detail in the present book.

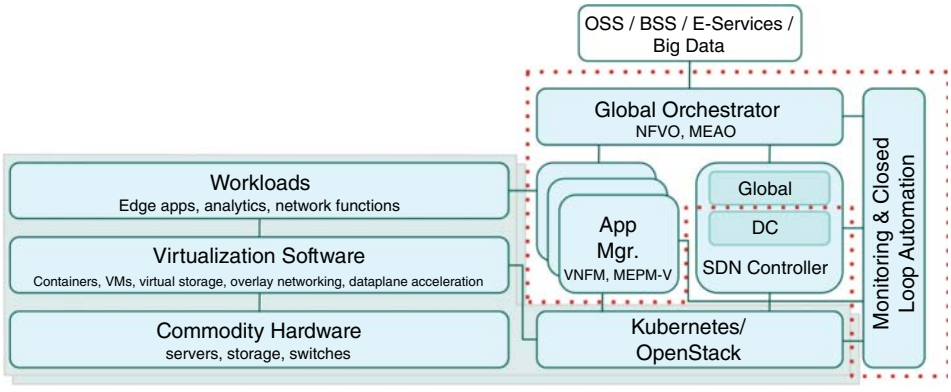



**Figure 6.2.5** ONAP framework (Source: Reproduced by permission of © ONAP).

- VNFs and analytics applications.

As of now, ONAP is primarily used in the core network; however, it may be extended to include at least parts of NG-RAN (e.g. virtualized gNB-CU).

### 6.2.5 NFV Evolution

Many operators have already deployed virtualized core networks for 4G and are now in different stages of their NFV evolution toward a cloud native virtualization platform for RAN. This process is expected to accelerate with 5G deployments. Generally, the following stages in deployment evolution toward virtualization can be envisioned:

1. Bare-metal (no virtualization).
   - In this stage, an application is run on bare-metal servers. The signaling and data plane are not decoupled. The system is often operated manually.
2. Virtualization.
   - In this stage, an application runs in a virtualized environment where software and hardware are decoupled. However, scaling capabilities are still manual and require reconfiguration when new network and computational resources are added.
3. Orchestration.
   - In this stage, orchestration is deployed to bring up the virtualized environments, with limited scaling, diagnosis, and healing.
4. Containers.
   - In this stage, the system is running on containers, where an application runs as a microservice where possible.
5. Automation.
   - In this stage, the system moves to a "Web-Scale IT"[4] platform, providing a fully automated system with the highest service assurance, analytics, and scaling.

4 Web-scale IT is an industry term referring to a modern networking architecture which is open, scalable, and implements intelligence in software.

Currently, most operators are in various stages of virtualized RAN deployments, depending on their vendors and ability to update their system rapidly. At least one "green field" operator is moving directly to a stage 4 cloud native platform, while most operators are at stage 1 for their RAN deployments as of 2019 but are rapidly moving to stages 4 and 5.

### 6.2.6 RAN Virtualization Platform

Before discussing how various split NG-RAN functions can be mapped to virtualization platform components we illustrate (Figure 6.2.6) the virtualization platform itself.

A RAN platform consists of hardware and software components that provide computing, networking, and storage capabilities to execute the NG-RAN functions. There are two types of functions – network functions (NFs), which run various parts of the protocol stack, and interworking functions (IWFs), which are used to communicate between the NFs and also with the external world. Where possible, standardized and open application programming interfaces (APIs) are used to communicate internally and externally to allow for multi-vendor support for the operators. As we show below, some NG-RAN functions can be implemented in software running on general purpose hardware, some functions are offloaded to hardware accelerators and/or graphic processing units (GPUs), some use FPGAs, and some use custom application-specific integrated circuit (ASIC).

Generally, a RAN virtualization platform needs to enable the following:

- Decoupling of hardware and software with acceleration as needed;
- Scaling of the solution and resource usage with workload;
- Support for various functional splits and deployment options in the RAN;
- Lifecycle management and network automation;
- Support for timing and synchronization in the RAN;
- High availability and redundancy in the network;
- A common platform that can be shared with the core network for edge applications.

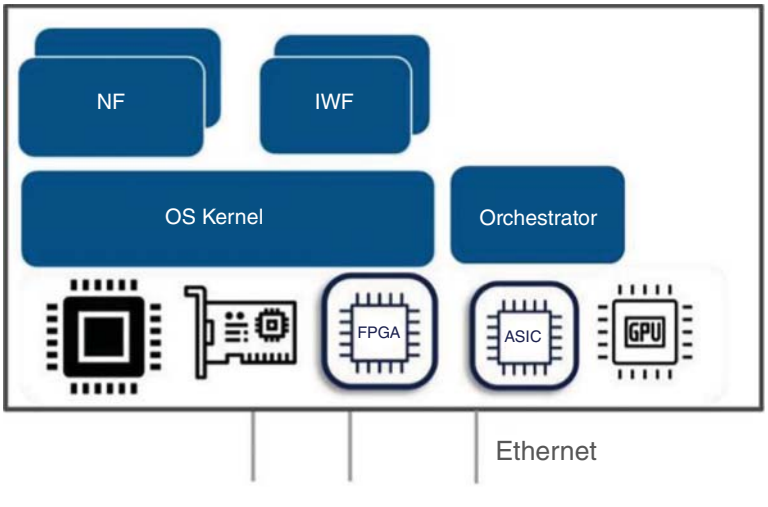



**Figure 6.2.6** RAN virtualization platform.

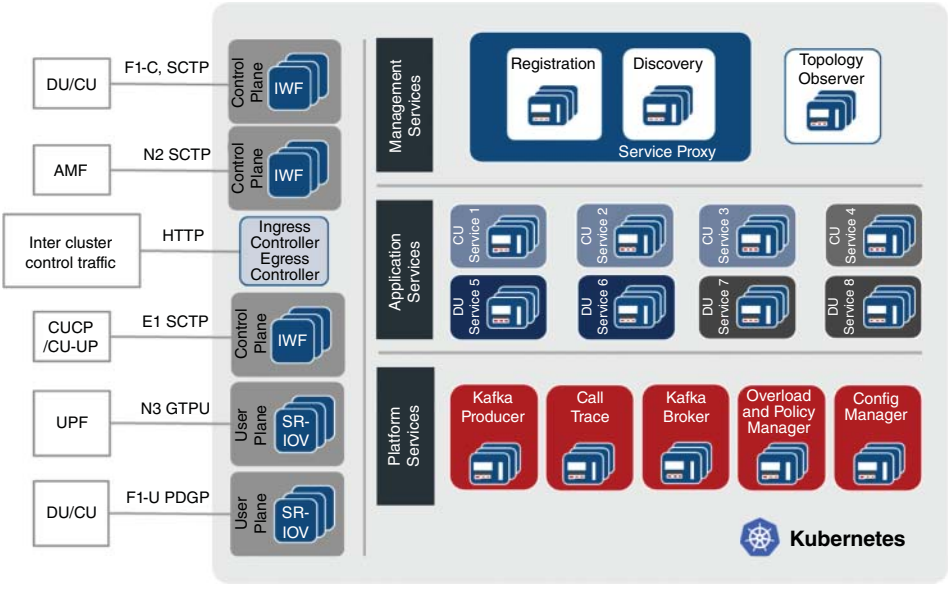



**Figure 6.2.7** Container-based RAN platform.

A detailed RAN virtualization example, with different functions mapped to different virtualized platform components, is shown in Figure 6.2.7. In this example, IWFs are used to convert Stream Control Transmission Protocols (SCTP) from 3GPP into HTTP2 commands, which is what is typically supported within a container platform. The ingress/egress controllers manage inter-cluster communications. The container platform comes with a set of platform services. These support features such as configuration and policy enforcement. The applications (e.g. NFs) run on top of the platform services. The orchestrator (e.g. K8S) provides management services such as registration and discovery. Each application microservice registers itself with service proxy indicating if the microservice has visibility externally to the K8S cluster. Service proxy provides application with services within the K8S cluster. The topology observer monitors the pod health and updates service proxy as needed. The IWF microservices terminate the RAN interface messages over Ethernet (SCTP/GTP) and translates to HTTP messages and vice-versa. HTTP is the native format for K8S.

#### 6.2.6.1 gNB-DU and gNB-CU Virtualization

For the purpose of illustrating RAN virtualization, we consider the NG-RAN to be implemented with functional splits as discussed in Chapter 4. For example, a NG-RAN with low-level split (Section 4.5) and high-level split (Section 4.2) may contain up to three network nodes:

- radio unit (RU), which consists of the radio and low PHY, connected to a
- DU consisting of the high-PHY, Medium Access Control (MAC), Radio Link Control (RLC), and scheduler, which in turn is connected to a

- CU where the Packet Data Convergence Protocol (PDCP), Service Data Adaptation Protocol (SDAP), and Radio Resource Control (RRC) layers are implemented.

Based on the use case and availability of transport bandwidth and latency, these nodes could be in a separate location or at the same location. Some or all of them may be virtualized and mapped to different components of a virtualized NG-RAN platform, as we describe below.

Figure 6.2.8 shows a representative gNB-DU container platform, which is assumed to be connected to RU, CU-CP, and CU-UP network nodes. The virtualized gNB-DU (vDU) CNF runs the high-PHY, MAC, RLC, and scheduler portions of the protocol stack in software. The operations, administration and management (OAM) interface is used for configuration and reporting/logging and connects to the network management system of the operator. Accelerators are used as required mostly for physical layer offload for functions such as forward error correction coding (FEC), which are compute intensive and latency sensitive.

The vDU microservice architecture needs to be designed carefully owing to strict timing requirements. The splitting of vDU functions into multiple microservices may impact the latency budget for features such as Hybrid ARQ (HARQ) and needs to be evaluated carefully based on processing power and the timing budget available. A timing/sync module (not shown) is also required to manage timing and synchronize the RU with the vDU. The timing/sync could come from GPS or from another timing/master clock that is transported to the vDU via IEEE 1588 (Precision Time Protocol [PTP]). PTP is also used to distribute timing from the vDU to all the RUs connected to the vDU.

Figure 6.2.9 shows a representative gNB-CU container platform, which is connected to a DU, user-plane function (UPF), and access and mobility management function (AMF). The CU container platform involves the control plane (CP) and user plane (UP) functions, which are separated. These may be on the same server or may be on different server locations via the E1 interface. The need for accelerators is reduced compared with the gNB-DU since these applications have less stringent real-time requirements. With the exception of encryption accelerators, it may be possible to implement a gNB-CU completely in software.

It is also more feasible to break the CU into multiple microservices due to lower real-time constraints, enabling better scaling for the CU.

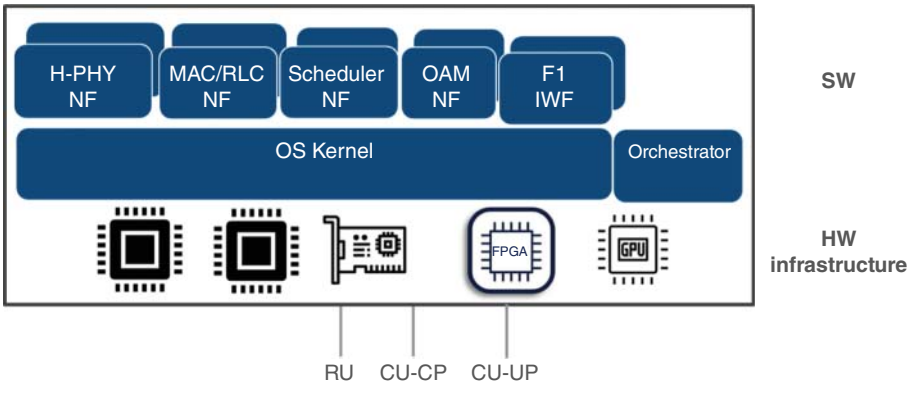



**Figure 6.2.8** DU container platform.

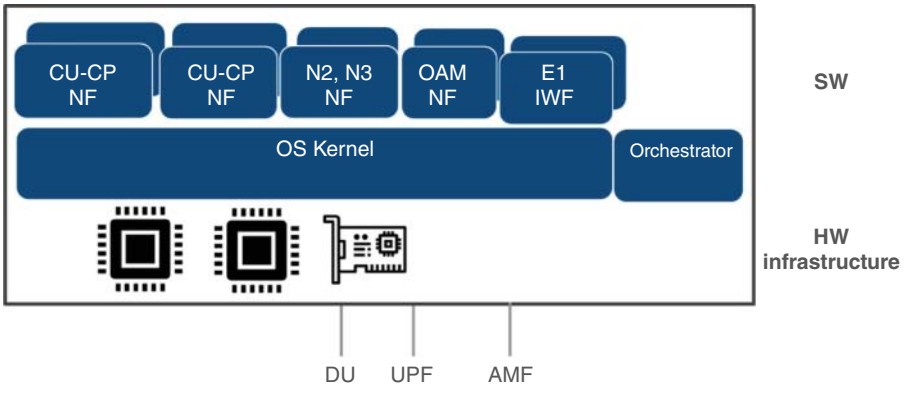



**Figure 6.2.9** CU container platform (assumes CU-CP and UP are co-located).

#### 6.2.6.2 Standardization of Orchestration and Cloudification in O-RAN

As we described in Chapter 4, one of the advantages of split NG-RAN architectures is that it makes it easier to deploy a gNB in a virtualized environment. For example, splitting gNB-DU functionality allows one to virtualize that network node as most of its functionality can be implemented in software. Having said that, while certain provisions have been made to enable that (e.g. a capability to add and remove SCTP associations on an NG-RAN control-plane interface to allow addition and removal of network and computational resources), 3GPP specifications do not explicitly define how NG-RAN can be virtualized.

Given the complexities of RAN virtualization and the need for inter-vendor interoperability, O-RAN Alliance has taken substantial efforts to provide a standardized model for cloudification of the RAN, and the orchestration of the resulting cloudified RAN, where:

1. Software and hardware are decoupled, which allows:
   - Multiple vendors of hardware platforms with CPU (e.g. x86, ARM), accelerators (e.g., FPGA, digital signal processor [DSP]/GPU), and abstraction layers which support the software providing RAN functionality.
   - A given hardware platform can support RAN software (e.g. for DU, CU, RU) from multiple vendors.
   - The mapping of software functions to hardware platforms can be done in multiple ways, and O-RAN is defining scenarios of interest, use cases within key scenarios, requirements, and reference designs.
2. Orchestration functionality is enabled for element discovery, lifecycle management, fault management (FM), and performance management (PM) for both physical network functions (PNFs) and VNFs that run on a cloudified RAN.
   - Management of a cloudified RAN requires a vendor-neutral approach to the above functions, as well as scale-out, slice management, fault tolerance, and hitless software upgrades.

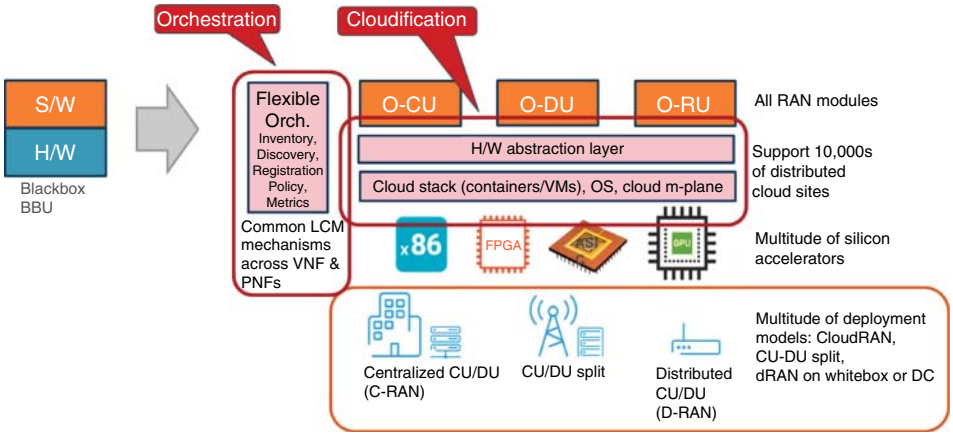



**Figure 6.2.10** O-RAN cloudification and orchestration work (Source: Reproduced by permission of © O-RAN).

- A common orchestration interface O1 is defined for the network management system (NMS) APIs across all the managed elements that support a disaggregated RAN.

Figure 6.2.10 shows the efforts being undertaken in O-RAN for cloudification and orchestration of the virtualized RAN.

## 6.2.7 Virtualization Challenges

While NG-RAN virtualization has significant benefits, as described above, there are certain challenges that need to be overcome in order to realize its full potential.

### 6.2.7.1 Accelerator Integration

Traditional virtualization does not include hardware accelerators, that is, it assumes that everything is implemented in software. However, with the compute-intensive workloads in 5G, accelerators may be needed to provide real-time support, reduce latency, and improve CPU utilization.

Most container frameworks provide a container network interface (CNI), which enables multiple containers to talk to FPGA via a single root input–output virtualization (SR-IOV) interface in a low latency and high throughput manner bypassing the kernel. One such example is the Multus plugin for the K8S framework.

Since there is a large range of deployment types for RAN with varying acceleration requirements, there needs to be a way to provide acceleration in a scalable manner. Accelerators typically operate in poll mode (as opposed to interrupt-driven software) and some of them need to respond in ten to hundreds of microseconds. The usage of multiple accelerators by multiple VNF instances in a single platform creates additional challenges, for example, in terms of accelerator sharing. Multiple accelerator models may need to be supported in the virtualized platform such as inline or lookaside models (see Figure 6.2.11). As of now, there is no uniform approach to the usage of hardware accelerators in virtualized RAN.

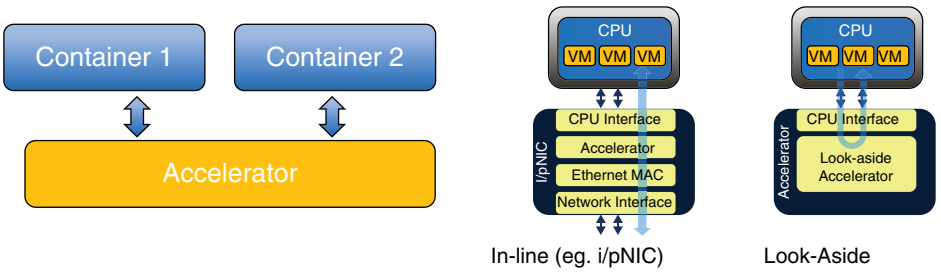



**Figure 6.2.11** Accelerator models for containers.

### 6.2.7.2 Timing and Synchronization

Most current (e.g. cloud and enterprise) workloads running on virtualized platform servers do not have as stringent real-time requirements as the RAN workloads. In order to for example run parts of the gNB-DU in a cloud infrastructure, accurate timing needs to be transported in the network. This is typically achieved via the PTP, which distributes timing from a centralized location to all radios where the timing accuracy of 100 ns or better is required. PTP provides frequency, phase, and timing accuracy. However, many traditional Network Interface Cards (NICs) do not support the hardware time stamping needed for PTP. In addition, many operators may also require synchronous Ethernet (SyncE) as a backup timing solution for holdover and to enable faster locking, which is again not very prevalent in data center solutions. The PTP accuracy can also vary on software-based implementations based on workload.

The O-RAN specifications define timing and synchronization models, protocols and topology (see Figure 6.2.12); however, as of now there is no well-defined approach for implementing this in a virtualized environment.

### 6.2.7.3 RAN Scaling with Workload

5G can be deployed in a wide variety of configurations, for example, in spectrum ranging from tens of MHz to hundreds of MHz, with number of antennae ranging from two to hundreds, in local (i.e. edge) sites deployments or cloud/data centers, etc. This requires RAN platforms to be sufficiently flexible and scalable.

Commercially available CoTS servers have a finite number of physical cores (e.g. <25 physical cores per socket is typical) – if multi-socket and core count is increased, CPU frequency starts going down. A virtualized gNB-DU should scale with hardware – most processing can be parallel – for example, processing different cells can go to different servers. While this works well for parallel workloads, it is not clear how to scale the edge hardware,

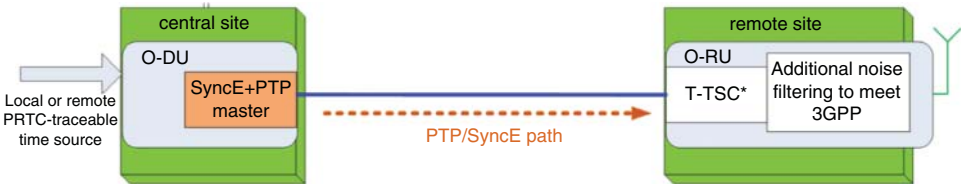



**Figure 6.2.12** Timing and synchronization (Source: Reproduced by permission of © O-RAN).

especially for the gNB-DU, which has strict real-time requirements, where the latency does not allow scaling it to off-chip to multiple servers.

Furthermore, the accelerators used for the L1 physical layer should scale as well; however, it may not be practical to put accelerators in all servers. Additionally, virtualization of accelerators with stringent requirements on latency and bandwidth is not well supported in CoTS platforms. In many cases, it is possible to scale the gNB-DU within a server only with some fixed mapping of radios to servers.

Another issue that has not been fully studied yet is live migration of the gNB-DU VNFs, which can be used for scalability as well as failure protection. At the gNB-CU, RAN scaling with workload is more feasible since the real-time requirements are less stringent and because most of the functionality can be implemented in software. However, as of now, there is no well-established approach to PHY (e.g. gNB-DU and/or RU) scaling in a virtualized environment.

#### 6.2.7.4 Inter-Process Communication

While containers with microservices enable significant flexibility and redundancy, this comes at the cost of an increased latency if a process is split among microservices. This additional overhead needs to be accounted for in the design of for example virtualized gNB-DU, with high data rates and stringent real-time requirements for HARQ. Therefore, inter-process communication can be a significant challenge in virtualized gNB implementations.

#### 6.2.7.5 Virtualization Overhead

All virtualization frameworks, that is both VMs and containers, have a certain processing overhead. The overhead is smaller in the case of containers, however it is not negligible in for example RUs and DUs, which are typically expected to be lightweight and low cost.

Furthermore, the choice of packet processing engines, for example, Open Virtual Switch (OVS) Data Plane Development Kit (DPDK) versus SR-IOV, can impact CPU resource utilizations as well. If OVS is used, the containers talk across systems using a virtual switch that is more flexible and scalable, while with SR-IOV the packets are sent to the container directly to provide higher throughput but may impact scalability.

If there are requirements for a VNF provider to support multiple platforms (e.g. both VMs and containers), there is an overhead involved in migrating applications, for example from VMs to containers. When an application is migrated to containers, there can be an additional latency overhead involved in splitting applications into multiple microservices. This latency overhead needs to be validated especially at gNB-DU where the timing is very critical.

#### 6.2.7.6 SCTP/GTP Interface Support

Most NG-RAN interfaces use SCTP (for control plane) or GTP protocol (for user plane). However, as of now all container frameworks only support HTTP. Therefore, an interworking function for protocol translation is required, which increases the processing overhead and latency, introducing an additional challenge.

#### 6.2.7.7 High Availability

High availability support also adds some overhead to provide redundancy which in turn impacts edge deployments in terms of cost, for example. However, high availability is not just about adding redundancy in, for example, hardware. It requires the system to perform even under hardware/software issues to provide 5 9's reliability. Containers are better suited for high availability, as they have faster platform orchestration support (e.g. auto spinning up new containers on failure) compared with VMs. Additionally, there are also geo-redundancy requirements in most Tier-1 operators. This requires careful design on the RAN to balance performance requirements for reliability versus increased amount of hardware for redundancy. This is further elaborated upon in Chapter 7.

#### 6.2.7.8 Power Consumption

Average power consumption for 2G–4G base station site is around 6 kW, which may raise to 10 kW at peak loads (Steven Carlini, 2019). This is expected to increase with 5G higher data rates, even though the average energy per bit will decrease. With increased densification of cell sites in 5G, power consumption is an important consideration with green initiatives from government and regulators.

The CoTS hardware (common in virtualization) is inherently less power-efficient compared with ASIC and dedicated hardware implementations. Furthermore, in order to meet stringent timing requirements, power-saving options are often disabled in hardware platform used for the virtualized NG-RAN. For example, enabling sleep mode in a CPU performing scheduling operations at sub-milliseconds is a challenge.

In order to make power-saving feasible, new basic input/output system (BIOS) and operating system settings need to be explored for virtualization of NG-RAN. Currently, all modern processors have power-saving features, which however have been designed for applications that do not have real-time requirements. For example the x86 Intel architecture supports P-state for dynamic power savings using Dynamic Voltage and Frequency Scaling (DVFS). For idle mode, there is C-state, which provides idle mode power savings including turning off the cache, phase-locked loop (PLL) flush, etc. However, it is not clear how the power-saving features mentioned above can be used in time-sensitive applications such as gNB-DU.

We expect that new power-saving features tailored for virtualizing time-sensitive applications will eventually emerge, alleviating this challenge to some extent.

#### 6.2.7.9 Distributed Cloud Deployments for RAN Nodes

Parts of the RAN application can be running in various locations based on the deployment. The choice of deployment locations can be based on latency and transport network bandwidth limitations, high availability requirements, geo-redundancy requirements, etc. The RAN applications could be hosted on a cell site, edge local clouds, or regional or hyperscale data centers. This implies the NFs and applications need to be supported on a distributed cloud platform model that is lightweight, flexible, and scalable for various applications. The platform should be able to coordinate across the distributed cloud to provide the containerized services for management and orchestration. As of now, there is no commonly established platform that addresses all these requirements.

### 6.2.8 Further Reading

While there is little literature available that explicitly addresses the RAN virtualization problem, there are a number of general virtualization-related resources available for further reading:

- ONAP documentation;
- Open Source MANO (OSM) documentation;
- Kubernetes container framework documentation;
- Docker container framework documentation.

## References

GSMA report (2018). Global Mobile Trends: What's driving the mobile industry?, September 2018.

White Paper (2018).The new mobile network economics, Why mobile operators are transforming to virtualized RAN for 5G, Mobile World Live. 2018

Webinar (2019). Building a better vRAN with Open Source Infrastructure, Intel Network Builders webcast, Paul Miller and Bejoy Pankajakshan.

ONAP (n.d.). Homepage. https://www.onap.org (accessed June 17, 2020).

ONAP demystified (n.d.). Automate Network Services with ONAP, Amar Kapadia.

Carlini, S. 2019, Massive 5G electricity costs are in focus ahead of the global build-out at the edge. Available at: https://blog.se.com/co-location/2019/11/11/massive-5g-electricity-costs-are-in-focus-ahead-of-the-global-build-out-at-the-edge (accessed June 17, 2020).

Docker (n.d.). Homepage. Available at: https://www.docker.com (accessed June 17, 2020).

Data Plane Development Kit (DPDK) (n.d.). Homepage. https://www.dpdk.org (accessed June 17, 2020).

## 6.3 Open Source

*Sasha Sirotkin*
*Intel Corporation, Israel*

Traditionally, RAN development has been driven by the standards (e.g. 3GPP) – that is, a standard is developed first, and then implementations from different vendors which adhere to that standard follow. That approach, if done right, ensures interoperability between different vendors, while also allowing vendor-specific enhancements.

An alternative approach that may lead to similar results is open source, where community members (vendors, academia, and volunteers) contribute source code to a common base, which everybody can then use to create products. These products would, at least in theory, be interoperable by virtue of the fact that they rely on the same code base (which may also include proprietary enhancements). Perhaps even more importantly, the open source approach helps reducing development and maintenance costs, as at least part of the cost is shared between multiple community members. This is generally in line with the desire of operators to lower NG-RAN deployment costs, and therefore we may see increased interest in open source as we move toward 5G.

The open source approach has been hugely successful in some fields, for example in the case of the Linux operating system, which is used on most enterprise servers, and which is also at the core of all Android-based mobile phones. Given the above, it is not surprising that eventually the idea to apply the same concept in telecom emerged. While there are many open source projects and initiatives for various components of a wireless network (e.g. core, management, etc.), in the present section we focus on open source RAN projects, specifically:

- Open Air Interface
- Telecom Infrastructure Project (TIP)
- O-RAN.

Before going into the details of RAN open source projects and related technologies, such as software-defined radio (SDR), it is important to discuss what is open source.

As the name suggests, open source refers to a product (which may be a pure software or device with software components) for which the end user is able to access (and most importantly – modify) the software. Initially the term was coined for software and, at later stages, it was also extended to include open source hardware. However, since for obvious reasons modifying software is substantially easier than hardware, most open source projects are software only.

There is a distinction to make between open source and free software. While open source software is often provided free of charge, there are also many commercial open source products. Furthermore, many closed source products are free (e.g. most of the Android apps and many "shareware" programs from the pre-mobile era).

On the other hand, the term free software is also often used in the context of "freedom," as in "free speech" as opposed to "free beer," to indicate that the main idea of free software is freedom not only to use it as one likes, but also to modify and extend it. This distinction between free and open source was perhaps one of the earliest controversies in this area

(e.g. between Richard Stallman[5] and Linus Torvalds[6]). Many other controversies followed, which however did not prevent open source projects from widespread adoption in servers, data centers, mobile phones, and many other appliances.

To illustrate the importance of open source, below we provide a very incomplete list of some of the most widely adopted open source implementations:

- Linux kernel, which is also what Android is based on.
- Linux operating system (including the kernel and the GNU software suite), e.g. RedHat and Ubuntu, which are what many cloud and enterprise servers use.
- LAMP (Linux, Apache, MySQL, PHP/Perl/Python), which is what the vast majority of small to medium websites run on.
- FreeBSD operating system, significant parts of which also power Apple macOS, iOS, and Sony Playstation.

Initially, the open source movement was primarily driven by volunteers; however, in recent years the vast majority of the open source code has been contributed by people employed by large corporations. Furthermore, while in the past the decision process was very much ad hoc (driven by community, with "maintainers" such as Linus Torvalds having the final decision power), it is now increasingly the case that the process is governed by organizations such as the Linux Foundation.[7]

Given the widespread adoption and huge success of some open source projects (e.g. Linux), it is natural to consider applying the same concept in the telecom industry. However, at the time of writing this book, most open source telecom projects focus on the core network and OAM, while open source RAN implementations appear to be in their infancy – that is, as of now there are no actual 3G, 4G, or 5G networks deployed based on open source implementations. Nevertheless, it is an interesting development that may become more important in the future.

### 6.3.1 Key Ideas

- Open source projects, e.g. Linux, have been tremendously successful in a number of domains: enterprise servers, cloud, and mobile phones to name a few. Inspired by that success, the telecom industry is interested in developing open source NG-RAN implementations.
- Open source software is usually distributed under a number of different open source licenses. These can be categorized as "permissive" and "copyleft." While most open source licenses do not allow patent royalties, the Open Air Interface (OAI) license does.
- Open source RAN is made possible through the implementation of an SDR concept. In an ideal SDR implementation, everything but the RF, A/D, and D/A is implemented in software, running on general purpose (typically x86) hardware.

---

5 Richard Stallman (rms), is a free software movement activist. He is the original author of the GNU Software project, which produced many software packages which are in wide use to this day, e.g. the GNU Compiler Collection (gcc).

6 Linus Torvalds is the original author of the Linux kernel.

7 Linux Foundation is a non-profit organization created to support and promote Linux. It is also involved with many other open source projects.

- A number of open source RAN projects exist, with the most advanced one being OAI, which is also the only open source project with an active 5G (i.e. NG-RAN) development. OAI is based on LTE Release-8, with select features from later releases, such as: LTE-M, NB-Internet of Things (IoT), and D2D. The code base contains the RAN, UE and EPC implementation. A 5G version of OAI is in development.
- Despite significant interest in open source, as of now there are no open source NG-RAN implementations yet that are close to being mature enough for commercial deployment.

### 6.3.2 Market Drivers

For operators, open source is yet another way to commoditize the RAN, with an ultimate goal of driving down CAPEX and OPEX and decreasing their dependence on a specific network equipment vendor.

For academia, open source RAN is a good opportunity to apply their research to what may eventually become an actual wireless network. Even if that promise does not materialize, open source implementations are perhaps one of the best ways for researchers in the field of wireless communications to test their ideas in the field.

While smaller network equipment vendors are likely to benefit from the availability of open source RAN implementation, which may shorten their time-to-market and generally decrease the entry barrier for them, it is not clear if there are benefits to be gained by the established market players. That being said, at least some of them are actively contributing to open source RAN projects.

### 6.3.3 Open Source License

As we mentioned above, different open source projects use different open source licenses, which in turn allow for different levels of freedom to use the software and different limitations imposed on its usage.

Perhaps not surprisingly, there are hundreds of open source licenses. The Open Source Initiative (OSI) is keeping a list of these, while the Free Software Foundation (FSF) maintains a list of what it considers free. These organizations are sometimes in disagreement about which licenses should be considered "open" and "free." Licenses that do appear on both lists are sometimes referred to as free and open source software (FOSS).

In addition to free vs. open, open source licenses are sometimes categorized as "permissive" or "copyleft." A permissive software license generally imposes few to no restrictions on the usage, modifications, and redistribution of the software. In particular, permissive licenses do not attempt to guarantee that future versions of the software will remain open and free. In contrast to the permissive license, a copyleft license grants a user full rights to modify and redistribute the software, with the limitation that any derivative work based on copyleft software must also be copyleft – that is open and free.

Here we elaborate on just a few of the most common open source licenses:

- Perhaps the most important (but not necessarily the most widespread as of now) open source license is the GNU General Public License (GPL) family. For example, GPLv2 is the license used by the Linux kernel. It is considered a "copyleft" open source license that

guarantees end users the freedom to run, study, share, and modify the software. However, it imposes a number of important restrictions; for example, that derivative work must be open source and distributed under the same license terms. It was initially written by Richard Stallman for his GNU project in 1989. The second, conceptually similar, version was released in 1991. However, after GPLv2 was released, several members of the open source community realized that GPLv2 can be exploited by allowing the inclusion of GPL-licensed software in custom hardware that prevents users from running modified versions of its software – the practice sometimes referred to as "tivoization" (Tivoization). GPLv3 was released to address that issue; however, it never became as popular as GPLv2, in particular because Linus Torvalds refused to adopt it for the Linux kernel. Once the most popular open source license, GPL has become much less widespread in recent years, perhaps in part due to the fact that most open source code is now written by people employed by large corporations, as opposed to volunteers.

- One of the earliest permissive licenses is the Massachusetts Institute of Technology (MIT) License, introduced circa 1990. Unfortunately, the term "MIT license" is somewhat ambiguous, as MIT used different flavors of that license at different times for different projects. It imposes very fewrestrictions and, as it is rather short, here we provide the full text of the one of the MIT license variants:

  > Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:
  > The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.
  > THE SOFTWARE IS PROVIDED "AS IS," WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

- The BSD License family is another example of a "permissive" license, conceptually similar to the MIT license. It was originally introduced for the BSD version of Unix.
- The Apache license is essentially similar to the permissive BSD license. It was written by the Apache Software Foundation (ASF) to be used for their extremely popular web server software. The language of the Apache license is somewhat more elaborate compared with BSD, making it more appealing to enterprises and therefore more popular. One of the important aspects of this license is that it ensures that people can use the software licensed under Apache terms without concerns for royalties.

- The OAI Public License, even though nowhere near as popular as the ones mentioned above, is important to mention as it represents a different type of open source license, which is incidentally used for one of the most popular open source RAN implementations. It is a modified version of the Apache license, with one significant difference – it allows contributing parties to charge royalties based on patents for commercial exploitation of the software. Therefore, some members of the open source community do not consider OAI as "true" open source. In particular, the OAI Public License contains the following terms not present in the Apache license:

> 3.2 Grant of Patent License for purposes other than study and research:
> For purposes other than study, testing and research, and subject to the terms and conditions of this License, You commit to be prepared to negotiate a non-exclusive, non-transferable, non-assignable license of Essential Patents with each Contributor and/or the Licensor on Fair, Reasonable and Non-Discriminatory ("FRAND") terms and conditions for the use of the Work or Contribution(s) incorporated within the Work.

### 6.3.4 Software-Defined Radio

One important difference between open source RAN and most other open source projects (e.g. Linux) is that RAN cannot be entirely implemented in software, as it has to interact with the physical world through radio. While the desire is to move as much functionality to software as possible, the radio hardware will always be a part of any wireless communications system. Ideally, such radio should be fully configurable by software, which is the concept of SDR, developed long before open source RAN was considered. Without SDR, neither virtual RAN (described in Section 6.2) nor open source RAN would have been possible.

Here we illustrate the concept by using the ideal SDR scheme, in which everything but the RF and A/D (or D/A) is realized in software, as shown in Figure 6.3.1.

In the ideal SDR receiver scheme an A/D converter is attached to an antenna. The output from A/D is sent to the digital signal processor, which runs the software implementing the rest of the receiver functionality. An ideal SDR transmitter is similar, in which a digital signal processor generates a stream of bits which are sent to a D/A converter connected to a radio antenna.

In practice, on the one hand the ideal scheme outlined above is currently not fully feasible due to hardware limitations. However, on the other hand, in most but not all baseband implementations significant parts of the functionality are implemented in software, which makes them highly configurable and at least partially aligned with the SDR concept.

Regardless of how close we are to the ideal SDR implementation, what is important is the availability of low-cost radio modules that have sufficiently large parts of their functionality implemented in software and are therefore highly configurable, making them a good fit for an open source RAN implementation.

Incidentally, the SDR concept lands perfectly in the virtualization paradigm (see Section 6.2), as the software component of SDR can be relatively easily virtualized.

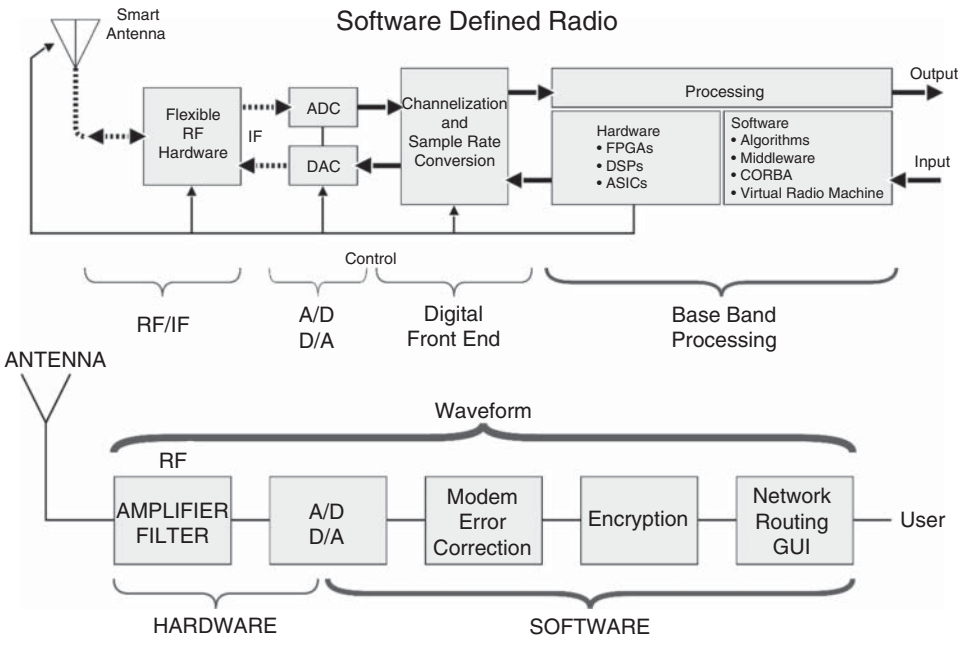



**Figure 6.3.1** Ideal SDR receiver and transmitter.

An important open source project to mention in the context of SDR is the GNU Radio, which is a free software (distributed under the GPL license) development toolkit that provides multiple building blocks that can be used to implement an SDR.

### 6.3.5 Open Source RAN Projects

All the RAN open source projects described below are still in their infancy – while some are more advanced than others, at the time of writing of this book none of them can be used to create a commercial 5G network. Nevertheless, some of them look promising and, in a few years, they may become a viable alternative to traditional commercial offerings. As of now, they provide a good platform for research and experimentation.

#### 6.3.5.1 srsLTE

srsLTE is an open source LTE implementation, produced by a commercial entity, however released under a permissive GNU Affero General Public License. It contains the following components:

- srsUE – an open source UE implementation;
- srsENB – an open source eNB implementation;
- srsEPC – a lightweight open source EPC implementation with mobility management entity (MME), Home Subscriber Server (HSS), Serving Gateway (S-GW), and Packet Data Network Gateway (P-GW).

The srsENB is based on Release-10 3GPP specification, implementing some basic features of an eNB, with S1 interface connectivity to EPC.

Since as of now it is unclear whether a 5G implementation will emerge based on this project, this is mentioned as background information about open source activities related to RAN.

#### 6.3.5.2 OpenLTE

OpenLTE is yet another open source implementation eNB, with a "built-in" simple EPC. Unfortunately, there has been little to no activity in that project since 2017 and it appears to be unlikely that new versions will be produced, let alone a 5G implementation.

#### 6.3.5.3 OpenBTS

OpenBTS is an open source 2G (GSM) base station implementation developed by Range Networks. It is distributed under the copyleft GNU Affero General Public License v3 and as such can be considered as a free software. It is perhaps the only open source cellular base station implementation that has been successfully deployed in the field.

That being said, OpenBTS is limited to 2G and as of now there appears to be no intention to evolve it into more advanced 3GPP releases. Therefore, it is also mentioned here for background information only.

#### 6.3.5.4 Open Air Interface

OAI is currently the most advanced open source RAN implementation and also the only one where active 5G development is taking place. It is based on LTE Release-8, with select features from later releases, such as: LTE-M, NB-IoT, and D2D.

From the architecture point of view, OAI contains both RAN (E-UTRAN) and core (EPC), which can be deployed together in one platform or on separate network nodes. Furthermore, the OAI project also has an open source UE implementation. The OAI RAN implements the PHY and all the air interface protocol stack layers (as described in Chapter 3); however, only the essential features required to build a functioning network have been implemented so far. Furthermore, the OAI RAN implements the X2 and S1 interface, once again with a limited set of functionalities – for example, X2 handovers are supported, but scheduling and resource coordination via X2 are not.

From the internal RAN architecture standpoint, the OAI RAN implements a number of split options (see Chapter 4 for details), as shown in Figure 6.3.2.

The splits shown in Figure 6.3.2 do not follow exactly the specifications and standardized architectures described in Chapter 4, but are rather "inspired" by these:

- The IF1 interface is conceptually similar to the 3GPP F1 (3GPP TS 38.470) interface (described in Section 4.2). Note that the 3GPP F1 interface currently only supports 5G, whereas OAI IF1 supports only LTE. In particular, this means that there are major differences between 3GPP F1 and OAI IF1. That being said, once 3GPP has specified the equivalent of F1 for LTE (referred to as W1), it is likely that OAI IF1 implementation will align to that W1 spec. (3GPP TS 37.470). Moreover, when OAI finally supports 5G, it is likely that OAI 5G IF1 will follow the relevant 3GPP specification.
- IF2 is similar to the Small Cell Forum FAPI interface (described in Section 4.6).
- IF4.5 is similar to the O-RAN low-level interface (described in Section 4.5).
- IF5 is similar to Common Public Radio Interface (CPRI).

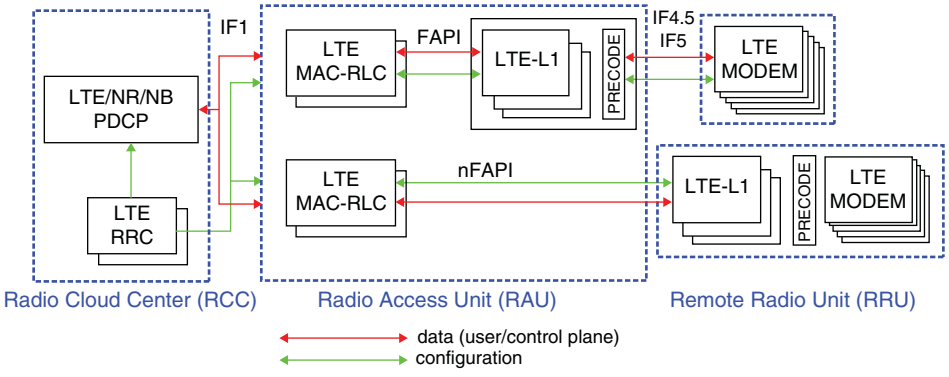



**Figure 6.3.2** OAI RAN architecture.

As mentioned above, one should not assume that OAI interfaces IF1, IF2, IF4.5, and IF5 closely adhere to the relevant standards. However, it is expected that over time OAI interfaces implementations may become more closely aligned to the respective 3GPP, O-RAN, and Small Cell Forum standards.

OAI has a number of options primarily designed to help development of new features, such as a simulator and "noS1" mode, in which a [preconfigured] RAN can be deployed without an EPC, which makes it a good platform for academic research.

At the time of writing this book, OAI RAN supports LTE only; however, 5G development is underway. It is likely that at least basic functionality of EN-DC (see Section 4.3) will be supported by OAI in 2020.

From the hardware perspective, parts of OAI implementation run on a generic CoTS x86 platform, which however also requires an SDR. A number of SDR hardware platforms are supported by OAI, such as: ExpressMIMO2, USRP, and LimeSDR.

It is important to mention that OAI source code is distributed under the OAI license, which is free for research; however, it contains clauses allowing intellectual property rights (IPR) holders who contributed to the project to charge royalties on OAI software usage.

#### 6.3.5.5 TIP

TIP has a number of projects (e.g. OpenCellular) that may produce software and hardware components, potentially to be released and licensed as open source. However, at the time of writing this book, there are no active open source RAN projects in TIP.

It must be noted that TIPs goal is to create a wide ecosystem of RAN solutions and component vendors, and there is no requirement for the implementation to be open source. Therefore, it is unlikely that a fully open source RAN implementation will emerge in TIP.

#### 6.3.5.6 O-RAN

The O-RAN Alliance, jointly with the Linux Foundation, have created the O-RAN Software Community (SC) collaboration project, with the goal of creating an open source 5G RAN implementation, to be released under the OSI-compliant Apache 2 open source license. The project is in the early stages and as of now no source code has been contributed yet.

### 6.3.6 Summary

In this section, we showed that there are some players in the telecom ecosystem attempting to replicate the success of open source projects (e.g. Linux) in the enterprise and cloud. We introduced a number of RAN open source projects that are at various stages of development with various levels of activity. The most advanced one is the OAI, which is also the only one with active 5G development. One must note that OAI is released under a special "open source license," which allows patent holders to charge royalties on commercial usage of OAI RAN.

Furthermore, several organizations (e.g. O-RAN Alliance and TIP) have active programs to develop open source 5G RAN, which are however yet to produce a working code.

We expect open source to become more important in the future; however, it is not clear whether actual commercial open source RAN implementation will eventually emerge or if the open source activity will remain contained to academia and research.

## References

GNU General Public License, version 2 (1991). GNU General Public License, version 2. Available at: https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html (accessed June 17, 2020).

Tivoization (n.d.). Tivoization. Available at: https://en.wikipedia.org/wiki/Tivoization (accessed June 17, 2020).

OAI Public License V1.1 (n.d.). OAI Public License V1.1. Available at: https://www.openairinterface.org/?page_id=698 (accessed June 17, 2020).

OpenCellular (n.d.). OpenCellular. Available at: https://telecominfraproject.com/opencellular (accessed June 17, 2020).

O-RAN Software Community (SC) (n.d.). Homepage. Available at: https://o-ran-sc.org (accessed June 17, 2020).

OpenLTE (n.d.). Homepage. Available at: http://openlte.sourceforge.net (accessed June 17, 2020).

srsLTE (n.d.). Homepage. Available at: https://www.srslte.com (accessed June 17, 2020).

Software Radio Systems (n.d.). Homepage. Available at: https://www.softwareradiosystems.com (accessed June 17, 2020).

OpenBTS (n.d.). Homepage. Available at: https://github.com/RangeNetworks/dev (accessed June 17, 2020).

Range Networks (n.d.). Homepage. Available at: https://rangenetworks.com (accessed June 17, 2020).

GNU Radio (n.d.). Homepage. Available at: https://www.gnuradio.org (accessed June 17, 2020).

Linux Foundation (n.d.). Homepage. Available at: https://www.linuxfoundation.org (accessed June 17, 2020).

## 6.4 Multi-Access Edge Computing

*Miltiadis Filippou[1] and Dario Sabella[2]*

[1] *Intel Germany GmbH, Germany*

[2] *Intel Corporation, Germany*

As mentioned above, 5G communication systems need to satisfy an increasing demand for data traffic communication, by connecting a vastly growing number of devices and heterogeneous traffic flows, with more stringent Quality of Service (QoS) requirements (Cisco17, NGMN15, ITU-R15). These challenges are in turn translated into a multitude of requirements on the capabilities of new communication and data processing systems, starting from early 5G deployments and expanding beyond current systems, since the trend is not going to stop over the years. To overcome these issues, not only wireless network performance improvements are needed, but also enhancements of processing systems.

Edge computing is commonly recognized as a key ingredient of future 5G systems (HPS+15) (TSM+17) that will provide operators and infrastructure owners the flexibility and reconfigurability required to satisfy the ever-increasing traffic demand and the related tight QoS requirements. More concretely the main advantages of edge computing are:

- Low latency communication due to the data proximity to end users;
- The availability of real-time network information, such as radio conditions and network statistics.

This technology is known as multi-access edge computing (MEC), which is standardized by the European Telecommunications Standards Institute (ETSI).

As of now, ETSI MEC is a leading international standard available for edge computing. MEC covers many 5G vertical market segments, such as the automotive, industrial automation, multimedia and entertainment, smart energy, smart transportation domains, etc. Some of the early MEC implementations were proprietary. Eventually, the ETSI MEC standard was developed to allow interoperable MEC deployments. Even though MEC is commonly associated with 5G, the MEC standard is access-agnostic, allowing deployment independent from the underlying RAN. MEC can therefore be supported with LTE, NR, and other non-3GPP access technologies such as Wi-Fi and fixed networks. When it comes to MEC deployment in cellular networks, certain provisions have been made in 3GPP specifications to allow the edge computing concept; however, the rest was left for ETSI MEC to specify.

In a nutshell, MEC provides an environment in proximity to an end user, in which a server application, which would otherwise reside in the remote cloud, can run. This environment provides ultra-low latency, high bandwidth, and access to added-value information through MEC APIs; for example, real-time access network information, context information, location awareness, etc. MEC is poised to open the cloud infrastructure to operators, service providers, and third parties (i.e. application developers and content providers), helping to meet the demanding QoS requirements of new 5G systems.

In this section, we present some aspects of the MEC technology and explain its relevance to the 5G architecture.

### 6.4.1 Key Ideas

In this section we describe the ETSI MEC platform as it is currently defined and a number of potential enhancements being considered, which may become part of MEC in the future.

- For end users, MEC can substantially improve latency, which is crucial for ultra-reliable low latency communication (URLLC) applications and new emerging services, such as augmented reality (AR). For mobile network operators (MNOs), MEC provides an opportunity to open up their networks to over-the-top service providers, creating a new business opportunity.
- MEC system architecture is defined in two flavors: standalone and NFV-based. Regardless of the flavor, a MEC host consists of virtualization infrastructure, a MEC platform, and a number of MEC applications. All entities in the MEC architecture are connected via standardized reference points.
- ETSI MEC defines a number of APIs, which can be categorized as: application enablement, service-related, and management- and orchestration-related. Furthermore, a set of RESTful APIs provide access to radio network Information, location information, UE identity information, vehicular-to-everything (V2X) related information, and others.
- Even though the standard does not explicitly define how a MEC can be deployed in a 3GPP network, one can assume that the MEC data plane shall be connected to a UPF (and therefore, from a purely formal viewpoint, in such architecture MEC is not part of NG-RAN). Application function (AF) in 3GPP architecture can be mapped to the MEC platform in ETSI MEC architecture.
- Inter-MEC system information exchange can be beneficial, to enable for example V2X applications, as V2X services in a single area are likely to be provided by multiple operators. A hierarchical framework is being considered for service discovery and consumption of inter-system communications at the following levels: between MEC orchestrators (used e.g. for MEC system discover), between MEC platform managers (used e.g. for MEC platform discovery), and between MEC platforms (used e.g. to exchange information between services running on different MEC systems).
- To ensure proper selection of the services to be consumed by a MEC application and to help the orchestrator to make decisions on the MEC application requesting said services relocation, a new message protocol is being considered. The proposed performance-centric design through edge host zoning enables flexible MEC service consumption in multi-vendor environments, including the case of inter-MEC system deployments.
- To facilitate MEC deployments in automotive scenarios, which are characterized by high mobility and hard-to-predict changes in network load and radio conditions, a solution based on cooperative information reporting and route-specific predictions is being considered. It utilizes the concept of radio measurement information partitioning in the MEC based on routes to obtain journey-specific QoS predictions.

### 6.4.2 Market Drivers

Most of the work on latency reduction (in research, standardization, and implementation) so far has been focused on air interface improvements, while the "end-to-end" (E2E) aspect has often been overlooked. Although air interface latency reduction is very important,

improving true system-level performance is only possible with the introduction of edge computing, as the application level communication endpoint is on the edge, thus providing benefits for all services requiring extremely low latencies, e.g. AR, virtual reality (VR), industry 4.0, etc. MEC is commonly recognized as the main way to ensure E2E low latency in communication systems.

However, latency is not the only key performance indicator (KPI) targeted by MEC. Client application offloading (i.e. from UE to MEC) is another example where moving processing load from a UE to the edge can potentially increase battery life at the terminal side. While such a scenario, which is unrealistic with cloud deployments, becomes possible with MEC, it comes at the cost of increased complexity and latency.

From a deployment point of view, MEC provides an opportunity for operators and infrastructure owners to expand their business and offer edge hosting services to application vendors and cloud service providers, thus opening up new markets and business models. Most of these stakeholders in fact recognize that the 5G market will not only bring value due to enhanced connectivity and network performance, but also will provide some suitable means to offer new revenue streams. According to some estimations (Ericsson5G), edge computing is opening up 25% of the total 5G market potential.

From a market perspective, vertical market stakeholders (e.g. from the automotive domain) also recognize the importance of edge computing as a natural solution for interoperability of the data exchange in multi-MNO domains. In fact, considering the case of the automotive market, car makers need to offer V2X services where the connectivity is not limited to a single operator, but also capable of including cars connected to different networks, or moving between areas covered by different MNOs (e.g. country boundaries).

Finally, the possibility to host an edge computing environment comes with the advantage of the ability to collect and process local data and context information, and offer them as added-value services that could be exposed as edge APIs to application developers, service providers, and end users.

### 6.4.3 MEC Standard

The MEC reference system architecture is defined by ETSI in a couple of "flavors":

- Standalone architecture that does not require NFV-based infrastructure.
- "MEC-in-NFV" architecture, where all MEC entities are placed together with their respective NFV elements, according to the ETSI NFV framework.

Below we explain both MEC architecture variants as defined in the ETSI MEC specifications (including standards related to the MEC platform and to MEC APIs).

#### 6.4.3.1 ETSI MEC System Architecture

The "standalone" variant of the ETSI MEC reference architecture is depicted in Figure 6.4.1. This architecture is divided into two layers: the MEC host level and the MEC system level. The first one is composed of several MEC hosts connected via the Mp3 reference point. In the upper (MEC system) level an orchestrator and OSS are connected to the external world (and device applications) through a proxy.

Every MEC system consists of the following three functional elements:

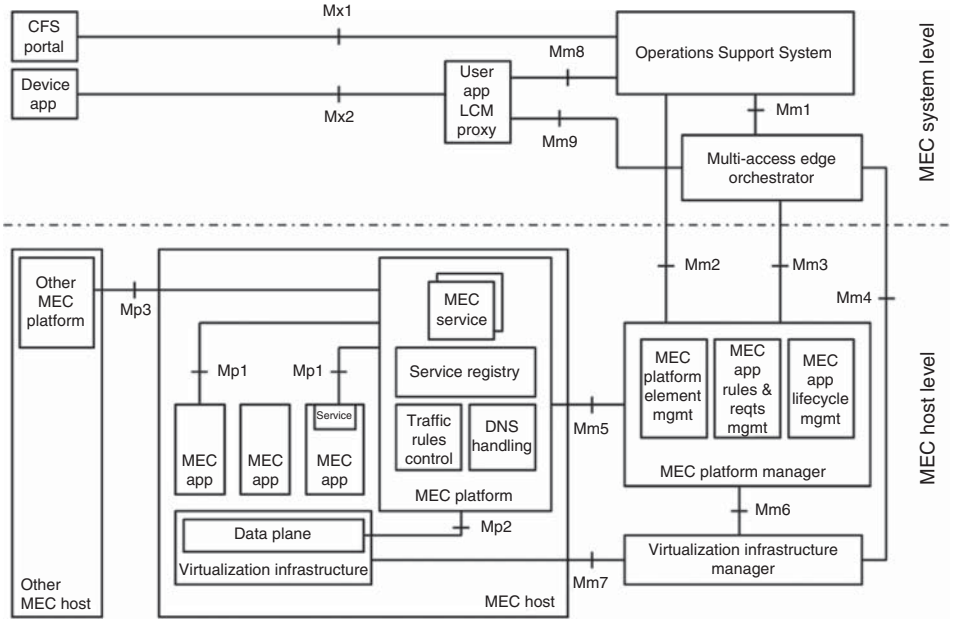



**Figure 6.4.1** Standalone variant of the ETSI MEC reference architecture (Source: Reproduced with permission from © ETSI).

- A MEC host, which provides storage, network, and other resources to a MEC platform.
- A MEC platform, which provides the environment where MEC applications can run.
- MEC applications, developed either by a MEC vendor or a third party, which provide some useful services.

There is an expectation that a market for MEC application will emerge, where multiple vendors will provide multiple MEC applications optimized to run in the edge. It remains to be seen whether the concept becomes popular.

Each MEC host is composed of virtualization infrastructure, a MEC platform, and a number of MEC applications. The virtualization infrastructure includes a data plane, which is in charge of traffic routing to MEC applications, based on configuration provided by the MEC platform at the control plane via the Mp2 reference point. The MEC platform also includes a number of MEC services, accessible to MEC applications through APIs via the Mp1 reference point. A MEC application itself can produce a service that could be exposed through Mp1 to other applications (via the Service Registry in the MEC platform), which is expected to help creating a MEC applications market. Mechanisms defined in MEC enable third parties to build their services (e.g. by exploiting location- and context-aware data available at the edge) and expose them to application developers, who could in turn use them to build new innovative services for end users.

As an important note, the MEC architecture standard (even in its "standalone" variant) has been conceived with NFV principles and definitions in mind. In order to enable a smooth MEC–NFV alignment, the following principles have been followed:

- MEC uses a virtualization platform for running applications at the network edge.

The NFV infrastructure utilized by MEC may be dedicated to MEC or shared with other NFs or applications.

- MEC uses the NFV infrastructure management entity where possible.

The two architectures (standalone and NFV-based) depicted in Figures 6.4.1 and 6.4.2 are similar; however, there are some differences, which are introduced in the "MEC in NFV" variant, specifically:

- The Mm3* reference point between MEC application orchestrator (MEAO) and MEC platform manager – NFV (MEPM-V) is based on the Mm3 reference point.
- The Mp1 reference point between an MEC application and the MEC platform is optional for the MEC application, unless it is an application that provides and/or consumes a MEC service.
- The MEC orchestrator (MEO), as defined in the MEC reference architecture (ETSI GS MEC 003), is replaced by a MEAO that uses the network function virtualization orchestrator (NFVO) for resource orchestration and for orchestration of the set of MEC app VNFs as one or more NFV network services.
- The MEPM, as defined in the MEC reference architecture (ETSI GS MEC 003), is replaced by an MEPM-V that delegates the life cycle management (LCM) part to one or more VNFM(s).

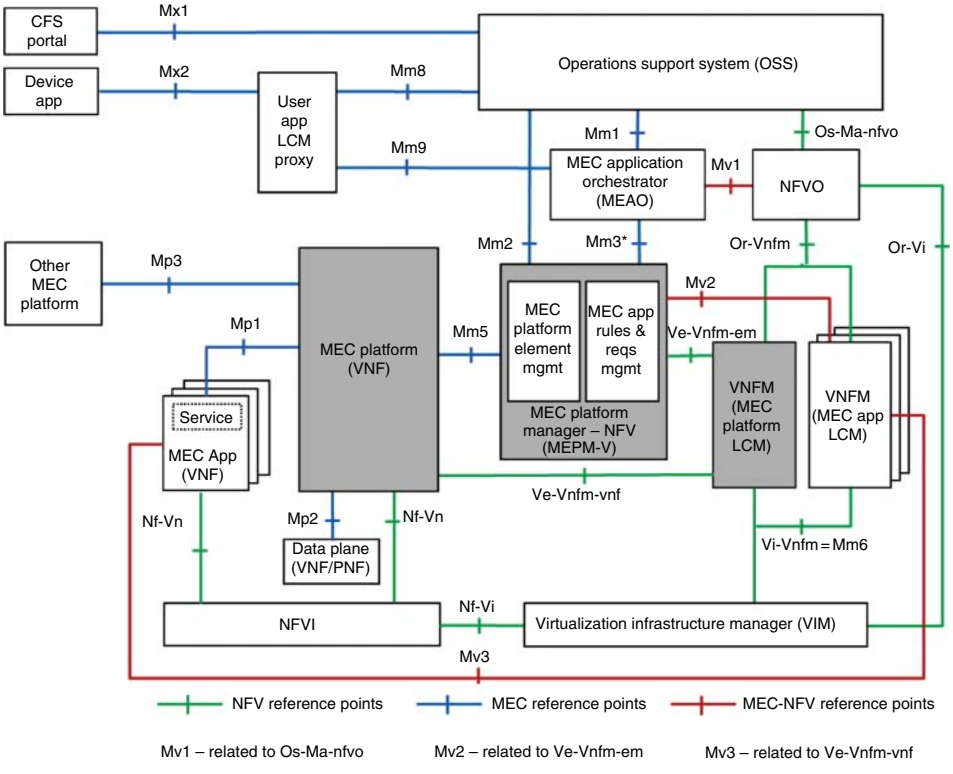



**Figure 6.4.2** MEC reference architecture: variant for MEC in NFV (Source: Reproduced with permission from © ETSI).

#### 6.4.3.2 ETSI MEC APIs

Having use case-driven requirements as a starting point (MEC002), the ETSI MEC Industry Specification Group (ISG) has defined a reference architecture in Group Specification MEC 003 (MEC003) and a set of APIs for key MEC interfaces. These include specifications related to the essential functionality of:

- Application enablement platform (API framework) (MEC009);
- Specific service-related APIs;
- Management and orchestration-related APIs.

Such APIs are designed to be application-developer-friendly and easy to implement so as to stimulate innovation and foster the development of different applications (GSS+18, GCA17) and ultimately the emergence of a MEC application market, so that it would be possible to deploy MEC applications from different vendors on MEC platforms from other vendors.

Furthermore, ETSI ISG MEC specifies a set of RESTful APIs as standardized interfaces, which can be used by application developers, in order to access radio network information (RNI API) (MEC012), location information (Location API) (MEC013), UE identity (UE identity API) (MEC014), as well as information for bandwidth management (MEC015) and other types of data pre-processed either by the MEC platform or by instantiated MEC applications. In particular, thanks to the RNI API, context information from the RAN can be provided to user-level applications or other services for network performance and QoS improvements. It is worth mentioning that while RNI provides a standardized API to an application to access RAN-related information, the retrieval of this information from RAN is not standardized and is left for vendor-specific implementations.

Recently, ETSI ISG MEC has expanded the scope of its activities to include additional access technologies besides cellular, as well as support for IoT deployments with low-energy support (ZZM+16) and connected cars. To reflect this change the term MEC was updated from mobile edge computing to *multi-access edge computing*. This is aimed at strengthening the engagement with original equipment manufacturers (OEMs) and service providers as key stakeholders exploiting MEC for their added-value product propositions. As an example of such activity, ETSI ISG MEC is introducing a "V2X Information Service API" to assist the MEC system in exposing information to applications allowing developers, car OEMs, and their suppliers to implement intelligent transportation system (ITS) services in an interoperable way, across different access networks, owned and managed by different MNOs and vendors (GSS+18).

It is important to mention that among the plethora of standards developed by ETSI MEC, the API standards mentioned above are arguably the most important ones. That is because these standards define interfaces toward applications which are likely to be developed by a third party and therefore having a standardized API is crucial to ensure interoperability.

Below we use the Location API standard (MEC013) as an example to illustrate the details of ETSI MEC APIs.

#### 6.4.3.3 Location API

The purpose of the MEC Location API is to allow the MEC platform and the authorized MEC applications to access a UE location. With this information, it is possible to track a

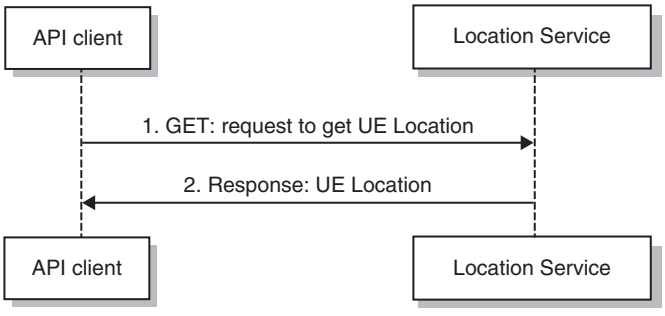



**Figure 6.4.3** UE location lookup procedure (Source: Reproduced with permission from © ETSI).

UE, provide location-specific services, etc. Furthermore, the service supports anonymous (i.e. without a UE identification) location statistics collection.

A MEC application can access location information using either:

- Location lookup procedures, i.e. "one-shot" location information request; or
- Location subscription procedures, which can provide periodic location information reporting to subscribed MEC applications.

The UE location lookup procedure is illustrated by Figure 6.4.3.

1. When a MEC application needs a UE location, it sends the "request to get UE Location" information message to the Location Service for one or multiple UEs (identified e.g. by IP address).
2. The Location Service responds with a "UE Location" message, carrying the requested location information of the UE(s).

Alternatively, if a MEC application is interested in receiving periodic UE location reports, it may use the UE location subscribe procedure, illustrated in the Figure 6.4.4.

1. If a MEC application needs periodic UE location reports, the application sends the "create UE Location subscription" message to the Location Service, indicating UE(s).
2. The Location Service responds with a resource URI containing the subscription ID.

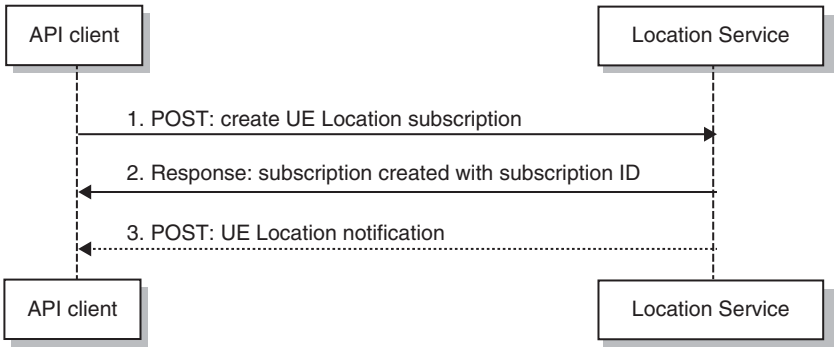



**Figure 6.4.4** UE location subscribe procedure (Source: Reproduced with permission from © ETSI).

3. The Location Service periodically reports the requested UE(s) location.

ETSI MEC API's data model is based on JavaScript Object Notation (JSON), which uses HTTP over TLS, also known as HTTPS, defined in IETF RFC 2818. Usage of HTTP without TLS is not recommended.

### 6.4.4 ETSI MEC Deployment in 3GPP 5G Systems

Even though the ETSI MEC standard is access-agnostic (and thus applicable to any kind of network), deployment in 5G systems is a key aspect in the success of this technology. Figure 6.4.5 illustrates an NG-RAN, in which multiple base stations (gNBs and ng-eNBs) are connected to the AMF/UPF elements in the 5GC through NG interfaces (for further details about NG-RAN architecture, refer to Chapter 4).

Even though MEC is expected to be deployed close to edge (i.e. in NG-RAN), formally it is not part of the NG-RAN architecture (and thus is not shown in Figure 6.4.5). In this architecture the user-plane data coming from terminals (and base stations) are forwarded to UPF, and the control plane to AMF elements. As the MEC data plane needs to be able to route user-plane packets to the MEC system, it can only be connected to a UPF. This, however, is not explicitly defined in the standard.

Having said that, according to 3GPP TS 23.501, the 5G system may also include non-3GPP Network Elements, one of which can be MEC. The following mapping between ETSI MEC entities and 3GPP entities can be envisioned:

- *UPFs* handle the user plane of Protocol Data Unit (PDU) sessions. Moreover, a UPF being the PDU session anchor may provide the interface to a data network (DN). Therefore, the logical UPF network node in the 3GPP architecture may correspond to some functionalities defined in ETSI for the MEC data plane (ETSI GS MEC 003).

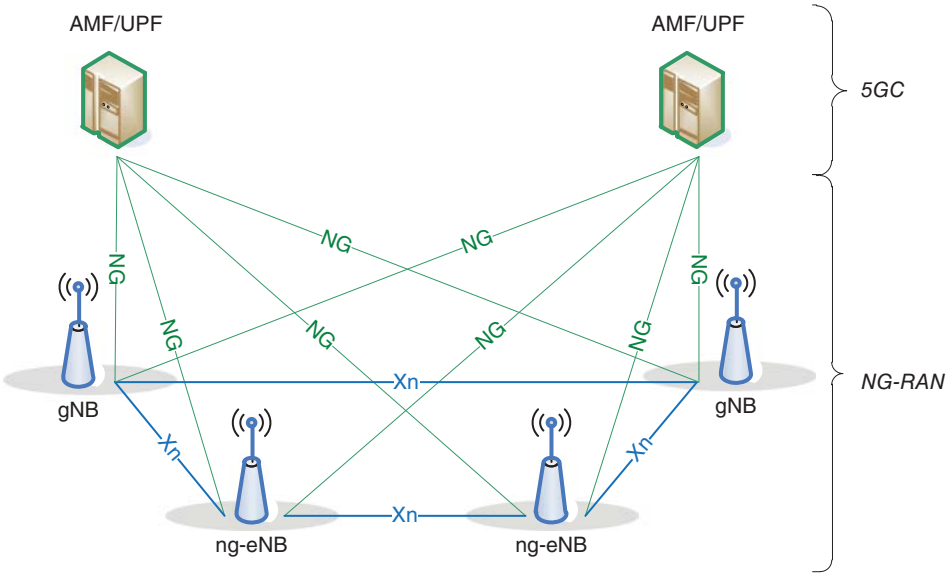



**Figure 6.4.5** NG-RAN architecture (Source: Reproduced with permission from © 3GPP).

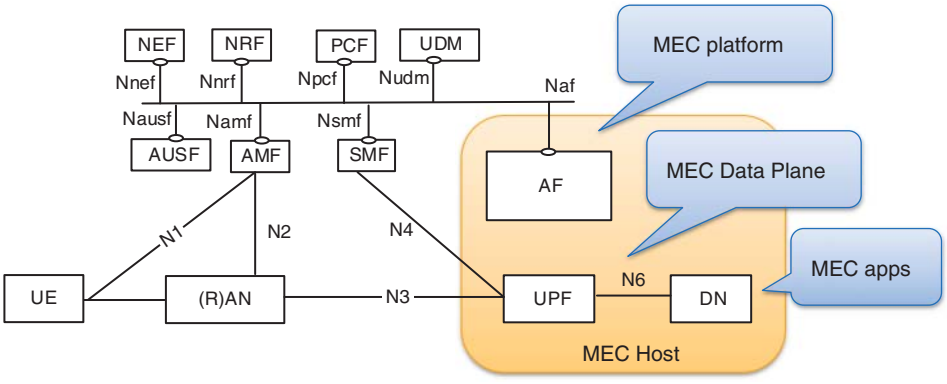



**Figure 6.4.6** Example of MEC mapping to the 5G system architecture.

- *AF* in 3GPP architecture contains the following high-level functionalities: application influence on traffic routing, access network capability exposure, and interaction with the policy. Therefore, the logical AF in the 3GPP architecture may correspond to some functionalities defined in ETSI for the MEC platform (ETSI GS MEC 003).
- Finally, the 5G system architecture introduces the *DN* receiving user-plane traffic from the UPF. Local DN deployments can be the perfect examples of environments hosting MEC applications, in contrast to a remote DN (or a central DN).

Figure 6.4.6 shows an example of possible MEC mapping to the 5G system architecture and the related correspondence of logical entities introduced by the two standard bodies.

A more detailed example of MEC deployment in 5G (based on fully virtualized network) is provided below.

#### 6.4.4.1 MEC Deployment in a 5G Network

Network slicing and management of a virtualized environment (see Section 6.3) are two important technologies in 5G. They have been well discussed within the 5G ecosystem and relevant descriptions have been provided in works such as the NGMN White Paper (NGMN) and the relevant 5G Americas (5GAWP), as well as GSMA White Papers (GSMA).

This section is focused on the role of MEC in efficiently supporting 5G network slicing. The considered communications system incorporates a MEC system, the architecture of which is specified in MEC003, deployed in a 5G network, the system architecture of which is specified in 3GPP TS 23.501. The assumption is to consider all logical functions (i.e. NFs and also AFs) as virtualized functions. The mapping of MEC entities into a 5G system is depicted in Figure 6.4.7. In particular:

- A MEC platform is implemented as a particular AF in 3GPP.
- The data plane in MEC architecture corresponds to a UPF in 3GPP.
- MEC applications are mapped to the local DN in 3GPP.

Assuming such a system setup, one observes that the E2E 5G system performance depends not only on the performance of the RAN and core network system components, but also on the performance of the MEC functional entities, as well as the reference points interfacing

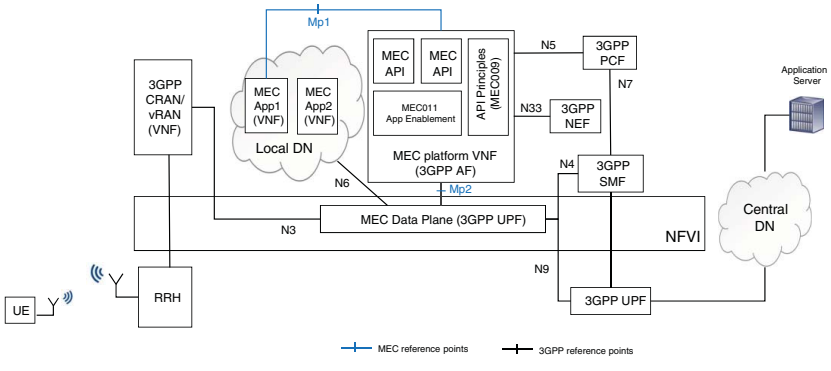



**Figure 6.4.7** 3GPP-based 5G system architecture and example of the mapping of MEC entities to AF and UPF 5GC network elements.

these entities. As an example, the E2E latency, that is, the two-way delay between the UE and the MEC application, is composed of the Packet Delay Budget (PDB), defined in 5G as the E2E delay between the UE and UPF, with a confidence of 98% (3GPP TS 23.501), and the additional delay between the UPF and the local DN, where the MEC applications are located. This second latency component is not taken into account by 5G QoS Class Identifier (5QI) characteristics in 3GPP, although it is important for performance optimization, as it is tightly related to the instantiation of the MEC applications. As a consequence, since the user traffic termination point is at the MEC application (located in the DN), slicing-relevant performance metrics (such as PDB) are not sufficient to describe the overall E2E performance. MEC application instantiation and the related VM allocation should therefore be carefully designed, as per network slice requirements, to satisfy E2E latency requirements.

Driven by the above considerations and focusing on the deployment of MEC in a fully virtualized 5G system involving the operation of multiple slices, one possible method of *slice-centric* system operation would be to optimize the allocation of MEC applications (modeled as VNFs) across the edge cloud. Allocation of MEC applications should be performed according to a slice-aware strategy, in order to meet the E2E performance requirements of a given slice, which are assumed to be part of a Service Level Agreement (SLA) between the network operator and a customer. A possible solution may involve an iterative procedure, involving both 3GPP (e.g. a mobile operator's OSS) and MEC system functional entities, as specified in MEC003 (i.e. the MEC system's MEAO, and the NFVO). The goal of such a solution is a slice-efficient allocation of virtualized resources, especially in the existence of user mobility. According to this procedure, measurements of delay components relevant to different system domains (3GPP, NFV, MEC) may be collected and then processed with the aim of identifying the performance "bottleneck" and allocating more resources to the entity/entities responsible for E2E performance degradation.

### 6.4.5 Inter-MEC System Communication

In this section, we focus on a deployment scenario involving multiple MEC systems, as specified in MEC003. An illustrative example (related to the automotive domain) considers a MEC-enabled V2X communication system, where a road operator (equivalently, an ITS operator) aims to offer V2X services in a cross-country, cross-operator and cross-vendor environment. Figure 6.4.8 illustrates an example scenario, in which there are two mobile operators deploying an ITS network, with MEC hosts co-located to the corresponding radio access nodes.

One of the most challenging (but also most frequent) situations is when the ITS operator has to provide the same V2X service to all vehicles connected to different mobile operators, even in the temporary absence of radio coverage. This case is often accompanied by the presence of multiple MEC vendors, and there is, consequently, a need to enable communication between different MEC systems. For example, different MEC systems should be able to communicate, as information (e.g. PC5 V2X-relevant information, a specific service, etc.) sought by a MEC application running at a host of a given MEC system may need to be directly available to MEC platforms of other MEC systems. Additionally, the availability of critical V2X-relevant information at MEC platforms of a given MEC system should be accessible to other nearby deployed MEC systems as well, since a vehicle may pass through

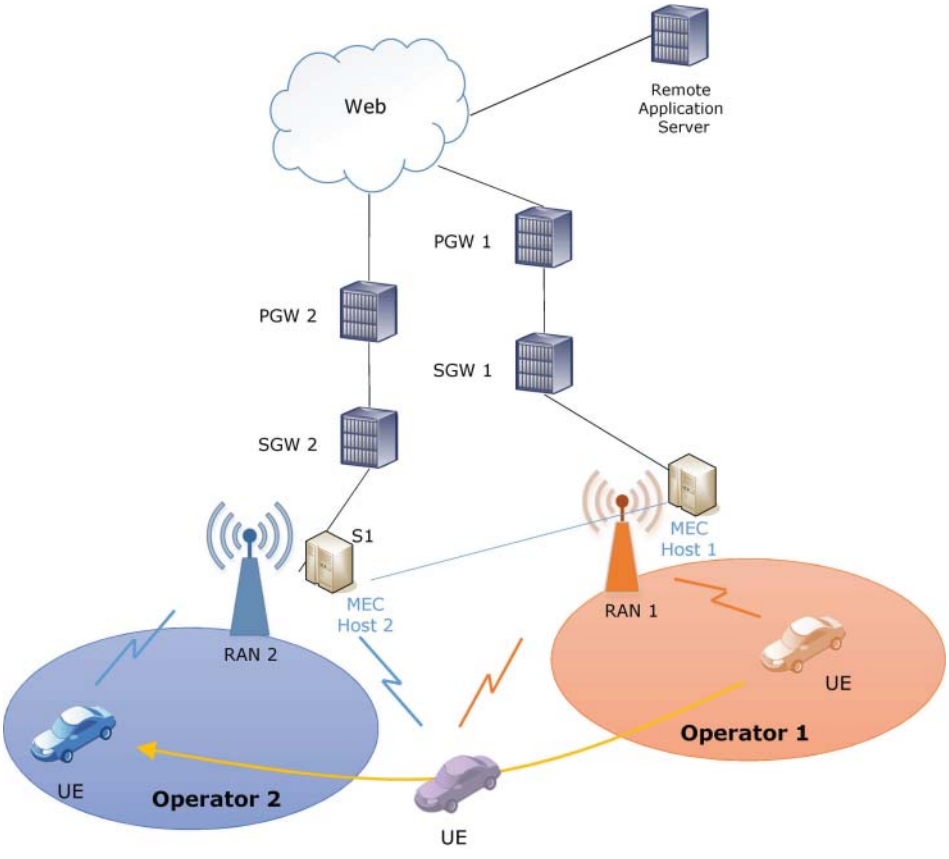



**Figure 6.4.8** A V2X communication setup involving two different mobile operators (Source: Reproduced with permission from © ETSI).

different radio connectivity areas co-located with different MEC systems. As a result, the following issues need to be addressed:

- How to provide inter-MEC system communication, e.g. for V2X services, in order to enable information exposure to MEC applications, potentially belonging to different MEC systems.
- How to enable a secure communication among MEC applications in different MEC systems.

Inspired by the above described scenario (see also the relevant clauses in MEC003), an inter-MEC system communication aims to address the following needs:

- A MEC platform should be able to discover other MEC platforms that may belong to different MEC systems.
- A MEC platform should be able to exchange information in a secure manner with other MEC platforms that may belong to different MEC systems.
- A MEC application should be able to exchange information in a secure manner with other MEC applications that may belong to different MEC systems.

To enable inter-MEC system communication, the following hierarchical inter-MEC system discovery and communication framework is assumed:

- MEC system-level inter-system discovery and communication.
- MEC host-level inter-system communication between the MEC platforms.

Below we describe a possible implementation of these hierarchy levels for inter-MEC system communication. At the time of writing this book, these mechanisms are not yet part of the ETSI MEC specifications.

#### 6.4.5.1 Possible Implementation

The information exchange required to support the inter-MEC system communications described above can be implemented in different ways, for example, at MEC orchestration, at MEC platform management, or at MEC platform level. Figure 6.4.9 illustrates the possible levels of interactions for the inter-MEC system communication.

- Communication among MEOs for MEC system discovery:
  Using a V2X communication system as an example, the ITS operator, or the mobile operator, has to rely on multiple different MEC systems deployed in some "ITS service areas," e.g. in a country or across the border, in order to provide consistent ITS services in that area. In this case, the operator may define a set of MEC systems (and their IDs) deployed in this "ITS service area." The set of MEC system IDs is communicated to all MEOs, e.g. by means of a dedicated reference point, so that every MEC system is aware of the set of other systems it may need to communicate with. Such inter-MEO communication may take place periodically, depending on the rate of deploying new MEC systems in the "ITS service area."
- Communication among MEC platform managers for MEC platform discovery:
  Once the communication interface between two MEC systems has been established, the next step is MEC platform discovery between these systems. In this step the MEPM of each system constructs the ID of available MEC host IDs under its control, e.g. those

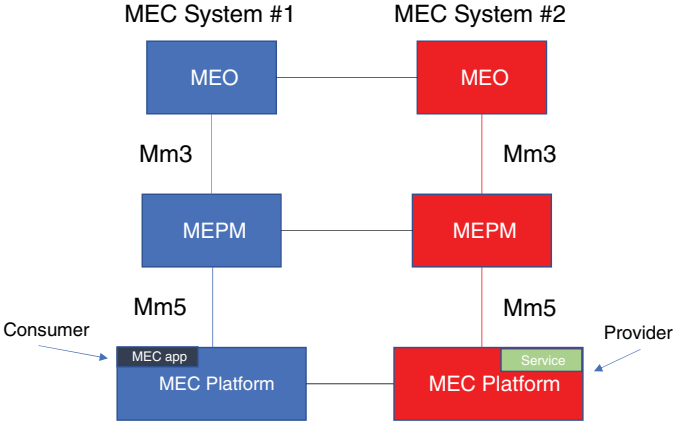



**Figure 6.4.9** Graphical representation of the layered/hierarchical approach for inter-MEC system communication.

which are not under maintenance and have sufficient processing, memory, storage, and potentially other resources. This set of available MEC hosts can then be communicated to the other system's MEPM, e.g. by means of a dedicated reference point. This enables the establishment of a direct communication interface between the two MEC platforms. The frequency of the inter-MEPM discovery message exchange may vary, depending on the frequency with which the availability of a MEC host changes (e.g. based on MEC load fluctuations caused by changes in vehicle spatial densification over time).

- Communication among MEC platforms:
  After MEC system and MEC platform discovery and establishment of the communication interfaces described above, every MEC platform may indicate for each service it supports whether it can be shared with other systems, e.g. by using a "public"/"private" tag. In this manner, services that should only be consumed locally (at the same host, or only within an intra-system "zone," e.g. due to privacy issues) will be excluded from inter-system inter-platform sharing. Therefore, only a subset of the supported services will be exposed to other MEC systems. The set of sharable services can be then directly communicated to the other accessible systems' MEC platforms, e.g. by means of a dedicated reference point.

### 6.4.6 Flexible MEC Service Consumption

In this section, we discuss the impact of MEC topology on the MEC service consumption and how it affects E2E performance. As mentioned above, a MEC application can consume MEC services, which can be running on the same or a different MEC host.

A MEC application decision about selecting the best MEC host for a certain service can be affected by different KPIs, for example, achievable communication round-trip time (RTT). Additionally, once the MEC application has selected the MEC host, the MEO may relocate the MEC application instance closer to the needed service, which may affect the KPIs used by the application in the initial selection decision. Efficient MEC application relocation may only be possible if the MEO has up-to-date performance measurements (e.g. delay), which is in turn affected by the VIM policy.

The MEC host selection process described above, performed by either MEC application or MEO, requires information about available MEC hosts and their locations. Below we describe a possible implementation of a flexible MEC service consumption framework, which addresses the issue described here. At the time of writing this book, these mechanisms are not yet part of the ETSI MEC specifications.

#### 6.4.6.1 Edge Host Zoning in Multi-Vendor Environments

In this example we consider a 5G communication system with MEC hosts deployed over a large territory. For the sake of simplicity, we consider only one MEC system composed of different MEC hosts, where each MEC host is associated with at least one gNB. Furthermore, there is a MEC application running on a MEC host, which needs to consume MEC services instantiated within the same MEC system. It is assumed that the required services are available in the MEC system; however, not necessarily running at the same MEC host. Figure 6.4.10 depicts the considered reference system.

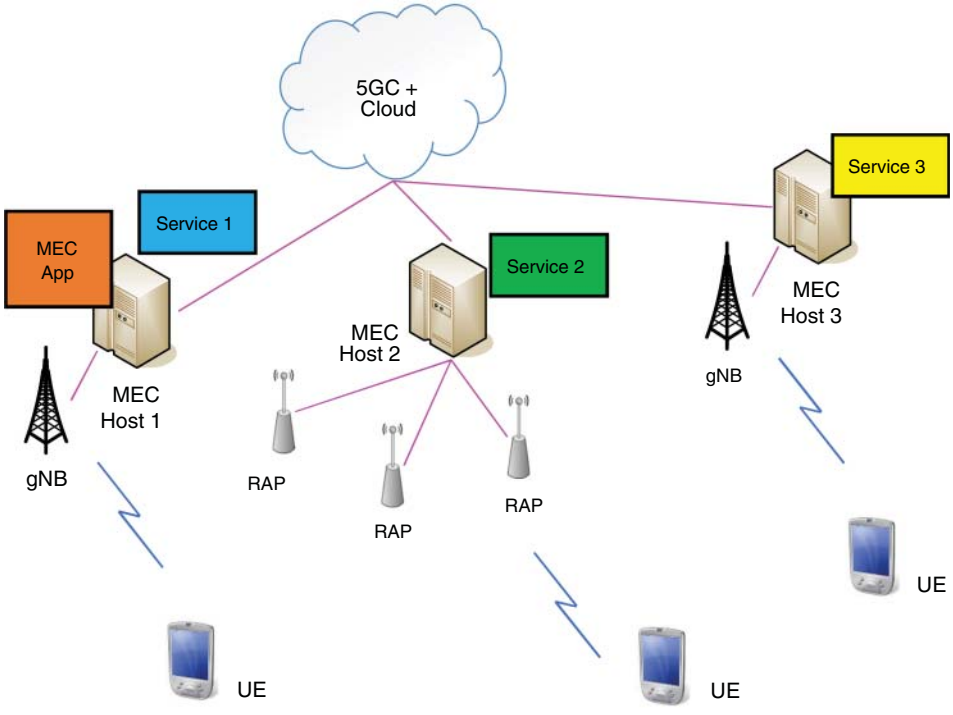



**Figure 6.4.10** A 5G system with MEC, where a MEC application attempts to consume services deployed at different locations.

One possible solution (FSR19) relies on the notion of *proximity zones* around MEC hosts hosting MEC applications, along with a zone-aware signaling protocol for delay-efficient MEC service consumption by a MEC application.

Figure 6.4.11 depicts an exemplary topology of a MEC system consisting of four MEC hosts within which MEC platforms run different MEC services. Furthermore, a MEPM, a MEO, as well as the various interfaces/interconnections between these entities explained above, are shown. Specifically:

- Mm3 interface connecting the MEO with the MEPM;
- Mm5 interface connecting the MEPM with the MEC platform;
- Mp3 interface inter-connecting the MEC hosts of the system;
- Mp1 and Mp2 within each MEC host.

It is further assumed that a MEC application is running on MEC host 1, which is potentially in need of consuming some of MEC services 1, 2, 3, and 4. To evaluate the cost of consuming a specific MEC service by the MEC application, the proximities of MEC hosts 2, 3, and 4 need to be measured (having MEC host 1 as a reference), classified according to a performance or a cost metric, and stored in the MEC system.

The MEO is the entity responsible for gathering, classifying, and storing the proximity measurements, as it has an overall view of the MEC system topology, the available resources, and the available MEC services (MEC003). Therefore, to accomplish that procedure, the

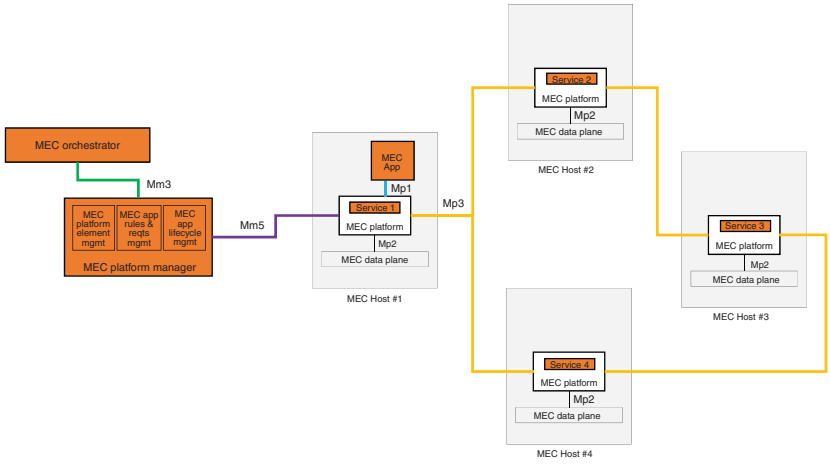



**Figure 6.4.11** Exemplary topology of a MEC system consisting of four MEC hosts; a MEC app is running at MEC host 1.

**Table 6.4.1** MEC hosts divided into proximity zones according to a certain criterion.

| Proximity zone | Minimum cost | Maximum cost (units) | MEC hosts of the zone |
|---|---|---|---|
| 1 | 0 | 5 | 1 |
| 2 | 0 | 10 | 1, 2, 4 |
| 3 | 0 | 20 | 1, 2, 3, 4 |

MEO will construct a table defining zones (i.e. clusters of MEC hosts), based on the latency (or, any other performance/cost-based utility) of reaching the reference MEC host running the MEC app (i.e. MEC host 1 in our example).

As an example, Table 6.4.1 together with Figure 6.4.12 provide an example of such proximity-based classification maintained at the MEO; it should be noted that the cost values and the number of MEC hosts in a zone are selected in arbitrary fashion in this example to illustrate the concept.

As shown in Figure 6.4.12, only MEC host 1 belongs to proximity zone 1, whereas proximity zones 2 and 3 incorporate MEC hosts, the hosted MEC services which can be reached at a higher cost (from the MEC app-to-service latency performance standpoint). It should be noted that the construction of the MEC proximity zones should be updated each time the MEC system deployment (topology) is altered; for example, when more MEC hosts are deployed in a given area, and/or, when the physical interfaces inter-connecting the MEC servers are upgraded.

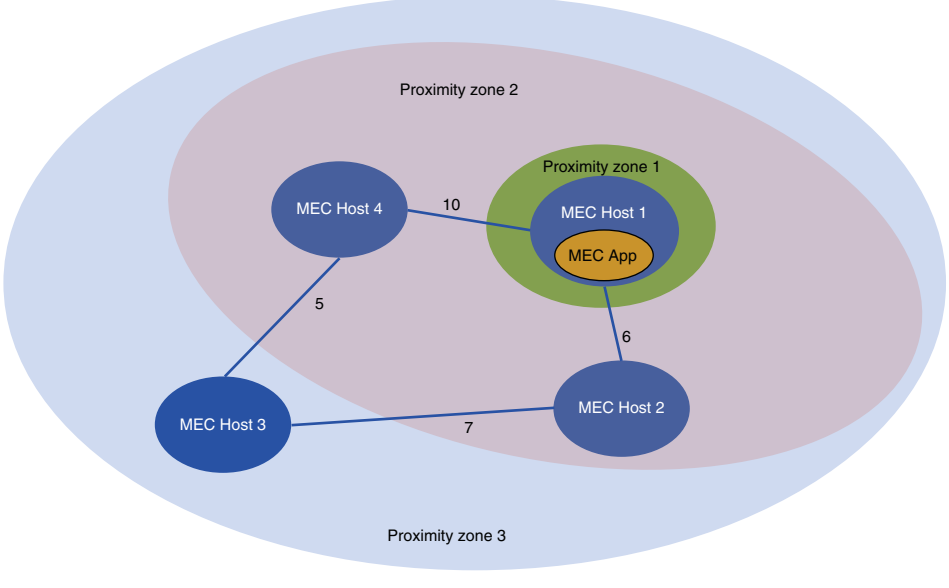



**Figure 6.4.12** Visualization of MEC host proximity zones (as seen by the MEC application), according to the utility-based classification of Table 6.4.1.

Based on such utility-based classification, the MEO can then make a decision on whether to recommend the relocation of a MEC application instance, or not. The decision would be affected by E2E QoS/cost requirements and the need to consume specific MEC services. To this end a signaling protocol (FSR19) can be defined among the various MEC system entities with the aim of achieving QoS-aware/cost-efficient service consumption by a given MEC application instantiated at a host of a given MEC system. The procedure should allow the MEO to evaluate whether a MEC application instance relocation would satisfy the E2E performance/cost requirements as the MEC service consumption delay is only a fraction of the E2E delay.

An exemplary version of the potential protocol to support the functionality described above is shown in Figure 6.4.13.

1. The MEC application triggers the procedure by requesting the MEC host deployment parameters from the MEO. The request should contain the list of service(s) to be consumed.
2. The MEO requests the parameters (e.g. expected delay) from the different deployed MEC hosts (hosting various services) by the MEPM.
3. The MEO gathers all needed MEC host parameters and defines performance zones, with reference to the location of the instantiated MEC application.
4. The MEO provides a recommendation for service consumption to the MEC application (e.g. suggesting direct service consumption or MEC app instant relocation); it is then up to the MEC application to accept/reconsider the recommendation, while considering the E2E nature of the performance metric.

The example shown above is not limited to a single MEC system and is applicable also in the case of consuming MEC services residing within different MEC systems (e.g. associated with different network operators).

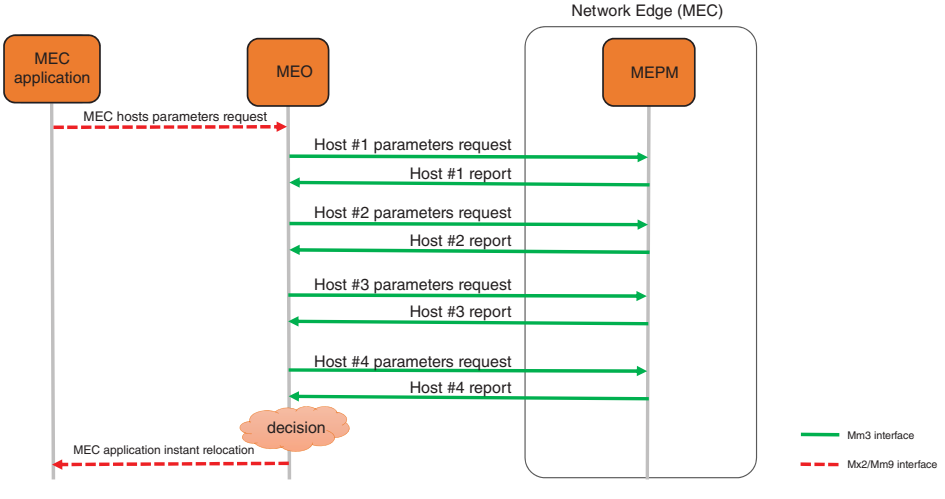



**Figure 6.4.13** Potential signaling protocol among functional entities of a MEC system to support proximity-based MEC service consumption.

### 6.4.7 High Mobility Automotive Scenarios

As we explained in the beginning of this section, the automotive sector is one of the key areas for edge computing applications. One particular difficulty of this application is the need for interoperability between multiple MNOs deploying ITS networks in the same area (MEC030). Below we describe a potential implementation of a MEC-based framework, where predicted QoS and related information can improve V2X services offered to drivers. The functionality described in the present subsection is not part of the MEC specifications yet.

#### 6.4.7.1 MEC-Supported Cooperative Information

While the issue of multi-MNO ITS deployment is bigger than just QoS, in this subsection we discuss the problem of efficiently predicting the QoS along planned vehicle trajectories, as an example of the type of problems that need to be solved for a successful ITS deployment. While it is generally beneficial to have both accurate and timely predictions of the radio environment at locations along the route of a vehicle, it is of particular importance in the context of MEC as these conditions can trigger:

- Initiation of certain V2X functionalities;
- Download of content delivery/software packages.

As explained above, MEC is a technology allowing applications to be instantiated at the edge of the access network, providing low latency applications in close proximity to user terminals. However, in V2X system scenarios characterized by high mobility, the information related to radio network conditions centrally collected by a MEC host may not be always up to date. The accuracy and the timeliness of the information collected are hampered by the environmental situation, for example, the occurrence of network congestion events when, for example, many vehicles attempt to provide radio measurements to the connected eNB/gNB, which is collocated with a MEC host. It is also affected by the deployment density of the cellular network, together with the capabilities of the deployed MEC infrastructure.

To illustrate the impact of the above-mentioned limitations on system performance we use an example of a vehicle en route from location A to location B and a related MEC application that would need to be informed of radio conditions along that route. The MEC application may need to have that information ahead of time in order to make timely decisions, such as:

- Enabling/disabling autonomous driving features;
- Downloading infotainment content;
- Scheduling software over the air/firmware over the air (SOTA/FOTA) updates, etc.

According to the current specification, journey-specific environmental/situational information will only be available to vehicles together with a bulk of other data, most of which is irrelevant to the planned route. This is suboptimal, as V2X communications links may become congested, which will result in that information not being available in time. It would therefore be beneficial to identify space/time correlations between radio quality data collected by different vehicles in a C–V2X system and a specific vehicle's planned journey

for better predictability of the quality of the communication along the designated route. Furthermore, how the predicted and journey-specific RNI can be exploited by vehicles to reliably schedule and complete SW package updates and/or content delivery tasks needs more consideration.

A possible way to address the above-mentioned challenges would be to design a framework based on a MEC infrastructure for cooperative acquisition, partitioning, and distribution of information for efficient and journey-specific QoS prediction. The main stages of such a framework can be as follows:

1. Each client application in the vehicle (which is assumed to be under cellular coverage) reports its planned journey information (i.e. map coordinates) to the MEC host (which is also running a geo-location service). The information reported by the client application may also optionally include the version of the installed SW/FW package, the versions of client applications in the vehicle, and other relevant data.
2. Each vehicle provides locally measured radio quality information to the MEC host. The reporting periodicity may vary, depending on the version of its running SW/FW package. Each message containing radio information is tagged with a time stamp and the vehicle's current location.
3. Based on the per-vehicle planned route, the composite information obtained at the MEC host is partitioned by routes as reporting vehicles are traveling. Historical partitioned information can be used to predict QoS characteristics.
4. Assuming the example of SW download, the availability of a new SW package at the network side triggers information partitioning by a MEC application. The MEC application takes the journey-specific information partitions, together with the SW package versions running at each of the vehicles of interest, as inputs for decision making. The outcome of the decision process may be a recommendation to trigger SW update, which is then communicated to the client application in the vehicle, which:
   a. Has a planned route with suitable radio conditions for the triggered action (e.g. SW downloaded);
   b. Has outdated SW subject to upgrade.
5. If the planned journey route changes, the procedures are repeated, starting from step 1.

Figure 6.4.14 illustrates the scenario described above. We assume that a MEC host is collocated with a road-side unit (RSU), for example, a gNB. We further assume that three vehicles, V1, V2, and V3, are under cellular coverage, where vehicles V1 and V2 have the same planned route from location A to location B, but V3 has planned a different route from location C to location D. Focusing on the use case of SOTA/FOTA updates, we additionally assume that vehicle V1 has the latest version (e.g. v1.3) of a specific SW package installed, whereas vehicles V2 and V3 have an outdated version (e.g. v1.2) installed.

The procedure may work as follows:

1. V1 reports its planned journey from location A to location B to the MEC host via transmission to the gNB-type RSU; it also informs that it has v1.3 of the SW package installed.
2. V2 reports its planned journey from location A to location B to the MEC host via transmission to the gNB-type RSU; it also informs that it has v1.2 of the SW package installed.
3. V3 reports its planned journey from location C to location D to the MEC host via transmission to the gNB-type RSU; it also informs that it has v1.2 of the SW package installed.

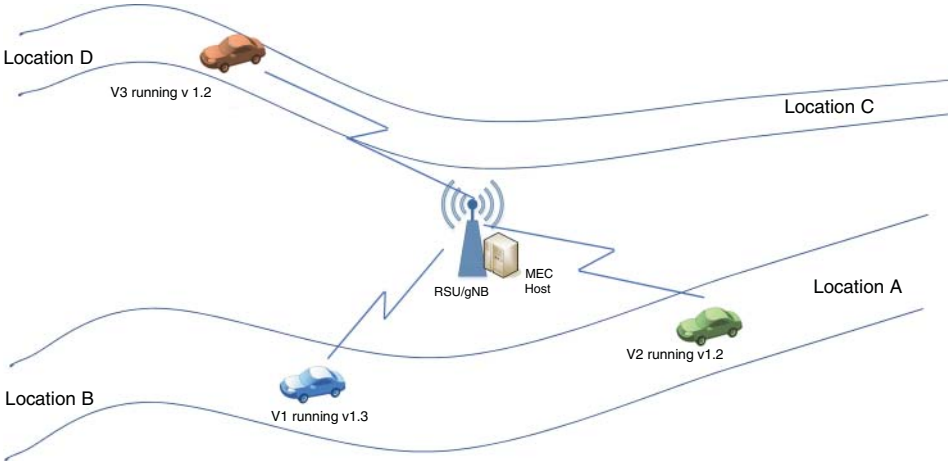



**Figure 6.4.14** Cooperative decision making for SOTA/FOTA updates with MEC.

4. All three vehicles report radio signal quality measurements, together with location information and time stamps; V1 uploads fewer tagged measurements as compared with vehicles V2 and V3, as it is not in need of a SW package update (for the moment).
5. The MEC host acquires the data reported and partitions it by route; two data partitions are created: one for the route from A to B and one for the route from C to D.
6. Using the partitioned data, together with statistics gathered previously, the MEC host predicts the radio signal quality that all vehicles are expected to experience and, taking into account the SW package's size, recommends V2 to start or postpone the download. If the recommendation is positive, the starting time of download is also recommended. In the present example, vehicles V1 and V2, which are grouped in the same partition, will receive the same recommendation, whereas the recommendation for V3 may be different.

### 6.4.8 Further Reading

For additional information about ETSI MEC architecture and APIs, please refer to the ETSI specifications provided below.

## References

1 5G Americas White Paper (2016). Network Slicing for 5G networks and services. Available at: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_Network_Slicing_11.21_Final.pdf (accessed June 18, 2020).
2 Cisco White Paper (2017). Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021. Available at: https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html (accessed June 18, 2020).

3 Filippou, M.C., Sabella, D., and Riccobene, V. (2019). Flexible MEC service consumption through edge host zoning in 5G networks. 2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW), Marrakech, Morocco, pp. 1–6.

4 Giust, F., Costa-Perez, X., and Reznik, A. (December 2017). Multi-access edge computing: An overview of ETSI MEC ISG. *IEEE 5G Tech Focus* 1 (4).

5 GSMA White Paper (2018). Network slicing – use case requirements. Available at: https://www.gsma.com/futurenetworks/wp-content/uploads/2018/06/Network-Slicing-Use-Case-Requirements-_-FInal-.pdf (accessed June 18, 2020).

6 Giust, F., Sciancalepore, V., Sabella, D. et al. (September 2018). Multi-access edge computing: the driver behind the wheel of 5G-connected cars. *IEEE Communications Standards Magazine* 2 (3): 66–73.

7 Hu, Y.C., Patel, M., Sabella, D., Sprecher, N., and Young, V. (2015). Mobile edge computing a key technology towards 5G, 1. Available at: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp11_mec_a_key_technology_towards_5g.pdf (accessed June 18, 2020).

8 ITU-R Working Party WP SO (June 2015). Draft new recommendation IMT vision – framework and overall objectives of the future development of IMT for 2020 and beyond. Doc. RI2-SGOS-C-0199.

9 ETSI (October 2018). Multi-access edge computing (MEC); phase 2: use cases and requirements. European Telecommunications Standards Institute (ETSI), Group Specification (GS) 002 v2.1.1.

10 ETSI (January 2019).Multi-access edge computing (MEC); framework and reference architecture. European Telecommunications Standards Institute (ETSI), Group Specification (GS) 003 v2.1.1.

11 ETSI (July 2017).Mobile edge computing (MEC); general principles for mobile edge service APIs. European Telecommunications Standards Institute (ETSI), Group Specification (GS) 009 v1.1.1.

12 ETSI (December 2019). Multi-access edge computing (MEC); radio network information API. European Telecommunications Standards Institute (ETSI), Group Specification (GS)012 v2.1.1.

13 ETSI (July 2017). Mobile edge computing (MEC); Location API. European Telecommunications Standards Institute (ETSI), Group Specification (GS) 013 v1.1.1.

14 ETSI (February 2018). Mobile edge computing (MEC); UE Identity API. European Telecommunications Standards Institute (ETSI), Group Specification (GS) 014 v1.1.1.

15 ETSI (October 2017). Mobile edge computing (MEC); Bandwidth management API. European Telecommunications Standards Institute (ETSI), Group Specification (GS) 015 v1.1.1.

16 ETSI (April 2020). Multi-access Edge Computing (MEC); V2X Information Service API. European Telecommunications Standards Institute (ETSI), Group Specification (GS)015 v2.1.1.

17 NGMN Alliance (2016). Description of network slicing concept. NGMN 5G P 1.

18 NGMN Alliance (February 2015). NGMN SG White Paper. February 2015.

19 3GPP (September 2019). 3rd Generation Partnership Project; Technical Specification group services and system aspects; System architecture for the 5G system; Stage 2 (Release 15). V15.7.0 (09-2019).

**20** Taleb, T., Samdanis, K., Mada, B. et al. (2017). On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration. *IEEE Communications Surveys & Tutorials* 19 (3): 1657–1681.

**21** Zhang, D., Zhou, Z., Mumtaz, S. et al. (December 2016). One integrated energy efficiency proposal for 5G IoT communications. *IEEE Internet of Things Journal* 3 (6): 1346–1354.

**22** Ericsson (n.d.). Switch on 5G for business. Available at: https://www.ericsson.com/en/5g/5g-for-business (accessed June 18, 2020).

**23** HTTP (n.d.). Over TLS. Available at: https://tools.ietf.org/html/rfc2818 (accessed June 18, 2020).

## 6.5 Operations, Administration, and Management

*Vladimir Yanover*
*Cisco Systems, Inc., Israel*

### 6.5.1 Introduction

The 5G system is expected to provide optimized support for a variety of communication services, different traffic patterns, and different end user categories.

To enable efficient network operations, 3GPP developed an OAM framework that can be applied to the 5G system and 3GPP legacy systems. The 3GPP management system directly manages 3GPP network components (e.g. NG-RAN, 5GC). For non-3GPP domains, such as the transport network, 3GPP management system needs to coordinate with the corresponding management systems of these domains.

The 5G management framework applies the novel concept of service-based management architecture, which includes NG-RAN and 5GC information models, new aspects of PM, and special arrangements for management of disaggregated RAN. For the networks with slicing support, the procedures for management of network slices are defined.

Furthermore, the self-organizing network (SON) concept evolves with 5G to address new radio (NR) technology and incorporates the concept of multi-domain optimization. SON evolution is driven by the evolution of 3GPP network management techniques for RAN and core network domains.

### 6.5.2 Key Ideas

- In LTE, the management architecture is defined in terms of logical network nodes: Network Elements (NEs) and Element Managers (EMs). In 5G, a service-based management architecture has been introduced. This is in line with the overall approach for describing the 5GC in the standards.
- 3GPP defines management information models, which are decoupled from the communication protocols (such as NETCONF and YANG) used to access the OAM information in a model.
- Management models defined in 3GPP cover all NG-RAN architecture options: monolithic gNB and various split NG-RAN architectures, described in Chapter 4.
- PM is an important part of the OAM framework, which consists of generation of performance measurements (e.g. by NG-RAN), their collection, and the corresponding management actions based on the collected measurements.
- The OAM framework supports network slicing through the Network Slice Instance (NSI) and Network Slice Subnet Instance (NSSI) models.
- The SON concept in 5G evolves from the 4G SON functionality and adds E2E aspects covering not just NG-RAN, but also 5GC and the transport network. Furthermore, the 5G SON framework has been extended to cover slicing, new QoS requirements, NG-RAN split architectures, and more.

### 6.5.3 Service-Based Management Architecture

The 3G and 4G network management reference model is illustrated in Figure 6.5.1 (3GPP TS 32.101 clause 5.1.1).

The model includes managed NEs, such as Node-Bs (3G) and eNBs (4G), and management functions such as (EMs, domain managers (DMs), and network managers (NMs). The Type 1 and Type 2 interface categories are defined.

One of the innovations in the 3GPP Release-15 is a transition to service-based management architecture (see Section 3.2 for details of using the same approach in 5GC). In this approach, the reference model with fixed roles, such as NE or EM, was replaced by the concept of management service (MnS), with basic roles being the management service provider and management service consumer. Sometimes the provider is referred to as "producer." An MnS of an MnS provider may have multiple consumers.

3GPP specifies only the management capabilities provided via the MnS, which is composed of individually specified components.

3GPP TS 28.533 defines three components of the management service referred to as type A, B, and C as follows:

- *Management service component type A* is a group of management operations and/or notifications. In standardized management services, a set of operations and set of notifications are typically uniform across a wide range of management services, so the services are differentiated only by the information models (see component type B). Definitions of specific instances of the component type A can be found in 3GPP TS 28.531 clause 6. Some of them are mentioned below.
- *Management service component type B* is the information model of managed entities. Specific information models are defined in 3GPP TS 28.541. In 3GPP, the information models are also referred to as Network Resource Models (NRMs). Some examples are provided below.
- *Management service component type C* is defined as the performance and fault information of the managed entity.

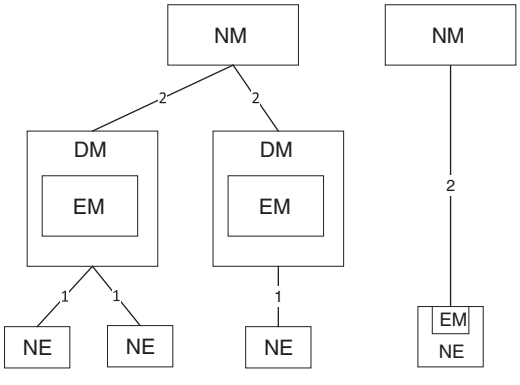



**Figure 6.5.1** 3G and 4G network management model (Source: Reproduced by permission of © 3GPP).

#### 6.5.3.1 Examples of Management Services

One family of the management services described in 3GPP TS 28.531 is the provisioning services; this family covers configuration management and LCM. It includes provisioning of:

- NFs
- NSIs

**Table 6.5.1** Operations and notifications for the provisioning MnSs for the NSI and NSSI.

| | | |
|---|---|---|
| Provisioning operations | For NSI and NSSI | • createMOI operation<br>• deleteMOI operation<br>• getMOIAttributes operation<br>• modifyMOIAttributes operation |
| | For NSI | • allocateNsi operation<br>• deallocateNsi operation |
| | For NSSI | • allocateNssi operation<br>• deallocateNssi operation |
| Provisioning data report operations | For NSI and NSSI | • subscribe operation<br>• unSubscribe operation |
| Provisioning data report notifications | For NSI and NSSI | • notifyMOICreation notification<br>• notifyMOIDeletion notification<br>• notifyMOIAttributeValueChanges notification |

Note: "MOI" stands for "Managed Object Instance."

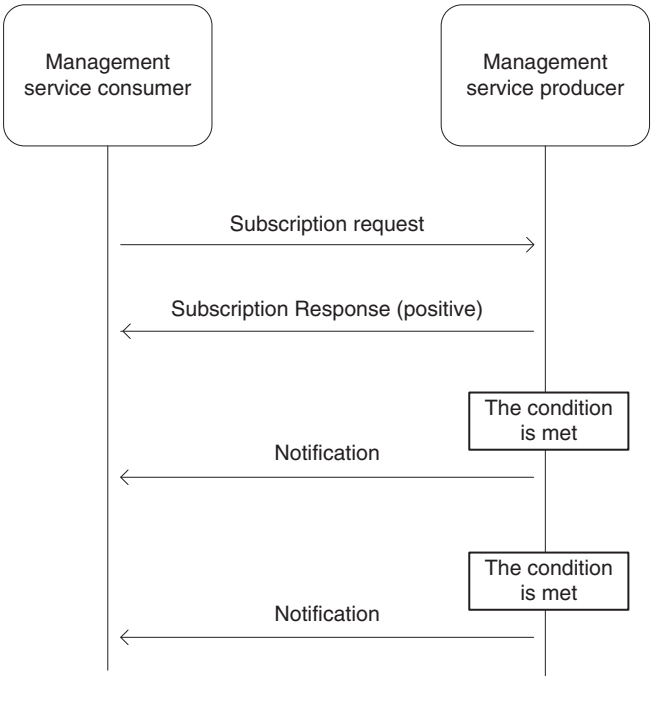



**Figure 6.5.2** Subscribe-notify communication paradigm (Source: Reproduced by permission of © 3GPP).

- NSSIs.

Component type A is typically common for a family of MnSs. Table 6.5.1 provides example of operations and notifications for the family of provisioning MnSs for the NSI and NSSI, further detailed in 3GPP TS 28.532.

Figure 6.5.2 outlines typical procedures associated with a subscription-based management service that provides notifications.

- Subscription request operation is initiated by the consumer.
- The provider replies with a positive response and the subscription is established.
- When the notification condition is met, the provider of the service sends a notification to the consumer.

#### 6.5.3.2 Management Service Exposure

The management service exposure concept is an important part of the service-based management architecture. Any entity can be a provider of the management service in the service-based architecture. A network entity consuming certain MnSs can be a "proxy" exposing (reproducing) it further.

Figure 6.5.3 shows an example of the management service X produced by entity A (NF) and consumed by another entity B (network management function) exposing it further to entity C (final consumer).

The concept of exposure provides additional deployment flexibility in the network management topology. For example, the network operator may decide that in a certain geographic area the provisioning service of RAN is directly consumed by the local SON component, while in another area the deployment will include concentrators (entity B) between the RAN nodes and centralized SON function(s).

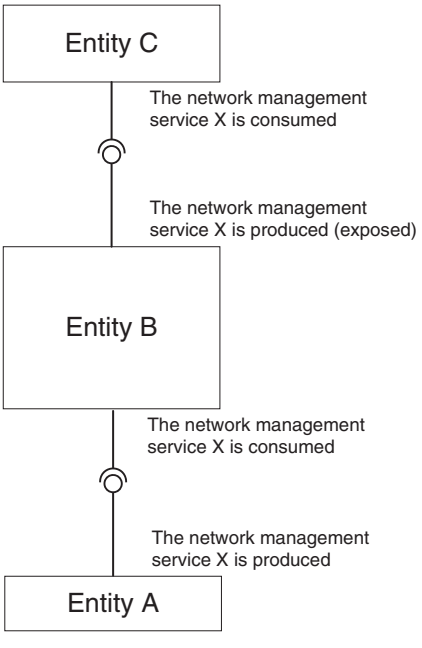



**Figure 6.5.3** The concept of exposure of network management services.

In the service-based management architecture framework, a management function (MnF) is an MnS consumer, but may also play the role of MnS provider. For example, in Figure 6.5.3, entity A is an NF producing the MnS; entity B is MnF playing both roles (MnS provider and MnS consumer)s; and entity C is an MnS consumer.

### 6.5.4 NG-RAN and 5GC Information Models

3GPP information models for 5G, also called NRMs,[8] are specified in 3GPP TS 28.541.

The NRM specification methodology includes separation of Stage 2 and Stage 3 standards, where Stage 2 specifies the semantics ("Information Service") while Stage 3 ("Solution Sets") specifies the syntax such as encoding used for information exchange between the MnS provider and MnS consumer. Protocol-independent Stage 3 definition decouples the NRM from the communication protocols. For example, an XML solution set can be used to support web service and NETCONF protocols, while a YANG solution set is not necessarily coupled with the NETCONF protocol.

The following 5G NRMs are defined in 3GPP TS 28.541:

- NR NRM (3GPP TS 28.541, clause 4)
- 5GC NRM (3GPP TS 28.541, clause 5)
- Network slicing NRM (3GPP TS 28.541, clause 6), which includes the information models of NSI and NSSI.

Examples of NRM elements described in the present section include basic elements of the RAN NRM "RRM policy."

An NRM can be used with any of the following solution sets:

- XML-based 3GPP NR and NG-RAN NRM solution set (3GPP TS 28.541, annex C);
- JSON-based 3GPP NR and NG-RAN NRM solution set (3GPP TS 28.541, annex D);
- YANG-based 3GPP NR and NG-RAN NRM solution set (3GPP TS 28.541, annex E).

### 6.5.5 Performance Management

Generation and collection of performance measurements is an essential part of the 3GPP management framework, particularly important for NG-RAN. The performance measurements are used as input for computation of KPIs for NFs.

The corresponding management services include:

- Measurement job control service
- Performance data file reporting service
- Performance data streaming service
- Performance threshold monitoring service.

The NRM-based measurements control framework introduced by 3GPP allows alignment of performance measurements control with the service-based management architecture. This control framework is based on generic provisioning services with operations, such as:

8 The term NRM has been historically used in 3GPP.

- Create Information Object Class (IOC);
- Get/modify IOC attributes, etc.

The corresponding NRM extensions include MeasurementControl IOC defined in 3GPP TS 28.622, which enables control over generation of measurements at relevant managed objects and their delivery to the network data collectors. This work started in Release-15 and continued in Release-16 with the specification of NRM for Trace Session Control.

3GPP TS 28.552 specifies an extensive set of performance measurements reported by the gNBs; the following are some important categories:

- Radio resource utilization statistics, including downlink/uplink total PRB usage, and breakdown to PRBs used for data traffic;
- UE throughput distribution;
- PDU session management statistics, including successful and failed PDU session setup requests;
- Mobility management statistics, including inter-(intra-)gNB successful and failed handover requests;
- Transport Block (TB) related measurements;
- DRB setup management statistics, including successful and failed DRS setup requests;
- RF measurements, including MCS distribution in PDSCH and PUSCH and Wideband CQI distribution;
- QoS Flow-related measurements, including setup and release statistics per cause;
- RRC connection establishment and re-establishment-related measurements, including number of successful and failed establishment requests;
- PDCP data volume measurements;
- Packet loss/drop rate;
- IP level throughput and latency measurements.

Two methods are defined in 3GPP TS 28.550 for the performance data reporting:

- *Performance data file method*: In this method the performance data are accumulated for a certain time before they are reported; the data will be delivered as a file. This method has been defined already in 3G and 4G.
- *Performance data streaming method*: In this method, the performance data streaming producer sends the performance data to the stream target when the data are ready. The stream target can be the consumer of the service or another network node. The volume of the performance data reported by streaming is expected to be small, and the Granularity Period[9] of the performance data stream is configurable and may be significantly shorter than with the performance data file method.

3GPP TS 28.554 defines E2E KPIs. Special attention was paid to supporting in RAN the measurements contributing to computation of the E2E KPIs.

The following KPIs are defined, grouped by categories:

- Accessibility KPIs;
  - Registered subscribers of network and NSI through AMF;

9 Time interval between consequent performance measurements.

  - Registered subscribers of network and NSI through unified date management (UDM);
  - Registration success rate of one single NSI;
  - DRB accessibility for UE services;
- Integrity KPIs:
  - E2E latency of 5G network;
  - Downlink latency in gNB-DU;
  - Upstream throughput for network and NSI;
  - Downstream throughput for single NSI;
  - Upstream throughput at N3 interface;
  - Downstream throughput at N3 interface;
  - RAN UE throughput;
- Utilization KPIs:
  - Mean number of PDU sessions of network and NSI;
  - Virtualized resource utilization of NSI;
- Retainability KPI;
  - QoS flow retainability.

The E2E KPIs can be used, for example, to evaluate the network performance in the slicing scenario. To this end, 3GPP provided specifications for KPIs specific to NSSIs and NSIs. One difference specific to performance measurements for slicing is that the performance data are received from multiple NFs, and the NSI performance data are computed based on the performance data received from multiple NSIs and possibly individual NFs.

### 6.5.6 Management of Split NG-RAN

#### 6.5.6.1 Background

In 5G, 3GPP defined not only the monolithic NG-RAN architecture, but also a number of split gNB architectures. In these architectures, a gNB can be split into a centralized node, referred to as gNB-CU, and multiple distributed nodes, referred to as gNB-DUs. Additionally, a gNB-CU may be further split into a control-plane node, referred to as gNB-CU-CP, and possibly one or multiple user-plane nodes, referred to as gNB-CU-UP. These architectures are described in detail in Sections 4.2 and 4.4.

#### 6.5.6.2 Information Object Classes

The NG-RAN NRM was designed to enable "separate" provisioning of gNB-CU, gNB-DU, gNB-CU-CP, and gNB-CU-UP. Figure 6.5.4 (3GPP TS 28.541) shows the "containment" relations between the IOCs in the gNB NRM. The gNB NRM is applicable to all deployment scenarios including monolithic gNB.

The model fragments are for the representation of both gNB and en-gNB (see Section 4.3).

The object classes used in NR modeling shown in Figure 6.5.4 are derived from the Managed Function class. There are object classes defined for gNB and en-gNB or their components, as explained in Table 6.5.2.

Furthermore, the following IOCs are defined in association with certain RAN components to serve some special needs. For example, the gNB-CU needs some information from the particular cells served by all connected DUs, as shown in Table 6.5.3.

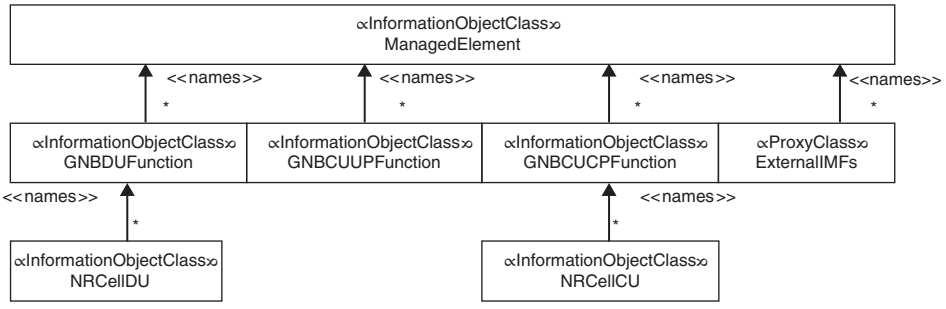



**Figure 6.5.4** gNB (en-gNB) NRM for all deployment scenarios (Source: Reproduced by permission of © 3GPP).

**Table 6.5.2** IOCs for the RAN components.

| Information object class | Description |
|---|---|
| GNBDUFunction | Represents the logical function DU of gNB or en-gNB |
| GNBCUCPFunction | Represents the logical function CU-CP of gNB and en-gNB |
| GNBCUUPFunction | Represents the logical function CU-UP of gNB or en-gNB |

**Table 6.5.3** IOCs for the DU components.

| Information object class | Description |
|---|---|
| NRCellCU | Represents the information required by CU that is responsible for the management of inter-cell mobility and neighbor relations via ANR |
| NRCellDU | Represents the information of a cell known by DU such as information of the resources realizing the cell |
| NRSectorCarrier | Represents the resources of each transmission point included in the cell. These in general have different physical locations (of the antennae), and possibly different frequencies or bandwidths. The UE is not directly aware of which NRSectorCarrier resources the network uses for its connection |

The standard also identifies the necessary endpoints required for the representation of gNB and en-gNB, of all deployment scenarios. The endpoints however depend on the split scenario. Taking the example of the gNB-DU, Table 6.5.4 provides three different endpoint definitions for three different split scenarios.

### 6.5.7 O-RAN Alliance Management Architecture

As mentioned in Chapter 4, in addition to 3GPP, the O-RAN Alliance also works on defining a number of NG-RAN architectures; for example, low-level gNB split (described

**Table 6.5.4** Endpoints IOCs.

| Req. role | Endpoint requirement for three-split deployment scenario | Endpoint requirement for two-split deployment scenario | Endpoint requirement for non-split deployment scenario |
|---|---|---|---|
| gNB | ≪IOC≫EP_XnC, ≪IOC≫EP_NgC, ≪IOC≫EP_F1C, ≪IOC≫EP_E1 | ≪IOC≫EP_XnC, ≪IOC≫EP_NgC, ≪IOC≫EP_F1C ≪IOC≫EP_F1U | ≪IOC≫EP_XnC, ≪IOC≫EP_NgC |
| en-gNB | ≪IOC≫EP_X2C, ≪IOC≫EP_F1C, ≪IOC≫EP_E1 | ≪IOC≫EP_X2C, ≪IOC≫EP_F1C | ≪IOC≫EP_X2C |

in Section 4.5), Non-Real-Time RIC, and Near-Real-Time RIC.[10] For OAM aspects of the low-level split architecture, please refer to Section 4.5.

### 6.5.8 Management of Network Slicing

#### 6.5.8.1 Basic Concepts of Slicing Management

One of the new 5G features is slicing, which is described in Chapter 3. Here we discuss the OAM impacts of this feature.

According to 3GPP TS 28.530, from a management point of view, an NSI is a subnetwork that includes all needed NF instances (including 5GC and NG-RAN), to provide a certain set of communication services to serve a certain business purpose. The management system controls the topology of the subnetwork together with the network resources allocated to the NSIs and associated QoS requirements.

The NSSIs are building blocks for NSIs. The NSSI represents a group of NF instances (and their resources).

NSIs and NSSIs can be composed of PNFs or VNFs or both. An important part of slicing management is management of virtualized and non-virtualized resources used by NSIs and NSSIs.

In general, NSI is an E2E construct and therefore contains not only NFs from NG-RAN and 5GC, but also elements of transport network, (S)Gi-LAN[11] facilities, etc. 3GPP, within its scope, defined the management aspects of the parts of NSI or NSSI that belong to the 5GC and NG-RAN. There is also some provisioning for the management of the non-3GPP part, containing the transport NEs including backhaul, midhaul, and fronthaul parts, links, routers, firewalls, etc. In a typical scenario, the 3GPP management system provisions the network slicing-related aspects to non-3GPP parts via their respective (non-3GPP) management systems.

NSSI may include NFs directly and via another constituent NSSI(s). Furthermore, an NF may be shared by two or more NSSIs. For example, a network operator may have most of

10 O-RAN RIC architectures are still work in progress and therefore are outside of scope of the book.

11 Gi-LAN (SGi-LAN) includes service enablers, which reside in the carrier's network beyond the Gi (SGi) termination point. Typical examples: firewalls, DPI, video optimization, TCP optimization, HTTP header enrichment, NAT, load balancers, caching, etc.

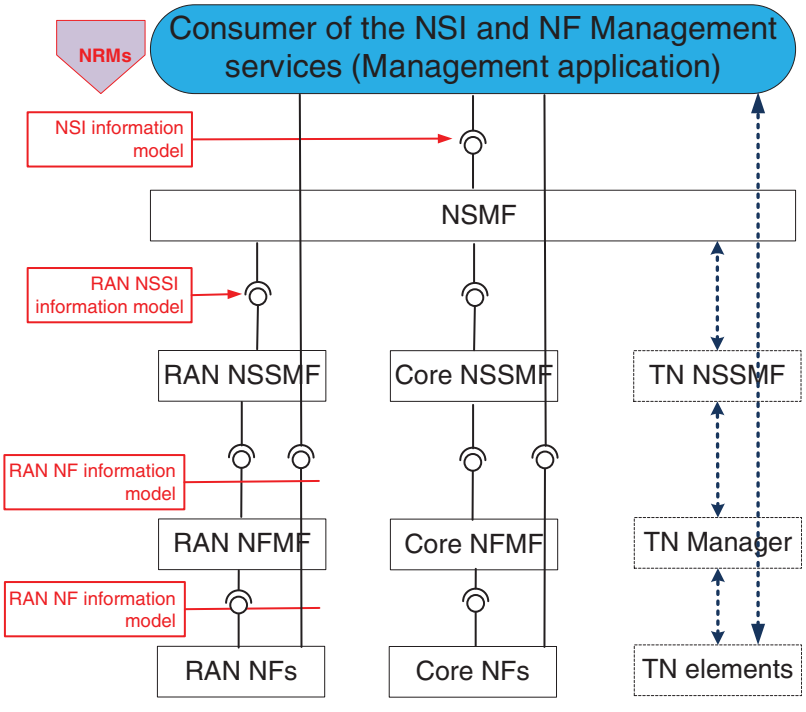



**Figure 6.5.5** An example of deployment scenario for management of a mobile network with slicing.

its NG-RAN nodes shared between all NSSIs supported in the area. An NSSI may be shared by two or more NSIs or NSSIs. An NSI (NSSI) that is not shared is sometimes referred to as "dedicated."

A slice consists of not only logical components, but also physical resources used by the NSSI. In the case of a shared NF (between NSSIs), the corresponding resources (physical or virtual) are split between the NSSIs. In particular for NG-RAN, it means a split of radio resources between the NSSIs; see example of "RRM policy" below. The NSSI information for transport networks also includes a set of links (connecting NFs) with their bandwidth and latency properties.

In a typical scenario (see Figure 6.5.5, which illustrates an example of slicing management), an NSI is composed of one RAN NSSI and one core NSSI.

This deployment scenario includes:

- Network Slice Management Functions (NSMFs), which provide for LCM of the NSIs;
- Network Slice Subnet Management Functions (NSSMFs), which provide for LCM of the NSSIs, such as RAN NSSI, CN NSSI, and TN NSSI;
- Network Function Management Functions (NFMFs), which expose the management services originally provided by the NFs.

Figure 6.5.6 also shows the consumer of the NSI and NF management services, which can be any management application, such as, for example, SON. The consumer may connect to the NFs directly, which is necessary in the case of NFs shared by different NSIs/NSSIs.

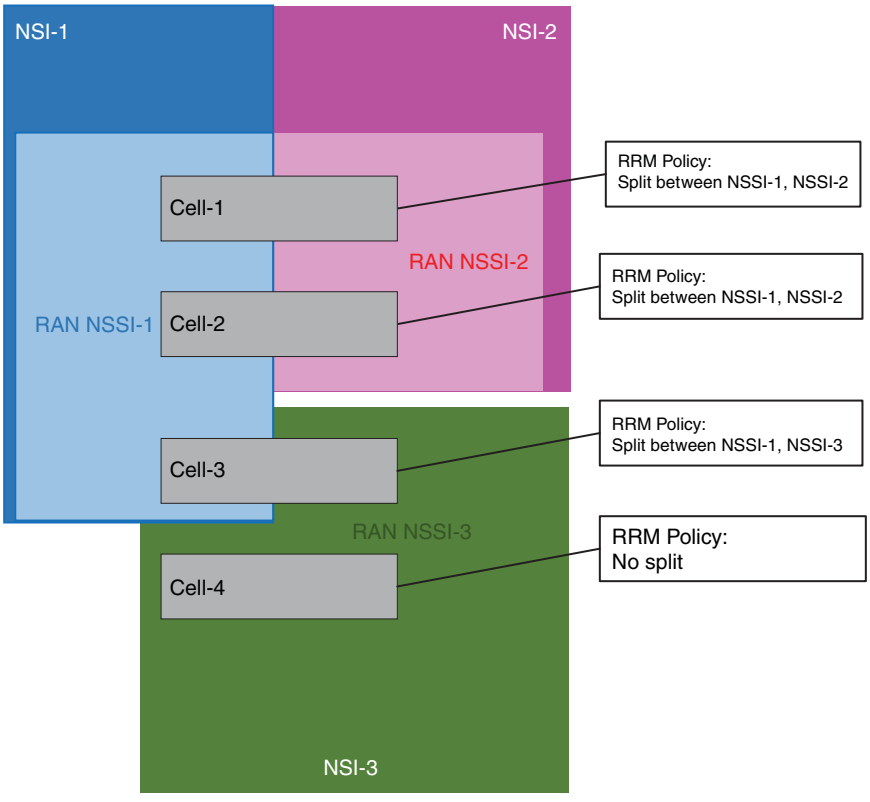



**Figure 6.5.6** Slicing support in RAN: radio resources management policy.

#### 6.5.8.2 Support of Slicing Management in RAN Provisioning Service

Figure 6.5.6 shows one example of slicing support in the RAN NF provisioning service, which is referred to as "RRM policy settings."

3GPP TS 28.541 defines a set of parameterized policies encoded in the attributes of the object NRCellCU, which belongs to the gNB-CU. The rRMPolicyType attribute designates the selection of the policy to be applied. For example, the policy with rRMPolicyType = 0 specifies a list of the S-NSSAI codes identifying certain NSIs, and the list rRMPolicyRatio contains the values interpreted as a target percentage of PRBs to be allocated to the corresponding NSIs. The target values are defined as average over an implementation-specific time interval. The sum of the configured values is not necessarily equal to 100.

Table 6.5.5 provides an example of RRM policy configuration. Cell 1, Cell 2, and Cell 3 provide support for two NSIs (two RAN NSSIs) each, while Cell 4 is dedicated to a single NSI (NSSI-3). In cells 1–4, the RRM policy settings can be configured to set the target percentage of PRB use for every particular NSSI.

Note that in Cell 4, not all of the bandwidth is allocated to the NSSI-3: the remaining part of the capacity will be used for the services not associated with any NSI.

**Table 6.5.5** Example RRM policy configuration.

| Cells | Target percentage | | |
|---|---|---|---|
| | NSSI-1 | NSSI-2 | NSSI-3 |
| Cell-1 | 60 | 20 | N/A |
| Cell-2 | 50 | 50 | N/A |
| Cell-3 | 20 | N/A | 40 |
| Cell-4 | N/A | N/A | 75 |

#### 6.5.8.3 Configuration and LCM of NSSI and NSI

The concept of management services (3GPP TS 28.533, clause 4) was applied to slicing management as well. In particular, the provisioning services described above enable configuration management (CM) and LCM of NSIs and NSSIs, specifically:

- a list of operations and notifications ("Component A") is defined for the provisioning management service for the NSI and NSSI as well as associated NFs;
- information models ("Component B") (alternatively called NRM), are defined for NSI and NSSI, as we describe below, see also 3GPP TS 28.541 for details.

In the case when an NF is shared between NSIs (NSSIs) there is a need to specify how the NF resources are shared. For this, slicing-related parameters (e.g. RRM policy attributes for NG-RAN) were specified in the provisioning of particular NFs.

One example of LCM is the NSI creation procedure, illustrated using the example of the slicing management deployment in Figure 6.5.6:

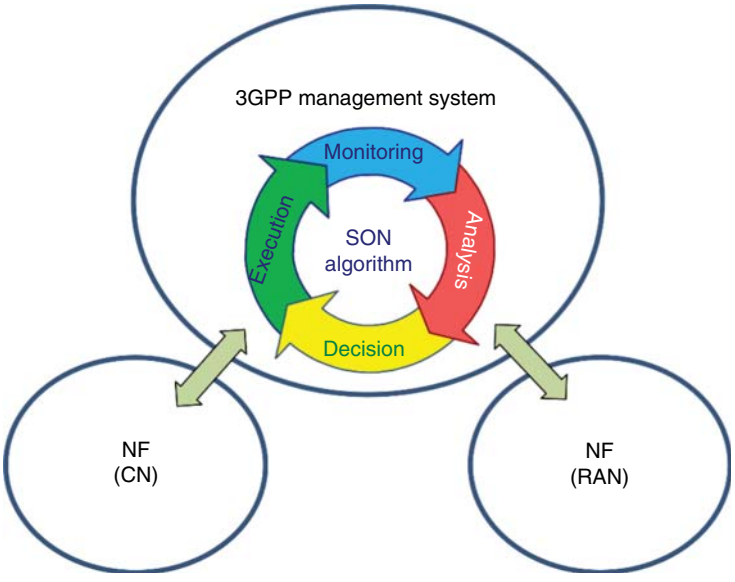



**Figure 6.5.7** Centralized SON solution (Source: Reproduced by permission of © 3GPP).

- The consumer (e.g. SON application) activates the allocateNsi operation on the NSMF; the activation operation carries the parameters that specify targeted NSI properties, such as service profile parameters.
- The NSMF determines the NSSIs to be created, such as CN NSSI and RAN NSSI, and activates the allocateNssi operations on the CN NFMF and RAN NFMF to incorporate the needed NFs into the corresponding NSSIs. Additionally, the NSMF calls the transport network NSSMF to establish the topology of the NSI and NSSIs and configure the links in-between.
- The CN NSSMF and RAN NSSMF determine the configuration parameters for the NFs to be assigned to the corresponding NSSIs and perform the configuration, with or without CN NFMF and RAN NFMF. Furthermore, the TN NSSMF configures the elements of the transport network in accordance with the transport network-related parameters received.

#### 6.5.8.4 NSI and NSSI Information Models (NRMs)

The NSI information model (NRM) defined in 3GPP TS 28.541 contains the following essential parts:

- The list of NSSIs that belong to the NSI, e.g. 5GC NSSI and NG-RAN NSSI;
- The list of service profiles supported by the NSI.

Similarly, the NSSI NRM includes:

- The list of the component NFs and/or the list of constituent NSSIs;
- The list of service profiles supported by the NSI.

The following information is included in the service profile:

- Extensive set perfReq of attributes representing the performance requirements.
- NSI (NSSI) capacity specifications such as maximum number of UEs (maxNumberofUEs) that the NSI should support.
- NSI availability.
- UE mobility level.
- The coverage provided by the NSI, specified in the format of the list of tracking areas.
- QoS properties, which may include indicators like E2E latency, along with such requirements as experienced data rate, area traffic capacity (density), and the information of the UE density. The set of QoS properties may differ from those of the standardized scenarios of eMBB, URLLC, and mIoT.

### 6.5.9 SON in 5G

#### 6.5.9.1 SON Evolution

Similarly to 4G, 5G supports SON functionality. SON functionality is often categorized as distributed SON (D-SON), carried out by NG-RAN nodes using control-plane messages, for example, Xn-AP, and centralized SON (C-SON), managed by OAM. C-SON is a closed loop optimization process, which performs collection of network performance data, analyses, and potentially network node reconfiguration, with the goal of improving network performance. In the current section we focus on C-SON, as it is performed by OAM.

5G SON contains similar functionality to that of 4G, which has evolved to address new features introduced in NG-RAN and NR. Furthermore, in contrast to 4G SON, in 5G the SON capabilities are extended to optimization of the E2E service quality indicators, in both slicing and non-slicing scenarios.

There are multiple aspects in 5G that require additional efforts compared with 4G for network automation, specifically:

- Increased network densification
- Challenging QoS requirements
- Complex radio technology
- Slicing
- New RAN architecture options.

As mentioned in Chapter 2, 5G network deployments are likely to be more dense, compared with 4G, at least when mmWave frequency bands are used. This is because in these bands the cell radius is significantly smaller (sometimes 0.1×) compared with lower frequency bands currently used in 4G. Therefore, to provide coverage in certain areas in the mmWave band, 5G may need 100 times more base stations than 4G.

Furthermore, the 5G network, including NG-RAN and 5GC, should provide support for a variety of network services with extremely challenging QoS properties. For example, 3GPP TS 23.501 specifies 10 ms latency for low latency eMBB applications such as AR and 5 ms for electricity distribution high voltage. Another example is real-time control for discrete automation with requirement of ≤1 ms E2E latency. To satisfy such stringent requirements, 3GPP selected significantly more complex radio technology compared with 4G (e.g. dynamic numerology and slot structure, massive multiple-input multiple-output (MIMO), beamforming, etc.). For details about NR physical layer refer to Section 3.5.

New features requiring SON enhancements are not limited to the physical layer. For example, the following NG-RAN and 5GC functionalities have SON impacts as well:

- Network slicing, which allows an operator to serve multiple tenants (with potentially very different QoS requirements) on the same NG-RAN and 5GC infrastructure.
- NG-RAN centralization with various split deployment options (described in Chapter 4) and virtualization (described in Section 6.2).

#### 6.5.9.2 "Legacy" SON Use Cases

The E-UTRAN SON functionalities introduced in Release-8 to Release-11 are still applicable to NG-RAN, as shown in Table 6.5.6.

Furthermore, the following "legacy" D-RAN SON functionalities remain relevant in 5G:

- Automatic Neighbor Relations (ANR) management (including automatic X2 and Xn interfaces setup);
- Physical cell ID (PCI) configuration and PCI conflict resolution;
- Load balancing;
- Inter-cell interference coordination;
- Random access optimization;
- Centralized capacity and coverage optimization;
- Self-healing;

**Table 6.5.6** Legacy SON functionality.

| Release-8 | Release-9 | Release-10 | Release-11 |
|---|---|---|---|
| Automatic inventory | Mobility robustness/hand over optimization | Coverage and capacity optimization | Handover optimization |
| Automatic software download | RACH optimization | Enhanced inter-cell interference coordination | Coverage and capacity optimization |
| Automatic neighbor relation | Load balancing optimization | Cell outage detection and compensation | Coordination between various SON functions |
| Automatic physical cell ID (PCI) assignment | Inter-cell interference coordination | Self-healing functions | |
| | | Minimization of drive testing | |
| | | Network energy saving | |

- Coordination between the C-SON and D-SON;
- SON for Active Antenna System (AAS)-based deployments;
- Trace and Minimization of the Drive Test (MDT);
- Mobility robustness optimization;
- NG-RAN energy saving.

Many of the D-SON features mentioned above have been defined in Release-16, with work to be continued in Release-17 to specify the remaining ones (3GPP SON/MDT enhancements). Naturally, the legacy SON features being introduced into 5G are extended to support NR technology specifics, such as flexible numerology, massive MIMO, beamforming, and network slicing.

#### 6.5.9.3 Multi-Domain SON with E2E Optimization

In contrast to 4G, which primarily supported RAN SON, the SON in 5G covers also the core network domain and, to some extent, the transport network; the latter including backhaul and fronthaul connections (see Section 6.6 for details about the transport network).

3GPP TS 28.861 defines three SON variants:

- C-SON, which is fully based on OAM;
- D-SON, which is based on control-plane messages exchanged between NG-RAN nodes;
- Hybrid SON combines both C-SON and D-SON.

Figure 6.5.7 shows C-SON architecture, which is the primary focus of this section.

As mentioned above, in the C-SON solution, the SON algorithm is running in the NMS. The management system collects network performance data and computes the KPIs. The received values are compared with the targets set by the network operator. As the result of comparison, certain actions can be performed by the C-SON system on the network nodes,

using the provisioning services. Closed loop optimization therefore includes network monitoring, analysis of the monitoring outcome, followed by execution of necessary actions.

Unlike 4G SON, the 5G C-SON is connected to both RAN and core network, so that the data collection and actions can be performed on both. With the addition of the transport network, this provides the SON algorithm with E2E visibility of the network and allows E2E service optimization. This approach can be also used for optimization of network slicing, if NSI-level KPIs are used.

E2E service assurance (e.g. for video quality indicators) is beneficial for the 5G networks; however, implementation of this requires support for cross-domain service optimization in SON. In addition to classic SON inputs, such as NE-level KPIs and performance measurements, the E2E SON additionally uses Quality of Experience (QoE) indicators and key quality indicators (KQIs).

Figure 6.5.8 shows the principles of the cross-domain service optimization function, in which E2E SON is connected to multiple network domains: NG-RAN, MBH, 5GC, Gi-LAN, IP services, NFV, and transport. E2E SON collects performance data and sends configuration commands to these domains, with the goal of achieving the optimization targets, for example, to monitor and optimize the SLAs for different services such as IoT, video, and voice, in an automated closed loop manner.

The reports collected from multiple domains, which may be related, for example, to the video data flows crossing all network domains, are aggregated into E2E QoS indicators. These indicators can be computed with granularity of the individual UEs (QoS flows), which can be further aggregated at the level of subnetwork or network slice. Analysis of the E2E QoS indicators may cause the E2E SON to execute certain actions performed on NFs in one or several domains.

Network performance metrics sampled at high frequency and with breakdown to individual services (flows), rather than aggregated per RAN cell, create a significant amount of

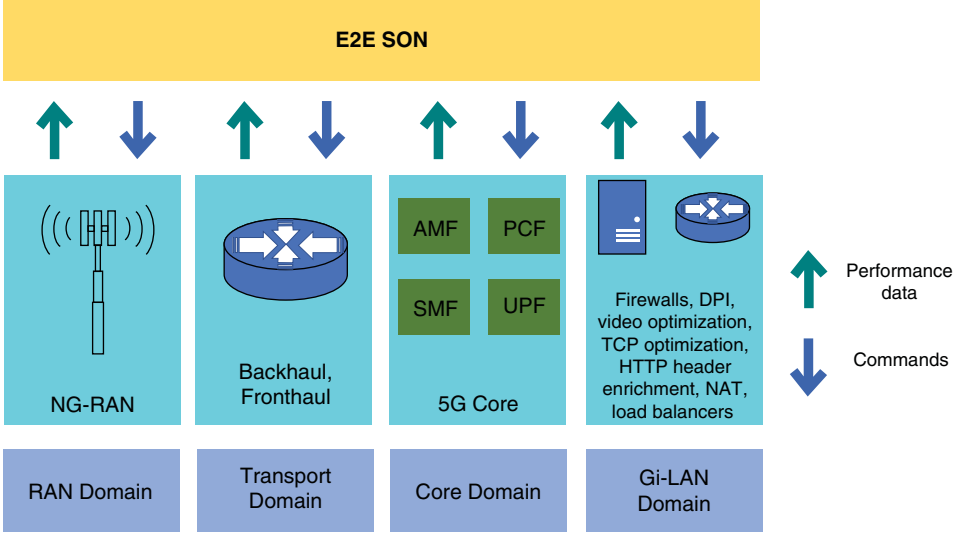



**Figure 6.5.8** End-to-end SON.

"big data" to be collected, stored, and processed. Processing of "big data" can be optimized by use of ML algorithms to identify and/or predict faults and/or performance degradation, isolate, and remedy them. These aspects of SON data processing are not specified in the standards and are left for vendor implementation.

The same E2E SON concept can be applied also to closed loop optimization of NSIs. Such implementation requires the E2E SON to be aware of the structure of the NSIs and NSSIs and of the targeted slice level performance. This information can be received from the network MnFs implementing LCM of NSIs, as explained above.

#### 6.5.9.4 SON Enablers in 5G System

5G SON mechanisms utilize the network data collection capabilities described above. In particular, both real-time performance data streaming and classic file-based reporting are used to collect performance measurements ("counters") and network traces.

Numerous data sources can be utilized by SON, for example:

- AMF, providing UE mobility event notifications and location change notification.
- NG-RAN, providing QoS parameter notifications generated, e.g. when the configured rate can no longer be guaranteed for a QoS flow.
- UPF, providing packet buffering and downlink data notification triggers.
- Charging Data Records (CDRs).
- Event Data Records (EDRs).

Additionally, the new generation of SON can use application-specific performance indicators provided by the client in the UE, for example a video player. A standard for collection of this information is under development in 3GPP (3GPP TS 28.405). Another source of application-specific performance data could be the application servers such as video streamers, for which northbound APIs defined in 3GPP TS 23.222 can be utilized.

Another important part of the 5G E2E SON support is the network data analytics function (NWDAF), which collects network data and may be hosting analytics applications for analysis of the collected information. The NWDAF makes the raw information and output of the analytics available to any NF, including SON. For example, there can be an analytics application computing the load level for a particular NSI, with input collected at the NFs and possibly transport network links constituting the NSI. The NWDAF capabilities can be used for optimization of services provided to end users, with or without slicing. When slicing is deployed, optimization of business-to-business-to-customer (B2B2C) services such as "Slice as a Service" and "NSSI as a Service" is important to make slicing commercially successful.

#### 6.5.9.5 Distributed SON

In addition to the C-SON described above, which is the main focus of this section, NG-RAN also supports D-SON. In contrast to C-SON, D-SON does not require a centralized management entity and is typically implemented in a distributed fashion between NG-RAN nodes over control-plane network interfaces (e.g. Xn-C and NG-C).

In Release-15, D-SON primarily enables dynamic configuration of Xn and NG interfaces, ANR, and network energy saving. It has been further enhanced in Release-16 to support Mobility Robustness Optimization (MRO) (intra and inter-system), mobility load balancing

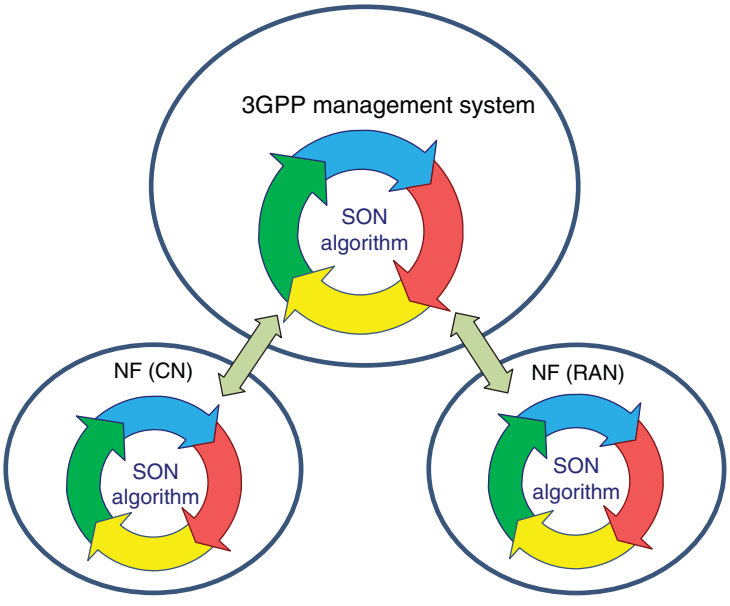



**Figure 6.5.9** Hybrid SON solution (Source: Reproduced by permission of © 3GPP).

(MLB) (intra-system), Random Access Channel (RACH) optimizations, and MDT, and additional enhancements are expected to be added in Release-17.

As the present section focuses on OAM, D-SON is considered outside of its scope. For further details on D-SON, refer to 3GPP TS 38.300 and TS 38.423.

#### 6.5.9.6 Hybrid SON

The hybrid SON solution expands the C-SON solution by integration of D-SON functions running in the network nodes, such as gNB. While C-SON has the advantage of visibility in multiple domains and centralized analytics capabilities, these come at the cost of certain delay, which may make it unsuitable for resolving problems requiring a fast response. D-SON on the other hand can address this issue, at least when the problem is local to a particular NF. As both capabilities are likely to be needed in a real network, hybrid SON can be used, as shown in Figure 6.5.9.

### 6.5.10 Further Reading

More information on RAN management can be found in the 3GPP specifications listed below. Here we provide a short explanation of which SON aspects are described in which specifications (out of many, which are relevant to SON):

- TS 28.530, Management and orchestration; Concepts, use cases and requirements.
- TS 28.531, Management and orchestration; Provisioning:
  - Management services for network slice provisioning;
  - Coordination with transport network, interaction with ETSI MANO facilities;
  - RESTful HTTP-based solution set of provisioning.

- TS 28.532, Management and orchestration; Generic management services:
  - Generic provisioning management service; operations and notifications; ManagedEntity role;
  - Generic fault supervision management service;
  - Mapping of operations and notifications onto the RESTful HTTP-based solution set of provisioning;
  - RESTful HTTP-based solution set of fault supervision;
  - OpenAPI specification: description of the capabilities of the management service; the OpenAPI document is represented in the JSON format code.
- TS 28.533, Management and orchestration; Architecture framework:
  - Management framework: definition of the management service and its components A, B, and C;
  - Management architecture reference model.
- TS 28.540, Management and orchestration; 5G Network Resource Model (NRM); Stage 1:
  - Concept of management support for NR and NG-RAN deployment scenarios: no split, two-split and three-split;
  - MR-DC;
  - 5GC support in network management.
- TS 28.541, Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3:
  - Information model definitions for NR NRM for all split options: information entities, definition of classes and their attributes;
  - Information model definitions for 5GC NRM;
  - Information model definitions for network slice NRM;
  - Solution sets with the code for:
    - XML-based 3GPP NR and NG-RAN NRM solution set (annex C);
    - JSON-based 3GPP NR and NG-RAN NRM solution set (annex D);
    - YANG-based 3GPP NR and NG-RAN NRM solution set (annex E).
- TS 28.545, Management and orchestration; Fault Supervision (FS):
  - Fault supervision management services components for NSI and NSSI;
  - Procedures for fault supervision management services.
- TS 28.550, Management and orchestration; Performance assurance:
  - Measurement job control-related operations;
  - Performance data streaming-related operations;
  - Performance assurance service components;
  - RESTful HTTP-based solution set of performance assurance specific operations and notifications: mapping of operations and notifications.
- TS 28.552, Management and orchestration; 5G performance measurements:
  - Performance measurements for gNB, AMF, SMF, UPF, PCF, and UDM;
  - Common performance measurements for NFs;
  - Measurements related to E2E 5G network and network slicing;
  - Virtualized resource usage measurement;
  - Monitoring of particular KPIs in NG-RAN and 5GC (annex A, informative).

- TS 28.554, Management and orchestration; 5G end to end Key Performance Indicators (KPIs):
  - E2E KPI definitions, per category:
    - Integrity KPIs;
    - Utilization KPIs;
    - Retainability KPIs.

## References

3GPP Technical Specification 22.261 (2019). Service requirements for next generation new services and markets. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.405 (2019). Management of Quality of Experience (QoE) measurement collection; Control and configuration. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.530 (2019). Management and orchestration; Concepts, use cases and requirements. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.531 (2019). Management and orchestration; Provisioning. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.532 (2019). Management and orchestration; Generic management services. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.533 (2019). Management and orchestration; Architecture framework. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.541 (2019). Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.554 (2019). Management and orchestration; 5G end to end Key Performance Indicators (KPI). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.552 (2019). Management and orchestration; 5G performance measurements. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 28.622 (2019). Telecommunication management; Generic Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS). Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Report 28.861 (2019). Telecommunication management; Study on the Self-Organizing Networks (SON) for 5G networks. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 32.101 (2019). Telecommunication management; Principles and high level requirements. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 38.300 (2019). Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Specification 23.222 (2019). Common API Framework for 3GPP Northbound APIs. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP RP-193255 (2019). New WID on enhancement of data collection for SON/MDT in NR, RP-193255.

## 6.6 Transport Network

*Yaakov (J.) Stein, Yuri Gittik, and Ron Insler*

*RAD Data Communications, Ltd., Israel*

In this section we discuss various transport network technologies used for backhaul (i.e. connecting NG-RAN to 5GC), fronthaul (i.e. connecting NG-RAN to RRG), and midhaul (connecting network nodes in split NG-RAN architecture). This is illustrated in Figure 6.6.1, where the usage of various NG-RAN functional split architectures (see Chapter 4) is also shown.

Figure 6.6.1 shows backhaul, midhaul, and fronthaul as straight lines connecting network nodes, but this is a bit misleading as in practice any haul (referred to as xHaul in this book) can have a complex network topology and many links, using many different technologies. This aspect is often overlooked in technical specifications (e.g. in 3GPP) and other network architecture documents. In the present section we discuss this topic (the transport network) in more detail.

### 6.6.1 Key Ideas

- The transition from 4G to 5G will strongly impact the transport network, due to requirements for higher data rates, lower latency, enhanced reliability, energy efficiency, and heightened dynamicity.
- Choice of 5G RAN segment defined by functional split (fronthaul through midhaul to backhaul) and RAN planning factors (density, distances, placement of Points of Presence [PoPs], etc.) deeply influence new requirements for transport.
- In addition, delivery of disparate services (eMBB, URLLC, and mMTC) requires different transport attributes.
- Distributed edge computing, whether for enabling virtual RAN (vRAN)/cloud RAN (cRAN) operation, serving the disaggregated transport network, or hosting end user

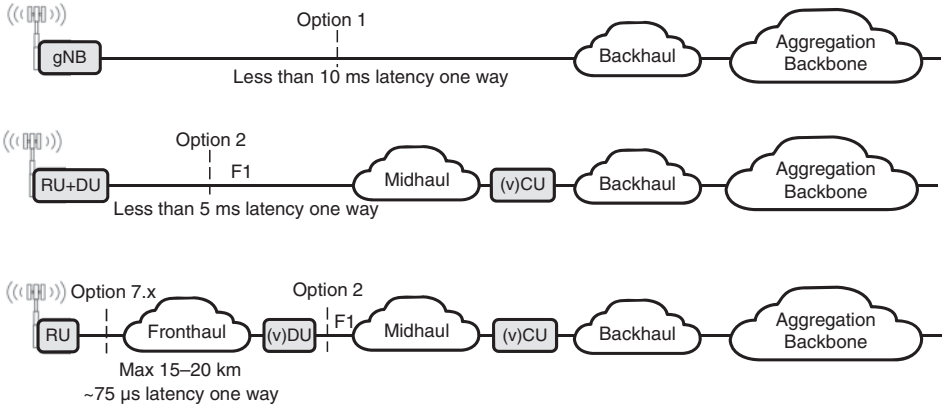



**Figure 6.6.1** Backhaul, midhaul, and fronthaul (xHaul) transport networks.

applications, drives yet more sophisticated transport network designs with compute resources and connectivity of distributed physical and virtual components.
- Integration of transport and computational components will produce a new xHaul infrastructure.
- To target higher data rates, new fiber Ethernet technologies (e.g. N*10G, 25G, and higher rates) will supplant the GbE links prevalent for 4G, although some passive optical network (PON) solutions may only be sufficient for the near term.
- To tackle lower latency, new time-sensitive networking (TSN) and deterministic networking (DetNet) technologies are being introduced.
- New protection switching, fast reroute, self-healing, loop free alternatives, and frame replication technologies may be employed to address reliability challenges.
- Co-existence of traffic with wildly diverging requirements mandates support for network slicing across the transport network.
- Backhaul networks for 5G may be based on wavelength division multiplexing (WDM), optical transport network (OTN), PON, Carrier Ethernet, multiprotocol label switching (MPLS), pure IP, segment routing (either MPLS or IPv6 variety), and vertical/horizontal combinations of these; and may be managed using distributed control protocols or centralized management (e.g. SDN).
- Introduction of 5G into existing (brownfield) transport networks will necessitate migration strategies.
- All of the above will transform transport networks, whether owned by the mobile operator or by a wholesale provider of xHaul services; in most cases the services will be terminated by an enhanced cell site gateway.
- The increased number of cells and the stricter time accuracy requirements will necessitate innovative timing solutions.
- Network dynamicity and slicing require tighter interworking between NMSs (and/or SDN controllers) of transport networks and mobile networks.

### 6.6.2 Market Drivers

To outline major market drivers that steer and shape the xHaul development, let's first analyze the key changes from 4G mobile backhauling to 5G xHaul (more precisely, combined 4/5G xHaul). Figure 6.6.2 depicts a high-level picture of such migration.

4G mobile backhaul (we can neglect the 4G fronthaul as it was commercially deployed in only a few countries) might be characterized by the following essential qualities:

- On the whole, connecting radio sites to centralized locations that host contents and applications. The exceptional case was the MEC, which was introduced to bring contents closer to users (e.g.with local CDN), but not successful in 4G.
- Single class-of-services for all applications.
- Static pre-provisioned connections ("pipes") with star or ring topology; typically, not direct X2 connections between eNBs, but via an aggregation node.
- Semi-static network management.

5G xHaul dramatically changes these basic features and entails new ones, specifically:

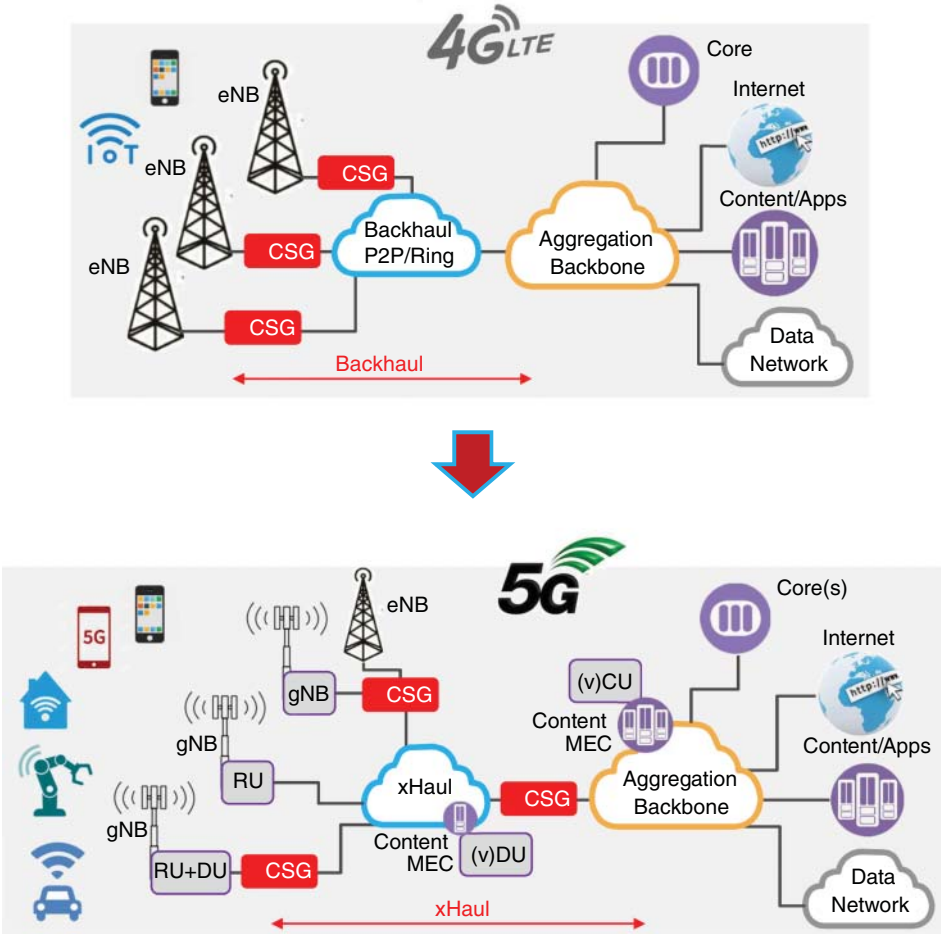



**Figure 6.6.2** Evolution of the backhaul transport network.

- Connectivity to highly distributed contents and applications that require various transport attributes:
  - RAN-related – fronthaul (e.g. CPR/eCPRI) delay, packet delay variation (PDV), HARQ delay, different data rates for different NG-RAN functional splits (see Chapter 4).
  - Application-related – eMBB, URLLC, mMTC use cases with higher data rates, lower latency, lower packet loss rate (PLR), and higher reliability.
- Multiple classes-of-service assured by E2E service and network slicing.
- Orchestrated dynamic connectivity for on-demand mesh topology embracing multiple physical and virtual components.
- Zero touch provisioning (ZTP) and automation for dynamic E2E slicing support.

Such substantial changes furnish major challenges (both engineering and economical) for a mobile operator to define a smooth migration path for their transport network. In most cases the existing 4G backhaul and aggregation networks will be incrementally enhanced

and upgraded to reach the 5G xHaul end-game. Recent practices indicate that such migration will usually be executed in phases; an example scenario consisting of the following phases:

1. Upgrading to support higher data rates.
2. Integrating the transport network with edge computing (i.e. MEC).
3. Decreasing latency and increasing reliability.
4. Adding support for slicing (different traffic types over a single network infrastructure).
5. Assuring higher density of UEs for IoT (optionally integrated with 4G IoT deployment).

### 6.6.3 Defining the Problem

At the highest level of abstraction, a 5G network consists of three entities:

- 5G UE;
- 5G NG-RAN, incorporating gNBs and ng-eNBs (described in Chapter 4) and other elements to be described later in this section;
- 5GC, described in Section 3.2.

In many cases there is also a catch-all fourth entity – external DNs or server platforms.

5G communications are carried out by inter-connecting these entities. 3GPP standards (3GPP 38.401) specify that the RAN consists of a radio network layer (RNL) and a transport network layer (TNL), where the TNL provides services for transport of both user plane and signaling. 3GPP specifications deal in minute detail with all aspects of RNL connectivity between the gNB and 5GC (and among gNBs), but severely underspecify[12] the TNL aspects of these, viewing connections from the cell site to the core as ideal transport pipes.

The connection between the 5GC and external networks or servers generally utilizes the IP suite, which is defined by standards produced by the Internet Engineering Task Force (IETF). It may additionally entail Ethernet protocols, defined by the Institute of Electrical and Electronics Engineers (IEEE) and other organizations.

The connection between the gNBs and the core is called the NG interface and gNBs can also be interconnected by interfaces named Xn (see Section 3.3). Transport of data over either the Xn or NG interface is conventionally known as *backhaul*, in line with the terminology of previous generations of mobile communications.

As described in Chapter 4, gNB may be further split into a number of logical network nodes, thus resulting in more interfaces for the transport network to carry.[13] For the purposes of this chapter we consider a 5G base station that can be decomposed into an RU, a DU, and a CU, in which case, we define the F2 interface (which is equivalent to the O-RAN fronthaul interface described in Section 4.5) between the remote RU and the rest of the gNB, and call the transport segment between these two units *fronthaul*. If the CU is physically detached from the DU, the two units are connected via an interface known as F1 (see Section 4.2), and the transport segment over this interface is sometimes termed *midhaul*.

12 It is generally assumed in 3GPP that such details (e.g. integration between a transport network and a 3GPP radio network) are addressed by each operator in a proprietary manner.

13 5GC is also split into logical network nodes (see Section 3.2), but these are not of interest here, as 5GC often runs in a data center where fiber is readily available.

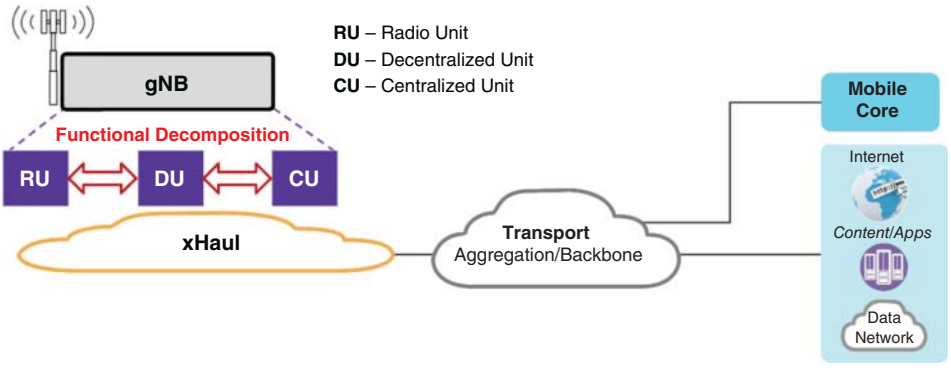



**Figure 6.6.3** Decomposition of the 5G base station and the resulting xHaul interfaces.

As shown in Figure 6.6.3, transport in a 5G system may involve only backhaul, but may additionally entail either fronthaul, or midhaul, or both, under the umbrella of the generic term xHaul. However, the transport network providing these 5G haulings will typically be required to provide additional transport services. Such services will generally incorporate those that bestow essential support for the 5G functionality, such as control and management of flows and time/frequency synchronization. In addition, the same network will often need to support 4G backhaul, and even previous generations (e.g. 3G voice). While there is a justifiable tendency to disregard such services as negligible in volume compared with that consumed by 5G, they may have specialized requirements, such as non-packet constant bit rate, specific physical interfaces, or stringent delay budgets. Finally, the transport network may also furnish services unrelated to mobile communications, such as Internet access or critical infrastructure communications.

3GPP standards tend to consider transport as a minor function that effortlessly delivers data over the named interfaces with no availability failures, data rate restrictions, burdensome latency, synchronization glitches, or other degradations. Unfortunately, this is not the case. Even well-engineered transport networks have limitations and occasionally fail to live up to designed requirements. The limitations of transport networks and the best practices to reach the highest levels of performance are the subjects of this section.

Now that we understand the different transport segments, we need to appreciate the challenges presented by each such one. There are several types of requirements, the most important of which are network topology (e.g. star, mesh), traffic capacity, traffic characteristics (packet size, burstiness, etc.), delay (E2E propagation latency), reliability (availability and time-to-repair), dynamicity (how quickly services need to be set up and torn down), and synchronization (frequency and time recovery). There may be additional requirements, such as provision of distributed computational platforms, and co-existence with or migration from existing network infrastructures.

### 6.6.4 The Physical Layer

In this subsection we will discuss a plethora of physical layer technologies that have been proposed to face the challenges presented by 5G transport. Figure 6.6.4 provides a preliminary overview of the proposed technologies and the challenges addressed by each.

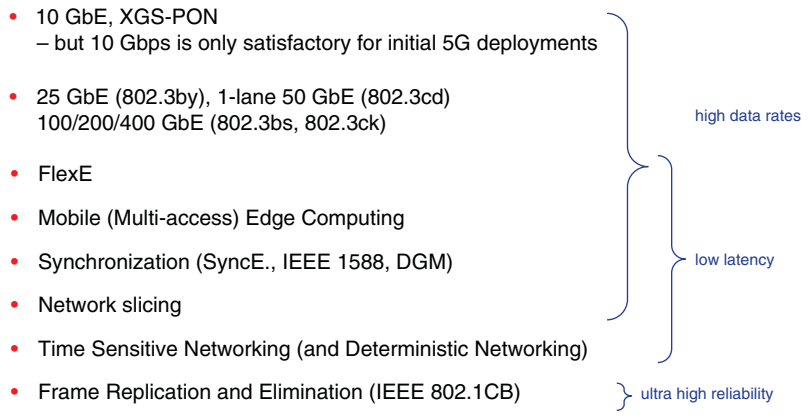



**Figure 6.6.4** Summary of mechanisms for upgrading the xHaul physical layer.

#### 6.6.4.1 Achieving the Required Data Rates

While achieving a 10- to 100-fold increase in data rate is well understood to be challenging for the air interface, it is far from trivial for xHaul transport as well. LTE backhaul is mostly based on GbE interfaces (whether fiber or point-to-point microwave) with even lower-rate synchronous digital hierarchy (SDH) and even plesiochronous digital hierarchy (PDH) still in common use. Physical links supporting data rates of up to 1 Gbps will not suffice for 5G needs (and even for LTE-A, which may approach 3 Gbps). The duct itself, whether active/passive optical or microwave, may require technology upgrading.

In the near term (Release-15 with eMBB traffic) it is expected that cell site backhauling (i.e. the NG interface) will peak at about 5 Gbps. Such data rates are readily handled by upgrading 1 Gbps Ethernet transport links to 10 Gbps ones, an upgrade that involves limited CAPEX (the existing fiber plant will support 10G, and 10G SFP+ [enhanced small form-factor pluggables] are no longer appreciably more expensive than 1G SFPs) and insignificant additional OPEX (less than 1 W difference at 10 or 20 km). However, this remedy comes with three caveats:

- The first is that those deployments that exploit TDM-based PON technologies will probably have to migrate to active networks. The first available PON standard is XGS-PON [ITU-T G.987.x] at 10 Gbps, which limits the ODN to a single 2 : 1 split. On the other hand, PONs that utilize wavelength-based multiplexing, including NG-PON2 [ITU-T G.989.3], which can reach 40G rates, will be viable for some time.
- The second is that 10 Gbps does not suffice for all functional splits, in particular, the F1 interface will require higher data rates even for initial NR deployments; and of course F2 fronthaul traffic may be much higher in volume.
- The third relates to futureproofing. The quoted 5 Gbps rate is for initial deployments; over time, and especially with deployment of mmWaves and system channel bandwidths of 200 MHz and above, the traffic to and from cell sites is expected to dramatically increase. 10 Gbps will probably suffice in most cases for the first two to three years of 5G deployment, and thus is an attractive upgrade option for existing networks. However, it is questionable whether it makes sense to design new networks (using 5 Gbps transport) that will require re-engineering in several years' time.

For those cases where 10 Gbps does not suffice, multiplexing of multiple 10G links may make sense. Such multiplexing can be accomplished via link aggregation (LAG) (IEEE 802.1AX), but only when there is some criterion that consistently maps flows and fairly distributes bandwidth between them. For backhauling, standard hashing techniques should suffice, but these methods may not be applicable at the low-level split (Section 4.5) with compressed headers.

The next rate to be considered could be the conventional 100 Gbps, although this capacity should not be needed for single cell sites. 100G is already used for LTE second-level aggregation networks, and will be required for 5G first-level aggregation networks. Advancing from 10G to 100G involves a major jump in CAPEX, as 100 Gbps currently comprises four lanes of 25 Gbps. Even for single-mode fiber where these lanes are instantiated as different wavelengths and not individual fibers, 100G requires a quad small form-factor pluggable (QSFP) with four lasers, making it significantly more expensive than 10G. In addition, 100G standards do not presently support bidirectional traffic on a single fiber (BiDi), and thus require twice the number of fibers when compared with 10G employing BiDi. In addition, NEs that can forward at 100G wirespeed are also significantly more expensive than comparable lower-rate ones. Power costs for 100G are not really that much higher than for 10G.

The IEEE, while standardizing 100 Gbps, included an intermediate rate of 40G (clause 80 of IEEE 802.3), but there seems to be little reason to consider this rate. For data centers 40G made sense, but was defeated in the market by 56G Infiniband. For backhauling it presents few advantages compared with deploying 100G, using four lanes and thus being about as expensive as 100G.

In 2016 the IEEE approved amendment to 802.3, which standardized a rate of 25 Gbps. A 25G link corresponds to a single lane of the 100G standard, and was thus born with operational experience. Like 10G, 25G interfaces are supported by inexpensive SFP+, and do not require a QSFP. It is thus reasonable to assume that 25G will supplant the 10G links used for initial 5G deployments. Support for 25G and higher rates has recently been added to OTN standards (ITU-T G.709) as well (ITU-T G.709-Amd3).

For yet longer-term cell site deployments, for aggregation networks, and for lower functional splits, multiplexing of 25G links will be used. The problem of distributing traffic over the 25G links can be solved here by using an emerging standard called FlexE. The Optical Internetworking Forum (OIF) published the original FlexE Interoperability Agreement in 2016 (and an updated one in 2017). Among a host of other features, FlexE enables *bonding* of an arbitrary number of 25G links.

Future developments will further increase data rates available for xHauling. The IEEE is currently working on enhancing the single-lane data rate to 50G (and hence the four-lane rate of 200 Gbps instead of 100G), and later to a full 100G (and hence a four-lane rate of 400 Gbps) (IEEE 802.3 cd, IEEE 802.3bs).

#### 6.6.4.2 Achieving the Required Latencies

The requirement for low latency for backhauling is ultimately derived from the E2E delay tolerated by the user application (unless this delay is unimportant, in which case the delay tolerated by control-plane signaling, for example RRC to PDCP configuration estimated at about 10 ms, dominates). For the general eMBB use case, this can usually be hundreds of milliseconds, while interactive voice or voice+video communications may suffer at more

than tens of milliseconds. More demanding applications, such as gaming, AR/VR, and V2x, may require one-way E2E delays as low as 1 ms. Some factory automation applications demand delays from sensor to programmable logic controller as low as 0.25 ms (in addition to extremely low PLRs); such low delays mandate local termination. Ultra-low delay will also be critical for recent innovations in manufacturing technologies, thought of as the fourth industrial revolution and thus called Industry 4.0, which introduce cyber-physical systems and cognitive computing. It should be noted that, from all these allowed E2E delays, one needs to deduct terminal processing times. Other important applications having stringent delay requirements, such as professional audio and video, are discussed in RFC 8578.

Fronthaul and midhaul have additional, usually more stringent, latency requirements, deriving from operational and technological constraints rather than from the user application. For example, for midhaul with functional split options from 5 and below (defined in Section 4.1), HARQ response times put limitations on the transport. CPRI fronthaul mandates latencies of up to 100 μs.

When discussing the transport network, we need to differentiate between low average delay and bounded delay. In certain applications, including interactive audio or video, having a sufficiently low average delay is sufficient, with occasional high delay packets considered lost and subject to packet loss concealment. In most cases of relevance here we need to focus on guaranteed (i.e. worst case) upper bounds to delay.

Achieving low network traversal latencies generally entails combining two strategies:

1. Finding potentially low-delay paths through the network (e.g. paths with short links, minimal number of active forwarding elements, etc.).
2. Ensuring low packet residence time for *express* traffic packets (i.e. packets whose forwarding must be expedited) at the forwarding elements.

The former strategy may be accomplished using SDN techniques involving maintaining a network topology graph at a centralized computational resource, and performing graph optimization algorithms. Once a feasible path has been found, it must be deployed, and protected against failures. Protection mechanisms may be E2E (requiring an alternative feasible *backup* path) or local (requiring bypass alternatives for all possible failures).

The latter strategy requires identifying and prioritizing *express* traffic packets, and may involve resource reservation at the forwarding elements.

One-way E2E transport latency is made up of several contributions. The first is the physical propagation latency of about 1 μs for 200 m of fiber or 300 m of point-to-point microwave. The second is the residence times in each active NE, which we can define as the interval from the first bit arriving at the NE until the last bit exits it. Each residence time is composed of packetization time (time for all bits to arrive at line rate), processing time (time for memory accesses, to read header fields, to classify packet, to look up forwarding information, etc.), queuing time (the time the packet waits its turn to be transmitted), and depacketization time (time to clock all the bits out). For low-priority packets, the queuing time dominates (especially in congested NEs); for highest-priority packets, the queuing time is reduced to *head-of-line* blocking time (the time for the currently transmitted packet to finish).

To understand these contributions, consider designing to minimize delay using standard technologies. We'll consider a hypothetical case of an RU directly connected over a 10 km

fiber link to a DU, which in turn is midhauled over 40 km via 10 hops to a CU, and a server directly connected to the 5GC but 250 km and 10 routers away from the CU. Adding the components we find 200 km of fiber contributing only a single millisecond. On the other hand, the 10 midhaul switches may contribute 20 µs each, and the 10 core routers 200 µs each, for an additional 2.2 ms. We are thus over 3 ms in the absolute best case. If midhaul or 5GC become congested the numbers will be much greater, and for servers not directly hooked up to the 5GC, the additional routers may add significantly higher latencies.

Physics tells us that the only way to minimize the propagation latencies is to reduce the distance traveled. One way to achieve this is to employ virtualized applications closer to the RU (e.g. with MEC, see Section 6.4 for details), where at least the first portion of the processing is placed at the cell site or at an aggregation point. Such placement allows for extremely rapid acknowledgements, which can, for example, enable battery-powered IoT devices to promptly return to sleep mode. MEC additionally reduces core data rate requirements, since large quantities of data may be combined or summarized.

For non-priority traffic the dominant contribution to residence time in a single NE is queuing time. Traffic shaping (MEF 10.3, RFC 2475) typically adds significant delay in its attempt to avoid exceeding packet loss objectives, which are considered more significant. For TCP-based traffic this "bufferbloat" (CoDEL) also expresses itself in reduced data rates, since the bandwidth-delay-product bounds its throughput.

For traffic of the highest priority the dominant contribution is head-of-line blocking time, namely the time a packet waits for draining of a packet whose transmission has already been commenced. Assuming a 1500[14] Byte packet just started transmission, a highest-priority packet needs to wait up to 1.7 ms (see Table 6.6.1).

To determine the contribution to E2E latency, worst case single-switch times in Table 6.6.1 need to be multiplied by the number of switches traversed.

It is, or course, possible to *runt* the outgoing packet (i.e. abruptly stop its transmission without computing a frame check sequence [FCS], and allowing the next switch to discard the errored frame), but this would require its full retransmission and burden the next switch along the path to parse and discard it.

**Table 6.6.1** Head-of-line blocking time versus line rate.

| Line rate | head-of-line blocking time |
|---|---|
| 10 Mbps | 1.7 ms |
| 100 Mbps | 170 µs |
| 1 Gbps | 17 µs |
| 10 Gbps | 1.7 µs |
| 100 Gbps | 0.17 µs |

14 2000 Bytes is the maximum Ethernet frame size (802.3as), however, jumbo frames of up to 9000 Bytes have also been standardized.

A new mechanism that ameliorates head-of-line blocking is frame pre-emption, whereby *express* Ethernet frames (i.e. ones requiring expedited forwarding) can pre-empt the transmission of *normal* frames. Frame pre-emption along with interspersing express traffic (IET) are defined in IEEE 802.1Qbu and 802.3br, respectively.

Frame pre-emption occurs over a single link (i.e. fractional frames do not propagate through the network, but are re-assembled by the following switch), and thus requires compliant switches at both ends of the link. When an express frame arrives and a normal frame is being transmitted, the packet transmission of the normal frame is temporarily suspended, the *neighboring* switch buffers the content already received, the express frame(s) are sent and forwarded, transmission of the normal frame is continued, and finally the neighboring switch reassembles the outgoing frame and forwards it.

Reflecting on Table 6.6.1, it is obvious that at high rates frame pre-emption is not really needed to reduce delay, and its real purpose is to avoid complete starvation of normal traffic when there is an abundance of express traffic.

Some TSN mechanisms assume that all (or at least most) network forwarding elements have access to high-accuracy (sub-microsecond) timing information, obtained, for example, by use of the PTP (IEEE 1588v2). Once the entire network is thus synchronized, a new set of mechanisms becomes available that can provide guaranteed upper bounds on E2E latencies.

The base mechanism of TSN is the time-sensitive queue defined in IEEE 802.1Qbv, which mimics time-division multiple access schemes by opening and closing at precise timeslots. Timeslot schedules may be dynamically computed by a centralized management system that configures the network nodes using the Stream Reservation Protocol (SRP). In this way express traffic classes are serviced without interruption, effectively eliminating queuing delay, and rendering residence time deterministic. This enables guaranteeing upper bounds on E2E latencies.

A readily understood, but non-optimal, method of exploiting time-sensitive queues without intricate signaling is called cyclic queuing (previously called peristaltic queuing) (IEEE 802.1Qch). In this scheme all switches simultaneously forward all packets of the same traffic class in the same timeslot. Before outputting, the priority marking is incremented, so that the packets exit the next switch in the following time slot. The *E2E latency is hence the number of switches traversed times the timeslot duration*.

#### 6.6.4.3 Achieving the Required Reliability

Mobile communications were originally considered relatively unreliable, due in large part to poor coverage and the customer's understanding of the limitations of the air interface. However, due to its ubiquity, people and businesses have become more and more dependent on mobile communications services, requiring upgrading the reliability of these services. In addition, over time more and more mission-critical services have migrated to the public mobile network, including first responders, hospitals, and more recently smart city applications. Initial studies of 5G identified ultra-reliable services as one of the vertical markets that needed to be addressed.

In this subsection we will consider two related topics, availability and packet loss of the transport network. Availability is beyond doubt the most important characteristic of any communication service, since a non-available service is of no use. The golden standard for

telephony service has always been 5 nines, which translates to less than 4 and a half minutes of downtime per month. Some cloud-based services now promise 6 nines, that is, less than half a minute of downtime per month. Additionally, the 4 minutes and 23 seconds of 5 nines, or even the 26 seconds of 6 nines, are not allowed to occur in a single duration. The golden standard here is 50 ms from failure detection to repair.

High reliability and fast repair is obtained in transport networks today by one of two self-healing mechanisms, which we may term automatic protection switching (APS) and fast reroute (FRR), respectively. With APS, supported by Carrier Ethernet and MPLS Transport Profile (MPLS-TP), one prepares an alternative disjoint E2E path, which is called the backup path in contrast to the working path (ITU-T G.808.1). A prevalent special case of APS utilizes rings, where one way around the ring is the working path and the opposite direction is the standby path (ITU-T G.808.2). Non-ring scenarios are often referred to as linear protection.

Upon detection of a failure of the working path (e.g. through physical layer indications or via loss of several consecutive OAM continuity check [CC] messages), the traffic is sent over the backup path. In order to conform to 50 ms repair times, CC messages are often sent at rates of 100 per s, and 30 ms without receiving a CC message triggers a failover switch.

There are four main APS variants (ITU-T G.808.1):

- In 1+1, APS-protected traffic is always sent over both paths, but the destination end consistently selects packets from one path until a failover is triggered. In 1+1 no APS signaling is required and failover time is little more than the detection time, but network bandwidth is wasted on redundant traffic.
- In 1 : 1, APS-protected traffic is sent only over the working path, leaving the backup path free to carry unprotected pre-emptible traffic. Upon detecting failure of the working path, the tail end signals the head end (over an APS signaling channel) to commence sending the traffic over the backup path. This mechanism is more efficient in use of network resources, but requires both an APS signaling channel and additional time before failover switching is accomplished. 1 : 1 APS systems may revert to the original state after the working channel has been repaired, although this may cause an unnecessary short service disruption.
- Yet more efficient is 1 : n APS, where a single backup channel is used to protect n working channels, with the drawback that two failures can't be handled. 1 : n APS requires two-phase signaling, where the tail end signals the head end of failure, and the head signals back that the backup channel is available and the switch has been made. 1 : n systems will almost always revert upon repair.
- Finally, in m : n APS m working channels are protected by a smaller number n of standby channels, enabling protection in the case of up to n simultaneous failures, at the expense of a three-phase APS signaling protocol.

The second prevalent self-healing mechanism, FRR, is frequently used in MPLS core networks (RFC 4090), although recent work has extended this method to IP networks (under the name Loop Free Alternates (LFA) (RFC 5286)) and to segment routing (under the name Topology Independent Loop Free Alternates (TI-LFA) (draft-ietf-rtgwg-segment-routing-ti-lfa)). Unlike APS where E2E backup paths are prepared, in FRR mechanisms local detours are prepared to bypass failed links or NEs. In

order to bypass a failed link one prepares an alternative next hop (NHOP), while bypassing a failed NE requires preparing a next next hop (NNHOP).

A related issue for packet switched networks is packet loss. When packet loss is low (say less than one packet in a million) its effects can be ignored except in the most demanding of URLLC applications, but high loss is essentially equivalent to service failure. Best-effort Internet connections may have a PLR of about 1%, while Carrier Ethernet services routinely specify $10^{-6}$. While cellular air interfaces have very variable PLR, depending on obstructions, speed, etc., backhaul transport paths tend to have relatively stationary PLR, almost entirely due to buffer overflows in NEs along the path.

A new mechanism called Frame Replication and Elimination for Reliability (FRER) for Ethernet and Packet Replication, Elimination, and Ordering Functions (PREOF) for IP/MPLS has recently been proposed to simultaneously achieve ultra-high reliability (better than 5 nines availability) and ultra-low packet loss (IEEE 802.1CB). FRER can best be explained by starting with 1+1 APS. Similar to 1+1, FRER simultaneously sends packets over alternative paths, but it is not limited to two paths (a working path and a backup one), but rather to as many as the planner desires. Unlike 1+1 APS, FRER does not consistently retrieve packets from a working path and only start retrieving from the backup path once a failure has been declared. Instead, it functions on a packet-by-packet basis. It utilizes a per-packet sequence number (adding one if necessary), and selects the first packet with the required sequence number to arrive. This not only automatically combats failures, but compensates for erratic packet loss. However, FRER goes a step further in order to protect against multiple simultaneous failures. Packet replication is performed not only at the head end, but at intermediate switches. In order not to completely overwhelm the network with duplicate packets, intermediate switches also perform an erasure operation, whereby after forwarding a given packet, additional copies are discarded and not forwarded.

#### 6.6.4.4 Frequency and Time Synchronization

Frequency and time synchronization requirements of the NG-RAN are critical in order to assure:

- Maximizing data rates on the air interface by minimizing guard frequencies/times in order to maximize spectral efficiency, and utilizing bandwidth-boosting technologies like Carrier Aggregation (CA) and MIMO/CoMP.
- Optimizing user experience, including smooth handover (significant reduction in call drops when sync is good), and reduced experienced delay.
- Supporting user applications that rely on highly accurate timing, such as location-based services.

While frequency and time (or phase) are definitely related, reference sources and dissemination at the highest accuracies employ very different technologies. Frequency references employ physical phenomena (such as narrow spectral lines of certain elements) and frequency distribution over communications links needs to be accomplished by the physical layer. Once a frequency reference is agreed upon, a time reference identifies particular moments in the periodic output of the frequency reference, and time distribution consists of sending data labeling these instants, and compensating for the propagation latency, which at high accuracy requires hardware time stamping, physical layer symmetry, and on-path support.

As a concrete example, a primary (frequency) reference clock (PRC) according to ITU-T G811 is required to have long-term frequency accuracy of $1 \times 10^{-11}$, which will lead to a drift of up to 864 ns/day or 26 μs/month. For more demanding applications, an enhanced PRC (ePRC) according to ITU-T G.811.1 is constrained for frequency accuracy of $1 \times 10^{-12}$, which implies one tenth of these time drifts, that is, 84.4 ns/day and 2.6 μs/month. This level of accuracy may be obtained by Rubidium atomic oscillators.

A primary time reference clock (PRTC) according to ITU-T G.8272 has its internal frequency locked to a PRC and is required to keep time to within 100 ns (for a PRTC-A) or 40 ns (for a PRTC-B) of the desired time standard (e.g. UTC). Recently, the ITU-T has specified even more stringent clock types (ITU-T G.8272.1).

In order to prevent gNB transmissions deviating from their allotted frequency allocations, frequency accuracies of 50 parts per billion (ppb) for macrocells and 100 ppb for small cells (observed over a 1 ms window) are required (3GPP TS 38.104). These accuracies seem lenient as compared with PRC/ePRC levels, but are actually not trivial to obtain at the cell site.

LTE TDD macrocells have a requirement for ±5 μs absolute time error, while the accuracy requirements for TDD small cells is ±1.5 μs and certain LTE-A features may require ±500 ns accuracies (ITU-T G.8271). With 5G's scalable sub-carrier spacing, the basic requirement becomes stricter, for example, ±780 ns for 30 kHz and ±390 ns for 60 kHz. CA may put even more stringent demands on the *relative* time error (i.e. the error between neighboring base stations) (3GPP 36.104), requiring 260 ns relative time error for the inter-band non-contiguous case (half that for the intra-band contiguous case, but this requirement does not impact the transport network). MIMO may drive down the relative time error to 65 ns, and CPRI interfaces require very strict transport delay accuracy of 16 ns.

Highly accurate location-based services will require even stricter time accuracy targets. Aiming for 1 m accuracy (3GPP 22.862) will necessitate less than 3 ns of relative time difference between participating base stations. While the latest time distribution standards (IEEE 1588v3) address sub-nanosecond accuracies, such tolerances do not come without a cost.

In some areas of the world macro cells have traditionally relied on non-network sources of timing, for example, via a Global Navigation Satellite System (GNSS) such as the GPS. With the increased number of cells expected in 5G, the cost of providing independent GPS-based timing will likely become prohibitive. Moreover, the accuracy of time recovery from GNSS is limited to ±100 ns, which is insufficient for the most demanding of 5G uses. The alternative is for base stations to obtain timing from the communications infrastructure that feeds it, namely from the backhaul transport, and the physical layer of the transport network is a critical element in delivery of high-accuracy timing information.

If the transport is carried over a natively synchronous infrastructure, such as PDH, SDH, OTN, or dark fiber, then highly stable frequency is automatically retrieved by the physical infrastructure, as the physical layer requires this frequency synchronization for its own proper functioning, and obtains it using a PLL. The most common asynchronous physical layer is 802.3 Ethernet, but even Ethernet, for rates of interest here, continuously transmits bits at a constant rate (sending idle codes when there is nothing to transmit), although that rate is not locked to a frequency reference with high accuracy. SyncE (ITU-T G.8262) remedies this deficiency by applying conventional mechanisms for locking the frequency of physical layers of synchronous networks to Ethernet. The first Ethernet switch in a chain

of switches has its transmit bit rate locked to a PRC; that is, a frequency reference with accuracy of $10^{-11}$ or better. Each switch in the chain locks its internal clock onto an input closer to the PRC, and transmits its outputs accordingly. The ITU-T has specified the Ethernet Synchronization Messaging Channel (ESMC) to indicate clock quality (closeness to the PRC) and to aid avoiding timing loops (ITU-T G.8262).

Distribution of time-of-day information over a packet network requires a packet time protocol, such as the network time protocol (NTP) (RFC 5905), or the PTP (IEEE1588) (IEEE 1588v2). The latter has the potential to be more accurate than the former due to its access to physical layer time stamping (e.g. it defines packet arrival precisely as halfway up the leading edge of the first bit) and to its defining on-path support, that is, PTP-specific mechanisms implemented along the path taken by the PTP packets. NTP is a client–server protocol (the client requests service from an NTP server, and the server maintains no information on the clients) while PTP is master–slave (the master sends information to the slaves).

In all such protocols the delay from the master or server to the slave or client must be measured in order to offset the time-of-day announcement. In NTP this measurement is merged with the announcements, while PTP separates the two functions in order to enable announcements at a higher rate (and even multicast). The standard technique measures round-trip delay and assuming symmetry between the two directions divides by two. Furthermore, the non-negligible residence time in the responding element must be taken into account. The calculation involves four time stamps. The client/slave initiates the exchange; t1 is the time (according to the client/slave's clock) that the protocol packet was sent originally; t2 and t3 are the time (according to in the server/master's clock) that the server/master receives that packet and transmits its response packet respectively; and t4 is the time (according to the client/slave's clock) that the client/slave receives the response packet. Assuming symmetry, the estimate of the one-way delay is:

$$\text{Delay} = \tfrac{1}{2}\,(t4 - t1 - (t3 - t2))$$

The problem is that this calculated one-way delay varies from packet to packet due to queuing delays in intermediate switches between master and slave, and infrequently due to routing changes. PTP enables cancelation of these effects through either of two mechanisms:

- A transparent clock (TC) is an intermediate switch that can measure its total residence time and add it to the accumulated time through the network.
- A boundary clock (BC) is an intermediate switch that like a slave disciplines its own internal clock, and like a master initiates its own PTP messaging.

In normal operation PTP is used at update rate sufficient to precisely set its internal clock frequency, so that time determinations can be extrapolated for some time. However, frequency adjustment made at packet rates is necessarily less accurate than that obtainable via a PLL working on the physical layer. That being said, it is possible to combine highly accurate frequency obtained from SyncE with less frequent time updates obtained from PTP (1588wSyncE).

It should be mentioned that in order to conform to the more stringent requirements of 5G, new techniques are still being developed. PTP has undergone a revision to 1588v3, which includes both new security mechanisms and higher accuracy through the so-called white

rabbit extensions (ITU-T 1588v3). In order to cope with brownfield networks that do not use SyncE and/or do not have PTP on-path support, the distributed GM (DGM) approach may be used (US 9,276,689). DGM breaks the paradigm of a single master clock located relatively remote from its slaves (and thus suffering from numerous uncompensated time errors from all the intermediate switches), and instead uses a large number of PTP masters close to the cell sites where the slaves are located. These devices are presently available in small form-factors to simplify deployment.

#### 6.6.4.5 Energy Efficiency

Studies estimate that between 0.5 and 1% of global electric energy consumption is directly attributable to mobile communications, of which about 20% is consumed by the transport segment. Power consumption can be taken to be linearly proportional to data rate (note that energy consumption of computation increases super-linearly with clock speed). Hence, 5G's striving to increase rates by a factor of 10–100 will lead to a dramatic impact on global power consumption, unless power efficiency is improved.

All physical layer technologies used for xHaul transport consume essentially constant power regardless of the true data rate, in fact regardless of whether data are being sent at all. For this reason, various green mechanisms have been proposed that save energy by putting ports into sleep mode when there are no data to be sent. Other proposed mechanisms include automatically adjusting transmitted power according to cable length, and automatically adapting transmission speed according to the amount of data that need to be sent.

IEEE has standardized the first mechanism as energy-efficient Ethernet (clause 78 of IEEE 802.3). When there are no data to be sent the physical layer sends low power idle (LPI) symbols for some specified time, and then enters sleep mode during which it only transmits periodic refresh signals to maintain link integrity, although all receive circuitry remains active. When new data arrive, the normal idle signal is sent for some time, and transmission is subsequently restored.

### 6.6.5 Higher Layers

As we have mentioned above, 3GPP specifications severely underspecify the transport segments within the NG-RAN and between the gNB and the 5GC, viewing connections from the cell site to the core as transparent transport pipes. In practice, this connection is implemented as a non-trivial collection of transport segments and NEs, often operated by multiple network operators, each with its own technologies, management systems, and business interests.

This state of affairs is the result of two sets of justifications. The first set consists of technological constraints that make it unappealing to simply connect base stations to the core network. Ports on core routers are limited in number and expensive – there are simply too many cell sites for it to be economically feasible to directly connect them all to core elements. Even were there sufficient ports, the number of fibers required for a star configuration from core edge routers would be prohibitive. Core elements have very high rate ports (e.g. 100 Gbps), which are unneeded and too expensive to implement for cell site equipment. Finally, provision of Xn interfaces between base stations without the exorbitant cost of direct fiber interconnect requires a non-trivial network architecture.

The second set of justifications consists of business constraints. Provision of transport services requires transport resources (fiber plant, microwave links, optical muxes, Ethernet switches, etc.) and transport expertise (operations, administration, maintenance, performance measurement, efficient APS, etc.), neither of which is at the heart of the mobile operator's business. It thus makes sense for the mobile operator to farm out the transport segment to a wholesale provider (or to an internal, but separate, transport division – see below) with the needed resources and expertise.

In many cases the mobile operator provides its own transport, but generally through a separate "transport" division or business unit. Quite often this transport network is used for multiple services – residential, business, and mobile. Although a mobile operator may decide to build out its own transport network, this case occurs most often when the mobile operator is the successor of an incumbent telephony provider, or the result of a merger of service providers with different specialties, or part of a diverse telecommunications company. Even in such cases the mobile operator may contract a wholesale provider to augment its footprint into areas where it needs to provide mobile coverage but has no transport network.

A wholesale provider is typically a network operator with extensive fiber and switching resources with which it delivers a variety of services, of which mobile backhauling is one. Mobile backhauling mandates certain specific requirements, such as synchronization, but is otherwise a relatively straightforward use of the wholesaler's network and expertise. A recent trend is for wholesale providers to additionally supply power and shelving for servers, in order to host (and perhaps even provide) virtualized services including virtual RAN functions (see Section 6.2).

Based on the above, one might be led to believe that the transport network extends from the base station (which in 4G and 5G necessarily has IP routing functionality) and the provider edge (PE) router at the edge of the mobile core. In fact, this is not quite true. Due to the organizational separation between mobile operator and wholesale provider or transport division, cell site gateways (CSGs) (BBF TR-221) are required for demarcation at the cell site, and aggregation site gateways (ASGs) at the edge of the mobile core. When transport is based on IP or MPLS technologies the CSG is sometimes called a cell site router (CSR), but this term will be avoided here since it is more often used for routers belonging to the mobile operator's network. 3GPP (which, as has been mentioned, is not concerned with transport network issues) does not define CSG or ASG functionality, leaving it to other standards organizations, notably the BBF and MEF. In any case, in this book we define the transport network to extend from the CSGs to the ASGs.

Demarcation is a function used in situations where a communication service is provided to an end user, or to a second operator providing an *over-the-top* service. The purpose of implementing demarcation in a separate gateway, rather than as a function in an existing NE, is to clearly delineate the boundary between the service provider (or transport division) and its client, in order to avoid finger-pointing arguments regarding provisions of SLAs in the case of a wholesaler or service level objectives (SLOs) in the case of a transport division. Thus, the wholesale provider will continuously monitor its QoS parameters using its OAM toolset, and automatically trigger corrective actions when an SLA objective is endangered. Note that the demarcation devices typically belong to the wholesale provider, not to the mobile operator, making it harder to absorb its functionalities into the base station.

Once in place, a cell site gateway may be used to provide additional functionalities as well, and more fully warrant its name. The most common function is aggregation of all traffic types originating in the cell, including 4G, 5G sub 6 GHz (FR1), 5G mmWaves (FR2), small cell transit traffic, Wi-Fi hot spot backhauling, etc. The CSG is responsible for homogenizing traffic across cellular generations and wireless technologies, minimizing transport expenses (including energy efficiency and multiplexing/duplexing in order to minimize OOF fiber expenses), constructing a single frequency and time-of-day reference clock, providing a cell-wide heartbeat to ensure connectivity to the cell and all its components, and initiating failover self-healing procedures as needed. It may additionally incorporate a micro-data center platform for edge computation.

#### 6.6.5.1 xHaul Network Topology

We concluded in the previous subsection that direct connection of base stations to the mobile core is infeasible, and that a non-trivial xHaul network, with various transport-specific functionalities, is required.

The classical backhaul network topology comprises a "core edge" switch or router connecting via an "aggregation network" to CSGs. The connectivity topology of this network can be star, tree, "hub and spoke," or a ring or collection of rings (ring–subring, ladders, etc.), as depicted in Figure 6.6.5. In geographically large jurisdictions, where cell count and distances overly burden such a topology, the connectivity between the core edge and the cell sites might be divided into "first aggregation" (or pre-aggregation) networks and second aggregation network (sometimes called "access" and "aggregation" networks, respectively). The first aggregation networks have hub and spoke or ring topology, while the second aggregation network comprises rings or some form of partial mesh.

This basic model supports various variants. For example, Xn interfaces between cell sites are supported by hair-pinning at the core edge (or second aggregation switch) rather than continuing on to the core. Fronthaul requires an extension to this model, where dark fiber is deployed from the cell sites to PoPs where the baseband unit is located, and the aggregation network commences at the PoPs. Local Internet breakout (if deployed) further muddies the waters.

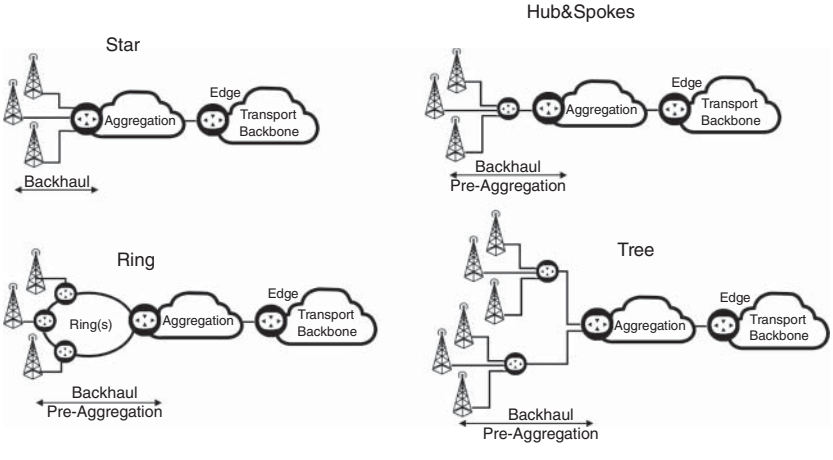



**Figure 6.6.5** Transport network topologies.

Migration from 4G to 5G will further impact the classical model in several ways. First, the rates will drastically increase, both due to the increased user data rates, and to fronthaul transport. While in principle rate does not affect topology, we shall see that in practice it does. More importantly, although at first 5G eMBB coverage may be collocated with 4G cells, to reap the benefits of 5G the number of cell sites will radically increase over time (if for no reason other than the limited range of mmWave propagation). The decomposition of the gNB into RU, DU, and CU also means that there will be more distinct types of physical or virtual elements to be connected. The automation of network slices will fundamentally alter the management plane, but will have limited effect on network topology. Additionally, the requirement for low-delay inter-cell site connectivity is greater, impacting the connectivity at some level. Finally, because of commoditization of pure transport services ("dumb pipes"), wholesale providers will need to move up the value chain. This means that they may attempt to additionally provide hosting of virtualized NFs (e.g. MEC) or even virtualized RAN components (e.g. vCU).

A generic depiction of the RAN network segments and their interconnection to the 5G core is given in Figure 6.6.6.

#### 6.6.5.2 Transport Protocols

There are a small number of packet forwarding protocols (Carrier Ethernet, MPLS, IP), and a much larger number of their variants (MPLS-TE, MPLS-TP, segment routing, EVPN, etc.) that can be exploited for xHaul packet transport. The choice between these is often influenced more by the transport provider and equipment vendor histories than by the true pros and cons of the technologies.

Generally, there are a number of different kinds of transport providers offering xHaul services:

- The first kind is a metro connectivity provider, who was originally PDH/SDH-based, then perhaps moved to asynchronous transfer mode (ATM), but has now overwhelmingly adopted Ethernet. Such service providers have enthusiastically embraced Carrier Ethernet, as espoused by the MEF forum, in order to provide SLA-based services.
- A second type of service provider covers large geographic regions, such as whole countries, or even worldwide service. Such service providers are traditionally MPLS-based, more specifically supporting the L3VPN (RFC 4364) variety of MPLS, perhaps supplemented with pseudowire services (RFC 3985).
- A third kind of provider is typified by electric utilities, who, having excess fiber plant, are willing to lease it at reasonable prices. Since these service providers are actually not networking experts, they tend to offer wavelength service, or piggybacking over whatever mechanisms they already have in place, e.g. flat IP or vanilla MPLS.
- Finally, there are those operators with an in-house transport division, whose technology toolkits vary according to their history.

Accordingly, there are two traditionally opposing worlds of backhauling based on Carrier Ethernet and MPLS, with a newer possibility of pure IP networks.

MPLS networks, with their local labels rather than network-unique addresses, tend to scale better, and thus second aggregation networks are most often MPLS-based. Ethernet technologies can be scaled up using Provider Backbone Bridge Network (PBBN) techniques

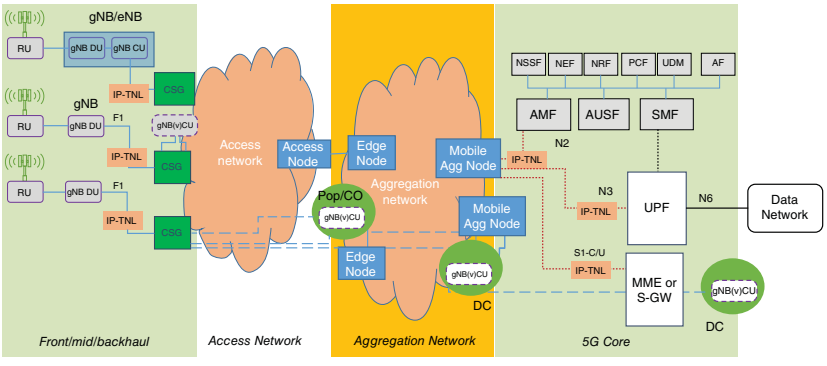



**Figure 6.6.6** RAN network segments and their interconnection to the 5GC.

(familiarly known as MAC-in-MAC in contrast to the standard service provider labeling called Q-in-Q), but these techniques have not been widely adopted. Hence, MPLS dominates the second aggregation network, when one is present.

On the other hand, Ethernet must be considered as a more secure technology for access networks (after all, MPLS packets have no source address that can be authenticated), and additionally (unlike IP and MPLS) define their own physical layer. Hence Ethernet has seen success in the first aggregation segment.

Since core networks also employ MPLS, the idea of seamless MPLS, that is, stitching together of two MPLS domains has been proposed (draft-ietf-mpls-seamless-mpls). However, concerns have been voiced about the security of this idea, and in any case the idea is unattractive when two different service providers are involved.

No matter which protocol or combination of protocols is utilized, the transport network must be "carrier grade," that is, it must be able to guarantee maintenance of specified quality levels. The QoS parameters may include data rate, one-way or round-trip delay, packet delay variation, etc. The contract between the customer (mobile operator) and the (wholesale) service provider specifying the precise QoS parameters, their required bounds, and the consequences of not conforming to these bounds is called an SLA. When the transport is not provided by a distinct business entity, the parameters and their values are often called SLOs. The toolkit for monitoring QoS parameters consists of OAM protocols (see Section 6.5), which in turn are classified as FM (CC, loopbacks, etc.) and PM (packet loss measurement delay measurement, etc.).

Ethernet was originally a LAN technology, and thus did not require specification of QoS parameters. When Ethernet became a service provided to business customers by metro-Ethernet service providers, it was upgraded to Carrier Ethernet by specifying QoS parameters, OAM protocols to measure them, management-based route specification, APS mechanisms (including for Ethernet rings) to rapidly recover from failures, and various other carrier grade features.

MPLS was originally a mechanism to accelerate forwarding of IP packets, and thus did not require specification of QoS parameters. It has since been extended in multiple fashions, and we can now distinguish several distinct flavors:

1. *Vanilla MPLS* is the variety deployed in the core of the Internet to accelerate forwarding. It usually employs LDP signaling to set up label switched paths (LSPs), but is not truly connection-oriented as these LSPs change with routing updates. It may employ FRR to locally bypass faults.
2. *RFC-2547* (obsoleted by RFC 4364) is the flavor used to implement Layer 3 VPN (L3VPN) services for business customers. It sets up LSPs using Border Gateway Protocol (BGP).
3. *MPLS-TE* is a true connection-oriented version that reserves resources in order to guarantee SLOs. It does not specify a full OAM suite or APS.
4. *MPLS-TP* is a transport network-specific flavor. Its defining characteristic is the definition of QoS protocols and true APS (instead of FRR); it may operate without IP forwarding or routing protocols.
5. *MPLS-SR* is the newest addition. MPLS segment routing is implemented by a stack of MPLS labels, which are popped in order to reveal the next hop, similar to (now all but deprecated) source routing but without the security issues of enabling an end user dictate forwarding behavior.

It must be noted that Ethernet defines both the physical (L1) and data-link (L2) layers, while MPLS (as part of the IP suite) does not define a physical layer. Therefore, it frequently occurs that MPLS or IP avail themselves of Ethernet for the lower layers. However, we differentiate between this use of Ethernet as a "dumb pipe," and the use of Carrier Ethernet as a carrier grade networking technology. Yet, it may happen that the differentiation is sometimes blurred, with some functionality being carried out in the Ethernet underlay network (e.g. timing, physical layer fault detection, etc.), and other functionalities carried out by the higher layer.

#### 6.6.5.3 Protocol Stacks for User Traffic

In this subsection we detail the construction of the forwarding plane packets as seen in various points in the transport network. For concreteness we will assume in the following that the PDU session type is IP, and that all the data-link and physical interfaces are Ethernet. We describe packet structure according to the layering convention rather than the packet order convention, so that higher layers appear first, and the rightmost header is the first to be transmitted.

NG-RAN user traffic (e.g. from either DU or CU) is delivered as GTP-U packets, which are themselves encapsulated in UDP/IP (note that 2G and 3G used TDM or ATM, and that these may still be required to be supported). Similarly, signaling traffic is encapsulated as SCTP layer 4 over IP. When these IP packets are the focus of the transport network, we speak of an IP TNL. If these IP packets are further encapsulated in Ethernet layer 2 frames and these are the focus of the transport network, we speak of an Ethernet TNL. For further details on NG-RAN protocols, refer to Section 3.3.

As a concrete example, in conventional backhauling, packets from a CU destined for the appropriate UPF, are delivered to the transport network as IP/GTP/UDP/IP/Ethernet. Here, the outer (first) "IP" represents the user's IP datagram including IP header with the UE's address as source address (SA) and the server or peer's address as destination address (DA), while the inner (second) one is the backhaul IP header with CU's source address and UPF's destination address. Note that a single CU may connect to multiple UPFs, and the appropriate UPF depends on the UE's session.

For the gNB-CU/DU split (described in Section 4.2), packets from a DU destined for its CU are delivered to the transport network as PDCP/GTP/UDP/IP/Ethernet, where PDCP represents the (encrypted) user traffic with ROHC compressed headers. The IP is the xHaul IP (SA = DU, DA = CU). Note that a single DU may only connect to a single CU.

The xHaul network must perform several functions. First and foremost control of forwarding behavior. While the outer IP designates the packet's ultimate destination, unless the network is trivial there will be multiple possible paths and path parameters (priorities, shaping/policing mechanisms, etc.). Most often the xHaul network will employ a transport tunneling mechanism (unrelated to the mobile network's GTP tunnels), such as MPLS or GRE, although more recently an approach utilizing segment routing has been proposed. As for any carrier grade network there must be resilience mechanisms, such as APS (possibly with rings) or FRR. Optionally there will be multiple gateways to exit the network. To trigger these mechanisms FM OAM is required, and PM is often needed as well.

It is generally agreed that the first aggregation network does not need to employ MPLS switching (although the packets may carry MPLS labels), so that it will most probably be

based on either plain Ethernet with ancillary mechanisms or on CE. For the latter case, S-tagged Q-in-Q VLANs will be used to control the forwarding, resulting in a stack of the type X/GTP/UDP/IP/S-tag/C-tag/Ethernet. However, if an MPLS label stack is inserted to identify tunnels, this instead will be X/GTP/UDP/IP/MPLS/Ethernet. In some cases the transport provider may wish to preserve the user's Ethernet, in which case an Ethernet pseudowire may be built, resulting in X/GTP/UDP/IP/Ethernet/PW[15]/MPLS/Ethernet (where the first Ethernet header is the user's and the second is the wholesaler's), or more generally X/GTP/UDP/IP/Ethernet/PW/MPLS/S-tag/C-tag/Ethernet.

For the case where CE is not used, required carrier grade features are often provided by IP mechanisms, for example, IS-IS for forwarding behavior, bidirectional forwarding detection (BFD) for OAM, IP-FRR for resilience, etc. This requires some tunneling mechanism, most commonly GRE (RFC 2890), although an MPLS label stack may be employed (RFC 4023). When using GRE the stack will be IP/GTP/UDP/IP/Ethernet/GRE/IP/Ethernet, or IP/GTP/UDP/IP/Ethernet/MPLS/IP/Ethernet where the first IP is the UE's, the second the BS, and the third the CSG. More complex cases may be found, for example using Virtual Extensible LAN (VXLAN) (RFC 7348) or other modern tunneling mechanisms.

In fiber-rich environments, and when higher data rates are required (especially for fronthaul), the lower layers may be circuit-switched instead of, or in addition to, packet switched. In such cases OTN (ITU-T G.709) and dense wavelength division multiplexing (DWDM) (ITU-T G.694.1) technologies will be employed. The ITU-T is presently studying the application of OTN to 5G transport (ITU-T G Suppl. 67). Similarly, point-to-point microwave may be used instead of fiber as a constant bit rate transport medium; although such technologies now frequently sport Ethernet interfaces.

Two families of PONs may also be employed at the physical layer. IEEE Ethernet passive optical network (EPON) (clauses 64 and 65 of IEEE 802.3) share much of the standard Ethernet physical layer structure, with modifications (e.g. in the preamble and additional K-codes for forward error correction) and augmentations (such as **M**ulti**P**oint **C**ontrol **P**rotocol frames). ITU-T PON flavors (e.g. ITU-T G.989.3) encapsulate Ethernet (and MPLS) in gigabit passive optical network (GPON) encapsulation method (GEM and XGEM) carried in a synchronous bit stream.

The physical layer very often avails itself of point-to-point microwave links (either native Ethernet- or TDM-based). An emerging elegant solution (attractive for mobile operators providing their own transport) is integrated access/backhaul (IAB) (described in Section 5.2) wherein the 5G air interface and the backhaul share the same wireless technology.

#### 6.6.5.4 Technology Comparison

Any comparison of the pros and cons of the different technologies needs to address the issues of scalability, multiservice support, controlling forwarding behavior, support for slicing, resilience, FM, performance monitoring, security, timing, and NFV/MEC, which are discussed in the following subsections.

##### *6.6.5.4.1 Scalability*

The problem of scalability in the transport network is much less acute than in the mobile network itself, since there is no awareness of individual end users or devices. In most cases the scale should not exceed hundreds of endpoints, including the CSGs and ASGs.

15 PseudoWire.

#### 6.6.5.4.2 Multiservice Support

No transport network is useful if it can't transport the required client traffic types. Ethernet carries a wide variety of traffic types via Ethertype marking, or logical link control (LLC), but does not natively transport TDM, which requires pseudowire extensions (MEF-8). IP carries different traffic types either directly (via the protocol number or "next header" field) or via layer-four port numbers. MPLS natively carries only IP or MPLS itself (the latter using the label stack) but may transport a wide variety of payloads via pseudowire mechanisms. It should be noted that MPLS packets are not self-describing, and thus there is no way of discovering the traffic type by packet inspection.

Although 3GPP defines three PDU session types for 5G, namely, IP, Ethernet, and unstructured, and the transport network may handle traffic from various gNB functional split options, in practice all 5G packet transport networks will be required to transport IP over Ethernet. The split option 8 fronthaul traffic (not being explicitly specified for 5G) will generally be transported as CPRI (CPRI) over dark fiber or OTN, and the low-level split defined by O-RAN (see Section 4.5) is expected to be much more popular due to its reduced data rate. This latter split may be encapsulated in IP using eCPRI (eCPRI). For CU/DU split (see Section 4.2) traffic will be GTP in UDP in IP; the split option 1 (also not being explicitly specified for 5G) may use be IP in GTP over UDP/IP; handoff to third-party packet networks will avail itself of pure IP. All of these may be carried over the aforementioned alternative lower layers, such as double-tagged Ethernet, MPLS, Ethernet pseudowires over MPLS, and often seemingly ridiculous combinations of these are regularly encountered.

For the foreseeable future CSGs will be required to support 4G (both fronthaul and backhaul), and perhaps 3G (IP or ATM) and even 2G (TDM). In addition, non-3GPP IP (e.g. Wi-Fi) and other sources of IP traffic (e.g. residential) may all be in the mix.

Some wholesalers, especially those with Carrier Ethernet networks, may prefer not to terminate the Ethernet underlay over which the IP traffic is delivered and employ Ethernet pseudowires (RFC 4448) instead. In such cases, user Ethernet PDUs may be handled natively.

#### 6.6.5.4.3 Controlling Forwarding Behavior

It is often the case that network traversal needs to be more nuanced than simply ensuring packet delivery. This is most often the case when there are alternate paths with quite different E2E QoS parameters, although other factors may also be influential (e.g. paths with quite different costs). Additionally, network slicing requires separation and appropriate forwarding of flows belonging to different slices, and issues specific to it will be discussed in the next subsection.

Two types of path computation are used in the control plane:

1. Distributed routing where forwarding devices exchange information between themselves and each independently builds a forwarding information base (FIB).
2. Centralized path computation (network management, SDN) where an omniscient "God box" uses graph optimization algorithms to compute paths, and disseminates these to the forwarding elements.

Independently of these, two types of QoS handling need to be considered in the forwarding plane:

1. Hard QoS (AKA IntServ, traffic engineering [TE]) where a combination of connection admission control (CAC) and resource reservation provides hard QoS guarantees.
2. Soft QoS (AKA DiffServ, traffic conditioning) where packets are afforded differential treatment according to their priority and discard eligibility, and scheduling/queuing/policing/shaping algorithms provide statistical QoS performance.

Ethernet, IP, and MPLS were all originally best effort with no QoS handling, but each developed such handling over time. Hard QoS was proposed for VoIP in the form of Resource Reservation Protocol (RSVP), but was never widely used. DiffServ IP is common based on the 6-bit DiffServ code point (DSCP) field in the IP header. Carrier Ethernet implements soft QoS based on the 3-bit priority code point (PCP, colloquially called priority bits or P-bits) and discard eligibility indicator (DEI). MPLS-TE adopted the traffic engineering approach by extending RSVP to RSVP-TE (RFC 3209).

Ethernet, IP, and MPLS have traditionally utilized various distributed control protocols to learn how to forward, while in Carrier Ethernet traditionally a NMS configured switch forwarding tables. In MPLS-TE a path computation element (PCE) (RFC 4655) was later added to optimize centralized path computation, and still later SDN advocated centralized control of IP and consequent simplification of the forwarding elements to become so-called whitebox switches. Use of SDN for 5G transport networks is described in ITU-T G.7702.

One advantage of the PCE approach over the related SDN one is that the PCE did not instruct the NEs, leaving the steering function to the source node, while the SDN controller needs to reach out and maintain state with every whitebox switch. Because of this, SDN controllers are single points of failure, suffer from scalability issues, and even minor programming bugs in SDN controller code can impact unrelated flows. Recently an alternative called segment routing has gained popularity for MPLS and IPv6.

Segment routing, similar to source routing, dictates forwarding by a list of path stations in the packet. Unlike source routing, this list of intermediate addresses is inserted by the ingress router, not by a source host, avoiding the negative security implications of source routing. In MPLS segment routing, the list of intermediate nodes is implemented as a standard MPLS stack, with each label switch router (LSR) popping the top of stack label, rather than swapping it.

#### *6.6.5.4.4 Support for Slicing*

A *network slice* is defined (3GPP TS 23.501, TS 28.530) as a logical network that provides specific network capabilities and network characteristics. *Hard isolation slicing* refers to the dedicating of resources to a slice instance (such as the assignment of a wavelength on a fiber), while *soft isolation slicing* refers to isolation of slice instances using shared resources.[16] Slice instances can't exchange packets or directly observe each other, but still may dynamically interact (e.g. due to resource contention), if soft slicing isolation is used. Soft slicing is achieved through logically multiplexing the data plane over a physical channel, by means of tunneling or pseudowires.

In order for a network slice to conform to its QoS criteria, it needs to be defined E2E, that is, on the air interface, the transport network, and in the core. To support slicing in the transport network the packets need to be classified as belonging to a particular slice.

16 Resource isolation aspects are not defined by 3GPP and are left for vendor implementation.

In the case of fronthaul this may not be possible, both because different slices are mixed together in the air interface, and because there may be no easily recognizable identifier for classification (unless slices are differentiated by RF band). In other cases the GTP headers or UDP ports must provide the necessary classification labels. Furthermore, for the split option 1, different slices may be directed to different UPFs, but at the split option 2 (the DU/CU split) there is only one CU for a given DU, and thus the different slices need all to be delivered upstream to the same CU, but may traverse different paths or incur different forwarding behaviors at intermediate NEs.

To support slicing using an Ethernet service, it is necessary to provide a mapping from a 3GPP-defined network slice identifier, that is, network slice selection assistance information (NSSAI), to some identifying fields in the Ethernet packet, such as 12-bit VLAN identifier (VID) or the 3-bit PCP. Ethernet VPNs would then constitute a form of soft network slicing. VPN technologies utilize tunneling, isolation of forwarding tables between different tenants, and overlay topology to provide connectivity between different sites of each virtual network. The VPN overlay and the underlay network resources are loosely coupled, and statistical multiplexing still functions to improve network utilization.

Carrier Ethernet supports a palette of isolation types, including (MEF 6.1):

- E-LINE – point-to-point Ethernet service:
  - Ethernet private line (EPL) – dedicated bandwidth E-LINE service, further subdivided by ITU-T G.8011.
  - Ethernet virtual private line service (EVPL) – shared-bandwidth E-LINE service (i.e. statistical multiplexing of user traffic) (e.g. VPWS).
- E-LAN – multipoint-to-multipoint Ethernet service:
  - EPLAN – dedicated-bandwidth E-LAN service.
  - EVPLAN – shared-bandwidth E-LAN service (e.g. VPLS).
- E-TREE (or Ethernet Virtual Private Tree) – point to multipoint Ethernet service.

To support slicing using MPLS the identifying fields can be either the 20-bit label (previously called LLSPs) or the 3-bit traffic class field (ELSPs). Using MPLS-TE one can guarantee performance (hard QoS) through resource reservation using RSVP-TE (RFC 3209), or by mapping each slice to a physical channel (e.g. wavelength or fiber).

Furthermore, since slicing requires supporting multiple logically self-contained networks over the same transport network (3GPP TS 28.530), the management systems of the mobile (described in Section 6.5) and transport networks need to function in harmony to economically attain the performance objectives of each slice instance. This will require cross network interconnection, alignment functions and security mechanisms, which have yet to be standardized.

#### *6.6.5.4.5 Resilience*

High availability mechanisms such as rapid restoration, rings, and FRER are discussed above in the context of the physical layer. In this section we focus on the impact on higher layer protocols. APS requires careful protocol work, planning, and proper configuration. Historically solutions for both linear protection (i.e. protection over general topologies) and ring protection have been employed.

Ethernet, due to its lacking a time-to-live field, disallows rings. Two solution strategies have been proposed:

- *Open loop* ring protection methods (e.g. G.8032), wherein at any instant in time one link in the ring is blocked, and upon a single link failure the protocol assures that the failed link is the blocked one.
- *Closed loop* ring protection methods, whereby some other mechanism, e.g. adding a TTL field, avoids loops.

Open loop mechanisms are generally incompatible with QoS assurance, and closed loop mechanisms have not gained wide acceptance.

For MPLS-TP the IETF has standardized linear protection (RFC 6378) and ring protection (RFC 6974), and the ITU-T has standardized alternative mechanisms (ITU-T G.8131, ITU-T G.8132). More prevalent is MPLS Fast ReRoute (RFC 4090) that provides a local detour around failed fibers or nodes at the cost of loss of determinism – the endpoints are not informed of the local route change.

IP recovers from failures by computing new routes, which is often a lengthy process. Loop Free Alternate Fast Reroute (RFC 5286) minimizes downtime by precomputing backup paths (called repair paths) that are guaranteed to be loop free. For IP (MPLS) an LFA to a destination with respect to an element (link/node) for a destination is a router that:

1. Is not the default next hop;
2. Is connected to the destination;
3. Does not forward through the element (and hence does not need to know about the failure).

In the context of MPLS segment routing, topology independent LFA (TI-LFA) allows the source LSR (which knows all the labels from the SR protocols) to immediately substitute and alternative MPLS-SR label stack. It is topology independent in the sense that a loop free backup path is found irrespective of the topologies before and after the failure.

The replication and erasure mechanisms (FRER, PREOF) discussed above have been specified by TSN for Ethernet (802.1CB) and are being specified for IP and MPLS by DetNet (RFC 8655).

#### *6.6.5.4.6 Fault Management*

Unless the failed element is physically connected to the destination, triggering any of the reliability mechanisms of the previous subsection relies on E2E continuity monitoring mechanisms. (Note that CC refers to verifying that information sent indeed arrives at the destination, while connectivity verification (CV) refers to verifying that information sent to a particular destination does not arrive somewhere else.) These continuously running OAM mechanisms, along with troubleshooting diagnostics (such as loopbacks) that are run when needed, are collectively called FM. Upon detecting a fault, FM may trigger control-plane functions such as APS or reroute, and management-plane functions such as collection of fault statistics, setting off alarms, notification of technical staff, etc. In the following subsection we discuss those OAM functions that monitor less critical operational parameters and which generally only trigger management-plane statistics gathering.

Ethernet, once without any OAM now has multiple standardized FM mechanisms. In particular connectivity FM (IEEE 802.1ag) and (ITU-T Y.1731) define CC heartbeats, as well as loop-back and link trace mechanisms. The ITU-T version further defines forward and backward defect indications, locking for diagnostics, and messaging channels for APS. Another ITU-T standard (ITU-T Y.1564) specifies the use of FM when commissioning a new service.

IP and MPLS have a FM protocol known as BFD (RFC 5880). Originally a simple keep-alive and loop-back (known in BFD as *echo*) mechanism between two adjacent routers, BFD has expanded to become a full-featured FM protocol, especially in the context of MPLS-TP, which additionally uses LSP-ping for on-demand diagnostics (RFC 6426).

#### *6.6.5.4.7 Performance Monitoring*

As mentioned in Section 6.5, 5G OAM has been extended to cover performance monitoring for measurement of QoS parameters of the transport network. Thus, while CC is critical for any application (no communications-based application can properly function without communications), such parameters as one-way delay, round-tip-delay, PDV, and PLR may be critical for proper functioning of some, but not all applications.

The ITU-T version of Ethernet OAM (Y.1731) supports both FM and performance monitoring, while the IEEE version (802.1ag) supports only FM. IP has one-way and two-way measurement mechanisms, known respectively as One-Way Active Measurement Protocol (OWAMP) (RFC 4656) and Two-Way Active Measurement Protocol (TWAMP) RFC 5357. MPLS-TP defines an extensive set of performance monitoring functions (RFC 6374 and RFC 6375).

#### *6.6.5.4.8 Security*

5G presents several new security challenges related to the transport network. The most obvious one is the radical change in trust model due to openness of the core toward third-party applications. Related to this are the new use cases and novel network architecture based on distributed telco cloud. Higher number of cells, higher data rates, and lower latencies all further impact security solutions.

Focusing on the transport network we find that performance issues lurk behind many of these challenges. Denial of service based on overloading physical bit per second rates, forwarding packet per second rate, cryptographic algorithm resources, or RAN-situated virtualization or computation infrastructure all need to be addressed in order to avoid bottlenecking the RAN or 5GC resources such as UPF. This involves both upgrading raw performance to handle worst case scenarios, and detecting and blocking threats at the edge. However, traditional Distributed Denial-of-Service (DDoS) mitigation approaches may be powerless against truly large-scale infected massive MTC attacks, and the conventional approach of redirecting suspicious packets to scrubbing centers may be irrelevant because of latency constraints.

RAN transport protocols play a part here as well. In the forwarding plane it should be noted that while IP and Ethernet packets have SAs, and thus allow for packet-by-packet source authentication, MPLS label stacks contain neither destination nor SAs, and hence necessarily rely on lower or higher layers for this function. In the control plane IP and vanilla MPLS conventionally rely on distributed routing protocols, which facilitate certain

attack vectors, while Carrier Ethernet and MPLS-TE traffic steering are configured from a central site in SDN fashion, which exposes them to different threats.

Attacks involving undermining virtualized RAN infrastructure (vDU/vCU, MEC) have yet to be adequately researched. Adopting cloud principles may lead to mobile and/or transport operator VNFs running on the same physical platform as third-party AFs, and even completely unrelated computational tasks. While economically attractive multi-tenant hosting introduces novel attack vectors, including denial of service, information leakage (potentially including discovery of passwords, shared secrets, and credentials), data manipulation, resource access, and even gaining complete platform control. Until lately virtualization technologies were thought to prevent these attacks, but the recent discovery of a plethora of attacks exploiting speculative execution, instruction pipelining, and paged memory have demonstrated that almost all modern CPUs are vulnerable.

#### *6.6.5.4.9 Timing*

We previously discussed the requirements and physical layer aspects of delivering highly accurate timing to base stations over the transport network. For the most stringent requirements the network's physical layer is utilized for stabilizing frequency (since bit rates are orders of magnitude higher than packet arrival rates) while the upper layers take care of time accuracy with some help from the physical layer.

MPLS and IP can't provide any of the required physical layer functionalities, simply because they don't define physical layers. Ethernet defines both layer 1 and layer 2, and is thus almost universally used for on-path support, even when forwarding is performed by a different protocol.

The IETF TICTOC working group seriously considered an MPLS PTP encapsulation, but this never progressed to standard status. The ITU-T telecom profiles (ITU-T G.8265.1, ITU-T G.8275.1, ITU-T G.8275.2) use UDP/IP PTP encapsulations in order to simplifying addressing, but still assume well-engineered networks and the use of on-path support as required, which in most cases is accomplished by means of an Ethernet underlay network (at very least to provide accurate time stamping). While it is extremely advantageous for PTP sync messages to be multicast in large networks, the current ITU-T telecom profiles mandate unicast. However, they do specify automatically configuring slaves with IP addresses of potential master clocks via 1588's optional unicast discovery mechanism.

#### *6.6.5.4.10 NFV/MEC*

As mentioned before (e.g. in Section 6.2), functionality virtualization is used in mobile networks for the following types of NFs:

1. 3GPP-defined network nodes, e.g. AMF and CU;
2. Networking functionalities required by the transport network itself, e.g. wide area network (WAN) optimization, FM and performance monitoring probes and reflectors;
3. User-centric functionalities such as firewalls, support for location-based services, and IoT aggregation;
4. MEC (see Section 6.4).

The first (and in many cases the fourth) of these involves functions belonging to the mobile operator, although the transport provider may provide the computational platform and

hosting services. The second function type focuses unambiguously on the needs of the transport provider (reducing costs, increasing automation), and is intended to be transparent to both mobile operator and end user. Items in the third category directly benefit the end user, and are frequently marketed by mobile operators, although they may be provided by third parties (a model facilitated by the 5G service-based architecture [SBA]).

The first type is obviously tightly coupled with the functional split option being used, while the second type is only indirectly influenced by the split (being directly susceptible only to traffic characteristics such as data rate or delay constraints). In some cases type 2 VNFs may be slice-dependent, in which case they must be able to classify traffic to a slice as has been previously discussed.

User functionalities are generally limited to backhauling (split option 1) where user IP packets are discernible. At NG interfaces these user packets are still transported in GTP-U tunnels, and there is a need to decapsulate (or at least snoop) the GTP. This GTP handling may be performed by a VNF hosted in a CSG, which then directs the user traffic to an appropriate server, or the entire functionality may be hosted in the CSG. Alternatively, a MEC platform (ETSI GS MEC 003) may be hosted in a CSG, or a vUPF implemented and afterwards (as part of the 5G core) 5G AFs may be similarly hosted [MEC 5G WP].

### 6.6.6 Conclusions

5G presents multiple highly inter-related challenges to transport. Numerous technologies, both physical layer and higher layers, are being proposed to meet these challenges. Increased data rates mandate newer physical interfaces, and/or new techniques for bonding physical interfaces, but also impact network topology, forwarding paradigms, and placement of edge computing (MEC) platforms. Reduced latencies necessitate time-sensitive forwarding mechanisms and efficient SDN-based routing, and in some cases dictate MEC processing. Both high data rates and TSN impose the requirement for yet more highly accurate time and frequency synchronization. High reliability impacts system design and obligates deploying new resilience mechanisms. Data rate, latency, and reliability are all QoS criteria that may need to be monitored and traded off (e.g. via network slicing). Many of the new requirements and proposed mechanisms entail increased energy consumption, which needs to be countered by yet other means. And all the above mechanisms need to be deployed without introducing new security vulnerabilities.

Non-standalone 4G/5G cases present yet further challenges to the transport segment; for example, integrated low PHY functionality for converged NEs. Network migration scenarios may be even more challenging as they typically attempt to leverage existing brownfield network infrastructure.

## References

1 Stein, Y.J., Geva, A., and Zigelboim, G. (November 2007). Delivering better time-of-day using synchronous Ethernet and 1588, ITSF-2007.

2 (2016). Feasibility study on new services and markets technology enablers for critical communications; Stage 1. Available at: www.3gpp.org (accessed May 29, 2020).

3 (2019). System architecture for the 5G System; Stage 2. Available at: www.3gpp.org (accessed May 29, 2020).
4 (2019). Technical specification group services and system aspects; Management and orchestration; Concepts, use cases and requirements. Available at: www.3gpp.org (accessed May 29, 2020).
5 (2019). Evolved Universal Terrestrial Radio Access (E-UTRA); Base station (BS) radio transmission and reception. Available at: www.3gpp.org (accessed May 29, 2020).
6 (2019). NR; Base station (BS) radio transmission and reception. Available at: www.3gpp .org (accessed May 29, 2020).
7 (2019). NG-RAN; Architecture description. Available at: www.3gpp.org (accessed May 29, 2020).
8 (2011). Technical report TR-221 technical specifications for MPLS in Mobile Backhaul Networks.
9 Nichols, K., and Jacobson, V. (May 2012). Controlling queue delay, ACM Queue.
10 Ericsson AB, Huawei Technologies Co. Ltd., NEC Corporation, Alcatel Lucent, and Nokia Siemens Networks GmbH (August 2013). Common Public Radio Interface (CPRI); Interface specification v6.0. Available at: www.cpri.info (accessed May 29, 2020).
11 Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B., Francois, P., Voyer, D., Clad, F., and Camarillo, P. (2019). Topology independent fast reroute using segment routing.
12 Leymann, N. (ed.) (2014). Seamless MPLS architecture. IETF.
13 Ericsson AB, Huawei Technologies Co. Ltd., NEC Corporation, and Nokia (2019). eCPRI specification V1.0 (2017-08-22). Interface specification.
14 (September 2014). Mobile-edge computing – Introductory Technical White Paper. Available at: www.etsi.org (accessed May 29, 2020).
15 (June 2018). ETSI White Paper No. 28, MEC in 5G networks, first edition – June 2018. Available at: www.etsi.org (accessed May 29, 2020).
16 (January 2019). Multi-access edge computing (MEC); Framework and reference architecture. Available at: www.etsi.org (accessed May 29, 2020.
17 (2008). IEEE 1588–2008 – IEEE standard for a precision clock synchronization protocol for networked measurement and control systems. Available at http://standards.ieee.org (accessed May 29, 2020).
18 (2019). IEEE 1588–2019 – IEEE standard for a precision clock synchronization protocol for networked measurement and control systems. Available at: http://standards.ieee.org (accessed May 29, 2020).
19 (2012). IEEE standard for local and metropolitan area networks virtual bridged local area networks amendment 5: connectivity fault management, now IEEE 802.1Q-2012 clauses 18–22.
20 (December 2014). IEEE standard for local and metropolitan area networks – link aggregation. Available at: http://standards.ieee.org (accessed May 29, 2020).
21 (September 2017). IEEE standard for local and metropolitan area networks – frame replication and elimination for reliability. Available at: http://standards.ieee.org (accessed May 29, 2020).

**22** (June 2017). IEEE standard for local and metropolitan area networks – bridges and bridged networks – amendment 29: cyclic queuing and forwarding. Available at: http://standards.ieee.org (accessed May 29, 2020).
**23** (August 2018). IEEE standard for Ethernet. Available at: http://standards.ieee.org (accessed May 29, 2020).
**24** (December 2017). IEEE standard for Ethernet amendment 10: media access control parameters, physical layers, and management parameters for 200 Gb/s and 400 Gb/s operation. Available at: http://standards.ieee.org (accessed May 29, 2020).
**25** (February 2019). IEEE standard for Ethernet – amendment 3: media access control parameters for 50 Gb/s and physical layers and management parameters for 50 Gb/s, 100 Gb/s, and 200 Gb/s operation. Available at: http://standards.ieee.org (accessed May 29, 2020).
**26** (February 2012). Spectral grids for WDM applications: DWDM frequency grid. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**27** (June 2016). Interfaces for the optical transport network. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**28** (March 2019).G.709 amendment 3. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**29** (May 2014). Generic protection switching – linear trail and subnetwork protection. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**30** (August 2019). Generic protection switching – ring protection. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**31** (September 1997). Timing characteristics of primary reference clocks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**32** (August 2017). Timing characteristics of enhanced primary reference clocks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**33** (2012). 10-gigabit-capable passive optical network (XG-PON) systems. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**34** (October 2015).40-gigabit-capable passive optical networks (NG-PON2): transmission convergence layer specification. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**35** (March 2018). Architecture for SDN control of transport networks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**36** (November 2018). Ethernet service characteristics. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**37** (July 2014). Linear protection switching for MPLS transport profile. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**38** (August 2017). MPLS-TP shared ring protection. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**39** (Janaury 2015). Timing characteristics of a synchronous equipment slave clock. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**40** (July 2014). Architecture and requirements for packet-based frequency delivery. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**41** (August 2017). Time and phase synchronization aspects of telecommunication networks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).

**42** (November 2018). Timing characteristics of primary reference time clocks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**43** (November 2016). Timing characteristics of enhanced primary reference time clocks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**44** (June 2016). Precision time protocol telecom profile for phase/time synchronization with full timing support from the network. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**45** (June 2016). Precision time protocol telecom profile for time/phase synchronization with partial timing support from the network. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**46** (June 2019). Application of optical transport network recommendations to 5G transport. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**47** (February 2016). Ethernet service activation test methodology. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**48** (August 2015). Operations, administration and maintenance (OAM) functions and mechanisms for Ethernet-based networks. Available at: www.itu.int/en/ITU-T (accessed May 29, 2020).
**49** (April 2008). MEF technical specification 6.2 Ethernet services definitions phase 2. Available at: www.mef.net (accessed May 29, 2020).
**50** (October 2004). Implementation agreement for the emulation of PDH circuits over metro Ethernet networks. Available at: www.mef.net (accessed May 29, 2020).
**51** (October 2013). MEF technical specification 10.3 Ethernet services attributes phase 3. Available at: www.mef.net (accessed May 29, 2020).
**52** Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and Weiss, W.. (December 1998). An architecture for differentiated services. Available at: www.ietf.org (accessed May 29, 2020).
**53** Dommety, G. (December 1998). Key and sequence number extensions to GRE. Available at: www.ietf.org (accessed May 29, 2020).
**54** Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and Swallow, G. (December 2001). RSVP-TE: extensions to RSVP for LSP tunnels. Available at: www.ietf.org (accessed May 29, 2020).
**55** Bryant, S. and Pate, P. (Eds.) (March 2005). Pseudo wire emulation edge-to-edge (PWE3) Architecture. Available at: www.ietf.org (accessed May 29, 2020).
**56** Worster, T., Rekhter, Y., and Rosen, E. (Eds.) (March 2005). Encapsulating MPLS in IP or generic routing encapsulation (GRE). Available at: www.ietf.org (accessed May 29, 2020).
**57** Pan, P., Swallow, G., and Atlas, A. (Eds.) (May 2005). Fast reroute extensions to RSVP-TE for LSP tunnels. Available at: www.ietf.org (accessed May 29, 2020).
**58** Rosen, E. and Rekhter, Y. (February 2006). BGP/MPLS IP virtual private networks (VPNs). Available at: www.ietf.org (accessed May 29, 2020).
**59** Martini, L., Rosen, E., El-Aawar, N., and Heron, G.(April 2006). Encapsulation methods for transport of Ethernet over MPLS networks. Available at: www.ietf.org (accessed May 29, 2020).
**60** Farrel, A., Vasseur, J.-P., and Ash, J. (August 2006). A path computation element (PCE)-based architecture. Available at: www.ietf.org (accessed May 29, 2020).

**61** Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and Zekauskas, M. (September 2006). A one-way active measurement protocol (OWAMP). Available at: www.ietf.org (accessed May 29, 2020).

**62** Atlas, A. and Zinin, A. (Eds.) (September 2008). Basic specification for IP fast reroute: loop-free alternates. Available at: www.ietf.org (accessed May 29, 2020).

**63** Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and Babiarz, J. (October 2008). A two-way active measurement protocol (TWAMP). Available at: www.ietf.org (accessed May 29, 2020).

**64** Katz D. and Ward, D.(June 2010). Bidirectional forwarding detection (BFD). Available at: www.ietf.org (accessed May 29, 2020).

**65** Frost, D. and Bryant, S., and Ward, D. (September 2011). Packet loss and delay measurement for MPLS networks. Available at: www.ietf.org (accessed May 29, 2020).

**66** Frost, D. and Bryant, S. (Eds.) (September 2011). A packet loss and delay measurement profile for MPLS-based transport networks. Available at: www.ietf.org (accessed May 29, 2020).

**67** Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and Fulignoli, A. (Eds.) (October 2011). MPLS transport profile (MPLS-TP) linear protection. Available at: www.ietf.org (accessed May 29, 2020).

**68** Gray, E., Bahadur, N., Boutros, S., and Aggarwal, R. (November 2011). MPLS on-demand connectivity verification and route tracing. Available at: www.ietf.org (accessed May 29, 2020).

**69** Weingarten, Y., Bryant, S., Ceccarelli, D., Caviglia, D., Fondelli, F., Corsi, M., Wu, B., and Dai, X. (July 2013). Applicability of MPLS transport profile for ring topologies. Available at: www.ietf.org (accessed May 29, 2020).

**70** Grossman, E. (Ed.) (May 2019). Deterministic networking use cases. Available at: www.ietf.org (accessed May 29, 2020).

**71** Finn, N., Thubert, P., Varga, B., and Farkas, J. (October 2019). Deterministic networking architecture. Available at: www.ietf.org (accessed May 29, 2020).

**72** Geva, A. and Stein, Y. (March 2016). US patent 'PLUGGABLE MASTER CLOCK.

# 7

# NG-RAN Deployment Considerations

*Andreas Neubacher*[1] *and Vishwanath Ramamurthi*[2]

[1] *Deutsche Telekom, Austria*
[2] *Verizon Communications Inc., USA*

## 7.1 Introduction

In the present chapter we describe various issues for an operator to consider when deploying the cellular network. With the understanding that the choices operators are facing are bigger than just RAN (e.g. spectrum and radio technology choices being the other obvious examples), here we focus on RAN-related decisions, as the main focus of the book.

Generally, the cellular network operator's objective is to provide services to the customer, fulfilling the required demands in the most efficient and economical way. To this end, operators acquire spectrum, build networks, and deploy services that they believe will be attractive to users. Here we attempt to illustrate how operators' goals can be fulfilled using various technologies described in the book, the practical constraints that operators may face, and how these affect the choice of NG-RAN deployment options and architectures described in the book. In reality, these decisions are more complex and often involve other non-technical considerations, which go beyond the scope of this book.

In the past, the range of services offered by operators was rather limited, with voice, SMS, and low-bandwidth data being the primary, if not exclusive, usage. At that time, optimizing a limited number of parameters has been sufficient to develop cellular network deployment plans meeting the required objectives.

Over the years, with growing volumes of non-human-originated traffic, an increasing number of additional requirements on the wireless access technology and radio access network have emerged. As a result, meeting "the customer's'" expectations has become significantly more challenging for operators; this trend is expected to continue in the future. Furthermore, the nature of traffic consumed by human users has changed as well, from voice and messaging, to video streaming and gaming, to possibly virtual reality (VR) and other applications in the future.

Therefore, one key deployment objective for NG-RAN is to expand beyond the existing 4G mobile broadband use case, which has been successful in providing users with wide coverage, average throughputs of several tens of Mbps, and peak throughputs of the order

*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

of 1 Gbps. This enabled proliferation of a wide range of digital applications for users and at the same time set a higher expectation of what can be achieved through 5G networks. With 5G, operators need to adapt to evolving customer needs and provide new services that enhance the customer's experience beyond what can be imagined today.

To this end NG-RAN should be designed and dimensioned to provide faster downloads and lower latency, and enable new experiences on mobile devices. However, NG-RAN deployments aim to offer Network-as-a-Service (NaaS) not only for use cases involving human-operated devices but also for use cases involving Internet of Things (IoT) devices with each use case having its own service criteria. Different 5G use cases may need different RAN deployment architectures in order to satisfy their service criteria.

The challenges are rooted, among other factors, in various aspects of differences between human-originated (or consumed) traffic versus machine-originated traffic.

In this respect coverage per population (CPP), which is the dominant benchmark for network deployments for humans, needs to be reconsidered. In the case of traffic relating to machines and connected things, coverage per area (CPA) may be more suitable. Consequently, network deployment choices to satisfy CPA might be quite different compared with those designed to satisfy CPP.

Furthermore, NG-RAN is expected to be deployed in some radically new scenarios, for example, factories and enterprise campuses, which will present new challenges for the deployment plans, which mobile operators (or non-operator entities deploying and using these networks) have not faced before.

This may be especially important for factory and enterprise campus deployments, where operators will face additional competition from new service providers. Furthermore, regional spectrum allocations to new market entrants are expected to increase the frequency coordination efforts among the operators.

One additional example of more complex (compared with 4G) network deployments is the ultra-reliable low latency communication (URLLC) case, which is an important application of 5G. URLLC cannot be expected to be supported ubiquitously by simply deploying a 3GPP-compliant air interface technology (i.e. NR). In practice, an NG-RAN deployment supporting URLLC will require careful planning of the whole communication chain end-to-end. In this context, end-to-end encompasses all elements and components of the network, ranging from reliable and redundant fronthaul transport links, to backhaul links, to power supplies, to power grid connections, to cooling, and many other components. This is applicable to all network sites along the communication path, as all of them have to fulfill the reliability and redundancy Service Level Agreement (SLA).

Furthermore, network planning and dimensioning may be complicated by the fact that mobile operators may not necessarily own and operate the whole network end-to-end, including NG-RAN, backhaul transport, transmission, 5GC, etc. In such a case, providing end-to-end reliability and adequate redundancy may be challenging. To make URLLC work, reliability of outsourced components needs to be guaranteed, measured, and enforced. Operators outsourcing parts of their networks will need to set up appropriate SLA with their partners in order to successfully deploy URLLC (and other technologies).

Finally, and perhaps most importantly, it remains to be seen whether all the new 5G applications and use cases will finally lead to an economically feasible deployment at all, once all the constraints are taken into account.

## 7.2 Key Ideas

- When considering where to provide 5G coverage, in addition to CPP, which has been the primary tool used in 4G networks, operators now must consider using additional key performance indicators (KPIs) (e.g. CPA), especially for IoT applications. Furthermore, mmWave frequency range, with radically different radio propagation characteristics, warrants consideration of new tools to determine what is an acceptable coverage level (in terms of throughout, reliability, and other KPIs).
- Air interface capacity planning for IoT is radically different from that for mobile broadband (MBB) voice-centric (and even data-centric) use cases; it requires new approaches for which calculations based on the Erlang-B formula are no longer adequate. As these new methods are not available yet, operators may start by providing certain initial capacity and then upgrading in an "ad-hoc" manner, when the initially planned capacity is exhausted.
- Edge computing resource planning requires a new criterion that needs to be taken into account in NG-RAN planning, which is similar to data center resource planning, where not only communications but also computational resources must be properly dimensioned. Furthermore, as reliability planning (e.g. for mission-critical IoT applications) becomes important, mobile network operators may need to consider adopting four-tier reliability criteria developed by the Uptime Institute.
- Capital expenditure (CAPEX) and operational expenditure (OPEX) reduction are among the most important factors driving 5G RAN architecture redesign. In particular, technologies such as virtualization and standardization of the fronthaul interface to create truly multi-vendor interoperable network interfaces are expected to drive down NG-RAN deployment and operational costs. However, in order to realize these benefits, operators may in fact need to invest more in CAPEX, taking upon themselves at least some integration and testing efforts, in order to drive down OPEX in the future.
- Networks operating in the high frequency range (i.e. mmWave) require substantially more dense deployments, that is, a significantly higher number of remote units (RUs) (or Remote Radio Heads [RRHs]) or small cells. Additionally, availability and cost of fronthaul transport network, along with transport network throughput and latency capabilities, will affect which NG-RAN architecture options an operator can deploy. Lower-level NG-RAN functional splits (see Section 4.5), while providing higher baseband pooling gains and better radio resource coordination, impose substantially higher transport network requirements and may not always be economically feasible to deploy, especially for the mmWave frequency range.

## 7.3 Deployment Objectives and Challenges

### 7.3.1 Where to Provide Coverage

In the past, operators primarily needed to provide coverage for human-operated terminals and user equipment. However, with the advent of 5G, millions of connected things will also require coverage. Predicting the origin of the traffic generated by humans was relatively

easy, as humans tend to concentrate in certain areas. On the other hand, predicting the origin of traffic caused by connected things can be challenging or at least requires a different approach.

Therefore, while MBB coverage-related KPIs (e.g. CPP) remain relevant for these use case that still needs to be supported, additional considerations need to be applied for network buildout for IoT/machine-type communications (MTCs), which are outlined below.[1]

MTCs are often used to control and monitor physical parameters of objects without human intervention. MTCs are characterized by sometimes unusual locations of devices generating traffic, and additional requirements unique to IoT, such as latency, lead to new challenges. Specifically, the following needs to be considered when planning an IoT network:

- Coverage extension to remote areas: additional cell sites may be needed to provide coverage to connected things such as burglar alarms in solitary houses, supervision of remote technical equipment, smart farming devices, etc. This poses additional challenges (and expenses) as some of these devices will be deployed in areas where there are currently no humans at all and therefore no coverage.
- Densification: some IoT devices may be deployed in locations with bad path loss conditions, e.g. devices deployed in sewerage or in basements, etc. In that case increasing the maximum allowed path loss for a specific device (as is done with NB-IoT or CAT-M) is not sufficient or may lead to inadequate data rates for the expected service, and therefore the network may need to be densified.
- Multi-access edge computing (MEC): to support URLLC applications with extremely low latency requirements, edge nodes with compute resources are expected to move closer to the base stations, reducing trunking gains (i.e. gains realized by sharing equipment to provide services to many users) and increasing network deployment costs.

Operator planning departments have established procedures, expertise, and tools for setting up deployment plans based on population maps and current traffic data. Furthermore, feedback from customers about coverage issues and service availability in the case of MBB provides an additional source of input, which can be used for network optimization and troubleshooting.

On the other hand, with IoT, and other URLLC use cases, the established methods may not always work and therefore new methodologies for network planning must be devised. Furthermore, business relationships in that case become more complex, as it can be seen as a business to business to consumer (B2B2C) operation. This provides new challenges in terms of getting customer feedback.

Moreover, in the case of IoT deployments, the customer requiring connectivity or services may not necessarily have a business relation with the operator who deploys the network at all. Hence operators need to consider using new methods to understand where potential traffic may be originated, how much network capacity and compute power will be needed to support it, and where edge compute resources needed to be deployed to set up economically successful networks.

Access to and proper usage of the data relevant for network planning will be a key factor determining success and economic feasibility of a network.

---

1 The terms IoT and MTC are used interchangeably in this chapter.

### 7.3.2 Network Capacity and Compute Resource Planning

#### 7.3.2.1 Air Interface Capacity

In voice-centric cellular networks of the past the Erlang-B formula (Eq. (7.1)) and its variants, which provides the Grade of Service (GoS) as a probability that a new call is rejected, has been the dominant tool for planning of respective capacity of a RAN site.

$$P_b = B(E, m) = \frac{\frac{E^m}{m!}}{\sum_{i=0}^{m} \frac{E^i}{i!}} \tag{7.1}$$

Erlang-B formula

Where:

- $P_b$ is the probability of call blocking/dropping;
- $m$ is the number of identical parallel resources such as servers or telephone lines;

$E = \lambda h$ is the offered load in Erlang.

When upgrading their wireless mobile broadband networks, operators could rely on already available statistics from their existing previous-generation networks (e.g. when upgrading from 3G to 4G). Operators used to add traffic growth predictions to those statistics to create a traffic demand forecast for their deployment and capacity planning of the next rollout.

In the case of SLAs with customers (if any), respective bandwidth overprovisioning factors are typically added to the collected statistics to ensure that the committed SLAs can be met.

In reality, however, these capacity-planning processes have always been used together with "rule of thumb" techniques such as ad-hoc upgrading of the links whenever they exceed a specified capacity. For the 5G MBB use cases the same process could be applicable too.

IoT traffic, on the other hand, will present new challenges. It is well understood (and explained briefly in the previous section) that IoT devices come with unique requirements, for example, in terms of latency. However, there is a common misconception that IoT traffic is necessarily low bandwidth and infrequent.

While it is true that most of the traffic is likely to be from low-bandwidth devices such as sensors, these sometimes require firmware/software updates, causing devices to consume orders of magnitude more traffic compared with the regular operations.

Furthermore, some IoT use cases require the sending of megabytes of machine-specific data (e.g. comprehensive telemetry and sometimes even video) to companies providing operation and predictive maintenance for machines. Hence it is reasonable to assume that IoT traffic will not only rely exclusively on technologies such as NB-IoT, but also on other radio technologies, some of which can provide high data rates.

In addition to the problem of varying data volumes, there is the risk of multiple concurrent transmissions from multitudes of IoT devices, as assumptions on traffic arrival rate for humans (where the Poisson process is often used) may not hold for IoT. In fact, unless special precautions are taken, it is rather likely that at least in some cases certain events may cause massive simultaneous transmissions from IoT devices.

Technologies and standards are being already developed, for example, by the oneM2M Partnership Project (oneM2M), to overcome these issues. OneM2M-based technology allows among other things a proper scheduling of IoT application data transmissions, to allow a battery-efficient communication. That avoids adverse effects on the cellular networks (oneM2M Technical Specification TS-0001). However, until oneM2M-developed (or similar) technologies gain sufficient market penetration, there will be always a risk of unintentional denial of service (DoS) from IoT applications, which may have been developed without taking the specifics of cellular connectivity into account.

#### 7.3.2.2 Compute Resources for Edge Computing Services

Edge computing is expected to play a major role in 5G deployments; for example, to ensure end-to-end low latency. For more details about edge in general and MEC in particular, refer to Section 6.4.

NR radio interface can provide an extremely low one-way latency of 0.5 ms. However, besides latency introduced by the air interface on the "last mile," latency introduced owing to data transmission over the backhaul links (as also discussed in Section 6.4) and through the core network to the other peer communication partner (e.g. an application running in a data center) contributes considerably to the end-to-end delay/latency budget.

The simple back-of-the-envelope calculation provided below illustrates the issue. Using the simple formula $s = c * t$,where:

$c$ represents the speed of light $c = 299\,792\,458$ m/s;
$t$ represents the time.

If 1 ms round-trip end-to-end latency is required (which is often assumed in 5G URLLC applications), that is t = 1 ms, then this translates to a distance of:

$$s = 299\,792\,458 \times 10^{-3} = 299\,792\,458 \text{ m} = 299.8 \text{ km} \approx 300 \text{ km}$$

In other words, once a round-trip delay of less than 1 ms is required, the distance between both communication partners must be less than $s_{max} = {}^{300}/_{2} = 150$ km, to allow ping times of 1 ms.

Assuming base stations are connected via fiber, an additional media delay should be taken into account. That is, just ~67% of the speed of light can be assumed for the information transfer.

In this case the distance between two communication partners need to be even less than $s_{max}$, that is, $150 \text{ km} \times 0.67 \approx 100$ km. Needless to say this delay cannot be mitigated by any engineering means.

Furthermore, there are likely to be additional unavoidable delays due to serialization of the data before transmission over the media, routing and switching delay, etc., which limits the maximal distance even further considerably below 100 km to meet a maximum allowed latency of, for example, 1 ms.

Considering the above, the only feasible way to provide low end-to-end latency communications is placing remote compute resources closer to the customer. This is what MEC and other related technologies have been designed for.

When it comes to network planning to realize the full benefits of URLLC applications, an operator or provider of such edge services needs to identify the location of customers

requiring such low latency services and to place the necessary compute resources within the required range to meet the requested requirements.

One of the new NG-RAN architectures, specifically the central unit/distributed unit (CU/DU) split described in Section 4.2, makes it easier to deploy edge applications such as MEC by concentrating more compute-intensive functions of the 5G radio in a central place (e.g. in the same cell site where gNB-CU is deployed). This allows sharing of compute resources to serve the load generated in the gNB-CU by several attached gNB-DUs, which can also be shared with MEC.

Concentration of compute hardware in larger data centers allows efficient reuse of other resources such as cooling, power, and broadband/backhaul connectivity, helping to provide URLLC services in a reliable, resilient, failsafe, and economical way.

As with many other examples of resource dimensioning, on the one hand there is a danger of overproviding edge compute resources. On the other hand, not providing enough compute capacity at the edge when it is needed may lead to the service expectations of customers being missed. This leads to a tradeoff between the risk of not meeting the SLA and the cost of building and operating a network to satisfy it.

In any case operators will need to deal with new KPIs, that is, cell edge compute resources, in their cell site planning and need to explore new parameters to measure customer satisfaction and resource usage.

At least initially, operators are likely to use simple heuristic rules to provision compute services, by deploying small scale and upgrading resources whenever a certain benchmark exceeds a specified demand.

#### 7.3.2.3 Reliability Considerations

Another important criterion for 5G deployments is reliability, which in itself is not new; however, different 5G use cases require different reliability levels. For example, in mmWave deployments supporting high throughputs, it is important to achieve a certain degree of reliability on par with that of lower frequencies (which is of the order of 99.99%) to provide consistent high average throughputs to users. For ultra-reliable low latency use cases, the reliability requirements may be even higher, for example, in the case of mission-critical applications.

In the past, the core network was the primary source of an operator's concern in reliability considerations. Failures in core network elements may cause considerable user and service impacts as they can affect large areas (and even potentially the whole network), while a failure in a single RAN cell site may only have a geographically isolated impact.

In 5G, at least with some mission-critical IoT services, RAN failures will be harder to tolerate. Furthermore, in addition to radio reliability considerations, which are well understood from previous generations of cellular deployments and have been properly addressed by NR design, there are additional factors related, for example, to compute resources reliability. That is because NG-RAN may be virtualized (see Section 6.2) and may provide edge computing functionality (see Section 6.4), all of which rely on the availability of edge compute resources. To account for these new features, which can also become additional failure points, the cellular industry may consider adapting a classification similar to the ones used in data centers. For example, the tier rating defined by the Uptime Institute (Uptime Institute Tier Standard) may be a helpful starting point.

Four tiers defined by the Uptime Institute are:

Tier I: no redundant IT equipment, 99.671% availability.
Tier II: some redundant infrastructure, 99.741% availability.
Tier III: additional data paths, more redundant equipment, which is all dual powered, 99.982% availability.
Tier IV: fault tolerance, all cooling equipment is dual powered, 99.995% availability.

Tier IV is designed to guarantee "zero single points of failure." A Tier IV provider needs to ensure redundancies for every process and data protection stream. Any outage or error must not shut down the system. 2N + 1 redundancy needs to be provided, meaning the provisioned infrastructure is two times the amount required for operation plus a backup. A Tier IV infrastructure needs to foresee a 96-hour power outage protection.

Even though Tier IV is economically challenging, it may be feasible for a centralized data center. However, for a distributed infrastructure such as a mobile network providing a full nationwide coverage, it is questionable whether providing Tier IV reliability will justify the investment in the long run. Hence operators are likely to focus on selected customers with a need for such high availability requirements in a small coverage footprint only, for example, factories and enterprise campuses, while most other deployments will have to rely on a lower tier.

### 7.3.3 Service Fulfillment Criteria

As discussed previously, for the MBB use cases in the past, one of the major benchmarks was CPP accompanied by certain (e.g. average and minimum) data throughput thresholds. This remains important in 5G, as one of the key criteria from deployment perspective is to deliver an enhanced MBB experience to users. However, 5G networks have the potential to deliver speeds many times faster than the current 4G networks based on wider spectrum bandwidths as well as better spectral efficiency. 5G is expected to ultimately enable peak data rates in the order of 10 Gbps and average throughputs in the order of several hundreds of Mbps. This will enable various applications such as AR/VR solutions, smart home, mobile gaming, remote healthcare services, autonomous vehicles, etc. It is reasonable to assume that CPP and minimum throughput requirements may need to be adjusted to reflect extreme 5G throughputs.

Furthermore, the mmWave spectrum provides a large swathe of bandwidths of several hundreds of MHz, which can be utilized to enable ultra-high speeds to users. However, these frequencies have propagation characteristics that limit the coverage distance to only a few hundred meters in non-line-of-sight (nLOS) conditions. Also, mmWave frequencies are subject to high losses due to foliage and other obstacles. Providing reliable coverage using mmWave requires careful network planning including overall RAN deployment architecture, selection of antenna technology, and placement of RUs taking into account the local geographical environment at a level of granularity that was never done in the previous generations of cellular networks. These include detailed maps considering foliage, buildings, roads, and other local infrastructure. Local and seasonal weather conditions also need to be taken into account during practical deployments to ensure that seasonal snow, weather, foliage growth, etc. do not impact customer experience.

Smaller wavelengths at mmWave enable use of several hundreds of antenna elements in a compact form factor that enables targeted beamforming to overcome the propagation difficulties at these frequencies. Analog and hybrid beamforming technologies are becoming a key in mmWave 5G deployments as opposed to just digital beamforming, providing another tool to help fulfilling required service criteria. In theory, CPP and throughput threshold criteria can still be applied to the mmWave spectrum; however, the mmWave propagation characteristics mentioned above pose some challenges in applying these simplistic criteria.

Therefore, as mentioned above, CPA is receiving more attention in the effort to address new 5G use cases. In the case of URLLC, completely new criteria are getting traction, such as latency and delay budget. When latency is considered, all network components contributing delays must be accounted for, including firewalls, switches, and routers.

In the past (e.g. 3G and 4G), typically only the terminals went through a formal certification process, while it was the responsibility of every operator to ensure that the network was built to provide adequate capacity and other KPIs. With 5G, in order to guarantee end-to-end reliability, it is possible that not only terminals need to undergo certification. It can be envisioned that certification of the whole network will be necessary to allow a customer (in case the customer is an enterprise) or a regulator to determine whether a certain operator has built the network, which is reliable and resilient end-to-end, and whether the internal processes are able to guarantee downtimes as specified in the "tier ratings" (if adopted by the cellular industry).

It appears that, at the time of writing, there are no widely adopted 5G service fulfillment criteria addressing all the above issues.

## 7.4 Deployment Considerations

### 7.4.1 Deployment Cost

In order to run a sustainable business, operators must deploy and operate their networks, NG-RAN in particular (which accounts for the largest part of CAPEX and OPEX of a mobile network), in a cost-effective manner while at the same time achieving the desired service objectives. One of the most sensitive parameters in terms of cost is an available and used spectrum for the deployment. The issue of spectrum pricing or other models of spectrum allocations to operators (and other entities) is generally up to a local regulator and is beyond the scope of the book.

Other important aspects of 5G RAN deployment include cost of RUs (or RRHs), cost of baseband unit (i.e. DUs and CUs), cost of transport network, which includes fronthaul, and real estate costs. Quite a number of NG-RAN standardization activities (described in Chapters 4 and 6) have been driven by the desire of operators to drive down NG-RAN CAPEX and OPEX.

For example, there is the expectation that virtualization will help driving CAPEX down by decoupling hardware from software and allowing operators to source these from different vendors. Furthermore, virtualization may help driving down OPEX by allowing higher levels of automation in the network (see Section 6.5). While there are reasons to believe that virtualization does drive OPEX down, as was observed in for example, data centers,

one must remember that it also increases CAPEX, as virtualized implementation tends to consume more compute resources and more power than ones that run on custom hardware. The latter may eventually become negligible thanks to Moore's law; however, one must also remember that at least as of now it appears to be hard to fully virtualize radio hardware, therefore it may take time until operators are capable to fully realize virtualization gains.

Additionally, opening up network interfaces by moving from proprietary Common Public Radio Interface (CPRI)-based fronthaul to more rigorously standardized interfaces defined for split architectures (see Sections 4.2 and 4.5) should spur competition in NG-RAN, potentially driving down costs by allowing true multi-vendor NG-RAN deployments. The caveat is that even fully standardizing network interfaces is not enough to achieve that goal, as non-negligible integration and interoperability testing effort still remain. In order to allow true competition between NG-RAN vendors, operators may need to take the integration effort upon themselves, or use the services of a third-party integration – thus bearing additional costs, at least in the initial phase.

### 7.4.2 Spectrum and Radio Propagation Considerations

In the present section we provide some basic considerations related to network planning, taking into account radio propagation. This is meant to provide just a glimpse into this complex topic, which deserves a book of its own.

The relation between the most fundamental aspects in the context of the used spectrum can be derived from the free space propagation formula as given below (Eq. (7.2)).

$$P_e = g_s.ge\frac{P_s}{4.\pi.d^2} \cdot \frac{\lambda^2}{4.\pi} = P_s \cdot g_s \cdot g_e\left(\frac{\lambda}{4 \cdot \pi \cdot d}\right)^2 = P_s \cdot g_s \cdot g_e\left(\frac{c}{4.\pi.d}\right)^2 \cdot \frac{1}{f^2} \tag{7.2}$$

Where:

$P_e$ = *power at the receiver input*; $P_s$ = *power at the sender output*;

$g_e$ = antenna gain at the receiver, $g_s$ = antenna gain at the sender;

$d$ = distance between sender and receiver; $\lambda$ = wave length of the carrier frequency;

$\lambda = \frac{c}{f} = \frac{3.10^8\, m/s}{f}$ where: c = $3.10^8$ m/s (speed of light) and f = frequency in Hz.

Transforming Eq. (7.2) leads to Eq. (7.3):

$$\frac{P_e}{g_e} = P_s \cdot g_s \cdot \left[\frac{\lambda}{4\pi \cdot d}\right]^2 \text{ and further to, } \frac{P_s \cdot g_s}{\frac{P_e}{g_e}} = \left[\frac{4\pi \cdot d}{\lambda}\right]^2 = L \tag{7.3}$$

Where $L$ represents the path loss between the sender and the receiver.

Using a logarithmic representation of Eq. (7.3) leads further to Eq. (7.4):

$$L_{[dB]} = 10\log\left[\frac{4\pi \cdot d}{\lambda}\right]^2 = 20\log\left[\frac{4\pi \cdot d.f}{c}\right] = 20\log(d) + 20\log(f) + 20\log\left(\frac{4\pi}{c}\right)$$

$$= L_{[dB]} = 20\log(d) + 20\log(f) - 147.56 \tag{7.4}$$

Using the equations listed above we attempt to derive some useful rules of thumb.

Assuming a certain constant maximum allowed path loss $L$, we are interested in calculating the maximum distance between the sender and the receiver, depending on the carrier frequency.

Using Eq. (7.3) we arrive at Eq. (7.5):

$$d = \frac{\sqrt{L.}}{4\pi} \cdot \lambda = \frac{\sqrt{L.}}{4\pi} c.\frac{1}{f} = K.\frac{1}{f} \tag{7.5}$$

Where the term $\frac{\sqrt{L.}}{4\pi} c$ can be regarded as a constant $K$.

This shows that the maximum distance $d$, between sender and receiver is indirectly proportional to the carrier frequency. In other words, doubling the carrier frequency from $f$ to $2f$ reduces the maximum distance by a factor of 2.

However, in order to use the above result for actual deployments and the related cost analysis, we need a more important parameter, which is the related number of base stations or RUs required to cover a certain area.

For reasons of simplicity, we assume a perfect circular coverage by a base station (even though in practice most deployments today use three-sector base stations). The coverage area can be calculated by:

$$A_{circle} = r^2 \cdot \pi$$

Where $r$ represents the radius.

Inserting $d$ from Eq. (7.4) into the formula for the area of a circle leads to Eq. (7.6):

$$A_{circle} = \left(K \cdot \frac{1}{f}\right)^2 \cdot \pi = K_1 \cdot \frac{1}{f^2} \tag{7.6}$$

Where the constant $K$ and $\pi$ have been combined to create a new constant $K_1 = K^2\pi$.

From Eq. (7.6), we further deduce that the coverage area is indirectly proportional to the square of the carrier frequency. In other words, doubling the carrier frequency from $f$ to $2f$, decreases the coverage area by a factor of 4, which means that for the case of deploying a technology on the double carrier frequency, four times more NG-RAN base station sites are needed. This is of course nothing more than a back-of-the-envelope calculation, but it does illustrate the network densification problem mentioned in previous chapters.

Another interesting aspect is the question about the actual loss or gain in dB, depending on the carrier frequency.

Assuming $d$ is constant, Eq. (7.4) can be transformed to:

$$L_{[dB]} = 20\log(d) + 20\log(f) - 147.56 = K_2 + 20\log(f) \text{ where } K_2 = 20\log(d) - 147.56$$

In other words, doubling the carrier frequency from f to 2f, leads to:

$$L_{[dB]} = K_2 + 20\log(2.f) = K_2 + 20\log(f) + 20\log(2) = K_2 + 20\log(f) + 6.02 =>$$

That is, a path loss increases in 6 dB.

This shows that technology advantages designed to increase the maximum coupling loss (MCL) (e.g. NB-IoT MCL increase from 144 dB of GSM to 164 dB) can be lost quickly, once a certain technology is being deployed on higher frequency bands without taking care of proper network planning with appropriate inter-NG-RAN site distances.

The examples above are just simple considerations based on the free space propagation formula (FSPF). Even though the FSPF isn't applicable for wave propagation in urban areas, it allows the estimation of the number of base stations required for the cellular deployments, independent of the technology used.

Detailed channel modeling for different NG-RAN deployment scenarios is available in 3GPP TS 38.901 (3GPP TS 38.901), which can be used to further refine the back-of-the-envelope calculations shown above.

### 7.4.3 5G Frequency Ranges

Currently the frequencies being considered for NG-RAN across the world include low, middle, and high frequency ranges. The propagation characteristics for each of these ranges are different and justify different kinds of NG-RAN deployment architecture.

In the lower ranges, a typical coverage of a macro cell is around 2–2.5 km. In the middle range, it is around 1–1.5 km and in high ranges (mmWave) it is around 100–500 m. In terms of bandwidth, lower frequency ranges offer tens of MHz of bandwidth, middle ranges up to 100 MHz of bandwidth, and mmWave several 100 MHz of bandwidth. This is illustrated in Table 7.1.

In other words, lower bands provide larger coverage with lower bandwidth, while higher bands provide larger bandwidth with lower coverage. These characteristics lead to different deployment architectures for high bands versus mid and low bands.

One particularly important conclusion is that while deploying 5G in the middle frequency ranges using the same cell sites as 4G (which can be considered low range) is feasible, the high frequency range (i.e. mmWave) will require much more dense deployment of small cells (or RUs), which inevitably leads to new cell site acquisition and drives up development costs.

If, for example, the coverage range (of e.g. mmWave) is 5 times smaller, then 25 times more RUs/small cells will be required to cover the same area (as shown above), which of course has a direct impact on CAPEX and OPEX. Covering an operator's entire service area using the high frequency range would require a huge number of RUs, therefore mmWave radios are expected to be deployed as a capacity booster in places of high throughput demand rather than as a blanket coverage layer, which is best provided for using low and middle range radios.

**Table 7.1** Typical bandwidths and coverage ranges for different 5G frequency ranges.

| | Typical bandwidth (MHz) | Typical coverage range |
|---|---|---|
| Low frequency range (<1 GHz) | 10–60 | 2–2.5 km |
| Mid-frequency range (1–6 GHz) | 100 | 1–1.5 km |
| High frequency range (>24 GHz) | 400–1200 | 100–500 m |

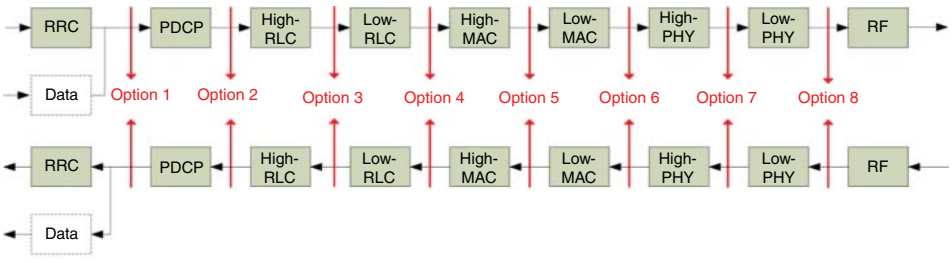



**Figure 7.1** Possible NG-RAN functional splits (Source: Reproduced by permission of © 3GPP).

### 7.4.4 Transport Considerations

Whichever band a NG-RAN radio is operating in, it needs to be connected to the core network and Internet through some kind of backhaul and often fronthaul and midhaul transport network (see Section 6.6). While the bandwidth required for the backhaul depends on user data rates only, the fronthaul bandwidth also depends on the RAN architecture protocol split. Figure 7.1 (3GPP TR 38.801) illustrates various functional splits considered in 3GPP and other organizations (e.g. O-RAN), which are explained in detail in Chapter 4.

When modeling these architectures, a NG-RAN node is typically split into a CU and a DU, with the functionality in each depending on the split option. In some cases, a DU may be further split into a DU and an RU. As a general rule, lower splits (confusingly designated by higher split numbers) have higher centralization of functionality at the cost of increasingly stringent fronthaul bandwidth and latency requirements. All these options are described in detail in Chapter 4.

Table 7.2 shows an approximate fronthaul bandwidth (FHBW) required for various protocol splits for low-, mid-, and high-band scenarios assuming 8 antennas per sector, 4 layers, 8 bits per I/Q sample, and an overhead of 25%. This is just an approximate back-of-the-envelope calculation, as in practice it depends on many factors such as quantization, especially for low-level splits.

**Table 7.2** Fronthaul bandwidth for low-, mid-, and high-band scenarios.

| | Low band | Mid-band | High band |
|---|---|---|---|
| **Total bandwidth (MHz)** | **40** | **100** | **800** |
| **Split type** | **Fronthaul bandwidth (Gbps)** | | |
| Options 2 (Packet Data Convergence Protocol–Radio Link Control [PDCP-RLC]) | 0.7 | 1.5 | 12.0 |
| Options 5 (High Medium Access Control [MAC] – low MAC) | 0.8 | 1.7 | 13.3 |
| Option 7-2 (Intra PHY) | 2.7 | 6.7 | 53.8 |
| Option 7-1 (Intra PHY) | 5.4 | 13.4 | 107.5 |
| Option 8 (PHY/RF) | 9.2 | 22.9 | 183.5 |

The required FHBW is the lowest for Option 2, being very close to actual user traffic with some necessary protocol overheads. The required FHBW increases as we split down the NG-RAN protocol stack with quantized modulation symbols being transported in Options 7 and 8. Low level split Options 7 and 8 provide the ability to carry out coordinated processing across multiple radios at the cost of larger FHBW.

The FHBW required for Options 7 and 8 for the high frequency range scenario is especially significant.[2] Therefore, transport network availability and cost becomes an important factor in selecting the right NG-RAN architecture. Large-scale mmWave deployment with lower-layer split would dramatically escalate transport network deployment costs, making it significantly less attractive for operators.

For low- and mid-band scenarios, FHBW using Option 7/8 might still be manageable considering the bandwidth and coverage in those bands.

In addition to fronthaul transport network throughput limitations, which impose restrictions on NG-RAN architectures an operator can use, latency is also a factor that needs to be considered.

Table 7.3 shows an example of typical fronthaul latency requirements for various protocol split options. The latency requirements for lower-layer split Options 7 and 8 are limited by Hybrid ARQ (HARQ) loop and coordinated processing requirements and therefore are tighter, while the higher-layer split option is more relaxed in terms of latency, which is limited only by end-to-end performance requirements.

To view the impact of fronthaul latency requirements on NG-RAN deployment architecture, consider an example circular deployment area of 100 km radius and assume that DUs can be placed anywhere within this area to provide the required coverage and capacity. The maximum distance from a CU to DUs (or RUs) is limited by the corresponding latency requirement. Assuming an optical fiber deployment for fronthaul network and assuming speed of light in fiber to be $2 \times 10^5$ km/s (two thirds of the speed of light in vacuum), a rough estimate of the number of CUs required for the deployment area for different CU–DU round-trip latency requirements is shown in Figure 7.2.

As the round-trip latency requirements become less stringent, the number of CUs required reduces significantly, for example, in Option 2. On the other hand, the lower-layer split Options 7/8 that have tight latency requirements would need a larger number of centralized sites (i.e. CUs).

**Table 7.3** Latency requirements for various split options.

| Split | Round-trip latency | Comment |
|---|---|---|
| Option 2 | 5–10 ms | Limited by end-to-end performance |
| Option 5 | 0.5–3 ms | Limited by real-time Radio Link Control functionality |
| Options 7–8 | ~50 µs | Limited by Hybrid ARQ loop and coordinated processing requirements |

2 Note that the FHBW requirements stated below are per RU.

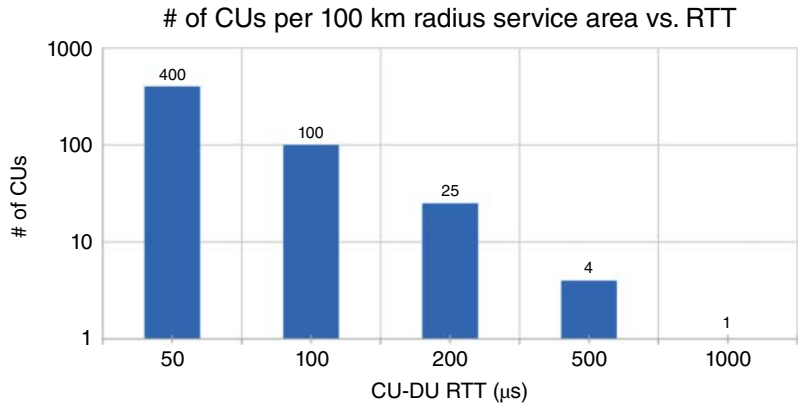



**Figure 7.2** Estimate number of CUs as a function of transport network round-trip time.

### 7.4.5 Baseband Pooling

Baseband pooling is one of the two main advantages of centralized NG-RAN (based on split architectures mentioned above), with the other one being centralized radio resource coordination and scheduling. In the past, most of the baseband compute, especially PHY/Medium Access Control (MAC)/Radio Link Control (RLC) parts of the protocol stack were done in specialized hardware specifically designed for that purpose. However, the processing capability of general-purpose processors has evolved, and specialized accelerators have been added to these. As a result, general-purpose hardware has evolved to such an extent that it can be used for baseband compute including MAC and RLC layers (and to some degree even PHY). These advancements, together with standardization of network interfaces between split NG-RAN nodes, open the possibility of virtualization (see Section 6.2) of the entire RAN protocol stack and the resulting savings for operators.

Lower-layer split can enable baseband consolidation through pooling of multiple basebands at a centralized location. Consider an example case where a base band unit (BBU), which can be a combined CU and DU using the terminology defined in this book, is implemented using a general-purpose compute that has the compute capacity to handle 100 sub-6 GHz cells (sector carriers) and assume that each deployment site has 10 cells. A centralized NG-RAN architecture based on Options 7/8 can lead to pooling of resources across 10 different sites and result in a 10 times resource reduction for BBUs. This shows that pooling of BBUs using lower-layer split can lead to heavy savings for operators by reducing the number of BBUs required for a certain deployment area.

On the other hand, high-band cells require much higher compute capacity per sector carrier. Therefore, the pooling gain using Options 7/8, if any, is negligible. Each high-band deployment site would typically need one or more BBUs/DUs using a lower-layer split. Additionally, since high-band deployments tend to be dense, the number of BBUs/DUs required would be higher leading to higher requirements in terms of space, power, and cooling at the BBU pool locations. This leads to higher real estate and infrastructure costs making lower-layer split less attractive for high-band deployment scenarios.

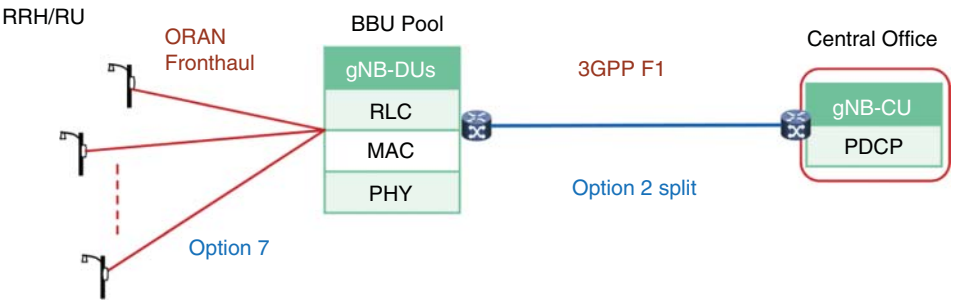



**Figure 7.3** Two-level split NG-RAN architecture suitable for sub-6 GHz frequencies.

### 7.4.6 Choice of a NG-RAN Split Architecture

Here we consider typical 5G RAN deployment architectures for sub-6 GHz (including low and mid-bands) and high-band scenarios considering the various factors discussed in the earlier sections.

#### 7.4.6.1 Sub-6 GHz Case

Sub-6 GHz (low- and mid-frequency ranges) radio deployments with larger propagation distances and respectable pooling gains can benefit from using both the higher-layer Option 2 split and lower-layer Option 7/8 split in a hierarchical architecture as shown in Figure 7.3. This would involve two levels of centralization: first sector carriers from multiple remote radio heads (or RUs) can be connected to a local BBU (i.e. DU) pool using Option 7/8, which can be then further centralized using split Option 2 in a telco central office.

Lower-layer split Options 7/8 are appropriate for sub-6 GHz frequencies where it is important to obtain the coordinated processing gains using mechanisms such as Coordinated Multi-Point (CoMP) Tx and Rx. The resulting fronthaul requirements in terms of FHBW and latency are manageable using current optical transport technologies. Furthermore, the benefits of baseband pooling can be realized and both CUs and DUs can be virtualized

#### 7.4.6.2 High-Band (mmWave) Case

As discussed above, high frequency range deployments are characterized by larger bandwidths and smaller propagation distances. These mmWave deployments are expected to be more appropriate for dense pockets in critical demand areas of an operator's overall coverage area. Considering high FHBW as well as tight latency requirements for lower-layer split Options 7/8, it might not be suitable for high-band mmWave deployments, which are typically expected to be used to provide ultra-high throughputs in certain critical areas. The costs of the lower-layer split option for a high frequency range scenario in terms of transport and real estate far outweigh any benefits in terms of cooperative processing and pooling, which are rather minor.

High frequency range deployments are thus more suited to single-level centralization using Option 2 as shown in Figure 7.4. In this architecture, the lower layers of the protocol unit (i.e. DU) are co-located with the RU (either integrated or connected via a short cable). The Option 2 split reduces the fronthaul transport network requirements to a reasonable

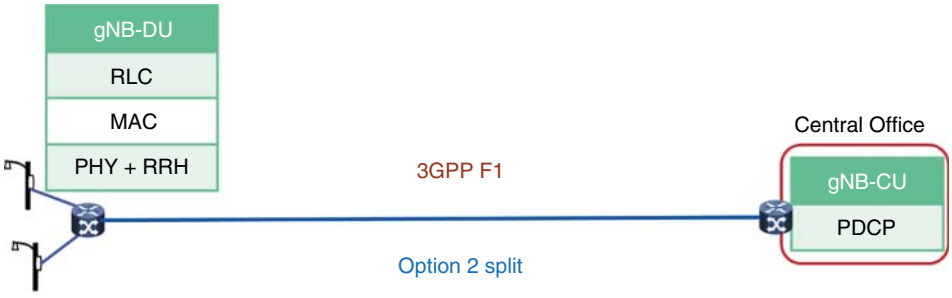



**Figure 7.4** Single-level split NG-RAN architecture suitable for mmWave frequencies.

level, making it easier to deploy. Furthermore, this architecture only needs a limited number of centralized sites, which can also benefit from virtualization.

## 7.5 Conclusions

In the present chapter we tried to provide a glimpse into a plethora of complex considerations an operator must go through when designing its networks. In relation to the choice of an appropriate NG-RAN architecture, these considerations range from geographical areas where a service needs to be provided, to applications that are expected to be used in the deployed network, to the spectrum frequency range used, and to availability and capabilities of the transport network. As one can see, while there are some well-known planning methods that have been used in the past (i.e. 4G), when it comes to 5G deployments there is no established methodology yet, which will have to be developed when 5G networks are rolled out and new services making use of these networks emerge.

## References

3GPP Technical Report 38.801 (2017). Study on new radio access technology: radio access architecture and interfaces. Available at: www.3gpp.org (accessed May 29, 2020).

3GPP Technical Report 38.901 (2019). Study on channel model for frequencies from 0.5 to 100 GHz. Available at:www.3gpp.org (accessed May 29, 2020).

Common Public Radio Interface (CPRI) (2013). Interface specification. Available at:www.cpri.info/spec.html (accessed May 29, 2020).

oneM2M Technical Specification TS-0001 (2016).Functional architecture. Available at: www.onem2m.org (accessed May 29, 2020).

oneM2M White Paper (2015). The interoperability enabler for the entire m2M ecosystem.

Uptime Institute (2014). Tier standard: operational sustainability. Available at: https://uptimeinstitute.com/resources/asset/tier-standard-operational-sustainability (accessed May 29, 2020).

# Index



*5G Radio Access Network Architecture: The Dark Side of 5G,* First Edition. Edited by Sasha Sirotkin.

## *d*



## *e*

## *f*



## *g*

## *h*



## *i*



## *j*



## *k*



## *l*

### *m*



### *n*

### *O*

## *p*

## *q*



## *r*

## *t*



## *u*