Understanding 5G

A Practical Guide to Deploying and Operating 5G Networks

Second Edition
by VIAVI Solutions

Understanding 5G

A Practical Guide to Deploying and Operating 5G Networks

Second Edition
by VIAVI Solutions

Limit of Liability/Disclaimer of Warranty: While VIAVI has used best efforts in preparing this book, VIAVI makes no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional when appropriate. VIAVI shall not be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, personal, or other damages.

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately identified.

Understanding 5G, Second Edition
By VIAVI Solutions

Printed in the United States of America

Table of Contents

Foreword ... xi

Preface ... xiv

Chapter 1: The 5G Evolution Story ... 1
- 1.1 Naming of Parts ... 3
- 1.2 Once Upon a Time ... 6
  - 1.2.1 The 1980s and Early 1990s: Here Cometh "GSM" ... 6
  - 1.2.2 Late 1990s to Early 2000s: The Almighty UMTS ... 10
  - 1.2.3 The 2010s: The Rise of LTE ... 13
  - 1.2.4 Virtualization and Telecommunications Lifecycle ... 16
  - 1.2.5 2017: The Game-Changer 5G ... 18
- 1.3 5G Use-Case Based Service Drivers ... 22
- 1.4 Independence of RAN from CN ... 25
- 1.5 Control and User Plane Separation (CUPS) ... 26
- 1.6 Service-Based Architecture in 5G CN ... 27
- 1.7 Virtualization and Orchestration ... 27
- 1.8 Disaggregation and Open Interfaces ... 28
  - 1.8.1 O-RAN ... 30
- 1.9 Network Slicing ... 31
- 1.10 Navigating the 3GPP Standards for Architecture ... 32
- 1.11 5G Architecture Blueprint ... 33

Chapter 2: 5G New Radio ... 37
- 2.1 5G to Simultaneously Reduce Latency by an Order of Magnitude and Increase Data Speeds by One to Two Orders of Magnitude over LTE ... 39
- 2.2 Frequency Bands for 5G ... 39
- 2.3 5G Waveforms, Numerologies, and Frame Structure ... 42
  - 2.3.1 Symbols and Modulation ... 42
  - 2.3.2 Subcarriers ... 44
  - 2.3.3 Guard Bands and Spectral Efficiency ... 45
  - 2.3.4 Flexible Subcarrier Spacing and Numerologies ... 45
  - 2.3.5 Frame Structure and Slots ... 47
  - 2.3.6 Bandwidth Parts ... 52
  - 2.3.7 Blank Slots ... 53
- 2.4 Channel Coding ... 53
- 2.5 NR Medium Access Control (MAC) layer and HARQ for flexibly balancing latency, reliability and energy efficiency requirements ... 53

2.5.1 An Introduction to HARQ: Reliably sending a signal through a noisy channel 54

2.5.2 Creating a configurable 5G NR MAC layer to balance reliability, latency and power efficiency 55

2.5.3 Exploiting gNB and UE processing time capability 56

2.5.4 Configuring PDSCH to trade-off latency, reliability and power efficiency 57

2.5.5 Configuring PDCCH monitoring to trade-off latency, reliability, power efficiency 58

2.5.6 Configuring PDSCH duration to trade-off latency, reliability, power efficiency 59

2.5.7 Configuring PUCCH duration to trade-off latency, reliability, power efficiency 59

2.5.8 Configuring PUCCH duration to trade-off latency, reliability, power efficiency 60

2.5.8.1 Grant-free operation for UL transmission 61

2.5.8.2 Pre-emption of UL and DL resources 61

2.5.8.3 Code Block Group-based re-transmission 61

2.5.9 Review of NR HARQ v LTE HARQ (with some examples) 62

2.6 MIMO, Massive MIMO, and Beamforming 65

2.6.1 History of Spectrum Reuse in Wireless Access 65

2.6.2 Introduction to MIMO 65

2.6.3 SIMO and Receive Diversity 67

2.6.4 MISO 68

2.6.5 MIMO 69

2.6.6 Introduction to Beamforming 74

2.6.7 Pure Digital Beamforming and Hybrid Analog/Digital Beamforming 75

2.6.8 Support for MIMO and Beamforming in 5G NR 76

2.6.9 Comparison of MIMO and Beamforming 81

2.7 Energy Saving Support in 5G NR 82

2.8 Standards Evolution in 5G NR 83

2.8.1 Evolution of URLLC 83

2.8.2 Evolution of mMTC 84

2.8.3 Evolution of eMBB 86

2.9 Navigating the 3GPP Standards for 5G NR 87

2.10 5G NR and Network Slicing 89

Chapter 3: 5G Radio Access Transport Networks 91

3.1 Fronthaul Networks 93

3.1.1 Optical Layer Technology 94

3.1.2 CPRI (Common Public Radio Interface) 97

3.2 Synchronization Networks 100

3.2.1 Synchronization Requirements and Technologies 100

3.2.2 Frequency Synchronization Standards 102

3.2.3 Time and Phase Synchronization Standards 107

3.3 5G Transport Networks 111

3.3.1 Fronthaul Challenges and Functional Split Options 111

3.3.2 Higher Layer Split (HLS), Lower Layer Split (LLS), and eCPRI 113

3.3.3 Timing Sensitive Network (TSN) 118

3.3.4 Emerging 5G Synchronization Requirements 120

3.3.5 Radio over Ethernet (RoE) and Fronthaul Gateways (FHGW) 121

3.4 5G RAN and Network Slicing, Summary and Outlook 122

Chapter 4: 5G Core 125

4.1 Mobile Core History 126

4.1.1 LTE EPC Components 126

4.2 The Case for Next Generation Core (NGC) 127

4.3 4G Evolved Packet Core (EPC) versus 5G Next Generation Core (NGC) 129

4.4 5G Service-Based Architecture (SBA) 130

4.5 5GS QoS Architecture and Models 141

4.5.1 PDU Connectivity Service and PDU Session 141

4.5.2 PDU Session is Network-Slice Native 142

4.5.3 5G QoS Model 142

4.5.4 5G QoS Profile 143

4.5.5 5G QoS Identifier and Characteristics 144

4.5.6 5G Reflective QoS 145

4.6 5G NGC and Network Slicing 145

4.6.1 Introduction to Network Slicing 145

4.6.2 5G Core Network Slices Topology 146

4.6.3 5G Network Slice Identifiers 148

4.6.4 5G Network Slice Selection 149

Chapter 5: Edge Computing 151

5.1 Standardization Overview 153

5.2 MEC Application Enablement Framework 156

5.3 MEC Services 158

5.3.1 Network Information Services 159

5.3.2 Location Service 161

5.3.3 Bandwidth Management Service 161

5.3.4 UE Identity Service 161

5.3.5 V2X Service 161

5.4 MEC Deployment Options 161

5.4.1 S1 Based 161

5.4.2 SGi Based 163

5.5 CUPS 164

5.6 Edge Computing in 5GS 165

5.7 MEC Deployment in 3GPP 5GS 166

5.8 5GS Common API Framework 168

5.9 MEC Impact on Operational Systems 170

5.10 It is Just the Beginning 173

Chapter 6: Strategic Importance of Network Virtualization 175

6.1 Thinking Differently 176

6.2 A Shift in Revenue Models 177
6.3 A Brief History of Virtualization 178
6.4 The Main Value of Virtualization 180
6.5 Key Technology Options 181
6.5.1 NFVI 182
6.5.1.1 Evaluating NFVIs 183
6.5.2 Virtualized Infrastructure Managers (VIM) 184
6.5.2.1 Virtual Machines (VM) 185
6.5.2.2 Containers 186
6.5.2.3 Software Defined Infrastructure (SDI) 187
6.5.3 Private and Public Cloud Environments 188
6.6 Open Source Projects and Standardization 190
6.6.1 ONAP 190
6.6.2 OSM 192
6.6.3 CORD 192
6.6.4 Standardization 193
6.7 Operational Considerations and Patterns 194
6.7.1 DevOps and NetOps 194
6.7.2 Radically Different Concepts 194
6.7.3 Model and Intent-Driven Orchestration 196
6.7.4 Reactive Versus Proactive Assurance 197
6.7.5 Hybrid Operational Models 198
6.8 Zero Touch – The Vision of Full Automation 199
Chapter 7: Cellular Internet of Things (CIoT) 203
7.1 Business Drivers 205
7.1.1 CIoT Ecosystem 206
7.1.2 Revenue Per Unit 209
7.2 Device Types 210
7.3 Battery Life 211
7.3.1 eDRX 212
7.3.2 PSM 213
7.3.3 Coverage Extension 213
7.4 CIoT Interfaces 215
7.5 CIoT Transfer Methods 217
7.5.1 Cat M1 218
7.5.2 Cat M1 CIoT Using S1-U Data Transfer or User Plane CIoT EPC Optimization [3] 218
7.5.3 NB-IoT Using CIoT EPS Optimization Data Transfer Using NAS PDU and IP-Based Tunnels [3, 5] 222
7.5.3.1 Mobile Originated 222
7.5.3.1.1 LTE Mobile Originated 222
7.5.3.1.2 5G Mobile Originated 225
7.5.3.2 Mobile Terminated 226
7.5.3.2.1 LTE Mobile Terminated 226
7.5.3.2.2 5G Mobile Terminated 231
7.5.4 NB-IoT Using Non-IP Data Communication (NIDD) 235
7.5.4.1 Mobile Originated 235
7.5.4.1.1 LTE Mobile Originated Data Transport 235
7.5.4.1.2 5G NEF Anchored Mobile Originated Data Transport 237
7.5.4.2 Mobile Terminated 238
7.5.4.2.1 LTE Terminated Data Transfer 238
7.5.4.2.2 5G NEF Anchored Mobile Terminated Data Transport Mobile Terminated 239
7.5.5 NB-IoT Switching Between IP-Based Tunnel and Non-IP Data 242
7.5.6 NB-IoT Using SMS 242
7.6 Implications 242
7.7 Management Philosophies 243
7.7.1 Managing Battery Life 243
7.7.2 Signaling Storms and Security 244
7.7.3 Indoor Deep Coverage Device Placement 244
7.8 5G Network Slicing and CIoT SLA Management 245
Chapter 8: 5G Automation and Optimization 249
8.1 Business Drivers for Automation and Optimization 250
8.1.1 Stakeholders in Automation and Optimization 253
8.2 Benefits of Optimization and Automation 254
8.2.1 Delivery of the 5G Service Enablers 254
8.2.2 Enabling Coexistence of Network Slices for Rich Service Ecosystems 254
8.2.3 A Platform for Frictionless Service Creation and a Market in Connectivity 255
8.2.4 Achieving Optimized OPEX to Manage Operational Complexity 256
8.2.5 Reduced or Deferred CAPEX 257
8.2.6 Reduced Energy Consumption 258
8.2.7 Cyber Security and Network Security 258
8.2.8 Meeting of Regulatory Requirements 259
8.2.9 Intent-Driven Optimization 259
8.3 Technology Enablers for Automation and Optimization 260
8.3.1 Disaggregation 260
8.3.2 Network Flexibility and Complexity 261
8.3.3 Network Programmability 261
8.3.4 Virtualization and Service-Based Architecture 262
8.3.5 Artificial Intelligence and Machine Learning 262
8.4 A Closer Look at Automation and Optimization 262
8.4.1 Optimization Timescales 263
8.4.2 Spatial Extent of Optimization 264
8.5 Network Tuning for Optimization 265
8.5.1 Beam Configuration 265
8.5.2 Neighbor Lists 266
8.5.3 Physical Cell Identity 267
8.5.4 Handover Parameters 267
8.5.5 Physical Layer Parameters 267
8.5.6 Layer Management 268

8.5.7 Scheduler Configuration ... 268
8.5.8 Idle and Inactive State Operation and Access Parameters ... 268
8.5.9 Placement and Configuration of Network Functions ... 269
8.5.10 Transport Configuration ... 269
8.6 The Role of Artificial Intelligence and Machine Learning ... 269
8.6.1 Various ML Algorithms ... 271
8.6.2 Models for Wireless Communications ... 273
8.6.3 Analytics to Deliver Value from ML Models ... 274
8.6.4 Applications of Machine Learning in Telecommunications ... 275
8.6.5 Network Slicing and Intent-Based Optimization ... 280
List of Abbreviations ... 281
About the Authors ... 297
Index ... 303

Foreword

We have come a very long way and with 5G at commercial stage, we still have a long way to go! We began this century with on the one hand, the great telecom crash that resulted from a global deregulation of the telecom industry coinciding with the rise and spread of the Internet, and on the other hand, the ascension of the second generation of mobile technology, GSM, which turned out to be a global success and paved the way for each successive generation.

Although GSM provided a comprehensive evolution framework for each generation, our telecom roots are coming from revolutionary explorative work performed in the 19th century. In 1876, Graham Bell was awarded the first US patent for the telephone we are still using today. People are still taking and the need to provide voice services on any G is crucial and will never disappear. In fact, our current wireless systems are grounded in 19th century notions of fundamental limitations of electromechanical propagation that we are still dealing with. 5G is no exception and remains in what is classified as the classical telecom domain discovered and set by Claude Shannon, the father of information theory, at Bell Labs in 1946. Every Shannon law applies to communications systems regardless of the propagation media (e.g., air, copper, fiber…).

In information theory, the Shannon–Hartley theorem tells the maximum rate at which information can be transmitted over a communications channel of a specified bandwidth in the presence of noise. This discovery happened 2 years prior to the creation of the Optical Coating Laboratory (OCLI) and 23 years after the creation in Germany of Wandel and Goltermann, a company that became rapidly famous for their world-class high performing communication analyzers, test sets, radio receivers, amplifiers, signal generators, and meters. Years ago, their company merged with TTC Test Equipment and formed a new global brand: Acterna. In 2000, at the peak of the optical networking wave fueled by the appetite for bandwidth in newly deployed fiber-optic networks, JDS Uniphase Corporation (JDSU) acquired OCLI, and in 2005, Acterna, turned the combined company into one of the world leading manufacturer of electronic test and measurement. As JDSU's goal was to sustain its cutting-edge competitive advantage and stay at the forefront of any telecom developments involving innovative technologies, the acquisition spree never ended. JDSU acquired Agilent Network's tools in 2010, Arieso's location intelligence and RAN analysis for service assurance in 2013, Network Instruments' Performance Monitoring and Diagnostics toolset in 2014, and renamed itself VIAVI Solutions after selling the Lumentum optical products. That way, the company is fully devoted to the development, design and supply of world-class electronic test and measurement equipment for telecommunications. It's worth reminding that before testing either a new telecommunications network or a new generation of mobile communications, VIAVI Solutions has to stay ahead of the curve by participating in all technology discussions to collect the requests from which it will have to design testing solutions that will later be used by vendors and service providers.

As I said at the beginning, we have come a long way; we are 144 years after Graham Bell's telephone patent and with 5G, are getting ready to explore new domains that go way beyond voice. We started 1G with analog telephone, 2G was digital telephone accompanied with text messaging that became a global success and is still heavily used today, 3G brought data to the mix and prepared the ground for mobile broadband, which was fully unleashed by all IP-based 4G LTE networks. Now, the rapid arrival of 5G with the promise of enabling infinite uses cases raises many questions that will be answered as the technology spreads and becomes more ubiquitous. In fact, 5G is nowhere near what we have seen before. It's no longer just about radio communications, it goes beyond and provides a new resilient flexible end-to-end service-based architecture (SBA) that enables quick service turn-up and easy monetization.

Early commercial 5G services remain in the mobile broadband domain because that's what we know the best and as we all know, it's always wise to start launching new services while staying in the comfort zone and introducing new ones on the side. Similarly, many components of the 5G technology are evolution of existing 4G ones. For example, for radio communications, we stay in the orthogonal frequency-division multiplexing (OFDM) domain despite a flurry of new potential modulation scheme candidates, but we added MIMO invented in 2010 at Bell Labs. In the core, most features are coming from the 4G evolved packet core, some like call detail records (CDRs) are eliminated and replaced with more advanced one such as the charging function (CHF) of monetization. Key information technologies (IT) functions including software defined network (SDN) and network function virtualization (NFV) along with edge computing and network slicing have become pillars of the SBA. And lastly, radio access transport networks are also seeing major architecture changes.

As we did with every G, we simplify the architecture: 4G LTE looks far more complex than 3G and 2G. However, with 5G envisioned as the network of the networks bridging many networks (e.g., fixed, wireless, mobile) and sub networks onto a single end-to-end SBA, we have also added some level of complexity that deserves close examination. This book examines with great details all the technologies involved in the 5G SBA, all the way from radio access to core network. It also gives an overview of the evolution from GSM to where we are today and reviews network optimization—another area VIAVI Solutions has deep expertise—as well as the role of artificial intelligence and machine learning in wireless communications.

It is my hope that this book, written by long-time experts and telecom veterans of VIAVI Solutions, will provide some helpful and comprehensive insights for everyone seeking a better understanding of 5G, and will encourage the development of the next G, providing there is one coming!

Stéphane Téral
Technology Fellow, Telecommunications
San Francisco, CA
January 2021

Preface

By the beginning of the nineteenth century, several inventions—such as the printing press and steel processing—had already changed the history of many industries and impacted life in significant ways. This golden era of the "Creative Thinking Age" was really the first technological revolution. We can argue this age started with the Renaissance era, while the accelerated pace of inventions and creativity occurred during the nineteenth century.

In the last half of the nineteenth century, the "Steam Age" marked the start of the second technological revolution. Many historians labeled this era as the first industrial revolution. Steam replaced wind and waterpower as well as horse- and human-power to produce smaller machines, which in turn improved productivity.

At the dawn of the twentieth century, the third technological revolution was enabled by the "Electricity Age." Many considered this era to be the second industrial revolution made possible by electric factories with thousands of production lines.

The end of the Second World War marked the beginning of the fourth technological revolution, the "Mass Production Age." Assembly lines and the exponential scaling of product manufacturing created the third industrial revolution. Many innovations during that era, such as mechanical robots and machine presses, impacted the pace of technological advancements.

Toward the end of the twentieth century, we witnessed the start of the fifth technological revolution, the "Information Age." This is the era we are still living in at the time of writing of this book. The technology of the "Information Age" continues to have significant industrial and commercial impacts, as well as a major impact on our day-to-day lives. Telecommunications and the Internet played a major role in shaping this era. The Information Age started in the 1950s with the first introduction of computer networks. In the 1960s, the US Department of Defense initiated the ARPANET project, marking the real beginning of the Internet concept. Then, in the 1970s, with the invention of TCP/IP (Internet Protocol), packet switched networks took off. In the 1980s, the World Wide Web was introduced, using the invention of hypertext documents that allowed

Internet sites to link to each other; information became far more accessible to the masses and kept growing, with no sign of slowing. This digital transformation, along with the introduction of several communication innovations and devices, created a new generation of knowledge-based society and fueled the growth of global economic markets led mainly by technology-related industries.

With this never-ending growth in information traffic—including voice, data, video, and the mass proliferation of business and social applications—the world is about to witness the sixth technological revolution, the "Artificial Intelligence Age." Precursors of this sixth technological revolution are the introduction of 5G and Internet of Things (IoT) technologies and the widespread use of machine learning and data sciences. Some believe this era is still twenty-five to thirty years out. However, recent advancements in IoT technologies, the maturation of virtualization and automation systems, and acceleration in the adoption of and plans for 5G indicate that we are closer to the onset of this sixth revolution. The term "Artificial Intelligence" or AI, is not new; it was coined in the late 1950s. However, the continuous year-over-year improvements in computational power, following Moore's law, and the recent explosion in all sorts of data collection and gathering (Big Data), are leading to the need for AI to assist and eventually take over manual decision-making processes. For mission-critical IoT, such as a connected car or surgical robot, AI will play an important role in replacing slow, error-prone, human decision-making with higher-confidence, real-time, and precisely automated decision-making processes.

For AI to scale and be effective everywhere and all the time, the data volumes and speeds needed are more than 100 times current network capabilities. Furthermore, communication latencies must reach a record low since some of these AI decisions will need to occur in fractions of milliseconds (think about a connected car avoiding hazardous conditions down the road). That is where 5G comes into play with a completely new architecture designed from the get-go to satisfy these two essential requirements, for the sixth technological revolution era to start.

For more than four decades, the first four generations of wireless technologies steadily led the expansion of access to information and life-impacting applications and services for billions of people. The first generation (1G) and second generation (2G) of wireless technologies spanned the last twenty years of the twentieth century and were both mainly designed for voice calls and limited data and messaging capabilities.

At the beginning of the twenty-first century, the third generation (3G and 3.5G) was designed to provide mobile broadband (mobile Internet) access to millions of mobile devices (smartphones, laptops, hot spots, etc.). Speeds upwards of several megabits per second (Mbps) were achievable; mobile video downloads and streaming were taking off.

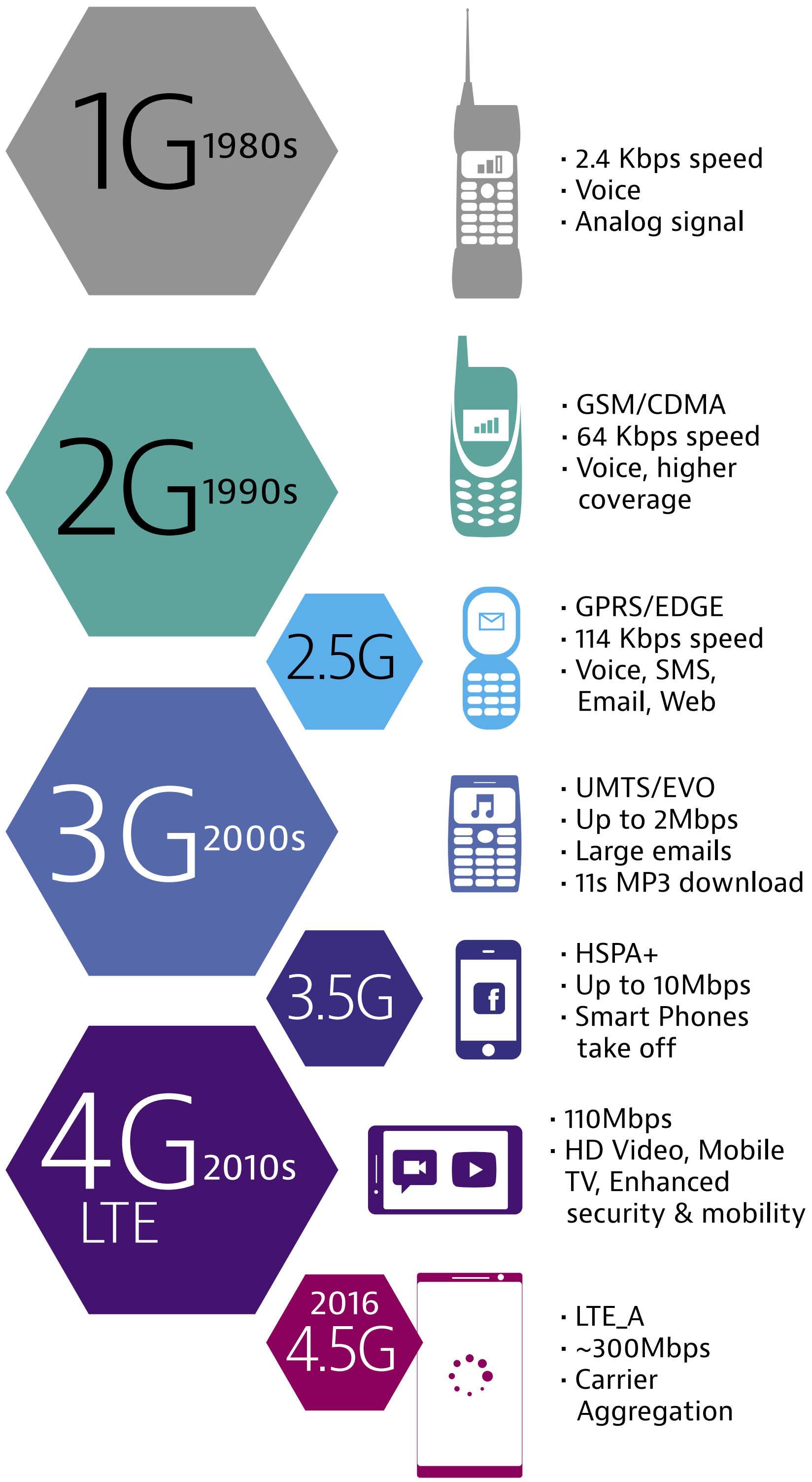


Figure I.1 The road to 5G: Wireless technology evolution from 1980's to 2017

The first fourth generation wireless technology (4G, also commonly referred to as Long Term Evolution or LTE) was commercially available in 2009 in Europe and in 2011 in the US. LTE provided speeds up to 110 Mbps, and with the introduction of LTE Advanced (LTE-A or 4.5G) speeds of 300 Mbps are achievable. It is worth noting that the deployment and adoption of LTE was one of the fastest in telecommunications history. It took only five years for LTE to reach 2.5 billion subscribers worldwide compared to ten years for 3G to reach the same number. We predict that 5G will take less than three years to reach 2.5 billion subscribers and 20 billion IoT devices connected.

In November 2017, the International Telecommunication Union (ITU) introduced the International Mobile Telecommunication system for 2020 and beyond, IMT-2020, standard (1), defining the first requirement for 5G networks. As a first step towards satisfying the IMT-2020 requirements, 3GPP (3rd Generation Partnership Project) introduced the 5G New Radio (NR) standard (2). While that date marks the official start of the 5G era, Verizon had worked on a pre-standard version of 5G prior to that date—the 5G Technical Forum (5GTF) (3)—and announced limited commercial availability in 2018 for fixed wireless networks in a few US cities. Also, KT worked with a variant of 5GTF, called 5G-SIG (Special Interest Group) (4) that was used during the PyeongChang 2018 Winter Olympics.

According to IMT-2020, there are three categories that define 5G. The first is enhanced Mobile Broadband (eMBB). This is a natural expansion to current LTE-A capabilities. The demand for more and more data bandwidths and speed is not wavering, and with new applications like 4K TV, Virtual Reality (VR), and Augmented Reality (AR), the trend will continue upward. The second category is Ultra-Reliable Low Latency Communications (URLLC). This is where the connected car and ambitious autonomous driving requirements are pushing. The third category is massive Machine Type Communications (mMTC), expanding the densification needed for IoT implementations.

While 3GPP and ITU focused on 5G NR, the 5G revolution involves all aspects of wireless and wireline communication networks. To understand this, we need to look at the two basic 5G promises: higher data speeds and low latency. With the current architecture of the telecommunications networks, these requirements are inversely correlated: to achieve one, you need to relax the other. For a network to achieve both simultaneously, this network has to be sliced, i.e., virtually programmed and disaggregated, at every segment of the network service chain, to deliver these two characteristics without the need to manually reconfigure or redeploy any physical connection or element.

Network slicing is possible and has been demonstrated, but to have it at the scale required for the new 5G use cases, the 5G network architecture, at all segments, must be different from all previous generations.

Table of Contents

To better understand this architecture revolution in telecommunications, let's draw a parallel with the history of computing and programming methodologies. The first digital computer, invented in the 1940s, used punched cards for all inputs and outputs and could only perform one computing task: solving large systems of simultaneous equations. The processing elements and memory were all hardwired to perform this one computing task. We can see the parallel here with early telephone systems that used hardwired connections to carry voice calls between finite numbers of phones.

In the 1950s, ENIAC (Electronic Numerical Integrator and Calculator)—a huge machine with many vacuum tubes and large power consumption—was born. In concept, this was the first general purpose computer that could be "programmed" to perform different calculations through the laborious act of reconnecting cables and switches. It is easy to see the parallel here with the first circuit-switched telephony systems.

With transistors replacing vacuum tubes and the invention of Integrated Circuits (IC) or microchips, smaller computers (called mainframes at that time) became possible. Programming using a computer language enabled the use of thousands of states within the microchip to perform a "programmed" sequence of computation instructions. In the 1970s, Intel integrated the first microprocessor and the Central Processing Unit (CPU) into a single microchip. This was a true general-purpose computer; using linear programming languages like Basic, COBOL, Pascal, and C, one could write sophisticated applications. However, these programs still needed to have well-defined input and output parameters, and any change in the format or number of these input and output parameters required rewriting and re-compilation of the code. There was no separation between the data and code in these programming methods.

We compare this to the IP-based wireline and wireless network, shown in Figure I.2, with IP routers and switches that can be deployed and configured to switch and route IP packets carrying different kinds of predefined services (voice, email, video, text, et al.). In essence, the evolution from 1G to 2G to 3G and 4G was all about the packetization of the mobile network. Still, control plane (signaling required to connect the network) and data plane (the packets containing the media of the service) were centralized together and each element of the network required a specialized hardware component to deal with the control plane protocols and the data related to the types of services that would be packetized. Every time a completely new type of service had to be added, manual reconfiguration and in many cases a forklift of the network elements was required.

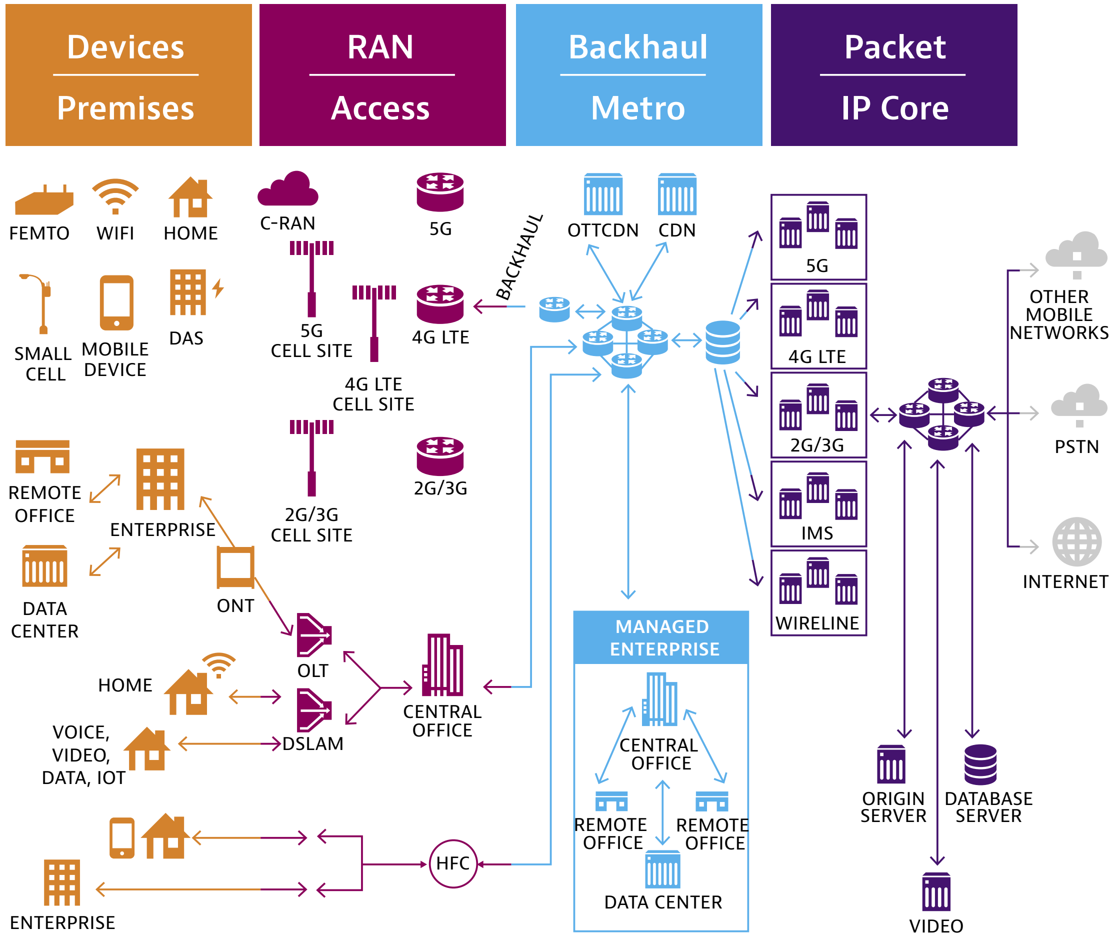


Figure I.2 Current packet switched networks

With the age of object-oriented and functional programming languages (e.g. C++, Java, C#, Lisp, Scala), the separation of code and data was achieved, and code behavior is virtually defined by the data it carries or serves. This is exactly what the 5G architecture is designed to achieve, as shown in Figure I.3, a cloud-native, fully virtualized network, orchestrated with complete disaggregation of control and data planes, and programmable in real time to deliver multitudes of network slices with corresponding characteristics.

In this book we will explore the new 5G revolutionary architecture and describe how each segment of the network is redesigned in 5G to provide the promised characteristics and the new use cases and applications that define the sixth technological revolution era.

The first chapter will go through the elements of this architecture revolution compared to previous generations. It will also describe the independence of the Radio Access Network (RAN) from the Core Network (CN), and Control and User Plane Separation (CUPS) in 5G. Details about RAN disaggregation and open interfaces will be explained, along with how 5G will enable network slicing using virtualization and orchestration.

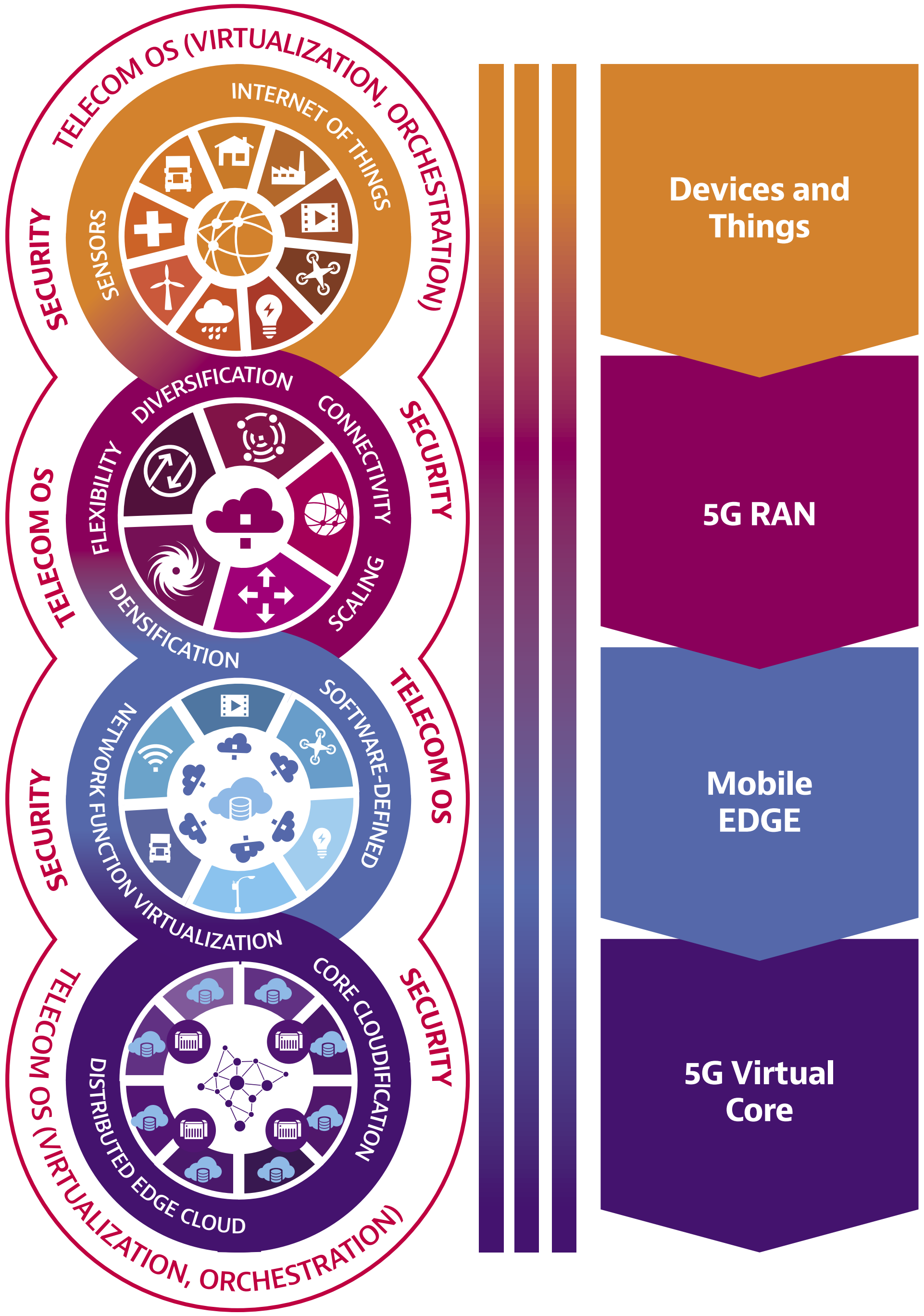


Figure I.3 Futuristic cloud-native slicing-enabled and programmable 5G network

The second chapter will cover the 5G New Radio (5G NR) specifications and characteristics, detailing the new 5G frequency bands, modulations, waveforms, numerologies, and frame structures. A major element of the 5G NR is the Massive MIMO antenna. Chapter 2 will describe the MMIMO elements and introduce the new 3D beamforming technology that characterizes 5G.

Chapter 3 will describe the evolution of the 5G Radio Access Network (RAN) and its related new transport networks. It will start with a brief history of RAN transport networks, then dive into the 5G transport networks with all their different functional split options. It will also touch on the Timing Sensitive Network (TSN), which is essential to achieving the URLLC requirements of 5G.

Chapter 4 will examine the new 5G virtualized core and draw the difference between it and the LTE Enhanced Packet Core (EPC). More about how 5G core is designed for network slicing is discussed in this chapter.

Another element of the new wave of applications that will be enabled by the 5G architecture is the Multi-Access Edge Computing (MEC) platform which will be discussed in Chapter 5.

Virtualization is key to 5G network slicing. There are a lot of lessons learned from the last decade when operators embarked on the virtualization journey. Chapter 6 will provide insight into the world of virtualization as it relates to 5G.

As mMTC is the third category of the 5G revolution, Chapter 7 will go into the relationship between IoT technologies and 5G.

Finally, as we prepare for the AI era, the topics of machine learning in optimization and automation will be explored in Chapter 8.

We focused this book on topics that will initially interest the operators, network equipment manufacturers, and network technologists who will be involved in developing, testing, and deploying the 5G related network and system elements. We try to give an overall view of the operational aspects of 5G and the impact they will have on reliability, performance, scale, and optimization.

One topic that will become very significant in the success of 5G deployments is security. With cloud computing and many cloud service providers delivering real-time services and mission critical applications, the security aspect of these services has been in the headlines recently. Security has always been a concern with all wireline and wireless operators, ever since the inception of telecommunication systems. Limited physical access to facilities, very strictly enforced digital access, and continuous monitoring of cyberattacks or efforts to compromise the network are different mechanisms that operators have used to ensure the security of their networks. However, with the transition to virtualization and cloud-native components of the network, many challenges to current security safeguards are present.

Since the topic of security and how it will be impacted by 5G requires a complete book to discuss it in greater detail, we decided not to include it in this book.

Table of Contents

This book was designed to span many domains in the 5G ecosystem so that the rationale for the inflections in the different domains can be understood in the context of the whole system. As such it can be a resource for those seeking a more detailed overview of 5G as well as experts wanting to understand why developments in their domain of expertise are made and how they interact with other domains.

CHAPTER ONE

The 5G Evolution Story

Table of Contents

It's All About the Architecture

Cellular mobile radio is an incredible story of exponential growth in subscriber numbers and data volume. As shown in Figure 1.1, subscriber growth slowed as saturation of the addressable part of the global population was approached around 2010, then accelerated again as new drivers for subscriptions appeared. These new drivers included multiple devices per individual and machine type communications.

However, the revenue for operators has not kept up with the growth in subscriptions and data volume. A significant reason why the exponential growth in data has not been accompanied by a similar growth in revenue for the operators is that the advent of smartphones, and their supporting ecosystem of apps, facilitated the emergence of the Over The Top (OTT) Players like Google, Netflix, YouTube, Facebook, Apple, etc., and the economic power shifted away from the operators. This revolution in market behavior was supported by the availability of open platforms to cultivate innovation in the application space that led to an explosion of applications for a huge range of functions from the frivolous to the timesaving. This was further shored up by the introduction of the open source Android platform that made smartphones available for all budgets. This is a great topic of discussion and debate; however, this book is not the place to posit how this situation arose, but we will look at its consequences and how it leads to the need for a new game changer in both the mobile network RAN (Radio Access Network) and core architectures, and in the monetization aspects of such architecture.

To understand why 5G will be that game changer, we need to go back in time and review the evolution of the mobile network architectures defining its radio, access, and core elements and functions.

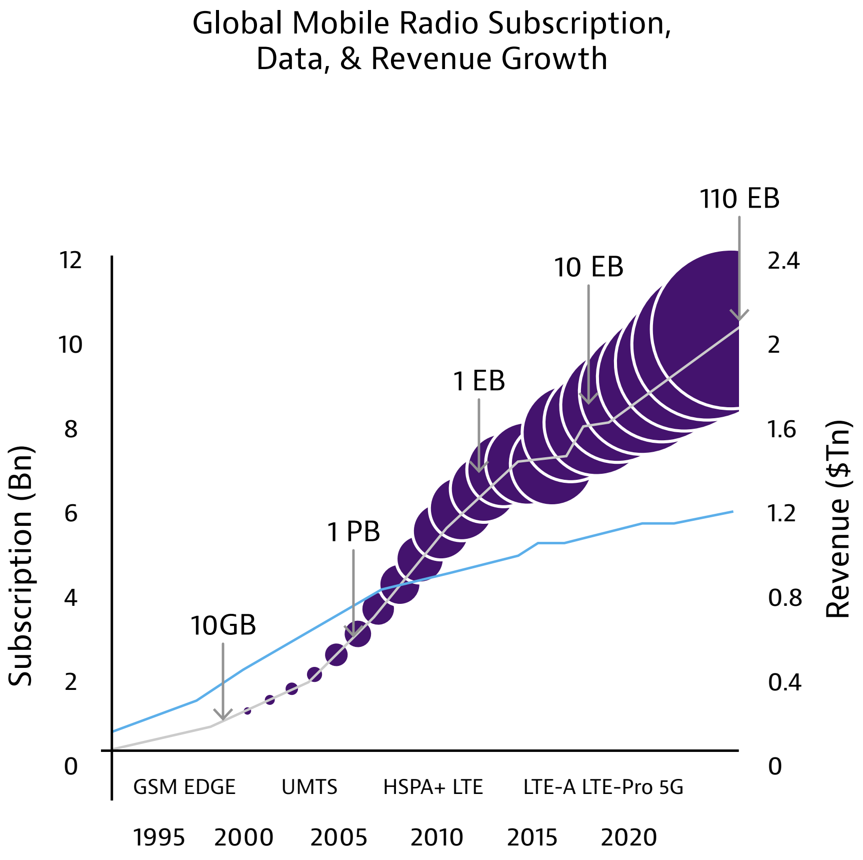


Figure 1.1 Global mobile radio subscribers, data and revenue growth

1.1 Naming of Parts

However, before we embark on our journey into the past, this section introduces some terminology that can be skipped or revisited at the reader's leisure. Telecommunication aficionados love acronyms and, sometimes, like an acronym so much that they use it for multiple purposes. For example, Inter-Operability-Testing was known exclusively as "IOT" until the Internet of Things "IoT" became a thing. Thus, naming the parts may be helpful in talking about the architectural revolution that 5G is introducing.

Figure 1.2 is a 10,000-foot view of the architecture of a cellular mobile network, comprising core network (CN), radio access network (RAN), and the user equipment (UE).

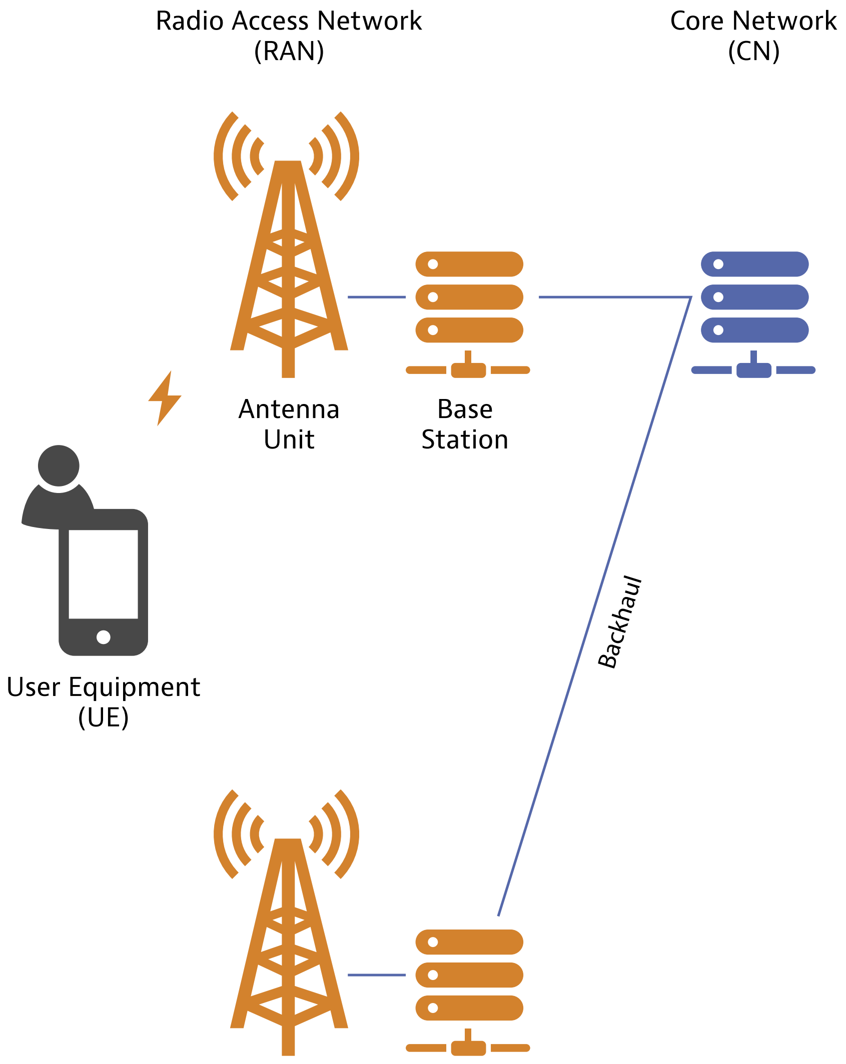


Figure 1.2 Simple mobile network architecture

The data exchanged between the elements in a communication system and the way in which these are interpreted are precisely defined by standardized protocol layers to ensure correct operation and interoperability between vendors. Figure 1.3 provides a protocol view of the architecture that is more or less generic across all generations of the cellular network, albeit the actual processing carried out along with the structure of the radio link between the antenna unit and the user equipment (which is known as the air interface) may be markedly different. Regarding the CN, however, the protocol architecture has undergone more significant changes, so the Figure only shows a generic CN; each major CN element is described below as it is introduced with its associated generation.

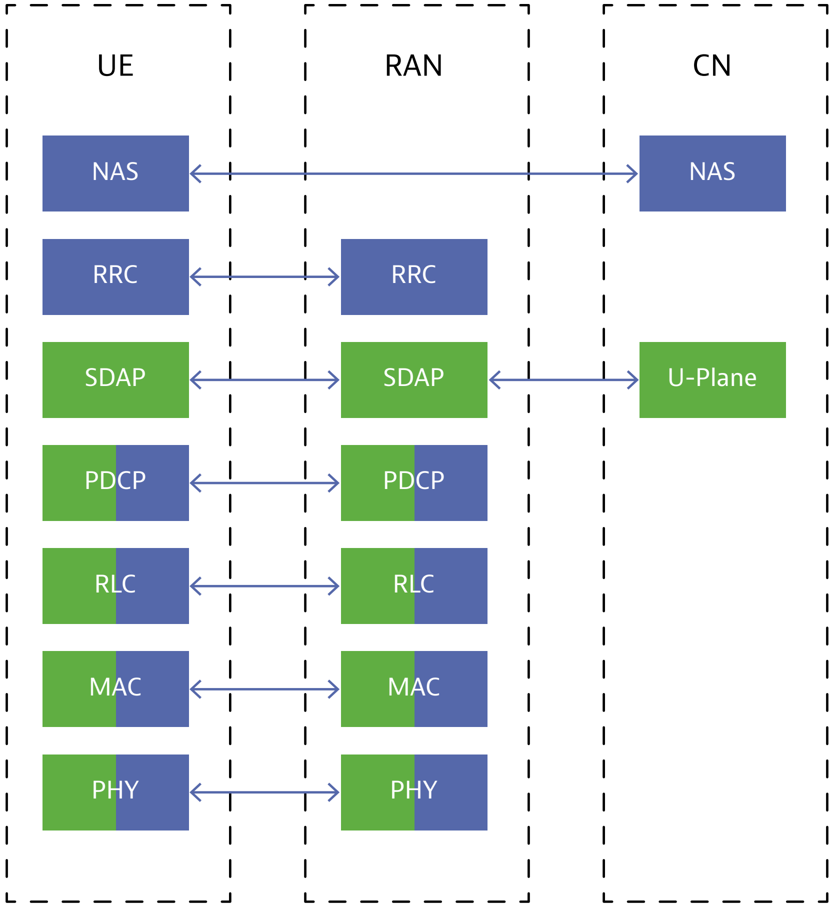


Figure 1.3 Protocol layers traversing the base station

The principal services provided by each layer are briefly introduced. The control plane (CP) sets up and tears down connections; protocol layers that handle CP are highlighted in blue. The user plane (UP) exchanges data between the UE and the CN; protocol layers that handle UP are highlighted in green.

The non-access-stratum (NAS) manages direct signaling between the UE and the CN to establish and maintain communication sessions with the UE as it moves through the network.

Radio resource control (RRC) manages the broadcast of system information: contacting UEs (paging); establishment, modification, and release of active RRC connections; handover between cells; selection of cells when not connected along with measurement; and reporting of the strengths at which the transmissions from different cells are received at the UE.

Service data adaptation protocol (SDAP), new for 5G, manages mapping of quality-of-service (QoS) flows to radio bearers, and provides QoS marking on data packets in the RAN so that packets can be prioritized appropriately. The SDAP communicates with the U-Plane entity in the CN.

Packet data convergence protocol (PDCP) manages ciphering, packet header compression, and sequence numbering.

Radio link control (RLC) manages packet segmentation and error correction with automatic repeat reQuest (ARQ). Prior to 5G, this layer also performed packet concatenation and reordering to maximize utilization of the air interface at the expense of increased latency.

Medium access control (MAC) manages multiplexing data from different logical channels into/from transport blocks for delivery on the radio physical layer, and error correction through Hybrid-ARQ (HARQ).

Physical (PHY) functions are dependent on the air interface, but generically include rate matching, modulation, resource element (RE) mapping, and mapping to antennas.

1.2 Once Upon a Time

If we were to travel four decades into the past, we would find that the drivers for the first-generation mobile network were quite simple: the need for good quality analog voice services anywhere at any time. However, with each subsequent generation, demand for the service grew, and mobile users wanted to send and read emails on the go. These new drivers led to major innovations, which in turn required complexity in the architecture to adapt in response to the never-ending mobile broadband usage expansions and user expectations.

To add another wrinkle, these new-generation architectures must carry the burden of supporting old generations operating simultaneously in the field.

1.2.1 The 1980s and Early 1990s: Here Cometh “GSM”

The GSM (global system for mobile communications) was a digital system using Time Division Multiple Access (TDMA) to share radio resources. It was developed primarily in Europe starting in the mid-1980s, replacing the existing “first generation” analog mobile radio systems, and hence is referred to as a second-generation system, “2G”. It became such a global success story after its initial roll-out in 1992 that the term GSM is now almost synonymous with 2G. There are other 2G systems, such as code division multiple access (CDMA) IS-95 in the US and personal digital cellular (PDC) in Japan. Like the VHS and Betamax conundrum, both CDMA and GSM were fighting for global dominance; however, this is all history and it doesn’t really matter which technology was better. Let us stick here with 2G (aka GSM) as it is relevant to our story of 5G evolution.

The drivers for 2G were initially to provide a voice service while introducing a moderate data service that could replace the legacy 1G systems. Another driver was to introduce interoperability between mobiles and network elements produced by different vendors.

Elements of the GSM network RAN include the base transceiver station (BTS) which contains the radio equipment needed to serve each cell in the network. A group of BTSs are controlled by a base station controller (BSC) that manages all the radio-related functions of a GSM such as handover, radio channel assignment, and the collection of cell configuration data. The GSM CN initially included the mobile switching center (MSC) that controls a number of BSCs, performs the telephony switching functions between mobile networks, and connects to the Public Switched Telephone Network (PSTN); see Figure 1.4.

GSM was based on a circuit-switched technology that later had to be re-engineered to provide a packet-based data approach, the so-called General Packet Radio Services (GPRS). GPRS introduced new CN network nodes to enable the transport of data and connecting the mobile network to the Internet. Serving GPRS support node (SGSN) and gateway GPRS support node (GGSN) formed what is called the GPRS CN. A high-speed circuit switched data (HSCSD) was a better fit to the existing GSM architecture. However, the market required a packet switched solution to efficiently support simultaneous data and voice services. The changes addressed the MAC, to allow air interface resources to be allocated in small chunks with successful delivery managed by HARQ, and the CN, to allow the connection to a packet data user to be dealt with as a series of temporary block flows (TBF) rather than as a continuous connection. The system was later improved with enhanced data rates for GSM evolution (EDGE) that introduced 8-phase shift keying (8PSK) modulation rate to the air interface. It can be argued that EDGE was the innovation that created the environment in which smartphones could evolve.

In fact, IBM introduced what is considered the first concept of a “smart” phone product called Simon Personal Communicator in 1992. We also know the story of the Blackberry phone that stormed the marked in the late 1990s to the extent that it seemed everyone who was anyone in business had one.

As GPRS was introduced and deployed, the network elements at that time were not dimensioned to support a mobility management scenario where all GPRS-capable mobiles on the system attached to the GPRS CN even though actual usage of GPRS data was low. A further problem with GPRS related to managing the latency for data services that could, for example, enhance user experience when browsing the web. Latency could be minimized by extending the duration of the TBF in case new data arrived in close succession. However, this was at the expense of capacity for other users to access the system, along with increased battery consumption.

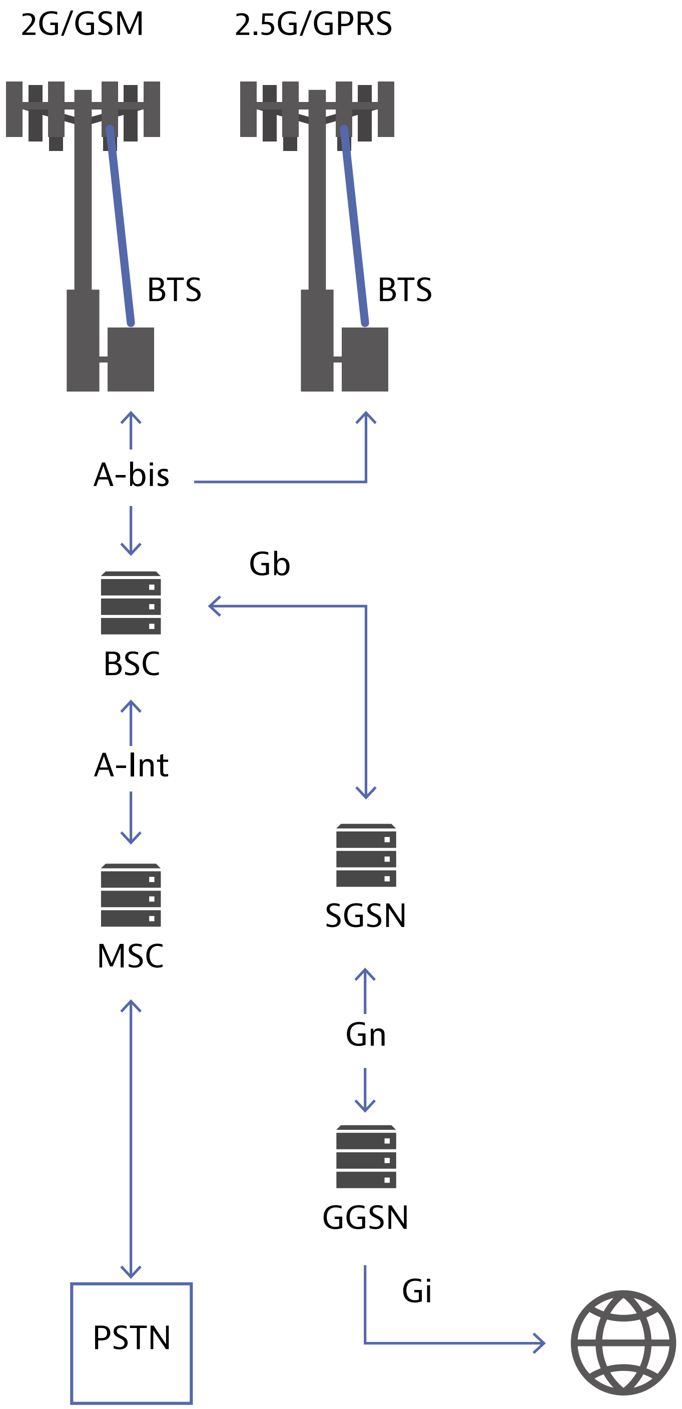


Figure 1.4 Simplified view of the 2G and 2.5G architecture

It is also interesting to note that the initial GSM system was delivered with a control channel mechanism to deliver short text messages over the air: the Short Message Service (SMS). Originally envisioned as a diagnostic tool, it became a mass-market success delivering messages in the billions and shaping a new generation of texters and shorthand languages (LOL).

Other features were added to the GSM standard during its life of active development and adoption into the 3rd Generation Partnership Project (3GPP). For example, capacity was increased by applying cell-splitting to go from omni-directional cells to sectored cells, and concentric cells, then to micro-cells in street canyons and pico-cells inside buildings. Frequency hopping was also used to introduce so-called fractional reuse where a "classical" frequency reuse was applied to the frequency layer that carries the broadcast channel (BCCH); frequency reuse of one was applied to the other frequency channels, but with a restricted occupancy of the channels.

We will discuss in a later chapter the emergence of machine type communications (MTC), but it is worth noting here that the need for authentication with point-of-sales devices, an early type of MTC service, was discovered to be well-suited to GSM/GPRS (sometimes called 2.5G) due to its ubiquity and low cost.

Such uses fed into the creation of the overarching concept of the Internet of Things (IoT). It was apparent that the requirements for IoT/machine type communications (MTC) differed from those of cell-phone subscribers. There were more stringent requirements for lightweight control signaling, to allow many more IoT devices than cell-phone subscribers to be connected in a given area; for very low power consumption, to support sensor-type use cases with no access to external power; for increased tolerance to path-loss, to support deployment in difficult radio propagation situations such as basements. On the other hand, requirements for latency were generally greatly relaxed.

Updates were made to GSM to better support IoT that sought to reduce control channel overhead and to ration random access channel (RACH) usage in the uplink for MTC devices, but the scope to retrofit them was limited and they were largely not taken up by the market, which retained use of legacy devices. Better technical solutions were available with other standards; nonetheless, GSM-based IoT solutions are available in the market at the time of this writing.

Other features such as the previously mentioned high-speed circuit-switched data (HSCSD), along with voice group call service (VGCS) and multimedia broadcast multicast service (MBMS), were added to the standard, but not widely adopted.

Other 2G systems that are relevant when considering 5G system requirements are those directed at specialized applications. For example, the TETRA system provided a group push-to-talk (PTT) service with a very fast (sub-second) call setup time, direct communication between UEs that bypasses the infrastructure, and high-power mobiles, which are required to support emergency services applications. The TETRA system did not keep pace with the changing needs of its users and the requirements were merged into long-term evolution (LTE), which is now in the process of replacing these legacy systems in some territories around the world.

1.2.2 Late 1990s to Early 2000s: The Almighty UMTS

The universal mobile telecommunication system (UMTS) was a 3rd-generation system developed across Europe, the US, and Japan starting in the mid-1990s to replace the existing 2G system. Again, as the Betamax vs. VHS story continued, there were other 3G systems, such as CDMA2000 in the US and TD-SCDMA in China. We will focus here on UMTS as the proxy for 3G.

Based on a wideband CDMA (WCDMA) radio physical layer, the UMTS Terrestrial Radio Access Network (UTRAN) was composed of a new element, the Node B, that replaced the BTS elements in the GSM/GPRS network. The Node B streamlined the radio-function interaction with the mobile devices. These Node Bs are controlled by the radio network controller (RNC) that carries out radio resource and mobility management functions. The UMTS CN is similar to that for GSM where the RNC switches the data services plane through the SGSN and GGSN CN subnetwork and the voice services through the MSC CN subnetwork; see Figure 1.5.

The driver for 3G at the design stage was quite simply extra voice capacity to support the growth in subscriber numbers, combined with an ability to support higher data rates for both Circuit Switched (CS) and Packet Switched (PS) domain services.

Additionally, there was a perception that GSM had very inflexible service definitions and consequently there was a desire to introduce a very flexible methodology resulting in the Radio Access Bearer (RAB)/Signaling Radio Bearer (SRB).

Finally, an additional driver was to minimize the latency for call setup at least for some classes of service. Flow establishment latency badly affects subscriber experience with web browsing. As described above, the solution adopted for this in GSM, which was to extend TBF duration, was inefficient. In contrast, UMTS introduced three additional connected mode states, intermediate between fully idle or fully connected state, which provided a more efficient, albeit complex, solution.

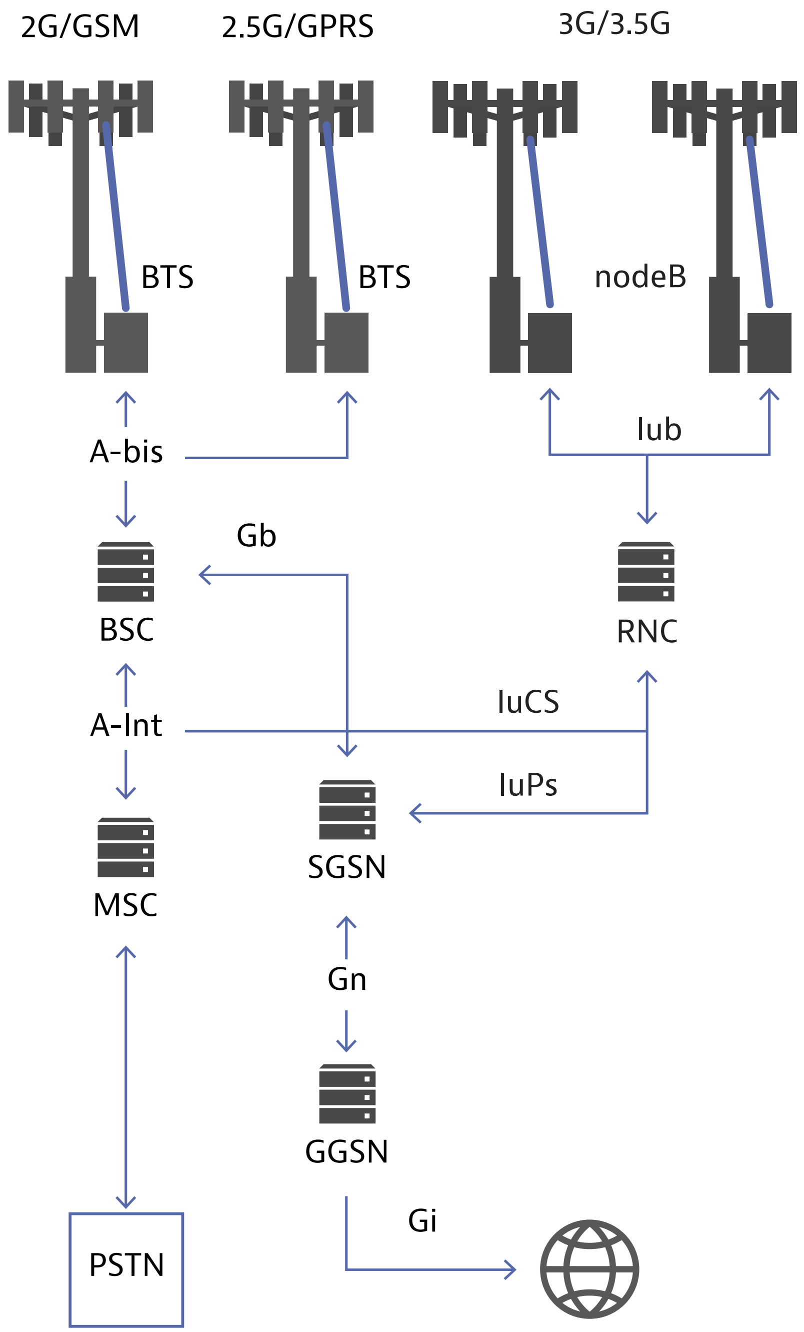


Figure 1.5 Simplified view of 3G and 3.5G architecture

The fully connected state is known as CELL_DCCH for the dedicated control channel (DCCH) on which resources for a given UE's connections are reserved for a period. The intermediate states are CELL_FACH, CELL_PCH, and URA_PCH. CELL_FACH, which work by keeping the network updated about which cell the UE is in and supporting limited data transfer on the shared forward access channel (FACH). Similarly, the CELL_PCH state also updates the network about which cell the UE is in, but it uses the paging channel

(PCH) rather than the FACH for downlink data. Finally, URA_PCH [1] also uses the PCH for data, but it differs from CELL_PCH in that it updates the network about the UE's location with the granularity of a UMTS routing area. These states allow a signaling connection to be maintained to speed setup of a dedicated connection. Consequently, they can limit congestion on control channels for social media and messaging type applications where many users simultaneously attach to the system and exchange small volumes of data with frequent updates. However, the design stage did not foresee the "smart-phone revolution" which created demand for "always-on" type applications with low average rates, but occasional high peak data rate with low latency. The UMTS system had several shortcomings in supporting this use case and was re-engineered to address it. Particularly, a fast shared-access channel was added, and the WCDMA approach was made more TDMA-like, by time-slicing allocations to higher modulation channels with lower spreading factor, to address the peak-user rate limitation by exclusively allocating resources to a few mobiles. Additionally, while the 10ms frame based HARQ period was maintained, an optional short 2ms HARQ period was introduced to reduce latency.

These enhancements started with high-speed downlink packet access (HSDPA) in 3GPP Rel-5 and high-speed uplink packet access (HSUPA) in Rel-6. These enhancements ushered in the 3.5G era. Later upgrades to these features known collectively as HSPA+ (in Rel-7 and beyond) were characterized by the best-in-class vendors being able to make them available as software upgrades.

One of the distinguishing features of the WCDMA-based UMTS system that was deprecated with 3.5G was soft handover. This feature creates a robust radio channel by enabling a mobile to simultaneously connect with transceivers from widely separated base stations, considerably suppressing the effects of multi-path propagation and shadow fading. This provided more resilient coverage for voice and circuit switched data services. However, it consumed extra network and air interface resources. In HSPA it was replaced with fast handover to maintain connection to the current best local base station. However, soft handover was retained for the shared control channel. Later, fractional DPCH was introduced to conserve/share downlink code tree resources between a larger number of connected users.

Other features were added to the UMTS standard during its life of active development and adoption into the 3GPP. Updates were made to UMTS to better meet the requirements for IOT/MTC, but the scope to retrofit them was limited. Those adaptations did not address the issue of mobile complexity and power consumption reduction, and better technical solutions were available with other standards. Nonetheless, UMTS-based IoT solutions are available in the market at the time of this writing.

1.2.3 The 2010s: The Rise of LTE

As mobile operators were busy expanding their 3G and 3.5G networks, a competing technology was rising and in 2010 made a huge debut in the communications market. This technology was WIMAX (worldwide interoperability for microwave access), based on the advancement of IEEE 802.16 standard and refreshed from earlier versions of the standard to satisfy the requirements for a fourth generation (4G) system defined by the International Telecommunication Union (ITU). WIMAX promised to replace the last mile of the communications link for both residential and enterprise broadband and provide speeds up to 100 megabits per second.

This WIMAX frenzy pushed the acceleration of the emergence of the 3GPP 4th-generation network and, in particular, the LTE (Long Term Evolution) standard. Also, the expansion of GSM and the need to support standardized worldwide roaming was one motivation for carriers like Verizon, who had bet on CDMA during the 3G phase, to quickly adopt the LTE standard. The fact that it took less than ten years from 3G to have the LTE standard ready for implementation was an indication of how smartphone take-up and higher data rates needed for the new apps frenzy exceeded initial planning and market predictions.

The central driver for LTE was to achieve higher data rates in the packet domain and simplify the network by eliminating the need to support a separate circuit-switched connectivity for voice. The LTE advantage was the creation of a "flatter" network architecture without the need for a base station controller (BSC/RNC). This was also accelerated by increases in microprocessor processing capability and cheaper RAM. So, what used to be a system of Node Bs and RNCs was repackaged into a system of single elements, the evolved Node Bs (eNodeB). This simplification also reduced the number of different states of the User Equipment (UE) activation that had been introduced with UMTS, which was to have reduced data latency. A totally new CN, the Evolved Packet Core (EPC), was introduced to support LTE. This included the Mobility Management Entity (MME), the Serving Gateway (S-GW), and the Packet Gateway (P-GW); see Figure 1.6. Other key elements of the EPC were the Home Subscriber Server (HSS) to manage subscriptions, Policy and Charging Rules Function (PCRF), and Authentication, Authorization, and Accounting (AAA) server for security.

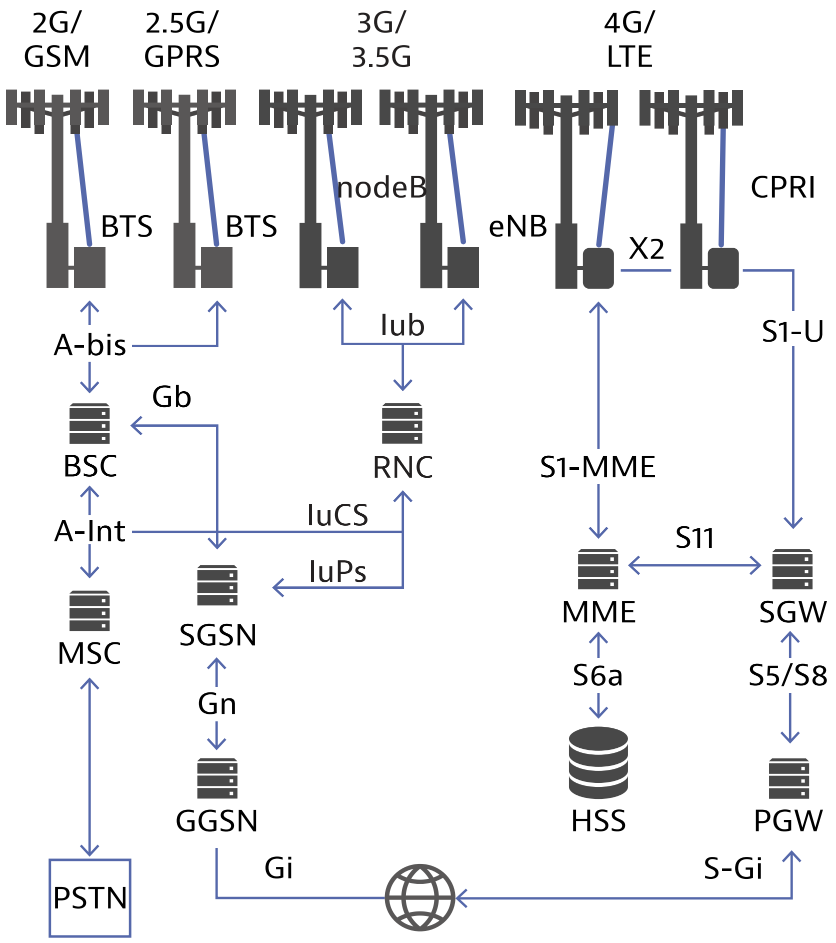


Figure 1.6 Simplified view of the 4G architecture

While LTE Rel-8 supported voice over IP [2], an acceptable solution to provide handover to legacy CS voice on 2G or 3G was not available when LTE was first deployed. This arose because the EPC did not support the CS domain and PS to CS handover was not available. Initially networks had to rely on circuit-switched fallback (CSFB). Consequentially, Voice Over LTE (VoLTE) deployment was delayed until after 2012 when the Single-Radio Voice Call Continuity (SRVCC) feature based on an upgraded core network became sufficiently available. To support this feature, the EPC CN is supplemented by the IP Multimedia System (IMS) CN to provide an "anchor" to support the signaling to permit seamless handover between PS domain and CS domain. The IMS CN includes the Call Session Control Function (CSCF), the Subscriber Location Function (SLF), Breakout Gateway Control Function (BGCF), Media Gateway Control Function (MGCF), and Media Gateway (MGW).

To make a more straightforward distinction from a marketing perspective, the LTE Rel-10 was dubbed "LTE-Advanced." This increased peak data rates through the introduction of carrier aggregation of up to five carriers (100MHz total band-width) and enhancement of multi-antenna techniques (8x8 MIMO downlink and 4x4 MIMO uplink). This is the LTE release that was submitted to the ITU to meet the international mobile telecommunication (IMT)-Advanced requirements [6]. The LTE air interface, which will be introduced fully in Chapter 2, sends multiple data transmissions in parallel using Orthogonal Frequency Division Multiplexing (OFDM), and delivers data to and from multiple users at the same time using Orthogonal Frequency Division Multiple Access (OFDMA). This scheme results in a multitude of discrete units of communication which can be allocated flexibly. The ability to allocate resources in a very granular way on the LTE air interface also meant that adaptations, as well as addressing the possible control channel congestion for delay tolerant devices in a similar manner to that done for UMTS, could also address the enablement of low complexity and low power MTC devices. In particular, the Narrowband Internet of Things (NB-IoT) feature, Rel-13, allowed devices to use a single OFDM subcarrier, simplifying implementation, and to use significant repetition coding, increasing the path-loss resilience by 20dB, and enabling low-power and in-building devices.

LTE Advanced Pro (LTE-A, also known as 4.5G) was introduced with 3GPP releases 13 and 14, and in 2016 was considered a precursor to the 5G evolution. 4.5G significantly increased the data speeds and bandwidth available. This was achieved using a number of different technologies, including carrier aggregation, which increased the number of simultaneous carriers supported from five to thirty-two; license assisted access (LAA) to include carriers in the unlicensed spectrum by using a listen-before-talk (LBT) mechanism to enable co-existence with existing users of the spectrum; advances in antenna systems with full dimension (massive) MIMO increasing supported antennas from sixteen to sixty-four to support two-dimensional beamforming; and higher order modulation up to 256 quadrature amplitude modulation (256QAM) [3].

The geographical distribution of traffic demand across a mobile network is very uneven. The coverage area of macro cells is large enough to provide spatial averaging. That is, one macro cell in a given location will generally experience a similar traffic demand to its neighbors. As cells get smaller, this no longer holds, and some cells will experience high traffic and adjacent ones considerably less. This sets the minimum useful size of a macro cell inter-site distance at about 200m. To further "densify" the network to provide extra capacity in "hot-spots" and in-fill coverage in localized "not-spots," LTE small cells and heterogeneous network (HETNET) access were introduced. This development led to an exponential increase in the number of RRHs (remote radio heads); it was not economical to attach a single eNodeB to each RRH. The centralized RAN (C-RAN, sometimes called cloud-RAN if virtualized), allowed the efficient management of thousands of RRHs with a central eNodeB pool (called BBU hotel); see Figure 1.7.

With the LTE era, operators have not seen the exceptional returns on investment they'd captured with earlier generations, and revenues on new LTE roll-out showed little or no growth. Another factor is the current saturation of the smartphone market and the increase in the expected replacement cycle for these smartphones from two to almost four years in 2019. These factors have contributed to mobile subsidy forming a significant proportion of network capital expenditure (CapEx).

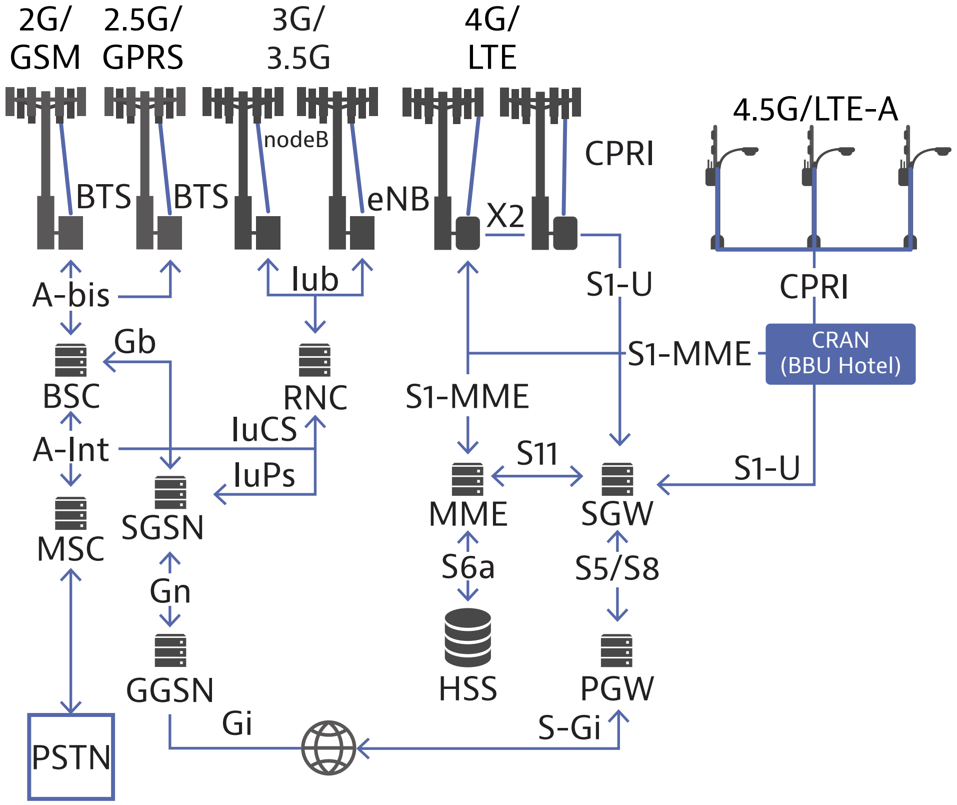


Figure 1.7 Simplified view of 4.5G architecture

1.2.4 Virtualization and Telecommunications Lifecycle

Telecommunication standards have traditionally taken up to fifteen years to make a major step forward to the next generation and delivery of a complete end-to-end standard. Innovations in radio communication is ultimately an analog game. Progress is held back by limitations of what can be achieved with technology that must span the digital and analog worlds while being high performing, requiring a reasonably low amount of power and, for the UE, fitting into a handheld device. Then there is the Shannon limit, which places an upper limit on how much information can be transferred over a given channel bandwidth of spectrum with a given level of interference and noise. This is an ever-present specter which, until the next innovation, appears to make progress in ever diminishing steps.

However, this is changing. LTE took less than ten years. Whether or not Androids dream of electric sheep, there has been a steady convergence between the worlds of information technology and telecommunication technology, leading to an increasing "softwarization" and "virtualization" of telecommunication infrastructure.

It is easier to virtualize CN functions as they are less processor-intensive than the RAN, so it is here that virtualization had its first significant effect with the introduction of the bearer-independent CN in 2004 in Rel-4 of the UMTS/ GSM Standard. This saw the division of the MSC into an MSC server (MSC-S) and a media gateway (MGW) that allowed separation of the functions related to control plane from the user plane. This was followed by the standardization of the IP Multimedia CN Subsystem (IMS) in Rel-6 in 2006 that had the vision, if not the actuality at the time, of transitioning support of all voice call related services into the IP domain. As noted above, this didn't really have an impact until 2012 when there was a market imperative to support voice service seamlessly across LTE and legacy CS networks.

The introduction in 2008 of the EPC in Rel-8, starting with a blank sheet of paper, had more freedom to disaggregate the CN into separate server-based entities including, as mentioned above, the HSS to manage subscriptions, the PCRF to administer service and admission rules, and the AAA for security.

3GPP Rel-14 in 2017 saw the introduction of separation between control plane (CP) and user plane (UP) entities in the LTE EPC called CUPS. This addresses the kind of issue that had occurred with the introduction of GPRS, where the mobility management function in the CN had to deal with all the users attached to the system while there was actually very little user data. Separation of CP and UP allows independent scaling, location, and upgrading of the functions.

Historically, telecommunication investment cycles have been slow with network infrastructure having an expected lifespan of ten to fifteen years, and a payback time typically of multiple years. Mobile devices have historically experienced replacement cycles of at least two to three years with associated mobile device subsidies forming a relatively small part of the overall CapEx of the network operator. These factors have supported the tendency for RAN to be deployed on bespoke, high-performance hardware that is complex for network equipment manufacturers (NEMs) to develop and is also inflexible, being purposed to a limited set of tasks, even while softwarization was starting in the CN. However, there have been considerable advances in generic computer server platform technology and functional abstraction technologies such as DPDK [4], and real-time operating systems (RTOS) that are increasingly enabling common off-the-shelf (COTS) servers to "softwarize" more aspects of the RAN, thereby eliminating the need for a bespoke hardware platform on which to execute. Additionally, technology developments are now making the power of devices such as application-specific integrated circuits (ASICs) and central processing units (CPUs) sufficient to address the complexity of OFDMA air interface. In concert, the evolution of transceiver/power amplifier (PA) design has enabled high-power transmission of higher modulation schemes with enough fidelity (low error vector magnitude (EVM)) to make them practical.

Virtualization is supported by disaggregation of the RAN, that is, dividing the functionality so that it can be scaled and located independently. However, it should be noted that, to a limited extent, RAN disaggregation has been available from GSM. For example, some implementations used a proprietary PHY-level split to allow the RF functionality to be located separately, on an antenna mast, from baseband processing at the base of the tower. This was developed further in 3G by introduction of the common public radio interface (CPRI) [5] which is closer to be an open interface. This allows a centralized RAN to be deployed where all base station functionality is centralized, and optical fiber is used to distribute in-phase and quadrature (IQ) samples of RF baseband to remote "dumb" radio units.

The faster-than-expected introduction of LTE demonstrates the evolution of the economics driving the business cycle. Furthermore, these developments in technology offer freedom to accelerate this trend, enabling a more agile opportunity-focused investment cycle by reducing the prevalence of monolithic forklift interchangeable network elements based on bespoke hardware. Potentially this offers significant savings in CapEx and in operational expenditure (OpEx) as RAN functions may be flexibly orchestrated, i.e., created, configured and deleted, as software functions. The topic of virtualization and its supporting ecosystem is addressed in more depth in Chapter 6.

1.2.5 2017: The Game-Changer 5G

Before we jump into what drove the acceleration behind the first 3GPP 5G Standard release in 2017, it is worth noting that there were other global drivers in the world economies that may have influenced such development, including the rise of China as a fast market economic leader; the need for the European and North American markets to reverse the GDP growth stagnation since 2007; and the unstoppable expansion of the non-traditional communication players like Google, Amazon, Facebook, Apple, Netflix, and Uber offering new, innovative services. All these factors are the precursor of the anticipated "Artificial Intelligence Age," with automation and everything connected to everything as explained in the introduction to this book. It is not hard to see why a new network generation design is needed at the heart of such technological revolution with demanding features like unlimited bandwidth and Ultra-Reliable Low Latency.

It seems apparent that the designers of previous generations of mobile telecommunication standards created systems that efficiently delivered the one or two services that they were designed for. However, the systems were not designed to be easily adaptable to the needs of new services. After deployment, the systems had to be substantially re-engineered to meet the evolving/developing service needs of the subscriber or wait until elements of the core network had been upgraded to allow service continuity. GSM supported voice and low-data-rate circuit-switched data but had to be re-engineered to support packet data. UMTS enhanced capacity to support voice and increased available data rate. However, it did not efficiently support "always-on" type services with occasional high peak rate and low latency requirements. It was subsequently re-engineered with HSPA+ to do this. Additionally, both GSM and UMTS air interfaces were restricted in their flexibility to adapt to the demands of MTC and were limited to introducing mechanisms to limit network congestion.

In contrast, LTE supported efficient packet data transport for both high-bandwidth services and small-packet voice-like services. However, adaptations to the core network were required to allow a seamless voice service across LTE and legacy networks, and these were not available when LTE was first deployed.

Additionally, problems were encountered when the control traffic and associated processing did not scale at the same rate as that of the user traffic. This occurred with the adoption of GPRS due to the prevalence of always-on type services, and with UMTS where many adaptations were required to circumvent the control "channel heavy" nature of the UMTS DCH channel. Adapting the systems to serve both high bandwidth and small-packet users with a minimum of control channel overhead and latency was challenging, and arguably non-optimal solutions were developed.

Finally, the economics of the telecommunication industry have changed. The straightforward model of wholesale replacement of the earlier standard as a forklift exercise has been made a less attractive proposition, as evidenced by the little-to-no growth in operator revenue after LTE deployment in some markets. Technological and ecosystem trends mean that this is no longer the only show in town.

An overarching trend against the backdrop of the respective waves of technology, depicted in Figure 1.1, has been ever-reducing cell sizes creating a "densification" of the network. This enables the available frequency resources to be reused over shorter distances to provide extra capacity. Furthermore, on the one hand, handsets are constrained by battery power, which limits the available transmit power; on the other, increasing data rates mean more bits must be exchanged. The shorter cell range reduces propagation loss and helps the system meet energy-per-bit/noise-floor requirements.

However, densification increases the number of elements required to be deployed and managed, which, all things being equal, increases CapEx and OpEx. This expectation is supported by an economic analysis by Frisiani, et al. (McKinsey) that, viewed globally, showed that CapEx and OpEx are consuming a greater share of operator revenue leading to reduction in cash-flow. Figure 1.8 shows a simple extrapolation [6] of Frisiani's model over the timespan used in Figure 1.1, which illustrates that it is imperative for the new fifth-generation Standard to enable operators to "reinvent" the way they manage network CapEx and OpEx.

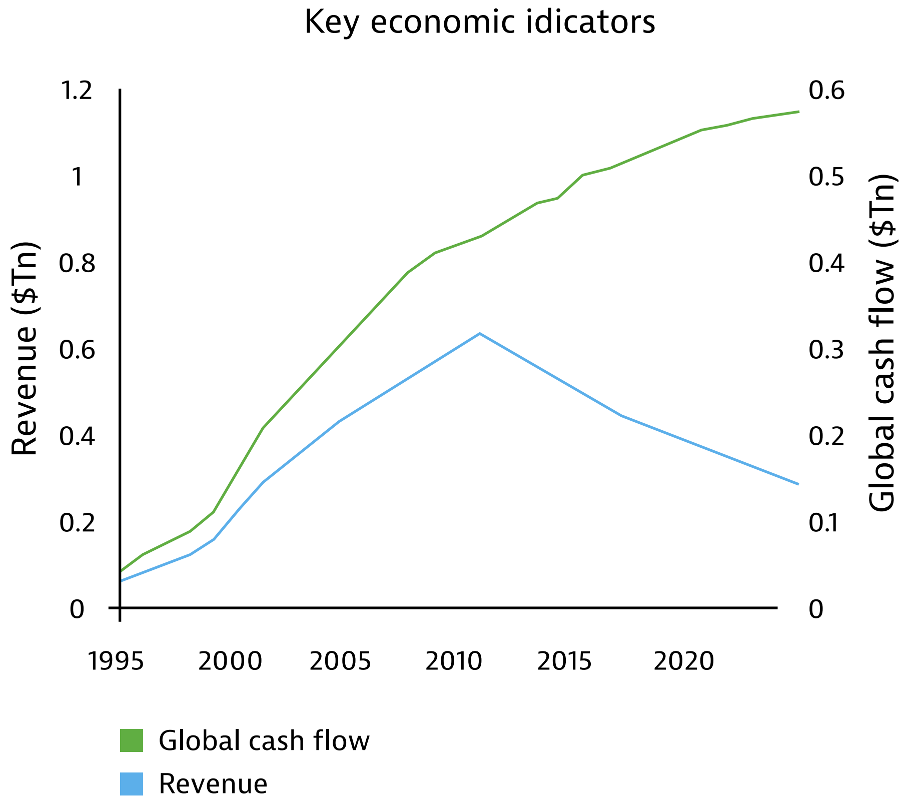


Figure 1.8 Global mobile operator revenue and cash-flow

The overriding driver from the analysis above is that 5G needs to be flexible to allow the system to:

- Adapt to unforeseen service requirements and thereby be better able to exploit new revenue opportunities.
  - Potential candidates are higher-definition on-demand video, smart cities, and IoT for vertical markets such as automotive and emergency services networks.
- Facilitate incremental deployments overlaid on legacy networks to tailor CapEx spend to emerging revenue opportunity to avoid the "forklift" problem.

  For this to happen, 5G has taken on several strategies to design-in flexibility. These include:

- Independence of RAN from CN
  - Aiding overlay deployment on legacy systems.
- Control and user plane separation in CN and RAN
  - Aiding flexible deployment and scaling of processing capability depending on emerging service need and helping support virtualization and "softwarization."
- Service-based architecture in CN
  - Moving from a point-interface based architecture to a service-based architecture that utilizes capability discovery and exposure, accelerates deployment of new services, and circumvents 3GPP standards release cycle time for interface updates.
- Virtualization and orchestration
  - Facilitating rapid reconfiguration of network functionality and low-cost network operation even given that the complexity of the network, as represented by the number of network sub-functions that need to be instantiated and interconnected, is increased.
- RAN disaggregation and open interfaces
  - Aiding flexible deployment and scaling of processing capability and helping support virtualization and "softwarization."
  - Open interfaces facilitating innovation and "best-in-class" sourcing at a disaggregated RAN sub-function level.
- Analytics and artificial intelligence
  - Facilitating optimization and network operations and generating new revenue streams through monetization of subscriber data.
- Network slicing
  - Facilitating flexible service definition and support for new vertical services such as IoT and emergency services networks (ESNs).

Some of these factors are directly addressed by the 3GPP standard, while others are addressed by industry groups that effectively "fill in the gaps" between the standard and its physical or virtual embodiment.

1.3 5G Use-Case Based Service Drivers

It is clear from the history of the various generations of mobile technology set out above that there is an inherent uncertainty in how the network will be operated and what it will be used for. Furthermore, some of the requirements presented by these service needs are contradictory. For example, IoT needs very low control overhead for minimized battery consumption and is delay tolerant. Conversely, ESNs demand very low latency and can tolerate relatively high battery consumption. However, previous systems focused on delivery of peak data rate and minimum call setup latency, which tended to "bake-in" a heavyweight control channel that subsequently limited the ability to add services optimized for minimum power, such as MTC services.

In the absence of a means to increase prescience, the 5G designers have approached this problem by defining a diverse-as-possible set of potential use cases that exposes upfront many of the contradictions in system design that caused problems with earlier generations of technology. This allows the designers to adopt a flexible system design that avoids a service-specific hard-wired approach and facilitates subsequent re-engineering to accommodate unforeseen services. Moreover, it supports the ITU design premise, "Depending on the circumstances and the different needs in different countries, future IMT systems should be designed in a highly modular manner so that not all features have to be implemented in all networks."

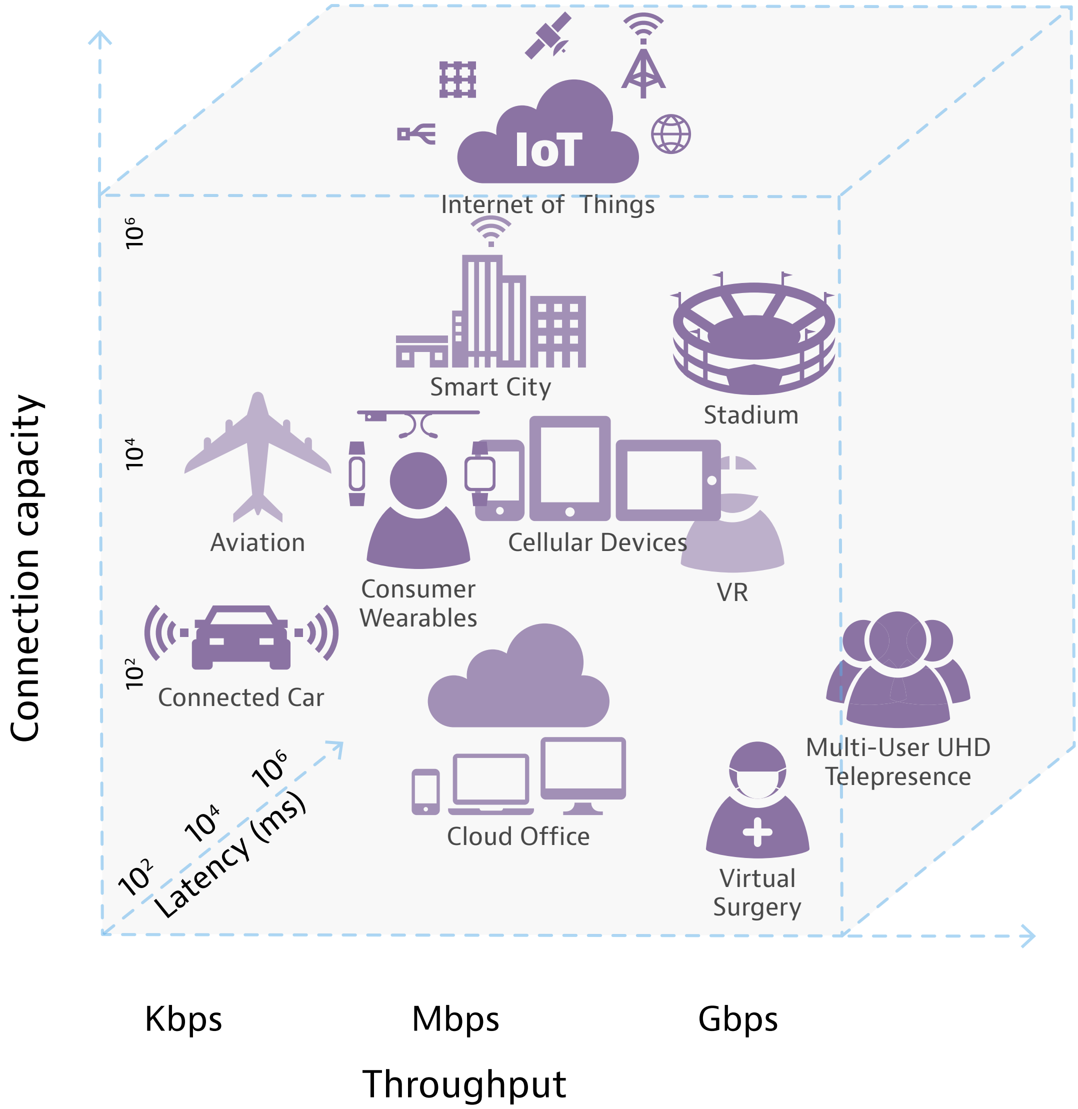


Figure 1.9 The main 5G use-case scenarios (ITU IMT Vision)

The primary usage scenarios are enhanced Mobile Broadband (eMBB), which follows the historic trend to provide ever greater bandwidth; massive Machine Type Communications (mMTC) that demands Ultra-Reliable low overheads to enable battery life in excess of ten years; and Ultra-Reliable Low Latency Communications (URLLC) that demands resilience along with sub-millisecond service latency. These were designed by the ITU to support IMT-2020; they were not designed to be exhaustive, and additional unforeseen use cases are expected to emerge. Figure 1.9 summarizes these scenarios and highlights some of the potential applications that could be supported; eMBB is toward the righthand plane of the Figure; mMTC is toward the top plane of the Figure; and URLLC is toward the front plane of the Figure. 5G is important for operators because it allows the greatest possible freedom of action to respond to market demands for new services with a single, flexible approach.

The approach highlights that these use cases, and the new revenue streams that they represent, are not standalone activities that fall under the conventional umbrella of mobile radio services; rather, they are associated with new industry verticals that have the potential to revolutionize how whole sectors of the global economy may operate. What is being envisaged is a wholescale re-engineering and rebooting of the global economy. The intent is to address how people see and experience reality; how cities operate; how goods are manufactured, distributed, and sold; how crops are grown; how disasters are prevented or managed; how society is protected and served.

The challenges that must be overcome by the single flexible standard to meet the diverse and conflicting service requirements represented by the IMT vision is made more apparent when the key capabilities required to support the use case scenarios are illustrated on a radar plot; see Figure 1.10. The mMTC use case scenario prioritizes connection density and, to a lesser extent, network energy efficiency, and is relatively insensitive to the other requirements. In contrast, URLLC prioritizes latency and mobility while eMBB broadly prioritizes the whole range of key capabilities. For example, this means that the building blocks of the standard—such as those that define how calls are set up and torn down and how the connection between the mobile and the network is maintained— must offer radically different modes of operation. So must the building block that determines the granularity with which chunks of resource are allocated to mobiles, to enable both tiny and massive allocations to be made with efficient signaling. For example, massive allocations may be made on up to thirty-two aggregated carriers to support eMBB, whereas single symbol allocations must be made on a contention basis to support URLLC, as described in Chapter 2 on NR.

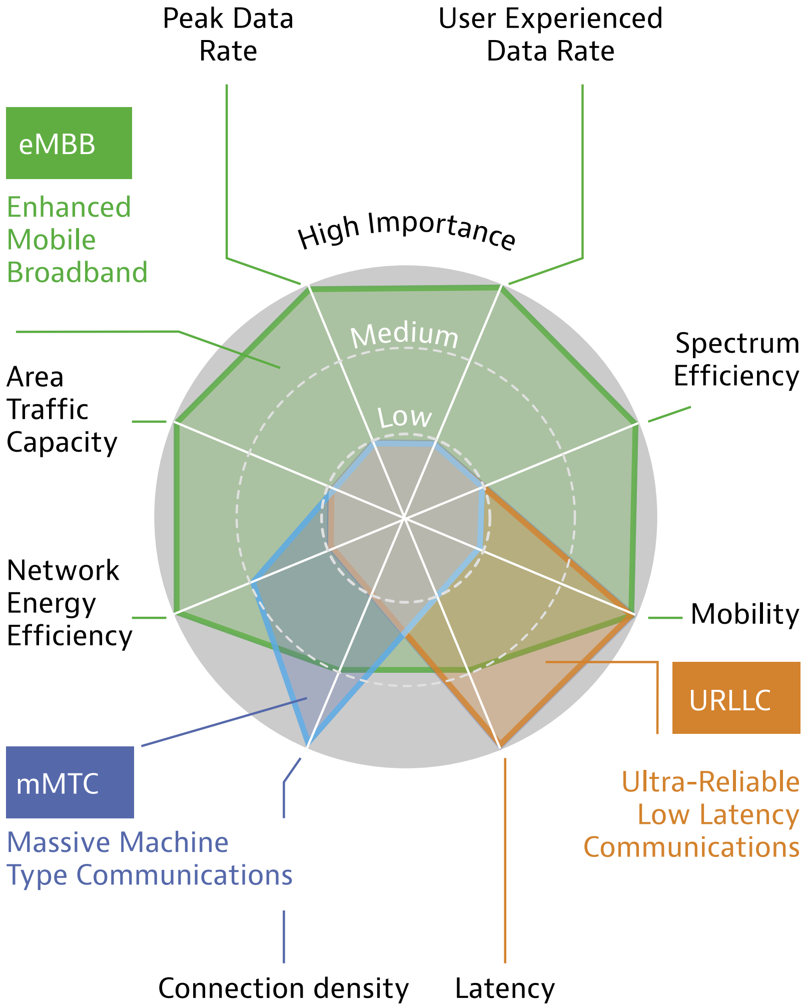


Figure 1.10 The importance of key capabilities in different usage scenarios (From ITU IMT Vision)

1.4 Independence of RAN from CN

This is quite a pragmatic objective. The initial focus of the 3GPP standardization activities has been on completing the so-called non-standalone (NSA) variant of the network architecture in Rel-15. This architecture enables the 5G NR to be supported by the existing RAN and EPC [7], which is the expected initial deployment as 5G is likely to be targeted at dense urban areas that have an existing LTE deployment. Additionally, it helps avoid dependencies of RAN features on CN capability. It is not an exact parallel, but it is interesting to compare this with the example of VoLTE introduction where the seamless handover feature, SRVCC, required a CN supporting IMS to be available. The standalone (SA) architecture cedes control from the EPC to the 5G core network (5G CN).

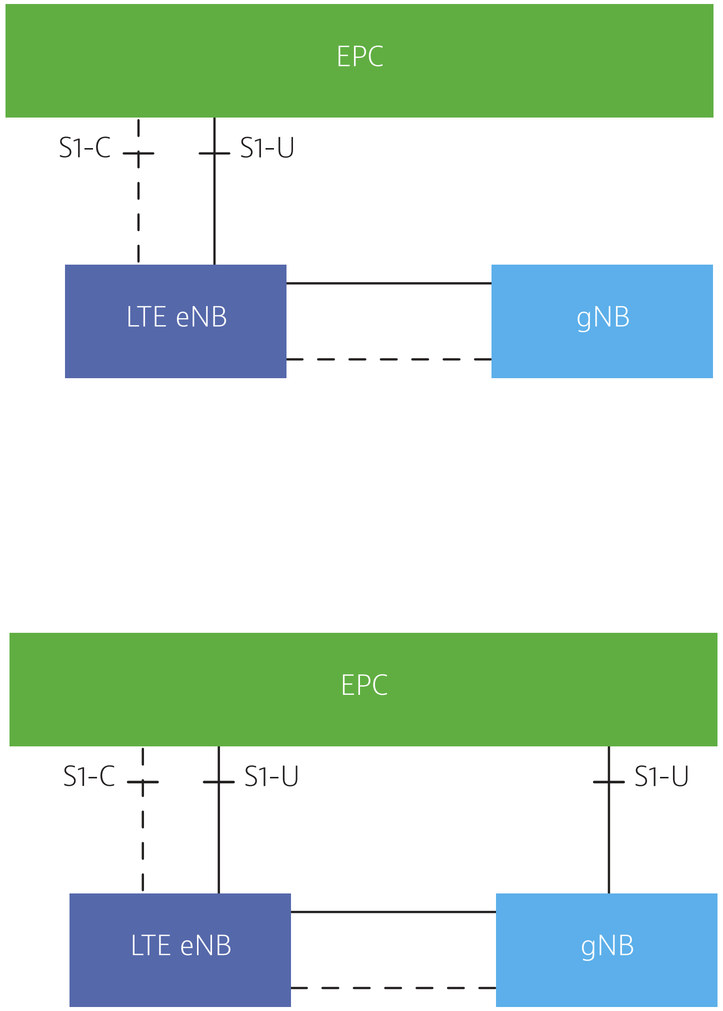


Figure 1.11 The NG-RAN NSA option 3/3a. Option 3x was added later (ex 3GPP 38.801)

The terminology about the potential NG architecture options was defined in the NR TR 3GPP 38.801 along with the possible NSA architecture for NR-RAN connected to the legacy EPC. In NSA Option 3, both the CP and UP from the gNB are managed by an LTE eNB that connects to the EPC using the S1-C and S1-U interfaces. Whereas in the NSA Option 3a and 3x, the control signaling for the gNB is managed by the LTE eNB that connects to the EPC, but the gNB has a direct S1-U connection to the EPC for the UP. The difference between 3a and 3x is with user plane handling in the gNB as with Option 3x the LTE eNB can be used for user plane transmission which is not possible for Option 3a. This is illustrated in Figure 1.11.

1.5 Control and User Plane Separation (CUPS)

As previously mentioned, separation of CP and UP, introduced to the CN with the CUPS feature in Rel-14, has the benefit of facilitating the scaling and even location of UP and CP functionality independently. This concept has been developed significantly in the new 5G CN, as we touch on below. However, with 5G, this concept has been extended into the RAN. 3GPP has defined the most central part of the gNB as the central unit (CU) and standardized the interface between the CU-CP and CU-UP as the E1 interface. This allows a single CU-CP entity to manage multiple distributed CU-UP entities. A CU-CP may manage the CP for multiple distributed unit (DU) entities and a CU-UP may manage the UP for multiple DU entities. A DU is managed by a single CU-UP but may have connections to others to provide redundancy. This is illustrated in Figure 1.12.

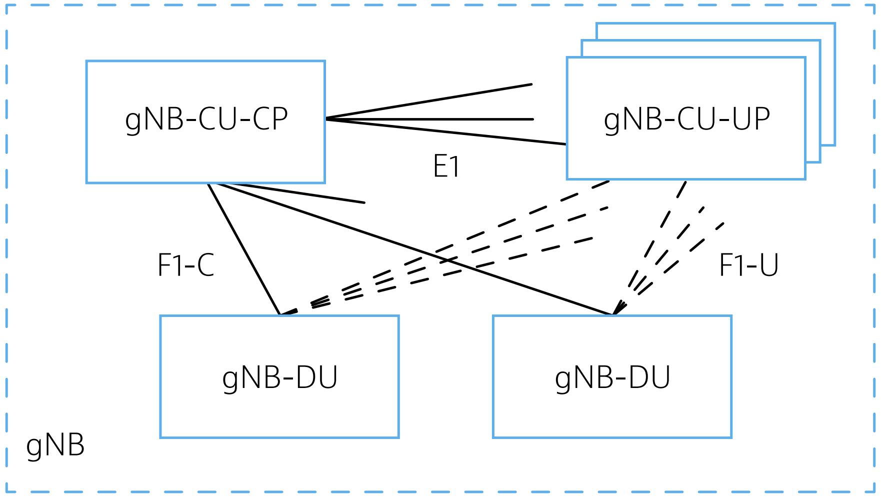


Figure 1.12 CP and UP separation in NG-RAN [from 3GPP TS 38.401]

1.6 Service-Based Architecture in 5G CN

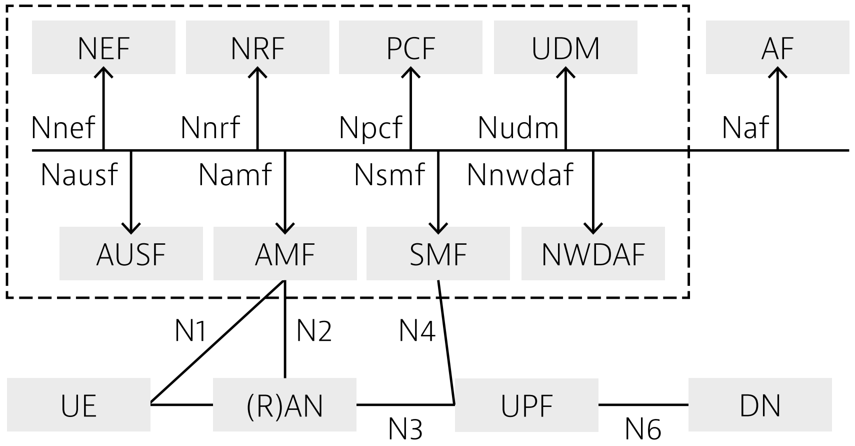
The 5G CN has extended the principle of CP-UP separation to incorporate the approach used to construct typical web services, that is, to define the functions as services. So rather than defining point interfaces between the different functions that require up-front definitions of protocol messages, which makes the definition of new services subject to the long cycle-time of 3GPP standardization, the process of service discovery and utilization is effectively self-defining and extensible. A point interface-based architecture is available for Rel-15 and Rel-16, but the evolution path is toward the Service Based Architecture (SBA), as illustrated in Figure 1.13. Rather than showing interface names connecting the functions, the figure shows each available service with proper capitalization.

1.7 Virtualization and Orchestration

Virtualization of 5G CN and RAN was introduced above. Orchestration is the process of instantiating the set of virtualized network functions (VNFs) and physical network functions (PNFs) across the requisite pools of virtualization infrastructure and physical hardware, to meet the required performance profile for each of the sets of network slices. The promise of virtualization and orchestration is to automate a great many of the processes involved with deploying and operating cellular networks. The various approaches to the topic and the players involved in each are considered in more depth in Chapter 6.

1.8 Disaggregation and Open Interfaces

Disaggregation of the RAN increases the flexibility of the available deployment options. It creates the prospect of allowing infrastructure to be constructed from the best available global services and products, reducing time-to-market, limiting vendor lock-in, and enabling more cost-effective roll-out solutions. In fact, some use cases are likely not achievable with monolithic base stations due to the inability to separately scale or locate functions.



NEF	Network Exposure Function
NRF	Network Repository Function
PCF	Policy Control Function
UDM	Unified Data Management
AF	Application Function
AUSF	Authentication Server Function
AMF	Access & Mobility Function
SMF	Session Management Function
NWDAF	Network Data Analytics Function
UE	User Equipment
RAN	Radio Access Network
UPF	Application Function
DN	Data Network

Figure 1.13 3GPP 5G Core Network Service-Based Architecture (SBA)

Potentially, RAN disaggregation makes delivery of RAN more amenable to "agile" style development processes, in the sense of being able to divide up technology into elements one minimal capability at a time. The Telecom Infrastructure Project (TIP) is promoting such an agile style development where requests-for-information (RFI) to potential vendors replace overarching standardization to develop new telecom products. In principle, this may help address the issues discussed above in connection with the mobile telecommunications business lifecycle.

In addition, by demarking the RAN into separate functional elements, RAN disaggregation facilitates extension of network function virtualization (NFV) into the RAN. This creates a virtualized deployment environment that enables a more agile delivery of incremental (and disaggregated) 5G functionality as the vendor is no longer tied to the delivery of a fixed-sized hardware platform.

Figure 1.14 summarizes the functional elements of each protocol layer of a 4G/5G RAN for both uplink and downlink; the functions are essentially generic at this level of abstraction. The figure illustrates the more popular split points within the RAN. The split between the PDCP and the RLC, referred to as high layer split (HLS), has been standardized by 3GPP as the F1 interface for 5G in the 38.470 Series Specifications, and the W1 interface for 4G. The splits between the MAC and the PHY and within the PHY are referred to as lower layer split (LLS). 3GPP studied the LLS, but standardization has been left to the market with the Small-Cell Forum and the Telecom Infrastructure Project (TIP) promoting Split 6, and the O-RAN operator-led forum promoting split 7-2x. The figure shows the names [8] of the split parts of the RAN, the central unit (CU), the distributed unit (DU), and the radio unit (RU). The shading indicates functions residing in RU and DU depending on the chosen split, which may differ for UL and DL.

Previous generations of the RAN have adopted elements of RAN disaggregation. For example, CPRI was introduced in the 1990s, and now would be classified as an option 8 split between the PHY and RF that transports digital samples of the baseband analog waveform at about 20x line rate. It enabled the RF to be mounted at the mast head, reducing energy consumption as cable loss is eliminated and heating, ventilation, air conditioning (HVAC) requirements are reduced. It eases many elements of deployment and troubleshooting and even RF interference analysis (apart from climbing the mast to inspect the RF). Additionally, it provided the flexibility to implement a C-RAN. However, the explosion in data rates and numbers of antenna elements makes it prohibitively expensive to scale to 5G.

The options for disaggregation and the resultant interfaces are addressed in more detail in Chapter 3.

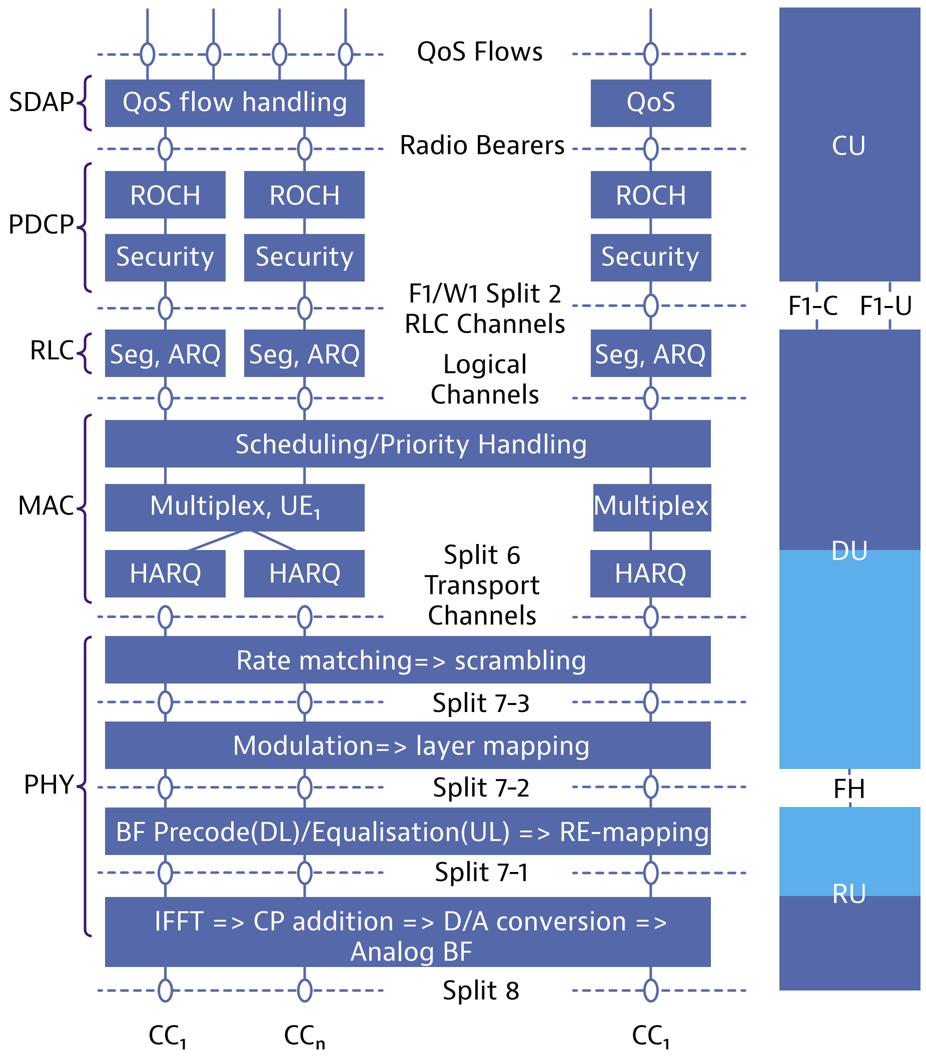


Figure 1.14 Telecom functions of the RAN showing Functional split options

(PDCP also provides its services to RRC layer, not shown here)

1.8.1 O-RAN

RAN disaggregation and open interfaces, as mentioned in the previous Section, facilitates flexibility and deployment options. However, it also brings more complexity.

A principal consideration for 5G is the scale and flexibility of deployment, optimization, management and orchestration of the network, and this is only made more pressing by the use of open RAN. Delivering new services and managing RAN capacity will no longer be practical if managed manually. Intelligence and automation must be integrated into all aspects of the network lifecycle to reduce both CAPEX and OPEX. Like RAN disaggregation, intelligence in every layer of the RAN architecture is at the core of open RAN technology.

This will allow operators to deploy a truly self-managed, zero-touch automated network. Consider an example, where base band capacity can become a bottleneck during an unplanned network event – with artificial intelligence and machine learning, this event can be detected and characterized in a short amount of time and additional capacity can be introduced quickly and efficiently on a white-box platform to overcome that challenge.

To achieve the above-mentioned goals of an open radio access network, operators founded the O-RAN Alliance to clearly define requirements and help build a supply chain eco-system that can foster an environment for existing and new vendors to drive innovation. As per the charter of O-RAN alliance, O-RAN Alliance members and contributors have committed to evolving radio access networks around the world. Future RANs will be built on a foundation of virtualized network elements, white-box hardware and standardized interfaces that fully embrace O-RAN's core principles of intelligence and openness.

The key principles of the O-RAN Alliance include:

- Lead the industry towards open, interoperable interfaces, RAN virtualization, and big data enabled RAN intelligence
- Specify APIs and interfaces, driving standards to adopt them as appropriate and exploring sources where appropriate
- Maximize the use of common off-the-shelf hardware and merchant silicon, thus minimizing proprietary hardware.

1.9 Network Slicing

Network slicing is not a new concept within the mobile telecommunications world. For example, mobile virtual network operators (MVNOs) exploit slicing in legacy networks. Typically, this is accomplished by reserving a set of subscribers IMSI [9] for the MVNO and slicing at the subscriber management/billing layer. Network sharing can also be considered a precursor for slicing. For example, the MOCN [10] shares a single RAN between different operators' CN; and MORAN [11] shares a single RAN with separate frequency allocation per operator and GWCN [12].

However, the control and user plane separation in 5G, particularly with the 5G CN SBA, allows a much finer granularity of slicing. The functions in the network become logical functions that may be instantiated in physical locations as service requirements and capabilities demand. This is further enhanced by Network Function Virtualization (NFV) that permits the logical functions to be instantiated on a virtualization abstraction layer hardware supported on COTS hardware.

In this new context network slice is defined as "a logical network that provides specific network capabilities and network characteristics," and a network slice instance is defined as "a set of Network Function instances and the required resources (e.g. compute, storage, and networking resources) which form a deployed network slice." Consequently, the slice instance will also determine the preferred control/user plane splits, function locations, and required telemetry to provide assurance of the SLA.

Example types include slices for emergency services networks (ESNs), mMTC, and enterprises. Moreover, an MVNO may be provisioned as having access to a subset of slices of the required types.

Network slicing effectively requires disaggregation of the 5G CN and RAN in the service/tenant domain. Ideally, the slices are independent and isolated from the point of view of SLA assurance, as this simplifies resource management and service of SLA. However, this arrangement requires a sacrifice of efficiency. Additionally, there are limits to the isolation that is attainable, for example, when it comes to meeting stringent latency and bandwidth requirements on the air interface. Resource allocation to the slices is generally dynamic and potentially contingent on priority, e.g. for ESN, and the concept of a broking service to manage this contention has been proposed.

1.10 Navigating the 3GPP Standards for Architecture

TS 22.261: 5G service requirements

TS 23.501: System Architecture for the 5G system describes the overall architecture placing the NG-RAN in the context of the 5G core using either a reference-point interface-based architecture and the future service exposure-based architecture

TS 23.251: Network sharing architecture and functional description

TR 36.576: Study on architecture evolution for Evolved Universal Terrestrial Radio Access Network: Discusses RAN LLS

TS 37.324: Service Data Adaptation Protocol (SDAP); specifies SDAP for a UE with connection to the 5G CN.

TR 38.801: Study on new radio access technology: Radio access architecture and interfaces Defines URLLC; 5G architecture options; Defines RAN functional split options

TR 38.816: Study on CU-DU lower layer split for NR

TS 38.300: NR; NR and NG-RAN overall description, Stage 2 Defines dual connectivity (DC)

TS 38.401: NG-RAN; architecture description

TS 38.410: NG-RAN; general aspects and principles

TS 38.420: NG-RAN; Xn general aspects and principles

TS 38.460: NG-RAN; E1 general aspects and principles

TS 38.470: NG-RAN; F1 general aspects and principles

TS 28.500: Management concept and architecture for NFV

1.11 5G Architecture Blueprint

Figure 1.15 illustrates a simplified SA 5G architecture overlaid on the legacy of previous generations. The technologies summarized in previous sections allow an operator to flexibly choose the appropriate level of aggregation/disaggregation, for example adopting one or another C-RAN approach to balance benefit from sharing, and advanced antenna techniques, with stringency of fronthaul bandwidth and latency requirements. Within that aggregation/disaggregation choice, separation of CP and UP allows the decision to be made separately for control plane and user plane processing to allow optimal performance and processing capacity scaling. Further, with the addition of RAN virtualization and network slicing, such choices can be made on a per-slice basis. For example, a URLLC service slice may combine CN and RAN functionality at the edge of the network to achieve sub-millisecond latency, whereas an mMTC service slice may aggregate CP functionality on a national basis. Moreover, the combination of open interfaces, network function virtualization, and ability to support edge computing allows for new models of RAN operation and optimization to be explored with the potential to deploy third-party RAN Intelligent Controller (RIC) algorithms into the RAN.

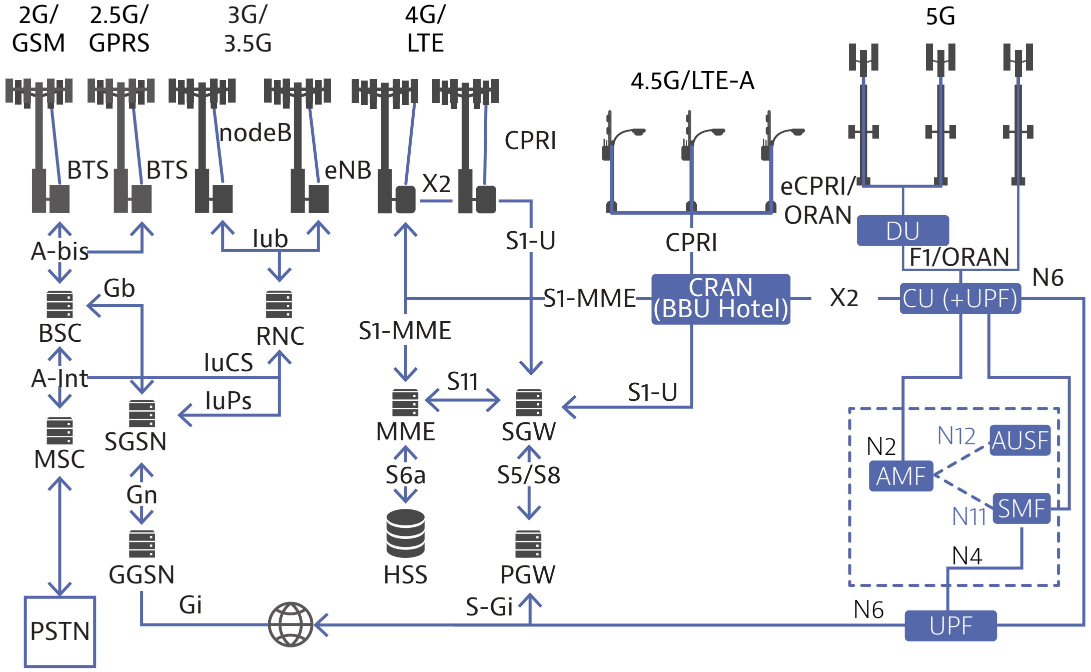


Figure 1.15 Simplified view of the 5G architecture evolution story

In conclusion, we quote the 3GPP "5G Stage 1 Service Requirements" to sum up the strategies that have been applied in the definition of 5G to enhance its flexibility:

"Flexible network operations are the mainstay of the new system. The capabilities to provide this flexibility include network slicing, network capability exposure, scalability, and diverse mobility. Other network operations requirements address the necessary control and data plane resource efficiencies, as well as network configurations that streamline service delivery by optimizing routing between end-users and application servers."

At the beginning of this chapter, we determined to set the evolution of the 5G CN and RAN architecture in context of the history of exponential growth in mobile subscriber volumes, explosion in data usage, and revenue capture by the OTT players. It became apparent that conventional network design and deployment approaches were no longer sufficient to maintain required levels of cash flow.

However, according to the analysis of Frisiani, et al. (McKinsey), the 5G approach to design-in flexibility offers a bright ray of sunshine that application of virtualization and softwarization techniques has the potential to dramatically reduce CapEx and OpEx costs.

Figure 1.16 shows the global revenue and cash flow (as shown in Figure 8), including an adjusted cash flow considering CapEx and OpEx enhancements arising from softwarization and virtualization according to the prediction of Frisiani, et al. In their paper they suggest that CapEx spend ratio can be reduced from 16-17% to less than 10% of revenue and that OpEx can be reduced from 50% [13] to about 20% of revenue. In the Figure, we assume these changes occur over a four-year period from initial deployment. The scope of the paper addresses all aspects of an operator's business, including customer care and sales, as well as acquisition and operation of a virtualized RAN, so the Figures should be regarded as indicative or even aspirational. However, notwithstanding these caveats, the graph, adjusted for the expected benefits of softwarization and virtualization, is encouraging in that it suggests that 5G will be able to meet its design targets at least from an economic perspective. In the subsequent chapters we address in more detail the potential and challenges of the 5G system.

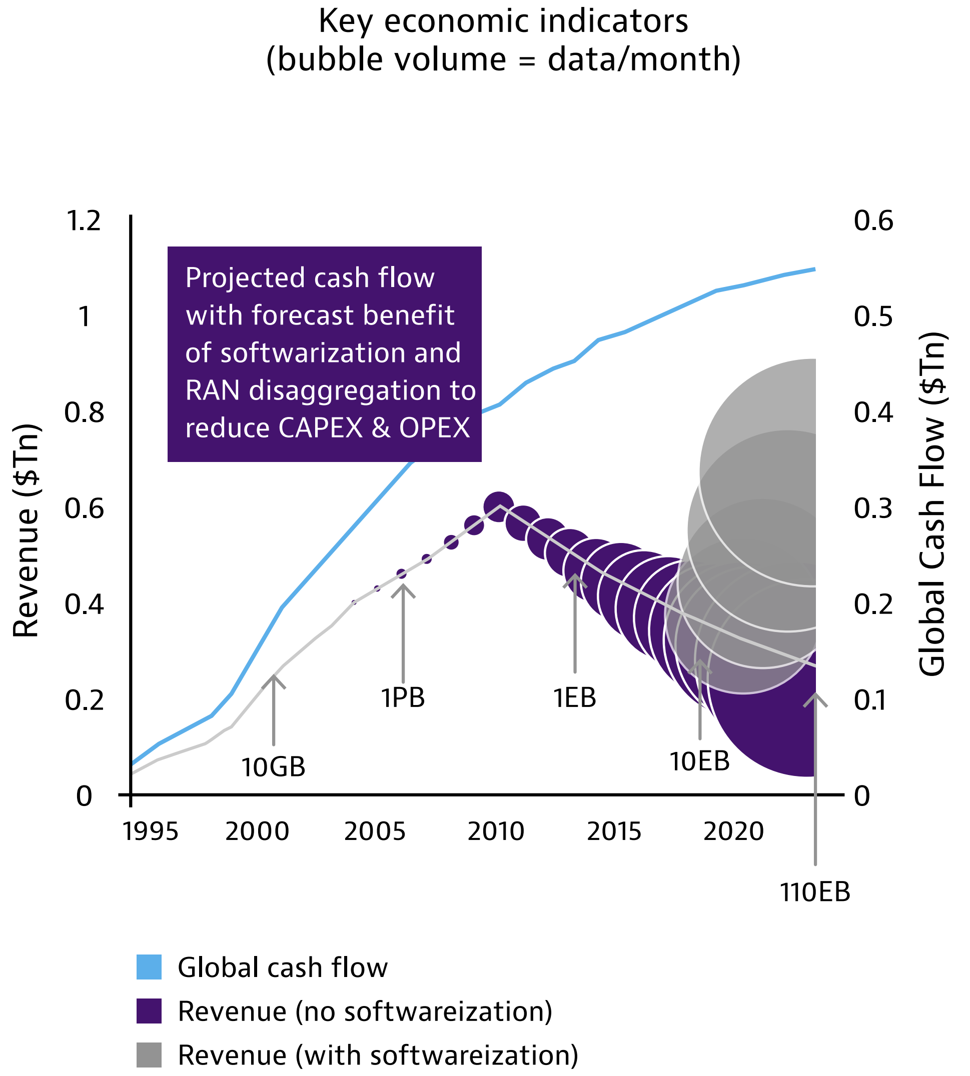


Figure 1.16 Global revenue and cash flow (adjusted for "softwarization" benefits)

Notes

1. UMTS Routing Area
2. Provided by the feature's semi-persistent-scheduling and TTI-bundling features, which support the steady stream of small packets with defined latency requirement that characterize the voice service. Available from LTE Rel-8 v8.5 in 2009.
3. 256QAM was introduced in Rel-12, but is considered as part of LTE-Advanced-Pro.
4. Data Plane Development Kit – an Intel technology to enhance packet processing for virtualized implementations.
5. This is referred to as Split-8 in the 3GPP study on NR split options.
6. Extrapolation backwards assumes cash flow tends toward 35% of revenue before 2003 and follows the simple 6% decline suggested by Frisiani after 2010.
7. EPC = Evolved Packet Core.
8. Various groups have played in this naming sandpit and a variety of terminology is used.
9. International Mobile Subscriber Identity.
10. Multi-Operator Core Network – single-shared RAN with independent CN.
11. Multi-Operator RAN – shared RAN with separate frequency bands per operator and independent CN.
12. Gateway CN – multiple shared CN and shared RAN.
13. Actual OpEx Figures are inferred from the ratio for CapEx in the paper and the revenue Figures available from the GSMA and the ITU.

CHAPTER TWO

5G New Radio

In the span of twenty-five years, the world went from a few million phones with only voice capability at the beginning of 1990 to almost five billion phones with more than twenty thousand exabytes of mobile data used by the end of 2015. Despite these amazing technological feats, not everyone on Earth has a mobile phone. In 2020, we estimate that between twenty-five and thirty percent of Earth's citizens still have no mobile device. On the other hand, we estimate upwards of twenty billion Machine Type Communications (MTC) connections are active in that same year.

One key element that will enable this exponential expansion in connections and usage is the efficient use of radio spectrum at higher and higher frequency bands.

The radio spectrum ranges from low-frequency waves at around 10 kHz up to high frequency waves at 100 GHz. The mobile network uses predefined bands of various widths at different frequency ranges in the radio spectrum. We can think about these frequency bands as highways: the wider the highway, the more cars (data) go through it. However, the higher the frequency, the more energy is needed to make it go through obstacles (walls, glass, trees, air, etc.), which means the shorter distance it can go before dissipating.

The first generation of mobile technologies used frequencies ranging from 450 MHz to 900 Mhz. The second generation, 2G, expanded the frequency ranges used up to 1900 MHz. With the enhancements made by the General Packet Radio Service (GPRS or 2.5G), mobile data speeds reached upward of 114 Kbps (kilobits per second). Enhanced Data Rates for GSM Evolution (EDGE) supported speeds of more than 200 Kbps.

Smartphones started to emerge and expand toward the end of the 2.5G era. While using the same spectrum frequencies as its predecessor, 3G went on to efficiently use the spectrum by splicing the data across different frequency channels, hence increasing the total throughput per single user. In 3.5G, data rates upwards of 40 Mbps (megabits per second) were achievable.

4G expanded the frequencies used to between 3 GHz and 4 GHz and added more spectrum efficiency by using a technology called OFDM (Orthogonal Frequency Division Multiplexing); speeds of 100 Mbps were attainable.

LTE Advanced (aka 4.5G) added carrier aggregation, a technology to combine two or more separate LTE carriers into one data channel, to support speeds up to 1 Gbps (gigabits per second).

2.1 5G to Simultaneously Reduce Latency by an Order of Magnitude and Increase Data Speeds by One to Two Orders of Magnitude over LTE

As we saw in Chapter 1, the aspirational capabilities of 5G are ultimately driven by a desire to create a network able to simultaneously satisfy a vast range of use cases with widely disparate performance characteristics. The adaptable and dynamic network that results will foster an ecosystem in which new services that were previously not possible can be created. The resulting technological inflection may even see the emergence of entirely new industries.

The 5G system needs to satisfy various performance goals, referred to by standards organization 3GPP as the 5G service enablers. These are the enormous data throughput rates required by enhanced Mobile Broadband (eMBB) along with Ultra-Reliable Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC). Satisfying each of these service enablers in isolation is challenging. Satisfying all these together simultaneously in the same network further compounds that challenge. An additional level of complexity is introduced by the desire to reduce power consumption of 5G. As we shall see in this chapter, the new radio aspects of 5G standardized by 3GPP take a significant step toward making these requirements a reality.

Here we examine why the performance goals being asked of 5G are so challenging, and how the 5G NR is addressing those challenges.

2.2 Frequency Bands for 5G

To put the challenge for 5G to deliver the eMBB service enabler into context, consider how enormously consumers and business fixed broadband services have evolved over the last two decades. Innovations in transmitting data at ever higher rates and over longer distances through xDSL, cable, and fiber connections has underpinned the explosion in a streaming media industry. The penetration of fiber links into these networks and the associated ability to transmit with less noise and signal attenuation has pushed data rates way up and unlocked more industry verticals to the retail subscriber.

The wireless broadband challenge stems from the inherently shared nature of the medium. In comparison to fiber, xDSL, or cable, the radio interface is an analog and shared resource subject to harsh conditions such as interference and noise. If the eMBB service enabler is to be delivered, the network must squeeze orders of magnitude more data through the same radio resources. While there are numerous hurdles to this, the principal limitation is that the spectral efficiency is reaching the Shannon limit, meaning we cannot do significantly better just by squeezing more data through the same spectrum resources without a radical rethink of how to utilize the resource more efficiently.

Although 5G does address this spectral efficiency limitation, it also addresses the limitation through allocation of new spectrum for 5G carriers. This immediately raises the challenge that the valuable resources below 6 GHz are in very limited supply. Older mobile communication technologies have relied on spectrum in this range because it has more attractive propagation characteristics. These characteristics include an ability to achieve non-line-of-sight propagation by combinations of reflection, scattering, or diffraction, and in some cases penetration of physical barriers. While some of this can be "re-farmed" from older technologies to 5G, this takes the spectral efficiency only a little closer to the Shannon limit rather than delivering the orders of magnitude improvements required by 5G. The spectrum is also still in short supply.

The solution to the capacity problem adopted by 5G is to open up the higher frequencies, including millimeter Wave (mmWave) bands, because they have the major advantage of significantly more plentiful spectrum, and spectrum regulators are more willing to make them available for use by operators of 5G services. The mmWave range also benefits from the availability of wider contiguous blocks of spectrum able to accommodate single 5G carriers with larger frequency bandwidth, which leads to less loss of spectral efficiency from overheads. Although 5G supports aggregation of different carriers together for communication to devices, larger contiguous blocks of spectrum can make for simpler management of the spectrum and less complex devices.

Use of spectrum above traditional blocks for wireless communication and up to the mmWave range imposes various challenges on the 5G standards. At these higher frequencies, propagation starts to become challenging since they typically require line-of-sight. Also, transmissions in this frequency range attenuate through solid objects, meaning that outdoor base stations cannot generally provide indoor coverage. These bands also experience less scattering, reflection, and refraction, presenting more challenges to the 5G network operator to plan and deliver reliable coverage.

This greater supply of spectrum is coupled with other technological solutions in the 5G standards to address the propagation problems at these bands. Received power reduces with the square of the wavelength (multiplied by $\lambda^2/4\pi$) owing to the smaller antenna aperture at lower wavelengths. The higher transmission losses at these bands also need to be addressed. Highly directive radiation of energy can overcome this problem and is a feature of massive MIMO beamforming that will be covered later in this chapter. In practice, although physical antenna dimensions can be smaller for these wavelengths, larger antennas are required to capture sufficient energy.

The 5G standards separate the spectrum supported into two main portions. These are consistent with the bands identified in the International Mobile Telecommunication-2020 (IMT-2020) of the International Telecommunication Union (ITU). This defines Frequency Range 1 (FR1) which occupies the range 450 MHz to 6 GHz. The frequency bands for FR1 defined by 3GPP are shown in Figure 2.1 below.

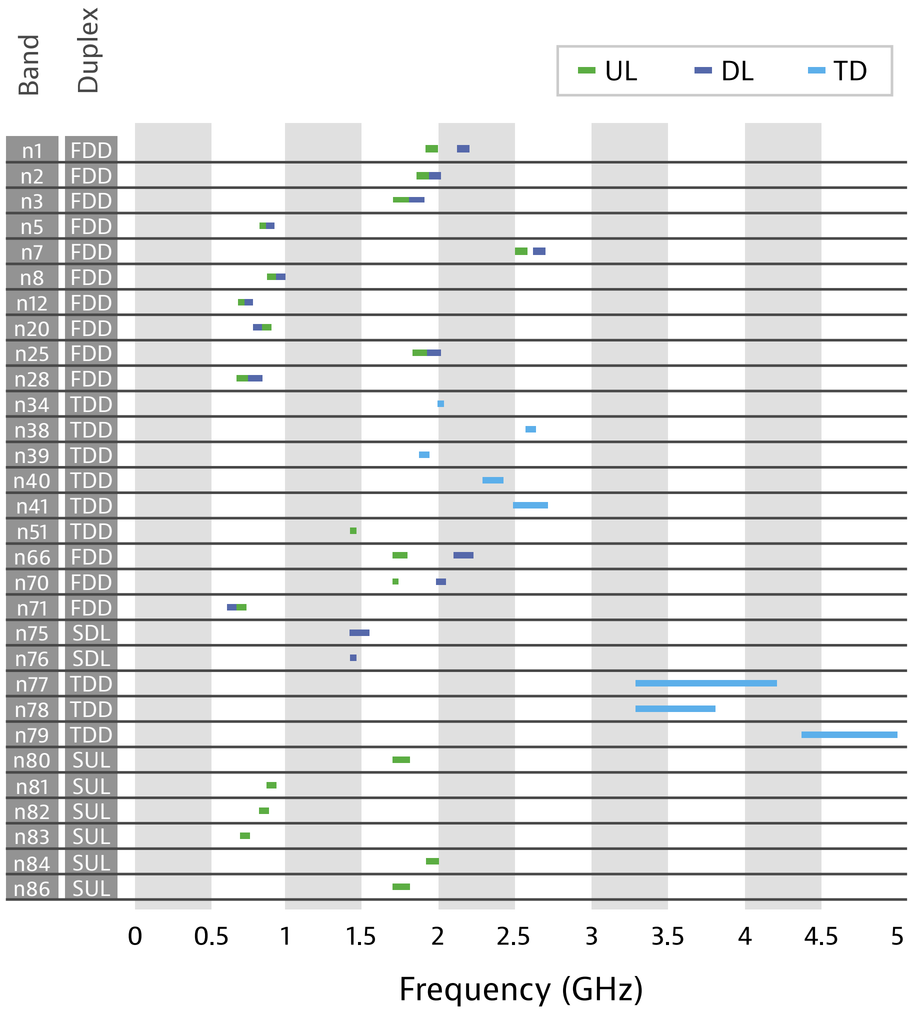


Figure 2.1 Snapshot of frequency bands defined for FR1 by 3GPP in release 15

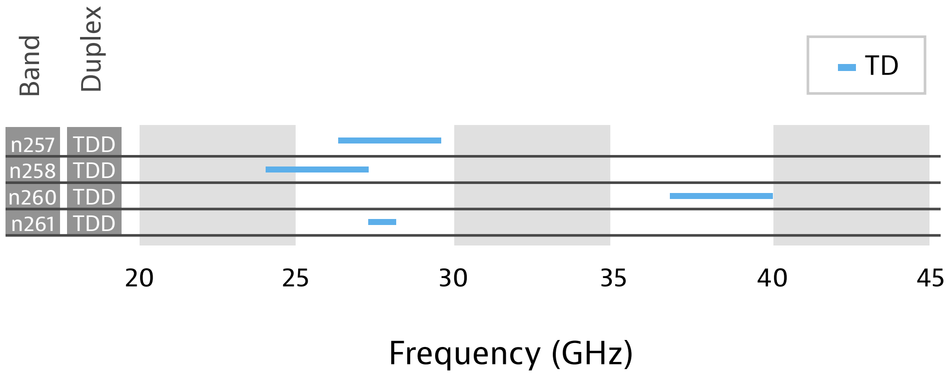


Figure 2.2 Snapshot of frequency bands defined for FR2 by 3GPP in release 15

There are also frequency bands defined as FR2, shown in Figure 2.2, occupying spectrum between 24.25 GHz and 52.6 GHz, although no bands above 40 GHz are defined in Release 15 of the standards.

This latter range is almost entirely in the mmWave range. Real-life deployments will generally depend on where the available spectrum is in the associated jurisdiction. The relatively high availability of mmWave spectrum together with technological solutions for the problems of transmission at these frequencies means that the spectrum strategy of 5G is a major cornerstone of the ability of the standards to deliver eMBB use cases.

2.3 5G Waveforms, Numerologies, and Frame Structure

The waveform choices for 5G were determined by balancing many factors. The resulting system had to be capable of delivering the massive data rates required by eMBB. This meant it had to be sufficiently compatible with massive MIMO. It also had to perform well over all the frequency bands under consideration for 5G. Transmitter and receiver complexity was also a consideration, even over large bandwidths.

The 5G waveform is based on OFDM. The downlink and uplink use Cyclic Prefix-Orthogonal Frequency Division Multiplexing (CP-OFDM) while the uplink (UL) can optionally use Discrete Fourier Transform-Spread-OFDM (DFT-S-OFDM). While CP-OFDM is inherited from LTE of which it was a cornerstone, DFT-S-OFDM is included in 5G because it improves uplink coverage. DFT-S-OFDM benefits from low peak to average power ratio (PAPR). This means that the RF amplifier can be simpler and require less power while avoiding the distortion associated with a large dynamic range.

OFDM has a high PAPR. The high dynamic range restricts the efficiency that the amplifier can achieve. It also depends on subcarriers that are stable and aligned with respect to each other in the frequency domain. Deviation from this will lead to some loss of orthogonality between subcarriers. Despite these disadvantages, the OFDM system has many significant advantages. Because subcarriers across a wide bandwidth can be selected dynamically for transmissions, it has resilience to frequency selective fading. For example, interference that is limited to part of the carrier bandwidth will only affect part of the transmission and in some cases can be avoided. Because the symbols in OFDM are relatively long, they are more resilient to inter-symbol interference than other systems with shorter symbols.

2.3.1 Symbols and Modulation

In 5G NR, data is transmitted in symbols, each of which carries one or more bits of information in a single tone. The number of bits conveyed is determined by the modulation scheme. The quadrature phase shift keying (QPSK) 16 quadrature amplitude modulation (16QAM), 64QAM, and 256QAM modulation schemes are supported by CP-OFDM for downlink (DL) and UL, and the DFT-S-OFDM in the UL. These modulation schemes convey two, four, six, and eight bits of information respectively. The inclusion of the 256QAM modulation scheme allowing eight bits to be transmitted per symbol in the best radio conditions, supports the eMBB service enabler.

The DFT-S-OFDM UL additionally supports the π/2-binary phase shift keying (BPSK) modulation scheme, which conveys a single bit of information for each symbol. The benefit of this modulation scheme is that the phase shift changes at each symbol transition, irrespective of whether the underlying conveyed bit is the same as or different from the previous bit. This makes it easier for the receiver to maintain phase lock and keep track of the boundaries between symbols. This can be particularly advantageous in complex propagation environments with delay spreading from multipath propagation. π/2-BPSK also benefits from a lower PAPR, which leads to increased efficiency of the power amplifier, particularly for lower data rates.

The supported modulation schemes are shown in Table 2.1 and the constellation diagrams illustrated in Figure 2.3.

Modulation scheme	Bits/symbol	DL	UL	
		CP-OFDM	CP-OFDM	DFT-S-OFDM
π/2-BPSK	1			Y
QPSK	2	Y	Y	Y
16QAM	4	Y	Y	Y
64QAM	6	Y	Y	Y
256QAM	8	Y	Y	Y

Table 2.1 Supported modulation schemes in 5GNR

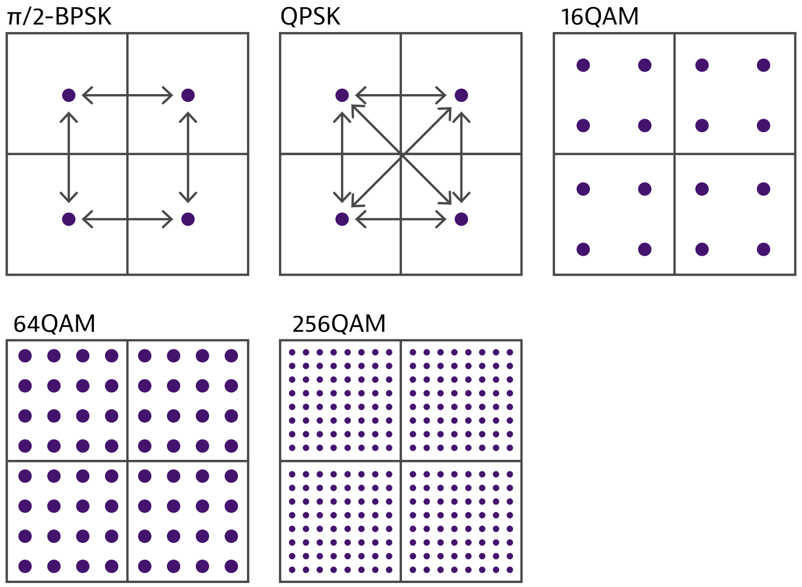


Figure 2.3 Constellation diagrams for supported modulation schemes in 5G NR

2.3.2 Subcarriers

OFDM systems pack many subcarriers close together in frequency. OFDM demodulators depend on fast Fourier transforms to separate out the various sub-carriers so that they can be demodulated. However, the different subcarriers are not orthogonal. If the lack of orthogonality is not addressed, it will impair the ability to demodulate the different subcarriers successfully. This orthogonality is at a maximum at multiples of λ_{scs} Hz from the subcarrier center frequency where λ_{scs} is the reciprocal of the time over which the symbol tone is transmitted, or symbol length for brevity. In the case of the 5G NR, the standard symbol length is 66.67 µs. The reciprocal of this is 15 KHz, which is the standard subcarrier spacing, identical to the subcarrier spacing for LTE.

The 66.67 µs symbol duration is preceded by a 4.7 µs cyclic prefix, identical to the approach for wideband transmission used in LTE. As in LTE, the purpose of the cyclic prefix is twofold. The main reason is to provide separation between adjacent symbols so that the transmission can be resilient to delay spread from multipath transmission. If delay spread does occur, then as long as it does not exceed the length of the cyclic prefix, it will not interfere with the next symbol. In practice, in 5G and in LTE, the cyclic prefix is not devoid of transmission. Rather, the end of the symbol about to be transmitted is replicated and transmitted in the cyclic prefix. In practice, this means that the transmitted tone lasts a little longer than it would otherwise. This has the beneficial side effect that if the delay spread is minimal, then the receiver can use the transmission in the cyclic prefix to support its effort to demodulate the symbol. This adds to the resilience of the receiver, lending support to the requirement for ultra-reliability.

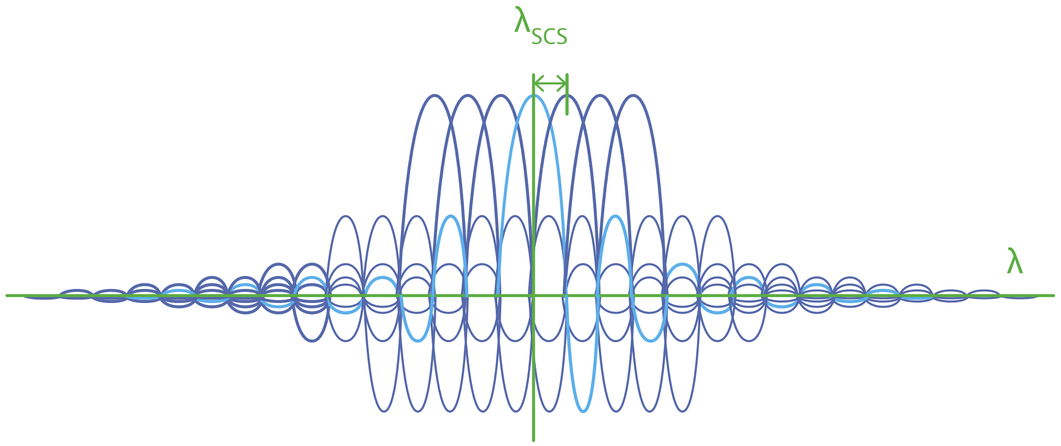


Figure 2.4 OFDM waveforms showing orthogonality between subcarriers with perfect waveforms.

Each subcarrier coincides with a minimum of inter-subcarrier interference.

2.3.3 Guard Bands and Spectral Efficiency

OFDM subcarriers cannot utilize the whole spectrum band. The power of the subcarriers will spread out around the center frequency and tail off. If this is ignored, then the power can leak out of the band in use whether licensed or unlicensed. OFDM, like many transmission schemes, requires a guard band at the edge of a carrier where no subcarriers are transmitted. While LTE was able to utilize ninety percent of the carrier for subcarriers, 5G does better than this, using techniques such as windowing and filtering to contain the transmissions. The guard band is a fixed overhead per carrier. The guard band must also be used at the edge of each carrier and eliminates otherwise useful spectrum from being used for subcarriers. Very large carriers are introduced in 5G NR. Carriers of up to 100 MHz are supported for FR1. In FR2 carriers of up to 400 MHz are supported as well. These carriers span large frequency ranges, meaning that the effect of the guard band overhead is reduced, and spectral efficiency is improved.

2.3.4 Flexible Subcarrier Spacing and Numerologies

The subcarrier spacing is chosen not only to defeat the interference between subcarriers but also to fix the problem arising from phase noise. This is caused by imperfections in the oscillators, meaning that the modulated tone is not perfect. This also means that the transmitted energy will be spread over an interval centered on the central frequency of the subcarrier with sidebands of phase noise interference. As the phase noise phenomenon increases, it can spread out over the frequency domain to the extent that it interferes with nearby subcarriers.

The problem of phase noise becomes more significant at higher frequencies and the width of the sideband interference increases. Therefore, phase noise becomes more serious in mmWave parts of the spectrum. If the 15 KHz channel spacing were maintained across all 5G carriers, phase noise would become significant and, coupled with the challenging propagation at higher frequencies, would result in a serious loss of capacity in mmWave.

To address the problem of phase noise, 5G NR deviates from its predecessors and introduces flexible subcarrier spacing. This technique allows the subcarrier spacing to be 15 KHz multiples of 2^{μ} where µ of 0, 1, 2, 3, or 4 results in subcarrier spacings of 15, 30, 60, 120, or 240 KHz. These are known as numerologies, with numerologies having a larger power of two being referred to as higher numerologies.

But increasing the subcarrier spacing brings a side effect. The symbol length must be reduced with higher numerologies owing to the need for the subcarrier spacing to be the reciprocal of the symbol length. If this were not the case, maximum orthogonality between subcarriers in the FFT as described above would not be maintained. The symbol length for a subcarrier spacing of 240 KHz is only 4.17 µs, which is 1/16th of the symbol length for 15 KHz subcarrier spacing. This appears to be a problem. As multipath propagation spreads the signal, the symbols can bleed into one another. At first glance, these much shorter symbols appear to be far more vulnerable to inter-symbol interference. In practice, the length of the symbol that can be used is dependent on the frequency band. As the frequency increases, the shorter symbols can be used. But this coincides with the frequencies that are less subject to inter-symbol interference. Delay spread becomes less significant at higher frequencies as propagation is predominantly line-of-sight and so the very shortest symbols can be used for very high frequencies.

The use of higher numerologies also has a secondary advantage. The mmWave carriers can span hundreds of MHz. This is demanding for the transmitter and receiver if the normal 15 KHz subcarrier spacing is maintained as larger Inverse Fast Fourier Transforms (IFFTs) must be used, leading to more complex devices. Spacing the subcarriers further apart and shortening the symbol length therefore results in resilience to phase noise while maintaining resilience to inter-symbol interference and the orthogonality between subcarriers in the receiver. It also demands less complexity in the transmitter and receiver for the widest carriers. Wider subcarrier spacing and the associated shorter symbols have implications for low latency communications. As we shall see, communication is broken up into self-contained units called slots, normally consisting of fourteen symbols. These slots become shorter in time as the numerology increases and the subcarrier spacing becomes larger, meaning that there are more frequent opportunities for scheduling data.

Some complexity arises with carrier bandwidth parts (BWP). These will be introduced fully later. Different BWPs can have different numerologies with varying subcarrier spacings and symbol lengths. If two BWPs adjacent in the frequency domain have different subcarrier spacing, the orthogonality between subcarriers in the receiver will be broken. This requires insertion of a guard band between the bandwidth parts, which impairs the efficient use of the spectrum. An alternative is filtering in the receivers which support BWP to maintain better spectral efficiency.

The 5G system must be able to support a wide range of environments including those with complex propagation and specifically those with high degrees of delay spread. As the complexity of the propagation grows, so must the length of the cyclic prefix to ensure that the receiver can reliably demodulate the symbol. Rather than set a cyclic prefix length long enough for every conceivable propagation environment that the system must support, 5G NR has the concept of the extended cyclic prefix for complex environments supporting the normal cyclic prefix for regular environments.

2.3.5 Frame Structure and Slots

The physical radio resources can be thought of as a set of subcarriers in the frequency domain and a set of opportunities for modulated symbols in the time domain. A single modulated symbol on one subcarrier is called a resource element. These resource elements are grouped together into logical structures that can be used for transmission and reception. As in LTE, there are ten subframes per frame and each subframe is broken down into a variable number of slots such that the number of slots per subframe is dependent on the numerology, as discussed above. Various numerology options are illustrated in Figure 2.5 for the lower numerologies and Figure 2.6 for the higher numerologies.

In the frequency domain, the resource elements are arranged in groups of twelve subcarriers that are called resource blocks. This is illustrated in Figure 2.7. Standard slots have fourteen symbols per slot, but mini slots can comprise seven, four, or two symbols. These mini slots are designed for low latency applications and allow transmissions to be rapidly acknowledged. Traffic pre-emption means that they can be inserted when required for low latency applications.

There are various engineering compromises in the choice of frame structure and how the resources are managed. The resources must be utilized efficiently. The management must be flexible and accommodate different mixes of DL and UL data demands. Ultra-Reliable Low Latency must be supported with data being scheduled when required without delay and acknowledged immediately.

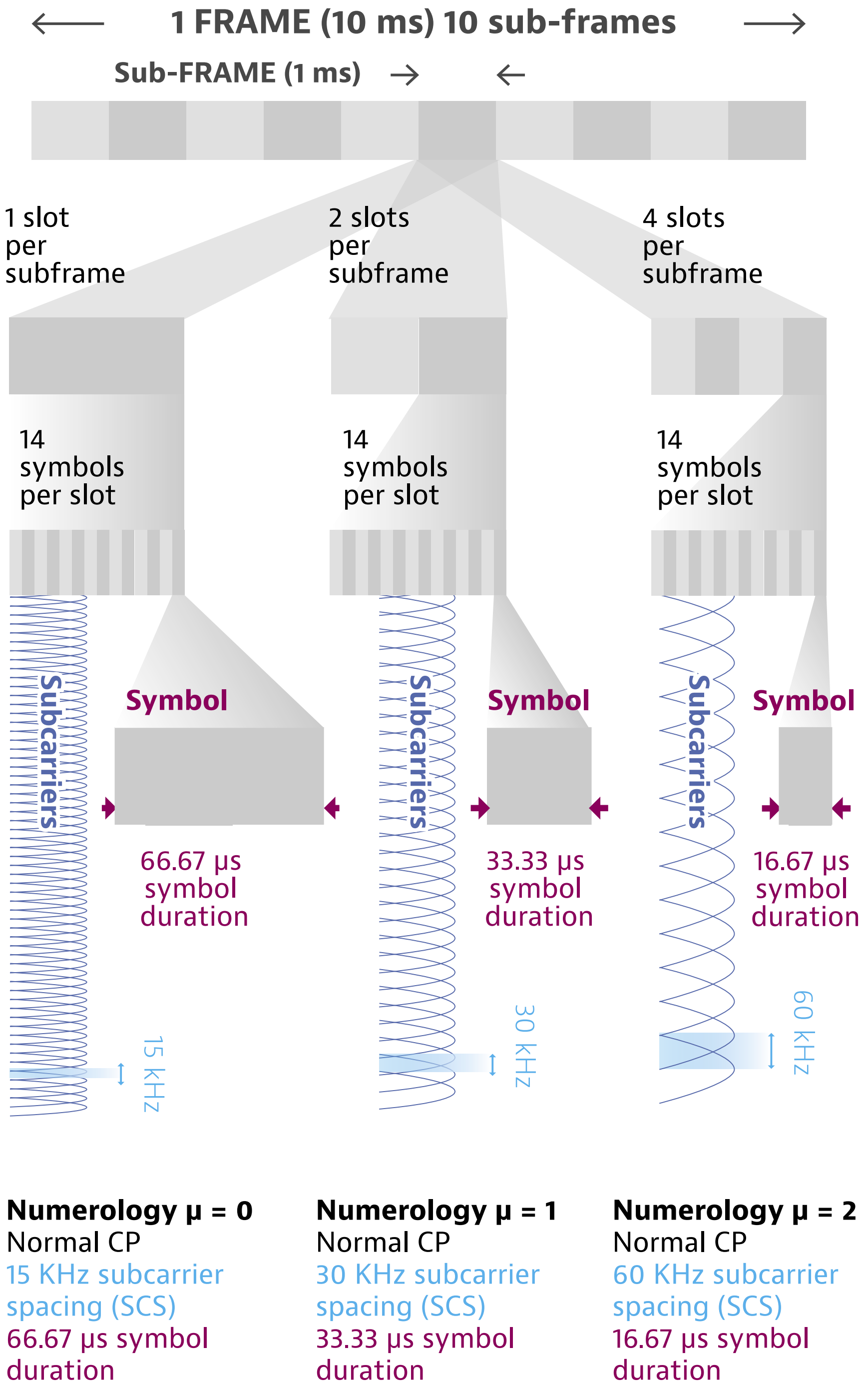


Figure 2.5 5G NR lower numerologies

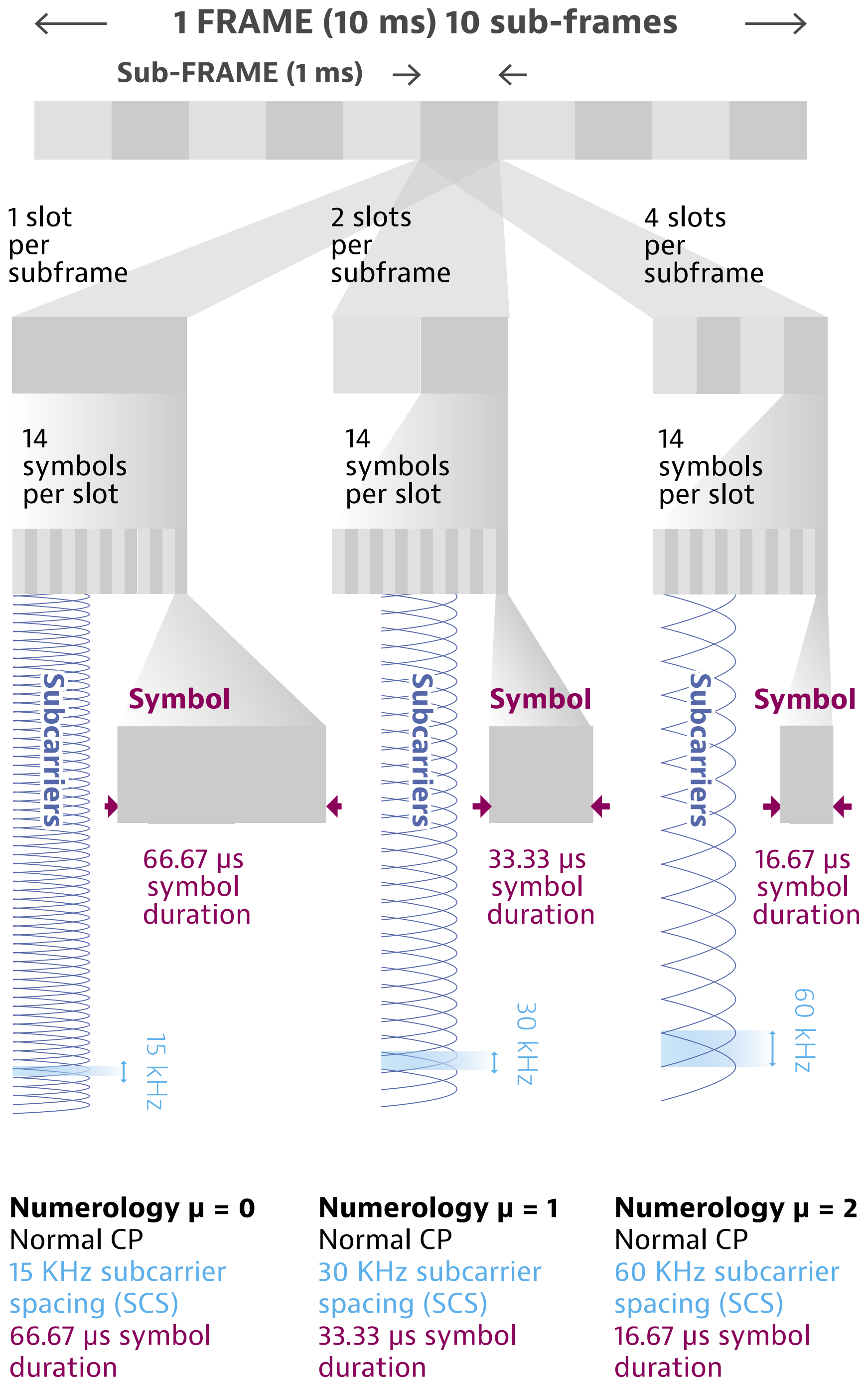


Figure 2.6 5G NR higher numerologies

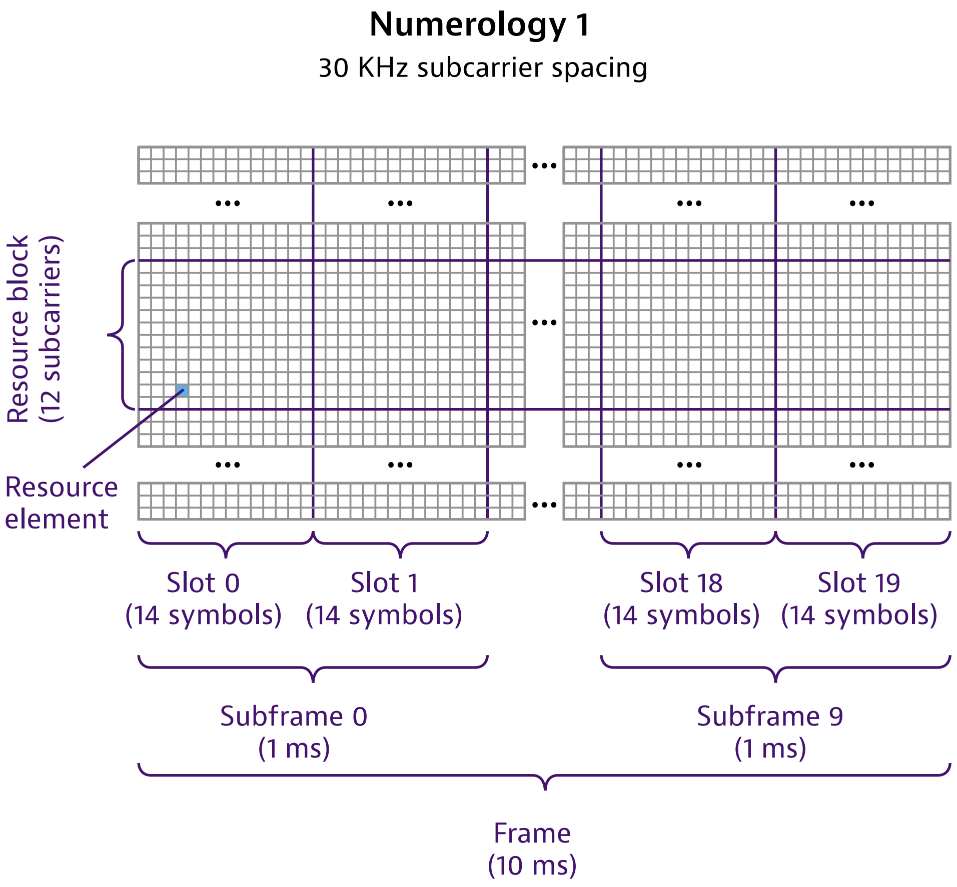


Figure 2.7 Resource grid, resource blocks and resource elements within frames for numerology 1

For example, network slicing is a key capability for the 5G network. This capability allows different classes of subscriber with wildly different quality of service requirements to be satisfied simultaneously with the same network. The frame must support network slicing and other mechanisms for separating users with different requirements such that the radio bandwidth can be controlled by different logical network functions simultaneously, potentially with different functional splits in operation.

Some 5G use cases require dynamic alternating demands on the downlink and the uplink. For example, there may be a predominance of eMBB during working hours and with massive Machine Type Communications (mMTC) devices synchronizing during the night. The eMBB usage would tend to have significantly more DL than UL. The mMTC devices would have the converse requirement. 5G allows for symbols in each subframe to be allocated flexibly to the DL or the UL depending on the need. This is applicable to the time division duplex (TDD) scheme, where the same carrier is shared for DL and UL at different times. Flexible allocation is performed using the Slot Format Indication (SFI). This can be statically allocated or even performed dynamically in order to respond immediately to changing relative needs of network slices. The flexible slot formats are shown in Table 2.2 and Table 2.3.

Format	Symbol number within slot													
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	D	D	D	D	D	D	D	D	D	D	D	D	D	D
1	U	U	U	U	U	U	U	U	U	U	U	U	U	U
2	F	F	F	F	F	F	F	F	F	F	F	F	F	F
3	D	D	D	D	D	D	D	D	D	D	D	D	D	F
4	D	D	D	D	D	D	D	D	D	D	D	D	F	F
5	D	D	D	D	D	D	D	D	D	D	D	F	F	F
6	D	D	D	D	D	D	D	D	D	D	F	F	F	F
7	D	D	D	D	D	D	D	D	D	F	F	F	F	F
8	F	F	F	F	F	F	F	F	F	F	F	F	F	U
9	F	F	F	F	F	F	F	F	F	F	F	F	U	U
10	F	U	U	U	U	U	U	U	U	U	U	U	U	U
11	F	F	U	U	U	U	U	U	U	U	U	U	U	U
12	F	F	F	U	U	U	U	U	U	U	U	U	U	U
13	F	F	F	F	U	U	U	U	U	U	U	U	U	U
14	F	F	F	F	F	U	U	U	U	U	U	U	U	U
15	F	F	F	F	F	F	U	U	U	U	U	U	U	U
16	D	F	F	F	F	F	F	F	F	F	F	F	F	F
17	D	D	F	F	F	F	F	F	F	F	F	F	F	F
18	D	D	D	F	F	F	F	F	F	F	F	F	F	F
19	D	F	F	F	F	F	F	F	F	F	F	F	F	U
20	D	D	F	F	F	F	F	F	F	F	F	F	F	U
21	D	D	D	F	F	F	F	F	F	F	F	F	F	U
22	D	F	F	F	F	F	F	F	F	F	F	F	U	U
23	D	D	F	F	F	F	F	F	F	F	F	F	U	U
24	D	D	D	F	F	F	F	F	F	F	F	F	U	U
25	D	F	F	F	F	F	F	F	F	F	F	U	U	U
26	D	D	F	F	F	F	F	F	F	F	F	U	U	U
27	D	D	D	F	F	F	F	F	F	F	F	U	U	U

Table 2.2: Flexible slot formats showing which slots are for DL, UL, or flexible (part 1)

Format	Symbol number within slot													
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
28	D	D	D	D	D	D	D	D	D	D	D	D	F	U
29	D	D	D	D	D	D	D	D	D	D	D	F	F	U
30	D	D	D	D	D	D	D	D	D	D	F	F	F	U
31	D	D	D	D	D	D	D	D	D	D	D	F	U	U
32	D	D	D	D	D	D	D	D	D	D	F	F	U	U
33	D	D	D	D	D	D	D	D	D	F	F	F	U	U
34	D	F	U	U	U	U	U	U	U	U	U	U	U	U
35	D	D	F	U	U	U	U	U	U	U	U	U	U	U
36	D	D	D	F	U	U	U	U	U	U	U	U	U	U
37	D	F	F	U	U	U	U	U	U	U	U	U	U	U
38	D	D	F	F	U	U	U	U	U	U	U	U	U	U
39	D	D	D	F	F	U	U	U	U	U	U	U	U	U
40	D	F	F	F	U	U	U	U	U	U	U	U	U	U
41	D	D	F	F	F	U	U	U	U	U	U	U	U	U
42	D	D	D	F	F	F	U	U	U	U	U	U	U	U
43	D	D	D	D	D	D	D	D	D	F	F	F	F	U
44	D	D	D	D	D	D	F	F	F	F	F	F	U	U
45	D	D	D	D	D	D	F	F	U	U	U	U	U	U
46	D	D	D	D	D	F	U	D	D	D	D	D	F	U
47	D	D	F	U	U	U	U	D	D	F	U	U	U	U
48	D	F	U	U	U	U	U	D	F	U	U	U	U	U
49	D	D	D	D	F	F	U	D	D	D	D	F	F	U
50	D	D	F	F	U	U	U	D	D	F	F	U	U	U
51	D	F	F	U	U	U	U	D	F	F	U	U	U	U
52	D	F	F	F	F	F	U	D	F	F	F	F	F	U
53	D	D	F	F	F	F	U	D	D	F	F	F	F	U
54	F	F	F	F	F	F	F	D	D	D	D	D	D	D
55	D	D	F	F	F	U	U	U	D	D	D	D	D	D

Table 2.3: Flexible slot formats showing which slots are for DL, UL, or flexible (part 2)

2.3.6 Bandwidth Parts

Carrier bandwidth parts (BWP) mentioned above allow for different parts of the bandwidth of a carrier to have different numerologies. This allows for different use cases to be supported simultaneously. Network slicing is also facilitated by this arrangement. It also permits high-complexity devices to enjoy the elevated throughputs of low subcarrier spacing while low-complexity devices can achieve a service with high subcarrier spacing. A single user equipment (UE) can have up to four bandwidth parts allocated to it on DL and UL with one active at any given moment. This allows for multiple use cases to be achieved by a single device, or for the device to be on multiple network slices.

2.3.7 Blank Slots

Not every slot needs to be schedulable. With a view to forward compatibility, blank slots can be reserved. This can allow a mix of 3GPP releases to operate in the same carrier. The new architectural options and network slicing means that this is entirely possible as different virtualized network functions could be creating the transmissions for different parts of the carrier for different network slices. The reservation of blank slots removes the requirement for all to be on the same release, allowing the operators to invest early with the knowledge that early investment will not become quickly obsolete by lack of forward compatibility.

2.4 Channel Coding

Channel coding must be able to deal with the massive data rates of 5G in such a way that it is efficient. In other words, it must be able to achieve a good throughput for a given coding rate (level of protection). It must also be able to do this with a sufficiently low complexity encoder/decoder and do this with low latency. The Multi-Edge Low-Density Parity-Check Code (ME-LDPC) coding scheme for eMBB data can achieve these goals and is capable of throughput rates that are significantly better than the turbo codes used in LTE, especially at higher coding rates. ME-LDPC also benefit from parallel decoding in hardware, which results in higher throughput.

Many control channels use the CRC-Aided Polar (CA-Polar) channel coding. These benefit from having low algorithmic complexity. They also have no error floor, which is a phenomenon in other channel coders where there is a plateau in the improvement in error rate as the signal-to-noise ratio improves.

2.5 NR Medium Access Control (MAC) layer and HARQ for flexibly balancing latency, reliability and energy efficiency requirements

The MAC layer in a wireless communication system provides data transfer and radio resource allocation services to the upper layers and provides data transfer, signaling of Hybrid Automatic Repeat Request (HARQ) feedback, signaling of scheduling request and measurements services, for example channel quality indication (CQI) to the physical layer.

In 5G NR the different requirements of the use case scenarios eMMB, URLLC and mMTC call for a flexible and configurable approach to the MAC. In the sections below we introduce the HARQ function which may be considered the heart of the MAC. We then analyze the latency contributions of the different signaling and processing functions within the MAC layer and describe how these may be configured in a flexible manner to meet diverse use case scenarios.

It should be noted that, as MAC is a software and signaling control layer, these enhancements may be ported back to LTE, albeit with gains limited by the constraints of the air interface. For example, short Transmission Time Interval (TTI) has been introduced to LTE.

5G seeks to make operation with an FDD and a TDD type frame as similar as possible, consequently, both air interface frame structures can be configured to support the range of use cases. However, FDD is more able to support a mix of different configurations simultaneously as low latency operation in a TDD frame requires frequent UL/DL switch points which is not compatible with use of different switch point configurations in the same cell or in different cells.

2.5.1 An Introduction to HARQ: Reliably sending a signal through a noisy channel

Sending signals over the mobile radio propagation channel is messy and subject to failure. The transmitted signal is scattered and obstructed or shadowed by all manner of objects including buildings, street furniture and people, which results in a received signal that consists of multiple attenuated and reflected copies of that transmitted. Other mobiles in the same cell or different cells, or even attached to different systems, transmit at the same time causing interference. Not to forget interference and inter modulation from other base stations in the same channel and base stations in adjacent channels. Moreover, background noise arises from thermal fluctuations in the circuitry of the receiver, impulse noise from car ignitions, cosmic background radiation and many other sources including, in at least one case, electrical noise from a malfunctioning beer fridge. Consequently, the received signal is composed of multiple attenuated and phase shifted copies of the transmitted signal which add in and out of phase as the mobile moves, or as scatterers in the radio environment move. This results in fast or Rayleigh fading, with imposition of slower shadow fading caused by obstructions and all underlaid with background interference and noise.

An analogy is trying to communicate over the hubbub of a school dining hall; conversing over any distance in such a place relies on repetition. That is, one party indicates to the other a negative acknowledgement (NACK) when they don't get the message and the other party tries again until they receive a positive acknowledgement (ACK). This strategy, Automatic Repeat Request (ARQ) is employed in radio systems.

However, it is possible to do better than simple repetition using a strategy called Hybrid ARQ (HARQ). In this case, the listening party stores a representation of the received signal they failed to decode when indicating a NACK. The counter party sends a new version of the message and the recipient combines this with one or more stored version creating a composite version more resilient to channel impairments and the process is repeated until successful ACK, or the maximum number of attempts is reached and a higher layer ARQ starts the process again or declares failure.

Different versions of HARQ are defined in the literature. Type I HARQ always sends the same sequence of data on each re-transmission; erroneous data may be discarded, or the soft decision data may be stored and combined with the subsequent transmission. This exploits time diversity between the subsequent transmissions. Type II HARQ sends a different subset of the same coded bits on each re-transmission with the effect, after combining, of increasing the effective coding rate (or redundancy) with each subsequent transmission. For example, if the 1st transmission has coding rate 1, the 1st and 2nd transmission together will have coding rate 1/2 and combining all three will give coding rate 1/3. Type III HARQ is like type II, but each transmission is individually decodable, so decoding is possible even if one of the transmissions is lost. LTE and 5G NR use type II HARQ.

HARQ is very beneficial for the efficient sending of digital data over the mobile radio channel as rather than having to ensure an acceptable frame error rate, say 1 in 1000 ($1x10^{-3}$), with a single transmission the system can tolerate an error rate of 10% or more knowing that subsequent re-transmissions will ensure a low frame error rate for those that didn't make it the first time. For example, 3 repetitions, if required, will take the error rate down to $1x10^{-4}$. In this way the system can strike a balance between the coding rate and frame error rate to optimally use the available downlink power (albeit with limited granularity).

There is a trade-off between power efficiency and latency. Latency is lowest if a message is delivered in a single attempt, but this necessitates use of very robust modulation and/or higher power which compromises power efficiency and the overall capacity of the system. Additionally, when success is not assured with a single transmission, one or more retransmissions will be required. The speed at which transmission, feedback and subsequent retransmission can be accomplished is determined by the round-trip time (RTT) between the UE and the gNB. RTT is affected by processing time in the UE and the eNB and by the configuration and time duration of the messages used to send the data and exchange feedback.

At the end of this section we illustrate how HARQ has evolved from LTE to NR. However, we first examine the other aspects of UE and gNB processing, MAC signaling and frame structure that have been made more configurable to facilitate this trade-off between latency, reliability and power efficiency.

2.5.2 Creating a configurable 5G NR MAC layer to balance reliability, latency and power efficiency

The cornerstone of 5G NR, as set generally throughout this book, is flexibility and configurability. Not surprisingly, 5G NR has taken many steps to make the air interface suitably configurable to allow a dynamic trade-off between reliability, latency and power efficiency.

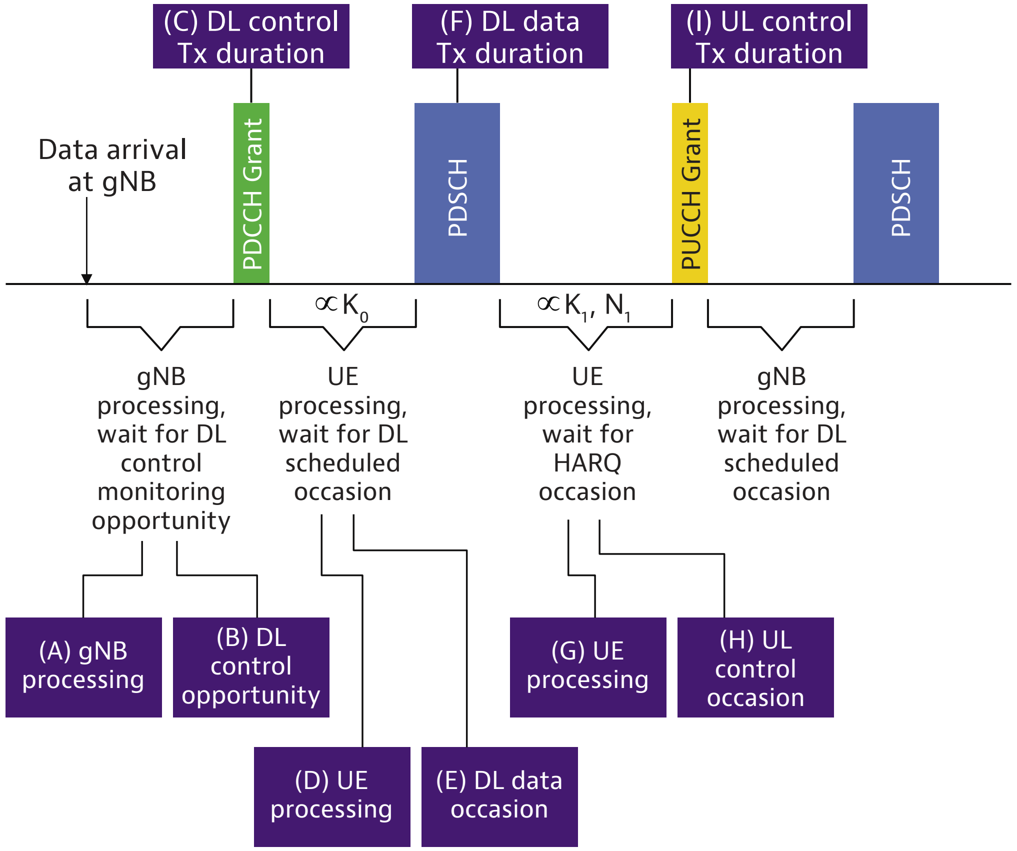

(C) DL control Tx duration
(F) DL data Tx duration
(I) UL control Tx duration
Data arrival at gNB
PDCCH Grant
PDSCH
PUCCH Grant
PDSCH
∝K0
∝K1, N1
gNB processing, wait for DL control monitoring opportunity
UE processing, wait for DL scheduled occasion
UE processing, wait for HARQ occasion
gNB processing, wait for DL scheduled occasion
(A) gNB processing
(B) DL control opportunity
(D) UE processing
(E) DL data occasion
(G) UE processing
(H) UL control occasion

Figure 2.8 Components of DL latency

Figure 2.8, inspired by "New services & applications with 5G Ultra-Reliable Low Latency Communications, 5G Americas, November 2018", illustrates the key factors that determine air interface latency. The Figure assumes semi-persistent scheduling (SPS) is used so a single grant may be used for both data transmissions in the Figure; it is up to the MAC as to whether new data or re-transmitted data is sent. Many of the issues addressed are common for the UL and DL, but UL specific issues are dealt with below.

Some of the configuration options that 5G NR introduces to minimize latency are deleterious to other aspects of performance such as reliability and power efficiency. Therefore, the approach taken in the following sections is to present, with respect to the highlighted callouts in the Figure, how choices may be made to enhance or optimally configure the air interface to trade-off latency, reliability and energy efficiency to best suit the use-case family at hand.

2.5.3 Exploiting gNB and UE processing time capability

LTE HARQ process imposed a hard-wired delay of 3ms to allow for processing and signaling opportunities in the eNB and UE which created a minimum RTT of 8ms. Albeit, backporting of the short TTI feature to LTE allows RTT to be reduced to 2ms. 5G NR removes this hard-wired restriction allowing device capability and channel configuration to be exploited to reduce latency.

The processing delay in the gNB is not defined in the Standard and may be as short as implementation allows. The value is accounted for in the HARQ scheduling processes in both UL and DL.

As is clear from Figure 2.8, UE and gNB processing times are elements in MAC latency. Thus, it is necessary for the gNB to know what the UE processing capability is to make appropriate scheduling decisions. Consequently, the UE capability is standardized. Two levels of UE capability are specified in Rel-15. The minimum required processing delay is dependent on the configuration of the channels that the UE receives and transmits on as well as the reported capability; it is calculated using look-up tables referenced from the parameters described in Table 2.4.

Parameter	Description	Comment
K_0	Delay in slots from PDCCH allocation to PDSCH data reception	3GPP 38.214 v15, clause 5.3, used to determine the minimum delay in slots from a PDSCH DCI allocating DL resource to a UE and reception of DL data by the UE
K_1	Delay from PDSCH to HARQ-ACK	3GPP 38.214 v15, clause 5.1, delay in slots from PDSCH reception by a UE to its sending of the associated HARQ-ACK. $K_1 = 0$ defines a self-contained bi-directional slot
K_2	Delay from UL grant to UL transmission	3GPP 38.214 v15, clause 6.1, delay in slots from reception of an UL Grant by a UE to its sending of UE PUSCH data transmission
N_1	UE PDSCH processing time	3GPP 38.214 v15, clause 5.3, UE processing delay, in OFDM symbols, from end of PDSCH reception to earliest possible HARQ-ACK transmission. Used to determine minimum value of K_1
N_2	UE PUSCH preparation time	3GPP 38.214, clause 6.4, UE preparation delay in OFDM symbols from end of the PDCCH scheduling UL to the earliest possible UL PUSCH transmission. Used to determine minimum value of K_2
processingType2 enabled	UE processing capability	3GPP 38.331, reported to gNB used for table look-up to determine N_1, N_2 in 38.214

Table 2.4 parameters reflecting UE capability to respond to a grant or produce a HARQ-ACK

2.5.4 Configuring PDSCH to trade-off latency, reliability and power efficiency

5G NR provides the capability to configure the data channel to minimize processing delay, affecting the latencies in Figure 2.8, where the gNB takes account of the UE minimum processing delay through parameters K0 and K1. Alternative configurations are also supported where performance in reliability or power efficiency are of greater importance than latency.

Demodulation reference signal (DMRS): LTE had both cell reference signals and UE specific DMRS. However, to support power efficiency cell reference signals are removed in 5G NR which makes the air interface less chatty and more "lean."

5G NR supports front loading of the DMRS so they arrive at the start of the transmission which allows the receiver to begin channel estimation immediately on reception which supports low latency.

However, this configuration compromises performance when the channel is changing rapidly, for example with high velocity UE, and reduces time diversity in the channel estimate which may compromise reliability. Therefore, additional DMRS may be configured across the channel.

Removal of time domain interleaving: In LTE data bits are interleaved in both time and frequency domains. Time domain interleaving provides time diversity which improves resilience against time varying fading and interference. In 5G NR interleaving is only done across the frequency domain which allows the receiver to perform de-interleaving on a symbol-by-symbol basis without having to wait for all the interleaved symbols to be received.

Frequency-first mapping: Additionally, 5G NR data bits are mapped to resource elements following a frequency-first mapping across the active Bandwidth Part (BWP) that further supports the ability to perform symbol-by-symbol processing at the receiver.

Polar channel coding: As discussed elsewhere in this Chapter, 5G NR replaces the Turbo coding employed by LTE with LDPC coding which is more amenable to parallel processing implementations that facilitate lower latency processing in both gNB and UE.

2.5.5 Configuring PDCCH monitoring to trade-off latency, reliability, power efficiency

The downlink control channel PDCCH sends the downlink control information (DCI) to schedule UL and DL transmission. As illustrated in Figure 2.8, an element of latency is the time that the eNB must wait for an opportunity to send the DCI to the UE. 5G NR uses several strategies to optimize this wait-time.

Defining frequent PDCCH opportunities that the mobile must monitor minimizes wait time. However, mobile battery life is one of the most precious resources in a mobile communication system as perhaps the most compelling selling proposition for a mobile device is being able to be in contact and contactable at any time. However, for the longest time, mobiles are in stand-by mode monitoring to see if there is any data for it to receive or to send. Thus, power consumption in stand-by mode is critical. Consequently, there is a direct trade-off between the frequency of the opportunities that the mobile is required to monitor and battery consumption.

5G NR can be configured to create PDCCH monitoring opportunities every few symbols, for example with mini-slot structures. Such a configuration clearly supports the low latency needed by some kinds of URLLC service. On the other hand, PDCCH opportunities may be configured less frequently and also monitored sporadically using a discontinuous receive (DRX) process that minimizes UE battery consumption at the expense of latency.

For the lowest latency use cases, the duration of the PDCCH (As illustrated by C in Figure 2.8) is important and it may be sent in a control-resource set (CORESET) spanning just a single symbol period. However, other use cases require a highly reliable PDCCH, and care should always be taken that the control channels are at least as reliable as the data channels that they seek to control. Correspondingly, in 5G NR, as with LTE, the resources used to send PDCCH may be aggregated. For example, rather than using a single control channel element (CCE) at aggregation level 1 to send the PDCCH, the message may be sent at aggregation level 2 using 2 CCEs and so on. In 5G NR aggregation levels up to 16 CCEs are supported.

2.5.6 Configuring PDSCH duration to trade-off latency, reliability, power efficiency

Figure 2.8 F, PDSCH may be flexibly configured with mini-slot configurations allowing durations as low as two symbols in the DL to meet tight latency constraints where transmission duration is a significant issue. Relatively high bandwidth may still be achieved by allocating resources across the available bandwidth. However, as discussed above, such an approach achieves good latency performance at the expense of robustness.

For less delay-sensitive use cases, PDSCH may be allocated across the remainder of the 14-symbol slots that NR supports. In addition, the PDSCH may be repeated, adding time diversity to the process. The number of repetitions is set by the PUSCH-Aggregation Factor and is monitored by a single HARQ process. This is touched on again at the end of this Section where several examples of NR HARQ are presented.

Although the PUSCH is not illustrated in Figure 2.8, it is sufficient here to note that the flexibility of the mapping and the use of repetition defined by a PUSCH-Aggregation Factor with respect to latency/reliability/power efficiency apply similarly for PUSCH.

2.5.7 Configuring PUCCH duration to trade-off latency, reliability, power efficiency

The Physical Uplink Control Channel (PUCCH) (I in Figure 2.8) is used to carry Uplink Control Information (UCI) including HARQ feedback and Channel State Information (CSI) as well Scheduling Request (SR).

In LTE the location, duration and timing are fixed. However, in 5G NR PUCCH may be flexibly configured in time, frequency and duration. PUCCH may be mapped to between 1 and 14 symbols and may use time or frequency multiplexing. The configuration options allow a trade-off between latency and robustness to enable the supported use case.

2.5.8 Configuring PUSCH duration to trade-off latency, reliability, power efficiency

Figure 2.9 shows the components of latency for an exemplary of UL data flow to illustrate how the elements of the NR MAC layer can be configured to manage latency with respect to reliability and power efficiency for the UL.

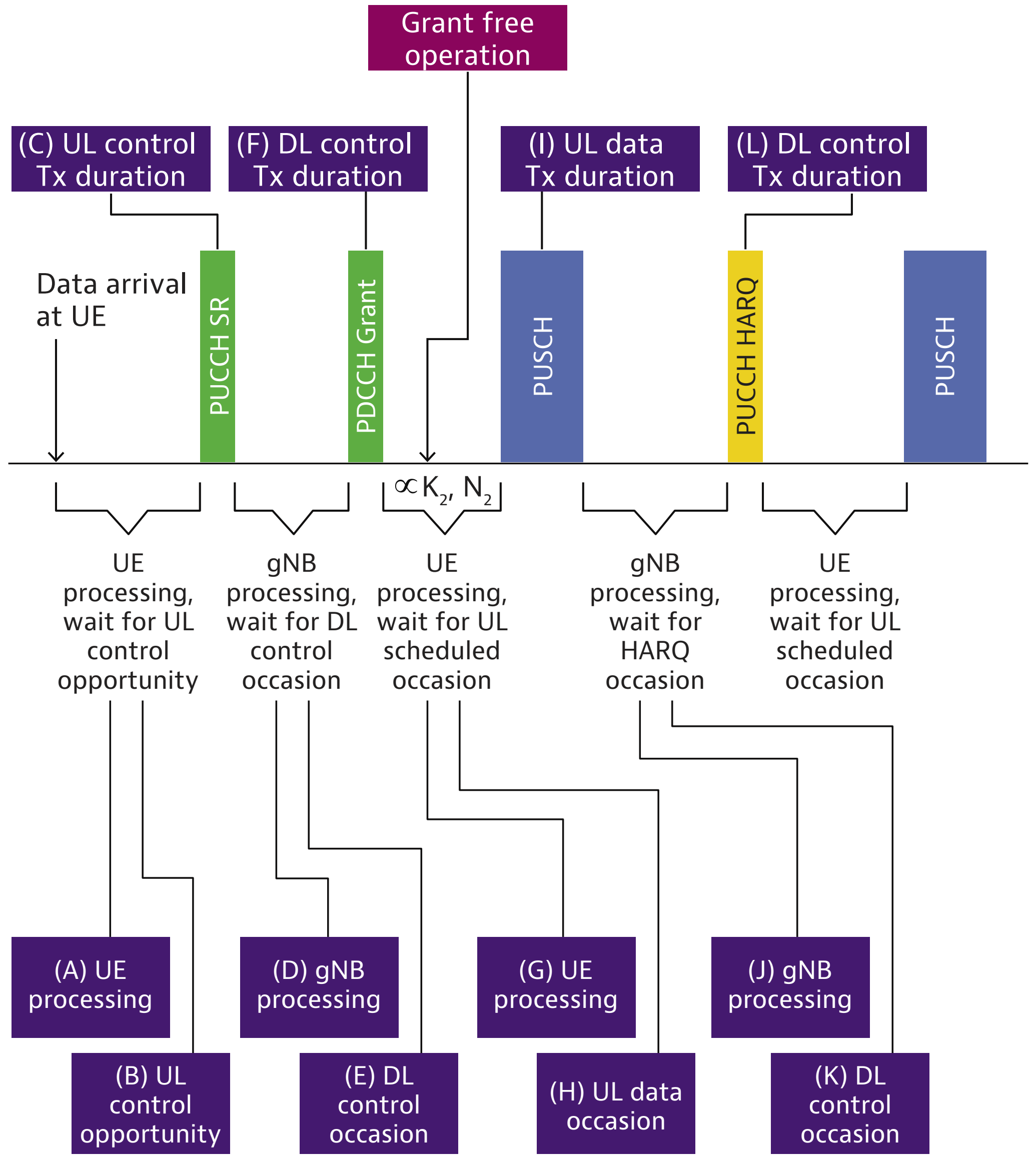

Grant free operation
(C) UL control Tx duration
(F) DL control Tx duration
(I) UL data Tx duration
(L) DL control Tx duration
Data arrival at UE
PUCCH SR
PDCCH Grant
PUSCH
PUCCH HARQ
PUSCH
∝K2, N2
UE processing, wait for UL control opportunity
gNB processing, wait for DL control occasion
UE processing, wait for UL scheduled occasion
gNB processing, wait for HARQ occasion
UE processing, wait for UL scheduled occasion
(A) UE processing
(D) gNB processing
(G) UE processing
(J) gNB processing
(B) UL control opportunity
(E) DL control occasion
(H) UL data occasion
(K) DL control occasion

Figure 2.9 Components of UL latency

There is a degree of similarity between the components of latency in DL and UL with similar configuration optimizations that may be made to trade-off latency, reliability and power efficiency. Consequently, similar types of optimization may be made for PUCCH and PDCCH frequency of opportunities/monitoring occasions and duration and PUSCH duration. For example, sending messages in a single symbol or spanning multiple symbol periods. And, as noted above, PUSCH also supports aggregation/repetition to further increase resilience.

2.5.8.1 Grant-free operation for UL transmission

A significant difference between DL and UL operation, which can be seen comparing Figure 2.8 and Figure 2.9, is the necessity to send a scheduling request (SR) to the gNB to request UL resources. And, following the SR, there is a wait to receive a PDCCH with a DCI indicating grant of UL resources.

The most extreme strategy for reduction of latency associated with uplink data transmission is to use grant-free operation which eliminates the need for SR and waiting to receive a grant, which eliminates step A through F in Figure 2.9. The UE will still have to process the data on arrival and wait for the pre-configured grant opportunity. A pre-configured grant is defined, for example, by using broadcast system information.

2.5.8.2 Pre-emption of UL and DL resources

The resource reservation associated with grant-free operation reduces efficiency as resources are ear-marked for use even if they are not used. 5G NR ameliorates this effect by allowing use of the ear-marked resources while reserving the right to pre-empt them. The gNB may indicate that on-going UL resources will be pre-empted allowing UEs to suspend transmission, reducing power consumption and interference.

Pre-emption is also possible in the downlink. UEs whose resources were pre-empted are informed in the following TTI allowing them to discard, rather than decode and store for use in incremental redundancy, any data apparently received during the pre-emption period. The gNB may then send the portion of the data that was not sent due to pre-emption. This strategy is suitable for the most latency sensitive use cases. Additionally, it is beneficial for UE power consumption as it eliminates at least one set of messages that the mobile would otherwise need to monitor frequently.

2.5.8.3 Code Block Group-based re-transmission

In LTE transmission and re-transmission is on a Transport Block (TB) basis. However, in cases where a TB is quite long, and errors occur only in a small portion, this leads to inefficient operation. 5G NR introduces the idea of Code Block Group (CBG) based re-transmission which, at the expense of a little extra signaling, allows large TB to be sent, but only the corrupted CBG re-transmitted.

CBG based re-transmission is particularly suited to the operation of eMBB, which demands large TBs, when combined with URLLC service employing pre-emption as only the CBG affected by the pre-emption need be sent.

2.5.9 Review of NR HARQ v LTE HARQ (with some examples)

This Section sets out the operation of LTE HARQ as a base-line, but things have changed with 5G NR and examples of tight sub-TTI latency and configurations with PDSCH or PUSCH aggregation for long range transmission are illustrated to show how 5G NR may flexibly configure the MAC and air interface to produce a variety of outcomes which result in different latency, reliability and, power efficiency trade-offs.

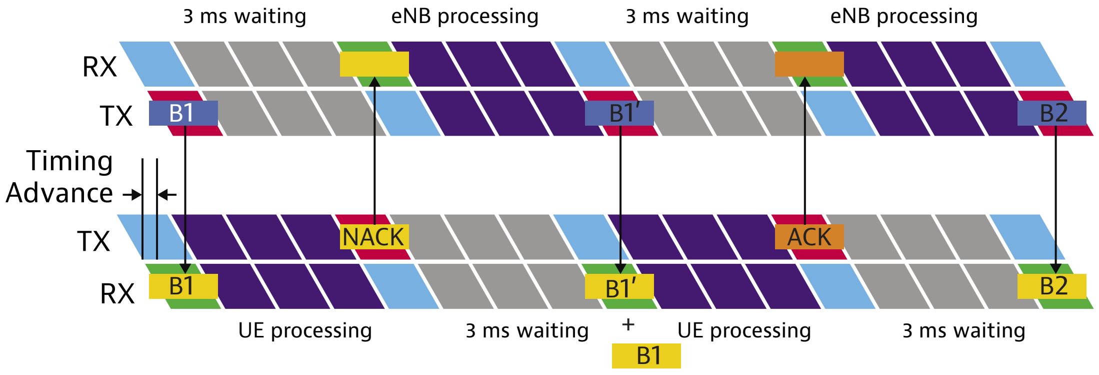


Figure 2.10 LTE FDD HARQ operation DL example

Figure 2.10 illustrates the operation of HARQ using the example of LTE FDD downlink data transmission. The eNB sends a data block B1 to the UE which arrives propagation delay (pd) later at the UE. The UE has a fixed timing budget of 3ms (3 timeslots) less timing advance (TA) to process the block and determine if it were correctly received. In the Figure B1 is not correctly decoded on the first transmission. The result of the decoding event is stored and a HARQ NACK is sent to the eNB at a time 3ms – TA later. The eNB has a time budget of 3ms (3 time slots) to process the UE response and, in case of HARQ NACK, to generate a second version of the data block B1′ with a different puncturing to support incremental redundancy. The second transmission is possible in the 8th timeslot after the initial transmission. Consequently, the HARQ round trip time (RTT) is 8 timeslots which is equal to 8ms for the LTE frame structure (without short TTI). In the Figure the UE is able to combine the initial decoding attempt of B1 with that of B1′ and correctly decode the data. Consequently, a HARQ ACK is sent to the eNB and in response the eNB send next data B2 block after the appropriate processing time. In LTE, there is a fixed processing/waiting time built into the HARQ process for both UL and DL. In addition, the UL uses a synchronous HARQ process that enforces a fixed timing arrangement between a downlink transmission and its corresponding HARQ-ACK. This reduces signaling at the expense of losing some scheduling flexibility. As described above, LTE HARQ has an RTT of 8 time slots. Consequently, to avoid gaps in transmission while waiting for a data block to be ACKed LTE supports 8-interlaced HARQ processes.

5G NR allows greater flexibility and may be configured where transmission, acknowledgement and retransmission can be accomplished within a TTI. However, this is not always the preferred action, particularly if there is an additional latency that the air interface must deal with, such as in the case of non-terrestrial access that is being standardized in Rel-16. Consequently, LTE Rel-15 allows up to 16 simultaneous HARQ processes.

Figure 2.11 provides 3 exemplary configurations. A is a configuration suitable for low latency use cases such as URLLC with transmissions confined to durations of a few symbols. Reliability in this configuration relies on the diversity provided by a wide bandwidth in the frequency domain. The configuration may use a pre-configured grant to eliminate the SR/grant delay. The resources reserved for the pre-configured grant may be shared with other applications pre-empting when necessary. For UL transmission, the gNB may indicate to mobiles' whose resources will be pre-empted allowing them to suspend transmission, and in the DL, the gNB may indicate in the following TTI that part of the earlier transmission has been pre-empted and should be discarded. B is a configuration suited to high reliability or low signal level configurations that uses an aggregation level of 4 for the PDSCH at the expense of lower efficiency as HARQ feedback is only provided after all re-transmissions are received, while C is a configuration for similar requirements with aggregation level of 3 for the PUSCH.

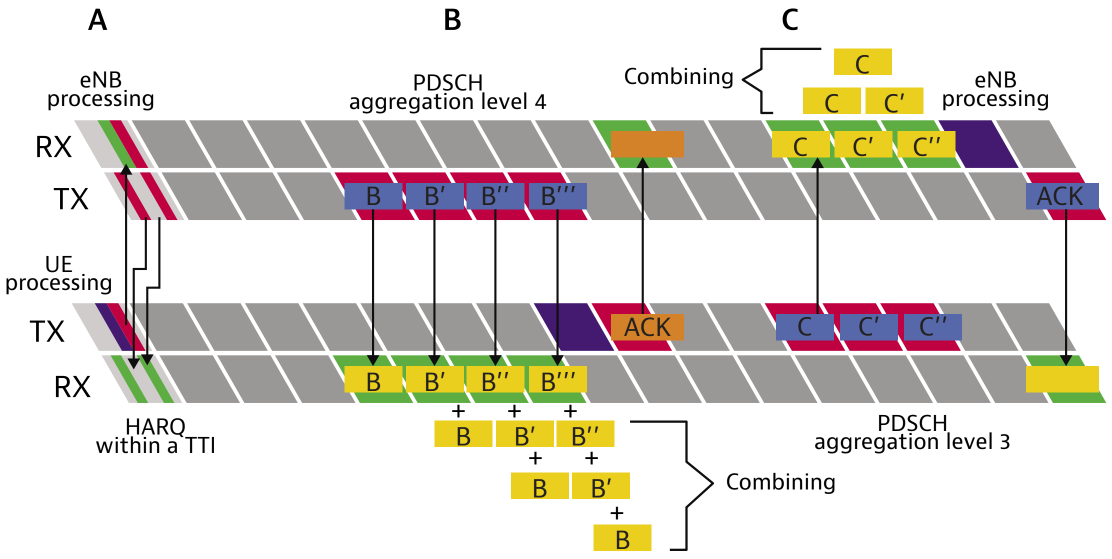


Figure 2.11 Examples of 5G NR FDD HARQ configuration: A self-contained slot for URLLC, B PDSCH with aggregation level 4, C PUSCH with aggregation level 3.

The Figure highlights that a wide range of RTT values may be achieved through flexible configurations and indicates that each configuration represents a trade-off between latency, reliability and power efficiency. An additional element of this trade-off that is not dealt with here but is dealt with elsewhere in the Chapter is the effect of variable numerology that allows reduction TTI.

Figure 2.12 is a qualitative illustration of the trade-off achievable with the 5G NR configuration parameters. It is provided for the case of the URLLC use case for the dimensions of capacity, latency and bandwidth. The Figure shows that URLLC capacity may be increased by reducing the latency target or by increasing the available bandwidth, whereas capacity is considerably reduced by making the reliability requirement more stringent based (Qualcomm reference https://www.qualcomm.com/media/documents/files/expanding-the-5g-nr-ecosystem-and-roadmap-in-3gpp-rel-16-beyond.pdf)

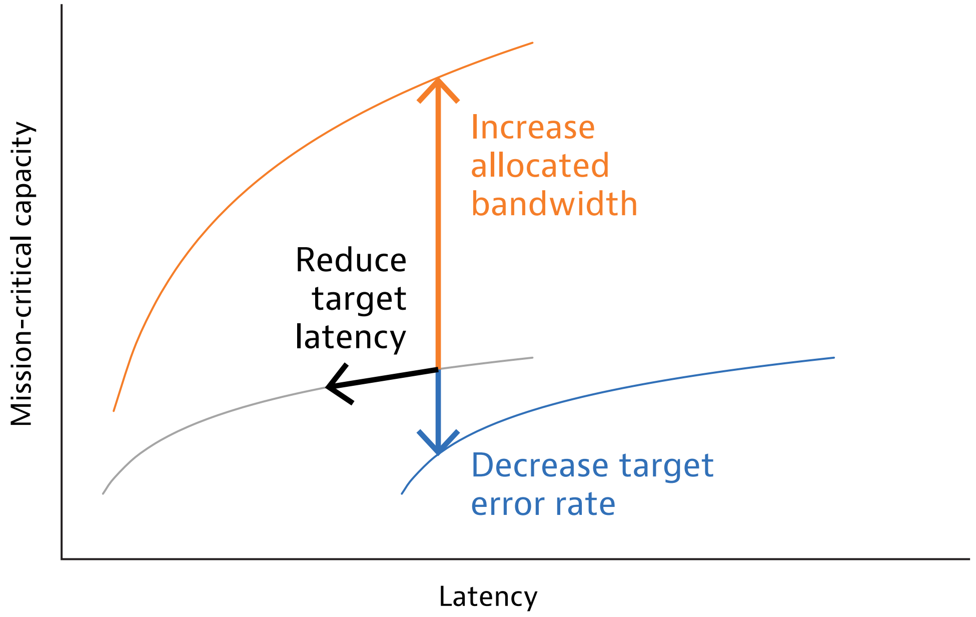


Figure 2.12 Trade-off of latency, capacity and bandwidth

Dual connectivity will allow NR carriers to be used simultaneously with LTE carriers. These support the eMBB use case by providing potentially massive bandwidths. Carrier aggregation is managed in terms of component carriers (CCs). Multiple CCs can be aggregated together where each CC can span up to 400 MHz and comprise a maximum of 3,300 active subcarriers. The actual allowed combinations of CCs are specified by 3GPP. In some cases, the duplex mode can vary between the LTE and NR carrier components, with LTE on FDD and NR on TDD, for example.

2.6 MIMO, Massive MIMO, and Beamforming

MIMO is a well-known cornerstone of 5G. It has wide implications, particularly for achieving eMBB. Here we set out to explain MIMO in terms of the underlying principles, the evolution through different flavors of multiple antenna technology, its similarities and differences from beamforming, and how it will play its part in delivering the promise of 5G.

2.6.1 History of Spectrum Reuse in Wireless Access

In GSM, neighboring cells using the same frequency would have resulted in substantial interference and dramatic loss of spectral efficiency. Therefore, the available spectrum had to be divided into chunks and nearby cells allocated different frequencies. This reuse was wasteful as the entire available bandwidth could not be used in multi-cell systems. In WCDMA, neighboring cells could use the same frequency bandwidths, but scrambling codes were used to separate the spread spectrum signals from different nearby cells. In this case the users at the edge of the cell would sometimes suffer poor signal-to-noise ratios and thus have to use more robust modulation and coding schemes which in turn lowered the spectral efficiency.

In LTE, the use of closely packed subcarriers allowed those parts of the full-spectrum bandwidth that were most suited for each user, in terms of the interference experienced, to be used for that user’s transmissions. However, in cases where there are many users per cell, the theoretical maximum data rate must be shared among the users, limiting the application bandwidths that can be achieved.

The holy grail of the communication system is that every user enjoys the full benefit of the bandwidth. 5G is taking a big step towards this vision, overcoming the limitations of previous generations. Massive MIMO and beamforming are major cornerstones of this by virtue of their ability to deliver multiplexing of multiple communication streams with fine spatial granularity. By breaking the coverage down into narrower beams, the distances over which the full spectral resources can be reused is shortened. While it may not be able to deliver a beam per user, it continues the trend to ever smaller granularity of frequency reuse and is a major leap towards this ideal.

2.6.2 Introduction to MIMO

MIMO stands for Multiple Input, Multiple Output. What does this mean? To what is the input going and from what is the output coming? The answer to both questions is the same: namely, the radio interface “channel” or medium between the transmitter and receiver. This additionally includes the RF domain components in the physical equipment at the transmitter and the receiver (e.g., cabling, RF amplifiers, and antennas).

It is common to think of RF communication taking place between a single transmit antenna and a single receive antenna. Data to be transmitted in the DL or UL is scheduled into resource elements, encoded and modulated, then undergoes digital-to-analog conversion, is amplified, carried to the antenna, and finally is transmitted. The receiver antenna captures the radiated energy, then demodulates and decodes the received signal and looks for the parts of the spectral resources (in frequency and time) that correspond to that user. This is Single Input, Single Output (SISO) transmission since there is only one input to the channel and only one output from the channel, as illustrated in Figure 2.13. Here the amount of data that can be conveyed to or from a given user is limited by the amount of spectral resources that can be devoted to that user. It is also subject to the harsh RF medium. Transmission is complex, particularly in lower frequencies. The success of the communication will depend on the combined effect of phenomena such as non-line-of-sight propagation, reflection, refraction, scattering, multi-path propagation, and constructive interference. In cases where there is no line-of-sight between the transmitter and receiver these phenomena can be very helpful, depending on the carrier frequency, because reflection, scattering, and refraction allow the signal to propagate via an indirect path.

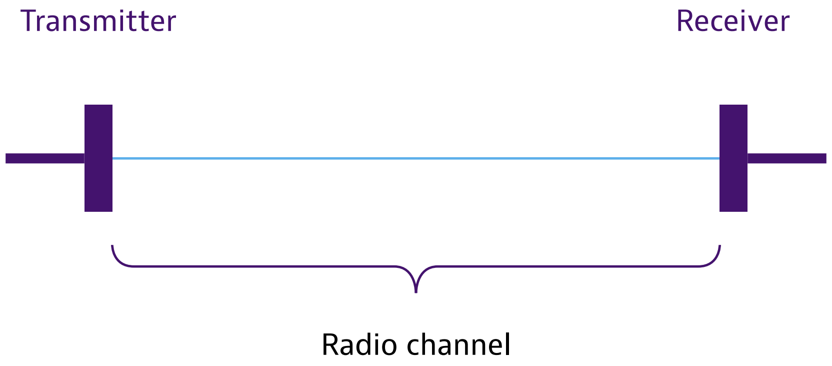


Figure 2.13 Single input single output (SISO) transmission

Transmitted signals can also be subject to less helpful effects such as shadowing and destructive interference or fading that conspire to defeat the transmission. The combination of these helpful and damaging phenomena together makes for a complex channel between the transmit antenna and the receive antenna as illustrated in Figure 2.14.

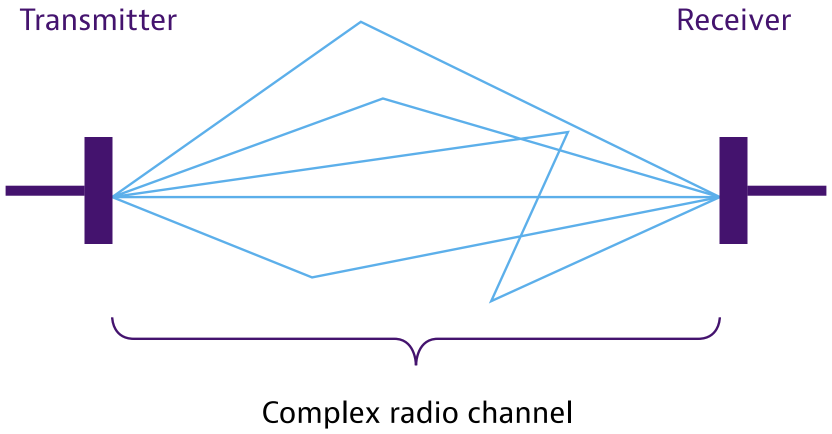


Figure 2.14 Single input single output (SISO) transmission over a channel with non-line of sight propagation

2.6.3 SIMO and Receive Diversity

Whether or not the received signal is useful will depend in part on the location of the antenna. It may be receiving a good strong direct signal, a reflected signal, a scattered signal, or in the presence of combinations of these, may be benefitting from constructive interference. The combined impact of these effects can vary significantly over short distances of even half a wavelength. Thus, if a second receive antenna is employed, and deployed at least half a wavelength from the first, then the chance of receiving a good signal increases.

The second antenna may additionally or instead be cross-polarized, where it is rotated so it is offset ninety degrees from the first. With the two antennas, the communication attempt will then only be defeated if the signals received at both antennas are seriously affected by some combination of destructive interference. Thus, this approach increases the likelihood that communication can be maintained.

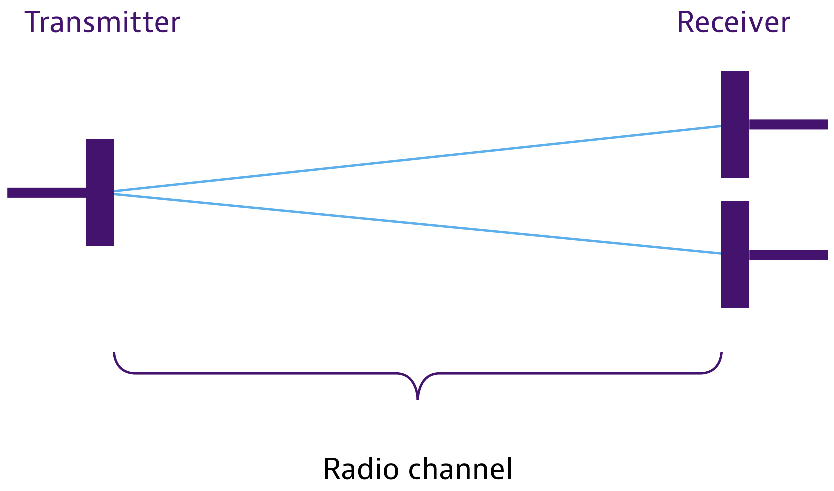


Figure 2.15 Single input multiple output (SIMO) transmission

This method of transmission is referred to as receive diversity. It is an example of Single Input, Multiple Output (SIMO) in its widest definition as there is one input to the channel at the transmitter and two (or more) outputs from the channel at the receiver. This is illustrated in Figure 2.15. This will increase the reliability of the communication link. The receiver can make use of the multiple signals in various ways. For example, maximal ratio combining will use the useful signal from all receive antennas and add them together for a stronger received signal. Another is switched diversity, which chooses the strongest received signal from any of the antennas.

2.6.4 MISO

An alternative to SIMO is to transmit the same signal from two transmit antennas to the same user using the same spectral resources. If done naively, the two signals might interfere with each other, the receiver would be unable to make sense of either, and nothing can be decoded. But this ignores the complex propagation environment discussed previously. As the distance between the transmit antennas increases, the channels will start to decorrelate. In other words, they will be subject to different combinations of non-line-of sight, multi-path propagation, among others. This provides more of a chance to get the signal successfully from the transmitter to the receiver.

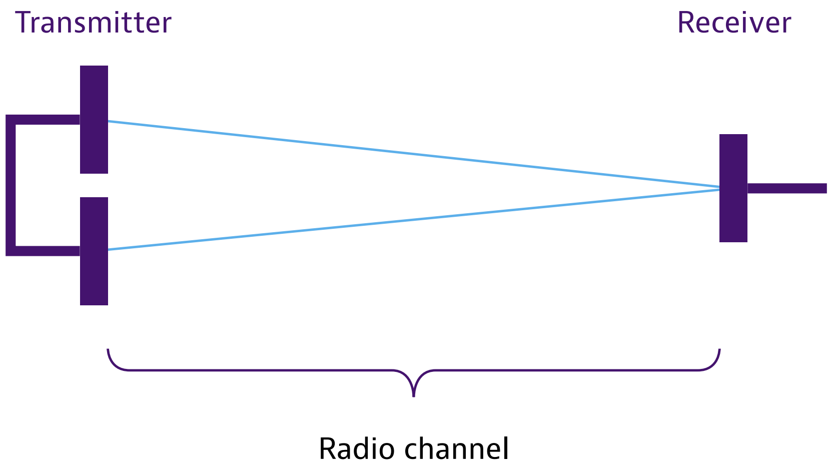


Figure 2.16 Multiple input single output (MISO) transmission

There are more chances for shadowing to be overcome if there are more transmissions. This method of transmission is called space time transmit diversity (STTD): sending the same signal to the same user over the same spectral resources at the same time. It is an example of Multiple Input Single Output (MISO) since there are multiple inputs to the channel but a single output as illustrated in Figure 2.16. This can increase the reliability of the channel by reducing the time that it is subject to the destructive characteristics of the channel. It can also affect the rate of data transfer that can be achieved. Sometimes there will be more destructive interference and the data rate will be reduced; at other times, there will be less destructive interference and the data rate will be increased.

The STTD flavor of MISO can also utilize more than two transmit antennas, although the benefit will diminish as the number of antennas is increased. STTD is not always appropriate. Its success increases as the antenna separation increases up to a point; ideally, they need to be at least a wavelength apart. In higher FR2 bandwidths, the channels are less complex and dominated by line-of-sight propagation. While adding a second transmitter might just mean that the receiver is no longer in the shadow of an obstacle by virtue of the modest separation, this is unlikely and the benefit from using STTD will diminish.

2.6.5 MIMO

As described above, STTD transmits the same signal to the same user over multiple antennas. Now consider the implications of transmitting a different signal (also known as layer or stream) from each antenna to the same user over the same spectral resources at the same time while introducing a number of receive antennas corresponding to (or exceeding) the number of unique transmissions. If the nature of the propagation, or channel state, between each of the transmit antennas and each of the receive antennas is sufficiently diverse, then the signals received at each receive antenna will be to some degree decorrelated. Although each of the transmitted signals will tend to interfere with each other at reception, if the correlation between them is low enough, then the signals can be separated by the receiver.

It may be that the signals must be transmitted with a more robust modulation and/or coding scheme, but in many cases, such as when the channel is sufficiently diverse, the conditions will be right and transmitting two streams of information at lower rate to a user will result in a higher overall data rate than transmitting a single stream of information at a higher data rate. This mechanism of transmitting two or more independent signals, or layers, to the same receiver is called spatial multiplexing or single-user MIMO (SU-MIMO) as illustrated in Figure 2.17. It can also be referred to as multi-layer MIMO as there are multiple layers of data transmitted on the same physical resources. We now have MIMO because there are multiple inputs to the channel and multiple outputs. A SU-MIMO mechanism can increase the data rate that a user can achieve.

SU-MIMO does not always give a gain. If the correlation between the combination of the different layers sampled in each receive antenna is high, then the equalization process will not be able to accurately separate and decode the symbols from each layer. Additionally, the data throughput that can be achieved overall will either drop or, in serious cases, the data may not be able to be decoded at all. In order to control this, the receiver must feed back to the transmitter the state of decorrelation between each of the channels. This is typically done by the UE with a metric referred to as the "channel rank." This indicates the order of decorrelation between the channels and thus how many unique transmissions can be sent between the transmitter and the receiver. Further improvement can be achieved if the receiver provides additional feedback on the overall quality of the channel. Even if they are decorrelated, the signal-to-noise ratio may be poor and the capacity of the channel limited, despite the use of SU-MIMO.

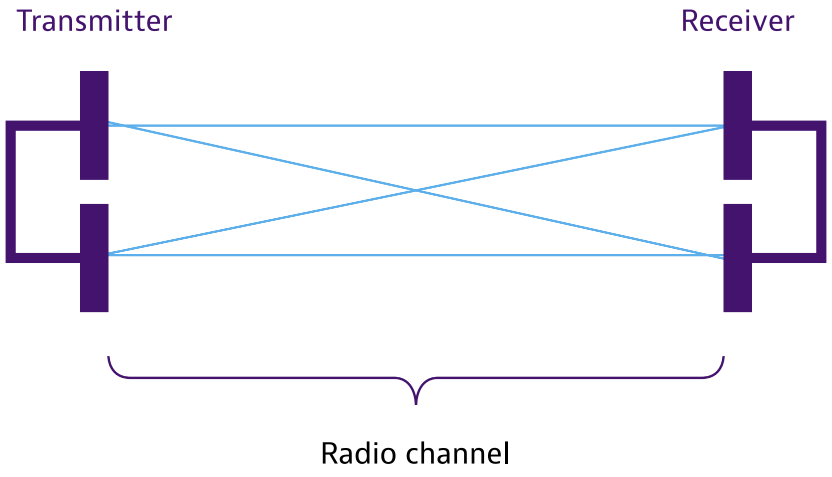


Figure 2.17 Single user-multiple input single output (SU-MIMO) transmission

A further piece of feedback can be provided by the receiver. Using special known signals sent by the transmitter, the receiver can provide an estimate of how the transmitted signals could be modified to maximize the separation between them. These are known as precoder weights. As part of the transmission process, usually in the digital domain, these weights can be applied to each layer in the transmission to optionally change the phase and gain of each layer. This is known as precoding and can deliver substantial performance gains.

SU-MIMO can be generalized from communication between one transmitter and one receiver to multiple transmitters or multiple receivers. This is multi-user MIMO (MU-MIMO), which is illustrated in Figure 2.18. This allows the multiplexing different RF streams onto the same spectral resources using different antennas. Again, multiple antennas at the receiver receive these signals and under the right channel conditions can separate them. Channel feedback can deliver better utilization of the channel and better orthogonality of the channels through precoding.

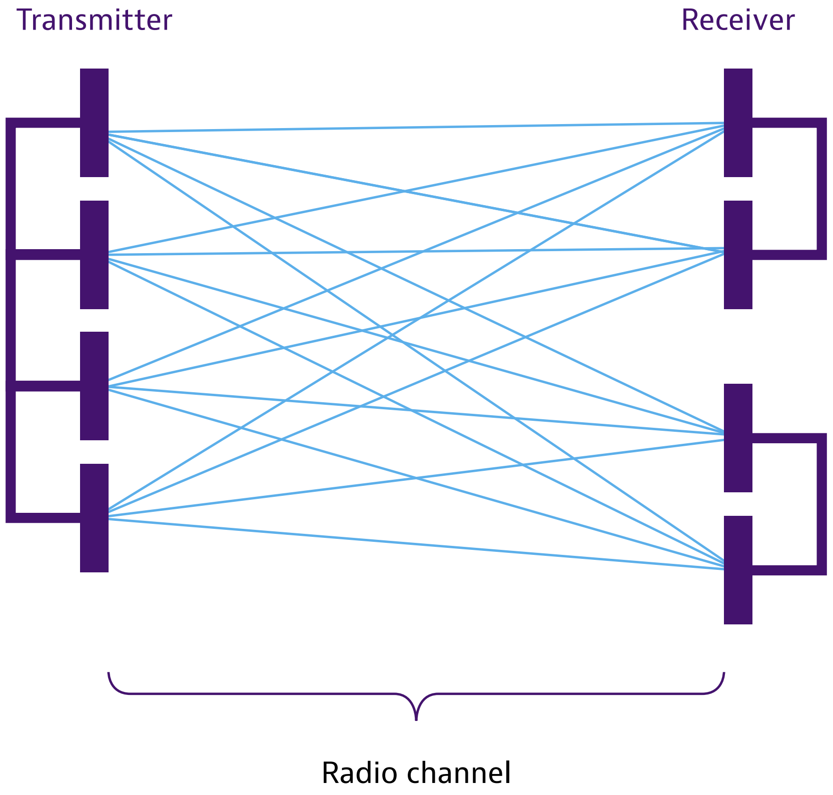


Figure 2.18 Multiuser-multiple input single output (MU-MIMO) transmission

An increased benefit of MIMO is achieved if the coupling between antenna elements is reduced. This is the phenomenon where power transmitted by one element in the array is absorbed by other antennas in the same array, impairing the transmitted signal. This requires the antenna elements to be spaced sufficiently far apart. A minimum spacing of the antenna elements of half a wavelength is generally accepted although larger spacings can increase the benefit up to a point. Increasing the spacing between elements in the array also lowers the correlation between the signals received at each, as the increased spatial separation allows different sampling of the channel between antennas.

These factors and constraints lead to the engineering contradiction that large arrays are desired to maximize the benefit of MIMO, but small form factors are also valued for the arrays to minimize cost, weight, wind shear, and other undesirable characteristics. The good news is that the ideal size of even complex antenna arrays decreases as the frequency increases, making massive MIMO viable and acceptable for FR2 bands in particular.

Although the underlying physics is the same, there are differences between the reality for MU-MIMO at sub-6 GHz and in the mmWave bands. At mmWave, the distances over which signals decorrelate are very short. The success of MIMO depends on the ability to provide feedback rapidly enough to update the precoding matrix to maintain good decorrelation between the channels most effectively. At mmWave frequencies, the precoding required at one location can vary dramatically from the precoding required at another, even millimeters away. Even for deployments serving static UEs, such as fixed wireless access with static antennas attached to the outside of the house, the dynamics in the environment such as people, vehicles, and other scattering objects have a similar effect as motion. This makes the job of maintaining the decorrelation between the channels hard. High dimension MU-MIMO is therefore less effective in mmWave spectrum.

Although the terminology is somewhat subjective, and the distinction arbitrary, low order MIMO involves up to around eight antennas, or channels, where the phase and amplitude can be controlled independently as illustrated in Figure 2.19. These can be dedicated to one user (SU-MIMO) or to a lower number of devices simultaneously. In contrast, massive MIMO supports more antenna elements, generally accepted to be significantly more than eight, and thus can transmit to more devices simultaneously.

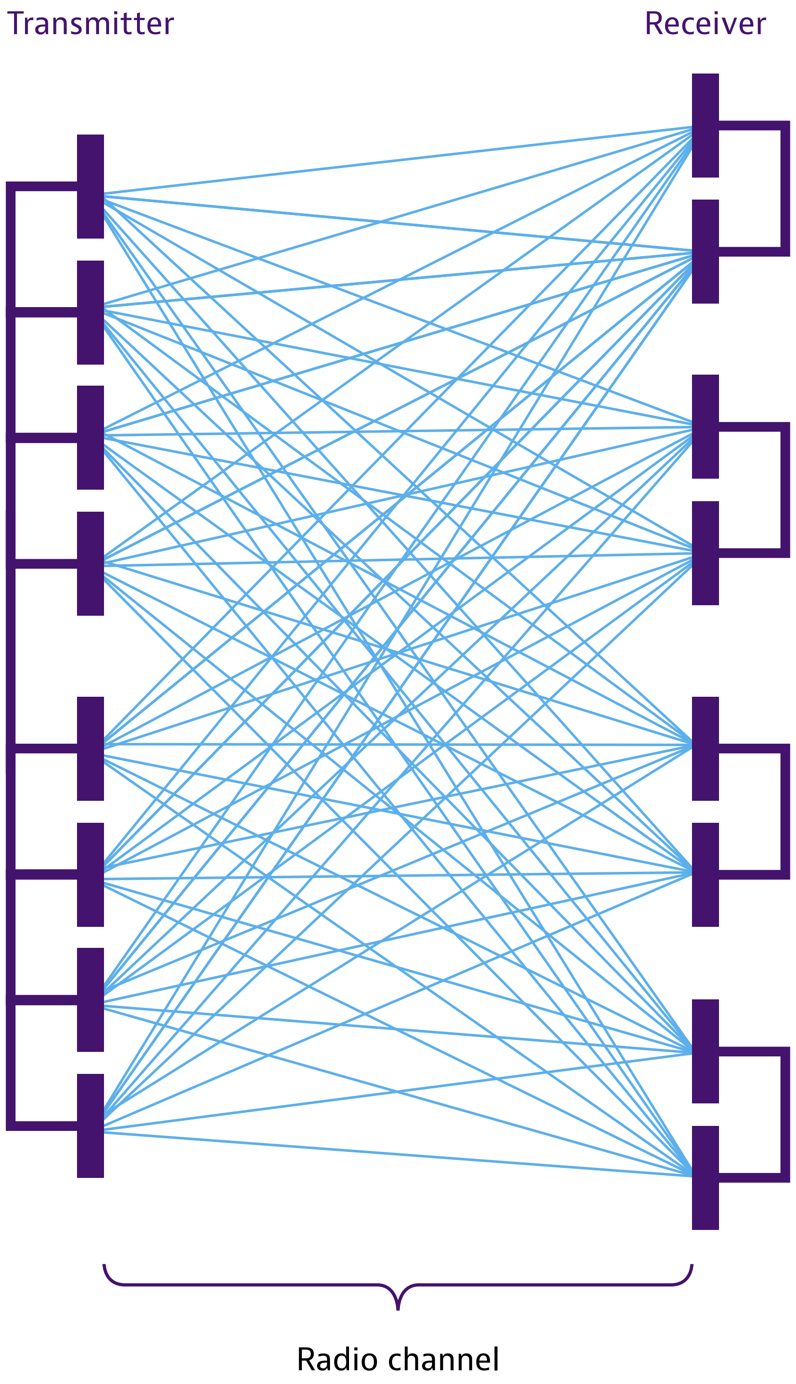


Figure 2.19 Massive MIMO

The LTE standards have supported high order MIMO since Release 13 which saw 16TX MIMO and Release 14 with 32TX MIMO. Support for at least 4x4 MIMO is essential for 5G NSA NR but Release 15 can support up to 32TX MIMO. As the number of antenna elements increases, the effective isotropic radiated power (EIRP) that can be achieved will increase.

Thus, M-MIMO can increase the gain by two orders of magnitude for hundreds of antenna elements. The 5G revolution that Release 15 brings is the flexible and dynamic combination of MIMO together with beamforming.

2.6.6 Introduction to Beamforming

Up to now, we have discussed the use of multiple antennas to increase the capacity and/or reliability of the channel to transfer more data between the transmitter and the receiver. Multiple antennas can also be utilized to shape how the energy is radiated from the antenna. A signal can be transmitted from each element in an antenna array. Each of these signals will tend to constructively or destructively interfere with the others with the exact result of this interference dependent on the location of the receiver.

Given a fixed physical arrangement of antenna elements, by changing the phase and amplitude of the transmission at each element, energy can be directed in one or more specific directions. This is known as beamforming and although it depends on multiple antenna elements, it is a concept distinct from MU-MIMO described above.

A two-dimensional panel antenna array presents the ability to shape and direct the beam in the vertical axis as well as the horizontal axis, facilitating fine control of beams so that coverage can be surgically directed to where it is required. This could be at different horizontal and vertical angles from the antenna boresight and at different distances, assuming the antenna is at an elevated location.

In order to control the beam direction effectively, a spacing of antenna elements of around half a wavelength is ideal. However, larger spacings between antenna elements above half a wavelength increase the occurrence of grating lobes where instances of higher power occur away from the intended beam. This will not only reduce the power in the direction of the intended beam; it will also increase interference and reduce the spectral efficiency and capacity. Having larger numbers of antenna elements results in the ability to focus the energy in narrower beams and reduce instances and effects of sidelobes directing energy in unwanted directions.

Beamforming can be operated in a passive mode, which is also known as switched beamforming. This means that the beams are static and a user in motion will move between beams. Which beam or beams provide service changes over time. Another mode of operation is active beamforming where the direction in which the energy is focused changes over time to track the user.

An important characteristic of beamforming is how tightly the energy can be focused. A system with a few wide beams will generally have less capacity than one with many narrower beams as the spectral resources can be fully reused in each beam. In a beamforming antenna system, one factor affecting the width of a beam is the number of antenna elements. As this number increases, the beam can be focused more tightly. The individual elements should ideally be around half a wavelength apart to allow the energy to be focused effectively.

At 1 GHz, this spacing is around 15 cm while at 30 GHz it is around 5 mm. This means that beamforming antennas at mmWave can be smaller, cheaper, lighter, and less subject to factors such as wind shear, which can disrupt communications. However, if smaller antenna sizes are used, the aperture is also smaller, and less energy can be collected.

2.6.7 Pure Digital Beamforming and Hybrid Analog/Digital Beamforming

Modification of the phase and amplitude of the transmitted signal is required for both MIMO and beamforming. In the case of MIMO, the amplitude and phase of the baseband streams are modified by a precoding matrix chosen for that receiver in its current conditions. In the case of beamforming, the phase of each signal must be aligned to achieve the desired direction for that beam.

In principle, both types of manipulation could take place in the digital domain where the digital signal processing could apply both the MIMO precoding and the phase adjustments to achieve the desired beam direction. In general, there are advantages to signal processing as much as possible in the digital domain with signal processing microprocessors, rather than converting to the RF domain and performing direct manipulation of the phase and amplitudes of the signals. As soon as the signal is converted to the RF domain, conveying the signal and processing it will incur a loss of power. The approach of performing digital-to-analog conversion close to the antenna unit with no further manipulation therefore holds a lot of appeal. However, there are issues with this pure digital beamforming. The complexity of the signal processing for large antenna systems with hundreds of antenna elements places high demands on the signal processing stage. In some deployments, there may be constraints on the availability of real estate for signal processing and the associated HVAC along with the bandwidth requirements between the baseband and RF components of the system.

Also, digital artifacts and intermodulation products can arise from purely digital signal processing. An alternative is to perform some aspects of the signal processing in the RF domain. For example, beam steering can be performed in the RF domain by applying analog phase offsets in the antenna array itself. This reduces the digital signal processing requirement but involves more manipulation in the RF domain. In this case, controlling the phasing and directivity of beams will be slower and more complex.

However, the ease with which these phase shifts can be performed in the RF domain means that this is an ideal solution, especially when dynamic beams are not required, and static or semi-static beams are sufficient. The system design must consider the cost of manufacture, HVAC demands, suitability for the use case, and availability of space for the various disaggregated components (CU, DU, and RU) and the distances between them, as well as the overall achieved performance of the resulting communication.

Beamforming is good for coverage-limited systems. This is important for mmWave which suffers from higher propagation losses. It is also important for allowing reuse of existing base stations designed for 2 GHz bands. Focusing the energy of beamforming means that these same sites can be used for 5G NR even with higher frequency carriers in FR1.

2.6.8 Support for MIMO and Beamforming in 5G NR

The 5G NR standard introduces various features to support multi-layer MIMO and beamforming. Fundamental to this is the concept of the beam, which is defined centrally within the standards. As in earlier technologies, the cell still exists as a logical entity. The cell has a primary synchronization signal (PSS), which indicates the start of the frame, along with the secondary synchronization signal (SSS), which indicates the locations of the start of the subframe.

The PSS takes one of three values, while the SSS takes one of 336 values. Together these determine the physical cell identity (PCI) for that cell for which there are 1,008 unique values. In addition to the cell logical entity, the beam is also a first-class logical entity in 5G NR, which eclipses the cell in terms of importance. Each beam has its own unique Synchronization Signal Block (SSB) which comprises the PSS and SSS, and thus the PCI from the parent cell, along with the Physical Broadcast Channel (PBCH) and its Demodulation Reference Signal (DM-RS). Although the beam exists as a logical entity and is named in a way that could be suggestive of some type of shaping, there is no requirement for the beam to be formed or shaped in any way.

The 5G NR standard defines the concept of antenna ports (See TS 38.211). These correspond to layers of transmissions of different data spatially multiplexed on the same physical resources. The standards define the use of 1, 2, 4, 8, 12, 16, 24, or 32 ports for flexible multiplexing dependent on the channel conditions (38.211). This is highly flexible and means that even where there is some combination of highly correlated channels, high interference, low signal strength, or other harsh conditions, the communication can be successful. This happens by using fewer ports, or low order MIMO, yet when the conditions are good the capacity can be maximized by increasing the number of ports. The CSI-RS symbols (defined below) are mapped onto physical resource elements in such a way that they can be decoded independently by the UE for each channel.

The standards define mechanisms for channel estimation, so that the optimal use of multi-layer MIMO and beamforming can be achieved. The mechanism is flexible; how it is configured in practice will depend on factors such as the frequency band and whether the channel can be assumed to be sufficiently similar in the UL and the DL, a characteristic known as channel reciprocity.

The Channel-State Information Reference Signal (CSI-RS) is transmitted on each beam by the gNB. This known signal allows the UE to estimate the degree to which the channels are decorrelated and thus the degree of spatial multiplexing, or the rank, that can be supported. Once the rank is known, the UE may then have to calculate the precoding required for its transmissions from the gNB, although this is not always the case.

The precoding must be applied by the transmitter in the digital domain by applying a precoding matrix. The matrix will be larger for higher order MIMO with many ports. The matrix itself does not need to be transmitted. That would require a lot of data transmission and would therefore impact capacity and add latency to the ability to respond to channel conditions. Instead of transmitting the raw precoding matrix, an index into a codebook of pre-defined precoding matrices is signaled. The number of codebooks available places the upper limit of ports that can be multiplexed onto the same physical resources. More codebooks would allow for higher and higher multiplexing, and it is the number of defined codebooks that places the limit of thirty-two ports.

The overall quality of the channel is important, as this will determine how robust a modulation and coding scheme should be used. Thus, the following pieces of information are transmitted by the UE to the gNB. The rank indicator (RI) is the number of independent layers that the UE can use. The precoding matrix indicator (PMI) is the index into the codebook of precoding matrices corresponding to the rank and what is most appropriate for the channel conditions. The channel quality indicator (CQI) indicates which combination of modulation scheme and channel coding is appropriate.

Another mechanism for channel estimation is possible. This can be used in conditions where there is channel reciprocity, e.g. when the TDD mode is used, such that UL transmission is on the same carrier as the DL. In this case, the sounding reference signal (SRS) can be used. This is transmitted by the UE to the gNB and allows the precoding matrix for the UE to be calculated directly by the gNB. In this case it becomes unnecessary for the UE to transmit the PMI to the gNB.

This has the advantage of the precoding weights being calculated where they are needed, and so are available with much lower latency and with a saving of UL bandwidth, compared to cases where the PMI must be calculated and then reported by the UE.
This has the further advantage that the choice of precoding weights is not limited to

a codebook of choices, but rather the weights can be fine-tuned more precisely to the channel conditions. Thus, this method of channel estimation is suitable for cases where the required precoding can change more rapidly, such as in FR2 carriers. The use of SRS is insufficient on its own in this case. The UE must still transmit the RI and the CQI to indicate the degree of MIMO that it can support and to allow the modulation scheme and channel coding appropriate for the channels to be calculated.

There is a further feedback mechanism beyond that of the channel. The same logical cell can have more than one beam. A cell with a single PCI can have multiple beams, each with a unique SSB. Each beam will also have its own CSI-RS, allowing the UE to establish the quality of the channels between each beam and itself. This feedback is transmitted from the UE to the gNB using the CSI-RS resource indicator (CRI), allowing the gNB to select the best beam for communication.

Now that we have established the way that channel conditions can be quantified, communicated, and transmission layers appropriately weighted, how does this relate to MIMO and beamforming? This is where some flexibility in MIMO and beamforming is established. Let's consider an antenna panel that has four rows of four cross-polarized pairs, making thirty-two discrete antennas in total. Initially, we consider the case where these are not beamformed. A UE close enough to the transmitter will be able to measure the CSI-RS on at least some of these independent transmissions and measure the channel state. If the conditions conspire in favor of the UE, it may be able to separate all thirty-two transmissions. So, the UE can report the channel state to the gNB using the CQI, the RI, and the PMI. This situation is illustrated on the left-hand side of Figure 2.20.

Now consider the situation where some beamforming is performed so that the thirty-two ports are separated into four beams of eight ports each. Now the CSI-RS is used by each mobile to calculate the relative strength of each beam and uses the CRI to report the information back, along with the CQI and RI as before. This situation is illustrated on the right-hand side of Figure 2.20. In contrast, consider these scenarios when there is channel reciprocity, with and without beamforming. This situation is illustrated in Figure 2.21. Note the PMI has been replaced by the SRS.

The 5G NR standards are thus very flexible in how beams can be formed and how spatial channel multiplexing can be configured.

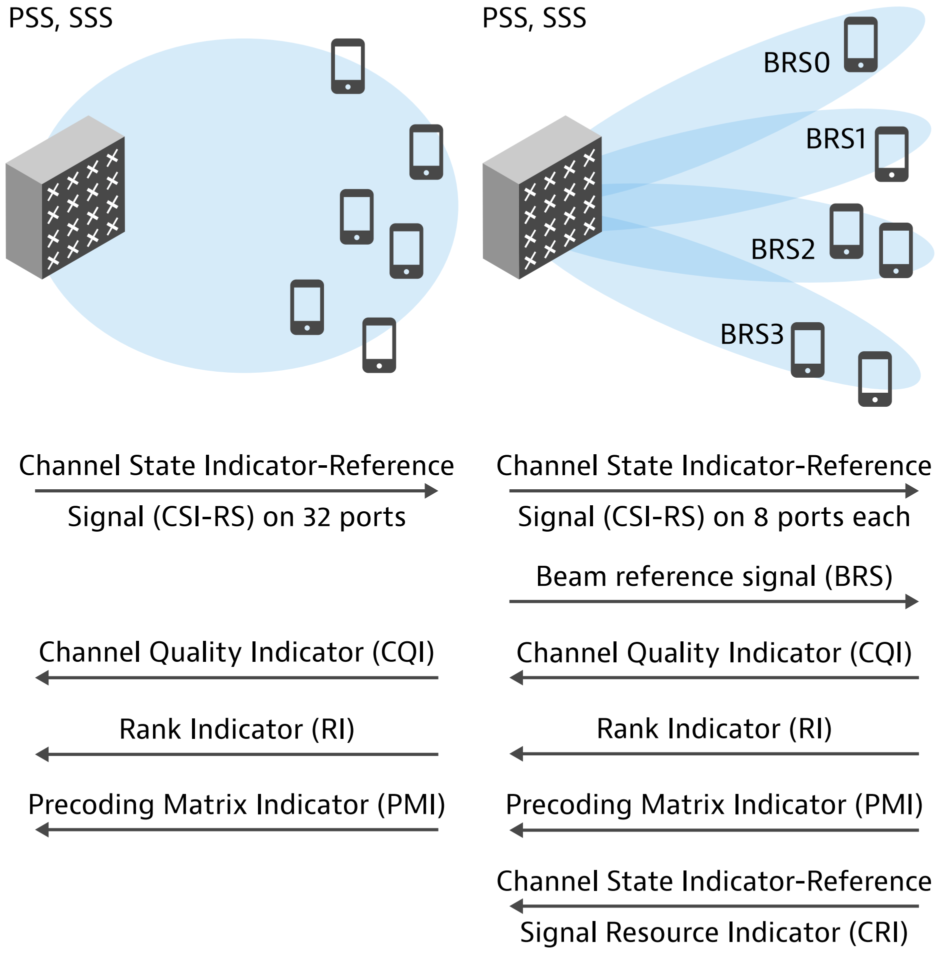

PSS, SSS
PSS, SSS
BRS0
BRS1
BRS2
BRS3
Channel State Indicator-Reference Signal (CSI-RS) on 32 ports
Channel Quality Indicator (CQI)
Rank Indicator (RI)
Precoding Matrix Indicator (PMI)
Channel State Indicator-Reference Signal (CSI-RS) on 8 ports each
Beam reference signal (BRS)
Channel Quality Indicator (CQI)
Rank Indicator (RI)
Precoding Matrix Indicator (PMI)
Channel State Indicator-Reference Signal Resource Indicator (CRI)

Figure 2.20 Channel and beam estimation and feedback for cases without channel reciprocity

Table of Contents

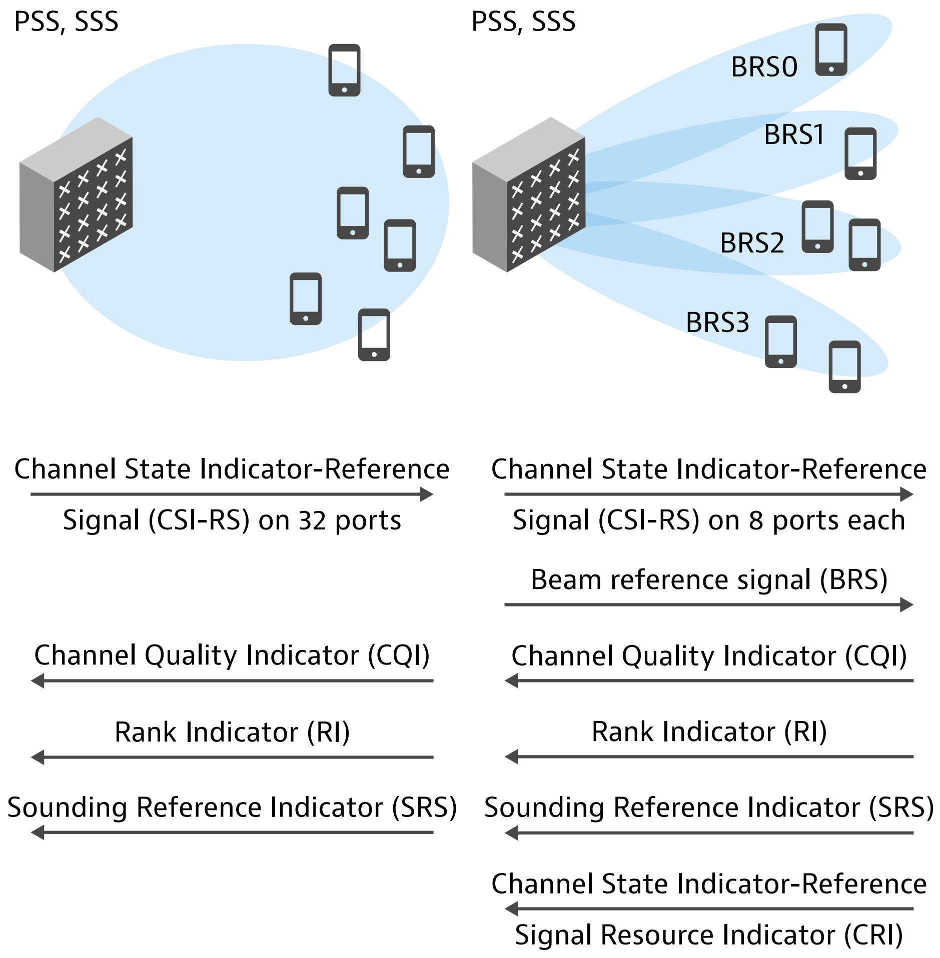

PSS, SSS
PSS, SSS
BRS0
BRS1
BRS2
BRS3
Channel State Indicator-Reference Signal (CSI-RS) on 32 ports
Channel State Indicator-Reference Signal (CSI-RS) on 8 ports each
Beam reference signal (BRS)
Channel Quality Indicator (CQI)
Channel Quality Indicator (CQI)
Rank Indicator (RI)
Rank Indicator (RI)
Sounding Reference Indicator (SRS)
Sounding Reference Indicator (SRS)
Channel State Indicator-Reference Signal Resource Indicator (CRI)

Figure 2.21 Channel and beam estimation and feedback for cases with channel reciprocity (e.g. TDD)

2.6.9 Comparison of MIMO and Beamforming

A comparison of MIMO and beamforming is provided in Table 2.5.

	MIMO	Beamforming
FR1 vs. FR2/mmWave	Higher order MIMO is most effective on FR1 carriers as the nature of the radio channels vary more slowly over distance in these bands, meaning that feedback can be provided in sufficient time to maintain orthogonality between the channels more effectively. Additionally, the channels themselves will have higher rank than FR2 channels. At best, lower order MIMO will deliver a benefit only at FR2 mmWave.	Beamforming can be used in FR1 and FR2. Beamforming is more challenging at lower frequencies as the longer wavelength requires larger antenna separations, and many tight beams require panels that are more expensive, subject to wind-shear and other disadvantages.
FDD vs. TDD	MIMO can work with FDD or TDD. While channel reciprocity can be exploited to calculate precoding weights, reciprocity isn't essential, as information on the precoding weights required can be fed back along with channel state and optimal MIMO rank. In cases where channel reciprocity can be assumed, such as TDD, the calculation of precoding weights is based on UL sounding reference signals. UL MU-MIMO operates the same in FDD and TDD because the base station can make the channel measurements directly in both cases.	Beamforming can be passive/ static which does not depend on any particular duplex mode, as the choice of how beams are configured is a system configuration. Active adaptive beamforming relies on measuring the phase shift between different antenna elements of the received signal in order to apply the same phase shift on transmission. This process is assisted in cases where there is sufficient channel reciprocity such as in TDD.
Antenna separation	MIMO requires separation of at least half a wavelength between antenna elements, but the gain will increase for larger separations. Cross-polarity is also important for maximizing channel diversity.	Antenna separation of around half a wavelength is optimal for beamforming to facilitate effective focusing of the energy. Two-dimensional panels are required for focusing in both azimuth and zenith planes.
Coverage limited vs. interference limited systems	MIMO can overcome some interference limitations by creating a more reliable composite communication channel.	Beamforming can overcome coverage limitations through its ability to focus the radiated energy into narrower beams. This focusing can also control interference offered to other cells and increase the capacity of interference limited systems.

continued

Table 2.5: Comparison of MIMO and beamforming

	MIMO	Beamforming
Open vs. closed loop configuration	MIMO either measures the DL CSI-RS and reports the PMI or else transmits the SRS in the uplink for precoding matrix selection.	Beamforming can be static to drive up the capacity or increase the coverage, which requires no closed loop feedback. Dynamic adaptive beamforming is based on phase offsets of received signals between different antenna elements.
Passive vs. active	Passive MIMO assumes that the channel conditions are sufficient to support spatial multiplexing. Active MIMO uses feedback about the channel conditions using specific messages.	Passive/static beamforming requires no feedback. Active beamforming measures angle of received signal and forms beam in that direction.
Digital vs. analog	Precoding needs complex amplitude and phase shifting and is typically done in the digital domain.	Simple phase shifts for groups of transmitters used for beamforming can be performed in the digital domain but can often be performed more cost effectively in the analog domain.
Single user vs. multiuser	Single user MIMO uses all channels for communication to the same user. Multiuser MIMO uses the channels for communication to multiple devices.	Beams can be used by multiple users although some scenarios with narrow adaptive beams will typically see one user per beam.
5G service enablers	MIMO supports the eMBB service enabler through increased spectral efficiency and capacity. It also supports the URLLC service enabler through the creation of more reliable channels with more resistance to multipath fading.	Beamforming supports the eMBB service enabler through the splitting up of spectral resources into smaller portions with reuse of the resources between them.

Table 2.5: Comparison of MIMO and beamforming

2.7 Energy Saving Support in 5G NR

The energy required for running a 5G gNB is reduced. In LTE, the eNB is required to transmit reference signal every 0.25 ms and a synchronization signal every 5 ms. Even in cases of low utilization where few or no UEs are actively using the cell, these must still be broadcast. In contrast, a 5G gNB is not required to transmit with the same frequency. 5G NR has no cell specific reference signal and the synchronization signal needs only be broadcast every 20 ms by default with the broadcast data. This periodicity is configurable and can be extended up to 160 ms.

Somewhat linked to energy saving is power saving in the UE. It can be costly for the UE to have to continuously monitor the physical downlink control channel (PDCCH) for scheduled DL data. This is inconsistent with the need for battery powered devices that must last for many years without replacement. This is addressed in 5G NR with the extended discontinuous reception cycle (eDRX) feature. When operating under eDRX the UE only need monitor the PDCCH during specific periods and can be dormant at other times as long as it has no data it needs to send in the meantime.

Bandwidth adaptation (BA) allows the UE to be configured to receive the PDCCH on only the active BWP. This also contributes to saving UE energy use.

2.8 Standards Evolution in 5G NR

The 3GPP Release 15 version of the standards was the first release referred to as 5G. However, the work will not be complete until the standards are sufficient to satisfy the requirements laid out by the ITU in IMT-2020. The focus of Release 15 was to deliver the requirements of the eMBB class use cases. There has been less focus on the URLLC, and very little attention has been given to the mMTC classes of use case. These other classes are receiving more attention in Release 16 and will continue to receive further attention beyond that release. To understand the direction that the 3GPP standards will take towards delivering IMT-2020 capability, we can look at the features and studies that are taking place in 3GPP as part of Release 16. Of course, if the 5G network is transcending a pure telecommunications industry focus, it is critical that 3GPP expands to include new industry verticals. As 3GPP developed Release 15, the ranks of 3GPP membership have swelled to include companies from industry verticals that previously had not taken part in 3GPP. These include the agriculture, satellite, automotive, aeronautical, and rail sectors. This is encouraging and will ensure that the standards evolution will be tailored to the new verticals for which 5G is developed. 3GPP plans to complete Release 16 in June 2020 and at that point is expected to meet the ITU requirements for IMT-2020 across all the service enablers; eMBB, mMTC and URLLC.

2.8.1 Evolution of URLLC

With the introduction of 5G into 3GPP with Release 15 came a drive to deliver the eMBB service enabler. The URLLC received less attention and was predominantly focused on TTI for low latency scheduling. In order to achieve the full IMT-2020 requirements for 5G, much attention has been focused on the URLLC service enabler in Release 16. There are studies being carried out in 3GPP to build on the ability of NR to deliver a wider range of URLLC use cases. For example, the study on physical layer enhancements for NR URLLC (TR 38.824) sets out to expand the modest aspirations of Release 15 for URLLC, whose scope was limited to entertainment type applications such as virtual reality. The goal for Release 16 is to address use cases with stricter requirements, such as factory automation, remote driving along with other vehicular applications, and power distribution. Within scope are enhancements to the physical layer including for example; PDCCH, PUSCH, UCI, MIMO feedback, and scheduling. It also includes UL prioritization and multiplexing of inter-UE transmissions and enhanced grant-free UL transmissions.

The PDCCH contains the Downlink Control Information (DCI) which is information that describes where in the OFDMA frame the downlink and uplink data channels are allocated. Until this is received a UE does not know where to look for the data it is to receive or when to transmit its data to send. Awaiting reception of this will naturally delay transmissions and add to latency. An enhanced monitoring capability studied in Release 16 means that this latency can be reduced in some cases.

The study on non-orthogonal multiple access (NOMA) for NR (TR 38.812) aims to reduce latency by removing the need for the uplink transmissions to be granted. There can be a significant delay if a device needs to send some data. In some circumstances if it does not have a sufficiently imminent opportunity for UL transmission, it will have to use the contention-based random-access procedure to establish connection with the network, prior to requesting resources for its transmission. NOMA aims to strip away this delay and cut the latency to an absolute minimum.

Other enhancements to support ultra-reliability are being considered. The Release 15 study on NR to support non-terrestrial networks (TR 38.811) envisages connectivity delivered by airborne or satellite access.

2.8.2 Evolution of mMTC

Delivering massive Machine Type Communications will place further demands on the standards. Use cases in this class will be expected to deliver a service in extraordinarily diverse subscriber densities. Densities of millions of devices per square mile are expected in some scenarios such as smart cities, campuses, factories, and seaport environments. This will require enhancements to the NR to enable, for example, the ability to access the network without congesting the random-access channels.

Conversely, some industry verticals will require very low densities of devices with very long-range communication. This will place a different set of challenges on the networks, such as the need for link budgets exceeding 160 dB. This might be achieved through some combination of robust channel coding and repetition, perhaps supported by beamforming. Many use cases will be viable only if the cost of devices can be driven down. This means that support for extremely simple devices with a single antenna will have to be delivered. Some devices will be battery powered and have decades-long service intervals; this will require extreme battery lives along with enhancements in the standards to support the devices' long sleep times.

Another aspect of mMTC is the connected car, which is the focus of a Release 16 study on vehicle-to-everything (V2X) (TR 38.885). 3GPP envisages a rich set of functionalities in 5G to support vehicular communications. Four use case groups are targeted for support in the standards. The first use case group is platooning. Vehicles in a platoon travel close together and can be regarded as a single entity while the platoon persists. This offers various advantages. For example, not all vehicles in a platoon may need an active human driver. One scenario is that only the lead vehicle needs a human driver. Another advantage of platooning is that the vehicles can drive closer together. This can increase road capacity and ease congestion as well as improving driving efficiency by reducing losses to air resistance.

Another use case group is extended sensors. It is common for vehicles to have numerous sensors for velocity, position, proximity, hazards and nearby objects, temperature, etc. Traditionally these sensors reside in the vehicle. Additional benefit can potentially come from sensors that are not attached to the vehicle but rather are in the vicinity of the current location of the vehicle. These can include video or photographic sensors that are part of the road infrastructure, sensors carried by pedestrians or nearby vehicles for example. The richer sensor data can provide a human or autonomous driver with a more complete view of the surroundings for safer decisions to be made. As some types of sensors are characterized by high data rates, the various 5GNR service enablers are ideal for satisfying this use case group.

The third use case group is advanced driving. This envisages multiple vehicles cooperating by sharing information, sensor data and intent. This goes beyond the extended sensors use case group in that it facilitates coordinated vehicle maneuvers in addition to sharing of sensor data.

The final use case group is remote driving. This includes use cases such as relieving a human driver, for example when they want to perform other activities, or where the driving conditions are challenging. It also envisages the guidance of vehicles that lack an on-board driver, either human or computer.

Many use cases within these use case groups have highly ambitious target latencies. Rather than relying on all communications going via the RAN infrastructure, these use cases are facilitated by a feature that allows radio connection directly between vehicles. This is known as the PC5 side link. The physical layer resembles that of regular NR; that is, it is based on CP-OFDM and supports a number of subcarrier spacings; 15, 30 and 60 KHz in FR1. As well as FR1, it is envisaged that FR2 carriers will be used and support 60 and 120 KHz subcarrier spacing. The close proximity of the vehicles in a platoon or advanced driving group for example will be ideally suited to these higher frequency carriers. Bandwidth parts are compatible with side link. This will facilitate the allocation of smaller portions of a larger 5G carrier to V2X activity.

Sidelink can operate in unicast, multicast and broadcast modes where multicast will be useful for supporting various use case groups such as platooning and advanced driving. In communication between a UE and a gNB, the gNB is in control of who can transmit at what times and in what subcarriers. In contrast, there is no inherent hierarchy between vehicles communicating via side link and therefore no obvious way of managing scheduling to achieve an efficient use of spectral resources. The standards identify two modes of coordination. In the first, a base station controls scheduling of side link resources. This of course requires there to be coverage by the network and so on its own would rule out autonomous of side link away from network coverage. To manage this scenario a second mode is identified. This encompasses autonomous selection of side link resources by UEs as well as coordinated scheduling. The coordinated schemes include UE scheduling other UEs and the assistance in side link resource selection of one UE by another.

2.8.3 Evolution of eMBB

Although the eMBB set of use cases was the central focus of Release 15, there are still active efforts to enhance this further. For example, the study on requirements for NR beyond 52.6 GHz (TR 38.807) is laying the groundwork for this and aspires to open even higher spectrum bands. This addresses challenges such as the increased phase noise encountered in these frequency bands, not to mention the increased propagation loss. These may even mean that new waveforms and spread spectrum utilization mechanisms are needed. Addressing the challenge of designing efficient power amplifiers for these bands is also being considered as part of the study.

This contribution of access to the higher bands to eMBB is complemented by work on NR-based access to unlicensed spectrum. This is particularly challenging as the radio access must co-exist with other transmissions. This leads to lower spectral efficiency and less reliability unless it is well managed. However, as a complement to licensed spectrum, unlicensed bands can bring significant benefits. This has implications for most aspects of the radio physical layer and associated procedures.

The use of NR operating in unlicensed spectrum, NR-U, is being developed in Release 16. This is targeted at 5 and 6 GHz. A design principle that is being followed by 3GPP is that of fair coexistence. That is, the addition of a 5G NR network to an unlicensed carrier should have an impact no greater than the addition of a Wi-Fi network on that carrier.

A number of scenarios for using unlicensed spectrum bands are envisaged across carrier aggregation, dual connectivity and stand-alone. The carrier aggregation scenario between licensed and unlicensed bands would see a licensed band NR primary cell and a NR-U secondary cell. Two dual connectivity scenarios are under consideration where NR-U is in dual connectivity either with LTE or NR licensed. The stand-alone scenario NR-U, connected to the 5GC, would be appropriate for applications not requiring high reliability or guaranteed data rates.

2.9 Navigating the 3GPP Standards for 5G NR

This section provides a quick guide to the key specifications for 5G NR.

TS 38.201: Physical layer; General description; very high-level description of NR physical layer in terms of the relationship to other layers, the choices for multiple access, physical channels, modulation, and channel coding, along with physical layer procedures and measurements.

TS 38.202: Services provided by the physical layer; high level description of the layer 1 functions along with the various uplink and downlink channels.

TS 38.211: Physical channels and modulation; covers physical channels and modulation, including physical channels, frame structure, modulation, random access, and synchronization symbols. Introduces antenna ports, resource grid, resource elements, resource blocks, and bandwidth parts. Introduces the UL and DL physical channels and signals.

TS 38.212: Multiplexing and channel coding; covers how logical channels are mapped to physical channels, how data are encoded, protected, and multiplexed onto the physical channels.

TS 38.213: Physical layer procedures for control; describes synchronization, radio link monitoring and recovery procedures, power control, random access, and procedures for exchanging control information.

TS 38.214: Physical layer procedures for data; describes the procedures concerning the physical layer that the UE must perform to maintain the connection. Includes receiving and sending transmissions and reporting the channel state.

TS 38.215: Physical layer measurements; describes the measurements that the UE and gNB are expected to make including received strength, quality, and signal-to-noise ratios of various signals, timing measurements, and measurements of LTE.

TS 38.300: NR and NG-RAN overall description; provides a comprehensive overview of 5G. Describes the interfaces within the 5G RAN and into the 5G core. Introduces the physical layer, medium access control (MAC), radio link control (RLC), packet data convergence protocol (PDCP), and radio resource control (RRC). Covers mobility in idle, inactive, and connected mode along with mobility between radio access technologies (RATs). Covers scheduling, QoS, security, self-configuration, and self-optimization.

TS 38.304: User Equipment (UE) procedures in Idle mode and RRC Inactive state; describes the various RRC states when the mobile is not connected and procedures for selection of the radio network, selection and reselection of NR cells, broadcast information, and paging.

TS 38.306: User Equipment (UE) radio access capabilities; describes the various categories of UE in terms of the capabilities they possess including maximum data rates and buffer sizes along with parameters of the PDCP, RRC, RLC, MAC, RF, and physical layers.

TS 38.321: Medium Access Control (MAC) protocol specification; defines the MAC. Includes coverage of random access, maintenance of timing alignment, scheduling of transmissions, retransmissions, MAC protocol data units, and control elements.

TS 38.322: Radio Link Control (RLC) protocol specification; defines the RLC along with the acknowledged-, unacknowledged-, and transparent-modes for RLC entities. Also covers procedures including those for management of RLC entities, data transfer, and automatic repeat request (ARQ). Defines how RLC protocol data units (PDUs) are constructed along with the RLC parameters.

TS 38.323: Packet Data Convergence Protocol (PDCP) specification; defines the PDCP in terms of its architecture, services, and functions. Also covers procedures including those for management of PDCP entities, data transfer, discarding service data units (SDUs) and data recovery, reporting on transmit and receive operation, header compression, ciphering, integrity protection, and PDCP duplication. Defines how data and control PDCP protocol data units (PDUs) are constructed along with the PDCP parameters.

TS 38.331: Radio Resource Control (RRC); Protocol specification; defines the RRC in terms of its architecture, services, and functions. Also covers procedures including those for management of PDCP entities, data transfer, discarding service data units (SDUs) and data recovery, reporting on transmit and receive operation, header compression, ciphering, integrity protection, and PDCP duplication. Defines how data and control PDCP protocol data units (PDUs) are constructed along with the PDCP parameters.

TR 38.912: Study on New Radio (NR) access technology; good overview of the deployment scenarios, forward compatibility, and protocol architectures of the user and control plane of the radio interface. Also includes the physical layer modulation, multiplexing, data and control channels, waveforms, multiple access, channel coding, multiple-antenna schemes, physical layer procedures, scheduling, power control, and random access. Also covers the protocol stack including MAC, RLC, PDCP, AS, and RRC. Also covers the architecture including the responsibility of core and RAN, the NG and Xn interfaces, quality of service, dual connectivity with LTE, and network slicing. Includes procedures for initial access, mobility, dual connectivity, and session management along with radio transmission and reception.

2.10 5G NR and Network Slicing

Network slicing is key to delivering mixes of use cases in the same RAN. The vision of many industry verticals being enriched by new services with different mixes of low latency, ultra-reliability, massive connectivity, and enhanced Mobile Broadband is most valuable if these can be delivered simultaneously in the same network. There are many design decisions in 5G NR that have been made with this desire to address numerous use cases together simultaneously.

Flexible numerologies and bandwidth parts are key to enabling network slicing. As described previously, the higher numerologies have larger subcarrier spacing and hence shorter symbols and slots can be scheduled more rapidly. As the slots are self-contained in terms of how data are scheduled on them, this is key to supporting low-latency applications. For the numerology with 120 KHz subcarrier spacing (the highest numerology that supports data), the slot lasts a mere 125 μs. This gives extraordinarily frequent opportunities for scheduling low-latency data in conditions where the channel conditions support it and can deliver the low latency required.

But it would be wasteful if an entire carrier had to be given up to a numerology with the shortest slots. Splitting a block of spectrum into two carriers could overcome this but would reduce the spectral efficiency with the need for guard bands. Thus, another key part of delivering network slicing is the use of bandwidth parts. The ability to use different numerologies within the same carrier means that services with vastly different characteristics can be supported, while preserving spectral efficiency.

While the use of high subcarrier spacing and short slots is consistent with low latency, it is also subject to the issue of inter-symbol interference. This will reduce the reliability of the transmission and could have an impact on Ultra-Reliable Low Latency Communication (URLLC). But even when high-order numerologies are used where the delay spread is significant and indicates against the use of such short symbols, there are mitigations in the 5G standard.

First, the most robust modulation and coding schemes may be sufficient to overcome the inter-symbol interference. But packet duplication can also assist in this regard. This is a mechanism that creates a second RLC entity along with a second logical channel on the radio bearer when more reliability is required. The data is then transmitted twice—once in each RLC entity. This significantly raises the chance that each packet will be delivered successfully.

Mini slots, described earlier, are also a key feature for low latency. Resource blocks in the resource grid can be reserved for mini-slots and not be used for transmissions scheduled as part of normal slot scheduling. Mini slots can start at any time in the period and don't need to be aligned to other slot boundaries. This is ideal for low-latency applications as the transmissions can start as soon as they are needed by the low-latency application with no constraints on delay. Mini slots can be used in the DL and UL and can be as short as one symbol in length.

The 5G standard makes it possible for data to be grouped into different logical channels. Logical channel prioritization (LCP) allows these logical channels to be assigned different prioritizations. This supports network slices that are able to respect the relative importance and latency requirements of the slices. Restrictions can be placed on the logical channels so that they must be restricted to being scheduled on specific combinations of configured cells, numerologies, or PUSCH transmission durations. Logical channels thus have a well-defined hierarchy of priority consistent with their importance, and the latency can be controlled by arranging for the latency to be low for slices that need it.

These various features, in combination with the other flexible uses of the 5G infrastructure, facilitate network slicing for delivery of the rich mix of heterogeneous services that 5G will provide.

CHAPTER THREE

5G Radio Access Transport Networks

In Chapter 1, we talked about the changes in the 5G RAN architecture and the emergence of new fronthaul, midhaul, and backhaul networks. In this chapter, we cover the transport aspects of these new networks for 5G. First, we'll go down memory lane and look at the RAN transport characteristics prior to 5G. We will also review the synchronization technologies as LTE-Advanced and 5G timing and synchronization requirements increasingly play a role in RAN network design and qualification. Finally, we conclude the chapter by describing the 5G service requirements and their implications for emerging transport networks, especially in fronthaul networks.

Let's start with a brief history of RAN transport networks. RAN is characterized by a large number of cell sites connected through backhaul networks with central offices that host elements of the core network. 2G RAN used T-carrier (T1/T3) or plesiochronous digital hierarchies (PDH) technologies for the backhaul. Networks moved to synchronous optical network (SONET)/ synchronous digital hierarchy (SDH) in larger macro cells and backhaul aggregation sites toward the end of the twentieth century as bandwidth demand remarkably increased, particularly in denser metropolitan areas. The increased offering of cellular data plans required significantly more bandwidth allocation in backhaul networks. However, this bandwidth increase was smaller than the incremental revenue associated with data plans. This challenge demanded a new cost-effective backhaul technology. The introduction of carrier ethernet in the early 2000s was just the solution wireless operators needed for the backhaul.

These technology improvements contributed to a quick deployment of 3G and 4G systems in initial rollouts and delivered large-scale coverage for wireless operators in their race to maintain market share. However, a drastic rise in smartphone use indoors and outdoors led to service bottlenecks in many areas.

In response, vendors offered small-size integrated cell solutions that delivered capacity in needed areas. They are also known as metrocells, microcells, nanocells, or more broadly small cells. They are connected through Ethernet links to central locations that can host macrocells, or core network elements. These links can be viewed as an extension or expansion of the backhaul networks. These products could be easily added to existing networks using the available backhaul infrastructure such as copper/DSL, microwave, or fiber. While being low cost and easy to deploy, small cells do not provide sufficient capacity for large-scale applications such as large indoor venues and dense metropolitan areas.

To combat this issue, some vendors offered proprietary solutions for large-scale indoor applications. Distributed antenna system (DAS) is one major category deployed in many venues such as stadiums and airports. DAS consists of multiple antennas connected through a system of coaxial and fiber links to baseband units located in central equipment rooms in the same facility or reasonably close to the same location. Over time, some vendors began to deploy standards-based digital fiber technology such as common public radio interface (CPRI) that is introduced later in this chapter.

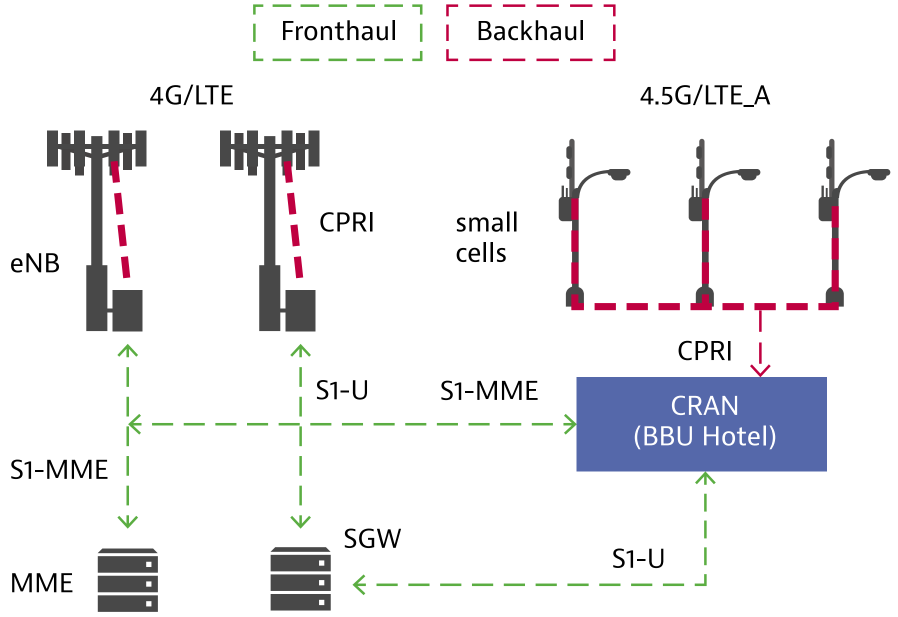


Figure 3.1 Evolution of 4G radio access networks

3.1 Fronthaul Networks

Parallel to the wireless operators' deployment plans, residential and enterprise applications drove a massive roll-out of fiber in metropolitan and suburban neighborhoods. Availability of fiber provided the wireless operators with a golden opportunity to scale their 4G systems for the drastic rise of capacity needs at reasonable costs.

By placing radios on rooftops or utility poles, they were able to multiply their cellular capacity without building large towers and ground-level shelters, or lease equipment rooms at every radio site. The power and cooling demands of baseband units and routers could be met by placing them in a central location, sometimes miles away from the radios. This centralization is also known as cloud or centralized radio access networks (C-RAN).

The last mile fiber-based network between the baseband unit (BBU) and remote radio head (RRH)/remote radio unit (RRU) was named the fronthaul network, in contrast to the backhaul that connects the BBUs with the core mobile network (Figure 3.2). The opposite side of the BBU connects it to the evolved packet core (EPC) network typically located in central offices and constitutes the backhaul network. The RRH is further connected to an antenna element at the top of a tower or rooftop with a coaxial cable. The RRH can be placed closer to the BBU at the ground level or near the antenna on a tower. Placement near the antenna makes the system more energy efficient as it reduces the amount of power loss on coaxial power. However, there might be limitations for placing the RRH near antennas due to proximity to power lines, concerns for wind loading, or preference for easier maintenance. A majority of systems place the RRH near the antenna. In older systems, such as BTS and Node B, the BBU and RRH were often placed in the same location and chassis (Figure 3.3). There are also remote radio units with integrated antenna systems (Figure 3.4).

3.1.1 Optical Layer Technology

Fronthaul networks are deployed in various configurations. The simplest variant involves using dark fiber where there is no equipment between the RRH/RRU and BBU. While being low cost, dark fiber does not make good use of precious fiber resources, which may have to be leased from a third party. As the number of radios increased, operators started to use wavelength division multiplexing (WDM) technologies using the same fiber for several antennas to increase the amount of traffic that could be sent through a network. WDM technologies come in multiple variants. They can be characterized in two categories: passive/active, and CWDM/DWDM.

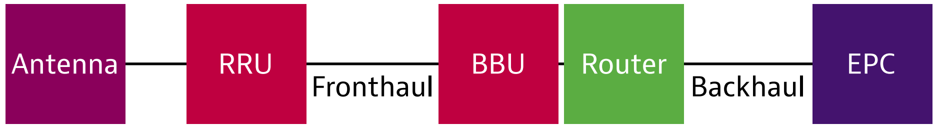


Figure 3.2 RAN functional blocks and interfaces

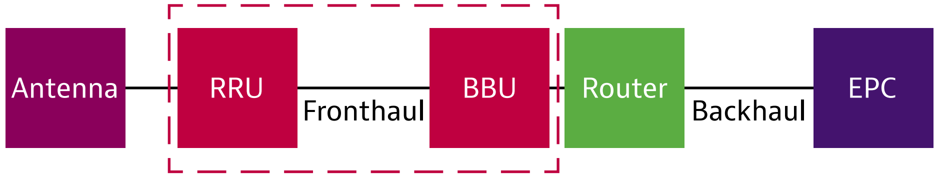


Figure 3.3 Integrated BBU and RRU Equipment

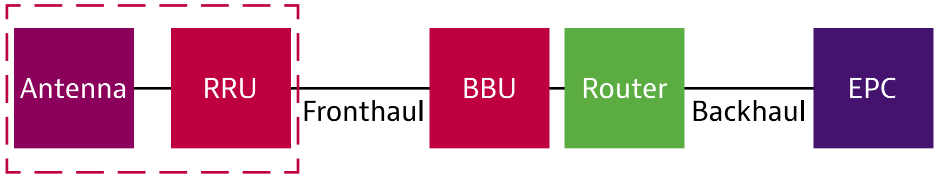


Figure 3.4 Integrated antenna systems

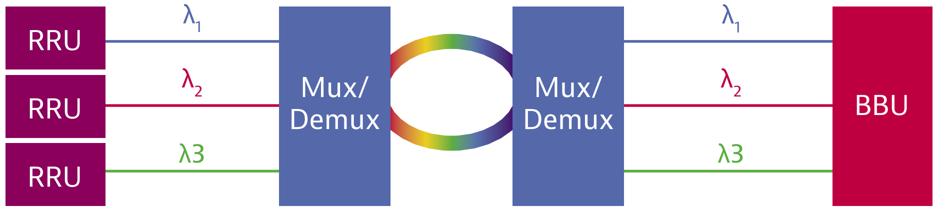


Figure 3.5 Simple wavelength multiplexing

Passive WDM systems (Figure 3.5), are composed of a wavelength multiplexer/demultiplexer (mux/demux) that is deployed as a pair near the BBU and RRUs. The BBU and RRU deploy colored optical transceivers that are distinguished by unique wavelengths assigned to the respective RRU. The passive WDM technology takes advantage of the optical energy in BBU/RRU transceivers and is more economical. Active WDM systems (Figure 3.6) accept grey colored optics as their clients and deploy wavelength translation functions to convert the client signals into colored signals before multiplexing them. Also known as transponder-based systems, they may incorporate optical amplifiers and dispersion compensation modules that allow the system to operate at much larger distance than the passive WDM systems.

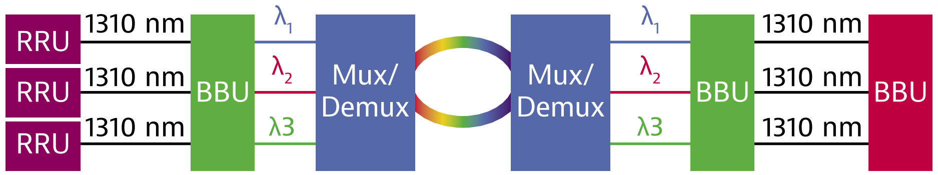


Figure 3.6 Enhanced multiplexing with wavelength conversion

The number of wavelengths that can be multiplexed into a single fiber is dependent on the type of colored optical transceivers used in the BBU/RRU. Coarse WDM (CWDM) typically deploys wavelengths from 1270 nm to 1610 nm with a channel spacing of 20 nm. With newer fibers (compliant to G.652.C/D), CWDM systems can support up to eighteen end points. CWDM transponders are not cooled in operation, thus temperature variations can cause a drift of the central wavelength. Therefore, each CWDM should be checked for possible wavelength shorts or power loss with an optical channel checker.

Dense wavelength division multiplexing (DWDM) was developed to take advantage of the capabilities of erbium doped fiber amplifiers (EDFA) to carry a large number of channels over larger distances. It uses C-Band (1525-1565 nm) or L-Band (1570-1610 nm). Wavelength spacing is much tighter in DWDM compared to CWDM, and therefore requires precision temperature control of the transmitters. DWDM is mainly deployed in backbone networks, but there is an emerging application for DWDM in radio access networks.

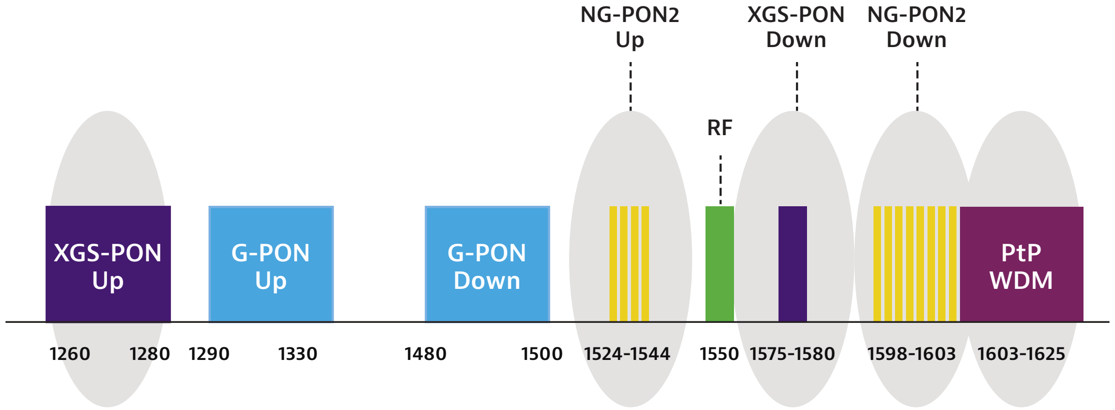


Figure 3.7 PON technologies

Beyond WDM technologies, passive optical network (PON) represents a possible solution for wireless transport networks. Several PON technologies are available, they are typically allocated a certain wavelength window on the optical fiber systems (Figure below). This fixed allocation of windows allows multiple PON systems to operate on the same fiber system at the same time. GPON technologies are characterized by relatively smaller amount of bandwidth (2.4 Gbps downstream/1.2 Gbps upstream). With XGSPON (X for 10, S for Symmetrical), the bandwidth increased to 10 Gbps for both downstream and upstream traffic. The above PON systems share the bandwidth in the downstream. For upstream, they have a time division multiplexing (TDM) approach that allocates a timeslot for each end user. While sufficient for residential, enterprise and wireless backhaul applications, this TDM scheme will pose limitations for wireless fronthaul networks. The TDM system implies a latency penalty. As we will see in the next sections, latency management is very crucial for fronthaul. This limitation has encouraged the PON community to work on improving the latency with multiple technologies that include dynamic bandwidth allocation (DBA) and Cooperative systems. When combined in Cooperative DBA, the PON systems “cooperate” with BBU/DU to dynamically allocate bandwidth to RU.

WDM PON or NGPON/NGPON2 systems combine the WDM and PON technologies to multiply the available bandwidth. The number of wavelengths may be same or different for upstream and downstream. If the number is same, the system is known as XGS-PON (S for Symmetrical).

PON takes advantage of the optical layer, which can help reduce the cost of fronthaul significantly but doesn’t provide some of the advantages of switched networks. Statistical multiplexing is one prominent example of these advantages that can drastically reduce the bandwidth needs, especially for traffic patterns that significantly change over time. The current switched Ethernet network technology is, however, not ready for strict timing requirements of fronthaul. Development work is under way in the time-sensitive network (TSN) community, and we will cover some of the highlights in following chapters. But before we talk about switched networks and TSN, let’s review the current transport technology in fronthaul networks.

3.1.2 CPRI (Common Public Radio Interface)

CPRI is the main protocol used over the fronthaul fiber links between BBUs and RRUs due to its efficient use of resources compared to other technologies. CPRI defines a digitized and serial interface between radio equipment control (REC) and radio equipment (RE). They can be considered as the equivalent of BBU and RRU, respectively. The protocol defines all necessary items for transport, connectivity, and control among these two entities. It represents layers 1 and 2 of the open systems interconnection (OSI) stack (Figure 3.8). The physical layer supports both electrical and optical interfaces, although optical is the main interface used.

CPRI protocol is defined for the physical line rates listed in Table 3.1. Lower-rate options deploy the 8/10B encoding scheme familiar from lower-rate Ethernet interfaces. The higher rate use 64/66B encoding is well known from the 10G Ethernet interface. The latter encoding is much more efficient than the former, which wastes 10 bits for every 8 bits of user traffic. For example, a CPRI rate 7 signal uses a 9.8 Gbps transport link carrying sixteen times the bandwidth of a CPRI rate 1 link. In contrast, CPRI rate 8 uses a slightly higher signal line rate (10.1 Gbps) but transports twenty times the capacity of a CPRI rate 1 link.

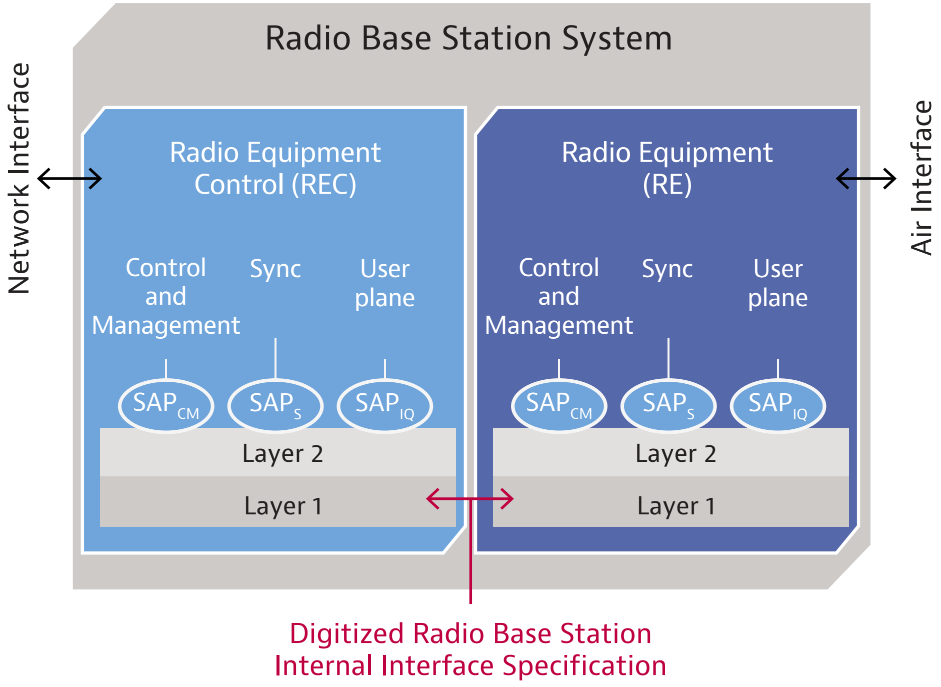


Figure 3.8 High level CPRI interface

Table of Contents

Option	Line Rate (Mbps)	Line Coding	Capacity
1	614.4	8B/10B	1 x 491.52 x 10/8 Mbit/s
2	1228.8	8B/10B	2 x 491.52 x 10/8 Mbit/s
3	2457.6	8B/10B	4 x 491.52 x 10/8 Mbit/s
4	3072.0	8B/10B	5 x 491.52 x 10/8 Mbit/s
5	4915.2	8B/10B	8x 491.52 x 10/8 Mbit/s
6	6144.0	8B/10B	10 x 491.52 x 10/8 Mbit/s
7	9830.4	8B/10B	16 x 491.52 x 10/8 Mbit/s
7A	8110.08	64B/66B	16 x 491.52 x 66/64 Mbit/s
8	10137.6	64B/66B	20 x 491.52 x 66/64 Mbit/s
9	12165.12	64B/66B	24 x 491.52 x 66/64 Mbit/s
10	24330.24	64B/66B	48 x 491.52 x 66/64 Mbit/s

Table 3.1: CPRI Options and Line Rates

Layer 2 of the OSI stack (Figure 3.9) is primarily composed of user plane traffic in digital in-phase/quadrature (IQ) data. Beyond IQ data, three channels are available for control and management information: Ethernet, high-level data link control (HDLC), and vendor-specific bytes. Finally, there is an allocation for exchanging layer 1 in-band protocol data that is used for link signaling for system start-up and maintenance purposes. Alarm indication signal (AIS) and remote defect indication (RDI) belong to the last category of CPRI layer 2 content.

The CPRI (Figure 3.10) signal is characterized by hyper frames composed of 256 Basic Frames (BF). Each BF starts with a control word. The remaining part of the BF can be used to carry IQ data. There is a total of 256 control words that are allocated to different purposes such as synchronization, Ethernet/HDLC Control and Management (C&M) channel, or L1 in-band protocol.

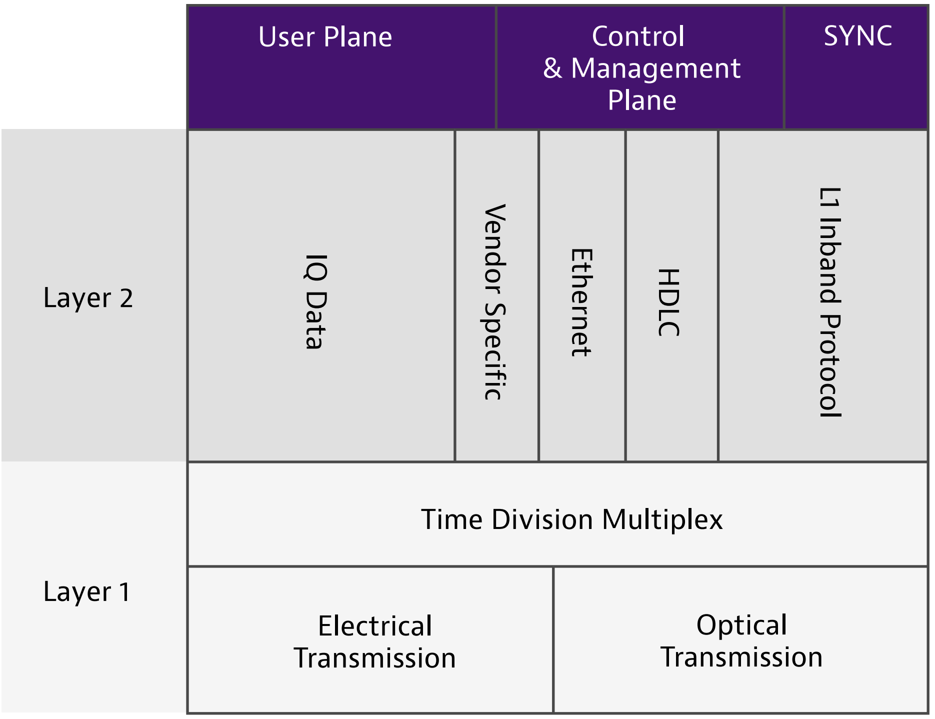


Figure 3.9 CPRI layers

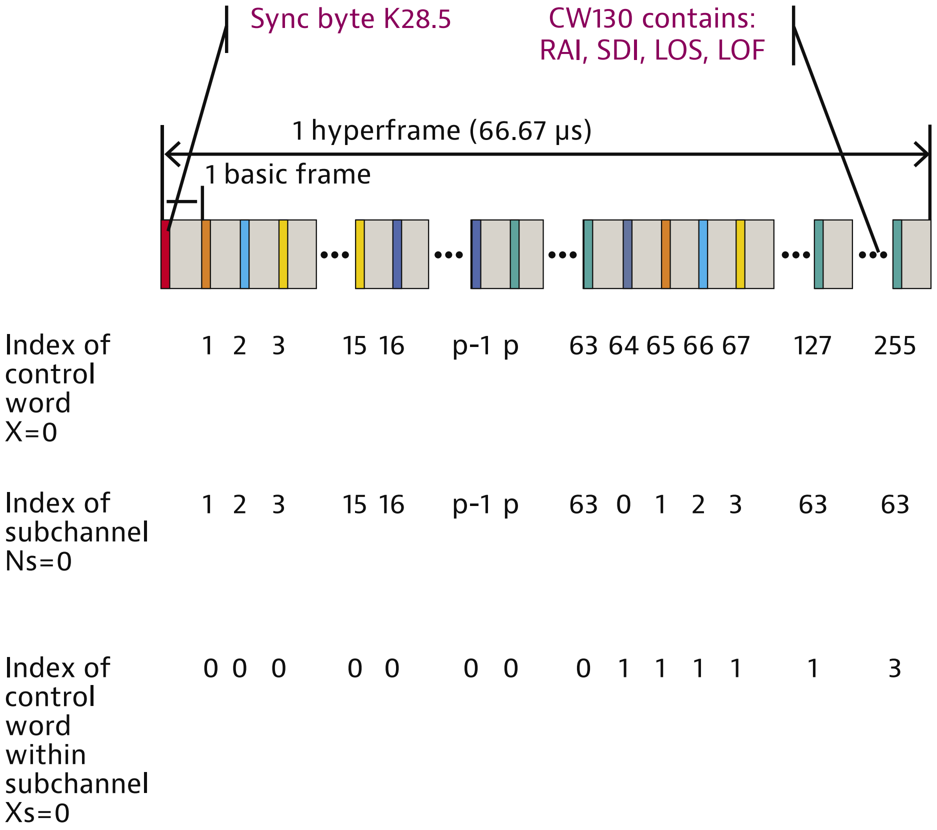

Sync byte K28.5
CW130 contains: RAI, SDI, LOS, LOF
1 hyperframe (66.67 µs)
1 basic frame
Index of control word X=0
1 2 3 15 16 p-1 p 63 64 65 66 67 127 255
Index of subchannel Ns=0
1 2 3 15 16 p-1 p 63 0 1 2 3 63 63
Index of control word within subchannel Xs=0
0 0 0 0 0 0 0 0 1 1 1 1 1 3

Figure 3.10 CPRI Hyperframe structure

3.2 Synchronization Networks

Synchronization is important to mobile networks as it works to get the most performance out of a network while preventing packet loss, which can degrade the quality of experience (QoE) for mobile users. Proper synchronization is critical for applications such as VoIP and video streaming and will be just as critical for 5G related services such as network slicing and IoT.

3.2.1 Synchronization Requirements and Technologies

Wireless services have relied on synchronization from the very beginning. GSM/ EDGE services (Table 3.2) depended on frequency synchronization for proper network operation. 3GPP standards mandate 50 ppb for the frequency stability of a macro BTS. Exceeding that range may create interference issues that ultimately lead to poor cell phone coverage. 3G (UMTS/WDCMA-FDD) and 4G (LTE-FDD) applied the same requirements. CDMA-2000, WCDMA-TDD, LTE-TDD networks drove the need for timing and phase synchronization. These requirements are partially driven by advanced wireless features such as coordinated multipoint (CoMP) or enhanced Inter-Cell Interference Coordination (eICIC). These synchronization requirements led to the deployment of GPS (GNSS) and PDH/T1/T3/BITS and SONET/SDH timing at cell sites and wireless switch/aggregation sites.

Radio Technology	Frequency Sync	Time/Phase Sync
GSM	50-100 ppb	
CDMA 2000		3-10 µs
UMTS/WCDMA-FDD	50-100 ppb	
WCDMA-TDD		3 µs
LTE-FDD	50-100 ppb	
LTE-TDD	50-100 ppb	3-10 µs

Table 3.2: Wireless Synchronization Requirements (2G/3G/4G)

While Global Positioning System (GPS) technology still represents the primary synchronization method for cell sites, network-based synchronization has been increasingly deployed in new networks around the world as a back-up or primary synchronization source. As operators increasingly moved to Ethernet instead of PDH/SDH or T-carrier/ SONET backhaul technologies, there was a need to use a packet-based synchronization and slowly consider retiring legacy backhaul technologies.

Prior to deployment of packet-based technologies in wireless networks, various packet-based technologies were deployed in smaller private networks such as enterprises or industrial automation applications. Network timing protocol (NTP) had been used for many years primarily for obtaining reference time from NTP servers and synchronizing various geographically dispersed end points. NTP was also selected as a technology for radio access networks, although its penetration slowed down with the introduction of IEEE 1588v2 standard (Table 3.3).

Also known as precision timing protocol (PTP), IEEE 1588v2 was proposed based on the IEEE 1588 standard that initially aimed to address the needs of synchronization in instrumentation applications in smaller networks such as production or laboratory test stands. Various capabilities were added to make it telecom grade for larger geographic distribution and meeting stricter timing requirements. Both NTP and PTP are based on a master-slave principle in which the master transmits its reference time and phase to downstream slaves. Between the master and slave there may be a number of packet processing switches or routers. These may be PTP "unaware" or "aware." Unlike the former, the latter process the PTP packets and are also known as boundary clocks (BC) or transparent clocks (TC). Taking advantage of PTP, ITU created a powerful set of standards that define network requirements, clock aspects, methods, and profiles for deployment in telecom networks. We will cover those in the next section.

Synchronous Ethernet (SyncE) represents another major synchronization technology. It merged the best of SONET/SDH synchronization capabilities with Ethernet. Stable oscillators, such as Stratum 3E with stability of 4.6 ppm, replace conventional oscillators used in asynchronized Ethernet switches (100 pm). Furthermore, a SyncE receiver can recover a clock from incoming Ethernet frames and synchronize it to its internal oscillator, thereby enabling a node to be synchronized to its upstream node. Finally, Ethernet synchronization message channel (ESMC) delivers a communication protocol for exchange of synchronization quality information between SyncE nodes. These capabilities enable SyncE to be a powerful technology for frequency synchronization.

Technology	Frequency	Time/Phase	Network based?
GPS	Y	Y	
PTP/NTP	Y	Y	Packet layer based
SyncE	Y	N	Physical layer based
E1/E3/DS1/DS3, 2/10 MHz BITS/SSU, SONET/SDH	Y	N	Physical layer based

Table 3.3: Synchronization Technologies

IEEE 1588v2 provides a powerful baseline for different synchronization applications. They include telecom/wireless, audio/video, and utility applications.

In this section, we are focusing on the applications in a telecommunications network. ITU-T has defined a series of standards for synchronization applications (Figure 3.11). Since IEEE 1588v2 covers a wide range of applications and attributes, PTP profiles are defined to permit specific selections of attribute values and optional features of PTP that when using the same transport protocol, interwork and achieve a performance that meets the requirements of a particular application. Examples for telecom profiles are mentioned further below.

ITU-T has defined a series of standards for frequency and time/phase synchronizations that are essential for reliable function of cellular networks (Figure 3.11). These standards capture network requirements, timing characteristics, architecture, and profiles.

G.8260 provides the definitions, terminology, and abbreviations used in frequency, phase, and time synchronization in packet networks. It includes the definitions of various metrics for time error function.

- Time error is defined as the difference between the time of a clock and a reference clock. It is characterized by a number of metrics such as:
  - Constant time error (cTE) which can be estimated by averaging the first M samples of the time error function
  - Maximum absolute time error (Max|TE|) is the maximum absolute value of the time error function

There are two categories of standards for frequency (G.826x) and time/phase (G.827x) synchronization (Figure 3.11).

3.2.2 Frequency Synchronization Standards

This section covers the standards for frequency synchronization. The following section will list those for time and phase synchronization.

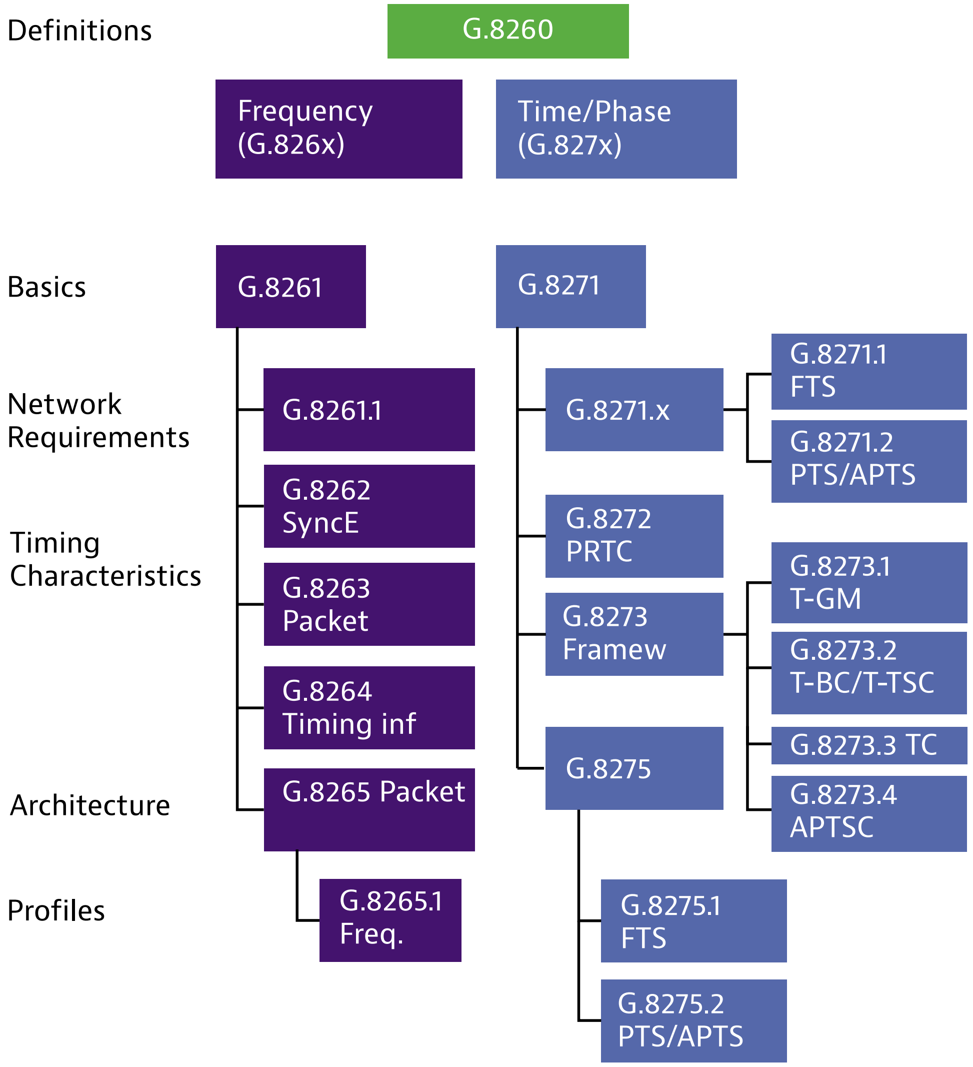


Figure 3.11 Overview of ITU-T synchronization standards

G.8261 (Timing and synchronization aspects in packet networks) focuses on frequency synchronization applications such as reference timing distribution over packet networks: packet networking timing (PNT) domain. The PNT can be delivered by using a distributed system of primary rate clocks (PRC) or by master-slave models such as IEEE 1588v2, or synchronous Ethernet.

G.8261.1 (Packet delay variation network limits applicable to packet-based methods) specifies the hypothetical reference models (HRM) and packet delay variation (PDV) network limits when frequency synchronization is carried via packet networks. Network limits are defined for various network reference points (Figure 3.11) such as:

- packet-based equipment clock master (PEC-M)

- packet-based equipment clock slave frequency (PEC-S-F) There are three categories for network limits in G.8261.1:
- Network limits at the input of PEC-M
- Network limits at the output of PEC-S-F
- PDV limits

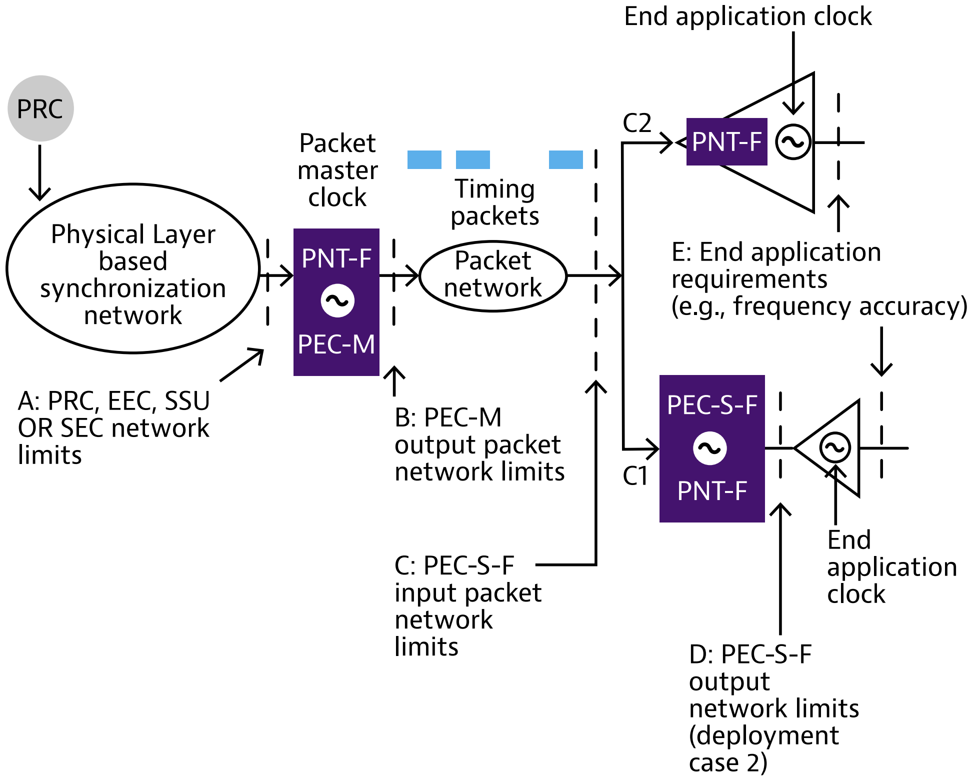

End application clock
PRC
Packet master clock
Timing packets
C2
PNT-F
Physical Layer based synchronization network
PNT-F
PEC-M
Packet network
E: End application requirements (e.g., frequency accuracy)
A: PRC, EEC, SSU OR SEC network limits
B: PEC-M output packet network limits
PEC-S-F
PNT-F
C1
End application clock
C: PEC-S-F input packet network limits
D: PEC-S-F output network limits (deployment case 2)

Figure 3.12 Reference points for network limits G.8261.1

The first two categories deal with clock interfaces and their limits are taken from other standards such as ITU-T G.8261 or G.823/G.824. The PDV network limit represents the maximum permissible levels of PDV at the interface C1/C2 (Figure 3.11). A network is qualified to carry frequency synchronization if it can generate a controlled amount of PDV. To determine the controlled amount of PDV, G.8261.1 introduces the concept of floor packet percentage (FPP). FPP is characterized by a window interval W, and a fixed cluster range δ. A network is qualified if for any window interval W of 200 seconds, at least one percent of transmitted packets will be received within a fixed cluster. The cluster starts at the observed floor delay for the respective window and has a range of 150 µs.

- FPP (n, W, δ) ≥ 1%

G.8262 (Timing characteristics of synchronous Ethernet equipment slave clock) outlines minimum requirements for timing devices used in synchronizing network equipment that supports synchronous Ethernet. The requirements are defined for the following metrics:

- Frequency accuracy
- Pull-in, hold-in, pull-out ranges
- Noise generation defined in terms of jitter and wander at equipment output
- Noise tolerance defined in terms of input jitter and wander that the equipment can tolerate
- Noise transfer
- Transient responses

G.8263 (Timing characteristics of packet-based equipment clocks PEC-M and PEC-S for master and slave functions) outlines minimum requirements for the timing functions of the packet slave clocks (Figure 3.12). The requirements are defined in four measurement categories:

1. Output frequency accuracy < 4.6 ppm (free running condition)
2. Output noise generation
3. PDV noise tolerance: the noise level that the PEC-S-F should tolerate. It is defined under G.8261.1 PDV limits (above).
4. Long-term phase transient response (holdover)

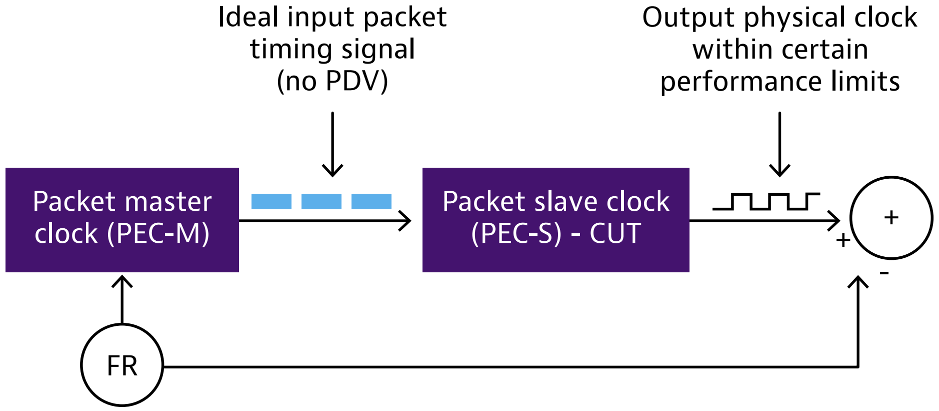

Ideal input packet timing signal (no PDV)
Output physical clock within certain performance limits
Packet master clock (PEC-M)
Packet slave clock (PEC-S) - CUT
+
+
-
FR

Figure 3.13 G.8263 testing procedure for noise generation

G.8265 (Architecture and requirements for packet-based frequency delivery) describes the architecture and requirements for packet-based frequency distribution in telecom networks. The recommendation covers:

- Architecture of packet-based frequency distribution (Figure 3.13)
- Timing protection
- Packet network partitioning
- Packet based protocols including network timing protocol (NTP) and PTP/IEEE 1588v2. It refers to G.8265.1 as the document describing the profile for PTP profile for telecom applications (next section).
- Security aspects

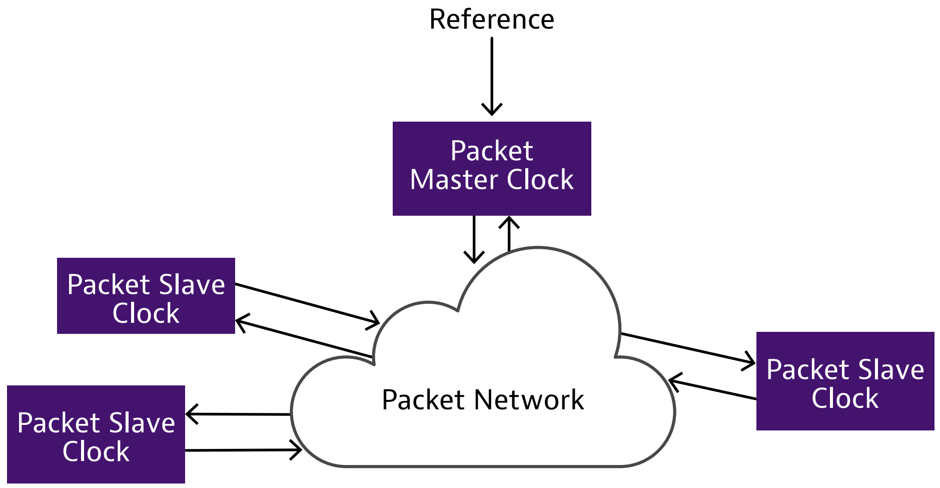


Figure 3.14 General packet network timing architecture

G.8265.1 (Precision time protocol telecom profile for frequency synchronization) specifies the PTP functions that are necessary to ensure network element interoperability for the delivery of frequency only. Some of the highlights of the profile are:

- No on-path support (no requirement for use of boundary or transport clock)
- IP/layer 3 network layer
- Announce message carry quality level (QL) that was defined in G.781 and are used in SONET/SDH and SyncE synchronization status messages (SSM)
- Unicast transmission
- Static provisioning (instead of best master clock algorithm BMCA)
- Message rates (Table 3.4 below)

Message Rates	Minimum	Maximum	Default
Announce	1 msg every 16 sec	8 msg/s	1 msg every 1s
Sync	1 msg every 16 sec	128 msg/s	Not defined
Delay request	1 msg every 16 sec	128 msg/s	Not defined

Table 3.4: G.8265.1 PTP message rates

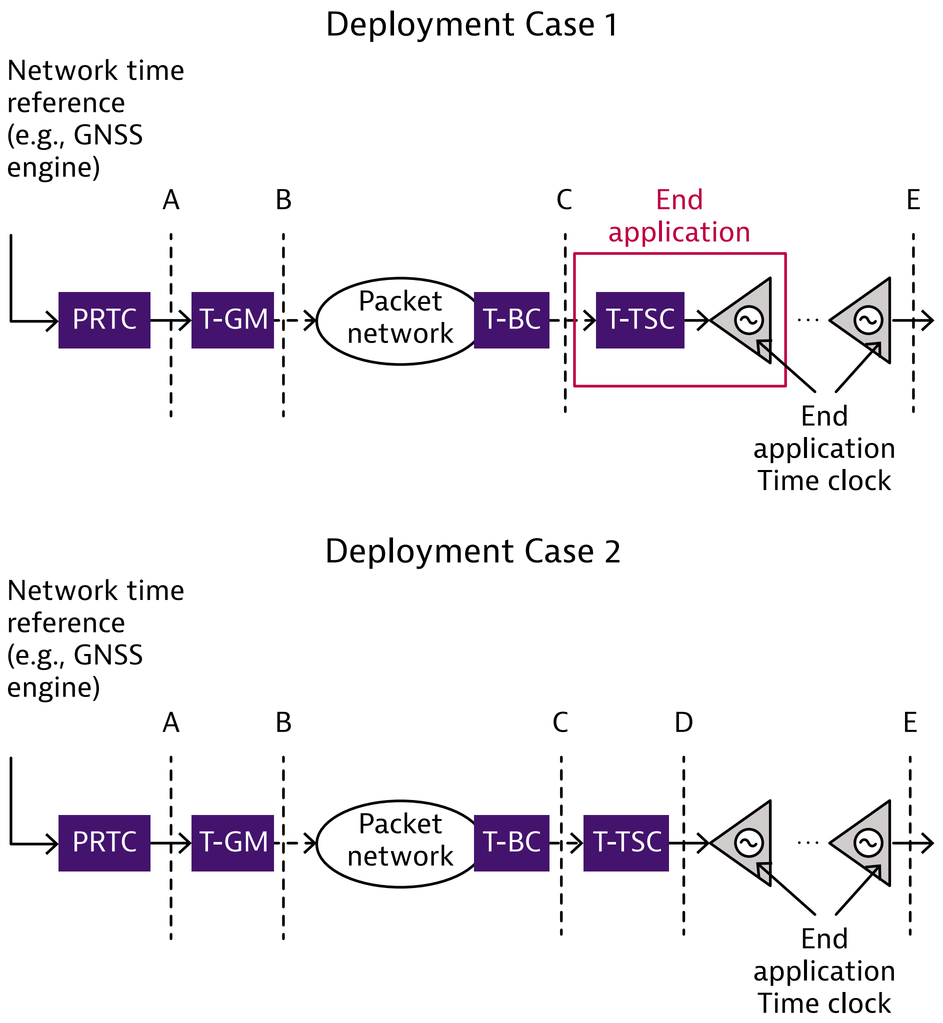
3.2.3 Time and Phase Synchronization Standards

The time and phase synchronization standards follow a similar pattern in numerology as frequency standards but cover a larger range of topics due to the broader range of equipment and system design aspects to be considered.

G.8271 (Time and phase synchronization aspects of packet networks) defines time and phase synchronization aspects in packet networks.

G.8271.1 (Network limits for time synchronization in packet networks) specifies the maximum network limits of phase and time error. It specifies the minimum equipment tolerance to phase and time error that shall be provided at the boundary of packet networks at phase and time synchronization interfaces. It also outlines the minimum requirements for the synchronization function of network elements. This recommendation addresses the use case for full timing support. Full timing support means that all network elements between the PTP master and the slave are PTP aware. The limits are defined for two main cases (Figure 3.14):

- Deployment case 1: the telecom time slave clock (T-TSC) is integrated in the end-application
- Deployment case 2: T-TSC is external to the end application



Deployment Case 1

Network time reference (e.g., GNSS engine)
A
B
C
End application
E
PRTC
T-GM
Packet network
T-BC
T-TSC
End application Time clock

Deployment Case 2

Network time reference (e.g., GNSS engine)
A
B
C
D
E
PRTC
T-GM
Packet network
T-BC
T-TSC
End application Time clock

Figure 3.15 G.8271.1 Time synchronization deployment cases

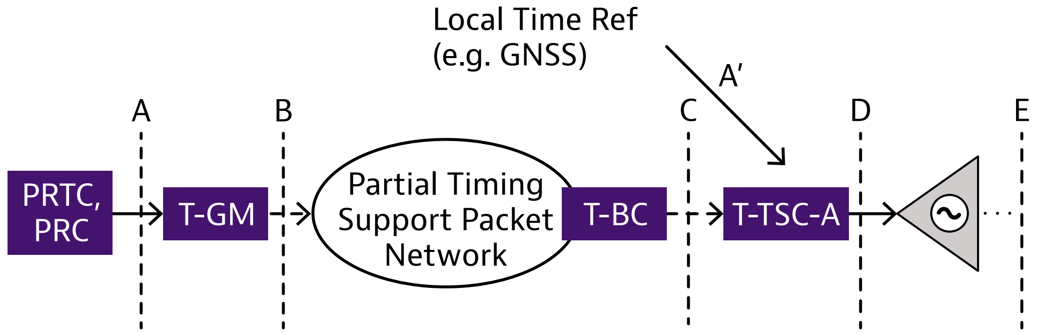
G.8271.2 (Network limits for time synchronization in packet networks with partial timing support from the network) defines the network limits for the case of a packet method with partial timing support that is characterized by two cases:

- Assisted partial timing support (APTS) in which case PTP is used as a backup to a local time refence based on the global navigation satellite system (GNSS) (Figure 3.15)
- Partial timing support (PTS) in which case PTP is the primary source of time (Figure 3.16)

Network limit at reference points A and A′: They are defined in G.8272 (below). In particular,

- Max |TE| < 100 ns

Network limit at reference point B: The limits are the same as for reference points A/A′, if the T-GM is integrated in the Primary Reference Telecom Clock (PRTC). For external Telecom Grandmasters T-GM, the limits are for further study.



Local Time Ref (e.g. GNSS)
A′
A
B
C
D
E
PRTC, PRC
T-GM
Partial Timing Support Packet Network
T-BC
T-TSC-A

Figure 3.16 APTS

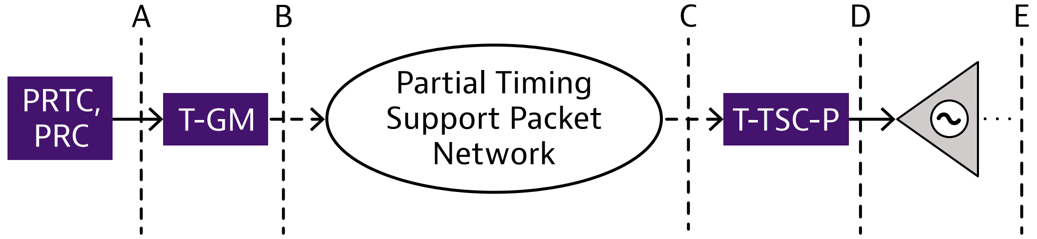


Figure 3.17 PTS

G.8272 (Timing characteristics of primary reference time clocks PRTC) specifies the requirements for PRTCs suitable for time, phase, and frequency synchronization in packet networks. A typical PRTC provides the reference signal for time, phase, and frequency synchronization for other clocks within a network or section of a network. This recommendation defines the PRTC output requirements. The accuracy of the PRTC should be maintained as specified in this recommendation. The recommendation also covers the case where a PRTC is integrated with a T-GM clock. In this case it defines the performance at the output of the combined PRTC and T-GM function.

The recommendation contains the time error, wander, and jitter requirements in locked mode, as well as the holdover requirements. In particular the recommendation defines:

- Time error of 100 ns against an applicable primary reference such as universal time clock (UTC)
- MTIE (maximum time interval error) and TDEV (time deviation)

G.8273 (Framework of phase and time clocks) is a framework recommendation for phase and time clocks for devices used in synchronizing networks defined in G.827x series of recommendations (above). This recommendation refers to a series of G.8273.x recommendations further described below. While the IEEE 1588v2 defines equipment such as OC, BC, TC, and GM, the ITU-T G.8273.x series define the following type of devices. The following devices include not only the specification of respective IEEE 1588v2 devices, but also contain additional performance characteristics outlined in sections further below in this paper:

- Telecom grand master (T-GM) in G.8273.1
- Telecom boundary clock (T-BC) in G.8273.2
- Telecom transparent clock (T-TC) in G.8273.3
- Telecom slave clock (T-TSC) in G.8273.2
- Assisted partial timing support slave clock (APTSC) in G.8273.4

G.8275 (Architecture and requirements for packet-based time and phase distribution) describes the general architecture of time and phase distribution using packet-based methods. This recommendation forms the base architecture for the development of telecom profiles for time and phase distribution.

G.8275.1 (Precision time protocol telecom profile for phase/time synchronization with full timing support A APTfrom the network) contains the ITU-T precision time protocol (PTP) profile for phase and time distribution with full timing support from the network. It provides the necessary details to utilize IEEE 1588 in a manner consistent with the architecture described in recommendation ITU-T G.8275. Some of the highlights of the profile are:

- Ordinary clocks (OC), boundary clocks (BC), and transparent clocks (TC) are used in this profile
- PTP over IEEE 802.3/Ethernet
- Multicast
- Alternate best master clock algorithm (BMCA)

G.8275.2 (Precision time protocol telecom profile for phase/time synchronization with partial timing support from the network) is pre-published at the time of this book.

The synchronization aspects can principally be applied in 5G networks, although with more demanding limits that fuel the need for higher performing synchronization equipment and network designs. Also, we need to keep in mind that 5G will not replace in one step all the 4G network. Hybrid networks will be the reality for the foreseeable future and so is the need for understanding 4G synchronization technologies.

3.3 5G Transport Networks

Understanding the requirements for emerging 5G transport networks necessitates a deeper look at service quality needs of emerging 5G services (Figure 1.9). They are broadly classified into three categories of service requirements:

- Enhanced Mobile Broadband (eMBB) provides greater data-bandwidth services with peak data rates of 10 Gbps and beyond. This data rate will enable new use cases such as augmented reality/virtual reality or ultra-high density (UltraHD) applications.
- Ultra-Reliable Low Latency Communications (URLLC) provides ultra-reliable capabilities with availabilities in the range of 99.9999%, and extremely low latency features in millisecond range. Vehicle-to-vehicle communication over 5G networks is one prominent use case for this category.
- Massive Machine Type Communication (mMTC) supports extremely large number of devices in the range of hundreds of thousands per square kilometer. For this application class, it is also essential to have battery lifetimes in the range of ten years.

These three categories pose different requirements for the underlying networks and applications:

- eMBB demands much higher bandwidth availability from the network for the UE
- URLLC necessitates extremely low latency in the network design for the relevant network components and their interconnecting transport network
- mMTC requires networks that can serve a very large number of end points in a power efficient manner

The following sections describe the implications of the above orthogonal requirements for the transport network design.

3.3.1 Fronthaul Challenges and Functional Split Options

Addressing the emerging 5G service requirements demanded a new way of partitioning the network functions in the radio access networks. To understand the reason for this partitioning requires an understanding of the limitations of the current 4G network technology. We will start with the fronthaul technology.

While CPRI continues to be a mainstream technology for fronthaul technology, it is bandwidth inefficient and cannot scale for 5G massive broadband services as the required bandwidth and antennas would push the CPRI bandwidth requirements above 100 Gbps (Table 3.5).

Antenna	10 MHz	20 MHz	100 MHz
1	0.49 Gbps	0.98 Gbps	4.9 Gbps
2	0.98 Gbps	1.96 Gbps	9.8 Gbps
4	1.96 Gbps	3.92 Gbps	19.6 Gbps
64	31.36 Gbps	62.72 Gbps	313.6 Gbps

Table 3.5: CPRI Bandwidth as a function of bandwidth and antenna ports (excluding line coding)

These bandwidth allocations would be extremely expensive for larger network rollouts. To develop an alternative solution necessitates an analysis of the key functional elements between a BBU and RRH (Figure 3.17). RRUs for 4G implement RF functions, while the other main functions are placed in the BBU. This functional distribution allows operators to centralize most of the functions at one location and have a basic lower cost radio at each end point (option 8).This centralization is also an enabler for resource pooling, which optimizes the utilization of resources. Furthermore, the architecture provides some key functions for advanced LTE technology. Being able to coordinate multiple radios from one location is a key enabler for implementing features such as coordinated multipoint (CoMP), which helps raise user bandwidth by aggregating traffic sourced from multiple cells at the user terminal. All these advantages come with a massive disadvantage for emerging 5G services: inefficient bandwidth use.

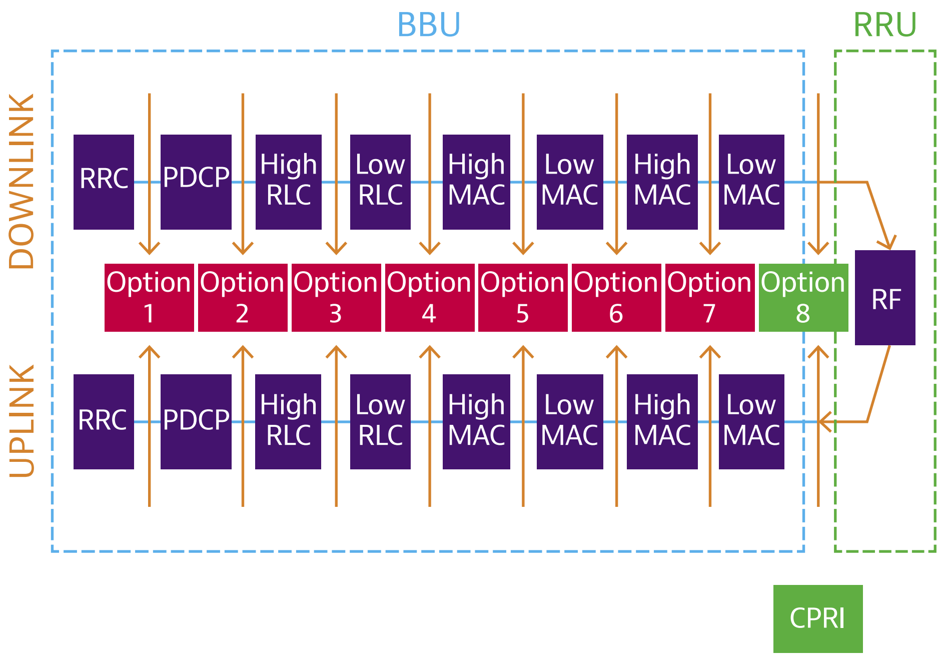


Figure 3.18 Functional split options

3.3.2 Higher Layer Split (HLS), Lower Layer Split (LLS), and eCPRI

Beyond the key disadvantage of bandwidth inefficiency there is another drawback. CPRI has a very limited delay budget. In practice, this means that the distance between BBUs and RRHs will be very limited. The distance is determined by the delay budget and the type of transport technology deployed in fronthaul. Dark fiber is the simplest one allowing for maximum distance. Transport equipment that contains some processing elements reduces the delay budget, sometimes substantially as with Optical Transport Networking (OTN). As is often the case, operators would need to look at the individual use case and conduct a trade-off analysis to determine the proper transport technology. Availability of fiber and equipment rooms, as well as the number and locations of radio end points, are some key factors in this trade-off analysis.

One use case of emerging 5G networks is fixed wireless access in which operators use 5G technology to deliver high-bandwidth broadband services to customers in fixed locations. This use case can be considered an alternative to other fixed wireline applications such as fiber to the home (FTTH), or residential cable services. In this application, coordination of multiple radios is not a necessity. The priority is delivery of high-capacity services that can require bandwidths in excess of 100 MHz.

For these applications, a higher layer split (HLS) option is recommended (Figure 3.18). This option places most of the functions inside the radio unit and can also be considered as a distributed unit (DU)/radio unit (RU) functional element. This placement significantly reduces the bandwidth at the HLS option interface. 3GPP recommends option 2 for HLS. This interface is also known as the new F1 interface. Beyond significant reduction of the bandwidth, the delay budget is in the range of several milliseconds, much higher than CPRI (fronthaul) interfaces. This budget allows the central unit (CU) to be located dozens of miles away from the DU/RU element. This segment of the network is called midhaul as it sits between fronthaul and backhaul.

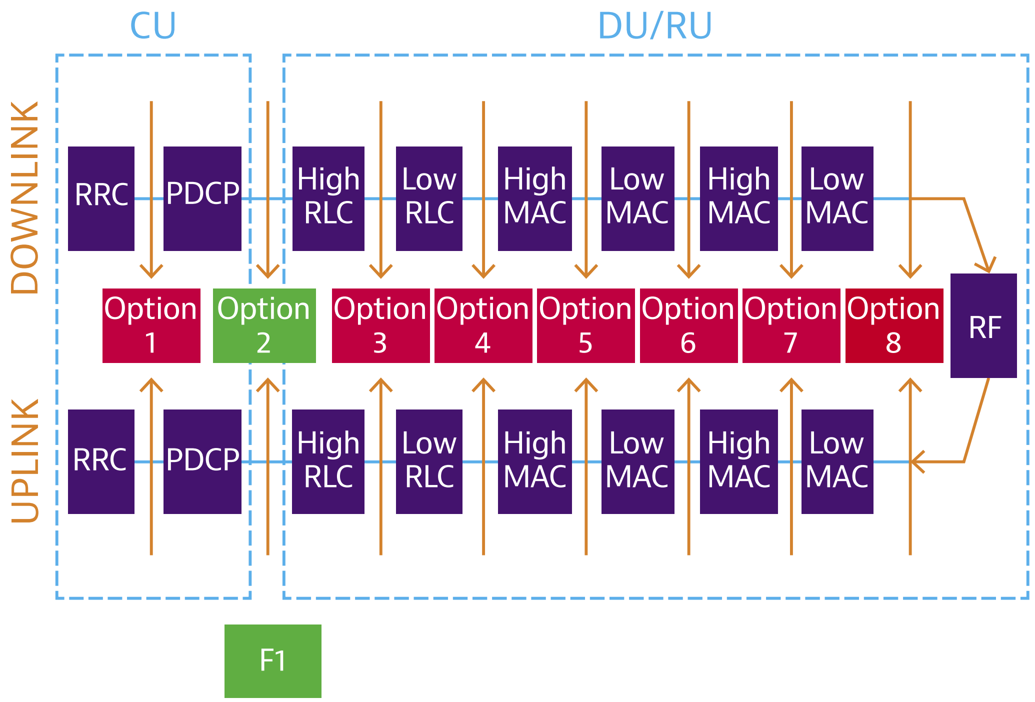


Figure 3.19 Higher layer split (HLS) option

Beyond fixed broadband services, massive mobile broadband services are expected to take advantage of advanced mobility applications that require coordination of multiple radios. This capability requires a lower layer functional split option that leaves most of the functional elements (Figure 3.19) in a centralized location coordinating the radios. Options 6 and 7 of the standards are currently being considered for this use. For this same use case, the CPRI organization published the first eCPRI specification in 2017.

The eCPRI technology is based on a functional split in the PHY component. PHY includes several functions as depicted in Figure 3.20. The eCPRI specification recommends that the split option I_U is used for uplink, and either II_D or I_D is deployed for downlink. In eCPRI, those entities are called eCPRI radio equipment control (eREC) and eCPRI radio equipment (eRE) as depicted in Figure 3.21.

Three planes are necessary for interaction between eREC and eRE: user plane, sync plane, and control and management (C&M) plane. The eCPRI standard defines the user plane and refers to other standards for the definition of the other planes. For example, an operator is free to choose precision timing protocol (PTP) or global positioning system (GPS) for synchronization, both in hybrid mode and other synchronization methods.

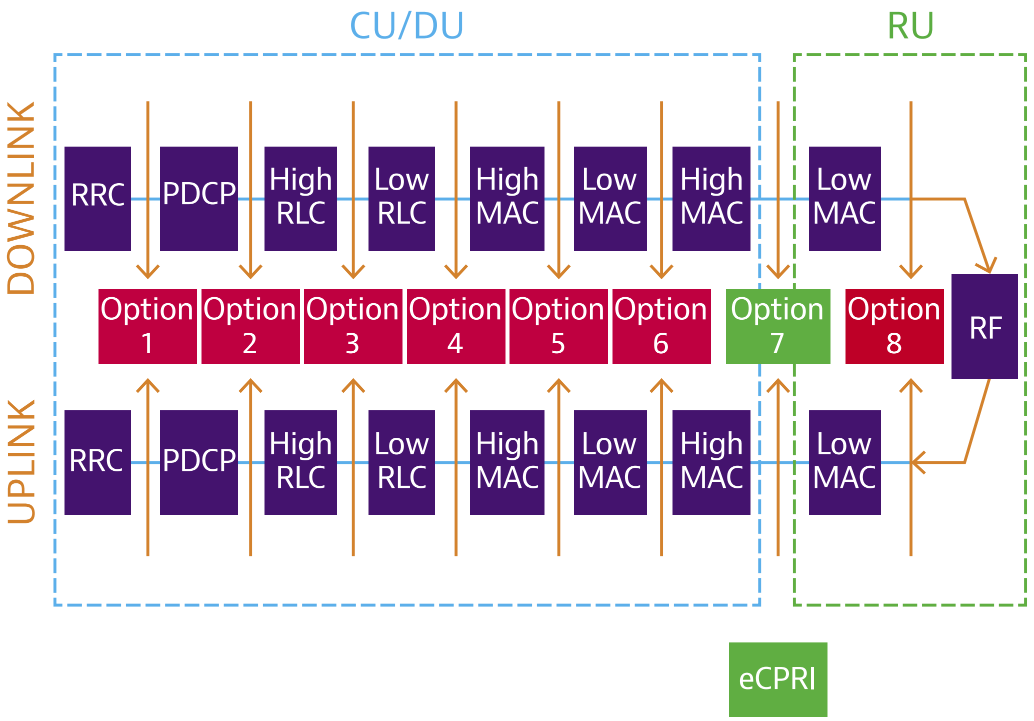


Figure 3.20 Lower layer split option

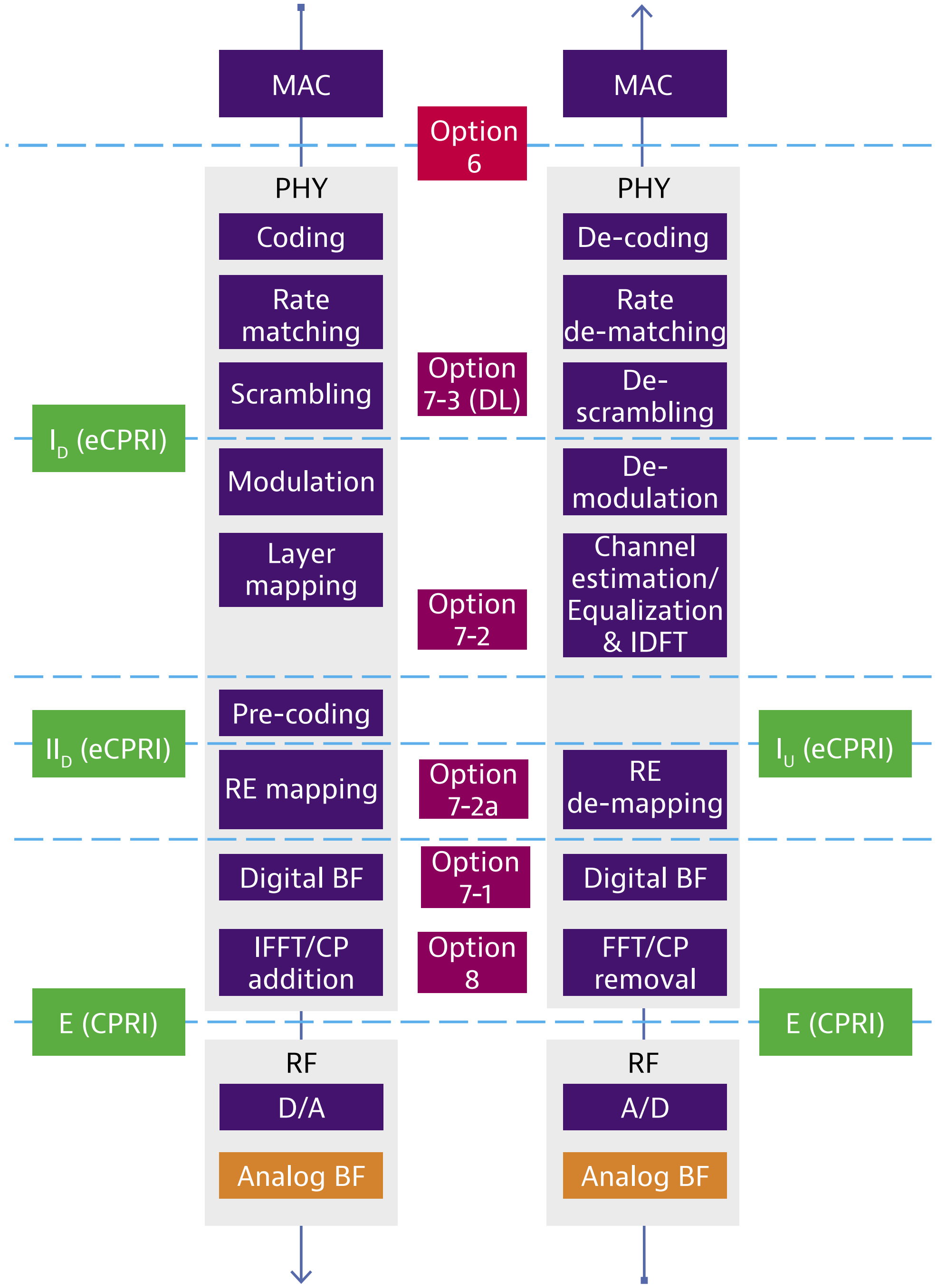


Figure 3.21 Functional split in Phy

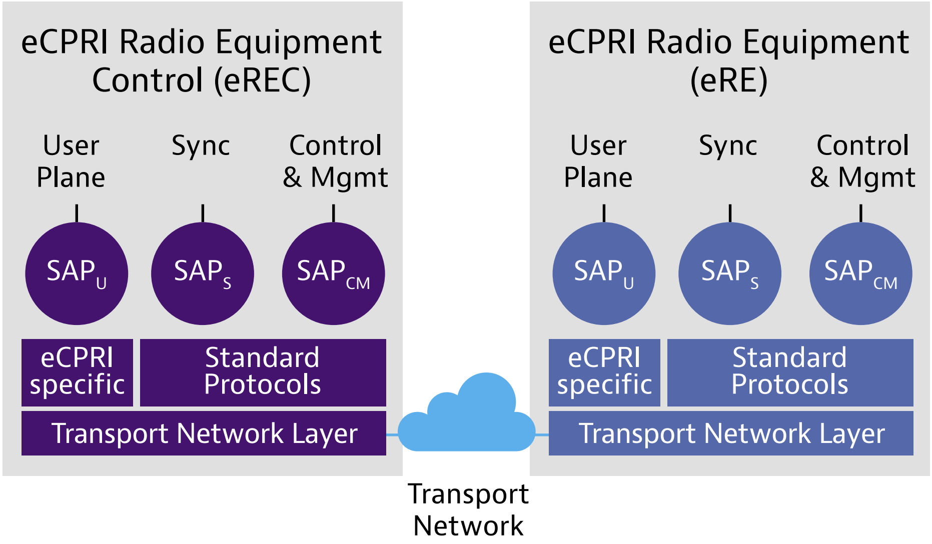


Figure 3.22 eCPRI protocol layers

eCPRI also mentions packet-based technologies for the transport of the user plane. Both Ethernet (layer 2) and Ethernet/IP/UDP (layer 2/3/4) are possible. For the physical layer, eCPRI refers to Ethernet rates 10 Gbps to 100 Gbps. The frame format is based on using an Ethernet or Ethernet/IP/UDP frame that uses the unique EtherType of $AEFE_{16}$. The frame includes an eCPRI header that follows layer 2 or layer 2/3/4 header and is followed by the eCPRI payload. eCPRI defines several message types for the payload listed in Table 3.6.

Message Type #	Name	Section
0	IQ Data	3.2.4.1
1	Bit Sequence	3.2.4.2
2	Real-Time Control Data	3.2.4.3
3	Generic Data Transfer	3.2.4.4
4	Remote Memory Access	3.2.4.5
5	One-way Delay Measurement	3.2.4.6
6	Remote Reset	3.2.4.7
7	Event Indication	3.2.4.8
8-63	Reserved	3.2.4.9
64-255	Vendor Specific	3.2.4.10

Table 3.6: eCPRI Message Types

The most significant part of the user plane is given by IQ data or bit sequence, with the former for split options I_U/I_D, and the latter for split option II_D. Since split option E is very bandwidth intensive, most IQ data deployments are expected to be based on split option I_U/I_D. The IQ data or bit sequence are carried in association with their respective real-time control data that contains vendor specific information between PHY processing elements in eREC and eRE. The above options rely on a single-split configuration. There are also good reasons to have a double-split-option (Figure 3.22). URLLC applications require extreme fast responses from a network. Vehicle to network (V2N) applications need response times in the range of a few milliseconds from vehicle to vehicle. This does not leave much budget for the cellular network if the two vehicles communicate over two RUs. This use case is a good example of cases that would benefit from a double-split design that separates the DU and CU. While the time critical functions in DU can be placed closely to the RU, and thereby help meet the low latency requirement, the non-time-critical functions can be placed farther away in a central location.

3.3.3 Timing Sensitive Network (TSN)

While the initial deployments of 5G transport networks are expected to deploy dark fiber and WDM technologies, these technologies will not be scalable for large-scale deployments. Given eCPRI use of Ethernet transport layer, switched Ethernet technologies seem to be a logical way to increase efficiency and reach scalability. Conventional layer 2 or layer 3 switched technologies are, however, not appropriate for the transport of fronthaul traffic due to stringent quality of service requirements.

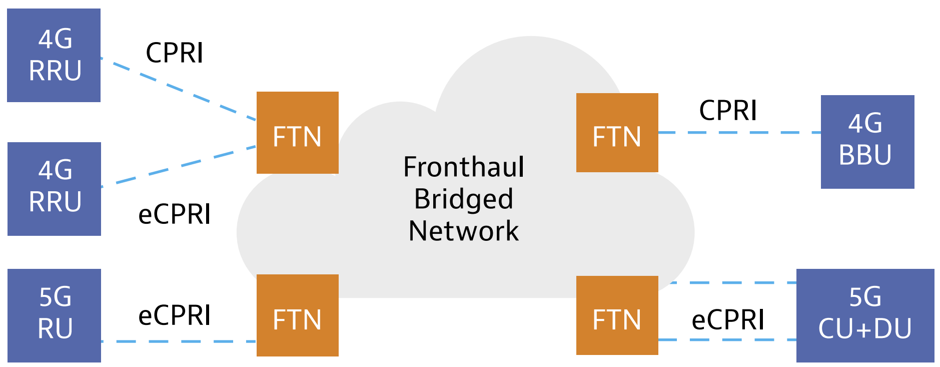


Figure 3.23 Fronthaul bridged network

To address these requirements, standards organizations have been developing standards for a fronthaul-friendly network design (Figure 3.23). IEEE 802.1cm selects features and options for fronthaul traffic by describing fronthaul requirements and synchronization requirements for two classes of networks that are distinguished by the functional split of the classes:

- Class 1: functional split 8 (CPRI)
- Class 2: functional split 7 (eCPRI)

Class 1 use cases revolve around transporting CPRI traffic. As we have seen in previous sections, CPRI traffic is composed of different flows such as IQ and C&M data. These flows have different quality of service requirements (Table 3.7) and can be prioritized and transported separately through a bridged network.

Flow	Latency	Frame Loss Ratio
IQ	100 µs	10^{-7}
C&M	No requirement	10^{-6}

Table 3.7: Requirements for Class 1 (CPRI) traffic (IEEE 802.1cm)

Synchronization is always provided independent of IQ and C&M flows. An example for synchronization mechanism is the PTP G.8275.1 (Full Timing Support) mentioned in the synchronization section above.

Class 2 refers to eCPRI (functional split 7) traffic. eCPRI specification lists two classes of user plane (Table 3.8). A subset of use cases may deploy the slow User Plane profile. However, the majority of use cases are expected to use the more stringent User Plane traffic profile. For C&M, there are two categories. The fast category has more stringent requirements than the regular C&M category. For the three use cases mentioned above, three distinct Class of Service (CoS) categories—low, medium, and high—are defined. These requirements are particularly listed for split options E, I_D, II_D, and I_U.

CoS Name	Flow	Latency	Frame Loss Ratio
High	User Plane (fast)	(see Table 3.9)	10^{-7}
Medium	User Plane (slow) and, C&M Plane (fast)	1 ms	10^{-7}
Low	C&M	100 ms	10^{-6}

Table 3.8: Requirements for split options E, ID, IID, and IU (eCPRI Transport Networks)

The high category is further refined in four latency classes, High25 through High500 (Table 3.9). The traditional LTE (Evolved Universal Terrestrial Radio Access E-UTRA) applications necessitate a tight budget for one-way delay in fronthaul networks as we saw in class 1 requirements above. The same is true for 5G NR radios. However, the transport networks are not necessarily designed only for mobile applications. For Ultra-Reliable Low Latency applications, the delay budget is even stricter. And yet other applications have less stringent requirements.

Flow	Latency	Use case
High25	25 µs	Ultra-Reliable Low Latency
High100	100 µs	E-UTRA and NR
High200	200 µs	For installation up to 40 km
High500	500 µs	Large latency installations

Table 3.9: CoS High Latency Requirements for split options E, ID, IID, and IU (eCPRI Transport Networks)

3.3.4 Emerging 5G Synchronization Requirements

The synchronization requirements are derived from several bodies including the 3GPP in its technical specification series 36.xxx and 38.xxx for 4G and 5G services, respectively. The technical specifications 36.104/38.104 represent two key documents that describe base station radio transmission and reception requirements. More specifically, section 6.5 (transmit signal quality) lists several requirements that are essential for synchronization network design including Time Alignment Error (TAE). TAE is defined as the largest timing difference between any two signals belonging to different antennas or transmitter groups. The requirements are categorized dependent on the wireless use case (Table 3.10). These use cases are assigned unique categories from A+ to A, B, and C. The use cases at the bottom of the table are being developed at this time and have not been assigned a category.

3GPP Feature	RAN	
	LTE	NR
MIMO or TX-diversity transmission	Category A+	Category A+
Intra-band contiguous carrier aggregation	Category A	BS Type 1: Category B BS Type 2: Category A
Intra-band non-contiguous carrier aggregation	Category B	Category C
Intra-band carrier aggregation	Category B	Category C
TDD	Category C	Category C
Dual Connectivity	Category C	Category C
COMP	Not specified in 3GPP	Not ready in 3GPP
Supplementary Uplink	Not applicable for LTE	Not ready in 3GPP
In-band Spectrum Sharing	Not ready in 3GPP	Not ready in 3GPP
Positioning	Not specified in 3GPP	Not ready in 3GPP
MBSFN	Not specified in 3GPP	Not ready in 3GPP

Table 3.10: Timing Accuracy Categories (eCPRI Transport Requirements)

Category A+ demands the most stringent synchronization requirements (Table 3.11) while category C's requirement is in line with current LTE backhaul networks. The requirements are identified in terms of relative and absolute Time Error (TE). The relative TE specifies the time error between any two RU or eRE. Absolute TE is the time error measured against a reference of the primary reference time clock (PRTC). In most cases the absolute TE requirements are in addition to the one for respective relative TE requirements (categories A+, A, and B). But there are some variations in the range of time error requirements. Those are necessary to account for different implementations of the telecom time slave clock (T-TSC). For example, if the T-TSC is integrated in an eRE/RU, then the time error limit is lower than the case of an external T-TSC function.

Category	Time Error
A+ (relative)	20-32 ns
A (relative)	60-70 ns
B (relative)	100-200 ns
C (absolute)	1100 ns

Table 3.11: Time Error Requirements

3.3.5 Radio over Ethernet (RoE) and Fronthaul Gateways (FHGW)

As with previous generations of wireless networks, the introduction of 5G radio access networks will overlap with the continuous deployment of 4G RAN. This overlap certainly raises the need for a converged transport network that can address the needs of both RAN technologies, e.g. transporting CPRI and eCPRI links over the same physical infrastructure (Figure 3.24). This example concerns bringing together the 4G and 5G fronthaul networks, but there is no reason why this convergence should be limited to fronthaul networks. It is reasonable to expect a convergence of fronthaul, midhaul, and backhaul networks, and in the future a convergence with fixed wireline service. In its simplest form, this converged network can be based on a WDM solution as depicted in Figures 3.4 and 3.5. Ethernet can also be considered as a cost-effective alternative. eCPRI is already based on Ethernet transport layer. To enable the transport of CPRI over Ethernet, the radio over Ethernet (RoE) standard was developed in IEEE 1914.3/1914.1 working groups.

RoE encapsulates CPRI signals into Ethernet frames; there are three variants of RoE:

1. Structure agnostic
2. Structure aware
3. Native encapsulation

Structure agnostic is the simplest method. It has basic knowledge of the client signal. It includes the type of line coding that is used. Structure aware is partially aware of the protocol used and is more efficient than the previous method. The most efficient method is RoE with native encapsulation, which transfers only the time or frequency domain IQ data along with control and management data. In summary, the complexity and cost can increase from structure agnostic towards native encapsulation but gains more efficiency with the latter approaches.

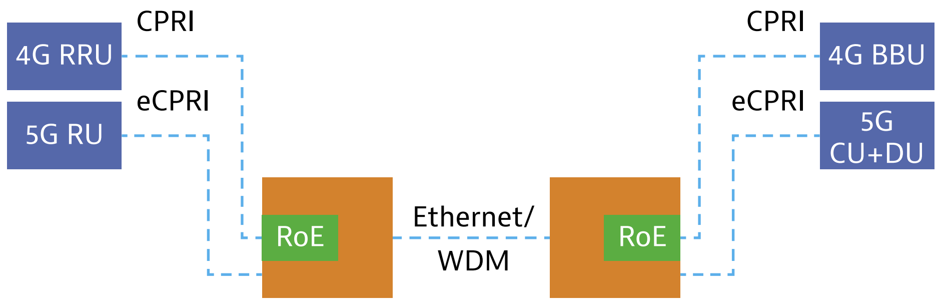


Figure 3.24 Converged Transport Network

With RoE, operators can create a converged fronthaul network for CPRI based and eCPRI based fronthaul networks. CPRI radios will remain in operator networks for many years due to their simplicity. On the BBU/DU side, the need for CPRI will yield to eCPRI over time as operators move towards a virtualized BBU/DU infrastructure with an eCPRI/Ethernet PHY interface.

But how can we connect an eCPRI BBU/DU with a CPRI radio? Fronthaul Gateways (next Figure below) promise to deliver an answer to this question by implementing a Lower Phy (L-PHY) function which translates the CPRI signal (split option 8) to eCPRI (split option 7-2a) as depicted in Figure 3-20 (Functional Split in Phy). This translation not only enables to connect a CPRI radio with a virtualized BBU/DU environment, but also significantly reduces the bandwidth on the fronthaul as the L-Phy function assumes several processor-intensive functions such as FFT/iFFT in the FHGW. This translation also creates a fronthaul link whose bandwidth is proportional to the user traffic (variable bit rate) unlike CPRI (constant bit rate) and therefore delivers an additional level of efficiency.

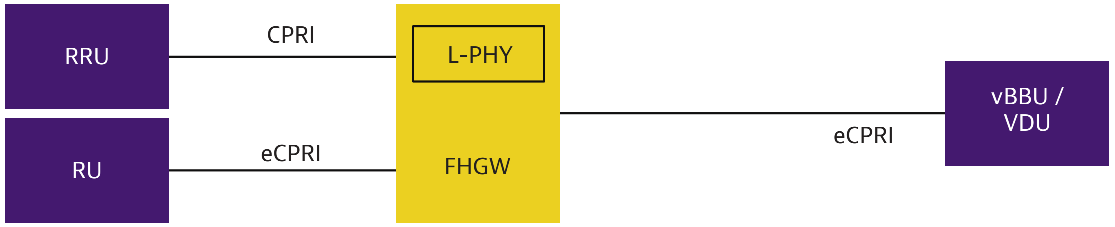


Figure 3.25 Fronthaul Gateways

3.4 5G RAN and Network Slicing, Summary and Outlook

Network slicing allows operators to offer different categories of services with a wide range of service requirements on a common, shared physical network. Figure 3.25 illustrates the example of deploying one physical network to serve an eMBB and an Ultra-Reliable Low Latency application at the same time. Whereas the former necessitates the use of a double split (options 2 and 7) architecture between the 5G core (NGC) and antenna, the later needs to place the core functions closer to the edge to meet the tight latency requirements.

Emerging 5G services demand different SLAs for eMBB, URLLC, and mMTC applications. While eMBB challenges the bandwidth inefficiency of existing fronthaul technologies, URLLC applications require Ultra-Reliable Low Latency networks, and mMTC demands a network that can manage a very large number of end points in a power-efficient manner. These new challenges have led to the consideration of new ways of splitting critical baseband and radio functions.

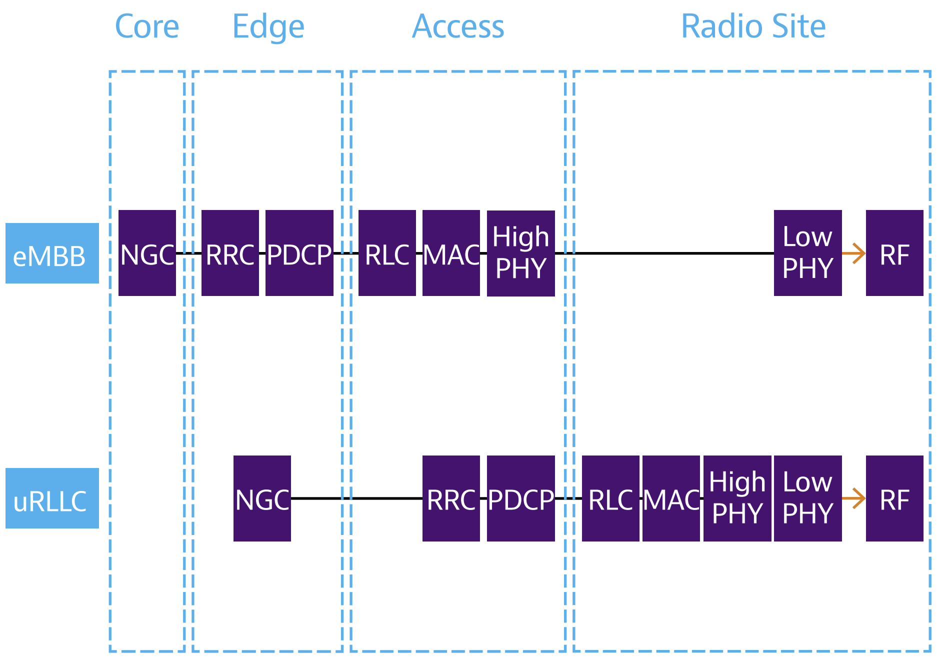


Figure 3.26 Network slicing for eMBB and uRLLC applications (Control Plane stack)

Proper network design requires careful analysis of various SLAs associated with the above functional split options and use cases. They are characterized by latency, frame loss ratio, and time error metrics. Considering the diversity of the use cases and SLAs, 5G transport networks can be economically viable only if they are designed on a single converged physical network. In addition, network slicing enables the deployment of multiple services with distinct SLAs on a single physical network.

Finally, the design of 5G radio access network can be scaled when considered in the context of a converged network that brings 5G and 4G fronthaul, midhaul and backhaul networks together. Initial converged networks are taking advantage of WDM technology. While sufficient for the initial deployment, massive deployment of 5G radios will necessitate an economic and ubiquitous technology such as Ethernet. To allow for the convergence of legacy CPRI-based and new Ethernet-based network technologies, the RoE standard can be deployed in fronthaul networks. Finally, an Ethernet-based technology can be most fruitful if its statistical multiplexing gains are effectively used. Taking advantage of this multiplexing gain can be realized only with a careful analysis of 5G latency requirements. Timing-sensitive networking is the ultimate goal of a cost-effective and massively scalable converged-access network.

CHAPTER FOUR

5G Core

4.1 Mobile Core History

The first-generation mobile network architecture relied on circuit switching (CS) to establish end-to-end service—in most cases a voice call—between devices. With GPRS, the addition of packet switching (PS) network architecture enabled efficient data transport in the form of packets. Voice and SMS were still carried over CS. In the third generation, the duality of CS and PS remained in the core network with dedicated elements for each mode.

With the fourth generation, the 3GPP community decided that Internet Protocol (IP) would be the best medium for transporting all services, eliminating the need for dual modes in the core network. This development was labeled the Long-Term Evolution (LTE). This new core was designed to enhance and evolve the last two generations' packet switching network and was called the Evolved Packet Core (EPC), introduced in 3GPP release 8.

4.1.1 LTE EPC Components

As shown in Figure 4.1, the main components of the 3GPP release 8 EPC are: MME (Mobility Management Entity): The MME is responsible for managing the signaling and control plane related to mobility and security of the LTE network. It tracks and handles the user equipment (UE) both in active and idle modes.

HSS (Home Subscriber Server): The HSS is a database of all subscribers' authentication data. It plays an important role in session setup to make sure that the user is authorized to use the service and set all parameters of the session related to the subscription profiles.

S-GW (Serving Gateway): S-GW serves as the main gateway between the UE (User Equipment) and the rest of the core network when it comes to transporting IP packets. It is also the main point for handover between different eNodeBs (eNBs) and also with other 3GPP access networks.

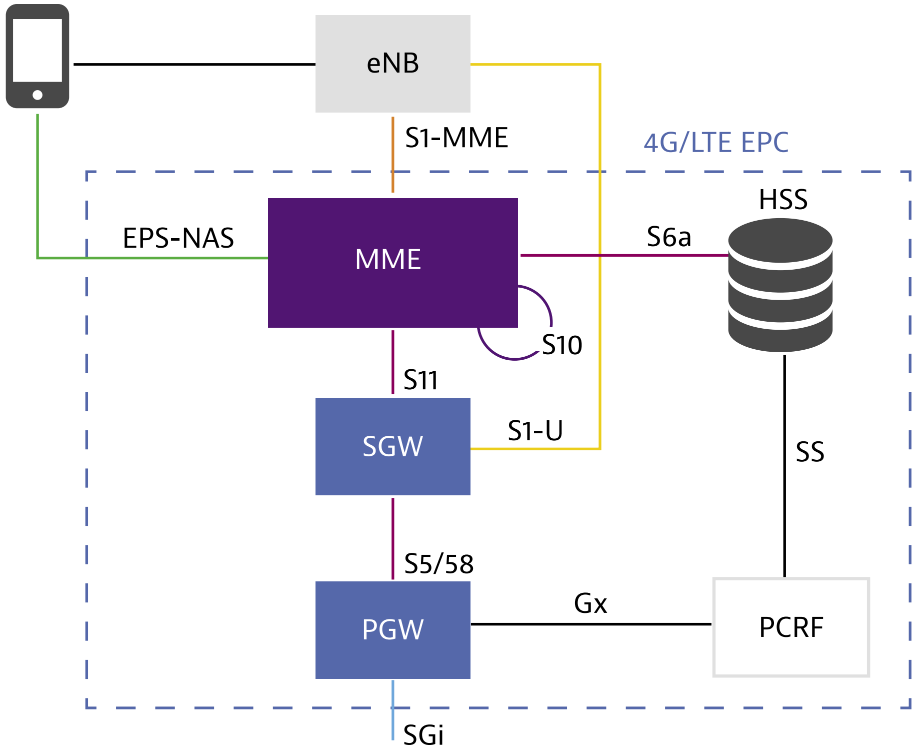


Figure 4.1 Simplified view of 4G LTE Evolved Packet Core (EPC)

P-GW (PDN Gateway): P-GW serves as the gateway between the User Agent and external Internet or data networks. It is the main router of IP packets to and from the external data networks. In many cases the P-GW and S-GW are combined into one physical element; however, they are still independent logical functions.

PCRF (Policy and Charging Rule Function): PCRF accesses the HSS and supports input into real-time enforcement of decisions and actions in accordance with policy rules related to the subscription profile, differentiated services, and charging premiums.

4.2 The Case for Next Generation Core (NGC)

Mobile core networks up to 4G have been designed and optimized for Mobile Broadband (MBB) services particularly around human communications—connecting humans, and humans to content.

In recent years, while MBB data rate demands have kept increasing, ARPU (Average Revenue Per User) from the subscriber-centric business has generally been declining. To overcome this decline, operators have no choice but to address new business opportunities through providing connectivity for the new IoT-enabled vertical industry segments, such as automotive, healthcare, smart city, smart utilities, and others.

As the Verticals have very different connectivity requirements compared to humans, connecting things is fundamentally different than connecting people. Also, different industry verticals will have different connectivity requirements in terms of latency, bandwidth, performance, densification, etc. Many of these requirements are also difficult to predict and anticipate as there are always new industries and use cases with previously unforeseen network requirements.

Operators globally have been custom designing, deploying, and maintaining multiple instances of 4G core networks, one each for applications with specific SLA requirements; this approach is neither scalable nor sustainable for addressing industry verticals needs. Such examples include mission-critical services networks, private networks for power grids, and massive Internet of things (MIoT) connectivity services for different applications.

There have as well been a few major attempts by 3GPP to enhance the 4G EPC in order to help operators provide a more efficient connectivity fabric optimized for specific vertical industries and use cases.

A few of the significant examples include addressing the MIoT opportunities needing operators to provide low-power, wide-area networks as a connectivity fabric for a massive number and density of low-power IoT devices (more than a million per km2). Smart meters are IoT devices performing infrequent packet transmissions and receptions, typically with relatively small payloads. This requires the mobile core to be highly efficient, reliable, and capable of scaling rapidly in a very cost-effective manner.

3GPP as part of Release 13 has standardized a variant of 4G Core—the Cellular IoT (CIoT) Serving Gateway Node (C-SGN) that has been optimized for MIoT use case requirements. The C-SGN is a simplified core network that can be deployed on cloud-based platforms in order to meet the rapid, cost-effective scaling requirements. Further enhancements were made in 3GPP Release 14 and15. CIoT and C-SGN will be further discussed in Chapter 7 “Cellular Internet of Things (CIoT)”.

Other examples of major efforts to extend and enhance the current 4G EPC core architecture to address new business opportunities include the 3GPP Release 14 control and user plane separation (CUPS) feature to help address user plane scaling challenges as mobile network data growth has risen rapidly in recent years.

Another major challenge the operators face is that the current mobile cores designed to date are neither cloud nor network-slicing native. Operators globally have invested substantial amounts of time and effort in virtualization, management, orchestration, and network slicing of the mobile core. The Return on Investment (RoI) has been minimal at best as the current generation of 3GPP mobile cores are not designed to be cloud native.

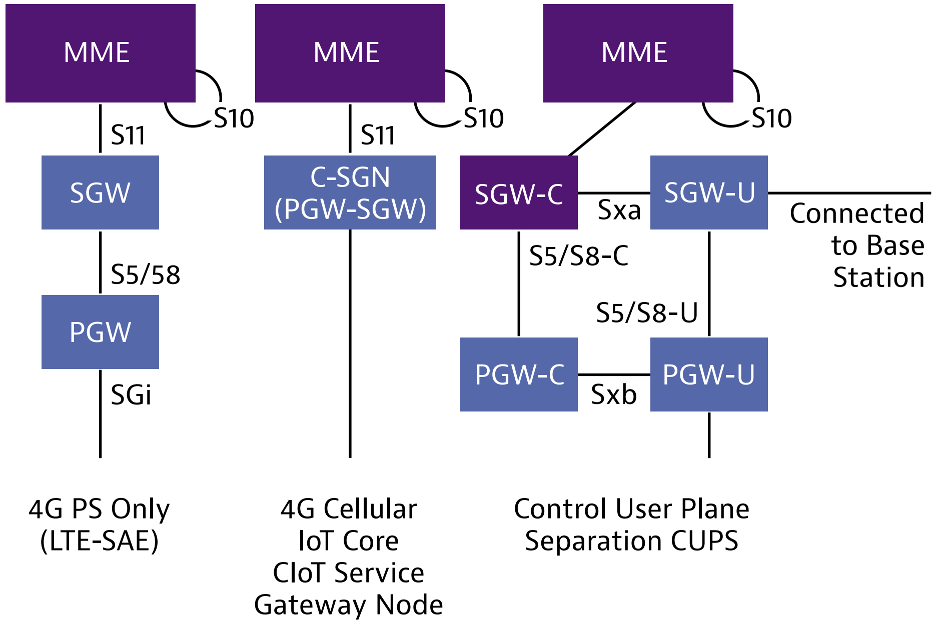


Figure 4.2 4G EPC Support for CUPS and CIoT

Figure 4.2 shows a simplified view of the major 4G core enhancements to the baseline EPC in 3GPP Release 13 and 14 to support CIoT and CUPS.

Through these efforts, the industry soon realized that they are facing diminishing return with further work put into enhancing and extending a mobile core network that has not been designed for the required modularity, flexibility, and programmability to support the widely diverse connectivity requirements of the vertical use cases from the beginning.

Industry was left with no choice but to start designing a clean slate 5G core.

4.3 4G Evolved Packet Core (EPC) versus 5G Next Generation Core (NGC)

The 4G EPC has been designed with distinct Network Elements and expanded in R13, R14, and R15 to support:

- Control and User Plane Separation (CUPS) in R14
- Dedicated Core Network (R13 DECOR, R14 eDECOR)
- Upgraded to support R15 Non-Stand-Alone (NSA) 5G EN-DC Dual Connectivity (EN-DC)
- Massive IoT (smart meters, etc.)

The next generation 5G core in comparison will be:

- Service-Based Architecture (SBA) and Interfaces (SBI)
- Disaggregated Core with Network Functions (NF)
- Native CUPS
- Native Network Slicing
- Supports both Stand-Alone (SA) and NSA deployment architectures
- Verticals Friendly

4.4 5G Service-Based Architecture (SBA)

The 5G core is designed from the ground up to be services friendly: utilizing state-of-the-art web and cloud technologies, which are extremely efficient in terms of extensibility and scalability, therefore making it fundamentally different from any of the previous generations' 3GPP mobile cores.

Traditionally 3GPP network elements communicate with each other via standards-defined interfaces and procedures, which is a static design and has proven to be challenging to extend to support new functionalities, network elements, and interfaces.

The 5G core, as shown in Figure 4.3, is designed to be cloud-native and is composed of network functions (NF). Each of the NFs has a set of mono (single) function NF services, which is therefore capable of offering a much higher granularity of programmability and configurability compared to the network element-based architecture.

5G Core NFs support native and true network slicing built with dynamic and precise control and automation of the network functions and services to match service requirements in real time.

The NF services-based design approach for the 5G core enables operators to take advantage of major advances in network function virtualization (NFV), management, and orchestration (MANO) techniques to assist with ease and speed of 5G core deployments and making progress with DevOps automation.

The 5G core consists of control and user plane network functions and services. The control plane network functions and services operate on the Service-Based Architecture (SBA) principle, while the user plane network functions and services within the 5G Core continue to use 3GPP classic interfaces and protocols. 3GPP has defined two representations for 5G core network functions interactions, namely: (a) reference point representation and (b) service-based representation. The service-based representation is applicable only for interactions between the control plane network functions in the 5G Core while the reference point representation is applicable for all the control and user plane network functions in the 5G System (5GS).

The reference point representation follows the traditional 3GPP representation in which reference points are defined for communications between two peer network functions. e.g. between the AMF and SMF—the N11 reference point is defined.

Figure 4.3 shows a simplified view of the 5G Core (5GC) with reference point representation.

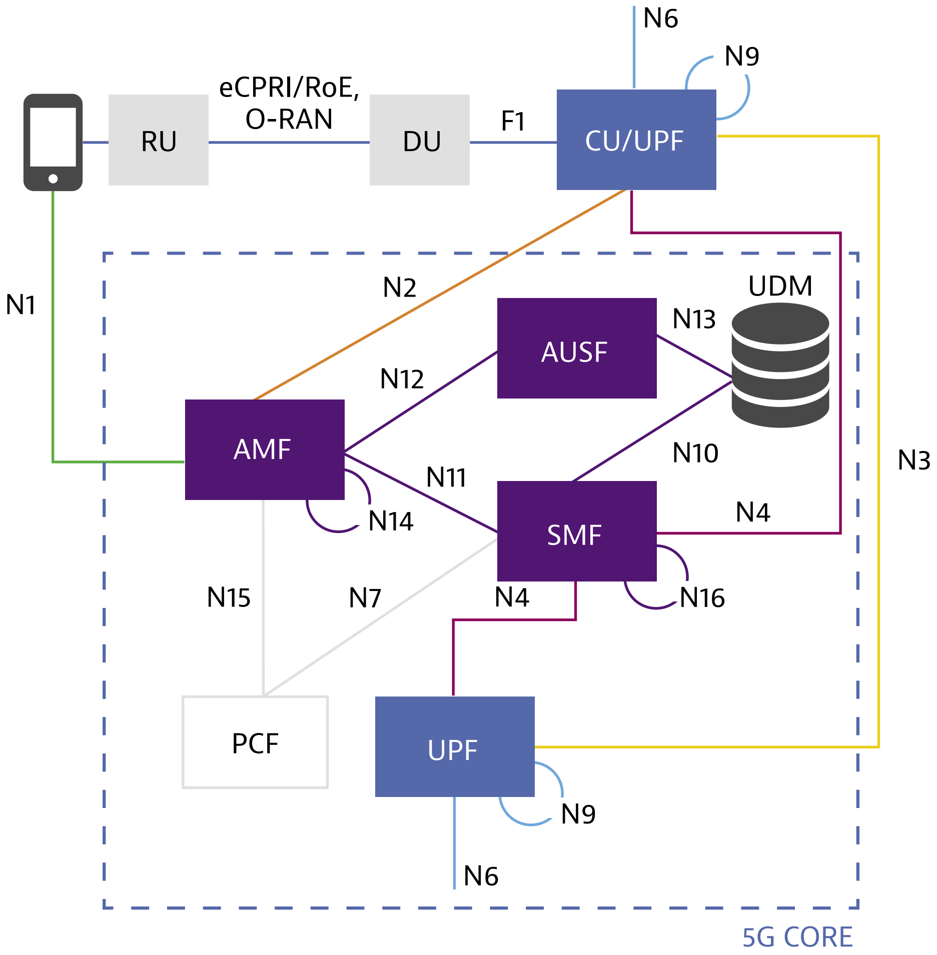


Figure 4.3 Simplified view of 5G Core (5GC)

For cross-referencing purposes, Figure 4.4 below maps out a few of the most common reference points (aka interfaces) in the 4G EPC and compares these against the somewhat equivalent reference points in the 5G core architecture.

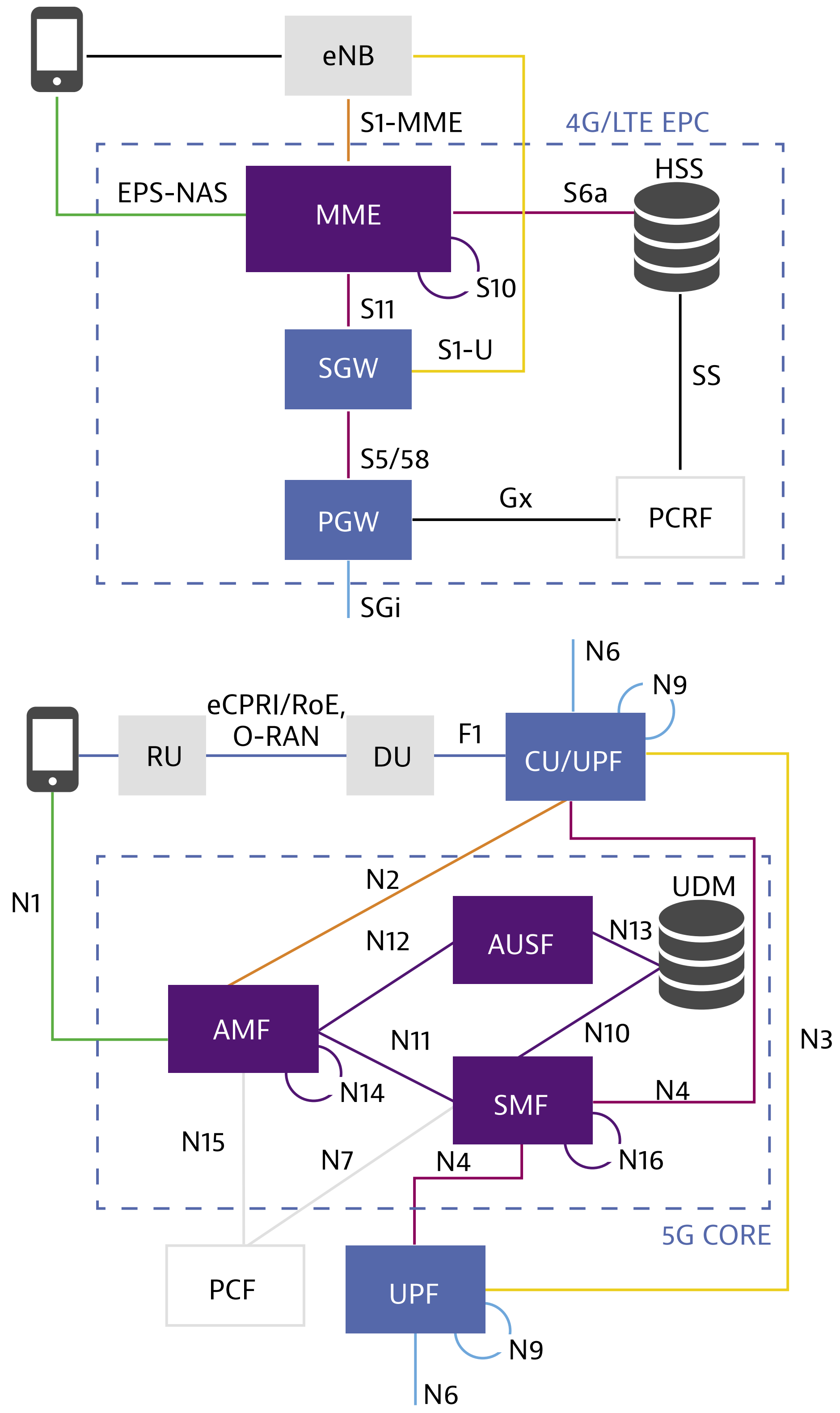


Figure 4.4 Reference point representation 4G versus 5G Core

What used to be the MME in 4G is now decomposed primarily into service-based functions, namely;

- Access and Mobility Management Function (AMF): this is where the NAS signaling terminates and where NAS ciphering, registration management, connection management, mobility management, authentication, and authorization are performed.
- Session Management Function (SMF): performs session management support, DHCP and IP address functions, data and traffic management.

What used to be HSS in LTE EPC is replaced by Unified Data Management (UDM) and Authentication Service Function (AUSF), for all encryption, authentication, policies, and credential management.

The EPC PCRF is replaced with the Policy Control Function (PCF).

The SGW-U and PGW-U are replaced by a User Plane Function (UPF) for packet handling and QoS management, which acts as the gateway for external Data Network interconnectivity. It is also involved in intra- and inter-RAT mobility.

The UPF instances can now be flexibly configured and deployed in multiple possible locations within the network slices to help best serve the requirements of the corresponding services utilizing the respective network slices.

For example, certain UPF instances can be deployed very close to the end subscribers for Ultra-Reliable Low Latency services (e.g., when dealing with Vehicle-to-Network (V2N) related applications) versus massive IoT services that can be better served with a centralized UPF, which does not require low-latency communications.

5G core network functions communicate using classic 3GPP protocols with the UE over the N1 interface, radio access network (RAN) over the N2 (AMF)/ N3 (UPF) interfaces, between user plane network functions over the N6 (DN)/ N9 (UPF) interfaces, and with the SMF over the N4 interface for control and user plane functions separation coordination.

The 5G Core SBA is designed to be stateless, open, and flexible. In the SBA architecture framework, each of these NFs will be offering their sets of NF services to other network functions over a common service-oriented framework, as shown in Figure 4.5, driven primarily by the needs of the services.

The Network Repository Functions (NRF) is designed to enable any network functions to register the services that they would like to offer to other network functions, and therefore these services can be discovered by other network functions seeking services offered by others.

The NRF maintains network function profiles of the registered network functions and the services that these network functions offer. The Network Function Profile (NFProfile) document includes information such as the type of the network function (e.g. SMF), network identification, and network slice identification (e.g. Single-Network Slice Selection Assistance Information (S-NSSAI), Network Slice Instance (NSI) ID, among others).

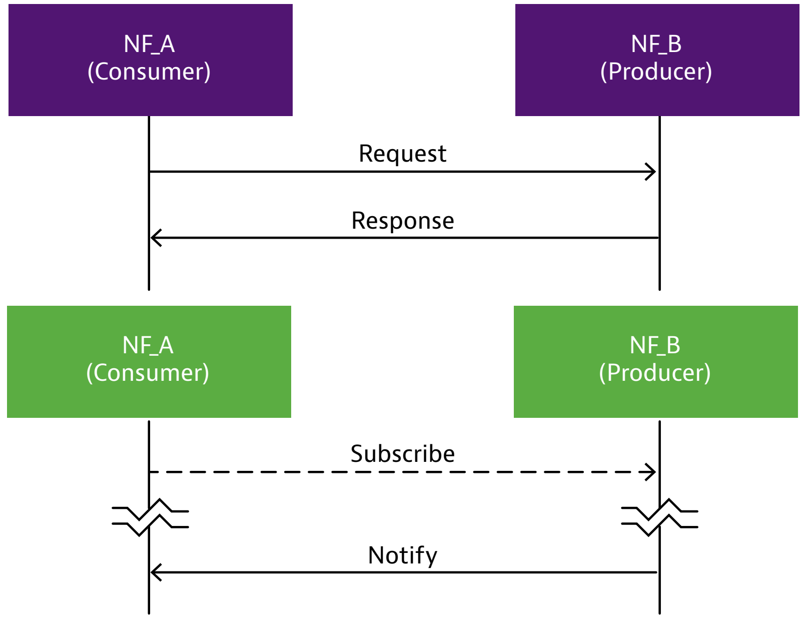
The same network function can be consuming network function services from other network functions while in other scenarios providing services that are consumed by other network functions. Network functions offering services to be consumed by other network functions are acting in a "producer" role with the other ends taking the "consumer" role.

Figure 4.5 shows the illustrations of network functions NF_A and NF_B assuming the Consumer and Producer roles performing the Request-Response and Subscribe-Notify interactions respectively.

5G core defines Service-Based Interfaces (SBI) adopting web protocols that are widely used web-scale network communications, replacing traditional mobile core networking protocols such as 3GPP GPRS Tunneling Protocol (GTP) and Diameter.

3GPP Release 15 adopts HTTP/2 protocol taking a RESTful approach in general and using JavaScript Object Notation (JSON) as a data format. Open API 3.0.0 (format for REST APIs) has been chosen as the interface definition language and Release 15 has as well created YAML (configuration files) as part of the 3GPP specifications.

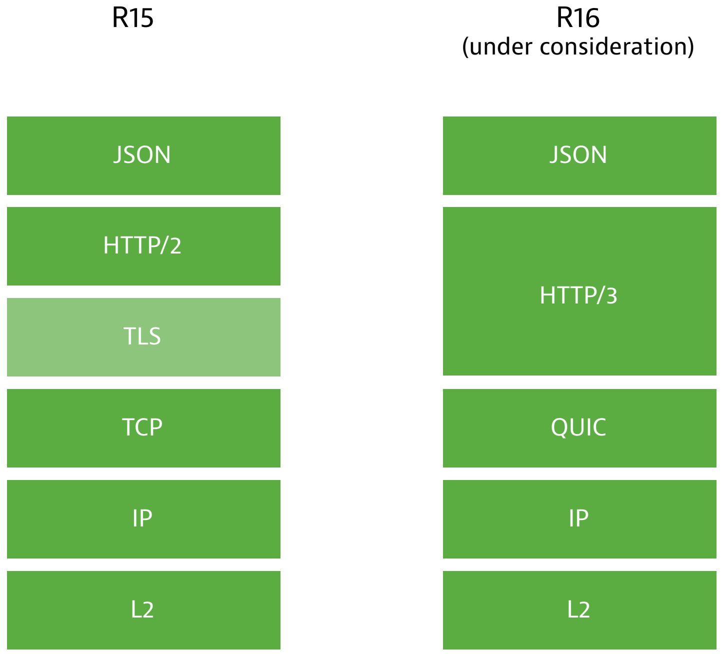
3GPP Release 16 considers Quick UDP Internet Connection (QUIC, initially developed by Google) as an alternative transport protocol. One of its key advantages is reduced time required to establish/re-establish connections compared to HTTP/2 over TCP, which is a crucial enabler for Ultra-Reliable Low Latency Communications.



3GPP TS 23.501

Figure 4.5 "Request-response" NF service, "Subscribe-Notify" NF service (Ref: 3GPP TS 23.501)

Figure 4.6 shows the Service based Interfaces control plane protocol stacks based on HTTP/2 and HTTP/3.



3GPP TR 29.893

Figure 4.6 Service-based Interfaces control plane protocol stacks - HTTP/2 and HTTP/3 (Ref: 3GPP TR 29.893)

Table of Contents

Figure 4.7 shows a simplified view of the 5G Core (5GC) with reference point representation of the classic and service-based interfaces marked out with Green and Blue colored lines respectively.

The service-based representation (Nx) is designed to be highly extensible as each of the network functions has its own service-based representation enabling other authorized network functions to access the services it offers. "x" is the network function name in lower case. For example, AMF uses Namf, SMF uses Nsmf to communicate to each other and other network functions via the common SBA framework.

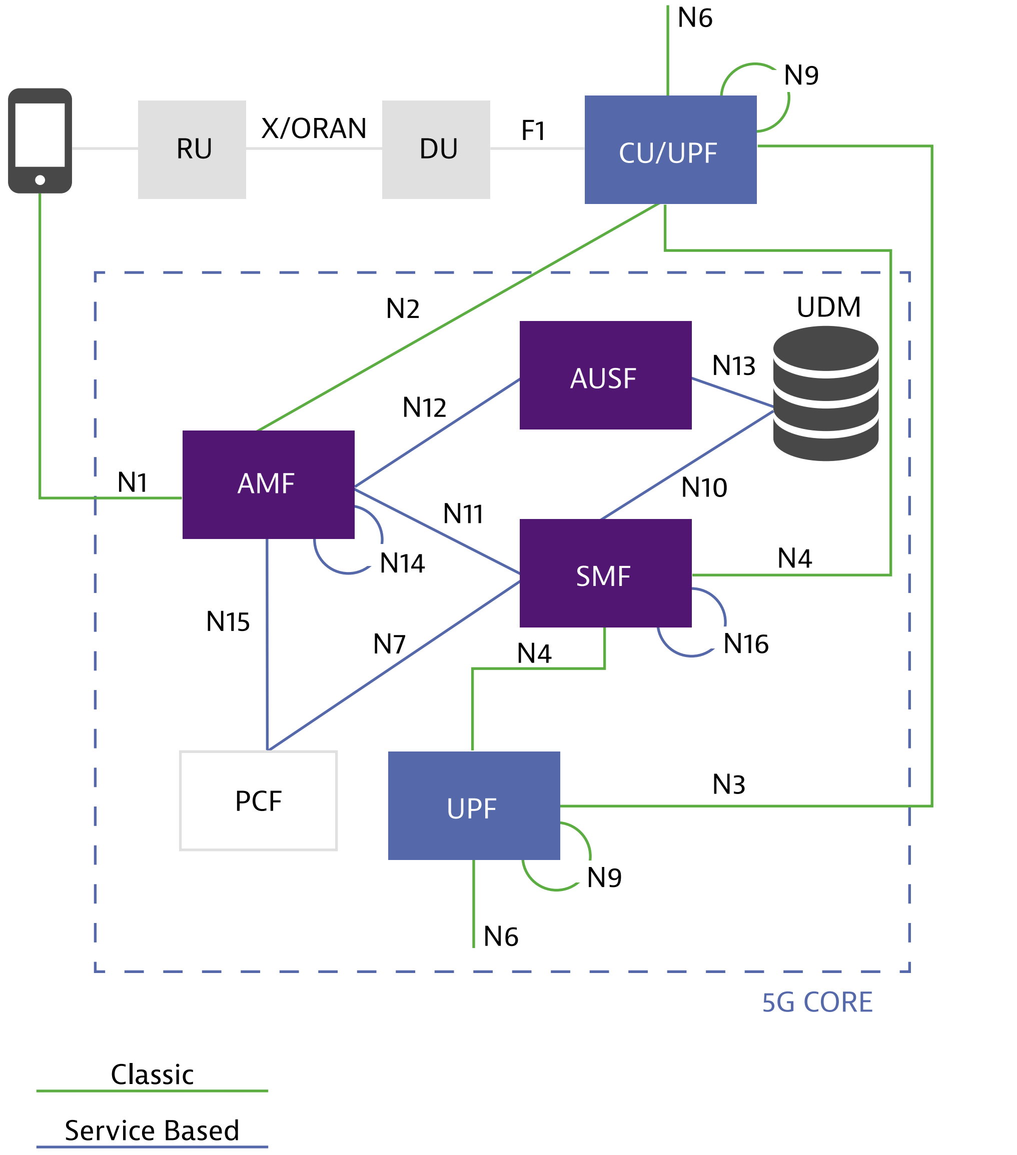


Figure 4.7 5G Core (5GC) with reference point representation of the classic and service-based interfaces

Figure 4.8 shows a 5G system architecture in the service-based representation with the classic and service-based interfaces marked out with Green and Blue colored lines respectively. Note that not all the 5G core network functions are shown in this Figure.

Figure 4.9 shows how the service-based representation is used. In procedure (3), the initial AMF request for UE's Slice Selection Subscription data from UDM and in procedure (4), the initial AMF requested for assistance on network slices selection from the Network Slice Selection Function (NSSF).

Table 4.1 lists the key 5G core 3GPP classic interfaces and service-based interfaces with their corresponding NEs/NFs, reference point representations, protocols and 3GPP specifications and compared against the somewhat equivalent of the 4G core network 3GPP classic interfaces.

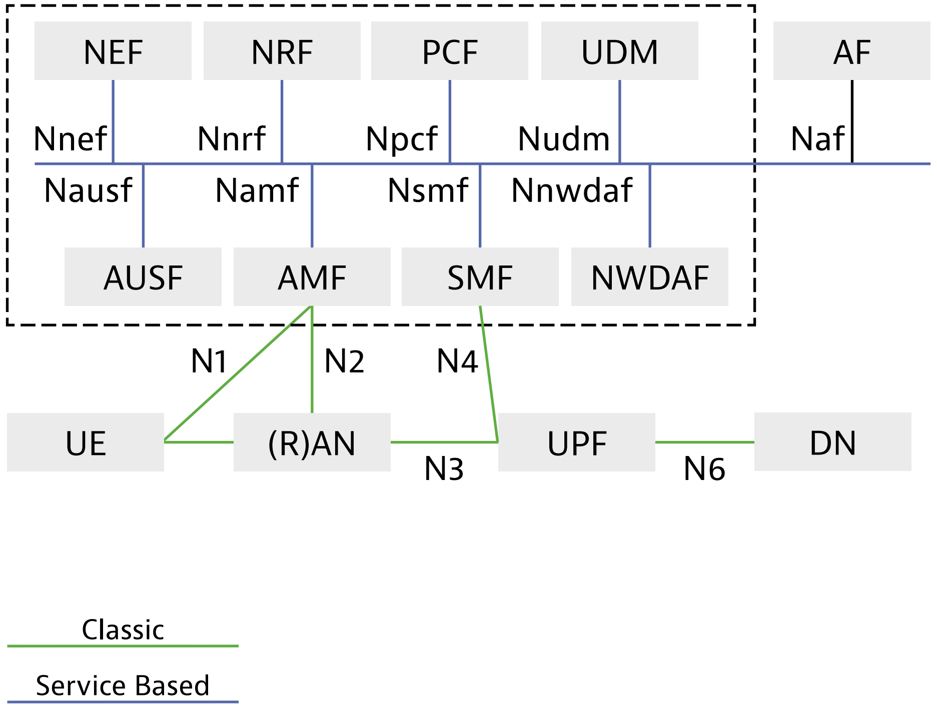
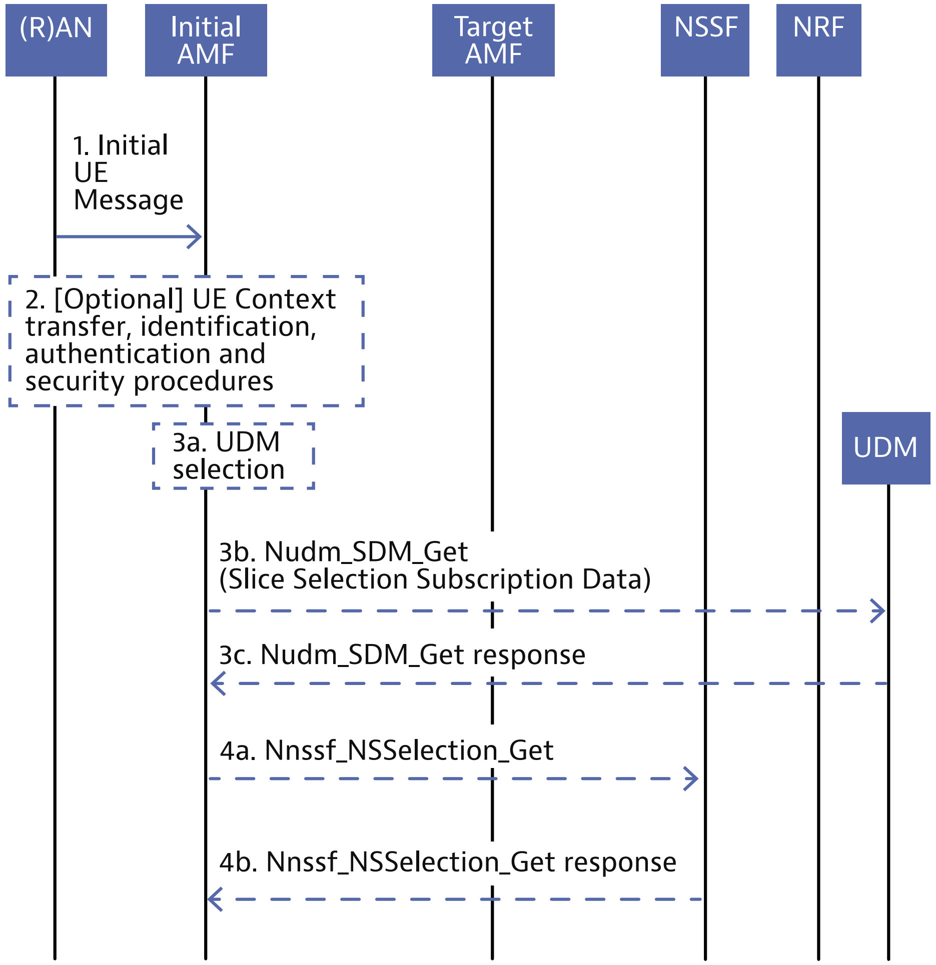


Figure 4.8 5G system architecture in the service-based representation with the classic and service-based interfaces


(R)AN
Initial AMF
Target AMF
NSSF
NRF
UDM
1. Initial UE Message
2. [Optional] UE Context transfer, identification, authentication and security procedures
3a. UDM selection
3b. Nudm_SDM_Get (Slice Selection Subscription Data)
3c. Nudm_SDM_Get response
4a. Nnssf_NSSelection_Get
4b. Nnssf_NSSelection_Get response

3GPP TS 23.502 Figure 4.2.2.2.3-1

Figure 4.9 Registration with AMF re-allocation procedure (Ref: 3GPP TS 23.502 Figure 4.2.2.2.3-1)

5GC Interface	5GC Protocol	3GPP TS	NE/ NF1	NE/NF2	4G Core Interface	4G Core Protocol	3GPP TS	NE1	NE2
N1	5GS NAS	24.501	UE	AMF	S1-MME	EPS NAS	24.301	UE	MME
N2	NG-AP	38.413	RAN	AMF	S1-MME	S1-AP	36.413	RAN	MME
N3	PDU UP GTP-U	38.415 29.281	RAN	UPF	S1-U	GTP-U	29.281	RAN	SGW
N4	PFCP, GTP-U	29.244	UPF	SMF	Sxa, Sxb	PFCP, GTP-U	29.244	SGW-C PGW-C	SGW-U PGW-U
N6	-	29.561	UPF	DN	SGi	-	29.061	PGW	PDN, other networks
N9	PDU UP GTP-U	38.415 29.281	UPF	UPF	Does not exist				
N26	GTPv2-C	29.274	AMF	MME	S3	GTPv2-C	29.274	MME	S4-SGSN
					Gn/Gp	GTPv1-C	29.060	MME	Gn/ Gp-SGSN

5GC Interface	5GC SBI NF Interface	3GPP TS	NE/ NF1	NE/NF2	4G Core Interface	4G Core Protocol	3GPP TS	NE1	NE2
N7	Nsmf NPCF	29.502 29.512	SMF	PCF	Gx	Diameter	29.212	PGW	PCRF
N8	Namf Nudm	29.518 29.503	AMF	UDM	S6a	Diameter	29.272	MME	HSS
N11	Namf Nsmf	29.518 29.502	AMF	SMF	S11	GTPv2-C	29.274	MME	SGW-C
N12	Namf Nausf	29.518 29.509	AMF	AUSF	S6a	Diameter	29.272	MME	HSS
N14	Namf	29.518	AMF	AMF	S10	GTPv2-C	29.274	MME	MME

Table 4.1 5G Core mapping to 4G core reference points and specifications

Figure 4.10 shows the protocol stacks for the 5G core reference points which use 3GPP classic protocols.

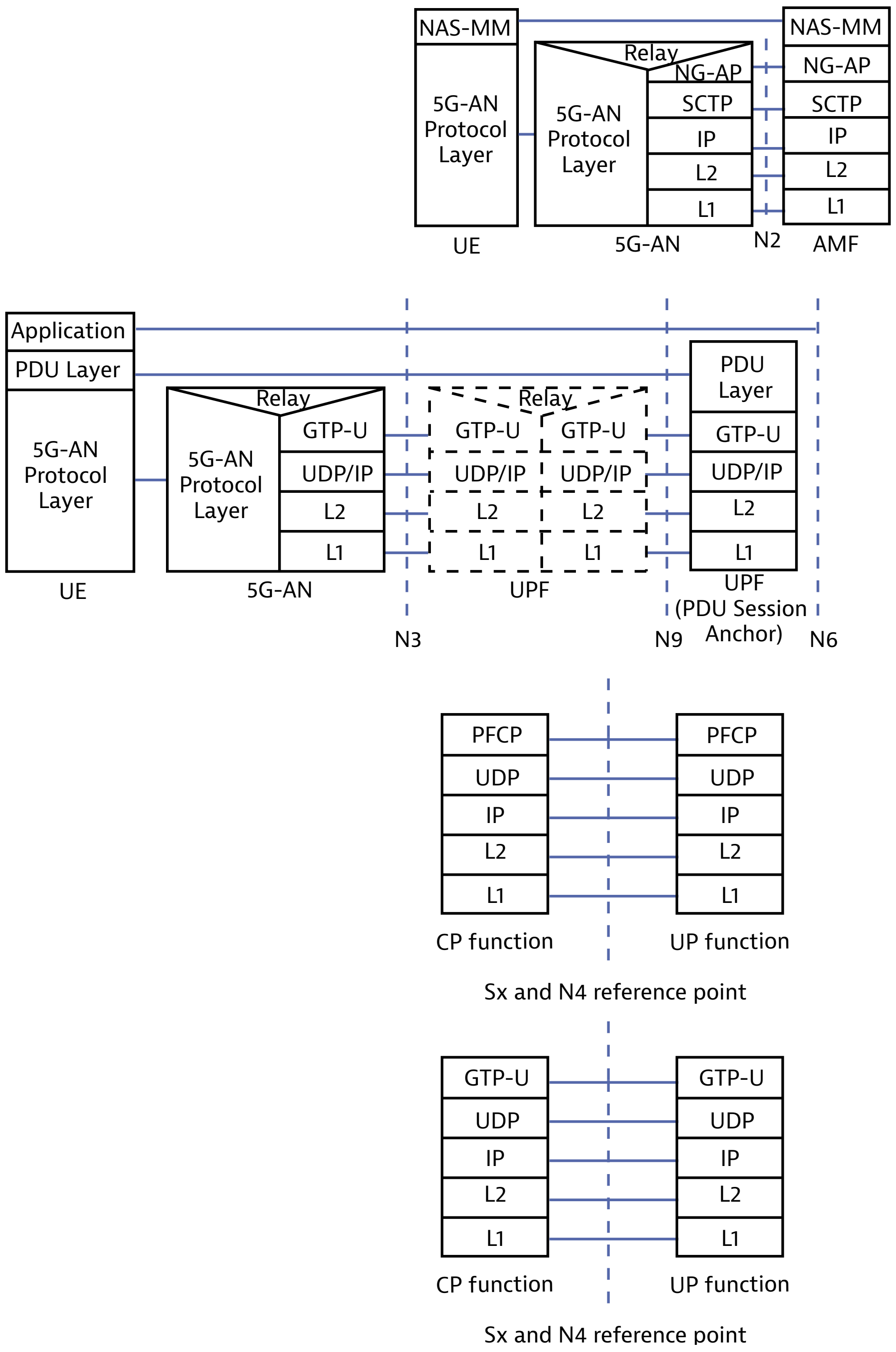


Figure 4.10 Control plane protocol stack between the UE and the 5G Core AMF, User Plane protocol stack between the UE and the 5G Core UPF, Control and User Plane protocol stacks for Sx and N4 reference points (Ref: 3GPP TS 23.501)

4.5 5GS QoS Architecture and Models

4.5.1 PDU Connectivity Service and PDU Session

The 5G system (5GS) is designed to provide Protocol Data Unit (PDU) connectivity service for exchange of PDUs between a UE and the Data Network (DN).

A PDU session is an association between the UE and DN providing the required PDU connectivity service.

A PDU session provides transfers of layer 3 IP packets, layer 2 Ethernet, or unstructured types. The Ethernet and unstructured types are newly introduced in the 5GS to help provide native support for industry IoT applications using the 5GS as the underlying transport.

A PDU session of Ethernet type enables flexible setup of industrial Ethernet using 5G core and 5G NR to cater for verticals needs. In this case the UE is part of the 5G "LAN" with the 5G UPF operating as a L2 switch. Packet filters can be created based on MAC addresses and QoS can be differentiated based on these parameters, similar to how it is done for Ethernet frames over wireline interfaces.

Figure 4.11 shows the ethernet packet filter sets matching and enforcement points in the UE and UPF.

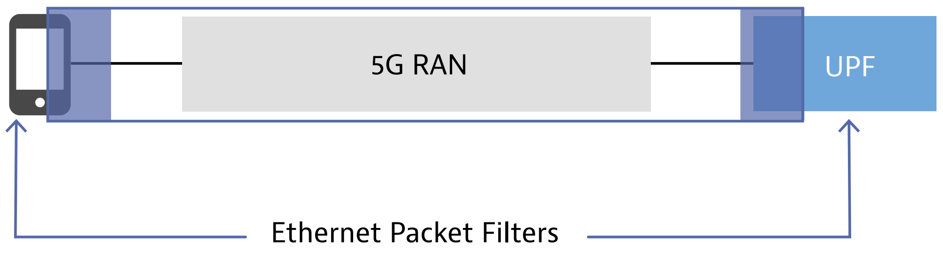


Figure 4.11 Ethernet packet filter set matching and enforcement points

3GPP Release 16 adds support of Ethernet Time Sensitive Network (TSN) over 5GS, as described in chapter 3, and this will enable Ethernet quality transmission in areas such as packet distribution, automatic address discovery, and QoS (Quality of Service). Combining TSN and 5GS in the industrial network architecture extends the potential application spaces for 5G into industrial Ethernet wireline applications.

A PDU session of unstructured type can be used to transfer any format/data structure unknown to the 5G system, which is a very common scenario as most of the IoT communications are currently performed using unstructured packet types. As these communications will continue to evolve over time, it is important to design the 5GS to be future proof by supporting unstructured data transfer.

4.5.2 PDU Session is Network-Slice Native

A PDU session is associated with one S-NSSAI and one DNN (Data Network Name). The DNN is equivalent to Access Point Name (APN) as shown in figure 4.14.

4.5.3 5G QoS Model

The 5G QoS model is based on QoS flows which is the finest granularity of QoS differentiation in a PDU session. As well as the guaranteed bit rate (GBR) and non-GBR QoS flows inherited from earlier technologies, 5GS introduces delay critical GBR QoS flows and reflective QoS.

Figure 4.12 shows the 5GS QoS architecture in the NG-RAN connected to the 5G Core.

QoS flow is identified within a PDU session by a QoS Flow ID (QFI) carried in the frame header over the NG-U, Xn-U, and N9 interfaces.

Figure 4.13 referenced from 3GPP TS 38.415 shows the QFI field in the frame header of the PDU Session user plane protocol.

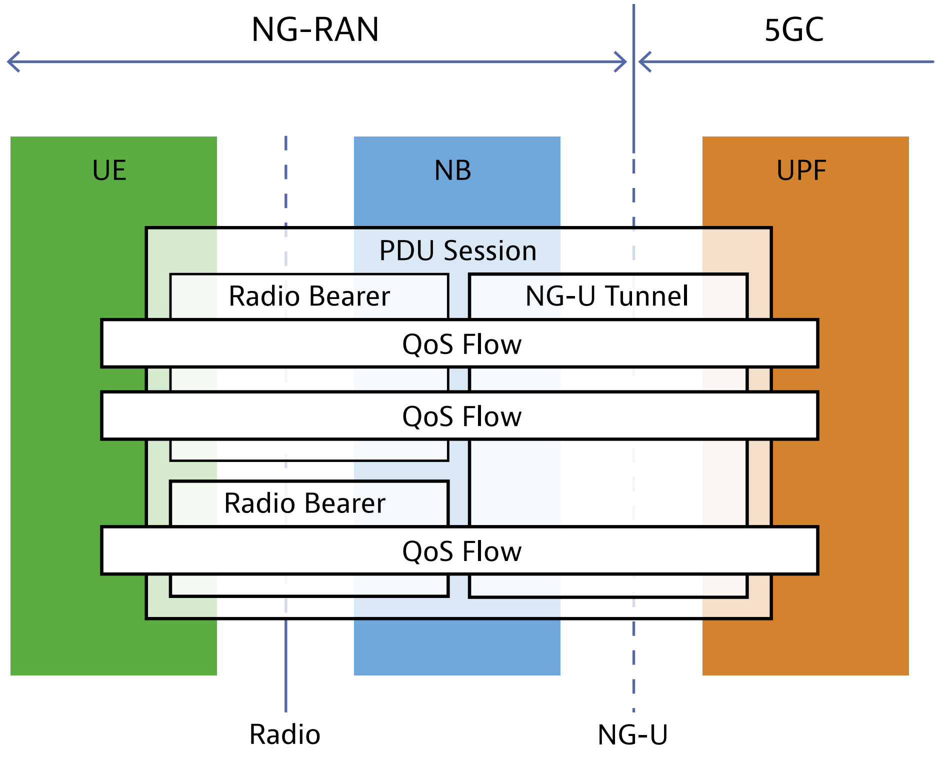


Figure 4.12 5G System QoS architecture (Ref: 3GPP TS 38.300 Figure 12-1)

Bits								Number of Octets
7	6	5	4	3	2	1	0	
PDU Type (=0)				Spare				1
PPP	RQI	QoS Flow Identifier						1
PPI			Spare					0 or 1
Padding								0 - 3

Bits								Number of Octets
7	6	5	4	3	2	1	0	
PDU Type (=1)				Spare				1
Spare		QoS Flow Identifier						1
Padding								0 - 3

Figure 4.13 Downlink PDU session information (PDU Type 0) format, uplink PDU session Information (PDU Type 1) format (Ref: 3GPP TS 38.415 Figure 5.5.2.1-1 and Figure 5.5.2.2-1)

4.5.4 5G QoS Profile

Each QoS flow has its defined QoS profile that corresponds to one QoS Flow identifier (QFI).

The QoS profile includes the following parameters:

- 5G QoS Identifier, 5QI, represents the QoS parameters of the specified QoS flow, which can be one of the standardized values listed in 3GPP TS 23.501 specification or a non-standardized value. The 5QI value may be used as the QFI of the QoS Flow.

- Allocation and Retention Priority (ARP) contains information about the priority level, pre-emption capability, and pre-emption vulnerability of the specified QoS flow.

For each GBR QoS Flow only, the following parameters shall be included:

- Guaranteed Flow Bit Rate (GFBR) specifies the bit rate that is guaranteed to be provided by the network to the QoS flow over the Averaging Time Window.

- Maximum Flow Bit Rate (MFBR) specifies the highest bit rate limit that is expected from the specified QoS flow.

For each GBR QoS Flow only, the following parameters may be included:

- Notification control can be requested from the NG-RAN when the GFBR cannot be fulfilled for the QoS flow during the lifetime of this QoS flow.
- Maximum Packet Loss Rate indicates the maximum rate of lost packets that can be tolerated for the specified QoS flow.

For each Non-GBR QoS Flow only, the following parameter may be included:

- Reflective QoS Attribute (RQA) indicates certain traffic for the specified QoS flow that can be subjected to reflective QoS treatment.

4.5.5 5G QoS Identifier and Characteristics

5QI is a scalar index used to reference a set of 5G QoS characteristics, which in turn describes the edge-to-edge (UE to UPF) packet forwarding treatment for the specified QoS flow. The 5G QoS characteristics are specified with the following parameters:

- Resource Type (GBR, delay critical GBR, or Non-GBR)
- Priority Level
- Packet Delay Budget (PDB)
- Packet Error Rate (PER)
- Averaging Window (for GBR and delay critical GBR resource type only)
- Maximum Data Burst Volume (MDBV) (for delay critical GBR resource type only)

The delay critical GBR resource type and MDBV parameters are newly introduced in the 5G QoS characteristics framework while these are not available in the 4G EPS QoS framework.

The main differences between delay critical GBR and GBR resource types are in the definition of PDB and PER for these resource types, and that the MDBV parameter applies only to the delay critical GBR resource type.

The MDBV parameter indicates the largest amount of data that the 5G RAN is required to serve within a period of 5G RAN PDB.

Table 4.2 is referenced from 3GPP TS 23.501 section 5.7.4, which shows selected Standardized 5QI to QoS characteristics mapping.

5QI Value	Resource Type	Default Priority Level	Packet Delay Budget	Packet Error Rate	Default Maximum Data Burst Volume	Default Averaging Window	Examples Services
82	Delay Critical GBR	19	10 ms	10^{-4}	255 bytes	2000 ms	Discrete Automation (see TS 22.261)
83		22	10 ms	10^{-4}	1354 bytes	2000 ms	Discrete Automation (see TS 22.261)
84		24	30 ms	10^{-5}	1354 bytes	2000 ms	Intelligent transport system (see TS 22.261)
85		21	5 ms	10^{-5}	255 bytes	2000 ms	Electricity Distribution- high voltage (see TS 22.261)

Table 4.2 selected Standardized 5QI to QoS characteristics mapping (Ref: 3GPP TS 23.501 Table 5.7.4-1)

4.5.6 5G Reflective QoS

The reflective QoS feature introduced in 5GS enables QoS differentiation over the downlink with symmetric differentiation in the uplink being achieved with minimal control plane signaling to the UE.

The UE derives the QoS rules based on the received downlink user plane traffic. It then maps the uplink user plane traffic to QoS flows, therefore not needing the SMF to provide QoS rules for these uplink QoS flows.

Reflective QoS and non-reflective QoS can be applied concurrently for the same PDU Session.

4.6 5G NGC and Network Slicing

4.6.1 Introduction to Network Slicing

Network slicing is a concept that enables operators to scale their network capabilities on an on-demand basis to help deliver the QoS service level agreements

(SLAs), meeting the needs of the industry verticals without needing to custom design and deploy dedicated networks for each of the vertical use cases. As discussed earlier, this approach has been proven to be inefficient.

Network slicing enables flexible on-demand creation of extremely efficient logical network partitions operating off a common physical infrastructure, with these network slices required to be dynamically managed and adapted to the diverse vertical services requirements, changing in real-time.

Network slices are required to be managed to meet the services QoS requirements across all network functions and inter-connecting segments from the end-to-end (E2E) perspective for both intra-slice and inter-slice scenarios.

E2E network slices in 3GPP standardization scope include the Radio, RAN, and Core aspects; however, in practical deployment environments, E2E network slices shall include the underlying transport infrastructure as it is providing the inter-connectivity between network functions and physical data centers.

Network slices are required to be completely isolated from each other, including protection from potential overload and security breaches from other network slices.

4.6.2 5G Core Network Slices Topology

5G core supports native network slicing, as all device connections and interactions through the 5G core are served and connected with network slices. The moment the 5GS-capable device is powered on, it is attempting to connect and register to a network slice.

5G network slices consist of 5G network functions that are deployed as common and dedicated network functions and service-chained dynamically.

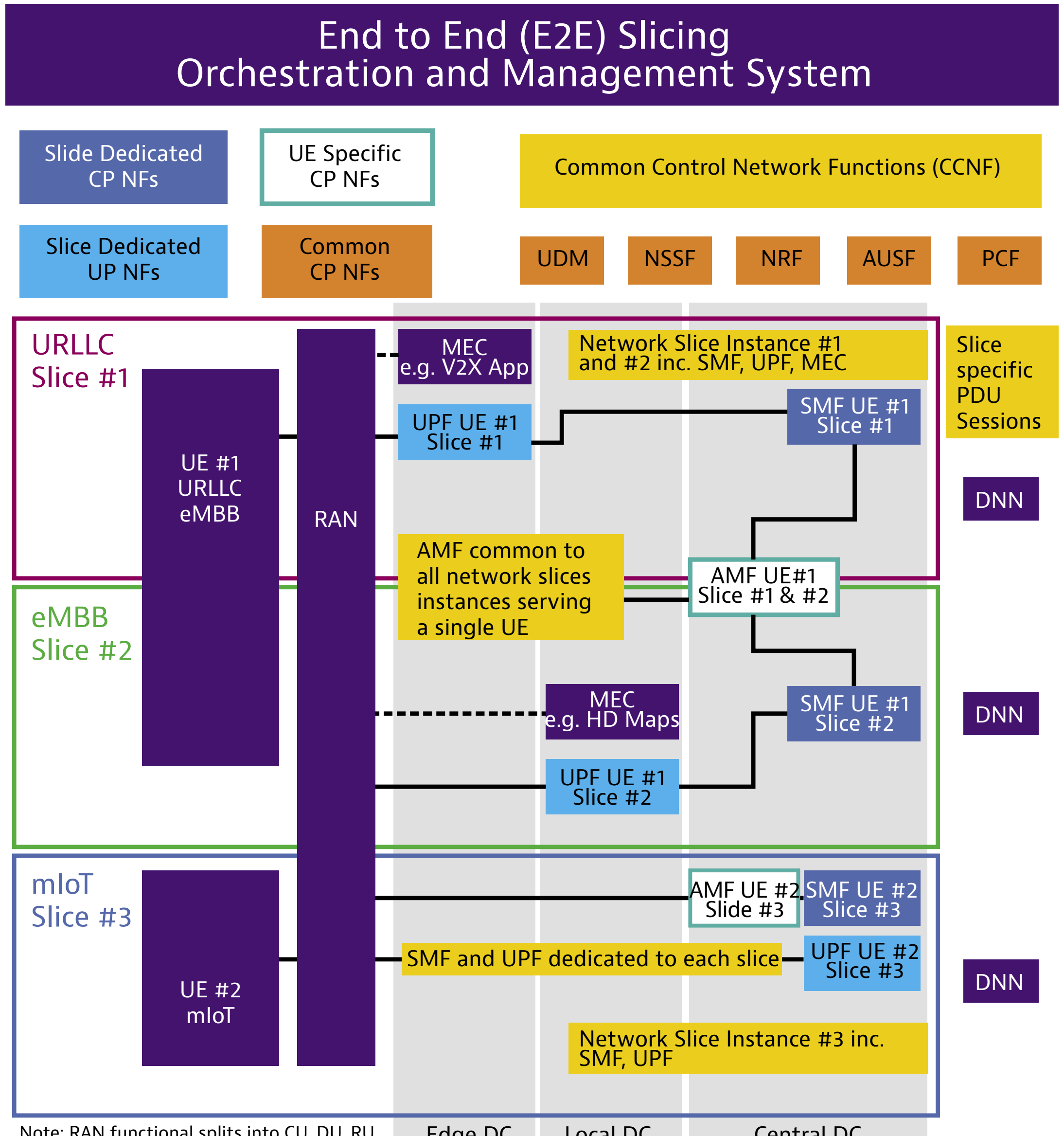
Common network functions such as a single common AMF are allocated to terminate a UE's NAS connection for all slices. This single AMF proxies session management messages to and from SMFs in the different network slices.

For dedicated network functions—such as in the case of user plane—each data connection of the UE is served by an SMF+UPF belonging to the same assigned slice.

The UE can have multiple PDU sessions in a slice to different data networks or multiple PDU sessions to the same data network via different slices, via the combination of slice identifier and APN.

A UE can establish and maintain connections to a maximum of eight slices in parallel.

Figure 4.14 shows a service deployment example for 5G System network slicing focusing on the internals of 5G core network slicing i.e. the UE, radio, RAN and transport slicing aspects are not shown in this Figure.



Note: RAN functional splits into CU, DU, RU components are not shown in this diagram

Figure 4.14 5GS Network slicing: service deployment example

4.6.3 5G Network Slice Identifiers

An S-NSSAI is used to identify a network slice.

An S-NSSAI is made up of:

- A slice/service type (SST) refers to the expected network slice behavior in terms of features and services. Examples of SST include eMBB, URLLC, MIoT.
- A slice differentiator (SD), which is optional information that complements the slice/service type(s) to differentiate among multiple network slices of the same slice/service type.

An S-NSSAI can have standard values or non-standard values. An S-NSSAI with a non-standard value identifies a single network slice within the Public Land Mobile Network (PLMN) with which it is associated. Standardized SST values enables global interoperability for slicing scenarios spanning multiple PLMNs, as roaming use cases can be supported for the most commonly used slice/service types.

Table 4.3 (referenced from 3GPP TS 23.501) shows the list of standardized SST values as defined in 3GPP Release 15 and their respective characteristics.

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

Table 4.3 Standardized SST values (Ref: 3GPP TS 23.501 Table 5.15.2.2-1)

The S-NSSAI can be associated with one or more network slice instances.

The NSI ID serves as an identifier for a network slice instance.

The NSSF may return NSI ID(s) to be associated with the network slice instance(s) corresponding to certain S-NSSAIs.

A PDU session, which is the 5G system (5GS) association between the UE and a Data Network that provides a PDU connectivity service, is associated to one S-NSSAI and one DNN (Data Network Name).

4.6.4 5G Network Slice Selection

Network slice instance selection for a UE is normally triggered as part of the registration procedure by the first AMF that receives the registration request from the UE.

The UE provides the requested Network Slice Selection Assistance Information (NSSAI) for network slice selection in 5G NR-RRC message, if it has been provided by NAS. The NSSAI is a collection of S-NSSAIs.

The AMF retrieves the slices that are allowed by the user subscription and interacts with the NSSF to select the appropriate network slice instance (e.g., based on Allowed S-NSSAIs, PLMN ID, etc.).

Figure 4.15 referenced from [1] shows an example of 5G network slice selection.

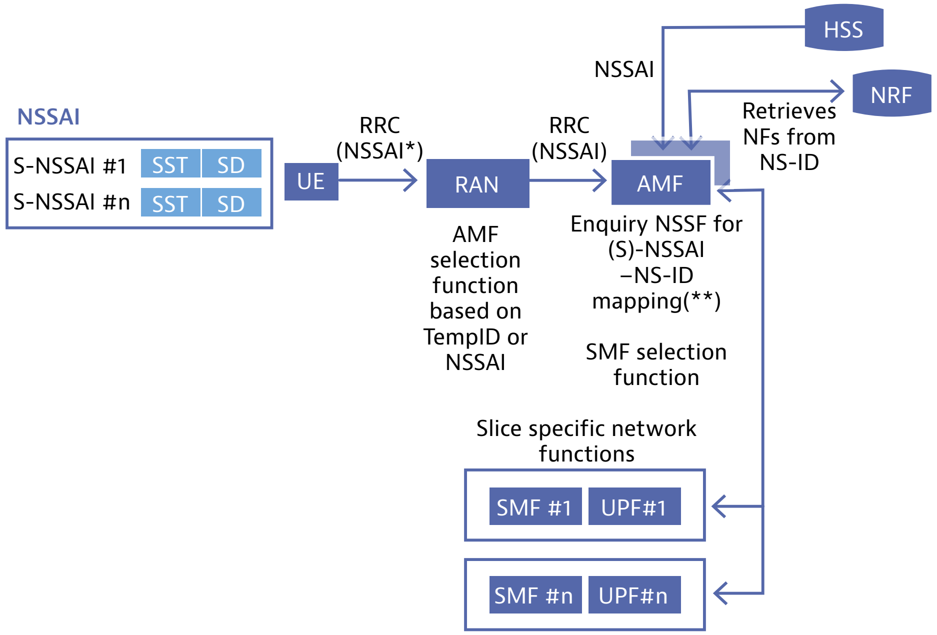


Figure 4.15 5G Network slice selection example

Notes

1. GTI 5G Network Slicing White Paper 1st version 2018
2. 3GPP TS 23.501 System architecture for the 5G System (5GS)
3. 3GPP TS 23.502 Procedures for the 5G System (5GS)
4. 3GPP TS 23.503 Policy and charging control framework for the 5G System (5GS); Stage 2
5. 3GPP TS 38.415 NG-RAN; PDU Session User Plane protocol

CHAPTER FIVE

Edge Computing

5G is an essential element enabling the "sixth technological revolution," referred to in the introductory chapter as the "AI era." This revolution is about connected things enabling new verticals and full automation using machine learning and artificial intelligence. 5G is designed and architected to enable this revolution and in particular the uses cases that will require Ultra-Reliable Low Latency Communications. Following the 5G architecture design as explained in Chapter 1, it can be concluded that many of the intelligent computations needed from the network to react and enable these new real-time services will move to the edge of the network, hence the term "edge computing."

More specifically, edge computing refers to the practice of offering cloud computing capabilities at the edge of a communications network, in close proximity to end-users and coupled with the service provider's network infrastructure. These computing resources can then be used to process user data close to where it is being generated, rather than relying on the centralized data-processing warehouses that form the traditional cloud environment. Since data is processed near its source, the transmission latency can be significantly reduced, while also minimizing the volume of data that must be hauled across the transport network to the cloud. Local processing also offers an additional level of security, since sensitive data need never enter the public cloud. This reduces the likelihood of it being compromised, or even corrupted, during transport.

To understand the significant advances enabled by edge computing, let us take a quick walk down memory lane of computing history and discover how it all began.

In the late '60s and early '70s, "mainframe" computers were invented. You needed a physical data center filled with these computers to be able to run your application. "Dumb" terminals were used to enter application data and retrieve results.

Then in the '80s and '90s, the invention and adoption of microprocessors led to the rise of local, powerful servers and desktop computers. Those computers were used by many companies to run their business and IT applications. We also saw the emergence of the personal computer (PC) and subsequently portable laptops.

With the turn of the century, the term "cloud computing" was being used to refer to applications that ran remotely in a data center, accessible through the Internet. This setup offered small- and medium-sized companies the ability to scale without having to spend much CapEx on office IT computers and applications. Then, with the introduction of smartphones and the associated app frenzy, cloud computing became the norm to manage access to a large amount of computational and storage resources in the cloud.

And now, with the need for real-time applications requiring Ultra-Reliable Low Latency Communications for decision-making (think connected cars), these cloud-based computational and storage resources need to move closer the edge of the network; we are witnessing the start of the multi-access edge computing (MEC) era. Applicable use cases for this era [1] include pervasive video; media sharing; tactile Internet; automated traffic control and driving; collaborative robots; and remote object manipulation. Subsequently 3GPP, the standards development organization (SDO) partnership project, developing the 5G System (5GS) specifications, has identified edge computing as a key technology enabler for achieving the required low latency.

This chapter focuses on edge computing standardization activities, including those within the European Telecommunications Standards Institute (ETSI) and the linkage to 3GPP. It then takes a more specific look at testing in this developing ecosystem, considering the transition from lab to field for edge-related products and services.

5.1 Standardization Overview

Established in late 2014, ETSI Industry Specification Group (ISG) MEC is the only SDO group solely focused on developing technical standards for edge computing [2]. The goal of the ISG is to specify a set of standardized APIs that enable application and content providers to utilize computing capabilities located at the edge of the network (mobile or otherwise). A high-level view of the architecture is provided in the framework depicted in Figure 5.1.

The overall MEC system provides the capability to run MEC applications within a service provider's network and is made up of the following components:

- MEC platform (MEP): A collection of essential functionalities required to run (locally hosted or remote) MEC applications on virtualization infrastructure, enabling them to provide and consume MEC services. It is deployed as a virtual network function (VNF) in a network functions virtualization (NFV) environment.

- MEC applications: Instantiated on the virtualization infrastructure based on configuration or requests validated by the MEC management. As with the MEP, these are deployed as a VNF in an NFV environment. The application packaging is inextricably linked to the run-time environment of the MEC system. It is run and packaged accordingly as a virtualized application, such as a virtual machine (VM), or as a containerized application, such as a Docker container or Kubernetes template. MEC applications form part of a new development model, which has three distinct locations: client, near server, and far server (Figure 5.2). The client location could be a traditional smartphone or other wireless-connected computer elements, for instance within a car, smart home, or industrial location. The near server refers to the edge component with a set of operations that the application performs at the edge cloud; for instance, for terminal device computational offload while still leveraging very low latency and predictable performance, or offloading bandwidth demands from the network backbone, or making use of local contextual information from the location or network information APIs. The remote components implement operations that are carried out in the remote data center/traditional cloud, where increased storage and database access is most practically supported.

- MEC service: Services offered by MEC applications, or the MEP directly and exposed via RESTful-based APIs.

- MEC platform manager (MEPM): Manages the MEP and MEC application rules, requirements, and lifecycle. Lifecycle management responsibilities are delegated to a VNFM for the deployment of MEC in an NFV environment.

- MEC system-level management: Includes the service provider's operations support system (OSS) and the MEC orchestrator (MEO). All requests for application instantiation or termination are made via the OSS, which makes the decision on whether to grant those requests. If successful, those requests are forwarded to the MEO for further processing. The MEO maintains an up-to-date view of the overall MEC system. It is responsible for on-boarding application packages and making the decision on where they should be instantiated. Components of the MEO are delegated to an NFVO for the deployment of MEC in an NFV environment.

- Virtualization infrastructure: the totality of all hardware and software components that build up the environment in which VNFs are deployed, managed, and executed.

- Virtualization infrastructure manager (VIM): Manages the virtualized (compute, storage, and networking) resources of the virtualization infrastructure, including the necessary preparation to run a software image on the infrastructure. It also provides virtualized resource performance and fault information to the MEPM.

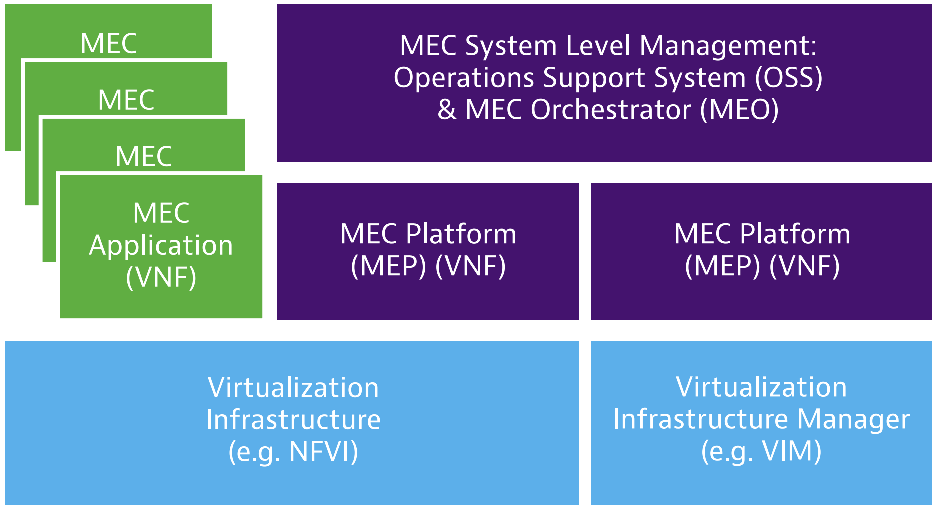

MEC
MEC
MEC
MEC Application (VNF)
MEC System Level Management: Operations Support System (OSS) & MEC Orchestrator (MEO)
MEC Platform (MEP) (VNF)
MEC Platform (MEP) (VNF)
Virtualization Infrastructure (e.g. NFVI)
Virtualization Infrastructure Manager (e.g. VIM)

Figure 5.1 Multi-access edge computing framework

As highlighted above, the MEC architectural framework has very much been developed with the NFV architecture in mind. Within MEC, there is a specific reference architecture variant for MEC in NFV, providing the function entity and reference point mapping between the two architectures [3].

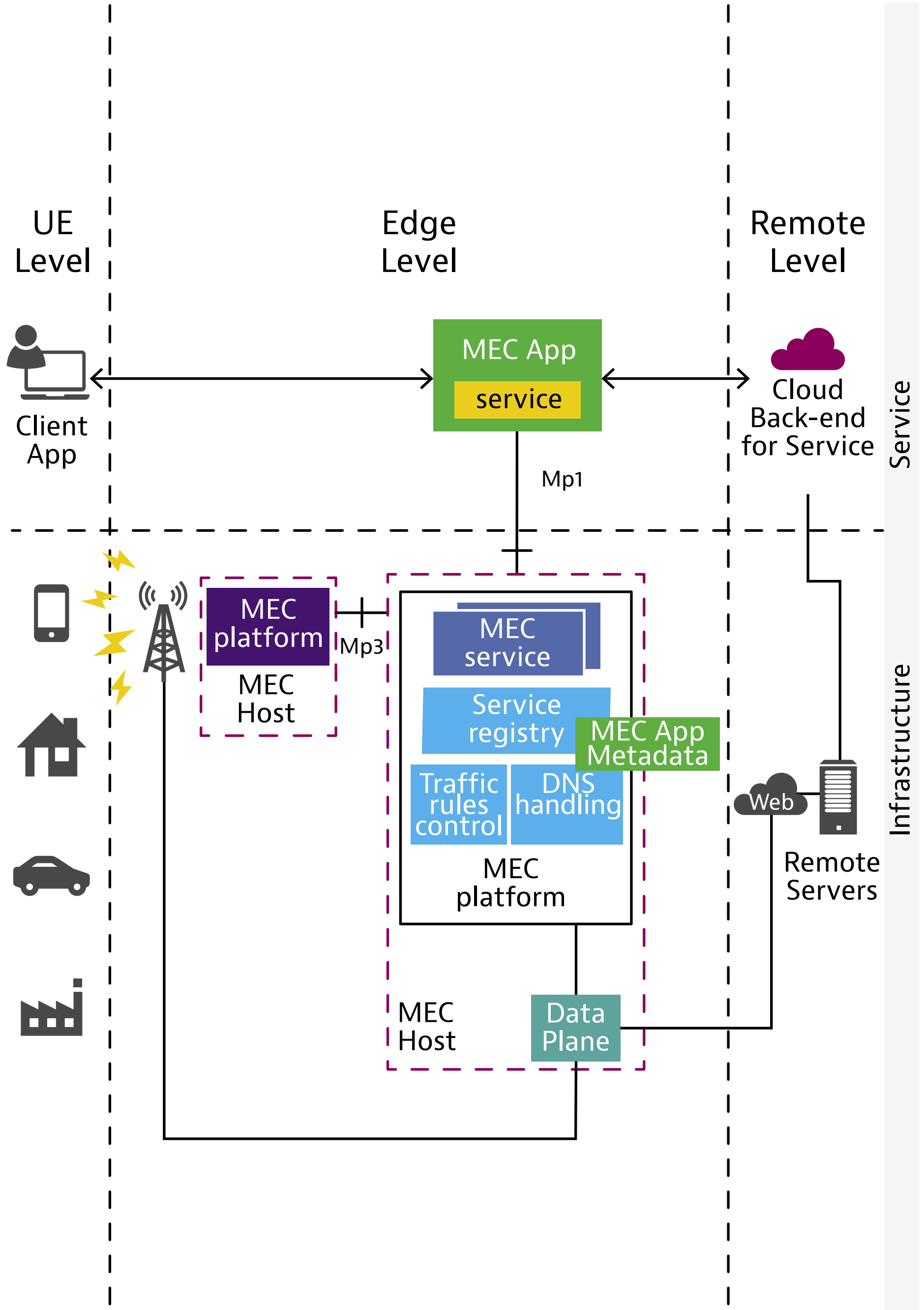


Figure 5.2 New application development paradigm introduced by MEC [4].

5.2 MEC Application Enablement Framework

The MEP is at the core of the standardized MEC application enablement framework [5], as depicted in Figure 5.3. This provides the Mp1 reference point, through which MEC applications can be authenticated and authorized. Such applications may consume MEC services or may even provide services themselves (more detail in the next section). The service-producing applications must be registered with the MEP, which maintains the service registry, via the Mp1 reference point. The registry facilitates service discovery, advertisement, and service-related notifications, e.g. state change.

Additional capabilities offered by the MEP are the ability to receive traffic rules and DNS records from the MEPM. The traffic rules may also be received from MEC applications or services and are applied to the underlying data plane. This allows IP traffic routing, or tapping, to the MEC applications, or to locally accessible networks (e.g. enterprise network, Internet access, etc.). Based on configuration or following an activation request from the MEC application, the DNS records are used to configure the DNS proxy/server, accordingly, providing the mapping between an IP address and its fully qualified domain name (FQDN). Overall, the MEC application enablement framework has wide-ranging applicability, since it is generic in nature and extendable. The result is that MEC can support any application and any application can run in MEC. However, MEC offers additional benefits to those applications that have been designed to be MEC-aware. Such benefits can be realized through MEC's ability to open up the service provider's network to authorized external applications and provide the means to expose pertinent information to them. This is considered a key value-add feature of the MEC specification, since it offers applications the ability to gain contextual information and real-time awareness of their local environment through standardized RESTful service APIs. The resulting service environment can be used to tremendously improve the user experience. For example, through the network information service APIs it is possible to precisely monitor events and performance of the network (e.g. radio) connection to the end-user device.

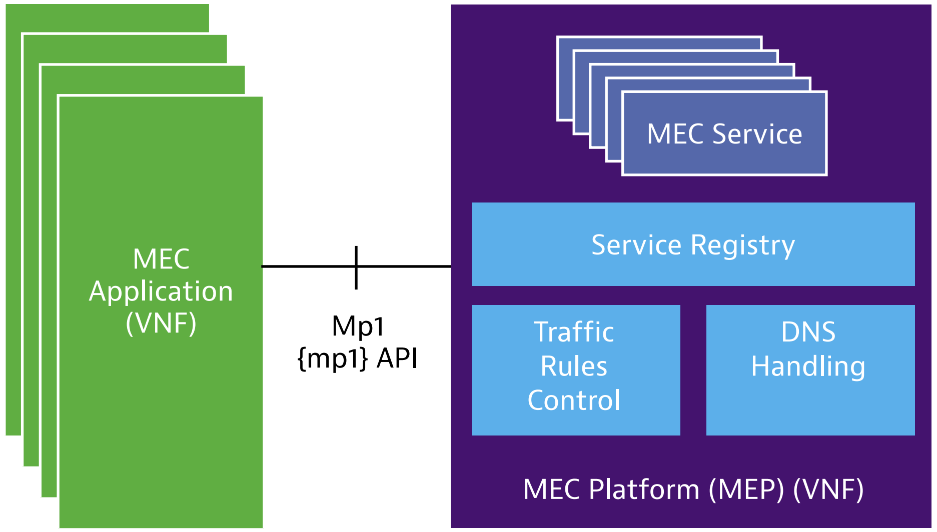


Figure 5.3 Multi-access edge computing service enablement

Such contextual information may be leveraged by a suitably designed MEC application to drive the behavior of the client application in the user device, as well as application components in a central cloud. The adjustments in behavior can be performed at runtime according to the prevailing conditions, where the proximity to the end users means the environment can be better predicted. In this manner, the network characteristics can be accounted for during the design of the end-to-end service. This permits edge applications to benefit from low latency and high throughput in a more predictable and controllable way. Knowing that this information is available means that it can be leveraged during service design to optimize the end-to-end service architecture during runtime. The overall network itself may also benefit from the MEC services provided by applications; for instance, a scheduler within the network could also make use of user behavior predictions based on edge analytics, to maximize the network efficiency.

The MEC-specified APIs follow a generic set of specified design principles and patterns [6]. Consistency has been achieved across the ISG MEC APIs through compliance to these principles, which helps facilitate APIs that are inherently application-developer friendly and straightforward to implement, particularly since OpenAPI specification [7] compliance descriptions have been made available [8]. Furthermore, the overall framework is extendable, since additional third-party service APIs may be offered as long as they adhere to these API guidelines.

5.3 MEC Services

It was highlighted in the previous section that MEC applications, supported by the standardized application enablement framework, may consume or produce services via RESTful-based APIs conforming to the MEC guidelines [6]. This is depicted in Figure 5.4, which highlights that such services may be offered by third parties:

- MEC Application A: consumes the third-party service offered by MEC Application C. It also produces an MEC-specified radio network information (RNI) service, which may offer supplementary information compared to the same service offered by the MEP (where not all MEP implementations will offer RNI). The MEC-specific services, including the RNI service, are detailed later in this section;

- MEC Application B: consumes the MEP provided Location Service (LS) and RNI service. It also consumes the third-party provided RNIs from Application A;

- MEC Application C: produces a third-party API, which adheres to the MEC API guidelines;

- MEC Application D: as with all MEC applications, makes use of the Mp1 application enablement API, but neither produces nor consumes MEC, or third-party services.

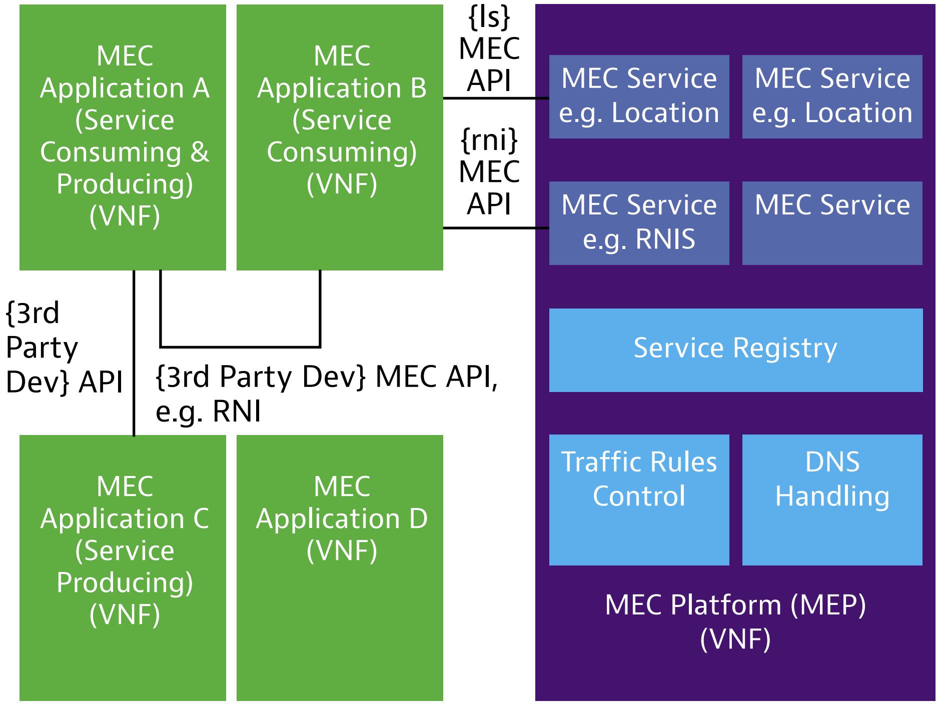

MEC Application A (Service Consuming & Producing) (VNF)
MEC Application B (Service Consuming) (VNF)
{ls} MEC API
{rni} MEC API
MEC Service e.g. Location
MEC Service e.g. Location
MEC Service e.g. RNIS
MEC Service
Service Registry
Traffic Rules Control
DNS Handling
MEC Platform (MEP) (VNF)
{3rd Party Dev} API
{3rd Party Dev} MEC API, e.g. RNI
MEC Application C (Service Producing) (VNF)
MEC Application D (VNF)

Figure 5.4 MEC service producing and consuming applications

5.3.1 Network Information Services

ETSI ISG MEC have specified a number of access network-focused information service APIs. The first among those was the RNI Service, specified in [9]. This provides near real-time 3GPP radio access network (RAN) information relating to the user equipment (UE) connected to the network. The RNI service, as with the majority of MEC service APIs, offers information via both approaches specified in the API guidelines: first, via direct query (request/response mechanism); second, via notification (subscribe/publish mechanism). The information available in each case is summarized in Figure 5.5, which collectively provides significant insight into the connections supported by the access network, including the prevailing channel conditions for UEs on the air interface and their mobility events. For maximum interoperability, the API may use JSON as the data format with the HTTP message exchanges. This approach is reasonable for simple queries. However, it is likely that scenarios will exist in which the amount of information, and update frequency, is so high that RESTful methods no longer scale. In this case, information may be shared using an "alternate transport," e.g. over-the-message broker of the MEC platform. This is also more efficient support for one-to-many communications. To discover and use such alternate transports, an MEC application queries the MEP for details on a message broker via the transport information query procedure as defined in [5]. Alternatively, such transport information may be pre-provisioned via configuration.

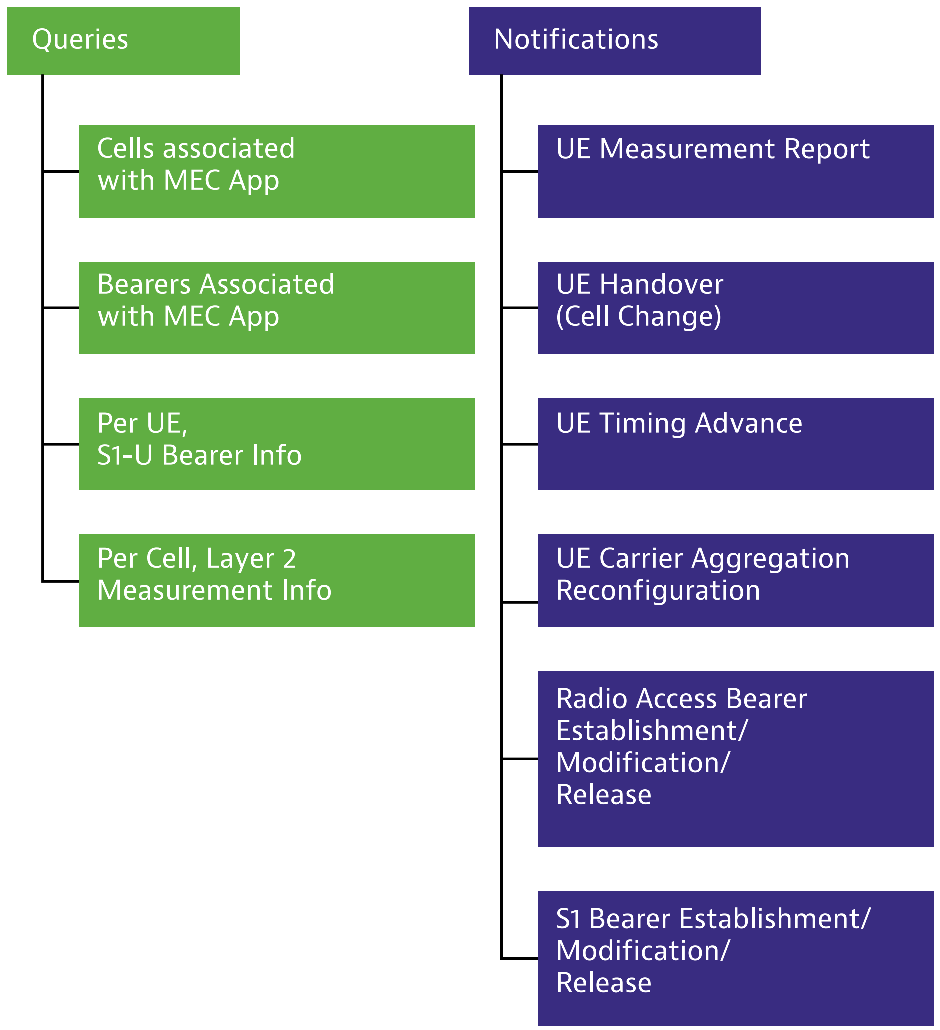

Queries
Cells associated with MEC App
Bearers Associated with MEC App
Per UE, S1-U Bearer Info
Per Cell, Layer 2 Measurement Info
Notifications
UE Measurement Report
UE Handover (Cell Change)
UE Timing Advance
UE Carrier Aggregation Reconfiguration
Radio Access Bearer Establishment/ Modification/ Release
S1 Bearer Establishment/ Modification/ Release

Figure 5.5 MEC radio network information service

A wireless local area network (WLAN) information service is also specified in GS MEC 028 [not yet published], which emulates the 3GPP mobile access focused RNI API, but for WLAN deployments. This provides a rich set of information from WLAN access points (AP) and stations (STA) based on that specified by IEEE and the Wi-Fi Alliance (WFA). The API uses filtering and attribute selection to allow selection of only the specific information of interest. For APs that information includes the AP ID, WLAN capabilities, Associated STAs, WAN metrics, BSS load, and neighbor information. For STAs that information includes STAS ID, associated apps, PHY rates, and fine time measurements (analogous to timing advance for 3GPP LTE radio access networks).

The last in the set is the fixed access information API GS MEC 029 [not yet published], which has a wider remit to cover fiber, cable, xDSL, and point-to-point fiber Ethernet access to MEC. The goal with this API has been to develop a generic API that provides access network-related information for the multitude of fixed-access technologies.

5.3.2 Location Service

The LS specified in [10] allows applications to exploit user proximity information (e.g., to retrieve and monitor the list of users connected to a specific cell or access point, or even provide their geographical coordinates).

5.3.3 Bandwidth Management Service

The bandwidth management service specified in [11] allows applications the ability to reserve networking resources via the MEP, thereby ensuring that quality of experience (QoE) requirements can be achieved.

5.3.4 UE Identity Service

The UE identity service specified in [12] allows applications to trigger user-specific traffic rules on the MEP, for example steering traffic to a local network.

5.3.5 V2X Service

There are also industry segment-focused APIs such as the V2X API being specified in GS MEC 030 [not yet published], which aims to facilitate interoperability in a multi-vendor, multi-network, and multi-access environment. The API builds are based on the recommendations resulting from the study on the ability of MEC to support V2X based use cases [13].

5.4 MEC Deployment Options

ETSI ISG MEC was established when the 3GPP LTE 4G architecture was already mature and already deployed in many networks. Therefore, the standardized reference architecture was designed to be agnostic to the evolution of mobile networks. This permits MEC deployments in 4G networks to be re-used in the support of 5G services, providing a smooth migration path for mobile network operators and a means to provide low-latency 5G-like services on their existing 4G networks. Even so, there are two main categories of deployment in a 4G network: S1 based and SGi based.

5.4.1 S1 Based

In this deployment scenario, the MEP is hosted either at the eNB (base station), on the S1 interface of the 4G architecture (Figure 5.6), or at the edge of the 4G core network. The specific deployment location can be selected based on the target latency, while offering the capability to steer traffic on a per session basis, or even down to the packet level.

If incorporated with the base station, plain IP packets can either be locally switched to/ from MEC applications or routed as GTP-encapsulated packets to/ from the serving gateway (S-GW). This deployment is well suited to CRAN-style deployments, where it is envisaged the MEC could share the same virtualization infrastructure used for the more centralized baseband unit (BBU), rather than the distributed remote radio heads (RRU) and be inserted prior to S1 encoding. A clear advantage with a CRAN-type approach is the centralization it affords. This allows traffic rules to be applied at the edge of the RAN across a wide area and for many UEs, rather than at an intermediate point along the S1 interface that would require S1 de-encapsulation and re-encapsulation.

If deployed on the S1 link itself, for instance at an existing aggregation point or at the edge of the core network (e.g. in a distributed data center, at a gateway), the MEC host's data-plane has to process user traffic encapsulated in GTP-U packets and therefore the encryption keys have to be made available. This specific option is known as "bump in the wire," with the wire referring to the S1 link.

For S1-based deployments there are further regulatory considerations, since operators are required to provide law enforcement agency (LEA) support. This includes lawful interception (LI) and retained data (RD) capabilities for traffic carried on their networks, which have typically been supported by core network elements, with all data passing through these elements. However, this approach is not appropriate for MEC where a portion of traffic will most likely not traverse the core due to local breakout of connections and local traffic generation. If Control User Plane Separation (CUPS) architecture is assumed for the underlying network, the LI support can be provided based on the available 3GPP standard. However, if the CUPS architecture is not available, a group specification [14] has been produced detailing the implementation of LI and RD collection functions at the edge of the network.

A significant factor with respect to an S1-based approach is that the mobile identity (IMSI) is not inherently available outside of the EPC. The IMSI is generally only exposed within the LTE EPC and communicated infrequently by the UE. The implication is that a secondary means is required to provide the information necessary to link the traffic flows supported within the RAN with specific UE Identities. The RAN is only aware of temporary identifiers such as the system architecture evolution serving temporary mobile subscriber identity (S-TMSI), associated Radio and S1-bearer IDs, and S1-application protocol (AP) UE IDs. Therefore, the MEC platform would need to be provided with the information to link a specific tag with the appropriate temporary identifiers. A solution would be to deploy probe-based agents in the EPC. Specifically, the role of the agent would be to extract IMSI/IMEI identifiers with their associated temporary identities for each connection session by probing key interfaces within the EPC. The agent would then provide this IMSI-based customer experience management (CEM) data feed with the necessary pairing information to the MEC platform. Using this information, the MEC platform could then fulfil the UE-specific traffic rules as required.

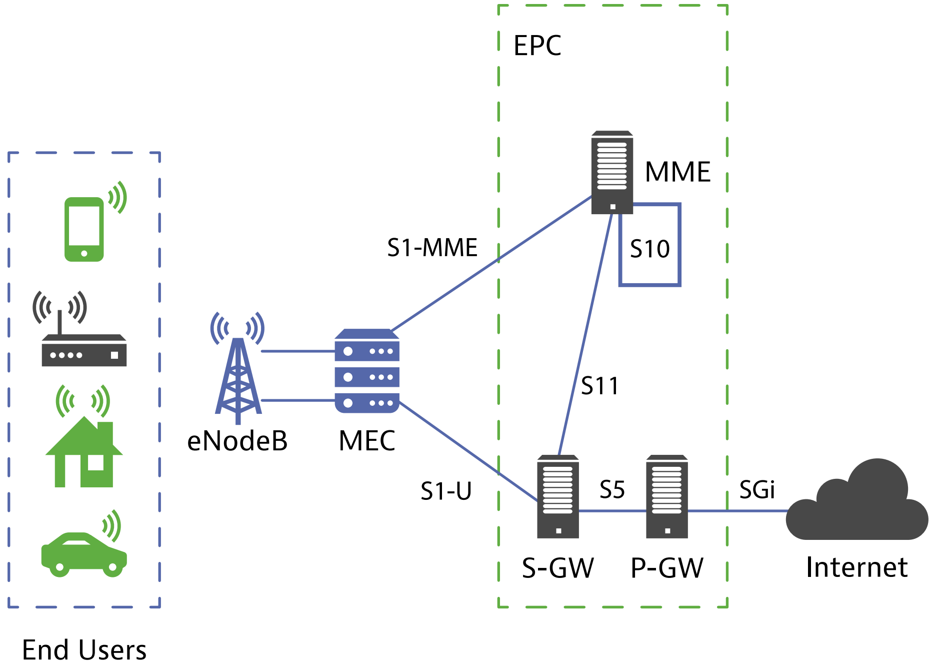


Figure 5.6 S1 MEC deployment

5.4.2 SGi Based

In order to achieve close proximity to the end user, this deployment is targeted to the situation in which the edge site logically includes all or part of the 3GPP evolved packet core (EPC) components, and the MEC data plane is placed on the SGi interface. This distributed EPC approach can be used to address enterprise-type use cases. For instance, for machine-to-machine-type networks, or 3GPP R13 mission-critical-push-to-talk (MCPTT), where the home subscriber server (HSS) could be collocated with the other network elements at the edge site. A working backhaul to a more centralized location is not required in order to maintain local connectivity.

With this scenario, once the UE subscribes to the EPC at the edge site, the P-GW assigns the IP address and local DNS information to resolve the MEC application's IP address, after having terminated the original PDN connection. In this manner, the appropriate UP traffic can be steered towards the MEC system. An advantage in this scenario is that fewer changes are required to the operator's network, since standard 3GPP entities and interfaces are leveraged for operations such as session management, charging, etc. However, it may not be suited to network wide deployment.

A variant of this deployment is to move only the S-GW and P-GW to the edge site (Figure 5.7), while maintaining the control plane functionality at the core. Local S-GW selection is then performed by the MME at the core. Traffic offloading is access point name (APN)-

based and therefore traffic for certain APNs may not be offloaded, e.g. roamers. Note that the S-GW is the mobility anchor, so this deployment may be problematic for moving users.

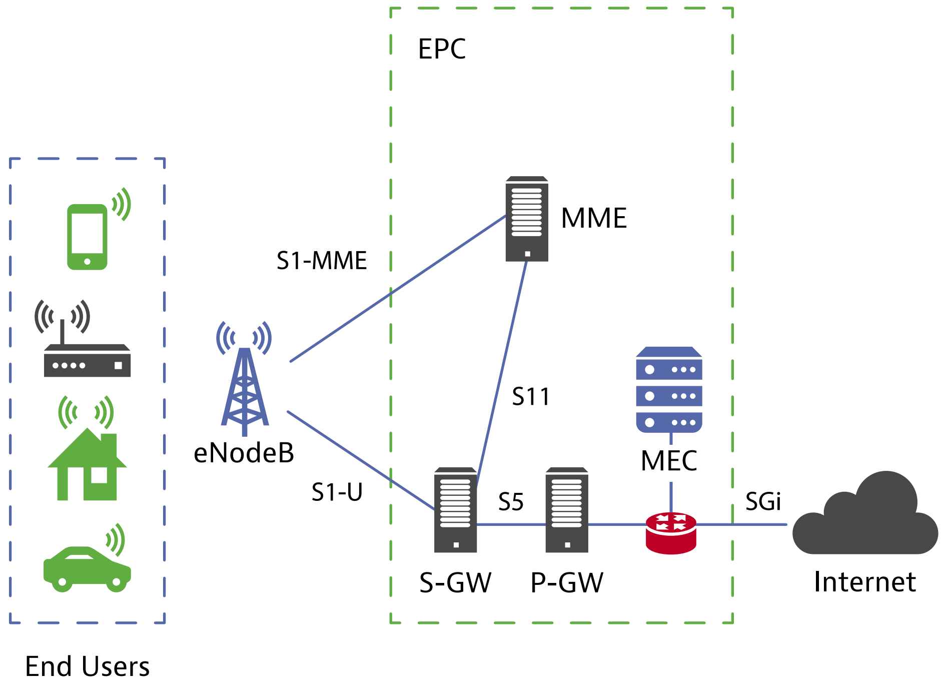


Figure 5.7 S1 MEC deployment

5.5 CUPS

Standardized by 3GPP in R14 and further developed in R15 [15], control and user plane separation (CUPS) specifies splitting the S-GW, P-GW, and TDF between their control and user plane components. This approach is considered a key enabler for edge computing by allowing a distributed user plane (UP) at the edge with a centralized control plane (CP). Here, a CP function can interface to multiple UP functions and a UP function can be shared by multiple CP functions. This allows for independent scaling of both the CP and UP functions, allowing edge-based UP deployments to be placed as required, and as demand dictates, with minimal impact to the core.

The CP functions continue to support mobility, charging, policy, and LI/ RD. Using rules, it also controls the local UP-based packet processing, namely: packet detection rules for packets inspection; forwarding action rules for packets handling (e.g. forward, duplicate, buffer, drop); QoS enforcement rules for QoS policing enforcement; and usage reporting rules for traffic usage measurements.

5.6 Edge Computing in 5GS

Next generation networks based on the 3GPP 5G System (5GS) specifications, detailed throughout this book, are considered a key target environment for MEC deployments. The service-based architecture (SBA) [16], introduced as part of the 5G specifications, leverages interactions between different network functions in order to align system operations with the network virtualization and Software Defined Networking (SDN) paradigms. This aligns with the approach taken in defining the ETSI MEC specifications.

While defining the 5G specifications, enablers for edge computing were considered from the outset. This has ensured mechanisms in place to allow MEC and 5G systems to collaboratively interact with one another on operations such as traffic routing and policy control.

The ETSI MEC-defined services introduced earlier in this chapter, together with the edge computing-focused technical enablers of the 5GS, facilitate system integration between the two systems and in doing so create a powerful environment for edge computing:

Routing and Traffic Steering: Mechanisms are provided to select PDU Session traffic to be routed to applications, e.g. MEC applications in a local area data network (LADN), via an N6 interface. Note that a PDU session may have multiple N6 interfaces, where a user plane function (UPF) that terminates such an N6 interface is said to support PDU session anchor functionality. Two UPF-based mechanisms are provided to steer traffic:

- The first is to use uplink classifier functionality, which relies on a set of traffic filters (e.g. based on IPv4 address) to identify the appropriate traffic to be steered to each PDU session anchor in the UL direction, and to be merged from multiple PDU session anchors in the DL direction.

- The alternative is to use branching point functionality, which routes sessions to a single DN, but via different PDU session anchors. This is applicable for PDU sessions associated with multiple IPv6 prefixes, which are referred to as multi-homed PDU sessions. A motivation for such an approach is to support make-before-break service continuity when switching between UPFs.

Application Function: External to the 3GPP network, an application function is offered the ability to influence UPF (re)selection and traffic routing directly via the policy control function (PCF) or indirectly via the network exposure function (NEF), depending on an operator's policies.

Session and Service Continuity (SSC): Three modes are provided for different UE and application mobility scenarios.

Local Area Data Network (LADN): A data network, in which applications (e.g. MEC applications) may be deployed, may be provided to serve a specific "local" service area.

QoS and Charging: Rules for QoS control and charging for traffic routed to the LADN are provided via the PCF.

5.7 MEC Deployment in 3GPP 5GS

As highlighted in [17], the 3GPP SBA, described in Chapter 4, and MEC API framework share a complementary approach and have adopted similar principles, with the SBA focusing on network functions and the services they offer and consume, while the MEC API framework focuses on MEC applications and services. Both frameworks include the functionality needed for efficient use of the services including registration, service discovery, availability notifications, de-registration and authentication, and authorization. Specifically, capabilities include the ability to authenticate the consumer, e.g. a network function, or external application function, and to authorize its service requests. Both frameworks also support flexible procedures to efficiently expose and consume services. A request-response model can be used for simple service or information requests, while a subscribe-notify model is provided for any long-lived processes. The remainder of this section now considers several of the key network functions offered by the SBA and MEC's relationship to them and potential interactions with them. This leads onto Figure 5.8, which shows how the MEC system can be deployed in an integrated manner in a 5G network.

First, there is the network resource function (NRF) that supports registration of network functions and the services they produce, while in MEC services are registered in the service registry of the MEC platform, a key component of the application enablement functionality. Network functions can directly interact with service-producing network functions if authorized, with the NRF allowing discovery of available services. Service accessibility may only be offered via the NEF, which is particularly relevant for untrusted entities that are external to the domain. In this manner the NEF acts as a centralized point for service exposure and plays a key role in authorizing all access requests originating from outside the system.

Next is the PCF, which handles the policies and rules in the 5GS. From an MEC perspective, acting as an application function (AF), it is the services of the PCF that are requested in order to influence the traffic steering rules. As with other network functions, depending on whether the AF is considered trusted or not, the PCF can be accessed either directly or via the NEF.

The role UPF plays with regard to routing and traffic steering has already been highlighted, where from the MEC system perspective it is considered as a distributed and configurable data plane. The difference is that the PCF is involved, rather than using the MEC Mp2 reference point to control that data plane for traffic rule configuration. It should be noted that in some specific deployments a local UPF may be part of the MEC system, as depicted in Figure 5.8.

Another key entity is the session management function (SMF), providing support for session management (including maintenance of the tunnel between the UPF and access network node), UE IP address allocation and management, dynamic host configuration protocol (DHCP) functions, charging (data collection and support for the charging interfaces), roaming, and lawful interception (event collection and interfacing to the LI system). It also performs UPF selection/ re-selection and configures the traffic rules for the UPF. This latter aspect is particularly important to MEC with regard to steering traffic to the local area data network (LADN). In addition, MEC acting as an AF can manage PDU sessions, control policy settings and traffic rules, as well as subscribe to notifications on session management events through the service operations exposed by the SMF.

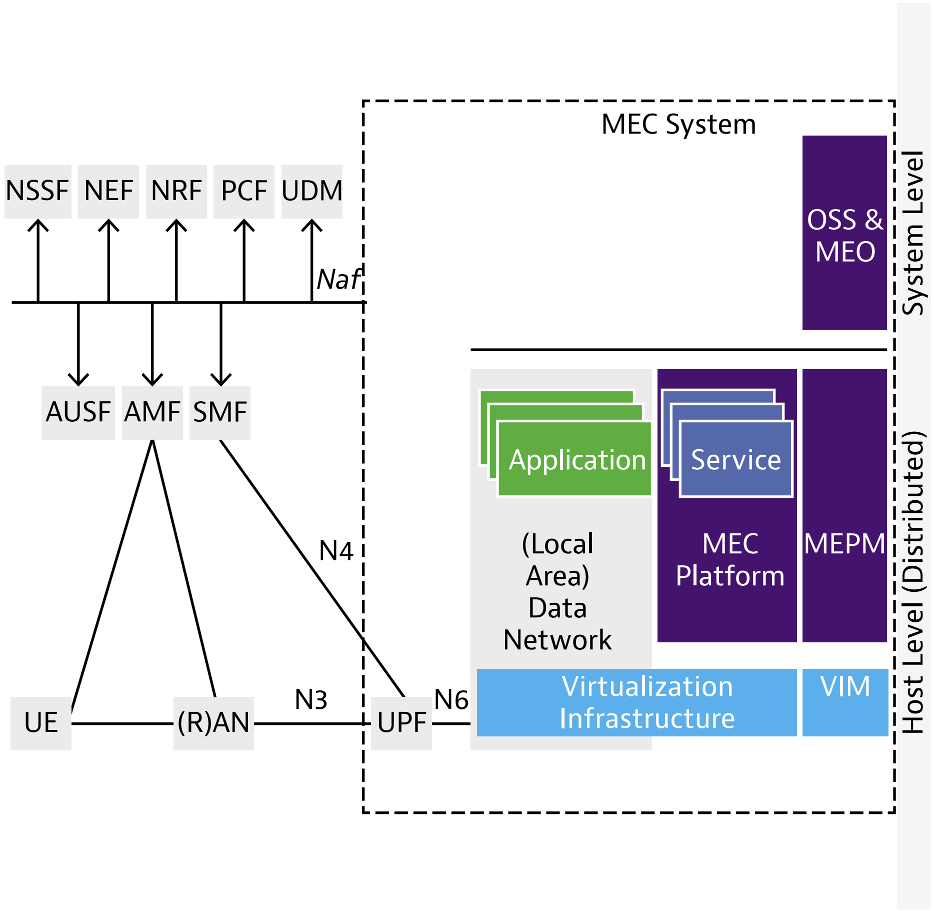


Figure 5.8 MEC deployment integrated in the 5G network

As part of the MEC system, reference Figure 5.8, the MEC orchestrator (MEO) is an MEC system-level functional entity that can interact with the NEF, or directly with the target 5G NFs if authorized. In this context it would logically act as an AF. On the MEC host level it is the MEC platform that can interact with these 5G NFs, again acting in the role of an AF. It is anticipated that the MEC host-level functional entities would be deployed in the 5G system's data network. Although the NEF, as a core network function, is a system-level entity deployed centrally together with similar NFs, it is feasible that an NEF instance could be deployed in the edge to allow low-latency, high-throughput service access from an MEC host.

5.8 5GS Common API Framework

Like ETSI ISG MEC, the 5GS SBA has adopted a RESTful design approach for its Application Programming Interfaces (API). These are being specified for the purpose of functionality exposure to third parties and well as other types of system internal communication. Functionality exposure to external entities is provided by northbound APIs. There are several northbound API-related specifications, e.g. the APIs for the evolved packet core (EPCs) service capability exposure function (SCEF) [18] and also the APIs for the 5GC NEF [16]. Therefore, to avoid potential duplication and an inconsistent approach between different API specifications, 3GPP has specified a common API framework (CAPIF) that includes common aspects considered applicable to any northbound service API.

It is possible to map architectural components of the MEC architecture onto CAPIF, which is presented in Figure 5.9. Specifically, MEC services produced by MEC applications, or the MEC Platform, can be mapped to the API provider domain in CAPIF. An MEC application (or MEC Platform consuming a service) is an API invoker in CAPIF. Finally, capabilities and functions of the MEC platform, such as service registration, can be mapped to 5G CAPIF core function.

In R15 there are five CAPIF reference points that map to MEC's single Mp1 reference point. In R16, a sixth is added for inter-Converged Control Plane Function (CCF) communication, analogous to MEC's Mp3 reference point. Note the "e" version of each CAPIF reference point refers to the variant covering interactions from outside a CAPIF's PLMN trust domain. Briefly:

- CAPIF-1: used for authentication, obtaining authorization, and discovery.
- CAPIF-2: for invoking the service APIs, e.g. SCEF (to AS/SCS), or NEF northbound interface (T8 and RESTful API respectively).
- CAPIF-3: provides service API access policy; provides authentication and authorization info of API invoker for validation; and for service API invocation logging and charging.
- CAPIF-4: covers publishing (facilitating discovery, via CAPIF-1) and storing the service APIs information.
- CAPIF-5: provides service API invocation log for auditing; provides monitoring information on service APIs status; and is used for exchanging API provider policy configurations.

What is not currently covered by CAPIF is the ability MEC offers to support traffic rule control and DNS handling, as described earlier in this chapter.

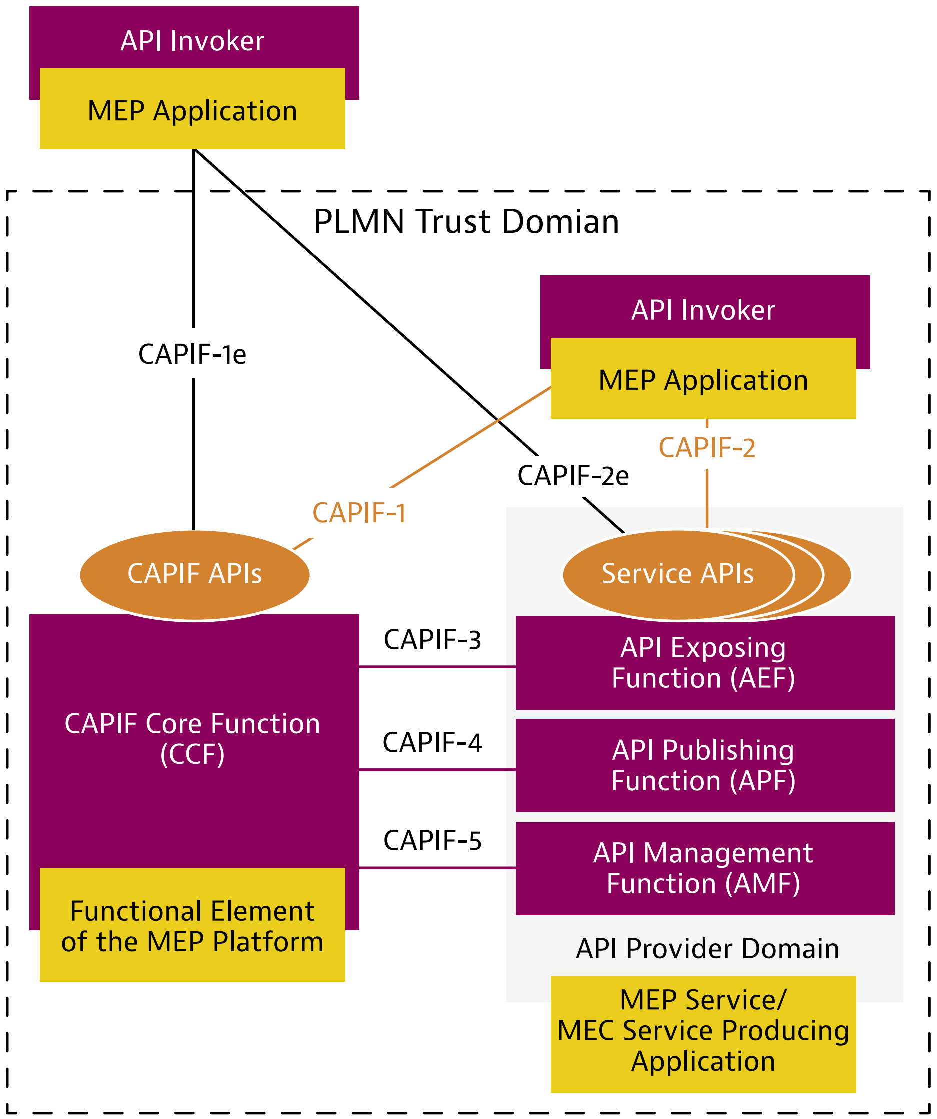

API Invoker
MEP Application
PLMN Trust Domian
API Invoker
MEP Application
CAPIF-1e
CAPIF-2
CAPIF-2e
CAPIF-1
CAPIF APIs
Service APIs
CAPIF-3
API Exposing Function (AEF)
CAPIF Core Function (CCF)
CAPIF-4
API Publishing Function (APF)
CAPIF-5
API Management Function (AMF)
Functional Element of the MEP Platform
API Provider Domain
MEP Service/ MEC Service Producing Application

Figure 5.9 Mapping architectural components of MEC onto the 5G common API framework

Traffic rule control is instead covered by the NEF, which supports AF influence on traffic routing [16]. This is a capability that MEC acting as an AF would utilize to influence the data plane, highlighting how MEC's Mp2 reference point between the MEP and data plane could in practice be realized.

5.9 MEC Impact on Operational Systems

How to operate and manage MEC must be given significant focus considering the architectural impacts of edge computing, which is anticipated to be typically deployed in a virtualized environment. Early functional and capabilities testing must span full lab-to-field deployment to ensure the desired behavior is achieved once the system is rolled out into a commercial network. Such test processes will cover conformance of the product components to the specifications; interoperability between those components, which may be provided by different vendors; a performance assessment of products under real-life load conditions, which will likely assess the reliability and stability of the system; and ongoing network assurance testing, providing insight into the operation of the product once deployed and the quality of experience of users of the system. Closely coupled are security and resilience testing, where the former is focused on identifying vulnerabilities of the system and the latter the system's ability to recover from failure.

The need for testing has been recognized within ETSI ISG MEC, with specific work items already underway to develop test framework guidelines [not yet published: ETSI GR MEC 025: "Testing Framework"], and also to define API conformance test procedures [not yet published: ETSI GS MEC 032: API Conformance]. This work focuses on testing methodologies to test devices implementing the standardized services and leverages similar activities in NFV{19} and existing ETSI best practices [20] & [21]. In this context, the two main and complementary testing methodologies are considered to be conformance testing and interoperability testing, which can be summarized as follows:

- **Conformance Testing** can be used to demonstrate that a product correctly implements a standard and its requirements. The product is formally referred to as an implementation under test (IUT), which forms part of the system under test (SUT). In this instance, the standards are those defined by ETSI ISG MEC and include testing a product's use of the MEC APIs. With respect to the APIs, the content and format of each message must be validated, as well as the sequence of message exchanges. Such tests are performed at the open standardized interfaces of the IUT and are generally executed by a dedicated test system that has full control of the SUT as well as the ability to examine all messages originating from, or sent to, the IUT. With its high degree of control over the SUT, the test system can detect both valid and invalid behavior according to a given sequence of messages with specific contents.

- **Interoperability Testing** can be used to demonstrate that individual functional entities within the MEC architecture will work together as expected. The functional entities are formerly known as functions under test (FUT), with each FUT being offered by a different provider. The collection of FUTs make up the SUT. The tests are aimed at proving that the end-to-end functionality between (at least) two FUTs aligns to the standards on which those functions are based. Tests are performed and observed at functional interfaces, including man-machine interfaces (MMIs), protocol service interfaces, and APIs. The high-level functions were portrayed in Figure 5.1 and include the MEC platform, the platform manager, the orchestrator, and the MEC applications that these entities enable. A limitation of testing at only functional interfaces is that the tests can only describe functional behavior, where sometimes it might not be possible to trigger or test protocol error behavior on the various interfaces Between the FUTs.

Although conformance and interoperability testing may be performed in isolation due to the distinct differences between them, the two techniques can be performed together to give combined results. This hybrid approach is described as interoperability testing with conformance checks.

ETSI ISG MEC has taken on the responsibility to define a testing framework and test conformance specifications. This includes the development of test purposes as well as the test cases, which constitute an abstract test suite. These provide the foundation for certification, for which there are dedicated bodies such as the Global Certification Forum (GCF). The GCF is the logical example since it is recognized as being the principal body in charge of certifying any product incorporating cellular mobile connectivity. Achieving GCF certification demonstrates that a device conforms to international standards for mobile technologies and helps ensure global interoperability between mobile devices and networks. It was founded in 1999 as a partnership among mobile network operators, mobile device manufacturers, and the test industry. It offers an independent certification program for mobile and IoT devices and helps ensure global interoperability and connectivity between mobile devices and networks based on international standards, such as those defined by ETSI and 3GPP.

Those programs continue to evolve, with new additions relevant to MEC use cases including 5G, automotive C-V2X, and mission critical services. Having an MEC product certification program would further encourage engagement of the ecosystem stakeholders in MEC technology and provide further evidence of the business viability of such products. GCF could be a potential candidate, expanding its remit beyond cellular mobile connectivity.

It has been highlighted that there is an expectation that MEC will be deployed in a virtualized environment, which offers new assurance challenges. Specifically, not only is the traditional performance monitoring and troubleshooting of the telecom functionality required, but also the performance of the underlying IT infrastructure, which affects the desired telecom functionality and must be included in the monitoring and troubleshooting coverage. To fulfil this requirement, data access agents must be able to operate in virtual environments to collect performance metrics by analyzing live traffic at virtual network interfaces, as well as collecting metrics on the performance of the virtual IT infrastructure. The MEC platform can aid such agents with its ability to route IP packets.

For instance, in tap mode, data-plane traffic is duplicated and forwarded to an MEC application that could implement a virtual network probe agent. Such an assurance agent can facilitate self-configuration of the virtualized environment by providing appropriate performance data/analytics (relating to both the telecom functions and IT infrastructure) to the policy control function responsible for reconfiguration decisions of the virtualized environment. With the agent running at the network edge, policy decisions can be made in near real time. This ensures optimal policy control based on the current state, enabling network reconfiguration decisions that are good business decisions as explored in the "Maximizing Profitability with NFV" TM Forum NFV Catalyst project [22]. The assurance solution must also self-configure in real time to match the reconfiguration of the network to ensure that there is no gap in network monitoring and troubleshooting. To receive such reconfiguration information, the assurance solution must have an interface with the orchestrator of the virtualized environment. In this manner, the assurance solution effectively becomes a part of the operational equipment chain, especially with respect to the assurance solution supplying the intelligence critical to ensuring the successful operation of the virtualized environment.

The ETSI ISG MEC-specified network information service APIs, such as RNI service, were introduced earlier in this chapter. These service APIs offer information that overlaps with the 3GPP defined trace data [35], for example RRC measurement report and handover event information. 3GPP refers to a trace collection entity (TCE) as the termination point for the trace feed, where the TCE may or may not be placed within the operator's secure zone. However, typically the TCE is located at an operator's centrally located network operations center (NOC). This is rather different from the MEC network information service APIs that offer a trace-like feed to authorized MEC applications at the edge of the network. These APIs also offer both query and subscription models, where the service consumer (MEC application) can cherry-pick the specific information of interest in a one-off request (e.g. current layer 2 performance information) or subscribe to receive event-triggered updates (e.g. handover or measurement events). This allows new use cases to be addressed, or existing use cases to be addressed in a new and innovative manner, such as distributed, localized, near real-time radio network assurance. Since the network information services provide connection level granularity, it is possible to offer subscriber-centric assurance solutions, with application-aware intelligence, that create a true understanding of the customer quality of experience (QoE). This in turn enables monetization of the network and automated edge-focused network optimization.

This section has highlighted that the edge computing environment has its own unique challenges and opportunities regarding testing. Such testing has a far wider scope than just the standardized services, particularly with regard to assurance once an MEC system is deployed.

5.10 It is Just the Beginning

This chapter has provided an overview of edge computing with an in-depth look at the ongoing activities within ETSI ISG MEC at the writing of this book. An architectural overview was provided, highlighting the key components and their purpose. Here the MEC application is perhaps of most interest, with the other entities primarily being made available to create an enablement framework for such applications. It should be clear that the MEC application is one component of the new deployment paradigm, which also has a client component and likely a backend central cloud component. The MEC application has the advantage of being in close proximity to the end user, i.e. client, and is able to exploit the contextual information available at the network edge to offer new and innovative services. Different deployment options for the new edge-based architecture also considered, including deployment in the 5GS. But this is just the beginning of the MEC era and the path in front of the full potential of edge computing is becoming clearer. Yet, as with any new technology innovation, there are new challenges ahead. Security is still a big concern. Distributed systems are always hard to provide security for compared to a centralized one, and, as we mentioned in the introduction, the topic of security in the new era of cloud computing is not included in this book since it deserves an entire book by itself.

Notes

1. Next Generation Mobile Networks Alliance, "5G White Paper," 2015.
2. ETSI, "Multi-access Edge Computing," [Online]. Available: https://www.etsi.org/technologies/multi-access-edge-computing.
3. ETSI, "GS MEC 003 V2.1.1: MEC; Framework and Reference Architecture," 2019.
4. ETSI White Paper, "Developing Software for Multi-Access Edge Computing," 2019.
5. ETSI, "GS MEC 011 V1.1.1: MEC; Edge Platform Application Enablement," 2017.
6. ETSI, "GS MEC 009 V2.1.1: MEC; General principles for Edge Service APIs," 2019.
7. OpenAPI Initiative (OAI), "The OpenAPI Specification (OAS) Repository," [Online]. Available: https://github.com/OAI/OpenAPI-Specification.
8. ETSI, "ETSI Forge," [Online]. Available: https://forge.etsi.org/.
9. ETSI, "GS MEC 012 V1.1.1: MEC; Radio Network Information API," 2017.
10. ETSI, "GS MEC 013 V1.1.1: MEC; Location API," 2017.
11. ETSI, "GS MEC 015 V1.1.1: MEC; Bandwidth Management API," 2017.
12. ETSI, "GS MEC 014 V1.1.1: MEC; UE Identity API," 2018.
13. ETSI, "GR MEC 022 V2.1.1: MEC; Study on MEC Support for V2X Use Cases," 2018.
14. ETSI, "GS MEC 026 V2.1.1: MEC; Support for regulatory requirements," 2019.
15. 3GPP, "TS 23.214 V15.5.0: Architecture enhancements for control and user plane separation of EPC nodes," 2018.
16. 3GPP, "TS 23.501 V15.4.0: System architecture for the 5G System (5GS)," 2018.
17. ETSI White Paper, "MEC in 5G networks", 2018.
18. 3GPP, "TS 23.682 V16.1.0: Architecture enhancements to facilitate communications with packet data networks and applications," 2018.
19. ETSI, "GS NFV-TST 002 V1.1.1: NFV; Testing Methodology; Report on NFV Interoperability Testing Methodology," 2016.
20. ETSI, "EG 202 237 V1.2.1: MTS; Internet Protocol Testing (IPT); Generic approach to interoperability testing," 2010.
21. ETSI, "EG 202 568 V1.1.3: MTS; Internet Protocol Testing (IPT); Testing: Methodology and Framework," 2007.
22. T. Poulos, "Maximizing profitability through NFV orchestration," 2016. [Online]. Available: https://inform.tmforum.org/features-and-analysis/2016/03/profit-maximization-needs-optimization-through-virtualization-and-nfv-orchestration/.
23. 3GPP, "TS 32.423 V15.0.0: Telecommunication management; Subscriber and equipment trace; Trace data definition and management," 2018.

CHAPTER SIX

Strategic Importance of Network Virtualization

Critical to the successful delivery of 5G is the programmability that will allow the dynamic assignment of resources to the corresponding tasks at any point in time and across the end-to-end network. However, in order to properly deliver both the incremental capabilities of 5G network programmability compared to legacy networks and to enable relevant cost savings, a disconnect of the peak and average utilization in terms of cost of delivering the service must be achieved. This means that the operator must be able to deliver service at peak utilization with the desired QoS but as usage scales down to a much lower level, the correlating resources assigned should be scaled back in order to align costs with the traffic levels. This is practically, perhaps only, achieved with the help of virtualizing the delivery platform.

Virtualization is not a new topic in the IT industry, or in the mobile telecom market, but to date, most of what is being discussed is how to lower the cost of owning and maintaining networks by sourcing lower-cost components than can be achieved with legacy networks. While this is important, it is only the beginning of what virtualization can and should do for telecom and mobile operators. Lowering costs is vital, but so is creating new revenue streams that are only possible because of virtualization and 5G. This point can't be stressed enough as operators are at a crucial point in history regarding the financial viability of their businesses. Operators that see the true possibilities of virtualization can greatly increase their average revenue per user (ARPU) and continue to have businesses with an operational profit model of forty-five to eighty percent in the future. Those that only see virtualization as a way to buy lower-cost components will most likely become a commodity business with a profit between five to eight percent, well-aligned with other utilities.

Virtualization is also at a critical point due to the number of employees retiring from the telecom workforce compared to the number who are entering. For example, a global mobile carrier in Europe has stated that by 2025, seventy-five percent of its workforce will have retired. At the same time, the top ten percent of students graduating university don't want to work in what they perceive as traditional telecom industry; rather, they prefer to work for OTT-leading providers who are perceived as cutting-edge in terms of automation and advanced business processes. All this creates a state of urgency to get to fully automated networks and business process as soon as possible, and most industry sources recognize that virtualization is a key enabler to automation.

6.1 Thinking Differently

Mobile operators have the advantage of reaching the public much faster than most companies can match. But how many operators plan to take advantage of this strategic advantage on the path to virtualization? Put another way, if operators are willing to think differently, what can they achieve? Can they gain a significant competitive advantage that lets their company thrive in the future?

Case in point: Reliance Jio completely disrupted the mobile market in India by thinking differently. Working in a greenfield environment, Reliance Jio was able to gain 100 million subscribers in the first 100 days of operations because they built an extremely cost-effective architecture that is fully virtualized and automated. They completely changed the market in what is considered the second largest mobile market in the world, and they did it all in less than one year of operations. Most important is that Reliance Jio caught most of their competitors off guard.

This is an important lesson. Virtualization will make winning and losing happen at a much faster pace than has been seen before. Also, an operator may do everything right on the path to full virtualization but still lose out because another company was more innovative in its thinking. Thinking differently is a big enabler for success in virtualization. Operators must go well beyond the technical steps toward virtualization and think of the end goal they want to achieve.

How can an operator, especially in this 5G technology inflection, become more agile as a company and create new revenue streams while also lowering the amount of risk by properly leveraging all that virtualization has to offer? This chapter will answer that question by giving operators the elements they need to consider so they can create a unique road map for the future. No two roadmaps will look the same, as each operator has its own unique challenges and advantages based on business models, network architectures, and business climates in its specific area of the world.

6.2 A Shift in Revenue Models

Total revenue has shifted over time from mobile operators to over-the-top (OTT) businesses such as Facebook, Netflix, Apple, Amazon, and Google. Mobile operators formerly used growth to determine when to invest in their networks and services, but that growth has largely stopped as OTT companies now provide the services and capabilities that appeal to subscribers. This can be seen in three main trends:

- Subscribers will pay for handsets that enable an app ecosystem that brings revenue to the owner of that ecosystem, such as Apple, and also to the creator of the app in a revenue share model. This enables individual app developers to take on the risk of success and allow the ecosystem owner to monetize the subscriber base.
- Companies like Google, Facebook, and Microsoft have created advertising-driven business models that draw from the same mobile subscriber base.
- Video-based companies such as Netflix, Amazon, and Hulu create incremental revenue growth from mobile subscribers as well.

For operators to succeed in the 5G future, they need a new way to manage revenue while also reducing risk. This concept isn't new to the world of mobile communications. In the past, DoCoMo had a service called FOMA (Freedom of Mobile Multimedia Access) that made money from apps and app brokering in the days of 3G. Subscribers used it through a non-web-based portal. It was successful because they had a distributed risk management system—much like Apple does today—that contained the infrastructure and the marketplace. They were not taking on the risk of developing the apps but shared in the revenue from the apps that were successful. Revenue-share models and a focus on innovation that provides value-added services will be critical for operators in the virtualized world.

One challenge toward innovation is finite scalability. Traditionally, mobile operators have bought solutions from other vendors as they try to evolve, such as buying a rack of special purpose hardware. In order for this to work, traffic forecast planning had to be perfect but accurate forecasting has proven to be challenging.

For example, SMS services grew much faster than operators had anticipated. Operators had a hard time keeping up with the network SMS assets to meet demand. The reverse happens as well. For example, operators overestimated the demand for rich communications services enhanced (RCSe) incremental voice/ video services. Operators built out capacity that gained no return based on how slowly demand grew for the services.

One of the great benefits of virtualization is that it gives operators a data center instead of dedicated special purpose hardware. With virtualization, operators can try out new services with their customer base without much risk. If the service is successful, operators can quickly add more capacity in the data center. If the service is not successful, operators can easily reallocate that capacity for something else. This model lets operators try out new services and apps without the dedicated risk as before.

Virtualization gives operators the starting point for this type of agility where traffic forecasting is not as critical. With the risks greatly reduced, are operators willing to bring their own OTT services and other innovative 5G use cases to market in this new 5G era? The industry would need verticals to buy into their new 5G ecosystems, which means that operators will need to provide services that offer a true value to subscribers for each vertical.

6.3 A Brief History of Virtualization

To fully understand mobile network virtualization, it is important to look outside the mobile industry and see what other industries have done. If operators think they are on the cutting edge of virtualization, they have already set themselves up for mistakes. There is a lot to learn from other industries.

Virtualization has been around for a long time in the form of enterprise virtualization. At the turn of the twenty-first century, the client server model connected remote desktops to a server as a way to better utilize resources throughout an enterprise. This was the first step towards full virtualization.

Fast forward to today and enterprise virtualization includes vast data centers and cloud computing that is able to scale up and down in real time to take advantage of every market opportunity. This has created agile and resilient networks that model what virtualization can achieve.

The mobile industry took the initial steps toward virtualization by adopting network function virtualization technologies developed by the standards organization ETSI, ETSI NFV, in 2012. A number of operators co-authored an ETSI NFV white paper that kicked off ETSI ISG with a focus on building a virtualization platform for operators. ETSI NFV plays an important part in replacing specialized hardware network nodes with virtualized network functions (VNF) that greatly reduce the cost associated with a network but offers little in the way of network scalability due to the isolated nature of the technology within a network. For virtualization to succeed, ETSI NFV for VNFs must incorporate an overall network and service management platform based on cloud native principles. This will help it gain the web-scale needed to offer on-demand services that subscribers want. ETSI is beginning to address this with ETSI ISG ZSM, which will be discussed later.

Another important aspect of virtualization is software defined networks (SDN). The concept of SDN was first considered by the Internet Engineering Task Force (IETF) in 2004. Practical SDN-type technology was originally designed and implemented by Google in 2010 for its own business. SDN is an architectural approach that separates the control and data planes of the network to maximize resource utilization. Mobile networks have incorporated SDN to some extent to replace MPLS, VLAN, or ATM. But to gain the most value from SDN, operators must connect the workloads within their networks with the SDN management agility. The reason for this will become more apparent as we discuss intent-based orchestration later in the chapter.

The commoditization of hardware has also played an important role in virtualization. Companies that host their own data centers have worked to disaggregate the network within those data centers as a way to lower the cost of equipment, lower power consumption, and create more flexibility in the way systems operate. One initiative, originally created by Facebook in 2009 as a way to collaborate with other Internet content providers (ICPs), has become known as the Open Compute Project (OCP). This was a response to similar strategies deployed by Google that was operating at a magnitude of scale higher than Facebook. Google has not chosen to make their designs public and thus Facebook and others had to start a new initiative to drive this development in the open.

OCP is now a global phenomenon and includes most of the ICPs through-out the world along with many hardware suppliers and several large financial institutions that host their own data centers, among others. The result has been a complete change to the data center model and has created an unprecedented amount of innovation—all of which is due to collaboration between groups that used to view each other as competitors.

Riding on the success of OCP within the data center, Facebook launched the Telecom Infra Project (TIP) in 2016 to create the same type of sharing and collaboration within the telecom market. The main goal is to create the same level of innovation and flexibility that was created by OCP. Facebooks says that it has saved $2 billion in CapEx through the efforts of OCP. TIP could produce similar benefits for operators as hardware commoditization allows for less specialization, which brings down costs and creates systems that can be used in multiple ways.

6.4 The Main Value of Virtualization

To date, the main value of mobile network virtualization has not been realized on a wide scale. In some cases, operators are currently thinking of lowering costs by piecing together the cheapest open source components or bits and pieces from their current suppliers. But to gain real value out of virtualization—and thrive as a company—operators must look at the real business need for such a network. Virtualization makes networks agile and dynamic, but operators must think of the business services they want to provide and then create a network and ecosystem to meet those needs. This is the only way to gain true value out of virtualization.

Let's take SDN as an example. In its effort to maximize resource utilization, the SDN controller watches for the right time where there is enough capacity for a certain application and then informs the business application that the capacity is available for a specific amount of time. SDN gives a specific capacity over a specific time frame for a specific business need, creating a very efficient network resource to perform a specific data transfer objective. This goes well beyond creating network optimization for the sake of network efficiency. By creating that optimization for the purpose of a real business need, operators gain a maximum utilization of the underlying hardware assets. As networks become fully virtualized, operators will be able to increase this utilization by automating their networks in a way that controls and orchestrates the network from a central location.

Mobile operator Elisa in Finland is a great example. Elisa had a business need to gain the highest data utilization rate possible while using the least amount of resources because its two to three million subscribers generate more data traffic than does the entire population of Germany. The operator sent its operations staff to training courses on Pearl and Python programming languages and asked that they write scripts that automate tasks they currently do manually. Elisa has now automated ninety percent of everything that happens in its network and within the operations team. The company moved from multiple people working in the network operations center to one person overseeing the system to make sure it runs smoothly. These tools have subsequently been made available commercially to other operators.

Through virtualization and automation, Elisa has created a network that has some of the highest data utilization rates in the world today. From a management perspective, they are already doing over 4G LTE what large operators around the world hope to achieve with 5G. The success of Elisa is based on determining the real business case for its unique situation and then creating a network to best meet that need. This is the mindset that operators need to have to be successful with virtualization and subsequent automation.

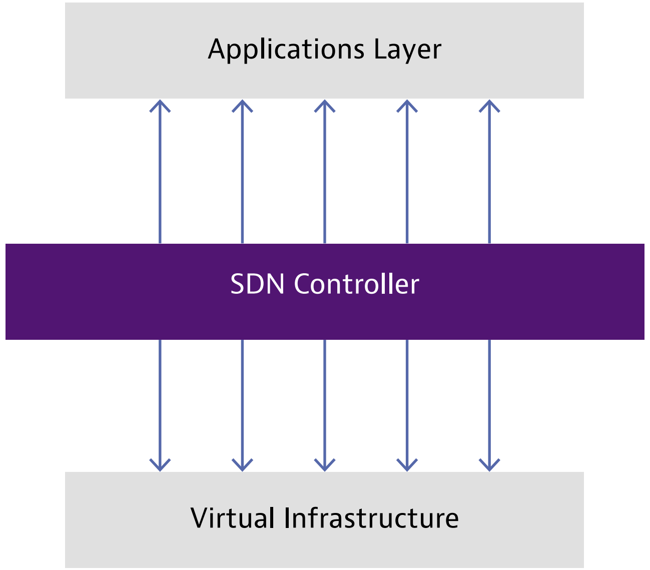


Figure 6.1 SDN controller maps virtual Infrastructure towards the applications Layer

6.5 Key Technology Options

To best evaluate how to virtualize a network and what parts of a network to focus on first, an operator needs to look at the key technology options available. Each technology has challenges and aspects that must be considered to gain the most value out of those technologies. In this section, we'll talk about the technology options as they relate to business strategy; some technologies are better suited to certain tasks, depending on the end goal an operator is trying to achieve.

We spoke of NFV before, which is a broad topic that encapsulates most aspects of virtualizing a network. Within NFV, there are three main areas: virtualized network functions (VNF), which are the specific functions that a network performs; management and orchestration (MANO), which orchestrates all of the functions within a network and will become critical as networks begin to automate; and NFV infrastructure (NFVI) that comprises the virtualized layer just above the physical hardware. The NFVI does its best to maximize the utilization of those hardware resources in order to allow the VNFs to perform their duties. We'll focus on NFVI here as MANO is covered in the automation chapter.

6.5.1 NFVI

From an operator perspective, NFVI can be any type of computing environment located anywhere across the network chain. This could be the home gateway that has a computing node, a mobile edge computing platform, an edge data center, a core data center, or even a huge centralized data center. In evaluating NFVI for a network, what does an operator need and what should be optimized for each of these environments? To answer these questions, an operator needs to know the design criteria for a network. From an internal development cost perspective, certain parts of a network will cost money to develop and other parts will be free based on the current network configuration of each individual operator. The best way to describe this is with an example from the Deutsche Telekom TeraStream design. Deutsche Telekom were redesigning their network based on two main criteria: 1) computing resources in the main data center are free to the company, and 2) bandwidth from the edge to the core is also free and unlimited from a cost perspective. Based on these design criteria, Deutsche Telekom came up with a network design that has no computing resources between the edge and the core. Also, all functionality to do traffic engineering and quality service management are generated inside the data centers that they already own. Obviously, this design has to stand the test of time in order to show that it can, and will, be widely deployed.

Regardless of what technology and architecture prevail in the mobile space, that technology is likely to be based on a set of core principles founded on a clear understanding of business strategy. Another operator might have other design criteria that influence how the network is designed. For example, an operator might already have a million edge locations with space, power, and networking that they can utilize. If so, then NFVI needs to be characterized so that these locations are free to the company. What is it going to cost to build out new areas of a network versus what the operator already owns? Operators need to use current assets as a strategic advantage. This is their starting point for setting design criteria and a competitive strategy for building out the rest of their virtualized network. Once an operator is clear on its unique starting point, they need to evaluate different components that will best get them to their end goal for virtualization.

6.5.1.1 Evaluating NFVIs

Vendors are rushing to supply the hardware and other components needed to make NFVI work, but with each operator having unique needs and uses for the components, how do vendors supply the right components to the right operators at the right time? In most cases, they don't. Although QuEST/TIA is working to change this, the group is still in its infancy. The reality is that most hardware components are designed with one major operator in mind. Those components meet the characteristics and the performance specifications that operator has for its unique network.

The hardware is readily available, but operators must ask themselves: Does this product suite meet my need and business case? If the requirements of the type of service being deployed are the same as what that major operator is doing, then most likely a similar component choice makes sense. But if not, then an operator must Figure out which specific combination of vendors will suit its needs. On top of needing to know their unique needs based on performance characteristics, costs, and capacity requirements, an operator must also know how to evaluate the different components.

When evaluating components, the data sheets for each vendor's offerings will show its specifications and performance levels, but doing an evaluation based on those data sheets can be misleading. QuEST/TIA set out to evaluate the different components vendors offer as a way to help operators get to the right end point with virtualization. The group found that information on the data sheet was not trustworthy and actionable. The vendors weren't lying, but the data sheet only represented using the component in a very specific environment. For anything outside that environment, the performance was much different, often worse.

This proves that operators must choose the components in their NFVI hardware and software configurations carefully based on the specific type of application they plan to run on that platform. For example, an NFVI receives a data packet on the network interface. The packet needs to be read on the interface and then forwarded up to the CPU, processed, and can then be written down to disk. If the NFVI doesn't have the correct capacity in all of these points, the operator may get good network and CPU performance, but the storage might not be sufficient because it can't be written down to disk fast enough.

In addition, just evaluating an NFVI based on items such as raw CPU performance or memory won't let an operator know how that NFVI will perform with specific applications. An operator must know the services and applications it plans to run in order to properly evaluate NFVIs. Once they do know, the person doing the evaluation can come back to the decision makers within an operator and make recommendations of which NFVI platform to source that meets the needs for each situation in the network. They can also recommend when it makes sense to have a vendor swap out certain components to make the platform perfect for that situation.

There are certain steps operators can follow to evaluate NFVI platforms for their unique situations. This list can include:

- Define the target functions to be deployed at each NFVI platform (edge vCPE, MEC, Core datacenter, Cloud datacenter) along with the anticipated traffic levels on each level and application (vFW, vEPC, vDNS, etc.).
- Define various permutations for each NFVI to evaluate relevant components such as the incremental performance given by SmartNICs or other acceleration techniques on the hardware level.
- Define various permutations of the software stack to be tested on each NFVI to understand the impact of the particular software components on the overall NFVI performance.
- Generate relevant traffic models to map against each platform and application.
- Onboard a representative VNF for each type on each NFVI, test the traffic profile towards the NFVI, and collect relevant NFVI metrics for CPU, memory, storage, and networking.
- Identify possible sweet spots for each application and NFVI platform combination in order to find good combinations of hardware and software platforms that deliver the desired performance at a reasonable cost.

North American operator Verizon went through a similar process and made the decision to use white box hardware from Dell along with ADVA Optical Networking's Ensemble Connector software to act as the brains of its NFVI platform. They decided this was the best solution based on their unique business and network needs, among many other choices. If other operators blindly follow this lead without sharing the same or similar business requirements, it is likely that those operators will make an unoptimized decision leading to possibly excessive costs or not enough resources to meet the localized requirements on their NFVI platform.

6.5.2 Virtualized Infrastructure Managers (VIM)

Once NFVIs have been evaluated and selected, they need to be managed along with software to deliver network services to customers. This is done by implementing a virtualized infrastructure manager (VIM). There are several types of VIMs and each has its strengths in certain situations. The oldest and most deployed to date are virtual machines (VMs).

6.5.2.1 Virtual Machines (VM)

VMs sit on top of a physical machine and create multiple virtual machines. The operating systems and any relevant applications share hardware resources from one physical server or a group of servers. One of the interesting facts about VMs is that they require their own operating system and the hardware is virtualized. This arrangement allows an operator to drive up the utilization of the hardware versus having one machine powering one application on top of it. Operators that use VMs tend to use a model that ties certain hardware infrastructure back into a specific virtual machine in order to leverage specific hardware-based acceleration techniques.

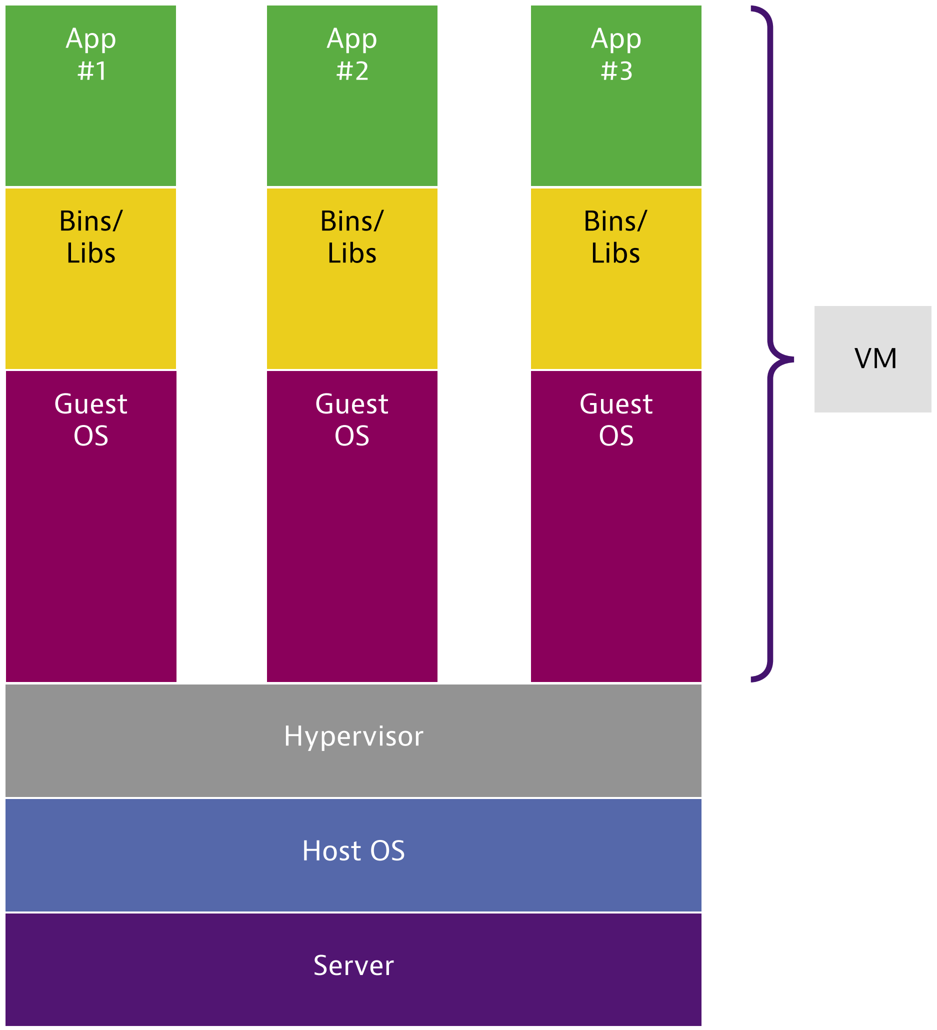


Figure 6.2 Virtual Machine (VM) stack with individual guest OS

The challenge VMs pose for operators is that traditional mobile networks use dedicated processors to perform certain tasks that need high performance levels. When those processes get generalized into a piece of software on a general-purpose CPU, the performance levels are generally much lower. Companies like Intel have tried to solve this problem by building acceleration capabilities into the general-purpose hardware. The data plane development kit (DPDK) is an example of this for network processing optimization. In this case, the acceleration is a functionality of the network interface card. This gives a server the ability to reach high performance levels in network applications. Without this feature, operators would get only a fraction of the performance out of the same machine.

If an operator has one to five network cards in a server, it can assign those resources to the VM and the VM then utilizes the acceleration capability. The problem is that a one-to-one ratio of server to VM rarely occurs in a virtualized network as it becomes too costly and reduces the dynamic scalability of a network. VMs are also notorious for using a lot of RAM and processing power since they need to run a virtual copy of all the hardware and operating systems individually.

It is more common to see multiple VMs sharing a server. For operators, this creates an issue for acceleration. For example, if a server has only one network card that has acceleration capabilities, but has two VMs sitting on top, then one VM will get the acceleration and the other won't. This means that the applications running on the second VM suffer. This creates network constraints that must be taken into consideration. Because of this, VMs are better suited in the core of a network where performance is needed and there are more hardware and computing resources available. Various industry groups are working to extract this further and provide non-blocking APIs into the acceleration components. It is yet to be seen if this delivers one further level of flexibility and agility or just another possibly cumbersome integration point for developers.

6.5.2.2 Containers

Containers are a newer technology than VMs and don't include the operating system. A container only contains the binaries and libraries needed to run an application. For example, if there is an application that runs on Linux, then an operator could run a container that runs a Linux-based host operating system without the need for any Linux environment setup. The container will reuse Linux from the operating system on the server. As a comparison, VMs have virtual hardware that has all of the characteristics of normal hardware. This means an operator would need to install a complete operating system on the VM with all of the environments needed.

A container is limited to what the host has for installations and capabilities but doesn't have the same overhead requirements as a VM. This makes a container more efficient as long as it doesn't need additional capabilities or another type of operating system. A container makes it easy to build something and is an agile way to deliver items such as micro-services.

For example, a subscriber wants to install a new app on their phone. It doesn't matter what operating system the phone uses because the platform hosting the app takes care of that in a container environment. This makes some implementation of containers a smart choice in areas such as subscriber-facing application platforms where flexibility is needed, since subscribers may be using different operating systems to access a certain app.

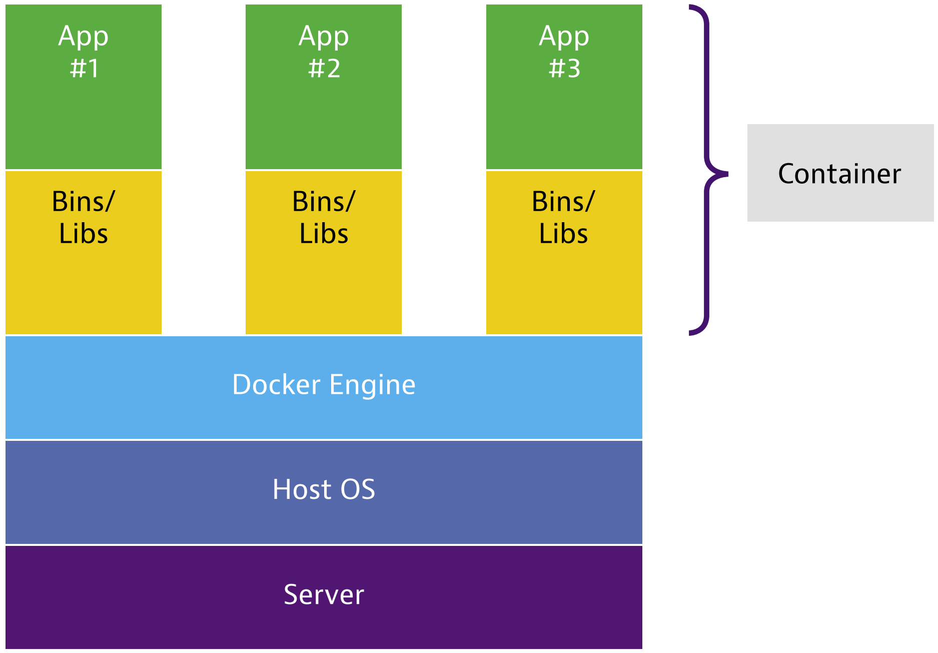


Figure 6.3 Container stack with reused OS from Host

For this reason, containers are becoming more popular. With Intel, among others, the industry is working on a type of API that goes into the hardware and will be able to then go into a container. Over time, this will give containers similar capabilities to a VM as the VMs capabilities will get exposed up through the layers of software making it possible for the container to use the same type of underlying APIs.

6.5.2.3 Software Defined Infrastructure (SDI)

One of the interesting aspects of both VMs and containers is that they try to optimize the finite components in a server such as CPU, memory, networking, or storage. This creates network constraints and service challenges depending on what a network is trying to achieve. As networks continue to become more dynamic and automated through virtualization, trying to optimize finite resources within a server will become a limiting factor on a network.

To resolve this issue, Intel has created a new concept called Software Defined Infrastructure (SDI). Instead of having a rack in a data center filled with individual servers, SDI utilizes a rack that has one part made up of memory, another part is disks, another part is

CPUs, and so on. Then based on a specific need within the network, these components would be assembled by a hardware orchestrator to create the optimal server for the network function. In this SDI environment, all of the components are interconnected by an extremely high-speed backplane enabling the instantaneous assembly of a specific configuration for a specific task.

SDI makes it easy to allocate components to certain tasks when they are needed, and each configuration looks like a stand-alone server to the operating system. From the software side, no one would be able to tell this is a software-composed hardware architecture.

In the near future, SDI will provide more flexibility to VMs and containers from a hardware point of view. Several networking vendors, including Ericsson and Nokia, are already bringing out data center solutions based on SDI. In the not so distant future, SDI may overtake containers and VM by providing the same capabilities in a more holistic environment if the cost overhead of an SDI model is less than the gained efficiency of the resource utilization.

In deciding the best way to move forward with NFVI and VIMs, operators need to take a step back and focus on how network functions can help drive their business goals. Operators usually focus their optimization efforts in areas of the network where they know they can make money. But a more important question is: Where are the areas in a network that an operator can't make money? These are the areas that are critical to optimize so that they take up as few business resources as possible.

6.5.3 Private and Public Cloud Environments

All of this talk about evaluating different parts of a network to optimize and the pieces needed to make it happen leads to discussions about private and public cloud environments. Just as what to optimize in a network will have an impact on the business model moving forward, so does the strategy used in determining whether to build a private cloud environment or buy into the public cloud.

The cloud provides the scalability and redundancy needed to create dynamic and resilient networks to handle all that 5G will bring. For operators that excel at evaluating NFVIs and matching those assets with real business needs, building a private cloud environment is the best solution as it allows the operator to create a network best suited to its individual business needs. They could also build a private cloud environment that leverages their competitive advantages. But some operators may find that they are good at providing services but not good at evaluating NFVIs to build the ideal network. They may also find that it costs too much to build such a network themselves. For these operators, buying into a public cloud environment might be a better solution. It should be noted that in many cases the large public cloud operators have added proprietary hardware acceleration capabilities to their servers in order to accelerate specific functions.

But each public cloud platform has its strengths and weaknesses, which means that an operator would need to evaluate each public cloud solution based on the VNFs or applications they plan to run. This is the only way to determine the value of each public cloud platform.

In the past, public cloud platforms have not been a good fit for operators to run the core or RAN functions due to latency issues. Low latency and on-demand services are the main goal of 5G, so operators might be better off using their own networks that include a Mobile Edge Computing (MEC) platform they can use and control to make low latency a reality.

Recently, Amazon has tried to address the low latency issue in its Amazon Web Services (AWS) cloud platform by buying grocery store chain Whole Foods. This may seem like an unusual move, but Whole Foods can provide a hosting location for state-of-the-art localized data centers in each of its stores that can serve as a platform for MEC-type services. Whole Foods has locations that are located in densely populated areas near subscribers and could serve the services that need low latency while sending the rest of the services back to the main data centers in the cloud.

AWS could also provide the compute environment or provide the complete service for mobile operators. AWS has been working to make its network more appealing to mobile providers by providing the Evolved Packet Core (EPC) as a service. In this case, mobile operators would only need to own the base stations in the RAN; AWS would provide the complete service to operate on the operator-owned spectrum or possibly in an unlicensed spectrum. Depending on the proximity between the public cloud host location and the end user's RAN node, this may not allow for 5G Ultra-Reliable Low Latency services. In such cases, the operator would need an MEC-like distributed NFVI in order to support certain autonomous driving scenarios, for example.

AWS is currently the largest cloud provider in the world, but other cloud providers are building the necessary infrastructure needed by mobile operators. For operators, decisions need to be made about how to put the right compute environment and software architecture around a certain amount of geographic assets and how to best utilize those assets. With this in mind, operators need to decide if a private or public cloud environment is the best solution for their business.

6.6 Open Source Projects and Standardization

The mobile industry has historically developed new technologies with each company trying to make money by getting their technology to be the new standard for the industry and then licensing that technology. This process made creating standards a lengthy process as each company fought to have their technology become the standard.

Mobile standards group 3GPP has a philosophy of standardizing everything from the big systems of a network down to each component. The members of the group must agree upon each standard created. On the other hand, standards group IETF has taken the position of standardizing the components but not the overall systems. Within this group, there is usually only one type of each component, but the systems are developed in an open source environment. At the same time, IP standardization happened by companies proposing an RFC to IETF that explained the value their technology brings to the industry. That RFC was then backed up with software implementation and was evaluated for a period of time and either accepted into the process flow or not.

But the push toward virtualization is happening too quickly for these traditional methods to be used. Virtualization is being led by open source projects where the ownership of the intellectual property rights (IPR) in itself does not lead to any commercial licenses. Instead, it leads to technology that is being developed for the overall good of the industry. That doesn't mean that everyone is working in the same direction though. Several open source groups are working on very different strategies to get to what they consider to be the ultimate environment for virtualization. It's worth looking at some of these projects to see if any of the various strategies can benefit what a specific operator is trying to achieve.

Most of the open source organizations relevant to this discussion were started as development projects within telecommunications companies. Then these companies brought that development to the open source community by going to the Linux Foundation and asking to make a project out of their development work. The Linux Foundation already has a structure in place for this; companies wanting to start a new project must adhere to that structure and the rules of the organization. This helps each project become beneficial to the mobile industry as a whole as everyone is using the same structure for development. Outside of telecom, this model of open sourcing projects has been common for a long time with significant contributions from companies such as Netflix, Facebook, and Google.

6.6.1 ONAP

Open Network Automation Platform (ONAP) is one of the projects to come out of this process. ONAP originally started as internal development work by AT&T. Called ECOMP internally and AT&T Domain 2.0 externally, it had written specifications of what needed to be built to enhance mobile networks. AT&T wrote close to eight-and-a-half million lines of code, most of which was later donated to the open source community.

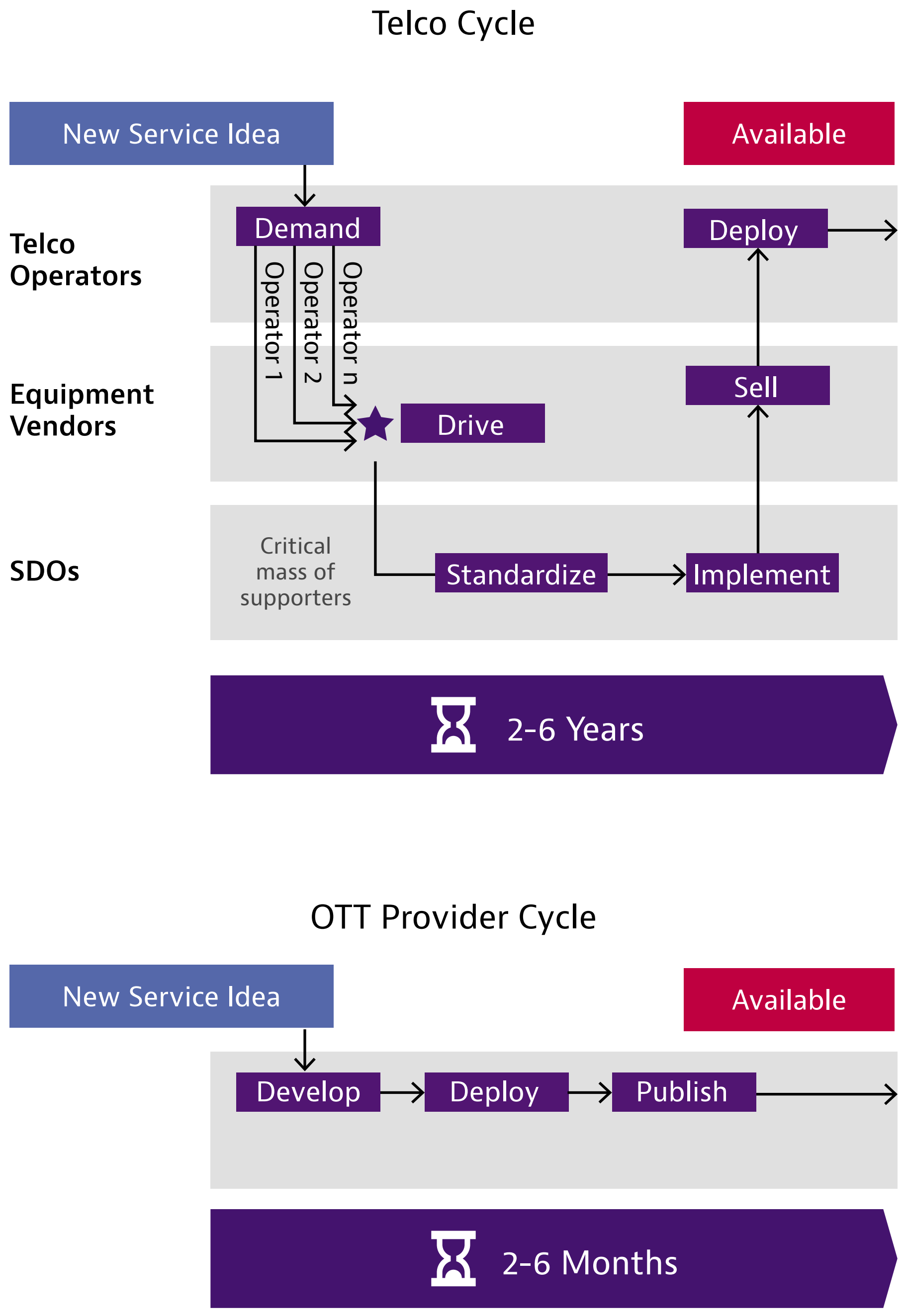


Figure 6.4 Comparing traditional telco cycle and the OTT provider cycle

At roughly the same time, China Mobile started the Open O project. The main goal of this project was to develop an orchestrator to drive other aspects of a network. China Mobile donated four million lines of code to the open source community. Open O and ECOMP eventually merged together to create ONAP. According to ONAPs website, the organization "provides a platform for real-time, policy-driven orchestration and automation of physical and virtual network functions that will enable software, network, IT and cloud providers and developers to rapidly automate new services and support complete lifecycle management."

The system was architected by multiple people and the quality of the code is not quite on target as it was originally written without the realization that it would be used in an open source environment. Significant portions of the code and architecture are now being redone to make it more viable.

One of the nice aspects of ONAP is that its architecture could be, after the re-architecture, broken down into smaller components with well-defined interfaces. This is useful as operators will be able to use only the parts of the architecture that suit their individual network needs. ONAP's design does not force compliance with the ETSI NFV project on the MANO level. One of the main reasons for this was due to the lack of standardization of the OSS process in MANO. This is today an effort undergoing research and standardization to manage. More information on ONAP can be found at https://www.onap.org/.

6.6.2 OSM

Open Source MANO (OSM) was created to focus on the MANO deliverable within ETSI NFV. MANO is technology that coordinates the efforts of virtual machines that deal with the VNFs in a network. For example, MANO can deal with how to control a machine that is in charge of spinning up a network slice. This group was originally started by Telefonica as an internal effort called UNICQA Project. At the time, Telefonica was trying to find a more efficient way to manage all of their OSS and base station subsystems (BSS).

Telefonica created the new architecture internally, but when the company decided to go down the path to virtualization, they donated large portions of the code to OSM. OSM now includes operators such as Telenor and Verizon, among others. Currently, OSM is considered by many industry analysts to be better architected with a better set of code than ONAP. But over time, the re-architected ONAP will not only be able to function with OSM, it will most likely have components that extend beyond the capabilities of OSM. This will allow operators to use OSM if they prefer but also use the components from ONAP that go beyond the capabilities of OSM to build their own ideal network. More information on OSM can be found at https://osm.etsi.org.

6.6.3 CORD

Some operators will go solely with ONAP as their main framework and others will use OSM only as a way to simplify their architectural strategies. This may change once ONAP is re-architected for areas beyond OSM. Other operators might choose a framework that is somewhere in between, such as CORD. Central Office Re-architected as a Datacenter (CORD) was started by bringing some components of its framework from the IT industry that had already virtualized their networks. This creates an interesting combination of IT and telecom from contributors such as Google, Verizon, and China Unicom among others.

CORD states that it "combines NFV, SDN and the elasticity of commodity clouds to bring datacenter economics and cloud agility to the Telco central office." More information on CORD can be found at: opencord.org.

6.6.4 Standardization

The road to creating standards for virtualization will look much different than the methods used for legacy telecom standards in 2G, 3G, or 4G networks. Open source plays a big role in this, but so does the difference between the IT and telecom industries. The IT industry is involved in the virtualization of mobile networks and have been doing automation and orchestration in the enterprise and web-scale environments for several years. Most telecom vendors are coming at the problem from a traditional telecom angle by trying to create an ecosystem similar to what they have done in the past and to leverage their current position of strength.

This is creating a lot of tension in the industry due to cultural differences between the two groups as well as differences in what the fundamental building blocks should be and what is required to make virtualization happen. The telecom group is more interested in the final outcome and how to get there. The IT group is more interested in patterns and repeatability so that they can create a system to be put in place that continues the evolution of a network beyond the initial virtualization phase.

For example, some of the companies from the IT industry want to create a data model that can be used over and over but provide that data model with different inputs to create the needed outcome for each individual operator. The telecom group would rather build an entire architecture each time a specific outcome is needed. Both have their positive aspects, but the IT model could provide a more flexible environment that can adjust with market conditions and as new technology is developed. For the IT model to work, an operator must define what specific outcomes are needed. The challenge is that the possible variations and outcomes are significant, and many in the telecom industry think it is close to impossible to know everything that is needed ahead of time. This difference of opinion is slowing down the progress towards creating standards for virtualization. It has become clear than certain operators are not waiting for a standard on every aspect of their network, and instead, are moving forward with their own implementations based on architectures from the open source community. Whether to wait for a standard or move forward using open source is largely up to each operator. Waiting could put an operator behind the competition but moving forward might take an operator down a less successful path as the technologies evolve.

6.7 Operational Considerations and Patterns

One way for an operator to get ahead of the competition is to rethink how it handles its development, operations, and network management internally. The network of the future will be less about trying to create the perfect network for each stage of that network's lifecycle and more about creating a flexible network and business organization that can easily evolve as subscriber needs change. Several aspects to this will be covered in this section.

6.7.1 DevOps and NetOps

DevOps is a term that was first coined by Facebook and Netflix. It's a management philosophy that says the same people that develop a capability or service must also be put in charge of the operations of that service. Traditionally, mobile operators have had separate development units that hand off a new service to an operations team that must deal with the challenges of that service on their own once it goes live.

Using the DevOps management approach, the team developing the new capability or service know that they are ultimately responsible for all aspects of the new service, including once it goes live. This creates an environment where the team can quickly validate the capability or service and make the needed changes, ensuring that what is built is ready for live operations. If anything doesn't work as planned, the team can easily make corrections at any time during the life of that service.

DevOps has the ability to create a nimbler organization with much smaller teams and is better suited for web-scale-type services and network configurations. It also lends itself to creating automated operations for repetitive tasks, which naturally leads to the automated operations environment needed once networks are fully virtualized and automated. NetOps is the same concept as DevOps. It is the term being used to apply the same management principles to the networking part of an organization.

6.7.2 Radically Different Concepts

The thought of a network going down due to a disaster is not something most operators want to think about. Currently, operators plan for disasters and then test and validate the resiliency of a network in the controlled environment of the lab. Once the network is live, tests are done once a year during off-peak hours to make sure a network can handle a potential disaster. What operators don't tend to do is test in real-life scenarios with peak traffic running on a network.

Of the mobile networks that have gone down, most happened in situations that were designed not to happen. For example, a nationwide European operator was upgrading three home subscriber servers (HSS/HLR) that held all of the subscriber information for the entire country. During the upgrade, the technicians accidentally tripped the main power supply and the secondary power supply failed to start for one of the HSS units. This forced all of the traffic onto the other two units as is the design in this scenario.

But the second HSS unit immediately failed as well, pushing all of the traffic onto the last remaining HSS unit. The third unit became overwhelmed and shut down, taking down the network. Subscribers were without service for several hours—all because the redundant power supply for one HSS unit failed. In another example, a Swedish operator suffered a power failure to its main data center. The backup generators failed to start, which created a nationwide service outage.

As virtualization and automation transform the industry, operators must test their networks in real time with peak traffic to make them as resilient as possible. Operators will no longer be able to design for the best-case scenario and hope for the best. Constant testing will be needed as networks become more dynamic due to virtualization and 5G services. It will be much easier for operators to start doing this now as they transition to fully virtualized networks, as networks will become more complex as they evolve.

An intriguing solution to this problem can be found in concepts like Chaos Monkey and the Simian Army, an open source project by Netflix. These solutions were created to answer the question: Will the network survive a disaster? The solutions travel through a network and randomly shut down parts of a network at any given time including peak and non-peak traffic times. Netflix does this an average of fifty times per month and the timing is completely random so that the company cannot prepare ahead of time. This helps Netflix make sure their networks can deal with any network problem that may arise and design the network accordingly. Over time, this way of testing creates a network that is resilient to any disruption that may occur. Several OTT providers are already using such solutions in their cloud networks. Operators need to decide what is the total cost of a major outage versus training staff to operate and test using these radically different concepts.

In addition, operators need to think about how they can reduce costs during a disaster by leveraging the architecture of a cloud environment. In a cloud environment, standardized data centers run software that is transferable between virtual machines or between containers. In this environment, there is no longer a need for support contracts that say emergency staff will be in place within minutes to fix the problem.

If part of a network goes down, the software and data are switched to another data center owned by the operator or to a third-party cloud provider so that service is not disrupted.

This gives the operator time to fix the problem during normal business hours without the need or expense of an emergency team or emergency service clauses in contracts on hardware as are common today. In this scenario, there could be a maintenance staff in the data center that gets a list once a month that shows what needs to be fixed from a hardware point of view. There is no longer a rush to make this happen given that another data center is already handling the traffic. This is one more way that the flexibility and agility of virtualization can help operators thrive in the future.

6.7.3 Model and Intent-Driven Orchestration

Another way to add flexibility and agility into a network is through model and intent-based orchestration. As discussed in the standardization section, the IT groups involved in network virtualization wish to create a data model or system that can be used over and over where different inputs are added each time to create the needed outcome for that specific situation. This model is not as concerned with the individual components needed to make this happen. Instead, it uses intent as the main goal and lets the system figure out the best way to reach that goal.

For example, an operator wants to deliver the best quality high-speed video of a sporting event. This becomes that operator's intent. This intent is given to the system, which then figures out the best way to make it happen based on the resources available at that time. In this model, humans no longer need to figure out exactly what needs to be done. The system figures out the most efficient way to make it happen on its own.

An operator must trust that the system is better at orchestrating and fixing a network than they could do it themselves. As networks become more complicated, there will simply be too many scenarios for humans to deal with manually. The nice aspect of intent-based orchestration is that it saves operators from needing to figure out every possible scenario that could happen in every situation, such as trying to deliver the best quality high-speed video. Instead, operators can make overall decisions as events unfold and let the system decide the best way to resolve a network problem or achieve an intent-based goal. For this to work, operators must become comfortable with handing over control of the decision making to a network.

What is right for each operator and how fast this intent-based orchestration can be put in place will depend on a few factors. Operators must look at their current installed base of vendors and see what can be retired from the network and replaced at what time. The faster this happens on the path towards full virtualization, the faster intent-based orchestration can be put in place.

The speed at which this model can be leveraged will also depend on the abilities of the current employees of the operator. In the new intent-based model, operators will need employees who require little direction and can act on their own once given a goal or directive. Depending on how that employee base currently looks will determine how quickly an operator can make the transition to intent-based orchestration.

Employees must also know what questions to ask the system so that it starts working on the correct solution. Going back to the example of best quality high-speed video, if an employee asks the system to deliver the best quality video, the system may slow down the service or other services to make this happen. If an employee asks the system to deliver the lowest-latency video service, the system might degrade the video quality to meet the goal of low latency. The employees must be skilled enough to know how to ask the question to get the desired results.

6.7.4 Reactive Versus Proactive Assurance

Up until recently, most assurance has been reactive in that a fault is identified that has already happened in a network. But proactive assurance solutions are now being developed that can proactively look for areas in a network that have a high probability of breaking and notify operations before the fault happens. As networks become more complex and dynamic, it will be impossible for operators to manually keep tabs on all aspects of a network regarding service assurance. Because of this, proactive assurance solutions are critical to a successful virtualization strategy.

If an operator knows how pieces of a network can break and has expectations of normal network performance that are accurate most of the time, then network failures can be predicted before they happen. This is how most proactive assurance works. For example, a node might become slightly slower in responding; recent past performance levels give the operator a good idea of how that node should be performing based on current network levels. If the node is reaching its maximum capacity, the node responding slower is a leading indicator that it could soon fail. An operator will have a small window of time to decide what to do.

More and more, software is taking the action of what needs to happen in this scenario, but if the operator is not yet at the level of intent-based operations, they will still need to apply a policy to that software to tell it what to do when this type of scenario arises. For example, should the system block some subscribers so that the subscribers already using the service can continue without degradation, or should it constrain the network, which degrades the service quality for everyone? Most operators would choose the first option in this example, but this is an area that operators must think about as systems become more complex and automated.

Another area of consideration is the data center. As mentioned previously, data centers can now scale up resources as they are needed by adding servers or components of servers, as is the case with SDI. In a fully virtualized and automated environment, policies should be put in place to tell the system what to do as leading indicators hit their trigger point to take action. For example, at what service level should the data center add more servers as traffic increases or take those servers offline as service levels decrease? This is known as scale in/ scale out; systems will be able to do this on their own but need an operator to decide what the trigger points should be. Once the trigger point is determined, the system can then take the action needed whenever that trigger point scenario arises.

This idea can be taken one step further in that a virtualized and automated network can also be predictive in when it scales in or out based on recent past trends. This ensures that a network proactively manages the capacity to accommodate upcoming usage. For example, a network might know that subscribers watch a lot of video at a certain time each Monday night based on past network usage. Knowing this information, a network could proactively scale in the resources needed slightly ahead of time to make sure the subscribers get the video quality they desire. Going back to the idea of intent-based systems, an operator in this scenario might tell the system to provide the best video quality to the most users. The system would then look at its options based on current available resources and initiate the needed network changes to make that intent happen.

As operators build out their 5G networks, proactive assurance should be incorporated into a network to create more flexibility and resiliency, because the systems will be able to make decisions that not only keep services at optimal performance level, but also keep network faults from happening in the first place.

When operators move from reactive to proactive assurance, they also need to change how they measure success. Today, operators look at the various KPIs and see that ninety-eight percent of calls on a network were successful, for example. In a proactive assurance model, the ninety-eight percent success rate is still important, but an operator also needs to ask if they got all the revenue they wanted from that service. What should a network do to accomplish both? This is where intent becomes very important. If the system is not given the correct intent, it might optimize a network for performance at the expense of revenue. Operators must give a network an intent such as "provide the best service while generating the most revenue possible," in order to be successful in a fully virtualized and automated environment.

6.7.5 Hybrid Operational Models

All of the items so far in this chapter can help an operator thrive in a fully virtualized and automated environment, but what can an operator do to ensure success as they transition from a legacy network to a fully virtualized one? As that transition happens, part of a network will be DevOps or NetOps and the other part will be legacy systems. This creates a very complicated environment that operators must get through as quickly as possible.

To do so, operators must have clear goals of what the final network and organization should look like. What parts of the network are virtualized and automated? What does the employee count look like and what are their responsibilities? How open are the employees to change and how fast can that change happen without the employees deciding they want to leave the company? There must be clear goals on what should be optimized and built first based on an operator's current network. Goals must also specify what gets optimized second, third, and so on.

The recommendation is that an operator take their best employees and have them build a virtualized environment in a restricted market that can be used as a development area. Most large operators already have trial markets that can be used for this purpose. If not, choose a city or region as a test market for virtualization. If it doesn't go well, then the damage is minimal and doesn't impact main markets or revenue streams. If it does go well, then it becomes a learning experience that can be applied to the network as a whole.

6.8 Zero Touch – The Vision of Full Automation

Getting to full automation is the ultimate goal of virtualization. But in the current operations models of most mobile operators, many tasks are both manual and repetitive. These are often managed by strict processes at the lower tiers of the operations staff. These models have been developed, tuned, and optimized for years. During periods of growth and high operational margins, the costs of operations were comparatively acceptable and most management attention was focused on other aspects of the business such as customer care, which often got significant scrutiny as each interaction was relatively expensive and led to customer churn if not properly managed.

As operational profits have declined further, cost scrutiny has become important over the last ten years. Most operators have worked step-by-step to identify significant cost drivers and have tried various methods to engineer them out one-by-one, often leading to stepwise improvements but not allowing the operator to move the needle enough to be comparable to the operational agility of OTTs. In order to achieve a comparable level of operational agility to an OTT, operators need to rethink everything from scratch—starting with a blank sheet of paper and building up from there. This requires significant soul searching of what is actually desired in operational qualities, and operators will need to be able to properly identify the patterns and models required to make this happen going forward.

By the end of 2018, it is still unclear if any established operator can effectively turn their operational organization around and achieve full automation. It has been shown that greenfield operators, such as Reliance Jio, have been able to achieve full automation of most if not all of their operational processes. But this has happened with the added expense of an expanded DevOps or NetOps team. In the case of Reliance Jio, they acquired the specialist firm Radisys in order to build and extend the overall OSS and BSS landscape as well as the ability to build out their own infrastructure.

The cost-effective solution to full automation might be ETSI's zero-touch network and service management (ZSM). In a white paper from December 2017 (http://bit.ly/2TTf8PS), ETSI says, "The goal is to have all operational processes and tasks (e.g., planning and design, delivery, deployment, provisioning, monitoring, and optimization) executed automatically, ideally with 100% automation and without human intervention." The white paper also points toward the critical nature of industry collaboration to drive standards, best practices, and open source of the overall industry requirements. This is a daunting task with a strong call to action. This publication has since drawn significant attention to the cause and has grown the membership of the ETSI ZSM industry specification group (ISG) to sixty-five companies and organizations (full list: http://bit.ly/2I5bW1h) with work ongoing. The first set of requirements has been published together with a draft target architecture (read the blog post: http://bit.ly/2TK6UJz) as seen in the diagram below.

The key driving principle of the ETSI ZSM ISG work is the separation of concerns into various domains. It is highly likely that any implementation of a zero-touch framework or architecture will be founded on cloud-based principles. The interested reader is encouraged to study the concept of the 12FactorApp by visiting https://12factor.net. Doing so will provide a better understanding of this model as well as the overall process.

An implementation following these principles would have the highest probability of success given the nature of the industry and the significant software development process maturity that has been driven from web-scale providers. Another significant resource to study to gain further understanding is the Google SRE book (http://bit.ly/2TKEFKZ). This provides a unique insight into how one of the world's largest environments is built and operated, and more importantly, provides insight into the process and methodology behind the current operational model. Not every operator will operate at the scale of Google and will thus not leverage all of the same characteristics, but the overall concepts and thought process are highly leverageable.

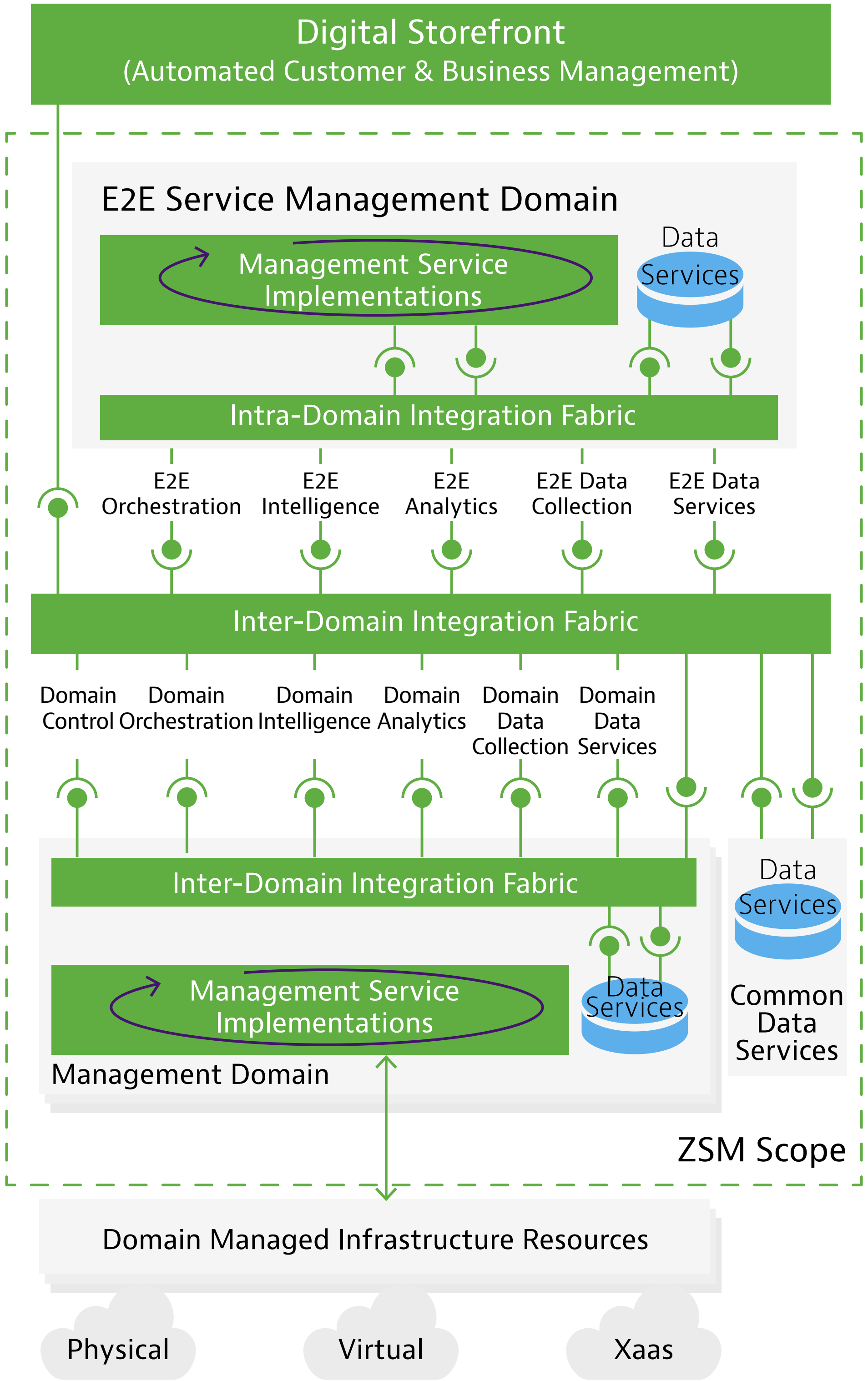


Figure 6.5 E2E ZSM framework according to the ETSI ZSM ISG

As previously noted in this chapter, it is possible to take a different approach to the concept of automation based on the current assets of an operator. For example, if a small or mid-sized operator has not been excessively process driven in their legacy operational model, but instead focused on building strong expertise and has a group of talented operations staff, then it is possible to automate current processes in an extreme manner. The example given earlier about the Finnish operator Elisa illustrates that by taking the right steps based on the individual strengths of the local organization, one can achieve significant results. The challenge is how larger, more siloed organizations will be able to succeed at this task. Many of the incumbent mobile operators that grew up via the GSM to 4G transition have a significant number of aging staff who are expected to retire over the next five to ten years. This, combined with the challenge to attract new competent staff members, forces operators to focus on "brutal automation," as Deutsche Telekom phrased it in October 2017. (Read about it in Light Reading: https://www.lightreading.com/automation/dt-brutal-automation-is-only-way-to-succeed/d/d-id/737111.)

As can be seen from the architecture proposed by ETSI or in the ONAP or OSM models, automation is based on various decision and control loops that seek to find an optimal state that is defined in one or the other policy. One of the challenges is to be able to define the policy and the states well enough. This has evolved to a new concept of intent-based management covered briefly in the earlier parts of this chapter. The intent-based model is yet a further level of separation of concerns and must be built on the trust that the underlying system is capable and able to manage inside its own domain, and that the problem statement at hand is possible to manage inside the domain.

At this time, artificial intelligence (AI) and machine learning are hyped to become the main driver of automation and zero-touch networks. But it is less important what the underlying system performing the tasks is as long as the operational efficiencies can be achieved, at an acceptable cost structure.

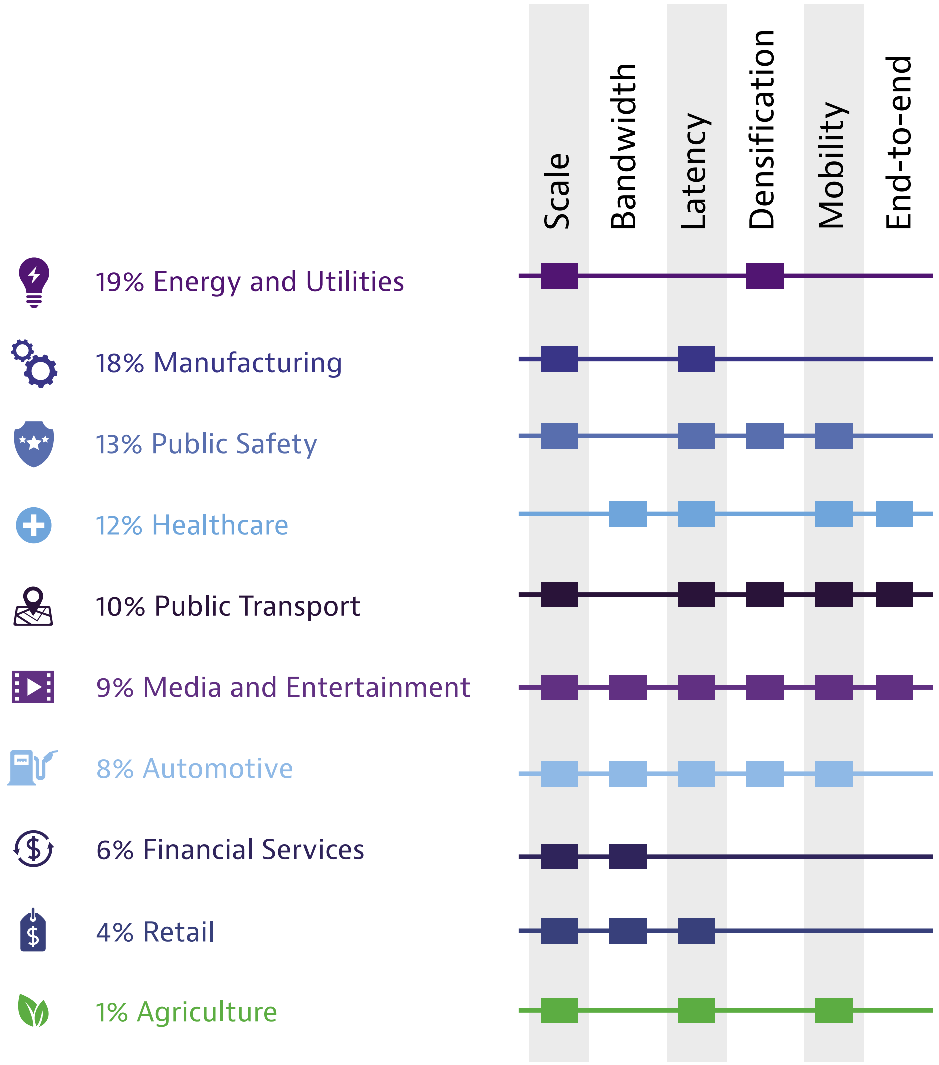
CHAPTER SEVEN

Cellular Internet of Things (CIoT)

Some may ask, what does the Internet of Things (IoT) have to do with 5G? It is true that IoT and machine-type communications (MTC) existed before 5G and cover wide varieties of communications technologies. But one of the main promises of 5G is that it is architected to deliver densification scale and low latency for many mission-critical IoT communications. The type of IoT that will benefit from 5G will be the cellular IoT (CIoT).

The opportunities provided by CIoT has excited businesses and consumers about the possibilities of what is to come. CIoT refers to machine-to-machine communications over a cellular network, and in most scenarios, happens without the involvement of humans. Why would we want machines talking to each other without our involvement? The main reason is efficiency. If a machine can communicate with another machine by itself, that saves us the time and trouble of gathering the information ourselves. Consider millions of devices communicating their data, and the efficiency and time savings becomes clear.

As an example, think of smart meters that keep track of the amount of energy a home uses each month. Millions of these meters are scattered throughout a country. Trying to get data from these meters usually means that someone must come to the home and read the meter each month. Some meters broadcast their readings wirelessly so that an employee of the power company can drive by the home and capture the reading, but it still takes a person driving past each home each month. With CIoT, a meter could simply send its reading to a main computer or application each month, saving valuable time and money. This is just one example. The opportunities seem unlimited; new ways of using CIoT are being discovered each day.



	Scale	Bandwidth	Latency	Densification	Mobility	End-to-end
19% Energy and Utilities	■			■		
18% Manufacturing	■		■			
13% Public Safety	■		■	■	■	
12% Healthcare		■	■		■	■
10% Public Transport	■		■	■	■	■
9% Media and Entertainment	■	■	■	■	■	■
8% Automotive	■	■	■	■	■	
6% Financial Services	■	■				
4% Retail	■	■	■			
1% Agriculture	■		■		■	

Figure 7.1 C-IoT different verticals

7.1 Business Drivers

For mobile operators, CIoT is a major business opportunity to dramatically increase revenue while also taking control of the entire CIoT ecosystem. Let's look at each of these in more detail.

Currently, most of a mobile operator's revenue comes from mobile handsets. This includes consumers using their smartphones and tablets, but also business-based handsets that are used for voice calls and data as employees work remotely conducting business. According to Ericsson [1], revenue growth in the handset market will remain relatively flat. Ericsson predicts that between 2016 and 2026, revenue from the handset market will increase only 1.5 percent to $1.736 billion USD. By adding CIoT, Ericsson thinks that growth could be 13.6 percent over the same period, reaching $3.458 billion USD.

The amount of predicted growth in the CIoT market has also changed significantly in the past two years. In 2017, 5G Americas predicted that CIoT connections would reach 1.4 billion by 2022. Just one year later, Ericsson predicted the number of CIoT connections would be closer to 3.5 billion during the same time period. The reasoning for the dramatic increase is twofold: 1) The mobile industry is finding new ways to leverage CIoT, and 2) China has become a major adopter of CIoT.

By 2023, the number of devices on mobile networks becomes staggering. According to Ericsson, the total number of mobile phones, fixed phones, PCs, laptops, and tablets on a mobile network will be 11.6 billion. By comparison, the total number of IoT-related devices on mobile networks will be 3.5 billion. As the mobile industry thinks of new ways to use CIoT in different vertical markets, these numbers will only increase. CIoT is not entirely new; it has been around for years for niche markets such as logistics and asset tracking in manufacturing where 2G networks are used due to their geographic reach. But as 5G is deployed worldwide, the latency of networks diminishes to a point that wide scale CIoT can flourish.

This opens opportunities to use CIoT for other verticals such as connected buildings for video surveillance and smoke alarms and connected industrial facilities for process and equipment monitoring. As 5G deployments become more widespread, further opportunities become available including smart cities, mobile health, and smart cars.

Some of the opportunities will be unique to each region of the world. For example, in Africa, mobile operators plan to track the location of specific wild animals using CIoT, such as elephants and lions which are major assets for tourism. Tourists would have a better chance of seeing these special animals. Animal tracking could also be used to prevent poaching. In this case, animals that traditionally live near an endangered species could be tracked. If that pack of animals start to behave in a way that is not normal, it could be a sign that poachers are in the area. Tracking the animals near the endangered species instead of the endangered animal itself gives authorities time to prevent the poaching from happening.

7.1.1 CIoT Ecosystem

As opportunities in new verticals become clear, mobile operators are seeing ways to take control of the entire ecosystem around each vertical. Instead of just providing the network or pipe as is done today in the handset market, operators could control the IoT device, the network, the platform that houses the data, and the vertical-specific application that captures and analyzes the data. In other words, mobile operators could become end-to-end solutions providers for CIoT. Ideally, mobile operators want to control the entire ecosystem for every vertical. This represents the greatest potential for revenue growth while keeping competitors at bay for certain aspects of a service. In reality, there will be too many verticals coming online at any given time to be a master at each vertical's unique needs. The ecosystem around CIoT will most likely come in several different variations.

In Europe, smart meters for utility companies pose a great opportunity for mobile operators to have control over the entire ecosystem since utility companies will most likely want to outsource the meters in the future. Utilities would hire the mobile operator to track and manage the devices and the communications, providing the utility with the data they need. This information would include the meter readings themselves but would also include other aspects such as if the meter is malfunctioning and communicating too often, or if the meter is not communicating at all.

The ability to determine what is happening with the meter and why it is happening will be important to the utility company, in addition to the actual data readings of the unit. This will enable the utility to more efficiently manage the lifecycle of the meters, helping to ensure the meters last as long as possible. Efficiently managing the meters means less truck rolls and maintenance to fix meters or change out batteries.

In other verticals, such as mobile health, it might be more advantageous for the mobile operator to partner with a provider that specializes in the healthcare field. Think of information such as real-time readings of a patient's vital signs that could be transmitted back to an application platform. The patient may be in a fixed location such as a hospital or might be an outpatient with health issues who is mobile. CIoT will allow the patient to wear a device that tracks vital signs and relays that information to the application platform no matter where the patient is located.

A doctor could then view those vital signs from their mobile phone and be notified if the vital signs become irregular. This would allow the doctor to call an ambulance for the patient and get to the hospital in time to treat the patient—saving valuable time and perhaps the patient's life.

Due to the critical nature of this type of vertical, a mobile provider might partner with a device maker that specializes in making CIoT devices which read patient vital signs and then partner with a company to interpret those readings. That application partner would be responsible for contacting the doctor in the example above. In this scenario, the mobile operator would provide the network and the platform for the application, but not the device or the application that sits on the platform.

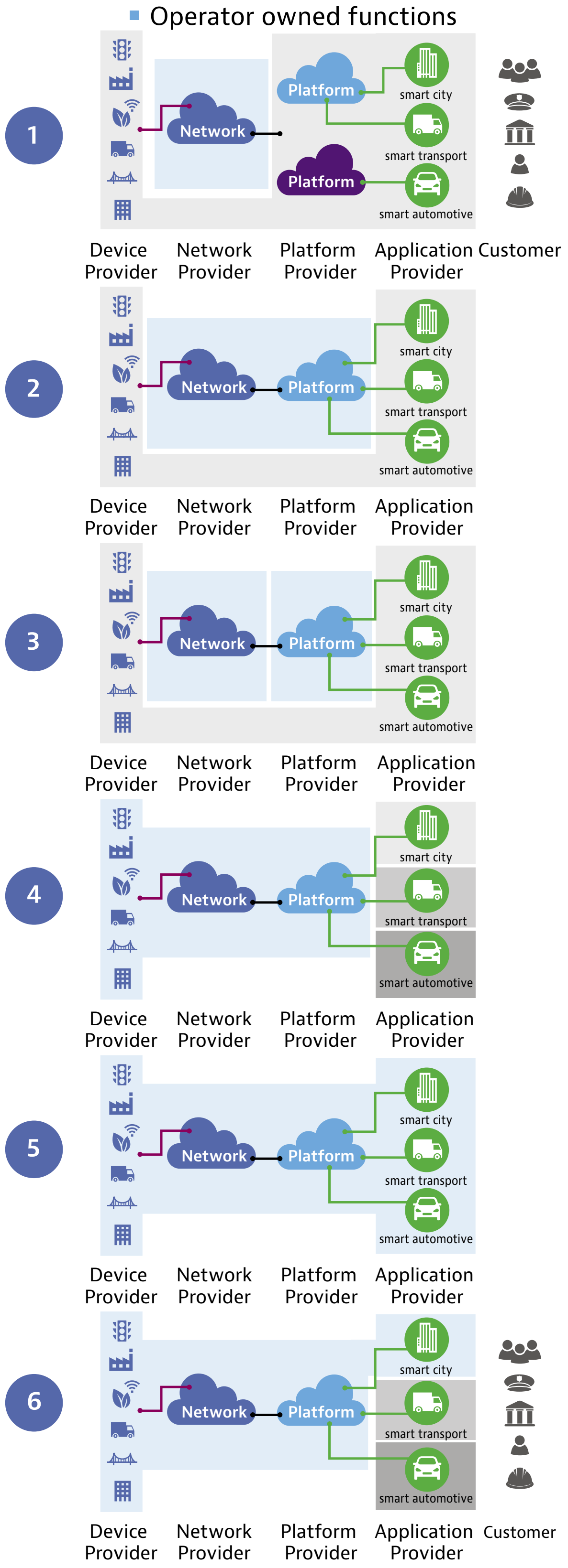


Figure 7.2 How to evolve to an end-to-end IoT service provider

For a setting such as smart cities, the mobile provider could partner with companies that make CIoT devices specifically for cities. This would include lighting, traffic signals, and train and bus management, among others. The mobile provider would then be responsible for the network, the application platform, and the application that captures and analyzes the data.

As an example of how this would work, think of a professional soccer game scheduled for a specific time. The lights to the stadium could be turned on automatically by computer, and the schedules for buses and trains could be adjusted to meet the surge in people attending the event. The timing of traffic lights would be adjusted to shorten the time drivers sit in traffic, and the police would be notified so they can send additional officers for crowd control.

All of this is known ahead of time, but what happens if the game ends early? This is where CIoT becomes invaluable. Monitors in the stadium would see that people are heading for the exits. This information would be transmitted back to the application that then determines the game ended early. The application would notify police, and once again change the timing of traffic lights and the schedules for mass transportation. If the stadium is close to additional night life, sidewalk lights could be turned on to assist pedestrians walking to get dinner. The stadium lights could be turned off as well. All of this would happen in real-time with machines talking to machines with very little human interaction. There is also the possibility for mobile operators to start with only the network for a certain vertical and then add to their ownership of the ecosystem as time goes on. As CIoT develops and continues to change over time, this type of opportunity will present itself. For example, asset tracking of shipping containers across land is already in place to a certain extent. But as CIoT advances, there could be an opportunity for mobile operators to take control of the CIoT device and the application. The operator would already have a robust platform for the application and a better understanding of what the industry needs in a tracking solution. At that point, it might be simpler and more cost effective for companies to turn the asset tracking over to the mobile operator versus trying to do it themselves or working with several different providers that must be managed.

7.1.2 Revenue Per Unit

The Revenue Per Unit (RPU) or revenue per CIoT device will also play an important role in how much of the ecosystem a mobile operator decides to take on. If the revenue is enough, it would make sense to take on more of the ecosystem. If the RPU is too low, scaling back the amount of the ecosystem an operator takes on might be in order.

The estimates for RPU range widely. 5G Americas [2] predicts the RPU will be $2.50 USD per device per month. Softbank estimates that the number is closer to $0.10 USD per month per device. AT&T has stated that the number could be as high as $10 USD per month.

To put this in context, let's take AT&T's estimate and compare it to the price of a typical 250 MB data plan. Based on market rates in the US and Europe as of this printing, the data plan would cost approximately $0.06 USD per Megabit. In contrast, the CIoT plan would cost approximately $2.00 USD per Megabit. In this case, the customer would choose the regular data plan over the CIoT plan, which would have negative ramifications on a network that will be discussed later in the chapter.

For CIoT to work, the RPU will need to be high enough to make it worthwhile for the mobile operator, but economical enough to make it useful to the end customer. How the RPU number works out will be different for each vertical because the needs of each vertical will be different. The number will also vary based on different regions of the world. In addition, the entire business model will need to be simplified to make it easy for customers to choose and use a plan. If the plans are too complicated, competitors will come in and capture key parts of the ecosystem and customers will look for alternatives.

7.2 Device Types

CIoT devices will need specialized chipsets in order to work efficiently on a network. Currently, two chipsets are gaining the most attention from the mobile industry. Both fit Release 13 governed by the 3rd Generation Partnership Project (3GPP) standards board. The first is Cat M1 and the second is Narrowband IoT (NB-IoT). Each has advantages depending on the network used and the vertical application.

Cat M1 is already available and works with the current network infrastructure. This makes it appealing to mobile operators that want to get a jump on IoT in order to quickly gain market share. The downside of Cat M1 is that it communicates with the network much like a handset does. For example, when a customer wants to download a video to a handset, the network must create an IP-based tunnel for the data transmission. The network then ensures that all the data packets of the video are sent and received properly. Once the video has been completely downloaded, the network must go through the process of breaking down that IP-based tunnel to make room for other communications from other devices.

For video, this process makes sense as it is designed to ensure the quality of that video download so the customer can enjoy the best viewing experience. For IoT devices that need to send only small amounts of data each month, the process of building up, transferring data, and breaking down an IP-based tunnel allocates important network bandwidth that is not fully utilized and reduces bandwidth for mobile applications. It may not sound like an issue until the number of tunnels being created for IoT devices is factored into the equation. For instance, if billions of IoT devices try to send data using this method, it could seriously impact mobile application quality of service.

As part of the migration towards 5G, an NB-IoT chipset will be available. NB-IoT is unique in that it can use the control plane of a network to send data instead of the user plane as discussed in the previous example. The control plane has traditionally been used to optimize the operation and performance of a network to make sure a network is available when customers need it.

The data sent by many IoT devices are so small at any given moment that transmitting them over the control plane makes sense. It creates less burden on a network and allows a network to handle many more data transmissions. But for NB-IoT to work, changes must be made to a network. Specifically, changes must be made to the Radio Access Network (RAN) with more substantial changes coming later to the core to increase the throughput and decrease latency. The RAN change will come in the form of a software upgrade but will take some time for operators to deploy. These changes are already planned as operators race to deploy 5G, but it means that IoT deployments will remain over Extended Coverage GSM for IoT (EC-GSM-IoT), LTE Cat 0, 1, and M1 in certain parts of the world.

Once the network changes are made, NB-IoT will be the prevailing chipset until a new chipset for 5G is developed. NB-IoT will still play a major role once 5G is deployed, as most mobile operators plan to keep their 4G EPC networks operating in a Non-Stand-Alone 5G mode (NSA). Some operators have indicated they will shut down their 3G networks and reallocate that spectrum to 5G.

EC-GSM-IoT will continue to play a small role in areas where a deeper network reach is needed and low-cost IoT solutions are required. This is because EC-GSM-IoT can use 2G to relay the data. Most mobile operators plan to keep their 2G networks running for this purpose.

7.3 Battery Life

One of the key technical considerations of CIoT is battery life. A majority of CIoT devices will need batteries to operate. This could be because the device isn't close to a major power source, or it would be too cost prohibitive to connect the device to the power source. Fire alarms are a good example where most alarms still operate using batteries for safety purposes even though they are in a building with a power source.

If a device uses too much power to operate, then IoT becomes too cost prohibitive on a mass scale. To combat this, mobile operators are looking at ways to use the least amount of battery power while still communicating the data needed for IoT. One solution is a low-power, wide-area network (LPWAN); LTE-M and NB-IoT are 3GPP's answer to this. To save battery life, the network reduces the amount of power needed to transmit data to and from an IoT device as well as the amount of communication to these devices.

Within 3GPP LPWAN are two innovations aimed at limiting the amount of times that a device checks in with the network. The less a device is active, the less battery it uses. This allows the devices to go for much longer periods without needing a battery replacement.

7.3.1 eDRX

The first innovation is extended Discontinuous Reception Cycle (eDRX). This decreases how often a device can be contacted by a network looking for commands and information.

A typical handset paging cycle is every 1.28 seconds, but eDRX slows that down to 5.12 seconds between pages and allows a device to sleep for 10.24 seconds between the paging cycles. However, this is not enough to extend the battery life to multiple years. eDRX therefore also allows the device to tell the network how many hyper frames of 10.24s it would like to sleep before checking back in. The hyper frame is of great importance to mobile operators because they will be able to decide how many hyper frames a device waits before paging a network.

For example, an IoT device may send data only once per day or once per month, but it still needs to be available to be paged to see if there are any new commands or data from the application. Mobile operators can set the number of hyper frames a device waits before paging a network. If that number were set to sixty hyper frames, then the device would wake up and page a network every ten minutes. At this rate, the device would use up two AA batteries in 4.7 years. Most IoT devices don't need to check in with a network often, so eDRX is a great way to save battery life.

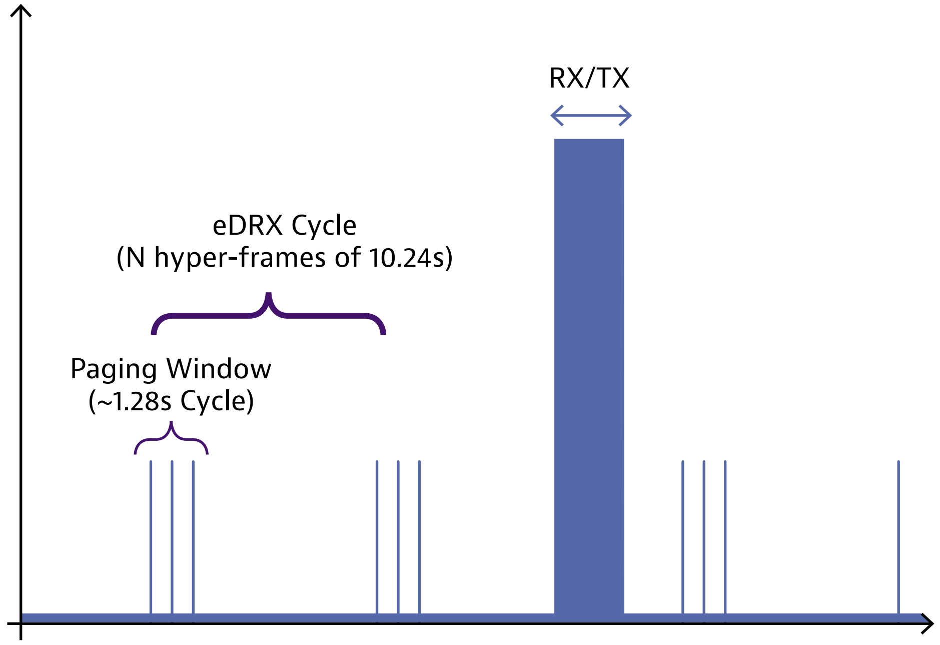


Figure 7.3 eDRX cycle

7.3.2 PSM

To extend battery life even further, Power Saving Mode (PSM) can be used. PSM tells the device to go completely dormant, either for a predetermined period of time, or until something happens such as a fire alarm detecting a fire. When the time period ends, the device wakes up, transmits to a network, and stays awake for a few seconds in case a network has any new commands or needs to reach the device. After that, the device becomes dormant again. If a device only wakes up once per day, two AA batteries could last ten years. This is a great alternative for IoT devices such as smart meters that need to transmit data only once per month and don't have a need to continuously page a network.

Softbank was one of the first companies to launch both Cat M1 and NB-IoT devices in April 2018 in Japan. The company is taking full advantage of power-saving innovations and has stated that it is using eDRX and PSM to improve device battery life.

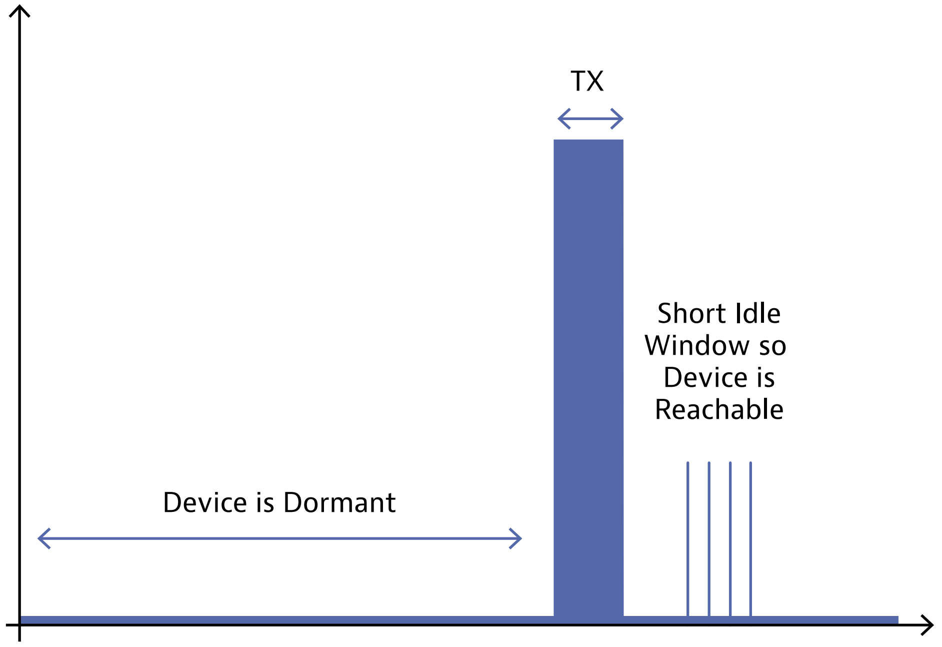


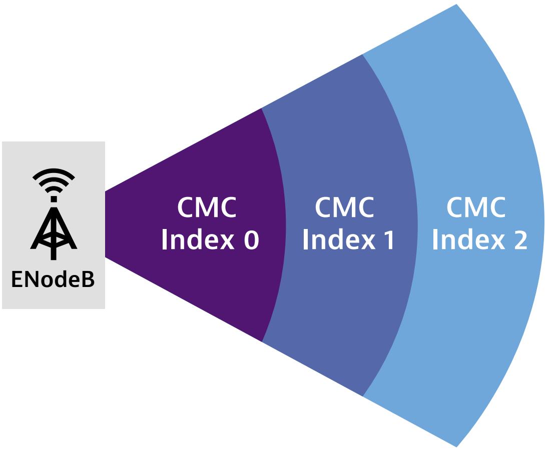
Figure 7.4 PSM cycle

7.3.3 Coverage Extension

Radio coverage is another key consideration for IoT services. Some IoT applications require devices to be positioned in areas that are not readily accessible by radio coverage, such as underground parking garages and in ground pits. The 3GPP Enhanced Coverage feature is an integral characteristic of NB-IoT, as it increases the depth of radio coverage to enable IoT devices to operate in locations that would otherwise not be possible, e.g. to provide network coverage for smart meters placed in basements.

The 3GPP Enhanced Coverage feature increases the power levels of signaling channels together with the ability to repeat transmissions. Repeated transmission improves the ability of receivers to correctly resolve the message sent. The trade-off is that repeating signal transmissions consumes additional power and the time between battery recharge or replacement may be reduced. 3GPP has defined three modes: EC0 for 0dB, EC1 for 10dB, and EC2 for 20dB for LTE. 5G RAN improves the utilization of current frequency bands (up to 6 GHz), and exploitation of new frequencies in the centimeter/millimeter wave bands (somewhere between 6 GHz and 100 GHz), to deliver the necessary capacity and coverage.

ManagedElement
+-ENodeBFunction
+-NblotCell
ceLevelNumber = 1 {1,2,3}
1 = CE_level1
2 = CE_level2
3 = CE_level3
cmcIndex = 0 {0,1,2}
0 = CMC Index 0
1 = CMC Index 1
2 = CMC Index 2



Channel (carrying)	Maximum Number of Repetitions		
	CMC Index 0	CMC Index 1	CMC Index 2
NPDSCH (NB-SIB1)	4	8	16
NPDCCH	2	16	64
NPUSCH (ACK/NACK)	2	16	64
NPRACH (preamble)	2	8	16

Figure 7.5 Coverage extension

7.4 CIoT Interfaces [3, 6]

CIoT introduces a number of new elements and interfaces into a mobile network to allow for efficient data communication between the IoT device and application.

The key interfaces involved in the communication between an NB-IoT device and an IoT application are as follows and are shown in the network diagram below:

LTE Interfaces

- S1-MME NAS enhancements allow for the communication of MO and MT between the UE to MME of encrypted small-data packets
- T6a/b between MME/SGSN and SCEF allows the SCEF
  - to receive reports of the monitoring events from the MME/SGSN configured via an HSS
  - to configure the monitoring events at an MME/SGSN that are not UE related in the non-roaming cases
  - to manage a connection between the MME and the SCEF on T6a
  - to send Mobile Terminated (MT) data on T6a from an NB-IoT device using non-IP Data Communication (NIDD)
  - to receive Mobile Originated (MO) data on T6a from an NB-IoT device using non-IP Data Communication (NIDD)

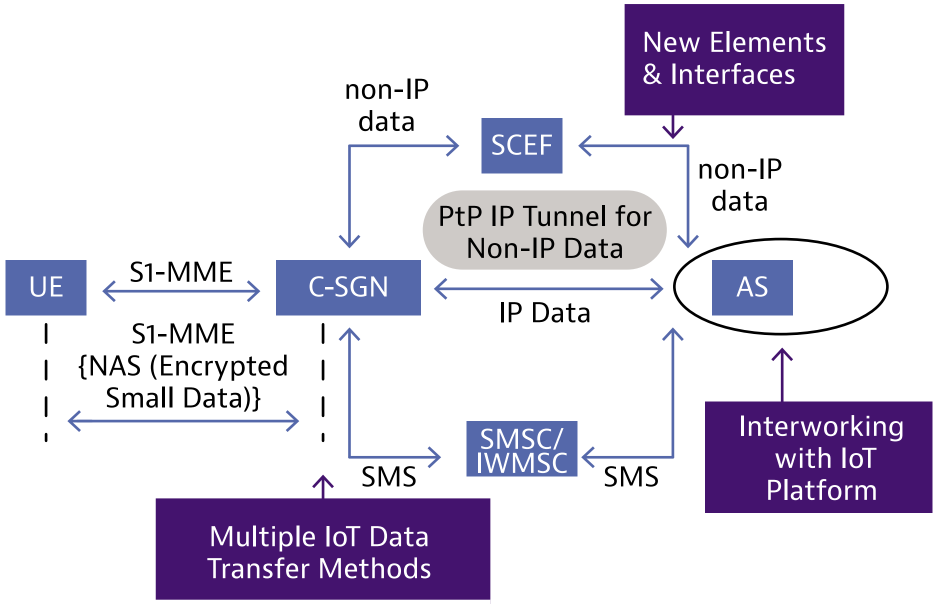

New Elements & Interfaces
non-IP data
SCEF
non-IP data
PtP IP Tunnel for Non-IP Data
UE
S1-MME
C-SGN
IP Data
AS
S1-MME {NAS (Encrypted Small Data)}
SMSC/ IWMSC
SMS
SMS
Interworking with IoT Platform
Multiple IoT Data Transfer Methods

Figure 7.6 C-IoT architecture enhancements

- SGd between MME and MSC for NB-IoT communication over SMS
- S6t between SCEF and HSS to configure monitoring events for a UE, communication patterns, and enhanced coverage restrictions.

5G Interfaces

- N1 NAS enhancements allow for the communication of MO and MT between the 5G UE to AMF of encrypted small-data packets
- N11 AMF to SMF to support transfer of decrypted small-data packets.
- N4 SMF to UPF to support CIoT EPS Optimization Data Transfer using IP-Based Tunnels in a 5G SA based network.
- SMF to NEF to support NIDD for delivering data via a PDU Sessions of type “Unstructured” in a 5G SA based network.

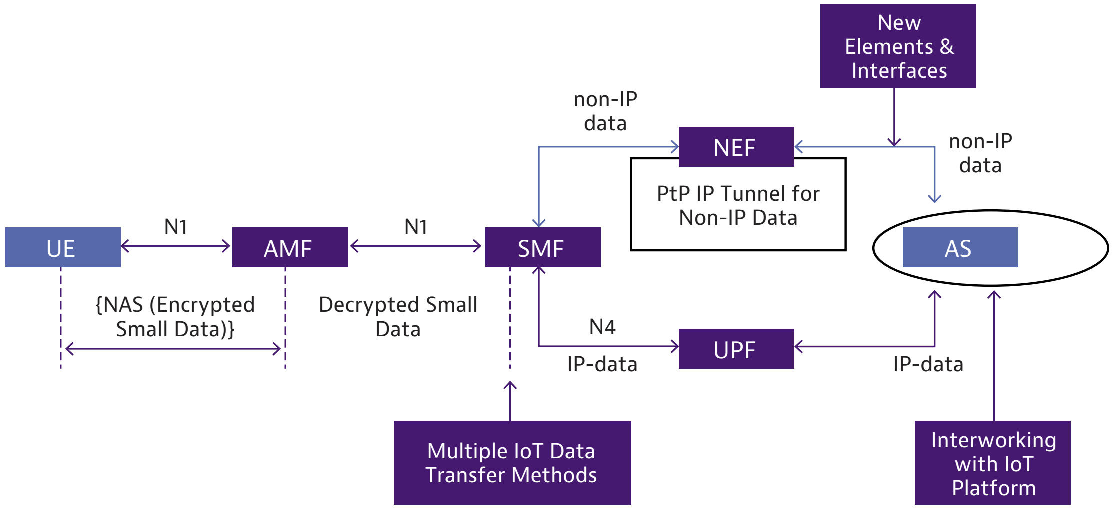

New Elements & Interfaces
non-IP data
NEF
non-IP data
PtP IP Tunnel for Non-IP Data
N1
N1
UE
AMF
SMF
AS
{NAS (Encrypted Small Data)}
Decrypted Small Data
N4
IP-data
UPF
IP-data
Multiple IoT Data Transfer Methods
Interworking with IoT Platform

Figure 7.7 C-IoT 5G architecture enhancements

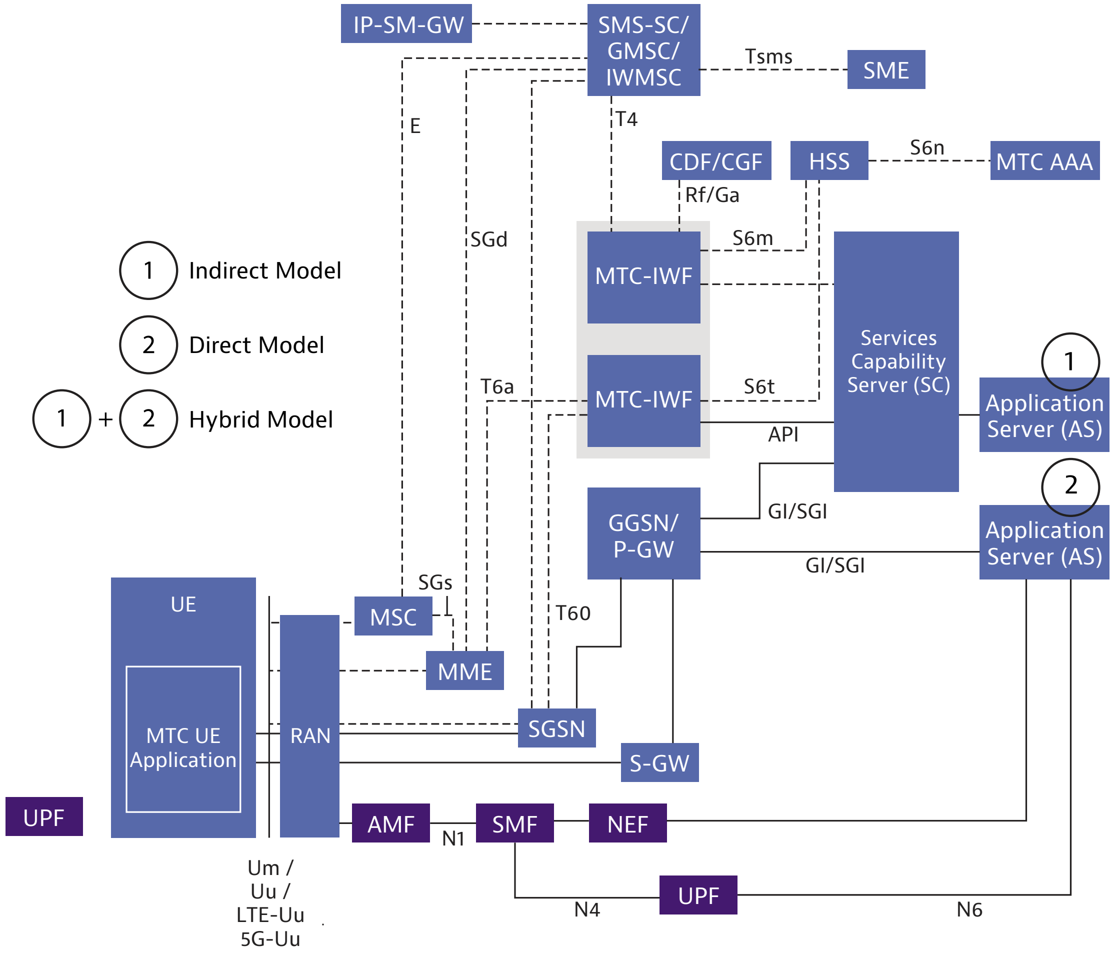


Figure 7.8 C-IoT architecture

7.5 CIoT Transfer Methods

The way in which an IoT device interfaces or communicates with a network will be different depending on the device being used and the situation for its use. This section provides a step-by-step explanation of the different methods by which devices communicate with a network. Before going any further, some clarification is needed around the terms IP-based tunnel and non-IP data. These terms are widely used in the mobile industry regarding IoT but can be misleading to those outside the industry or new to IoT.

All communications, no matter the IoT device, use an IP outer layer that acts as a wrapper around the transmission. This wrapper protects the communications and helps it move through a network. Think of it as a box moving through a delivery company’s logistics network to get to its destination. Inside the box is data going to its predetermined location.

Within the outer IP wrapper, data can travel within another inner IP layer or without this IP layer. The inner IP layer is called an IP-based tunnel since it uses a GTP tunnel to transmit the data. Cat M1 uses this method for the entire communications from one end to the other. This is also how data are transferred on a traditional handset such as a smartphone. This is a very reliable way to send data if the data are traveling over long distances or if there is a lot of data to be sent. The network builds a tunnel specifically for the data, sends the data, makes sure it was received in its entirety at the destination, and then breaks down the tunnel. This method ensures the quality of the communication but creates a burden on a network if many devices are communicating at the same time.

If data are traveling over shorter distances or there is not much data to send, the data are still wrapped in the outer IP wrapper, but may not use an IP-based tunnel. Instead, the data are sent over the control plane of a network. NB-IoT can work in this way so as to not overburden the resources of a network. When this method of transmission is used, it is called non-IP data delivery (NIDD).

7.5.1 Cat M1

Cat M1 uses an IP-based tunnel for all communications, regardless whether the device has a lot of information to send or not. It can be used on current LTE networks without modifications, making it an appealing choice for some operators. As networks advance, Cat M1 will focus on applications where the IoT devices are spread over vast geographic areas such as rural environments. The network frequencies for 5G and 4G LTE may not reach these locations, so 2G is used. Networks using 2G can reach farther geographically but can only send data using IP-based tunnels.

7.5.2 Cat M1 CIoT Using S1-U Data Transfer or User Plane CIoT EPC Optimization [3]

To see how this type of communications happens in a network, let's re-examine the smart cities example from earlier where a soccer game ended early and the crowd began moving toward the exits. The surveillance cameras saw the fans moving toward the exits and notified the application. The application would want to continue monitoring the flow of traffic as it makes other changes such as adjusting the timing of the traffic lights and the schedules for mass transit. In this case, the surveillance camera (IoT device) originated the transmission and wants to send a live feed of what it sees. An IP-based tunnel is more appropriate since there is a lot of data in the form of video.

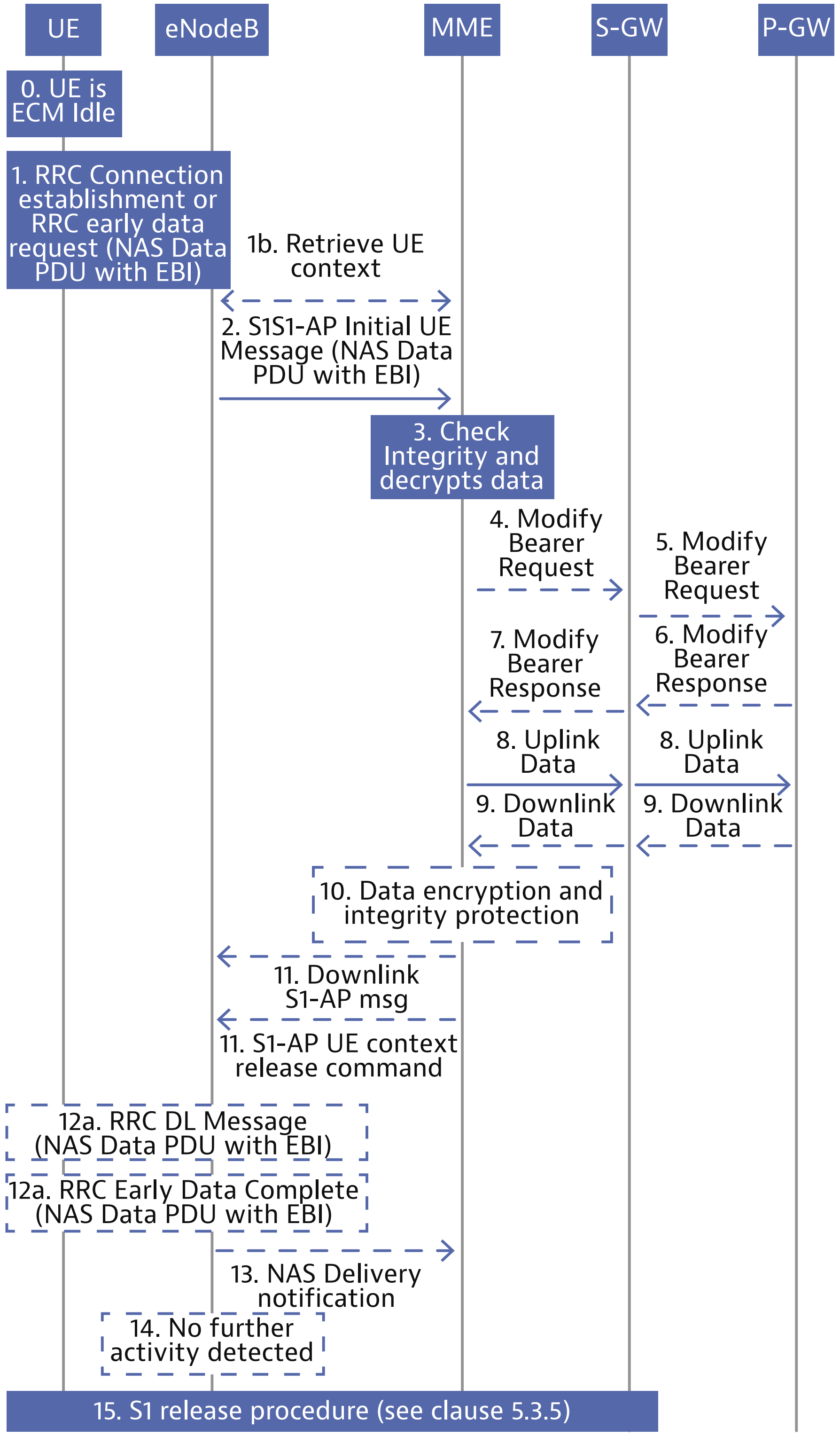

UE
eNodeB
MME
S-GW
P-GW
0. UE is ECM Idle
1. RRC Connection establishment or RRC early data request (NAS Data PDU with EBI)
1b. Retrieve UE context
2. S1S1-AP Initial UE Message (NAS Data PDU with EBI)
3. Check Integrity and decrypts data
4. Modify Bearer Request
5. Modify Bearer Request
7. Modify Bearer Response
6. Modify Bearer Response
8. Uplink Data
8. Uplink Data
9. Downlink Data
9. Downlink Data
10. Data encryption and integrity protection
11. Downlink S1-AP msg
11. S1-AP UE context release command
12a. RRC DL Message (NAS Data PDU with EBI)
12a. RRC Early Data Complete (NAS Data PDU with EBI)
13. NAS Delivery notification
14. No further activity detected
15. S1 release procedure (see clause 5.3.5)

Figure 7.9 Establishment of S1-U bearer during data Transport in Control Plane CIoT EPS optimization

In the example above, the security camera initially communicates images using CIoT EPS optimization data transfer using Non-Access Stratum (NAS) Network data Protocol Data Units (PDUs) (Step 1 below). However, as the crowd starts to leave the stadium, the application decides it needs more detailed information and requests a live video feed. In order to initiate the larger data transfer, the surveillance camera determines it needs to transfer to S1-U data transfer or User Plane CIoT EPC Optimization and initiates the procedure below (Steps 2 and 3).

This procedure is used where either the User Interface (UE) or Mobility Management Entity (MME) determines that S1-U-based data transfer, IP-based tunnel, from the UE to the P-GW is preferred due to the size of data to be transferred in uplink (UL) or downlink (DL), such as the video transfer as in the example above.

The MME receives the NAS message Service Request with active flag via the eNodeB as defined in Steps 2 and 3 from the UE (surveillance camera) and triggers the establishment of S1-U bearer(s). The MME checks if the UE can support establishment of the required number of additional user plane radio bearers based on the maximum number of user plane radio bearers indicated by UE in the UE Network Capability IE. The eNodeB then establishes the radio bearers to transfer the data to the UE. Uplink data can now be forwarded from the UE to the Packet Data Network Gateway (P-GW) and the application.

Step 1: The UE is sending and receiving data in NAS PDUs using the Control Plane CIoT EPS Optimization.

Step 2: The UE triggers the establishment of user plane bearers and sends a Control Plane Service Request with an active flag toward the MME encapsulated in an RRC message to the eNodeB.

Step 3: The eNodeB forwards the Control Plane Service Request with active flag to the MME. The NAS message is encapsulated in an S1-AP UL NAS Transport Message (NAS message, TAI+ECGI of the serving cell, S-TMSI, Closed Subscriber Group (CSG) ID, CSG access Mode). If the MME receives the Control Plane Service Request with active flag defined in Steps 2 and 3, it shall establish S1-U bearer(s) and execute the transfer.

Step 4: The MME sends any remaining UL data over S11-U, and in order to minimize the possibility of out-of-order DL data that could be caused by earlier DL data sent on the Control Plane, the MME may send a Release Access Bearers Request message to the Serving GW that requests the release of all S11-U bearers for the UE.

Step 5: If the Serving GW receives the Release Access Bearers Request message it releases all MME related information (address and downlink TEIDs) for the UE and responds with a Release Access Bearers Response message to the MME.

Step 6: The MME sends an S1-AP Initial Context Setup Request (Serving GW address, S1-TEID(s) (UL), EPS Bearer QoS(s), Security Context, MME Signaling Connection Id, Handover Restriction List, CSG Membership Indication, Service Accept) message to the eNodeB for all PDN connections where the MME has not included a Control Plane Only Indicator in the ESM request. The MME responds to the UE with a Service Accept message. The eNodeB stores the Security Context, MME Signaling Connection Id, EPS Bearer QoS(s), and S1-TEID(s) in the UE RAN context.

Step 7: The eNodeB performs the radio bearer establishment procedure. The user plane security is established at this step and the user plane radio bearers are set up.

Step 8: As the user plane radio bearers are set up, the UE uses user plane bearers to transfer data PDUs, except for EPS bearers where the MME has included Control Plane Only Indicator in the ESM request and for which Control Plane CIoT EPS Optimization is still to be used. The uplink data from the UE can now be forwarded by eNodeB to the Serving GW. The eNodeB sends the uplink data to the Serving GW address and TEID provided in Step 6. The Serving GW forwards the uplink data to the PDN GW.

Step 9: The eNodeB sends an S1-AP message Initial Context Setup Complete (eNodeB address, List of accepted EPS bearers, List of rejected EPS bearers, S1 TEID(s) (DL)) to the MME.

Step 10: The MME sends a Modify Bearer Request message (eNodeB address, S1 TEID(s) (DL) for the accepted EPS bearers, Delay Downlink Packet Notification Request, RAT Type) per PDN connection to the Serving GW.

Step 11: The Serving GW shall return a Modify Bearer Response (Serving GW address and TEID for uplink traffic) to the MME as a response to a Modify Bearer Request message, or a Modify Access Bearers Response (Serving GW address and TEID for uplink traffic) as a response to a Modify Access Bearers Request message.

Table of Contents

7.5.3 NB-IoT Using CIoT EPS Optimization Data Transfer Using NAS PDU and IP-Based Tunnels [3, 5]

Instead of using IP-based tunnels through the entire communications process as Cat M1 does, an NB-IoT device can send its data through the control plane of a network from the device to the MME or AMF at the edge of the core of a network. From there, the data can be sent using an IP-based tunnel or it can continue to use the control plane. This mechanism is much more resource efficient for small data transfers. There are several scenarios NB-IoT might want to send a transmission using NAS PDUs. Let's look at two different examples—one where the device initiates the communication and another where the application reaches out to the IoT device.

7.5.3.1 Mobile Originated

For this example, let's use an alarm system in a building. A breach has been detected and the alarm system needs to notify the IoT platform.

The IoT device, or UE, establishes an RRC connection with the closest network node. The UE sends its unique ID letting a network know which device it is and that it is authorized to use that network. The UE then encrypts and sends its data to the node in an integrity-protected NAS PDU containing the encrypted Uplink Data. In certain situations, the NB-IoT device might be configured to communicate through the node to the MME or AMF in the core of a network to retrieve a Quality of Service (QoS) profile. This level of service is agreed upon between the customer of the IoT service and the mobile operator.

For LTE, upon receiving the NAS Data PDU, the eNodeB forwards to the MME, which checks the integrity of the data and decrypts the data before forwarding the message over an S11-U connection to the P-GW. For 5G, upon receiving the NAS Data PDU, the gNodeB forwards to the AMF which in turn forwards the PDU to the SMF, which checks the integrity of the data and decrypts the data before forwarding the message over N4 connection to the UPF.

7.5.3.1.1 LTE Mobile Originated

The UE is ECM-IDLE.

Step 1: The UE establishes an RRC connection and sends as part of it an integrity protected NAS PDU. The NAS PDU carries the EPS Bearer ID and encrypted Uplink Data. The UE may also indicate in a Release Assistance Information within the NAS PDU if no further Uplink or Downlink Data transmissions are expected, or only a single Downlink Data transmission (e.g. Acknowledgement or response to Uplink Data) subsequent to this Uplink Data transmission is expected.

1b: In the NB-IoT case, the eNodeB, based on configuration, may retrieve the EPS negotiated QoS profile from the MME, if not previously retrieved.

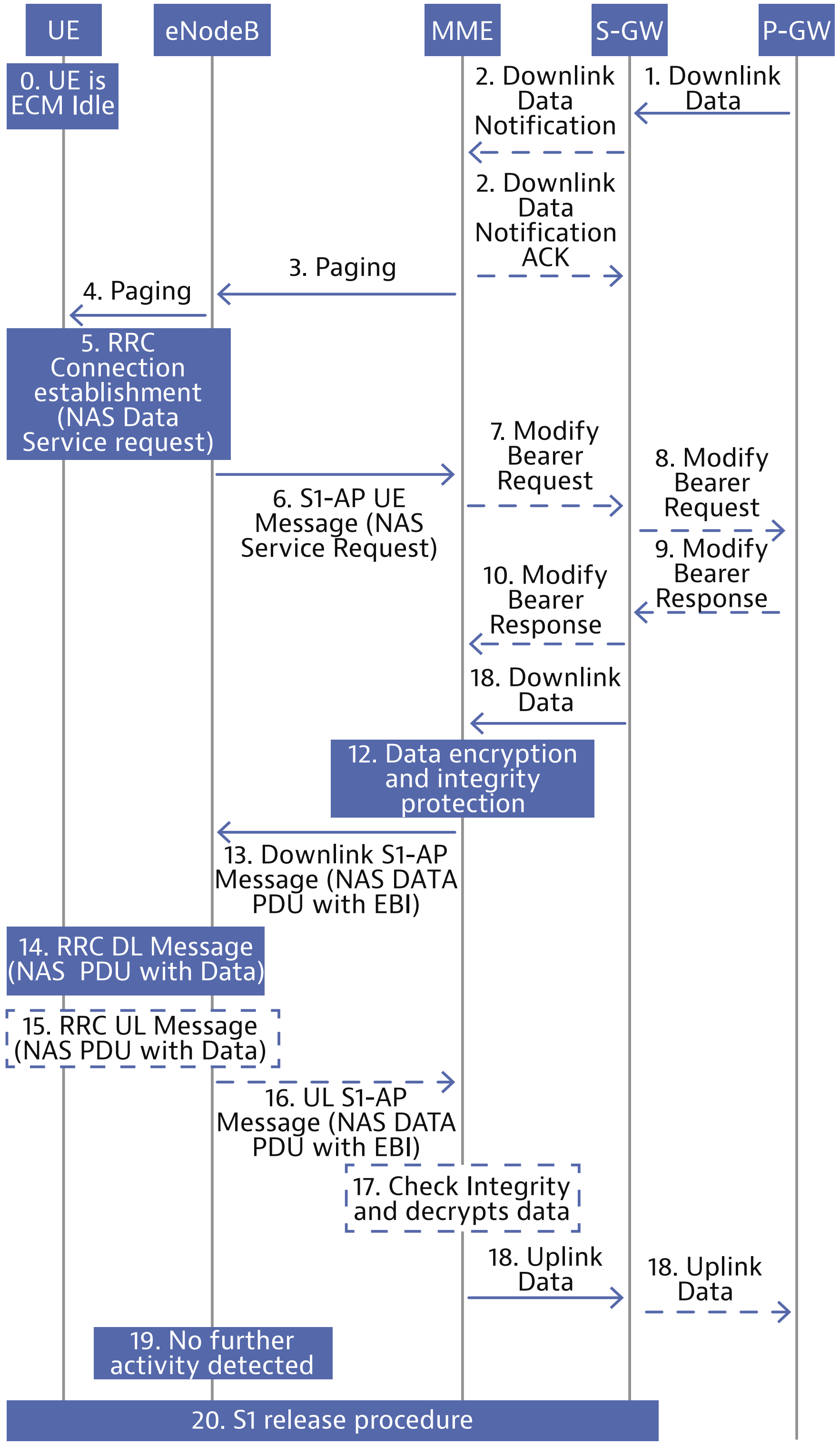

UE
eNodeB
MME
S-GW
P-GW
0. UE is ECM Idle
2. Downlink Data Notification
1. Downlink Data
2. Downlink Data Notification ACK
3. Paging
4. Paging
5. RRC Connection establishment (NAS Data Service request)
7. Modify Bearer Request
8. Modify Bearer Request
6. S1-AP UE Message (NAS Service Request)
9. Modify Bearer Response
10. Modify Bearer Response
18. Downlink Data
12. Data encryption and integrity protection
13. Downlink S1-AP Message (NAS DATA PDU with EBI)
14. RRC DL Message (NAS PDU with Data)
15. RRC UL Message (NAS PDU with Data)
16. UL S1-AP Message (NAS DATA PDU with EBI)
17. Check Integrity and decrypts data
18. Uplink Data
18. Uplink Data
19. No further activity detected
20. S1 release procedure

Figure 7.10 Mobile originated data transport in Control Plane CIoT EPS optimization

(NAS PDUs) with P-GW connectivity

Step 2: The NAS PDU sent in Step 1 is relayed to the MME by the eNodeB using an S1-AP Initial UE message.

Step 3: The MME checks the integrity of the incoming NAS PDU and decrypts the data it contains.

Step 4: If the S11-U connection is not established, the MME sends a Modify Bearer Request message (MME address, MME TEID DL, Delay Downlink Packet Notification Request, RAT Type, LTE-M RAT type reporting to P-GW flag, MO Exception data counter) for each PDN connection to the Serving GW. The Serving GW is now able to transmit Downlink Data towards the UE.

Step 5: If the RAT Type has changed compared to the last reported RAT Type or if the UE's Location and/or Info IEs and/or UE Time Zone and Serving Network ID are present in Step 4, the Serving GW sends the Modify Bearer Request message (RAT Type, MO Exception data counter) to the PDN GW (Steps 6 and 7).

Step 8: The MME sends Uplink Data to the P-GW via the S-GW.

Step 9: If no Downlink Data are expected based on the Release Assistance Information from the UE in Step 1, this means that all application layer data exchanges have completed with the UL data transfer, and if the MME is not aware of pending MT traffic and S1-U bearers are not established, Step 10 is skipped and Step 11 applies. Otherwise, Downlink Data may arrive at the P-GW and the P-GW sends them to the MME via the S-GW.

Step 10: If Downlink Data are received in Step 9, the MME encrypts and integrity protects the Downlink Data.

Step 11: If Step 10 is executed, then Downlink Data are encapsulated in an NAS PDU and sent to the eNodeB in a S1-AP Downlink NAS message.

Step 12a: The eNodeB sends an RRC Downlink Data message including the Downlink Data encapsulated in NAS PDU.

Step 12b: If End Indication with no further data is received in S1-AP message from the MME, the eNodeB may send the RRC Early Data Complete message with any NAS payload received from Step 11 (either NAS data PDU or NAS service accept). Step 14 is skipped in this case.

Step 13: The eNodeB sends an NAS Delivery indication to the MME if requested.

Step 14: If no NAS PDU activity exists for a while, the eNodeB starts an S1 release in Step 15.

Step 15: An S1 release procedure according to clause 5.3.5 is triggered by the eNodeB or MME.

7.5.3.1.2 5G Mobile Originated

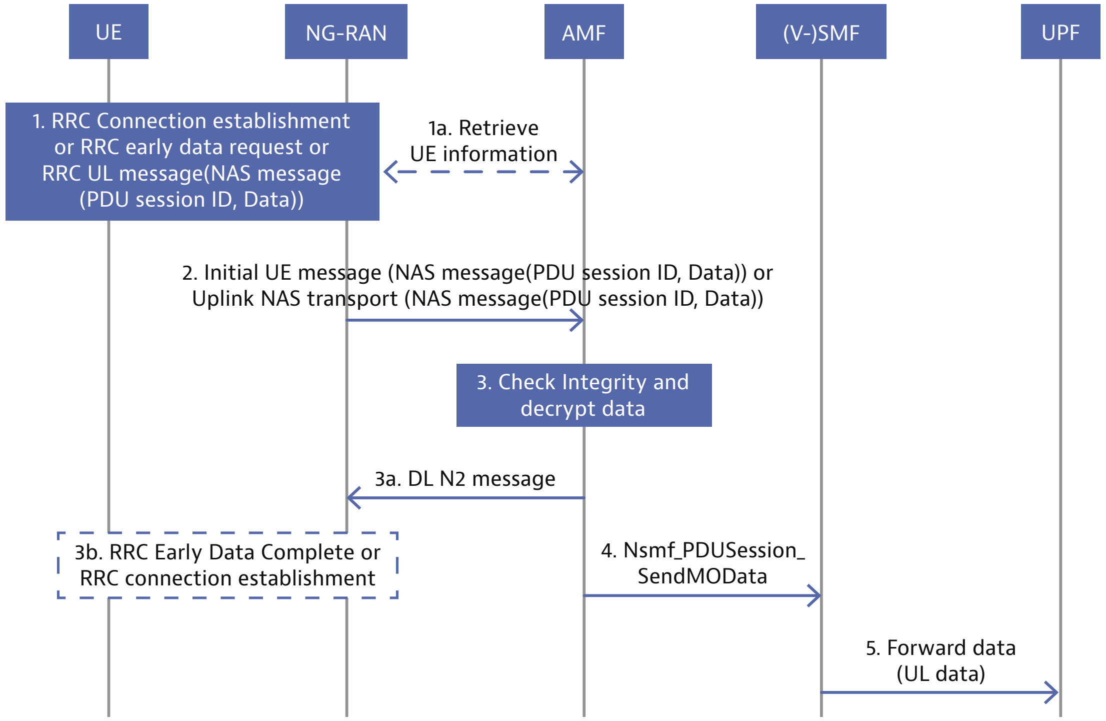

UE
NG-RAN
AMF
(V-)SMF
UPF
1. RRC Connection establishment or RRC early data request or RRC UL message(NAS message (PDU session ID, Data))
1a. Retrieve UE information
2. Initial UE message (NAS message(PDU session ID, Data)) or Uplink NAS transport (NAS message(PDU session ID, Data))
3. Check Integrity and decrypt data
3a. DL N2 message
3b. RRC Early Data Complete or RRC connection establishment
4. Nsmf_PDUSession_ SendMOData
5. Forward data (UL data)

Figure 7.11 UPF anchored Mobile Originated Data Transport in Control Plane CIoT 5GS Optimization

1. If the UE is CM-CONNECTED it sends a NAS message carrying the ciphered PDU session ID and ciphered uplink data as payload. If the UE is in CM-IDLE, the UE first establishes an RRC connection or sends the RRCEarlyDataRequest message and sends a NAS message as part of this.

1a. In the NB-IoT case, during step 1 the NG-RAN, based on configuration, may retrieve the NB-IoT UE Priority and the Expected UE Behavior Parameters from the AMF, if not previously retrieved. Based on such parameters, the NG-RAN may apply prioritization between requests from different UEs before triggering step 2 and throughout the RRC connection. The NG-RAN may retrieve additional parameters (e.g. UE Radio Capabilities).

2. NG-RAN forwards the NAS message to the AMF using the Initial NAS message procedure (if the UE was in CM-IDLE before step 1) or using the Uplink NAS transport procedure (if the UE was in CM-CONNECTED before step 1). If RRCEarlyDataRequest message was received in step 1, the NG-RAN includes "EDT Session" indication in the N2 Initial UE message.

3. AMF checks the integrity of the incoming NAS message and deciphers the PDU session ID and uplink data.

3a. If the AMF received "EDT Session" indication from the NG-RAN in step 2, the AMF sends an N2 message to the NG-RAN.

3b. If 3a was executed, the NG-RAN completes the RRC early data procedure

4. AMF determines the (V-)SMF handling the PDU session based on the PDU session ID contained in the NAS message and passes the PDU Session ID and the data to the (V-) SMF by invoking Nsmf_PDUSession_SendMOData service operation.

5. The (V-)SMF decompresses the header if header compression applies to the PDU session and forwards the data to the UPF. In the home-routed roaming case, the (V-) SMF forwards the data to V-UPF then to the H-UPF. The UPF forwards the data to the DN based on data forwarding rule, e.g., in case of unstructured data, tunneling may be applied.

7.5.3.2 Mobile Terminated

In the last example, the IoT device reached out to the application. In this example, the application reaches out to the IoT device. Mobile health is a good example where a patient's life is dependent on the IoT device. A patient has a life-threatening problem and goes to the nearest emergency room in a rural area. The facility can't help the patient due to lack of resources, so the patient is transferred to a larger hospital in the nearest city. A wearable IoT device is placed on the patient to track vital signs in real time during travel from one hospital to the other.

The application notifies the doctor at the destination hospital that the patient is coming, but the doctor also needs to see the patient's vital signs on his or her mobile device. The application reaches out to the wearable IoT device and tells it to start sending vital signs data.

7.5.3.2.1 LTE Mobile Terminated

The UE is Evolved Packet System (EPS) attached and in ECM-Idle mode.

Step 1: The application sends a message through the P-GW to the S-GW that it needs to reach the patient's wearable IoT device. The S-GW buffers that information and identifies which MME serves that UE.

Step 2: The S-GW then sends a message to the corresponding MME that it needs to reach the UE. The MME sends back a message that it has received the request. Here is where it gets interesting. If the S11-U is established, meaning that all the important information to properly create an IP-based tunnel between the MME and the P-GW is established and is the same as the last time the tunnel was created, then the process jumps to Step 11, and the IP-based tunnel is established. The S-GW sends its Downlink Data (the message to start sending the patient's vitals) to the MME while the MME tries to contact the UE or IoT device.

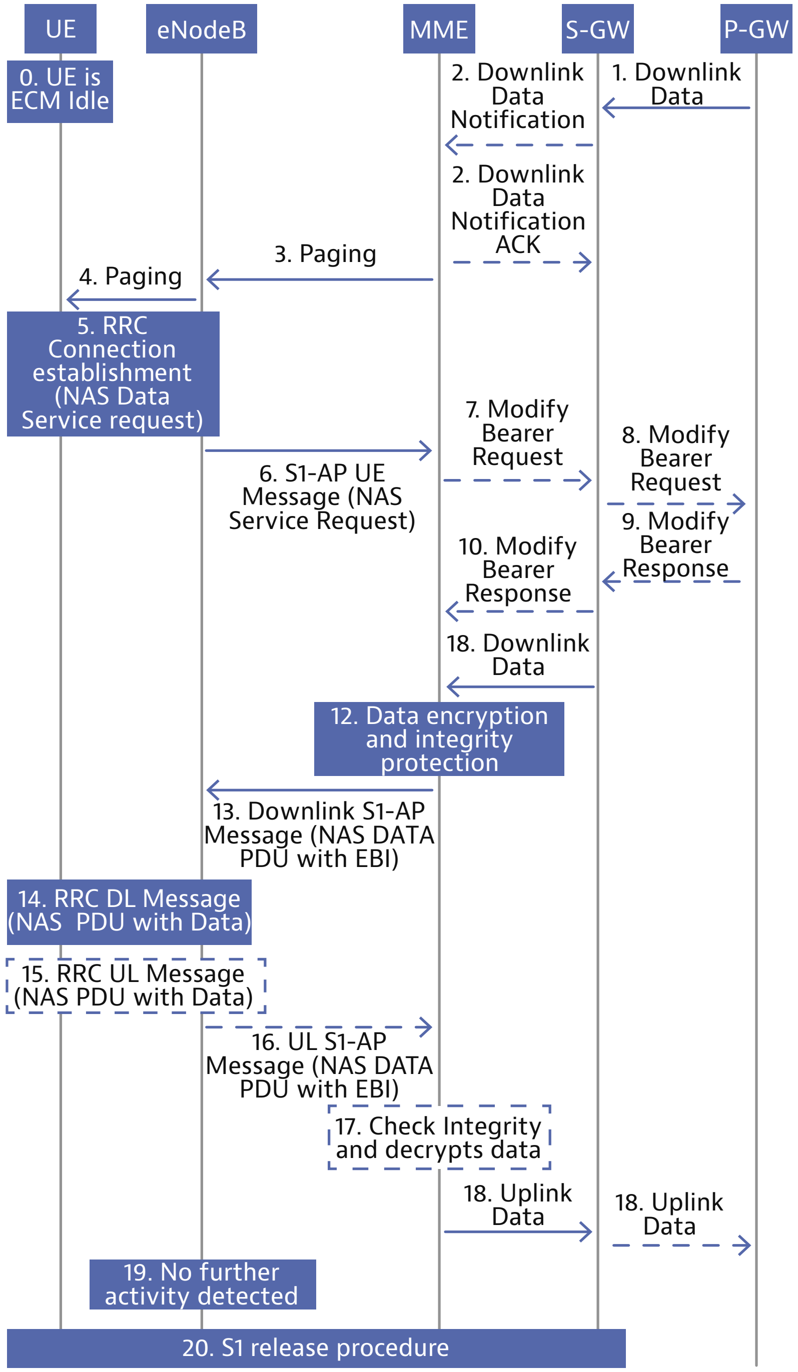

UE
eNodeB
MME
S-GW
P-GW
0. UE is ECM Idle
2. Downlink Data Notification
1. Downlink Data
2. Downlink Data Notification ACK
3. Paging
4. Paging
5. RRC Connection establishment (NAS Data Service request)
7. Modify Bearer Request
8. Modify Bearer Request
6. S1-AP UE Message (NAS Service Request)
9. Modify Bearer Response
10. Modify Bearer Response
18. Downlink Data
12. Data encryption and integrity protection
13. Downlink S1-AP Message (NAS DATA PDU with EBI)
14. RRC DL Message (NAS PDU with Data)
15. RRC UL Message (NAS PDU with Data)
16. UL S1-AP Message (NAS DATA PDU with EBI)
17. Check Integrity and decrypts data
18. Uplink Data
18. Uplink Data
19. No further activity detected
20. S1 release procedure

Figure 7.12 Mobile terminated data transport in Control Plane CIoT EPS optimization (NAS PDUs) with P-GW connectivity

Step 3: The MME pages the node nearest the UE requesting contact with the UE. If the UE is registered in the MME and considered reachable, the MME sends a Paging message (NAS ID for paging, TAI(s), UE identity based DRX index, Paging DRX length, list of CSG IDs for paging, Paging Priority indication) to each eNodeB belonging to the tracking area(s) in which the UE is registered.

Step 4: The node nearest the UE pages the UE. If eNodeBs receive paging messages from the MME, the UE is paged by the eNodeBs.

Step 5: This step through Step 10 are very similar to the process followed in Steps 1 through 7 in the previous mobile-originated example. The UE is saying that it is available and that the proper address and routing information is being updated by the system so that the UE and the application can communicate. This confirms the route the data will take through a network.

Step 6: If the MME receives no response from the UE after this paging repetition procedure, it uses the Downlink Data Notification Reject message to notify the S-GW about the paging failure. The MME performs (and the UE responds to) any EMM or ESM procedures if necessary, such as security-related procedures. Steps 7 to 11 can continue in parallel to this, however, Steps 12 and 13 must await completion of all the EMM and ESM procedures.

Step 7: If the S11-U is not established, the MME sends a Modify Bearer Request message (MME address, MME TEID DL, Delay Downlink Packet Notification Request, RAT Type) for each PDN connection to the S-GW. The S-GW is now able to transmit Downlink Data towards the UE via the usage of the Delay Downlink Packet.

Step 8: If the RAT Type has changed compared to the last reported RAT Type or if the UE's Location and/or Info IEs and/or UE Time Zone and Serving Network ID are present in Step 7, the S-GW shall send the Modify Bearer Request message (RAT Type) to the P-GW. User Location Information IE and/ or User CSG Information IE and/or Serving Network IE and/or UE Time Zone are also included if they are present in Step 7.

Step 9: The P-GW sends the Modify Bearer Response to the S-GW.

Step 10: If a Modify Bearer Request message was sent at Step 7, the S-GW shall return a Modify Bearer Response (S-GW address and TEID for uplink traffic) to the MME as a response to a Modify Bearer Request message. The S-GW address for S11-U User Plane and S-GW TEID are used by the MME to forward UL data to the S-GW.

Step 11: Once the critical information about the route and the identity of all the players is correct and the route has been established that the data will take through the network, the S-GW sends the Downlink Data to the MME.

Steps 12-13: The MME encrypts the data, adds integrity protection, and sends it as an S1-AP message to the node. The MME encrypts and integrity-protects Downlink Data and sends it to the eNodeB using a NAS PDU carried by a Downlink S1-AP message. If the eNodeB supports acknowledgements of Downlink NAS Data PDUs and if acknowledgements of Downlink NAS Data PDUs are enabled in the subscription information for the UE, the MME indicates in the Downlink S1-AP message that acknowledgment is requested from the eNodeB.

Step 14: The node sends the message on to the UE. The NAS PDU with data is delivered to the UE via a Downlink RRC message. This is taken by the UE as implicit acknowledgment of the Service Request message sent in Step 5. If header compression was applied to the PDN, the UE shall perform header decompression to rebuild the IP header.

Step 15: The node sends a delivery notification to the MME. The eNodeB sends an NAS Delivery indication to the MME if requested. If the eNodeB reports an unsuccessful delivery with an S1-AP NAS Non-Delivery Indication, the MME should wait for some time until the UE has potentially changed cell and re-established contact with the MME, by which MME should resend the Downlink S1-AP message to the eNodeB, otherwise the MME reports an unsuccessful delivery.

Step 16: At this point, the patient's wearable IoT device has received the message to start sending the patient's vital signs. For this example, the RRC connections would stay up since the vital signs need to be continuously sent. Data are sent from the UE to the node. While the RRC connection is still up, further Uplink and Downlink Data can be transferred using NAS PDUs. In Step 17 an Uplink Data transfer is shown using an Uplink RRC message encapsulating an NAS PDU with data. At any time that the UE has no user plane bearers established, the UE may provide a Release Assistance Information with Uplink Data in the NAS PDU.

Step 17: The patient's vital signs are sent in an Uplink S1-AP message to the MME. The NAS PDU with data is sent to the MME in an Uplink S1-AP message.

Step 18: The MME decrypts the data and checks the integrity. If header compression was applied to the PDN, the MME shall perform header decompression to rebuild the IP header.

Step 19: The MME sends the vital signs data through the S-GW to the P-GW. The vital signs would be in the application at this point and the doctor of the destination hospital would be able to view the vital signs through his or her mobile device. Once the patient reaches the destination hospital, the wearable IoT device may be turned off. This would lead to Step 20. The MME sends Uplink Data to the P-GW via the S-GW and executes any action related to the presence of Release Assistance Information as follows:

Step 20: If no NAS activity exists for a while the eNodeB detects inactivity and executes Step 21.

Step 21: The eNodeB starts an eNodeB-initiated S1 release procedure according to clause 5.3.5 or Connection Suspend Procedure defined in clause 5.3.4A. The UE and the MME store the Robust Header Compression (RoHC) configuration and context for the Uplink/Downlink Data transmission when entering ECM_CONNECTED state next time.

7.5.3.2.2 5G Mobile Terminated

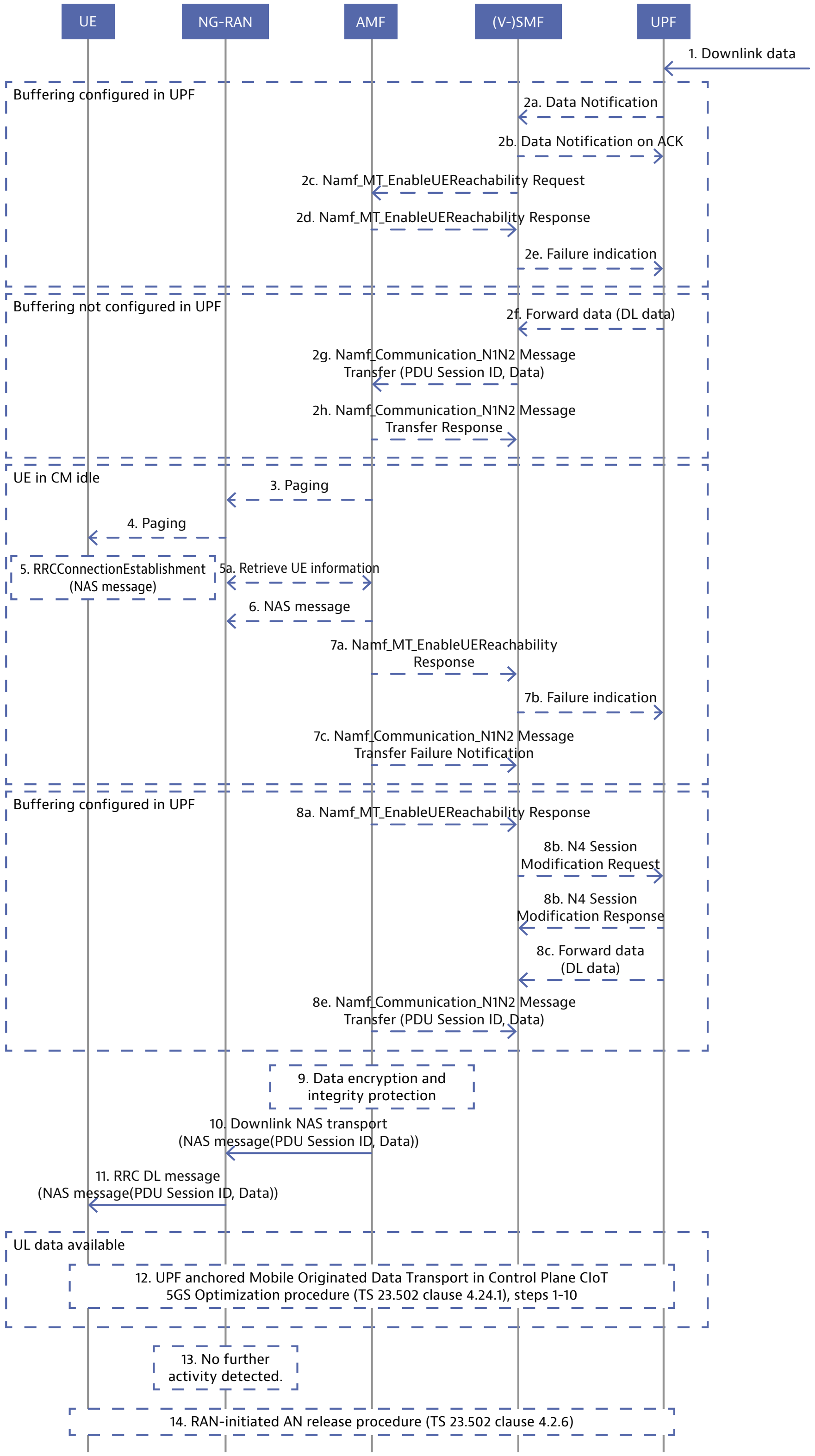

UE
NG-RAN
AMF
(V-)SMF
UPF
1. Downlink data
Buffering configured in UPF
2a. Data Notification
2b. Data Notification on ACK
2c. Namf_MT_EnableUEReachability Request
2d. Namf_MT_EnableUEReachability Response
2e. Failure indication
Buffering not configured in UPF
2f. Forward data (DL data)
2g. Namf_Communication_N1N2 Message Transfer (PDU Session ID, Data)
2h. Namf_Communication_N1N2 Message Transfer Response
UE in CM idle
3. Paging
4. Paging
5. RRCConnectionEstablishment (NAS message)
5a. Retrieve UE information
6. NAS message
7a. Namf_MT_EnableUEReachability Response
7b. Failure indication
7c. Namf_Communication_N1N2 Message Transfer Failure Notification
Buffering configured in UPF
8a. Namf_MT_EnableUEReachability Response
8b. N4 Session Modification Request
8b. N4 Session Modification Response
8c. Forward data (DL data)
8e. Namf_Communication_N1N2 Message Transfer (PDU Session ID, Data)
9. Data encryption and integrity protection
10. Downlink NAS transport (NAS message(PDU Session ID, Data))
11. RRC DL message (NAS message(PDU Session ID, Data))
UL data available
12. UPF anchored Mobile Originated Data Transport in Control Plane CIoT 5GS Optimization procedure (TS 23.502 clause 4.24.1), steps 1-10
13. No further activity detected.
14. RAN-initiated AN release procedure (TS 23.502 clause 4.2.6)

Figure 7.13: Mobile Terminated Data Transport in Control Plane CIoT 5GS Optimization

1. Downlink data is received by the UPF. If buffering is configured in the UPF, then the flow continues in step 2a, otherwise the flow continues in step 2f.

2a. [Conditional] If buffering is configured in the UPF, then the UPF sends a Data Notification to the SMF.

2b. [Conditional] The SMF sends a Data Notification ACK to the UPF.

2c. [Conditional] The SMF sends a Namf_MT_EnableUEReachability request to the AMF.

2d. [Conditional] If AMF determines the UE is unreachable (e.g., if the UE is in MICO mode or the UE is configured for extended idle mode DRX), then the AMF rejects the request from the SMF with an indication that the UE is not reachable. If the SMF included Extended Buffering support indication, the AMF indicates the Estimated Maximum Wait time in the response message. If the UE is in MICO mode, the AMF determines the Estimated Maximum Wait time based on the next expected periodic registration timer update expiration or by implementation. If the UE is configured for extended idle mode DRX, the AMF determines the Estimated Maximum Wait time based on the start of next Paging Time Window. The AMF stores an indication that the SMF has been informed that the UE is unreachable.

2e. [Conditional] If the SMF receives an "Estimated Maximum Wait time" from the AMF and Extended buffering applies, the SMF sends a failure indication to the UPF with an Extended Buffering time and optionally a DL Buffering Suggested Packet Count. The Extended Buffering time is determined by the SMF and should be larger or equal to the Estimated Maximum Wait time received from the AMF. The DL Buffering Suggested Packet Count parameter is determined by the SMF and, if available, the Suggested Number of Downlink Packets parameter may be considered. The SMF may also indicate to the UPF to stop sending Data Notifications. The procedure stops after this step.

2f. [Conditional] If buffering is not configured in the UPF, then the UPF forwards the downlink data to the (V)SMF in non-roaming and LBO cases. In the home-routed roaming case, the H-UPF forwards the data to the V-UPF and then to the V-SMF.

2g. [Conditional] The SMF determines whether Extended Buffering applies based on local policy and the capability of the SMF.

If user data is received in step 2f and Extended buffering is not configured for the SMF, then (V-)SMF compresses the header if header compression applies to the PDU session and creates the downlink user data PDU that is intended as payload in a NAS message. The (V-)SMF forwards the downlink user data PDU and the PDU session ID to the AMF using the Namf_Communication_N1N2MessageTransfer service operation. If Extended Buffering applies, then (V-SMF) keeps a copy of the downlink data.

If user data is received in step 2f and Extended Buffering applies, the SMF includes "Extended Buffering support" indication in Namf_Communication_N1N2Message Transfer.

2h. [Conditional] AMF responds to SMF.

If AMF determines that the UE is reachable for the SMF, then the AMF informs the SMF. Based on this, the SMF deletes the copy of the downlink data.

If AMF determines the UE is unreachable for the SMF (e.g., if the UE is in MICO mode or the UE is configured for extended idle mode DRX), then the AMF rejects the request from the SMF. The AMF may include in the reject message an indication that the SMF need not trigger the Namf_Communication_N1N2MessageTransfer Request to the AMF, if the SMF has not subscribed to the event of the UE reachability.

If the SMF included Extended Buffering support indication, the AMF indicates the Estimated Maximum Wait time, in the reject message, for the SMF to determine the Extended Buffering time. If the UE is in MICO mode, the AMF determines the Estimated Maximum Wait time based on the next expected periodic registration timer update expiration or by implementation. If the UE is configured for extended idle mode DRX, the AMF determines the Estimated Maximum Wait time based on the start of next PagingTime Window. The AMF stores an indication that the SMF has been informed that the UE is unreachable.

If the SMF receives an "Estimated Maximum Wait time" from the AMF and Extended Buffering applies, the SMF store the DL Data for an Extended Buffering time. The Extended Buffering time is determined by the SMF and should be larger or equal to the Estimated Maximum Wait time received from the AMF. The SMF does not send any additional Namf_Communication_N1N2MessageTransfer message if subsequent downlink data packets are received.

3. [Conditional] If the UE is in CM Idle, the AMF sends a paging message to NG-RAN.

4. [Conditional] If NG-RAN received a paging message from AMF, NG-RAN performs paging.

5. [Conditional] If the UE receives paging message, it responds with a NAS message sent over RRC Connection Establishment.

5a. [Conditional] In the NB-IoT case, during Step 5, the NG-RAN, based on configuration, may retrieve the NB-IoT UE Priority and the Expected UE Behavior Parameters from the AMF, if not previously retrieved. Based on such parameters, the NG-RAN may apply prioritization between requests from different UEs before triggering step 6 and throughout the RRC connection. The NG-RAN may retrieve additional parameters (e.g., UE Radio Capabilities).

6. [Conditional] The NAS message is forwarded to the AMF.

7a. [Conditional] AMF to SMF: Namf_MT-EnableUEReachability Response.

If the SMF used the MT_EnableUEReachability request in step 2c and the UE has not responded to paging, then the AMF sends a response to the SMF indicating that the request failed.

7b. [Conditional] SMF to UPF: If the SMF has received a Namf_MT-EnableUEReachability response from the AMF indicating that the request failed, the SMF indicates to the UPF to discard the buffered data and the procedure stops after this step.

7c. [Conditional] AMF to SMF: Namf_Communication_N1N2Transfer Failure Notification.

If the SMF used the Namf_Communication_N1N2MessageTransfer service operation in step 2g and the UE has not responded to paging, the AMF sends a failure notification to the SMF based on which the SMF discards the buffered data. The procedure stops after this step.

8a. [Conditional] AMF to SMF: Namf_MT-EnableUEReachability Response.

If the SMF used the MT_EnableUEReachability request in step 2c, then the AMF indicates to the SMF that the UE is reachable.

8b. [Conditional] SMF to UPF: N4 Session Modification Request.

If the SMF received an indication from the AMF that the UE is reachable, then the SMF indicates to the UPF to deliver buffered data to the SMF.

8c. [Conditional] UPF to SMF: N4 Session Modification Response.

8d. [Conditional] Buffered data is delivered to the SMF.

8e. [Conditional] (V-)SMF compresses the header if header compression applies to the PDU session and creates the downlink user data PDU that is intended as payload in a NAS message. The (V-)SMF forwards the downlink user data PDU and the PDU session ID to the AMF using the Namf_Communication_N1N2MessageTransfer service operation.

9. The AMF creates a DL NAS transport message with the PDU session ID and the downlink user data PDU received from the SMF. The AMF ciphers and integrity protects the NAS transport message.
10. The AMF sends the DL NAS transport message to NG-RAN.
11. NG-RAN delivers the NAS payload over RRC to the UE.
12. While the RRC connection is established further uplink and downlink data can be exchanged. In order to send uplink data, the procedure continues as per steps 1-10 of the UPF anchored Mobile Originated Data Transport in Control Plane CIoT 5GS Optimization procedure.

13. [Conditional] If no further activity is detected by NG-RAN, then NG-RAN triggers the AN release procedure.
14. [Conditional] The UE's logical NG-AP signaling connection and RRC signaling connection are released.

7.5.4 NB-IoT Using Non-IP Data Communication (NIDD) [4, 5]

This is where NB-IoT devices excel. Many of the IoT applications will be devices that need to send small amounts of data infrequently. For this type of communications, NB-IoT devices can send data all the way to the application using the control plane of a network instead of setting up an IP-based tunnel. The number of steps involved to send that data is reduced compared to setting up and breaking down an IP-based tunnel, reducing the need for network resources.

Also, by eliminating data such as header information required in IP-based data communications, the electric power needed for communication is reduced and, in addition to extending battery life, a broader area can be covered. NIDD also enables users to transmit data to IoT devices without allocating an IP address. By not using an Internet protocol in transmission, the risk of being subjected to a malicious attack targeting an IoT device is low, making it possible to build a highly secure network. Just as before, there will be two different examples—one where the device initiates the communication and another where the application reaches out to the IoT device.

7.5.4.1 Mobile Originated

Smart meters are a great example of an IoT device that stays dormant for long periods of time, wakes up, sends its utilization data, waits for a response from the application, and then goes dormant again. For this example, the smart meter is sending data on the amount of electricity a building has used over the past month. Setting up GTP tunnels for these small packets of data is inefficient both in the RAN and the CORE, therefore the non-IP based method of communication is ideal for these use cases.

7.5.4.1.1 LTE Mobile Originated Data Transport

Step 1: The IoT device (UE) wakes up and transmits the electricity usage data along with its International Mobile Subscriber Identity (IMSI) to the MME.

Step 2: The MME forwards that data to the Service Capability Exposure Function (SCEF). The MME also sends along its own identity, known as an EPS bearer identity, so that the network knows that it is authorized to use the network and knows how to send information back to the UE.

If the UE is located in an area that is only served by roaming for that mobile operator, then the MME of the visited network (VPLMN) sends the data to the Interworking SCEF (IWK-SCEF) of that visited network, which then forwards the data to the SCEF of the home network after figuring out routing and how to charge the home network for the service.

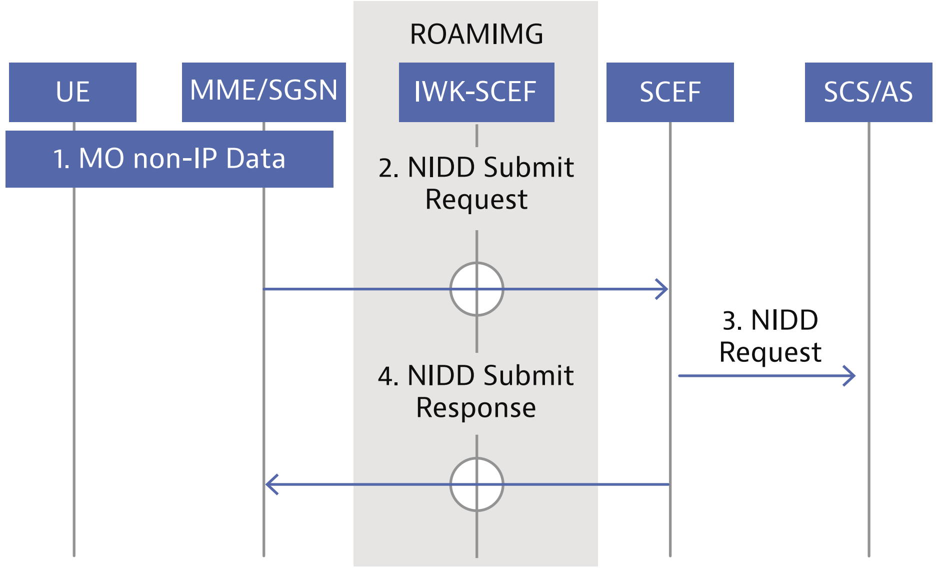


Figure 7.14 Mobile originated NIDD procedure

Step 3: The SCEF then forwards the electricity usage data to the application server that serves that IoT device.

7.5.4.1.2 5G NEF Anchored Mobile Originated Data Transport

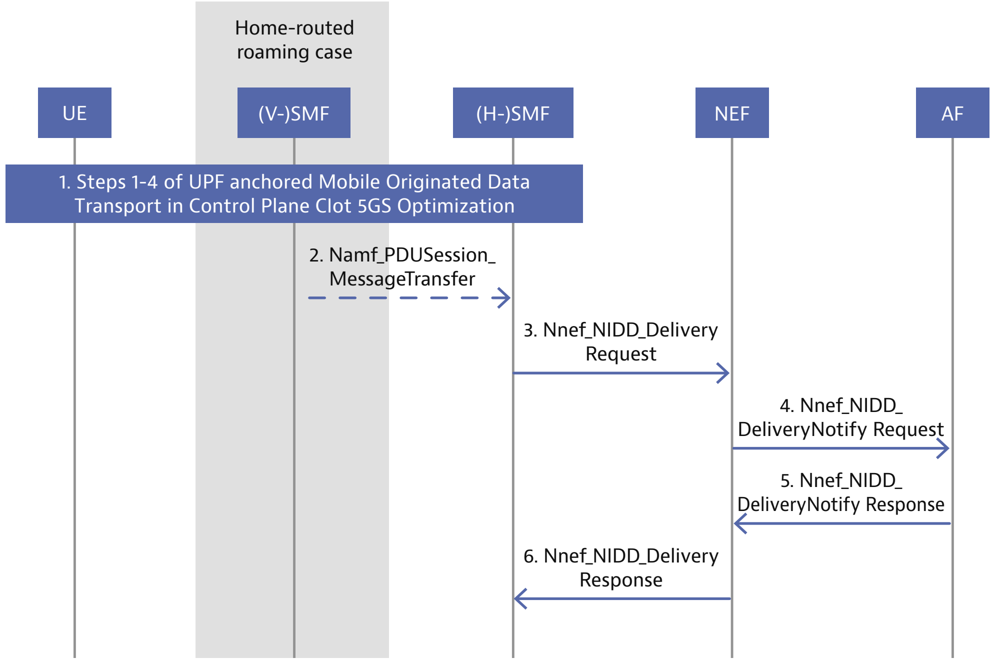

Home-routed roaming case
UE
(V-)SMF
(H-)SMF
NEF
AF
1. Steps 1-4 of UPF anchored Mobile Originated Data Transport in Control Plane Clot 5GS Optimization
2. Namf_PDUSession_ MessageTransfer
3. Nnef_NIDD_Delivery Request
4. Nnef_NIDD_ DeliveryNotify Request
5. Nnef_NIDD_ DeliveryNotify Response
6. Nnef_NIDD_Delivery Response

Figure 7.15 5G Mobile originated NIDD procedure

1. The IoT device (UE) wakes up and transmits the electricity usage data along with its International Mobile Subscriber Identity (IMSI) in a NAS message with unstructured data to the SMF.
2. In case of home-routed roaming the (V-)SMF forwards the data to the H-SMF.
3. The (H-)SMF sends the Nnef_NIDD_Delivery Request (User Identity, unstructured data) message to the NEF.
4. When the NEF receives the unstructured data and finds an NEF PDU Session context and the related T8 Destination Address, then it sends the electricity usage as unstructured data to the AF that serves that IoT device, identified by the T8 Destination address in a Nnef_NIDD_DeliveryNotify Request (GPSI, unstructured data, Reliable Data Service Configuration).
5. The AF responds to the NEF with a Nnef_NIDD_DeliveryNotify Response (Cause).
6. The NEF sends Nnef_NIDD_Delivery Response to the SMF. If the NEF cannot deliver the data, e.g. due to missing AF configuration, the NEF sends an appropriate error code to the SMF.

7.5.4.2 Mobile Terminated

The application may want to reach out to the IoT device using non-IP communications in situations where the application needs to know that the device is still there and functioning properly. An alarm system is a good example. The application sends non-IP data to the alarm system asking for the system to send back information on whether the system is functioning properly.

7.5.4.2.1 LTE Terminated Data Transfer

Step 1: The application sends the data request to the SCEF.

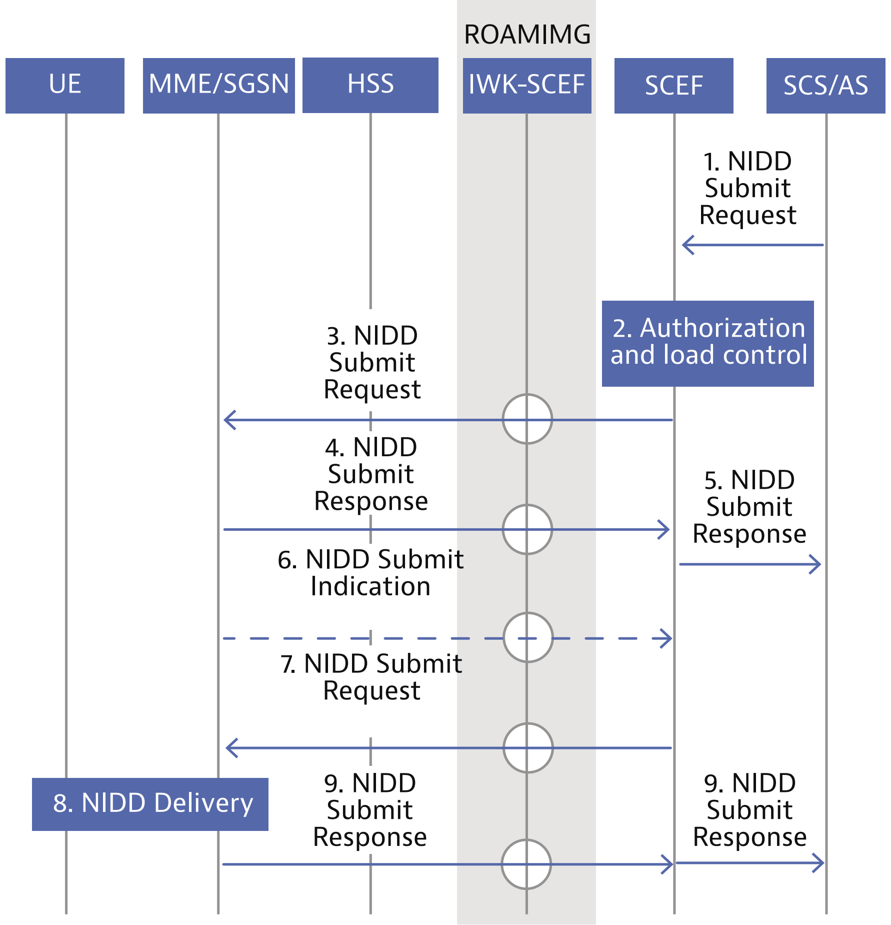


Figure 7.16 Mobile terminated NIDD procedure

Steps 2-7: The SCEF finds the MME that serves the IoT device (UE). It can do this based on the EPS Bearer Identity the MME sent on its last communication. The data is then sent to the MME.

If the IoT device can only be accessed through a visited network in the case of roaming, the SCEF sends the data to the IWK-SCEF in the visited network that then sends the data to the MME in the visited network. The same UE identity and EPS Bearer Identity information applies.

Steps 8-9: The MME checks for the UE's identity and whether there is a valid EPS Bearer Context for the UE. If the answer is yes for both, the data is sent to the UE. If the UE successfully receives the data, the MME sends an acknowledgement back to the SCEF.

7.5.4.2.2 5G NEF Anchored Mobile Terminated Data Transport Mobile Terminated

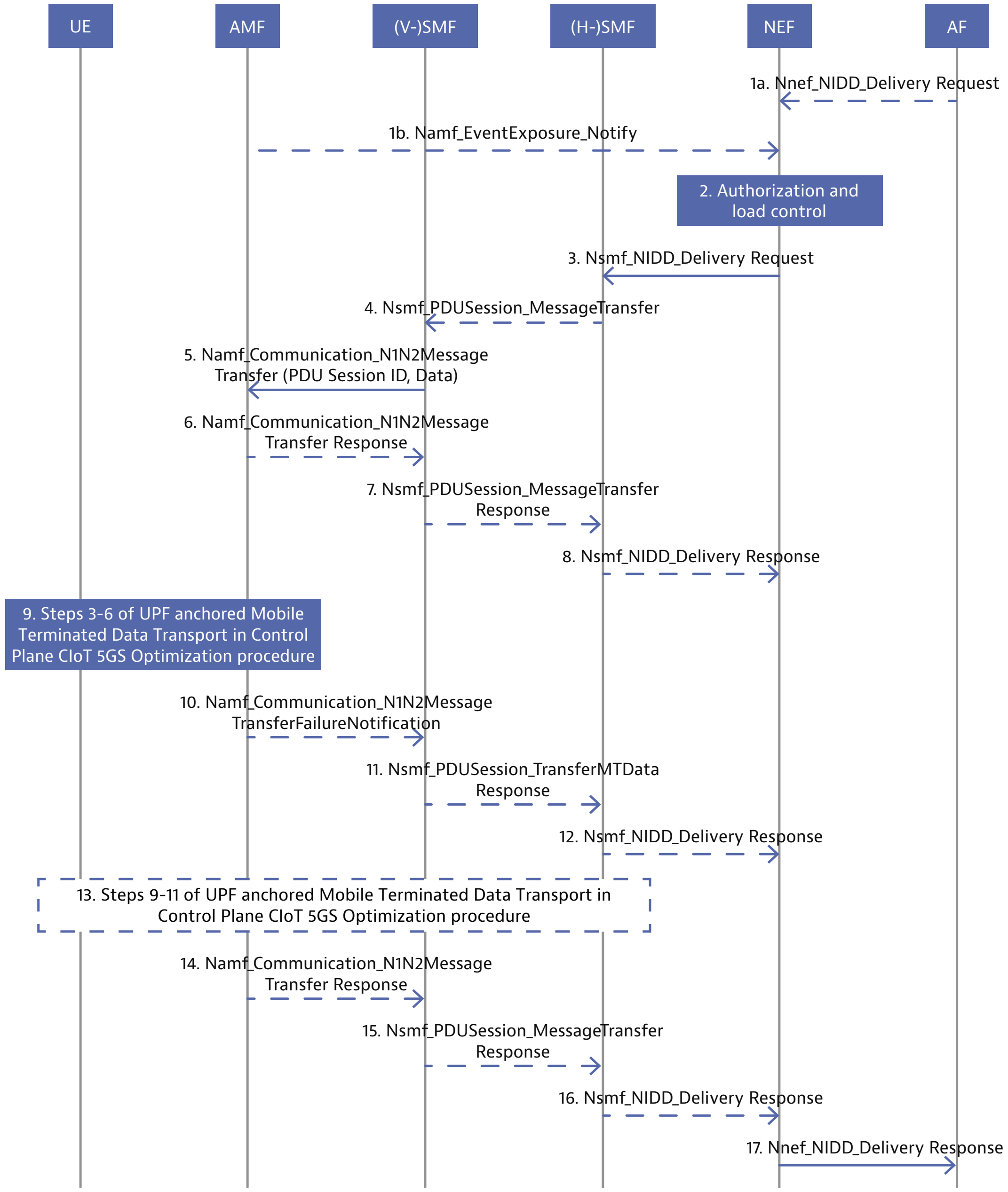

UE
AMF
(V-)SMF
(H-)SMF
NEF
AF
1a. Nnef_NIDD_Delivery Request
1b. Namf_EventExposure_Notify
2. Authorization and load control
3. Nsmf_NIDD_Delivery Request
4. Nsmf_PDUSession_MessageTransfer
5. Namf_Communication_N1N2Message Transfer (PDU Session ID, Data)
6. Namf_Communication_N1N2Message Transfer Response
7. Nsmf_PDUSession_MessageTransfer Response
8. Nsmf_NIDD_Delivery Response
9. Steps 3-6 of UPF anchored Mobile Terminated Data Transport in Control Plane CIoT 5GS Optimization procedure
10. Namf_Communication_N1N2Message TransferFailureNotification
11. Nsmf_PDUSession_TransferMTData Response
12. Nsmf_NIDD_Delivery Response
13. Steps 9-11 of UPF anchored Mobile Terminated Data Transport in Control Plane CIoT 5GS Optimization procedure
14. Namf_Communication_N1N2Message Transfer Response
15. Nsmf_PDUSession_MessageTransfer Response
16. Nsmf_NIDD_Delivery Response
17. Nnef_NIDD_Delivery Response

Figure 7.17 5G Mobile terminated NIDD procedure

1a. The application sends the data request to the NEF. If AF has already activated the NIDD service for a given UE and has downlink unstructured data to send to the UE, the AF sends a Nnef_NIDD_Delivery Request (GPSI, TLTRI, unstructured data, Reliable Data Service Configuration) message to the NEF.

1b. AMF indicates to NEF that the UE has become reachable. Based on this the NEF re-starts delivering buffered unstructured data to the UE.

2. The NEF determines the 5GS QoS Flow Context based on the DNN associated with the NIDD configuration and the User Identity. If an NEF 5GS QoS Flow Context corresponding to the GPSI included in step 1 is found, then the NEF checks if the AF is authorized to send data and if it does not exceed its quota or rate. If these checks fail, then steps 3-15 are skipped and an appropriate error code is returned in step 17.

3. The NEF forwards the unstructured data to the (H-)SMF using Nsmf_NIDD_Delivery Request.

4. In the roaming case, the (H-)SMF forwards the data to the (V-)SMF.

5. The (V-)SMF determines whether Extended Buffering applies based on local policy and based on whether NEF has indicated support of Extended Buffering in Nnef_SMContext_Create Response during SMF-NEF connection establishment. (V-)SMF compresses the header if header compression applies and forwards the data and the PDU session ID to the AMF using the Namf_Communication_N1N2MessageTransfer service operation. If Extended Buffering applies, then (V-)SMF includes "Extended Buffering support" indication in Namf_Communication_N1N2Message Transfer.

6. If AMF determines the UE is unreachable for the SMF (e.g., if the UE is in MICO mode or the UE is configured for extended idle mode DRX), then the AMF rejects the request from the SMF. The AMF may include in the reject message an indication that the SMF need not trigger the Namf_Communication_N1N2MessageTransfer Request to the AMF, if the SMF has not subscribed to the event of UE reachability.

   If the SMF included Extended Buffering support indication, the AMF indicates the Estimated Maximum Wait time, in the reject message, for the SMF to determine the Extended Buffering time. If the UE is in MICO mode, the AMF determines the Estimated Maximum Wait time based on the next expected periodic registration timer update expiration or by implementation. If the UE is configured for extended idle mode DRX, the AMF determines the Estimated Maximum Wait time based on the start of next Paging Time Window. The AMF stores an indication that the SMF has been informed that the UE is unreachable.

7. In the roaming case (V-)SMF sends a failure indication to (H-)SMF. If the (V-)SMF receives an "Estimated Maximum Wait time" from the AMF and Extended Buffering applies, the (V-)SMF also passes the "Estimated Maximum Wait time" to the (H-)SMF.

8. If the (H-)SMF receives a failure indication, (H-)SMF also sends a failure indication to NEF. If (H-)SMF has received the "Estimated Maximum Wait time" and Extended Buffering applies, the (H-)SMF includes Extended Buffering time in the failure indication. The Extended Buffering time is determined by the (H-)SMF and should be larger or equal to the Estimated Maximum Wait time. The NEF stores the DL data for the Extended Buffering time. The NEF does not send any additional Nsmf_NIDD_Delivery Request message if subsequent downlink data packets are received. The procedures stop at this step.

9. If the AMF determines the UE to be reachable in Step 5, then Steps 3 to 6 of the UPF anchored Mobile Terminated Data Transport in Control Plane CIoT 5GS Optimization procedure apply.

   If the Reliable Data Service header indicates that the acknowledgement is requested, then the UE shall respond with an acknowledgement to the DL data that was received.

10. If the UE has not responded to paging, the AMF sends a failure notification to the (V-)SMF. Otherwise the procedure continues at step 13.

11. In the roaming case, if (V-)SMF has received a failure notification from AMF, then (V-)SMF passes the failure notification to H-SMF.

12. If (H-)SMF receives a failure notification, then SMF indicates to the NEF that the requested Nsmf_NIDD Delivery has failed. If Extended Buffering applies, then NEF purges the copy of the data. The procedure continues at step 17.

13. Steps 9 to 11 of the UPF anchored Mobile Terminated Data Transport in Control Plane CIoT 5GS Optimization procedure apply.

14. AMF informs (V-)SMF that data has been forwarded.

15. In the roaming case, (V-)SMF indicates to (H-)SMF that the data has been forwarded.

16. (H-)SMF indicates to NEF that the data has been forwarded. If Extended Buffering applies, then NEF purges the copy of the data.

17. The NEF sends a Nnef_NIDD_Delivery Response (cause) to the AF.

The Reliable Data Service Acknowledgement Indication is used to indicate if an acknowledgement was received from the UE for the MT NIDD. If the Reliable Data Service was requested in step 1, then the Nnef_NIDD_Delivery Response is sent to the AF after the acknowledgement is received from the UE or, if no acknowledgment is received, then the MT NIDD Submit Response is sent to the AF with a cause value indicating that no acknowledgement was received.

7.5.5 NB-IoT Switching Between IP-Based Tunnel and Non-IP Data

Certain NB-IoT devices will have the ability to switch between IP-based tunneling and non-IP data depending on how much data needs to be transferred. An autonomous car is a good application for this. When the car is idle at a traffic light, it doesn't need to send and receive as much information as when it is moving. Non-IP data could be used in this situation so as to not burden a network.

When the traffic light changes, the car will need almost constant contact with the application to get vital information to drive the car. When this happens, the NB-IoT device switches to an IP-based tunnel to ensure the timeliness and integrity of the data.

7.5.6 NB-IoT Using SMS

NB-IoT devices can also send data using SMS technology, much in the same way that a text is sent through a network. As of this printing, none of the major mobile operators in the world are planning to use SMS since it is not as reliable as sending non-IP data through the control plane. SMS data could also impact user texting services by adding congestion to a network.

7.6 Implications

For CIoT to work and scale properly, there are several implications that must be considered as stated below. It will be up to the operators to keep these in mind as they build out their CIoT infrastructure.

- **Radio Efficiencies** – Are the devices using IP-based tunneling when they should be using the control plane to send data? Making sure the IoT devices are using a network efficiently will help the performance of a network. Mobile operators will need a way to monitor this.
- **Radio Quality** – The IoT device should be positioned and configured correctly to maximize radio performance and minimize power requirements.
- **New Infrastructure** – Using the control plane of a network for data is new. Mobile operators will need a way to properly monitor this part of the network for Quality of Service (QoS). To do this, more visibility into this portion of a network will be needed as well to accurately orchestrate and automate changes since so many devices will be connected at the same time.
- **IoT Coexistence with Other Services** – Operators should ensure that the IoT data and its network usage doesn't slow down or interfere with the performance of a network RAN, Control plane, or User plane which could disrupt non-IoT users on a network. They should also ensure that IoT device density and traffic profiles don't interfere with the QoE for handset users.
- **End-to-End Troubleshooting** – Mobile operators will need new ways to rapidly pinpoint and identify IoT issues in an environment where there are a massive number of heterogenous devices and networks in both functional and load scenarios.
- **New Quality Metrics** – Since networks are changing and being used in new ways for new vertical markets, operators will need new metrics that provide verticals-centric QoE that targets specific groups of IoT devices.
- **Managing Battery Life** – This ensures that an IoT device battery is going to last as long as predicted and not require early intervention through costly site visits.

7.7 Management Philosophies

Mobile operators are creating new business divisions to focus solely on IoT since there is so much opportunity. Even so, certain philosophies must be considered to make IoT a success for their businesses. One area for operators to focus on is the management of the devices and how those devices interact with a network to ensure the best efficiencies and long-term business profitability.

7.7.1 Managing Battery Life

One of the most critical aspects is managing the life of the batteries in an IoT device. As discussed earlier, there are technologies that can help lengthen the life of batteries, but that is only one part of the equation. For example, what if a device is improperly placed in a factory where it is located on the edge of a cell tower's range? The IoT device will have to increase its power output to properly communicate with a network or bounce between cell towers. This will quickly drain the batteries. A truck roll will be needed to replace the batteries, increasing the costs of running the device.

A device could drain its batteries if the device is too chatty, communicating with a network more than is necessary. Battery life would also be in jeopardy if the IoT device is using the wrong technology to communicate. Is it using IP-based tunneling when it should be using non-IP data? Is it using a standard LTE data connection in the user plane of a network when it should be using non-IP data in the control plane? These are all considerations that must be studied and managed to get the most life out of the batteries of each device.

If the mobile operator and the application provider are two different companies, who is responsible for managing the battery life and how do they obtain the information needed to see the problem? An application vendor might be able to see what the battery level is and how fast it is depleting but might not know why it is depleting so fast. Some of the early IoT deployments didn't include GPS trackers, so finding the exact location of the device might be a problem. This would make the example where the device is improperly located in a factory close to impossible to solve.

A mobile operator could see the RF parameters around the device in this situation and might be able to figure out at least part of the problem. It would come down to how well the application vendor and the mobile operator work together to find the problem. If the mobile operator is also the application vendor, then there are third-party solutions that can track and monitor the IoT device. This is the best possible solution since all aspects of the device would be visible to the monitoring solution and the problem could be quickly resolved as long as GPS information is provided. The importance of a monitoring solution will only grow as the number of devices continues to grow. Mobile operators won't have time to manually troubleshoot each device in the future.

7.7.2 Signaling Storms and Security

If IoT devices are not programmed correctly, they could communicate too frequently on a network. The sheer volume of devices communicating could create a perfect signaling storm that brings a network down. When the iPhone 6 first became available, it was programmed to contact a network much more than was needed. This had a negative impact on the control plane of LTE networks, slowing networks down. A software update to the phone was needed to fix the issue but figuring out what was causing the problem on networks took some time.

Part of transitioning from 4G LTE to 5G is the virtualization of networks where software will play a much greater role in how a network functions. This opens up the potential for security breaches. For example, if a network operating system is compromised, hackers could change the communications pattern of smart meters from once per month to once per second. This could bring a network to its knees. The integrity of the devices is a critical part of managing IoT to protect against security breaches.

7.7.3 Indoor Deep Coverage Device Placement

The indoor placement of devices in deep coverage situations will be important as these devices leave little room for error, since they are further from the core of a network. If the placement of the device in a building is off, even by a small amount, the device may drain its battery, overtax a network, or not be able to reach the application.

Some questions that should be considered:

- Can the device communicate with the application?

- Is it placed for the best cell coverage?
- Is the device using the correct amount of power?
- Is the output of the device on the right frequency?
- Is it using the correct enhanced coverage mode?

To answer these questions, a mobile operator must know several aspects of each IoT device. But most operators cannot currently recognize every IoT device on their network, much less what that device is doing. Most operators can categorize only ninety percent of the devices on their networks. They know that the device has an RF chipset, for instance, but they are not sure what type of service the device is using.

To answer the questions above, mobile operators will need to find better solutions to gain the visibility needed to not only see each device and know that it is indeed an IoT device, but to also see all aspects of that device such as what type of vertical it is supporting. A monitoring solution looking across the RAN and core parts of a network, as well as all the IoT devices' communications, could see where the devices are located, see if any of the devices are having issues, and be able to drill down to each individual device to see exactly what the issue is so it can be corrected. As networks become more virtualized and automated, monitoring solutions should be able to make corrections to the device on their own—only notifying humans if something physical must happen to fix the problem. For example, if the device is not located properly in the building or has run down its battery.

7.8 5G Network Slicing and CIoT SLA Management

If a mobile operator knows the specific information about what type of IoT device is in use, where it is located, and what that device is doing, the operator can start to implement different network slices and apply different Service Level Agreements (SLAs) to different scenarios. Each vertical being served by IoT will have specific needs such as how much coverage is needed, the amount of network bandwidth needed, how sensitive is the device to communication delays, and how available the IoT device and application need to be.

What is the energy need for the device in that vertical? Does the device have a full-time power source, such as an autonomous car, or does it have a finite battery life such as a smart meter? What is the vertical's sensitivity to delays in communication? Every vertical will be different.

For example, the needs of the mobile health vertical using IoT devices in pacemakers will be vastly different than a utility company needing information from smart meters. Smart water meters are not sensitive to delay, availability, or bandwidth requirements but do require deep coverage as they are often located in underground locations and are high-energy efficient to lengthen battery life. Conversely, a pacemaker application does not require deep coverage but is sensitive to delays and network bandwidth because decisions need to be made quickly. Therefore, the SLAs will be vastly different for these two examples.

For pacemakers, each device will need to be managed individually to a stringent degree by the application provider with no delays in network communications. For smart meters, an application vendor might sample a geographic region to see if there are any meters in the area and determine if any of the meters are malfunctioning. In this case, delay and individual monitoring is not important. For every vertical, automated identification of a device will be critical for IoT to succeed on a mass scale.

The Internet of Things will add billions of devices to the Internet and stands to trigger the next industrial revolution. However, IoT's demands are challenging for current cellular networks: These applications require high data rate and lowest possible latency as "things" need to communicate with each other continuously. 2G networks were designed for voice, 3G for voice and data, and 4G for broadband Internet experiences. 5G with its Network Slicing capabilities is going to be the biggest facilitator of IoT.

5G will be the first network designed to be scalable, versatile, and efficient in terms of energy consumption. In other words, each device and network created that is based on IoT will use only what it needs and when it needs it, instead of consuming anything and everything that's available.

5G promises to deliver latency and improve data rate and coverage. It will follow what is called a "non-orthogonal multiple access" model that supports multiple users to share limited bandwidth channels. This will allow myriad devices to share data in a timely manner without having to wait for other devices to use a bandwidth channel and release it. Consequently, we will be able to add indefinite numbers of devices to the Internet without having to worry about scalability issues.

With 5G, we'll see computing capabilities fused with communications everywhere, so billions of devices don't have to worry about computing power because 5G networks will bring computer processing to devices that need it. 5G networks will be faster but also a lot smarter.

5G networks are designed not only to better interconnect people, but also to interconnect and control machines, objects, and devices. They will deliver new levels of performance and efficiency that will support a broad set of industries. 5G is not just about multi-Gbps peak rates, it also brings Ultra-Reliable Low Latency, high reliability, and massive IoT scale. Allowing industries to tap into these new capabilities will help to facilitate the next industrial revolution. Mobile operators can ensure the success of CIoT and 5G through the management of the devices and how those devices interact with the network.

Category	Application Example	Availability	UL Bandwidth & Data Size	DL Data Size	Frequency	Power	Delay Sensitivity	Coverage
Automotive	Connected Car	HIGH 99.9%	HIGH 10Mbps 200bytes	HIGH 200bytes	HIGH Continuous	LOW	HIGH 1ms	NORMAL
Industrial Control	Switch on/off, device triggered to send	HIGH 99.999%	HIGH 50Mbps 0-20bytes 50% of cases require UL response	HIGH 20bytes	MED 1 day (40%) 2hrs (40%) 1 hour (15%) 30mins (5%)	LOW	HIGH 1ms	DEEP
Utilities/ Meters	Smart Water Meter	LOW 99%	LOW 50kbps 20bytes with cut off of 200 bytes	LOW 50% of UL data size	MED 1 day (40%) 2hrs (40%) 1 hour (15%) 30mins (5%)	HIGH 10yrs & 4800mAH	LOW 5sec	DEEP
Security	Smoke alarm detectors, power failure otification, tamper notifications	HIGH 99.9%	LOW 50kbps 20bytes	LOW 0 ACK payload size is assumed to be 0bytes	LOW Every few months, Every Year	LOW	MED 1sec	DEEP

Figure 7.18 CIoT SLA criteria

Notes

1. Ericsson Mobility Report November 2018
2. 5G America CIoT report 2017
3. 3GPP TS 23.401: General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access
4. 3GPP TS 23.682: Architecture enhancements to facilitate communications with packet data networks and applications
5. 3GPP TS 23.502: Procedures for the 5G System (5GS)
6. 3GPP TS 23.501: System architecture for the 5G System (5GS)

CHAPTER EIGHT

Automation and Optimization

We are asking a lot of our 5G systems. 5G must deliver a wide range of challenging use cases with extraordinarily diverse characteristics in terms of data rates, latency, reliability, and the number of connected devices. Some of these requirements contradict each other, yet the system must be able to deliver these use cases simultaneously with the same network.

Earlier chapters showed the innovations of radio technology for spectral efficiency, antenna arrays for massive MIMO and beamforming, and new architectural flexibility and distributed computing environments that will play a major role in meeting these requirements. In those chapters, we saw how 5G is architecturally designed and implemented so that it can deliver these use cases.

However, technology development is only one part of the story. The way it is operated, managed, and optimized has a critical role in delivering the ITU 5G service enablers. In this chapter, we explore the areas of automation and optimization, including what can be optimized along with the beneficial outcomes of optimization. The chapter will also cover artificial intelligence and machine learning and how these will play a major role in delivering on the promise of 5G.

8.1 Business Drivers for Automation and Optimization

Optimization has been a cornerstone of wireless networks and has facilitated increased performance since the earliest technology generations. Even from the early analog days of wireless communication, the network could be tuned to offer a capacity and coverage that was better able to meet the needs of the users. This was achieved through the choice of assignment of specific carrier frequencies to cells and adjustment of powers, for example. As the digital era dawned with 2G narrowband time division multiple access systems such as GSM, the challenge of frequency planning continued, although it was somewhat mitigated by frequency hopping and the introduction of flexible voice coder-decoders (CODECs).

The introduction of handover between cells introduced the need to plan the lists of candidate neighbors that mobiles could seek as potential handover targets. With spread-spectrum systems such as 3G UMTS, the planning of frequencies became less important and was eclipsed by the need for management of interference to maximize capacity while maintaining the desired coverage and achievable data rates.

With 4G LTE came the concept of self-optimizing and self-organizing networks (SON), and various capabilities were built into the standards in order to facilitate the automation of what had previously been manual. This included various procedures for initial configuration of parameters, from the choice of physical identities to prevent close-proximity clashes, or the choice of what cells a mobile would be directed to search for, as candidates for handover. Other aspects of SON were concerned with the parameters that controlled when a handover would take place between cells so that the risky transition process is more robust. SON also included facilitation of the balancing of the often-conflicting objectives of coverage, capacity, and quality.

Many of these features have yet to make their way into the 3GPP 5G standards. If and when they do, there are still many aspects of operating a network that require optimization. We could think of the concept of optimization as the configuration and operation of the network in a way that best delivers the services required, at the capacities required, in the locations where they are required, while meeting certain criteria such as characteristics in the level of service, reliability, or energy consumption.

Manual optimization today performed by engineers includes the analysis to identify problems, troubleshoot the root cause, and perform manual interventions to change various aspects of network configuration. This manual process is slow and requires laborious workflows to reach a conclusion. To scale this up naively requires more linear scaling in time, engineering resources, and ultimately cost.

Automation takes the principles of manual optimization and encodes the process in software that can be performed automatically with minimal or even no human interventions. This vastly reduces the cycle time for interventions and frees the engineers to manage other aspects of network operations. The workflows of manual optimization can be automated but as we shall see, automation opens the door to entirely new approaches that are not even possible with the current manual approaches.

There are other various motivating drivers for optimization. One such driver is the fact that the network is being designed to satisfy the connectivity needs of multiple vertical use cases with sometimes strict requirements in terms of data rates, latency, reliability, and device density. These requirements in many cases are inversely correlated to one another; to satisfy one you need to relax another. Ultimately, optimization will be expected to find a state for the various aspects of a network configuration that delivers a service that is the best compromise between the various conflicting goals.

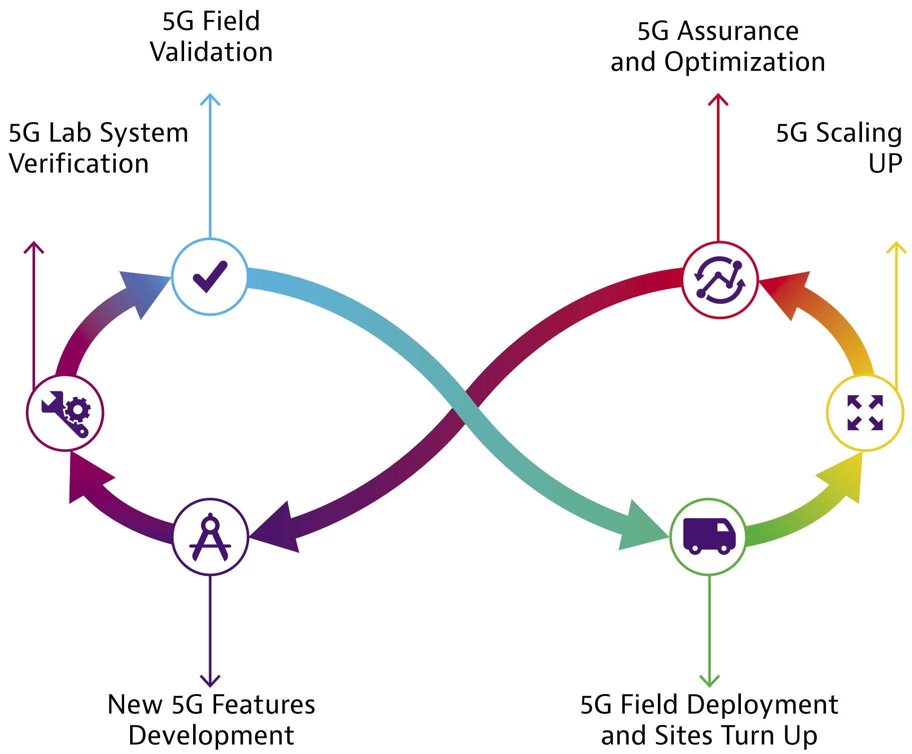

5G Field Validation
5G Assurance and Optimization
5G Lab System Verification
5G Scaling UP
New 5G Features Development
5G Field Deployment and Sites Turn Up

Figure 8.1 5G technologies life cycle

An additional driver for optimization and automation is the new dynamic nature of the networks. The ways that networks are configured changes constantly. While some Internet of Things (IoT) devices are expected to be stationary in street appliances or smart meters, and fixed wireless access (FWA) access points stay attached to the side of buildings, many UEs will be moving. People with smartphones traveling around as part of their daily routine and connected cars on highways driving at higher speeds place varying demands on the network.

This dynamic property goes well beyond spatial distribution as the use of applications varies. Critical communications for connected cars or smart manufacturing will vary by time of day such that the usage by application and network slice will be much more dynamic. This presents a challenge for the operator. At one end of the spectrum, a network can be optimized statically for average demand. At the other end of the spectrum, it can be constantly reparametrized so that it is always perfectly tuned for the needs of the users where they happen to be located, using the applications they want and need. While this latter vision may never be achieved, it has seen progress in recent years. It envisions a network that has greater capacity to serve subscribers since it is able to tune its parameters to adapt to what connectivity is required, what quality of service mix is required, where it is required, and when it is required. This results in a network that delivers more return for a given investment.

There is also a trend toward a greater mix of network infrastructure and virtual function vendors along with more disaggregation into logical components with well-defined interfaces in a network. This means that there is less scope for optimization to be performed within the ecosystem of a single vendor. Or rather, if nothing else is done, the push for openness and commoditization will be in vain unless the ability to automatically manage and optimize a system of heterogeneous components is also delivered. The alternative is a strong incentive to select a single vendor who has a proprietary intelligence layer for controlling the performance by optimizing at the system layer, with all the cost and vendor lock-in that this entails. An automation layer is needed to deal with complexity and abstraction of network system behavior irrespective of the vendors.

Within the operator, keeping the operational cost of running the network under control is an important factor in optimization. This may be achieved by choosing an optimization paradigm that allows cases where there are performance issues arising from poor optimization, and then root-causing these and resolving them with network reconfiguration.

8.1.1 Stakeholders in Automation and Optimization

The drivers for optimization have stakeholders who care about the outcome, and there are various outcomes of optimization and automation that they care about. Here we examine the various stakeholders in network behavior and the nature of their interest.

Clearly the users of a service are stakeholders in its performance. Their satisfaction with the service will depend on how well the network is optimized for delivering the applications they want to use in the locations and at the times that they want to use them, along with sufficient quality of experience.

The operator of the network has a stake in the optimization and automation. As well as ensuring that their customer base is happy with the service, the CFO will want to ensure that the capital and operational costs are balanced with the revenue a network generates. These are all outcomes on which optimization and automation will have a bearing.

Connectivity services for specific industrial and commercial verticals may be provided by third-party specialists who procure network connectivity from the operator, typically in the form of a network slice, and market the service to their customers. As well as having an interest in the outcomes of optimization and automation in terms of network performance, these entities need ease of service creation along with assurance and confirmation that service level agreements (SLAs) are met.

The regulator also has a stake in the outcomes of the optimization and automation. They will typically attach conditions to the right to use spectrum. These obligations can be in terms of the performance, where the network can be used, and what applications can be used. Whether these criteria are met will often depend on optimization and automation.

8.2 Benefits of Optimization and Automation

Optimization and automation are performed for a reason: to satisfy the needs of the stakeholders as outlined in the previous section. Let us explore what motivates these outcomes in more detail. Naively, the intent of these activities is to enhance performance of the network. But what exactly is the performance that we seek to improve? This has many dimensions and layers, including what services are available in what locations and with what capacities. Additionally, there are other objectives that are beyond what can be characterized as performance.

8.2.1 Delivery of the 5G Service Enablers

The success of 5G will be measured on its ability to deliver on the promised 5G service enablers, namely, enhanced Mobile Broadband, Ultra-Reliable Low Latency Communication, and massive Machine Type Communications, and optimization has a major role to play in making these a reality. The services built on these slices will succeed or fail depending on the ability of the network to adapt to delivery of the vertical use case slices in concert.

The success with which a service is delivered will often be measured in terms of SLAs. These are typically enshrined in the contract between the network operator and another party such as the provider of the vertical use case or the end user. These SLAs can be expressed in a variety of ways but will typically include measures of latency, capacity, and reliability, and may vary by geographical area. If these performance measures are not met, then there will typically be penalty clauses in which the network operator will receive reduced fees for access to their network or, in extreme cases, must pay compensation to the provider of the vertical service. Clearly it is desirable that optimization and automation enhance the ability to meet the SLAs. This can be achieved most directly in terms of the measures of application performance. But optimization can also deliver improved resilience in being able to recover quickly from an impairment, outage, or other interruption to the service.

8.2.2 Enabling Coexistence of Network Slices for Rich Service Ecosystems

5G networks will be expected to deliver slices of connectivity for multiple vertical industries simultaneously. Each of these will have its own applications with a unique set of characteristics or network performance that constitute what is a good, acceptable, or poor user experience. Each may have its own SLAs associated with it that will typically differ from slice to slice. Where these characteristics of performance and SLAs vary between slices, optimization and automation will play a major role in balancing these requirements together.

There may be cases where there is a conflict between the requirements of the various slices that cannot be satisfied simultaneously. In this case, the optimization process will have to balance the requirements between the slices while some lower priority slices are denied a full service in order that the higher priority slices can meet their service requirements fully. This becomes particularly important in the presence of impairments. Malfunctioning network functions or degraded connectivity between them will temporarily restrict the ability to deliver the service for all users on all slices.

8.2.3 A Platform for Frictionless Service Creation and a Market in Connectivity

Network slices with specific expectations of performance and associated SLAs may often be operated by third parties within an industry such as automotive, utilities, transportation, media, or entertainment. In order to foster a rich ecosystem of services that are engaging to the consumer along with being profitable to both the operator of the network and the slice, this partitioning of a network into slices will require some degree of automation. Any manual steps will add friction and present some barrier to bringing innovative services to the market. Some manual aspects of the lifecycle of service delivery, such as commercial negotiations, may be hard to remove entirely. But operators can aspire to remove even these steps.

By whatever means a new service vertical is accepted onto a network, that network will generally have to respond to ensure that it is optimally configured for the new service along with the existing network customers and slices. The reduction in friction to new services will depend on the ability to make the right decisions on how to reconfigure a network and automate the deployment of the reconfigurations.

But this reduction in friction goes beyond just the step at which the new service is integrated into a network. That optimization step will generally depend on models of a network. These models are discussed later in the chapter, but in summary, they will model various aspects of a network including how a network responds to stimuli and change. As well as underpinning the ability to change the configuration of a network for more optimal performance, these models can also be used earlier in the lifecycle of a service. At the stage of negotiation between the network operator and the operator of the vertical service, the network operator can assess the impact of the proposed new service on the network, whether the service performance targets can be achieved, whether the SLAs can be met, and at what cost, if any, to the other users of a network.

The answers to these questions can help with the decision of whether the service can be supported, and how much should be charged for the connectivity. If these questions can be answered by an automation and optimization system using the underlying models of a network, then the decisions themselves are more amenable to automation. This reduces further the friction of the entire service creation lifecycle.

At its limit, it is plausible that a market in connectivity may emerge which will see a rich and dynamic ecosystem of services with the provider of the connectivity rewarded for their investment in the next generation of high performance and flexible communication infrastructure, and the spectral resources on which these depend.

Together, these capabilities will result in a more responsive network able to deliver new services and retire legacy services with very low friction. The network of the future, when supported with the right automation and optimization, will be a platform for innovation with low cost of entry for new services, fostering a thriving industry of innovation.

8.2.4 Achieving Optimized OPEX to Manage Operaional Complexity

As we have seen, delivering a flexible ecosystem in which innovative and valuable services can be created easily to enable new applications is a key benefit of automation and optimization. But there are other outcomes to which we can aspire that go beyond the ability of a network to deliver service to subscribers and meet SLAs. Some of these can be addressed in part by optimization and automation. We explore these here.

As we saw in Chapter 6, we risk driving up the operational costs substantially unless there is a transition away from the traditional approach to managing a network. This change of mindset must take account of the complexity of a 5G network and delivering what is expected of it. A shift in approach is not just about delivering a set of network slices that coexist in harmony in a cost-effective way. Rather, it is an imperative born of the fact that no group of human network engineers will be able to achieve what is required. The entity or entities managing a network must understand the operation and dynamics of the network in sufficient detail to identify suboptimal performance, impairments, and interruptions to service, and cases of SLA infringement.

The managing entity must be able to troubleshoot the identified issues with enough resolution to discover the root cause quickly and accurately. It must be able to identify the changes that might resolve the issues and select from these the change or changes that most effectively achieve the resolution. And it must do all this across the entire end-to-end network from the edge through the transport network to the centralized functions, from the application layer to the physical layer and in sufficiently short timescales to minimize the impact of suboptimal performance and impairment.

This vision of operating a 5G network must be achieved to deliver on the full value of 5G. But this vision cannot be delivered as it has in the past, with teams of engineers working in manual ways. They will have to depend increasingly on more sophisticated solutions to be able to rapidly assimilate large quantities of heterogeneous information from across a network and make decisions that are substantially more autonomous than before. This vision is hugely ambitious, but as we saw in Chapter 6 it is also achievable and has precedent in other industries. The keen CFO may hope that the need for expensive operational staff will instantly evaporate, leading to the dream of a more performance-based network combined with dramatic reductions in staffing costs. Unfortunately, the latter of these is unlikely to happen immediately. Although the staff may be making fewer decisions about when and how to reconfigure the network parameterization, there will be plenty of other concerns that will require their attention.

This is because the systems that optimize and control a network will themselves require management and maintenance, along with expansion and upgrading. The DevOps role will become critically important initially, much as it has in web-scale companies. The need to incorporate new automated ways of creating and operating network slices will keep these roles busy as the transition to automation progresses. As new classes of service using a network come along, new ways to automate their happy coexistence alongside the other ecosystem of slices will be paramount. Unless and until robots will be expected to replace faulty hardware and upgrade obsolete parts, staff will still play a key role in ensuring the network operates efficiently, especially keeping track of the physical location of additional physical resources in more distributed locations with all the many issues that can befall a physical plant.

8.2.5 Reduced or Deferred CAPEX

The early days of a commercial 5G network are characterized by the rapid buildout of new sites in the race to cover as many potential subscribers as possible. The attention will shift to keeping abreast of the capacity demand once the initial coverage targets are achieved. This increased capacity generally cannot be achieved through a single mechanism but typically will be delivered through a mix of methods. These include network densification, addition of carriers, re-farming of spectrum from older technologies, or upgrading network infrastructure to support more advanced versions of the standards. In general, these will incur spending on a network infrastructure and the connectivity to deliver the transport. Also needing to be factored in are the associated costs of installation, commissioning, integration, spectrum licensing costs, and other activities associated with network growth.

Any mechanism to avoid these expenses, or at least to defer them, can result in substantial savings for an operator. This is where optimization process can have a role to play. For example, optimization of the capacity can squeeze more performance out of a network with the same infrastructure by delivering a system that has better signal-to-noise ratios and thus is able to utilize higher order modulation schemes and channel coders. In some cases, the cleaner RF and lower interference arising from optimization will mean that coverage is extended and could avoid some of the coverage expansion work.

A significant component of the capital expenditure will be for physical compute platforms for logical functions and the fiber links to provide the transport that will connect them. This will include the costs of laying fiber and acquiring sites to house hardware along with the costs of fiber and computer-associated hardware. The choice of where to physically place logical functions and how to route user plane and control plane traffic will have impacts on how much of each logical resource is required and how much physical transport is required. The choice of where to break out the user plane will also be a factor in these costs. An optimization suitably guided by these considerations can consider the capital expenditure so that physical compute or transport is avoided or deferred.

8.2.6 Reduced Energy Consumption

Reducing the energy consumed by a network leads to a reduction in operational costs. A major operator will incur annual electricity bills totaling many hundreds of millions of dollars, so a reduction in energy demand will produce significant savings. But such a saving also has peripheral benefits. More and more, companies can brandish their credentials as environmentally friendly enterprises. Having lower carbon footprints is seen as a competitive advantage as consumers migrate to reduce the impact of their consumption. One of the design considerations of 5G has been to keep power demands under control even in the face of delivering vastly more data to greater multitudes of devices. For example, the design of the 5G NR frame structure means that synchronization signals are transmitted only intermittently, so that when demand is low the radios can save power by only transmitting for small portions of the radio frame and shutting down for the remainder.

Optimization can augment these advances in the standards in a variety of ways. Accurate prediction of demand for services at different locations around a network facilitates the ability to take offline entire beams, cells, or entire base stations, or even logical functions allowing the underlying physical compute resources to be powered down. When objectives for power consumption are introduced into the control functions directing activity, the power required can be balanced against the coverage and capacity along with likely demand, such that a resilient service can match the demand and meet SLAs while minimizing energy requirements.

8.2.7 Cyber Security and Network Security

Smooth operation of a 5G network will be about far more than identifying and dealing with impaired services arising from the demand on a network and its ability to satisfy that demand. There will also be threats to a network, from subscribers seeking to achieve a service without being a subscriber, to unauthorized agencies attempting to eavesdrop on communications. It will include those determined to take computing cycles without authorization or disrupt service for malicious reasons, be it through jamming the physical spectral resources with interference or causing viruses to propagate a network. All of these can affect the integrity of a network, not just to the extent that the service is impaired but also that the users are threatened by loss of data or personal information.

There is an intersection between the cyber security and network security interests and the optimization and automation of the network. Many of the same approaches used for network optimization also have application in detection of security issues along with establishing the root cause. Models developed for the demands placed on a network and the way it responds to satisfy those demands will be able to detect deviations from normal operation. The anomalies in network behavior arising from nefarious activity will in many instances be discernible by the same models. In other cases, similar models using the same approach and infrastructure can be deployed at relatively low cost to further the detectability of this type of activity and attenuate its ability to impact network performance.

8.2.8 Meeting of Regulatory Requirements

Some regulators around the world are attaching stricter requirements for mobile operators to deliver high-grade services. This typically includes the requirement to deliver coverage to ever increasing proportions of the population. In some cases, there are requirements for certain data rates or latency at specific places, such as on roads. Meeting these regulatory requirements is yet another constraint to be considered in managing a network, alongside the delivery of capacity and coverage, meeting SLAs for the various network slices, and keeping capital and operational costs under control. Again, a mature optimization and automation system plays a significant role in helping the operator meet the constraints imposed by regulators.

8.2.9 Intent-Driven Optimization

The theme of these drivers and benefits for optimization is that control of a network will experience a shift. There will be an inexorable moving away from simply changing configuration settings with the objective of achieving lower-level network goals such as coverage, capacity, and quality. Instead, the focus will shift to specifying increasingly abstract policy objectives for the network in the form of intent, which can then be translated into optimization decisions that must be made. This was discussed in detail in Chapter 6 in the context of intent-driven orchestration. Orchestration of virtual network functions configured on physical infrastructure is a major component of this. It also encompasses the configuration of the functions themselves—i.e., the physical aspects, including radio functions—and of the transport network, for example.

Ultimately the operator will be setting policies about the services and user experience, and how a system will offer a variety of rich services on a set of network slices operating together. These intents and policies will translate into what sorts of services will be allowed on a network, in which locations, and how close to capacity to allow these services to operate, for example. This will mark the transformation towards intent-driven operation of a network that tells the optimization what objectives it should be achieving, not how they should be achieved.

8.3 Technology Enablers for Automation and Optimization

5G is anticipated to have several characteristics that will benefit from automation and optimization and create an atmosphere that allows mobile operators to thrive.

8.3.1 Disaggregation

The 5G network has more logical entities than previous generations. The gNB, which is the base station in 5G systems, is now defined as three logical elements. The central unit (CU) manages the processing of the higher network layers above the RLC layer. This is connected to the distributed unit (DU), which manages the lower network layers. The CU and DU are defined by 3GPP and connected by the F1 interface. The remote radio unit (RRU) is additionally defined and is connected via the lower-layer split (LLS) as defined by the Open-Radio Access Network (O-RAN) industry consortium.

One of the motivations for disaggregating the gNB into these logical elements was to address the challenge of network densification. As more cells are added to a network it will become increasingly hard to find locations where base stations can be placed. In many of these new locations—e.g., streetlights, sides of buildings—there will be insufficient power and space for a full base station.

To address this problem, disaggregation allows the minimum functionality to be placed by the radio in cases where power and space are limited. The rest of the functionality can then be placed in a more central location such as a local aggregation center. Disaggregation can be different for different entities such as carriers and other parts of the spectrum, and for the user plane and control plane. Although centralization has distinct advantages in facilitating network densification, it comes at a cost. The transport network for the lower-layer split has high data rate and very strict timing alignment requirements in order for the radio interface to function correctly. Thus, high grade transport connectivity, typically using costly fiber optic links, is required. Another disadvantage of centralization is that some very low-latency applications will require the user plane data to be broken out as close to the edge as possible. However, the user plane data is not available at network elements closer to the edge than the CU. Disaggregation and the resulting centralization clearly have benefits but also have the disadvantages identified above. Finding the right balance of cost and performance is an essential objective for optimization.

8.3.2 Network Flexibility and Complexity

The way that 5G has been defined permits vastly more flexibility in how it is constructed and configured. But with this flexibility comes potential complexity. As long as the complexity can be tamed, the flexibility can be exploited and used as an enabler for innovative ways of constructing the resulting system.

The disaggregation described in the previous section is one example of the way that 5G systems are more flexible. Beyond this there are choices of which carrier bands to use within the mmWave bands and the lower FR1 band. There is also the choice about how to break carrier bands into carriers and bandwidth parts, and which numerologies to use for these. How different UEs are scheduled onto these physical resources and how the resources are distributed between network slices are examples of the increased flexibility and complexity of the 5G system. When systems are flexible and complex, they can be constructed and reconfigured in a multitude of ways. As a result, ever larger sets of objectives can be achieved through careful choice of the configuration. Thus, the flexibility that we see in 5G systems is an enabler for a potent optimization layer that can drive performance to new limits, maximizing the return on investment for an operator.

8.3.3 Network Programmability

Flexibility and complexity are one side of the equation for automation and optimization. But unless the system is also programmable, the power of flexibility is locked away from the system operator. Parameters must be exposed and configurable in order to influence how the network operates. Traditionally, making changes has involved navigating graphical user interfaces in network element managers or operation support systems to change parameters manually. In extreme cases, changing a parameter or characteristic has involved visiting a site to manually change the direction in which an antenna is pointing, for example. While optimization is possible with these constraints, automation is not. There is a trend to make it possible to change the network via well-defined programmable interfaces such as APIs. It is this openness of exposed interfaces that brings the power of optimization to life with automation. It means not only that changes can be performed autonomously but can be made faster and more dynamically.

There will always be some level at which vendors of network functions make decisions and implement some aspects of a capability with no possibility to influence the operation externally. But these vendor-specific decisions will become less common as more disaggregation of functionality takes place in initiatives such as the O-RAN consortium, for example. Vendors can draw comfort from the fact that even if the opportunities for differentiated intelligence in the components become more limited, this will be counterbalanced by increasing opportunities for intelligent control systems to automate and optimize the 5G systems.

8.3.4 Virtualization and Service-Based Architecture

As we saw in Chapter 6, virtualization of network functions brings decisions about which functions to instantiate on what physical resource, in what physical locations, and with what connectivity. This allows more control over how the objectives of a network are achieved, with what performance characteristics, and with what resilience. Thus, virtualization is an enabler for optimization.

With the evolution to 5G, the 3GPP has decided to define many of the network functions as services. Not all functions are part of the service-based architecture, however. In particular, the RAN functions generally are not defined in this way. But many of the core functions are defined in terms of services. These have advantages such as well-defined interfaces and discoverability. With this transition, these functions start to resemble the services and microservices that are used to construct the typical web service. This architecture facilitates a network that is flexible where the logical entities used to construct a service end-to-end are chosen based on the service, the network slice, the service KPIs, and SLAs. This also means that the service functions can be placed in the physical locations, toward the edge or more centralized, where they are best able to deliver the latency requirements given availability of compute and other physical resources.

8.3.5 Artificial Intelligence and Machine Learning

Artificial intelligence (AI) and machine learning (ML) are another cornerstone enabler for optimization and automation. Modern wireless communication systems that convey vast quantities of data for the users are also awash in data generated by test and assurance systems, state information from elements, and other telemetry supporting network operations. Many machine learning algorithms and models thrive on data in large quantities, and these algorithms and models in turn underpin the artificial intelligence required to automate the optimization processes. This area will be covered in more depth later in this chapter.

8.4 A Closer Look at Automation and Optimization

We have seen that there are many drivers for automation and optimization, as well as multiple stakeholders in the outcomes. Moreover, we have seen that there are many technological enablers for these. Here we explore the different resolutions over which optimization decisions are made in 5G systems. We consider the temporal and spatial scales of decisions along with how and where the data are collected, and computation performed.

8.4.1 Optimization Timescales

Different types of optimization are performed over a vast range of timescales and cadence. Some aspects of the operation of a network are essentially optimized only once or are costly to change once deployed. The choice of where cells are placed is an example of this. Procuring sites along with power and transport connectivity is a costly business; there must be a compelling reason to change this once the site is deployed.

While there is significant friction to changing some characteristics of a network in cases where decisions must be made at the time of network design, other factors are much less of a barrier to modification. These can have diminishing or zero marginal cost in making a change. Some decisions can be made with great effect even over non-real-time timescales.

Examples of aspects of the network that can change more easily include the choice of the parameters that control the interactions in a network. These parameters control the transmission powers and beam directions that affect coverage and interference. Other examples include the parameters that control mobility between beams, cells, and radio technologies. Also, in this category are the configurations that control how traffic is routed through the transport network and where virtual functions are instantiated. Decisions about these factors can incorporate data and knowledge from multiple parts of a network. Decisions can be made considering the emergent system behavior and how this can best be aligned to the objectives of optimization or intent of the network.

Where static decisions made at the time of network design are at one end of the spectrum of timescales for optimization, the opposite end of the spectrum includes the near real-time decisions that must be made almost instantaneously in response to the changing dynamics of the network, the state of the radio, transport and network functions, along with the demand from the subscriber devices. Examples include which data to schedule in what order and on which parts of the OFDMA resource grid, or which beams to utilize for communication with specific UEs. These decisions must be made in a matter of a few milliseconds or less and must necessarily be made at or near the edge of a network where the resource mapping and other functionality is performed. This means that the real-time data that the algorithms have access to is limited; generally, only fresh data collected or generated at or near the network element making the decision can be used. However, the algorithms can rely on models and other aggregated information that have been gathered over longer timescales based on historic data. In this sense, the algorithms can learn what strategies are successful and which are not. We shall cover this in more detail later in the section on machine learning.

8.4.2 Spatial Extent of Optimization

In addition to different timescales, optimization can be performed on different spatial scales. Decisions of where a 5G NR beam should be directed and with what power are important, because they will affect which UEs can access the network from that cell, what degree of modulation they can achieve, and thus what data rate is possible. For a network containing a multitude of UEs and beams, the problem is to find the direction and power of each static beam such that as many users as possible achieve coverage while minimizing the interference that non-serving beams cause to those UEs that they do not serve. This is generally a problem best solved at the system level over larger clusters of beams, cells, and gNBs.

As we change the power or directivity of a single beam, other system parameters mean that UEs will now elect to get access from different beams; beams that were once beneficial by offering service will become detrimental as an interferer, and vice-versa. The beams can be thought of as interdependent. They must help each other by providing enough coverage but not to the extent that they have excessive overlap; that would result in excessive interference and loss of capacity. Thus, optimization becomes a problem of how to balance the beam powers and directivity of many beams over an area or cluster of cells. A piecemeal approach of localized decisions will be unable to find the unique mix of changes across a cluster of cells. But considering the cells in a cluster together facilitates the optimal trade-off between coverage and interference, allowing delivery of the capacity profile in the places across the geographical area that best satisfies the demand from users of the applications offered by a network.

The fact that some optimization decisions must be made within the context of a group of entities such as gNBs in a network raises the question of where and how these decisions should be made. Making such decisions requires data from across multiple entities over a finite geographical area of the network. Typically, the data must be used to create models that reduce the volume of the data into much smaller aggregates that capture the essence of the network, its performance, subscribers, or state. These models must be used to support the optimization decisions and must have a scope that includes an area of the network that is sometimes significant. This impedes the ability to make the decisions at points in a network that implement the decision, such as the gNB. Thus, some degree of centralization is required for this type of decision.

However, any drive towards centralization should be balanced against the need for data to use as inputs to the algorithms. More centralization can mean that data is transferred in large volumes from the network edge to the more centralized data center, placing extra strain on the transport network. This need to move data around can also impact the latency of decisions and lead to slower optimization dynamics. A hybrid approach that overcomes these challenges is to perform processing as close to the edge as possible. For example, models that underpin the decisions can be constructed at or near the edge, where the data required to make the models are available. Once these distributed models exist, they can be used in a variety of ways. One alternative is to use them in situ, being called upon to produce outputs which are then transferred to where the decisions are made. Alternatively, the models that capture the essence of the network can themselves be transferred to the more central locations where optimization decisions are made. The optimal choice for this distributed modeling and decision-making architecture will depend on many factors, such as the volumes of data required to create the models, the availability of distributed compute, and the latency required for decisions to be made.

8.5 Network Tuning for Optimization

In this section we build on the principles of the previous sections and describe the details of what can be configured, tuned, or otherwise changed in order to deliver optimization in the next generation 5G network. We review some of the specific ways that a 5G network can be adjusted or tuned in order to achieve a change in various aspects of performance. A thorough analysis of this would require a much longer treatment, so this section contains only a selection of the many ways to effect change, giving a flavor of the breadth of changes that are possible.

8.5.1 Beam Configuration

The power with which each beam is transmitted will naturally be a significant factor in determining how far from the transmitter it can be received by different

UEs. The transmit power will also contribute to what level of service can be received at which locations. Increasing the power will tend to increase the range and the coverage, but there are limits to this. If the beam is directed toward a structure that is opaque to the electromagnetic spectrum at the frequency of the carrier, then increased power will not achieve further coverage. In contrast, if the power is increased excessively then it will create interference with other beams using the same spectral resources. This will reduce the ability to support UEs in the area of interference. In the case where the capacity will be reduced, the types of service that can be supported may be restricted or the service may become entirely unavailable. There will also be constraints on the power at which the beams can be transmitted. These include the capacity of the power amplifier that prepares the signal for transmission. There are also sometimes regulatory constraints that must be respected.

In addition to the power with which the beams are transmitted, the direction in which they are transmitted is also important. Together with the transmitted power, the beam direction can be used to carefully deliver coverage where it is needed. The objective is a system that does not have undesirable holes in coverage yet manages interference and thus achieves a high level of capacity. In some cases, particularly for time division duplex (TDD) carriers, dynamic beamforming may be used to direct the beams in response to where the users are located, or even track them as they move. In this case, coverage beams facilitate the initial system access and handover between cells. These coverage beams must also be configured to achieve the required coverage while maximizing capacity.

Some systems will deliver retail or commercial broadband connectivity service only to static customer premises equipment (CPE). This is a simpler scenario in general, as the areas where coverage is required are dynamic only to the extent that new subscribers are added to the service. In order to have a more resilient service, some operators may impose targets for beam redundancy so that if one beam is occluded by a transient body such as a truck, the service is maintained through the other beams. But it is the systems that deliver full mobility for the services where coverage becomes so important. In this case, for those beams that are not dynamically formed, the task of optimization is to maintain enough overlap between static beams such that the users can move between the coverage areas of beams on the cell (intra-cell beam mobility) and between beams on different cells (inter-cell beam mobility). This must be done in a way that keeps the beams spatially separated enough to avoid excessive interference and the associated loss of capacity. When coordinated scheduling across multiple beams serving the same UE is used, data can be received on the same spectral resources from multiple beams for the same UE. This coordination can also be exploited for optimization.

There is flexibility in the way that feedback is configured for beam management and mobility. The optimal configuration will depend on whether the carrier uses frequency division duplex, time division duplex, or supplemental uplink or downlink. But there will be flexibility in the feedback configuration that will impact the system's ability to respond to dynamics in a timely manner.

8.5.2 Neighbor Lists

One of the functions of the UE is to measure beams that are not currently serving it, but which are candidates for providing service. When certain conditions are met, the UE will send a measurement report to the network including the quality of various candidate beams. The network may then elect to direct the UE to receive service from a new beam. To help the UE perform this, it is sent a neighbor list by the network, telling it what beams to attempt to measure. The network must decide how to populate the neighbor list and how many candidate handover target beams to include. Excessive numbers of neighbors can cause the UE to perform unnecessary measurements and also delay the process of transition to service from new cells. If good neighbor candidates are missing from the neighbor list, the risk of the connection being interrupted increases.

8.5.3 Physical Cell Identity

The beam mobility and handover process described above is facilitated by cell and beam identities. Each cell in a network is given an identifier that is unique in the network, and each beam on a cell is assigned a unique identifier within that cell. As the cell identifiers are unique within the network, they have to be represented by many bits of data to avoid collisions between identifiers. The process of decoding the full cell identity is a relatively costly operation for the UE. To speed up the process of measuring and reporting the strength with which beams from different cells are received, cells are also identified by a more concise physical cell identity (PCI). This PCI for a cell is determined by the combination of the identities conveyed by both the Primary Synchronization Signal (PSS), which can take one of three values, and the Secondary Synchronization Signal (SSS), which can take one of 336 values. Together there are 1,008 unique PCIs.

These 1,008 PCIs cannot be assigned uniquely across a network once the number of cells exceeds 1,008. To overcome this lack of uniqueness, there must be sufficient spatial separation between cells with coincident PCIs. It is important to maintain this PCI integrity such that when a UE reports a PCI associated with a measurement, the target cell can be unambiguously identified. This may require ongoing management of the PCIs, such as in the case where network densification requires the addition of new cells to the network. This can lead to the distribution of PCIs to become suboptimal over time.

8.5.4 Handover Parameters

As well as deciding which neighbor cells and beams to consider for handover, the system must be able to decide when to hand over service to new cells. Handing over too early can result in ping-ponging between cells, which can interrupt service and make connections less stable, increasing the chance of a service drop. Handing over too late can also cause instability and potential loss of coverage. This process is typically controlled using a variety of parameters that determine what quantities to compare between the serving beams and candidate target beams along with timers and hysteresis parameters.

8.5.5 Physical Layer Parameters

The physical layer is characterized by flexibility in the numerology; namely the choice of what subcarrier spacing to use. To some degree this choice is limited by the frequency carrier and the bandwidth, but often some flexibility is possible. The objective here is to balance the risk of inter-symbol interference with the risk of phase noise causing inter-subcarrier interference. In general, the risk of inter-symbol interference diminishes as the carrier frequency increases. In contrast, phase noise is less of a problem at lower carrier frequencies.

Where numerology 2 with a 60 KHz subcarrier spacing is used, there is the option of an extended cyclic prefix for cases where there is very complex propagation with excessive non-line-of-sight propagation. The degree to which mini-slots and slots for pre-emptive uplink transmission (such as non-orthogonal multiple access) can be configured balances the low-latency requirements with the capacity of the air interface.

In the case of the time division duplex, the limits on the ratio between downlink and uplink transmissions will be controlled by the choice of slot format. The configuration of bandwidth parts (BWPs) will balance what parts of the spectrum can be accessed by which UEs and consequently how the capacity is shared between the mix of device capabilities.

8.5.6 Layer Management

5G is capable of aggregating multiple carriers from 5G low band, mid band, and high band (mmWave) along with LTE carriers together in different combinations, as well as splitting 5G carriers into BWPs with their own numerologies. Managing which carriers and which BWPs are used in what circumstances will determine what performance characteristics such as data rates and latencies are achieved, in addition to the overall system capacity. Managing load between lower frequency coverage layers and higher frequency bandwidth layers will be important.

8.5.7 Scheduler Configuration

Transmissions from or to different UEs must be multiplexed onto the spectral resources by the scheduler. How this is done and how the streams from different users, bearers, classes of device, and network slices are prioritized will determine the quality of service experienced by the various applications, network slices, and subscriber classes, as well as the capacity of the overall system.

8.5.8 Idle and Inactive State Operation and Access Parameters

How the UEs operate when not in active state can be configured. Parameters determine which cell is monitored by a UE in case there is data for it to receive. Other parameters determine when this cell should change. The configuration also determines how often the UE should communicate with the network and in what circumstances it should indicate that it is monitoring a new cell. This will affect many outcomes such as how rapidly the UE can be reached when there is data to send, and how much system capacity is consumed for these communications. This configuration will also determine how much UE battery is consumed. The optimal configuration for these will depend on the type of device. For example, a battery life spanning many years is a paramount requirement for a battery powered IoT device such as a smart utility meter, while being able to establish connectivity very rapidly is unimportant. This is very different from the requirements for a commercial smartphone, the users of which will generally value apparently instantaneous connectivity and will be happy to charge their devices regularly.

8.5.9 Placement and Configuration of Network Functions

As discussed above, the 5G RAN is already disaggregated into the CU, DU, and RU logical functions, with further disaggregation planned by initiatives such as O-RAN. These bring choices of where these new elements are hosted, apart from the RU, which must be placed at the point of transmission. There is additional functionality concerning where mobile core functions are placed, along with where the user plane is broken out.

In situations where interference is excessive or extra capacity is required, coordination of beams and carriers between sites will encourage more centralization. The degree to which transmissions can be coordinated will depend on the nature of the transport for the fronthaul links. In cases of high-performance ideal transport, coordination can be performed on a phase-coherence basis to overcome interference. Less ideal transport will facilitate transmit and receive diversity giving an engineering benefit with a single transmission and reception chain consuming fewer resources. The placement of functions may vary by radio carrier or even by BWP offering yet more flexibility.

Mobile edge computing and distributed compute allow application functions to be deployed flexibly: either centrally or towards the edge. These decisions can vary depending on various factors. The network slices with low latency needs can have their functions pushed as close to the edge as the available compute will allow. Different connections can be assigned to different logical functions for application or core functions, and the degree of redundancy can be balanced against the amount of compute resource consumed depending on the resilience required.

8.5.10 Transport Configuration

The transport network has its own part to play in delivering performance of the overall system. The quality of services with poor tolerance to latency will depend on how effectively the traffic is prioritized by application, network slice, and device. The route taken through the transport network will also play a part. The resilience of the overall system will depend on whether redundancy is built in and how the routing is configured to take advantage of this.

8.6 The Role of Artificial Intelligence and Machine Learning

Many industries have seen substantial benefit arising from the appropriate application of artificial intelligence (AI) and machine learning (ML). Applications are numerous but include speech recognition, language translation, image classification and automated description, detection and prediction of disease, drug discovery, energy and carbon emission saving, automated driving, and many more. It makes sense to ask if the mobile telecommunication can similarly benefit from this technology area. In fact, in order to deliver the promise of 5G, AI and ML will play a critical central role.

AI has a variety of definitions. In the context of automation, we regard artificial intelligence as using an automated system to solve a problem that humans intuitively can solve but that computers typically find hard. Classifying a picture as containing an image of a cat is easy for a three-year-old child. In contrast, a programmer requires advanced skills and diverse tooling to reproduce the same capability on a computer. This is what characterizes AI. AI often involves ML, but this is not always necessary.

In contrast to AI, ML is the creation of models that describe relationships between inputs, typically called features, and outputs, typically called targets. A well-constructed model will be able to make predictions of the targets given an example of the features. The model may make a variety of different types of prediction. For example, targets may include a number in a continuous range, also known as a regression. A concrete example of this is a model that predicts how many people will be using a particular mobile service in one hour. Another model may make categorical predictions, such as whether certain characteristics are represented in the features. An example of this is whether users of a particular 5G MIMO beam are experiencing congestion.

Generally, ML models have parameters or state, and an optimal configuration for these must be found for the model to be able to predict the targets accurately given the features. A set of parameters that achieves this accurate predictive power is found using a process known as training. Training includes the presentation of real examples of the features to the model along with corresponding targets. The model is then allowed to find a configuration to better describe the relationship such that it becomes better at classification or regression.

Models can suffer from limitations. Typically, a model will not perfectly predict the targets based on the features. Systematic differences between model predictions and expected targets are known as bias. Bias can be reduced by training but carries a risk: If the training is allowed to progress too far, or there is too much flexibility in the model state, then overfitting can occur. Overfitting means that while the model is able to predict the targets based on the features in the data used to train the model, the predictions based on unseen data not in the training set will be limited and will generally have a larger bias. This is known as variance in the model. It can arise because the training data do not cover the whole space of features or the targets used for training contain errors. The best models will combine low bias with low variance for the most flexible and accurate predictions.

8.6.1 Various ML Algorithms

There is much literature available on the various types of ML algorithm and we do not propose to discuss these in detail here. However, we will briefly introduce a selection of approaches that are used to deliver results, and which can play a part in optimization and automation of wireless communication networks.

Parametric models are used when there is some knowledge about the form of the function underlying the relationship between features and targets. For example, some targets are related to features with linear relationships. Others have higher order polynomials or cyclical relationships such as some trigonometric functions. In this case the model can be chosen and training the model involves finding the values for the parameters that best describes the relationship between the features and the target.

In contrast to parametric models, non-parametric models do not depend on detailed knowledge about the form of the function underlying the relationship between features and targets. Rather the model is allowed, through training, to discover this form itself, along with the associated parameters and other components of state. The terminology can be confusing; non-parametric models do typically have parameters. It is just that non-parametric models allow more freedom in the relationship. A parametric model in the form of a fourth-order polynomial will model a quartic relationship between features and targets very well. It will do this with only five parameters. However, that model will generally be poor at modeling a cyclical relationship, or a logarithmic relationship, or a relationship that is a hybrid of these. A suitable non-parametric model could learn the relationship between features and targets for a wide range of function classes. Neural networks are an example of non-parametric models. They comprise discrete simple units of computation, linking their inputs to their outputs by simple algebraic manipulation and simple functions. These units are connected, sometimes in multitudes. Although each discrete computation unit has very simple predictive power, together the system of units has emergent power to model arbitrarily complex systems. In this sense it mimics the construction and operation of the human brain, although to date no neural network has come close to the scale or complexity of the organ by which it was inspired. The neural network generally has many parameters, typically comprising weights and biases, associated with it. Weights control how important each input is in each unit of the network. Biases offset the output of a computation unit to boost or suppress that part. These parameters must be optimized as part of the training process using a mechanism called back-propagation.

We turn our attention from neural networks to decision trees. These are ML models that repeatedly partition the vector space of features into finer and finer portions. Each portion is then assigned to portions of the target space. Decision trees can be used for classification or regression. The partitions are typically made by logical comparison involving components of the feature space. The number of decisions that are made gives the model increasingly discriminative power, although allowing too many decisions, or excessive depth to the model, risks overfitting the model and increasing variance.

Reinforcement learning is a way to allow a computer agent to interact with a system and learn how to achieve a goal. It has been used as the basis for Google's AlphaGo system for playing the ancient Chinese game Go. By making exploratory moves in a game or changes in a system where the objectives are known, the reinforcement learner is able to develop a strategy for what is more successful at achieving an objective and what is less successful. Reinforcement learning could be considered to deviate from pure ML in the sense that in practice it comprises several discrete models internally for delivering the capability. However, externally it makes predictions about what is the best change to make given the system state in order to maximize the chance of achieving the goal. In that sense it is a predictive machine learning model.

In systems that have many variables, or features, building models able to accept all these features as inputs presents challenges. Typically features have varying degrees of importance and the importance of some may be very small or even vanish. This can cause the model to be overly complex and expensive to train or evaluate. One solution to this is dimensionality reduction. This can include sensitivity analysis to determine which are the least important features that give negligible predictive power so they can be eliminated from the model. Even when features that do not contribute to the predictive power of the model are eliminated, the remaining features could be found to have more dimensions than necessary. In this case transformation to a lower-dimensioned vector space, typically using linear translations, can be performed. The outputs of the transformation can be used as features instead of the untranslated features without impacting the predictive power of the resulting models. Principal component analysis is an example of dimensionality reduction that finds an optimal set of orthogonal vectors for mapping the data, which maximizes the variance between the dimensions. Auto-encoding is another method for dimensionality reduction that uses neural networks to map the full dimensional feature data to a smaller dimensional layer in the neural network and then map it back again to the full dimension. It is trained to approximate the identity function and so to predict its own inputs. If it can do this sufficiently accurately while squeezing the features through a lower dimensionality layer, then the network layers beyond the compression layer can be discarded and the output of the lower dimensionality layer can be used as a dimensionality reduction layer.

For a non-parametric model such as a neural network to perform well and have good predictive power, it must be sufficiently complex and flexible to represent the underlying relationship. However, excessive complexity will not increase the predictive power but will add to training and prediction cost and is therefore undesirable. While a neural network has weights and biases that are optimized as part of the training process, it also has other characteristics controlling its architecture. These are not trained by back-propagation but must be configured when the network is constructed, including the number of layers, how many components comprise each layer, whether they include convolutional layers, and other aspects. These characteristics are typically referred to as hyperparameters. Optimal values of these can be found in a variety of ways. Intuition and experimentation by the network designer are one such method, but a systematic approach to discover the best ways to construct the network can also be employed. This is referred to as hyperparameter optimization. It is also possible to build neural networks that are able to change their architecture as part of training and operation in order to adapt to the specific challenges of the features being modeled. Inspired in part by the brain, these are known as self-organizing neural networks.

8.6.2 Models for Wireless Communications

Having introduced a selection of ML algorithms, here we explore how to bring the theory of ML modeling into the domain of wireless communications. ML gives us the power to build predictive models, but what models can we aspire to build to bring value? We can think in terms of the wireless communication network being a system with stimulus and response. In this case the stimulus is the demand placed on the network by the variety of users. These users are from a variety of subscriber classes or network slices. They are attempting to access a variety of services in various locations at various times. The wireless network itself will have a state. This includes the physical resources available, along with any configuration state and any impairments in terms of malfunctioning infrastructure, software, or transport links.

The network experiences the demand for services placed on it by the subscribers. Moreover, the network has a state. Together the demand and the state will result in a response by the network. At a fundamental level the response will be the signal strength and ratio of signal-to-interference experienced by each subscriber device at any instant along with whether they can access the network at all (and if so, for what services and whether they experience a disruption in service). The response will also include which performance metrics they experience such as data rate, packet loss rate, delay, and jitter. The response will include more abstract aspects such as how satisfied the subscribers are with the service they receive. These are harder to measure as well as to model. The response will also include more indirect characteristics such as the revenue generated from the subscribers, how likely the subscribers are to cancel their subscription, and the energy consumed in operating the system.

We can aspire to model these different aspects of the wireless network: the demand stimulus, the state, and the response. We can seek to model them in isolation. We can also seek to build models that capture the essence of the relationships between these aspects. For example, a model of the geographical distribution of demand in isolation can be used to understand whether the demand is anomalous at any instance. We can also model the relationship between the geographical distribution of demand and the resulting response of the network. A more sophisticated model would also include the current state of the network when predicting the response given the stimulus.

As we have seen, we can model the various characteristics of the network and their interactions. But how do we use these to perform functions that deliver business value? Models alone cannot do this; we need the branches of analytics that they enable to deliver value. We explore these next.

8.6.3 Analytics to Deliver Value from ML Models

We have seen how we can build models of the characteristics of a wireless network and the relationships among different aspects. To understand how ML models deliver value, it is useful to think in terms of the various branches of analytics. The simplest level is descriptive analytics. Often implemented with an ML model, descriptive analytics is typically used to recognize a situation or phenomenon. For example, descriptive analytics can be employed to recognize that congestion is occurring or that UEs are experiencing poor signal strength, or that demand for the service is unusually high. The problem of determining how the subscribers are geographically distributed is also an example of descriptive analytics. Descriptive analytics will recognize various situations, phenomena, or other characteristics, but they will not reveal why these situations are arising. For this we need diagnostic analytics. Diagnostic models will assign a root cause to a phenomenon. For example, descriptive analytics may tell us that congestion is occurring, but diagnostic analytics will help us to define if this is because there are more subscribers than usual or because a transport link is impaired and carrying less data than normal, or alternatively because a gNB has failed for example.

Another stage of analytics is predictive analytics, which helps us to anticipate what will happen in the future. If we develop a model that is able to predict that the physical fiber that conveys a transport link will fail within the next month, or that demand will be fifty percent higher than normal in a geographical area in an hour, then these are examples of predictive analytics.

Another facet of predictive analytics is the ability to estimate what will occur as the result of a change prior to that change being made. One example of this is predicting if there will be sufficient coverage to support certain applications in a specific area as a result of changing some parameters in the network.

Bringing these together, we can identify problems we have now and problems we will have in the future with descriptive analytics and predictive analytics respectively. Furthermore, we can determine the root cause of these problems with diagnostic analytics, and then anticipate what might happen as a result of changing the network, again with predictive analytics. This is a lot of information and power afforded us by our models, but we still need to know what to do. This is where prescriptive analytics comes in. Prescriptive analytics takes information about the system along with knowledge of the policies and objectives of what is desired to be achieved and yields a course of action. With the constraints of the model accuracies and completeness of the information available, the course of action recommended is what should be done to maximize the chance of the objectives being achieved most completely. It is prescriptive analytics that underpins the intent-based optimization mentioned earlier and intent-based orchestration introduced in Chapter 6.

8.6.4 Applications of Machine Learning in Telecommunications

Above we introduced some examples of ML models and associated algorithms. We showed how models of wireless communication and associated infrastructure can be envisaged along with the demand placed on them by the users and how they respond in terms of services and other outcomes. We have also seen how models can be employed to deliver different types of analytics. Now we introduce and discuss classes of use cases for wireless communications that are enabled by machine learning.

Detection of anomalies is a key class of use cases for machine learning. Anomalies include changing behavior or characteristics. Anomalies can be indicative of an underlying problem or impairment with the network, or they may indicate a change in the way the network is being used by the subscribers or how it is responding to the demand. Detecting an anomaly is not the same as diagnosis of why the anomaly has occurred, which we cover later, but it can be a powerful vehicle to understanding network dynamics. Underlying issues will manifest as anomalies and alert the operator to the existence of problems. Here the operator could be an engineering team or a control system responsible for automated operational management of the network. Once the anomaly is detected it will open the door to understanding the root cause, triaging, and ultimately resolving any problems thus uncovered.

Potentially, many anomalies can be discovered with machine learning anomaly detection algorithms. When the normal patterns are understood, or modeled, then anomalies can be found in terms of deviations from that modeled behavior. The characteristics that can

Table of Contents

be modeled include what services and applications are being used on the network, where they are being used, and by what network slices. Anomalies can be found in the state of the radio: Where coverage is good or poor will have a natural cycle from which it can deviate. Interference can similarly be nominal or anomalous. The times and locations when and where different subscribers are able to connect to the network will have their own characteristics. Blocking of subscribers is a strategy to avoid overloading the network; this will also have its own characteristic patterns which may be normal or anomalous.

The behavior and state of the transport network can be measured, and anomalies detected. For example, measures such as data volume carried, packet loss rate, retransmissions, jitter, and delay will have characteristics behavior from which they can deviate. Core network functions and applications also exhibit anomalies. This can be in terms of performance measures such as response times, resource utilization, or physical locations of functions. The behavior and characteristics of the physical infrastructure on which virtual functions are hosted can also exhibit anomalies.

Anomalies can be detected on different timescales. Some need to be detected and reacted to in near-real time to avoid serious adverse impacts. An example is congestion of a transport link. If the link is congested, subscribers will be experiencing poor performance. If these are important subscribers or they are using a critical communication network slice, the operator will want to diagnose the problem and resolve it as quickly as possible.

Detecting anomalies in near-real time is more demanding for a machine learning anomaly detection algorithm. It must be able to characterize behavior as anomalous with very few samples spanning a very short time, sometimes much less than a second. A retrospective analysis is more discerning and will be able to detect more subtle anomalies. While this type of anomaly detection cannot allow instantaneous anomalies to be reacted to, this non-real time analysis can also have value. For example, a failing piece of network hardware, or a network function nearing capacity, may exhibit transient performance changes as a leading indicator of future failure. Non-real time anomaly detection can be a useful tool for detecting this.

While the detection of anomalies is important, it is generally followed by diagnosis as this will allow the underlying cause to be understood and potentially resolved. Here machine learning can assist us. Some anomalies are their own diagnosis. For example, patterns of what applications are being used by which subscribers in what locations can have patterns that are subject to change and anomaly. But other characteristics are indicative of an underlying issue or problem that can be diagnosed. An increased packet loss rate on a transport link can have various causes. It may be indicative of more demand on the network by subscribers, or it could be due to degrading of the fiber that conveys the transport. This in turn could be caused by a fiber connector performing poorly. Alternatively, a network switch could be degrading. The best action to take to resolve the packet loss issue will depend on which of these potential triggers is causing it.

Another example of an effect with multiple potential causes is the radio interface performance. To achieve a high data rate depends on the signal-to-noise ratio (SNR) being high enough to support the more spectrally efficient modulation and coding schemes along with the higher order MIMO that will allow more data to be multiplexed over the same channel. SNR falling at particular locations or times has various causes. It can happen when the ability of the signal to penetrate to the subscribers is impaired, e.g., when an object occludes the signal. It could also be due to a transient phenomenon such as a vehicle parking in the line-of-sight between the transmitter and the receiver. It could be caused by a seasonal cycle such as foliage on trees. It may experience a more permanent degradation resulting from new construction that impedes the signal transmission. Another cause of lowering SNR is an increase in interference. This can happen when signals on nearby beams intended for other subscribers are attenuated less than they were previously, or are transmitted with more power, or utilize more of the spectral resources than they had previously. How this loss of SNR is addressed will of course depend on the underlying cause. Machine learning models can be trained to detect these underlying causes. This can be done directly, so that the characteristics of each cause can be recognized directly. Alternatively, some root causes can be diagnosed by creating and training models for normal operation of each aspect of the network. Then those anomalies that are correlated together in time can be found and organized into causal relationships.

Subscribers must be able to access the network initially when the device is powered on if they are to be able to achieve a service. While anomalies in terms of what proportion of subscribers are able to access the network at different locations and times can be found, there are many and varied potential root causes for this. These range from poor SNR discussed above, with all the potential root causes that this can entail, to constraints of admission control. Problems with the transport network can also be the culprit as can the core network functions such as the access and mobility management function (AMF). The virtualized function may themselves be the problem, or the physical infrastructure on which they reside. Again, machine learning models can support troubleshooting by finding correlated anomalies or directly recognizing the patterns of behavior that relate to specific diagnoses.

Anomaly detection will find situations potentially indicative of a poor performance or other adverse issues either during the issue or after it has taken place. There is also value in predicting when adverse situations will happen in the future. For example, prediction of future demand on the network will allow the network capacity to be planned accordingly. New demand can lead to increased congestion or interference and reduced coverage. New sites, cells, or carriers can be added in anticipation of the growing demand for data for new services and increases in subscriptions. The adverse impact on services can thus be avoided. Modeling the cyclical aspects of the demand on daily, weekly, and seasonal cycles along with the secular trend can be coupled with models of the relationship between new services and the corresponding change in demand for data.

The transport network is critical to the integrity of the network. Service in parts of the network can be lost if a break occurs in a physical fiber. If a fiber passes through ground subject to subsidence, this can cause strain and result in partial failure or a complete break. Predicting these ahead of time using ML models allows the problem to be pre-empted and the fiber replaced. Models optimize this activity by allowing the replacement to be performed in time to avoid disruption but not so early that there is an unnecessary increase in operation expenditure.

Just as the physical fiber can degrade and fail, so too can the computer platforms and other physical assets in a virtualized network. Some parts will generally not be virtualized, such as the radios and high-performance signal processing hardware for generation of the modulated radio. Prediction of when these will fail through hardware faults in the future can allow them to be replaced or repaired prior to an interruption in service. Again, ML models can learn the leading indicators that will predict what will fail and when it will fail.

As operators become more reliant on revenue from delivery of services to industry verticals with strict SLAs attached, so there is a need to predict when these SLAs will be infringed, because they come with financial penalties and reputational costs. Similarly, prediction of when and how requirements imposed by the regulator will be infringed is critical to minimizing the damage and costs associated with infringement.

If a UE has a GPS or global navigation satellite system (GNSS) module, it can measure its own location. There are mechanisms in the mobile standards to support the UE to make location estimates using satellite location systems or other techniques that do not rely on satellites. There are also mechanisms in the mobile standards to allow location fixes to be reported to the network. These location fixes can be used for various applications such as responding to emergency calls or limiting service to certain geofenced areas. But these estimates of location, coupled with measurements of the characteristics of the radio network at that location, can be valuable for managing and optimizing the performance of the radio network. Some locations, such as within buildings or in very dense urban areas with high-rise buildings, can have poor visibility of the GNSS satellites. Some types of UE lack the GNSS subsystems and cannot make their own location fixes. Additionally, the number of mobile connections when the mobile reports its location routinely can be a tiny proportion of all the mobile connections. Increasing the number of devices and connections where the UE is requested to report its location is possible but has limitations. Running the GNSS subsystem will increase the drain on the UE battery. Reporting the location to the network will use some of the network capacity, leaving a little less for use data. These constraints limit the visibility of the geographical distribution of network performance and can bias that distribution to certain locations or device types. A solution to this is to estimate the location of devices without relying on GNSS fixes. This can be achieved using measurements that the UE routinely reports to the network as part of the normal management of radio resources. These measurements include signal strength and timing information. However, these measurements are not designed for location estimation. But they can be correlated with the location of the UE. This geolocation approach is heavily dependent on machine learning, because measurements for radio resource management and external data must be combined and correlated with geographical location. As such, UE location can be estimated even when there are very low levels or no GNSS locations reported.

Once locations of the UEs can be calculated, the geographical distributions in the radio conditions can be measured and correlated with other factors such as terrain, building density, etc. ML models can be created using this information to answer questions about how to expand the network, for example. The capacity that can be achieved with cells using MIMO and beamforming will depend in part on the environment in which they are installed and what locations are chosen. Thus, ML models can support the decision of where to place MIMO cells in order to balance the system performance with return on investment.

We have introduced various concepts where descriptive analytics is used for detecting anomalies, diagnostic analytics can rapidly identify the root cause of an issue, and predictive analytics can tell us about an adverse event prior to it becoming an issue. Now we turn our attention to prescriptive analytics: what to do to deliver the best performance that satisfies the requirements placed on the network. Various aspects must be optimized to achieve this. Coverage optimization encompasses optimization of the coverage beams and static user beams to ensure that the network can be accessed in the required locations while also meeting any targets for beam redundancy, for example. ML can help us here. Models for what performance would be achieved as a result of changing various parameters can be built. These can model the impact of parameters such as transmit powers and beam parameters, such as direction. The models can predict coverage along with other targets of interest such as interference, capacity, beam redundancy, spectral efficiency, and data rates, along with how these vary by location. Such models can then be used to underpin a system to find combinations of parameter configurations that best achieve what the network is being asked to do. Thus, the effectiveness of parameterizations can be evaluated prior to being deployed to the network. Thus, the parameterization predicted to best meet the needs can be chosen.

8.6.5 Network Slicing and Intent-Based Optimization

As models become more sophisticated, and as network slicing becomes more of the normal choice to provide competing requirements within the same infrastructure, more aspects of the network parameterization choices and the corresponding performance will be required as inputs and outputs for these models. Some of these will include choices such as the numerology, choice of cyclic prefix, RACH channel configuration, and configuration of BWPs. These will consider the variations in UE capabilities of the device population including which numerologies are supported and what channel bandwidth can be accessed on which carriers, so that the parts of the spectral resources that can be accessed by which devices is modeled. Thus, the overall composite capacity achievable for the mix of devices and subscribers over multiple network slices will be resolved by the model. The most sophisticated models will not only model the dynamics of UEs as they move around a network on a specific spectral resource; they will complement this with model components that include the interactions between network layers. Network layer modeling will include which carriers and BWPs are used to deliver service to each class of subscribers in the various locations. This will also include the interworking between LTE and 5GNR carriers in dual connectivity mode, respecting the system parameters that control the management of spectral resource layers.

Such powerful models will capture more faithfully the complex system of interactions that is the NR and LTE radio interface and will underpin optimization of the many parameters and choices that can be tuned in the next-generation networks. The resulting networks will utilize the physical resources more efficiently and deliver the best balance of coverage and capacity for each class of subscriber on each network slice.

These AI and ML and models that possess ever more powerful predictive capability will be in the ascendancy, as initiatives such as disaggregation in the RAN, service-based architecture in the core, and the O-RAN Alliance open up the network into more discrete components. As these components become more programmable, so will the data that they expose to fuel the next generation of advanced AI models. These open programmable networks and the transport networks that fuel them, when combined with sophisticated models, optimization, and prescriptive analytics, will be a potent combination. Drawing on many areas of advanced technology, with autonomic monitoring, self-regulation, and intelligent adaptability, we will have the most sophisticated hybrid digital and physical systems ever created by humans. From these new artifacts will emerge a communication system that delivers a richness of experience with breadth and depth well beyond what we dare to imagine today.

List of Abbreviations

3GPP	3rd Generation Partnership Project
5G CN	5G Core Network
5QI	5G QoS Identifier
5GS	5G System
5G-SIG	5G Special Interest Group
5GTF	5G Technical Forum
8PSK	8-Phase Shift Keying
AAA	Authentication, Authorization, and Accounting
A-bis	GSM interface between BTS and BSC
AF	Application Function
AI	Artificial Intelligence
A-int	GSM interface between BSC and MSC
AIS	Alarm Indication Signal
AMF	Access and Mobility Management Function
AP	Application Part
API	Application Programming Interface
APN	Access Point Name
APTS	Assisted Partial Timing Support
APTSC	Assisted Partial Timing Support Clock
AR	Augmented Reality
ARP	Allocation and Retention Priority
ARPU	Average Revenue Per User
ARQ	Automatic Repeat reQuest
AS	Access Stratum
ASIC	Application-Specific Integrated Circuit
ATM	Asynchronous Transfer Mode
AUSF	Authentication Server Function
AWS	Amazon Web Services
BA	Bandwidth Adaptation
BBU	Baseband Unit
BC	Boundary Clock
BCCH	Broadcast Control Channel

BF	Basic Frame
BGCF	Breakout Gateway Control Function
BMCA	Best Master Clock Algorithm
BPSK	Binary Phase Shift Keying
BSC	Base Station Controller
BSS	Base Station Subsystems
BSS	Business Support System
BTS	Base Transceiver Station
BWP	Bandwidth Part
C&M	Control and Management
CapEx	Capital Expenditure
CAPIF	Common API Framework
CA-Polar	CRC-Aided Polar
CBG	Core Block Group
CC	Component Carrier
CCF	Converged Control Plane Function
CDF	Charging Data Function
CDMA	Code Division Multiple Access
CEM	Customer Experience Management
CFO	Chief Financial Officer
CDF	Charging Gateway Function
CIoT	Cellular Internet of Things
CMC	Cell Maximum Coverage
CN	Core Network
CODEC	Coder-Decoder
CoMP	Coordinated Multipoint
CORD	Central Office Re-architected as a Datacenter
CORESET	Control-Resource Set
CoS	Class of Service
COTS	Common Off-The-Shelf
CP	Control Plane
CPE	Customer Premises Equipment

CPU	Central Processing Unit
CP-OFDM	Cyclic Prefix-Orthogonal Frequency Division Multiplexing
CPRI	Common Public Radio Interface
CQI	Channel Quality Indicator
CRAN	Centralized Radio Access Network
CSGN	CIoT Serving Gateway Node
CRC	Cyclic Redundancy Check
CRI	CSI-RS Resource Indicator
CS	Circuit Switching
CSCF	Call Session Control Function
CSFB	Circuit-Switched Fallback
CSG	Closed Subscriber Group
C-SGN	CIoT Serving Gateway Node
CSI-RS	Channel-State Information Reference Signal
cTE	Constant Time Error
CU	Central Unit
CUPS	Control and User Plane Separation
CUT	Clock Under Test
CWDM	Coarse Wavelength Division Multiplexing
DAS	Distributed Antenna System
DC	Dual Connectivity
DCCH	Dedicated Control Channel
DCH	Data Channel
DCI	Downlink Control Information
DECOR	Dedicated Core Network
DFT-S-OFDM	Discrete Fourier Transform-Spread-OFDM
DHCP	Dynamic Host Configuration Protocol
DL	Downlink
DM-RS	Demodulation Reference Signal
DN	Data Network
DNN	Data Network Name
DNS	Domain Name Service

DPCH	Dedicated Physical Channel
DPDK	Data Plane Development Kit
DRX	Discontinuous Reception Cycle
DSL	Digital Subscriber Line
DU	Distributed Unit
DWDM	Dense Wavelength Division Multiplexing
E2E	End-to-End
EC-GSM-IoT	Extended Coverage GSM for IoT
eCPRI	Evolved Common Public Radio Interface
eDECOR	Evolved Dedicated Core Network
EDFA	Erbium Doped Fiber Amplifiers
EDGE	Enhanced Data Rates for GSM Evolution
eDRX	Extended Discontinuous Reception Cycle
ECGI	E-UTRAN Cell Global Identifier
ECM	EPS Connection Management
ECOMP	Enhanced Control, Orchestration, Management, and Policy
EDGE	Enhanced Data Rates for GSM Evolution
eDRX	Extended Discontinuous Reception Cycle
EEC	(Synchronous) Ethernet Equipment Clock (ITU-T G.8262)
eICIC	Enhanced Inter-Cell Interference Coordination
EIRP	Effective Isotropic Radiated Power
eMBB	Enhanced Mobile Broadband
EMM	EPS Mobility Management
eNB	Evolved Node B
EN-DC	E-UTRAN New Radio – Dual Connectivity
ENIAC	Electronic Numerical Integrator and Computer
EPC	Evolved Packet Core
EPS	Evolved Packet System
eRE	eCPRI Radio Equipment
eREC	eCPRI Radio Equipment Control
ESM	EPS Session Management
ESMC	Ethernet Synchronization Message Channel

ESN	Emergency Services Network
ETSI	European Telecommunications Standards Institute
E-UTRA	Evolved Universal Terrestrial Radio Access
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
EVM	Error Vector Magnitude
FACH	Forward Access Channel
FDD	Frequency Division Duplex
FFT	Fast Fourier Transform
FOMA	Freedom of Mobile Multimedia Access
FPP	Floor Packet Percentage
FQDN	Fully Qualified Domain Name
FR	Frequency Range
FTTH	Fiber to the Home
FUT	Function Under Test
FWA	Fixed Wireless Access
GBR	Guaranteed Bit Rate
GCF	Global Certification Forum
GDP	Gross Domestic Product
GFBR	Guaranteed Flow Bit Rate
GGSN	Gateway GPRS Support Node
GHz	Giga Hertz
Gi	Interface between GGSN and PDN to convey packet data
Gn	Interface between SGSN and GGSN to support mobility
gNB	5G NodeB
GNSS	Global Navigation Satellite System
GPRS	General Packet Radio Services
GPS	Global Positioning System
GR	Group Report
GS	Group Specification
GSM	Global System for Mobile Communications
GSMA	Global System for Mobile Communications Association
GTP	GPRS Tunneling Protocol

GW	Gateway
GWCN	Gateway Core Network
HARQ	Hybrid-Automatic Repeat reQuest
HDLC	High-level Data Link Control
HETNET	Heterogenous Network
HLR	Home Location Register
HLS	High Layer Split
HRM	Hypothetical Reference Model
HSCSD	High-Speed Circuit Switched Data
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High-Speed Uplink Packet Access
HTTP	Hypertext Transfer Protocol
HVAC	Heating, Ventilation, and Air Conditioning
IC	Integrated Circuit
ICP	Internet Content Provider
ID	Identity
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IFFT	Inverse Fast Fourier Transform
IMEI	International Mobile Equipment Identity
IMS	IP Multimedia System
IMSI	International Mobile Subscriber Identity
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
IPR	Intellectual Property Rights
IP-SM-GW	Internet Protocol Short Message Gateway
IPT	Internet Protocol Testing
IPv4	Internet Protocol version 4
IQ	In-Phase Quadrature

ISG	Industry Specification Group
IT	Information Technology
ITU	International Telecommunication Union
IUT	Implementation Under Test
IWF	InterWorking Function
IWMSC	InterWorking Mobile Switching Center
IWK-SCEF	Interworking SCEF
JSON	JavaScript Object Notation
KHz	Kilo Hertz
KPI	Key Performance Indicator
KT	Korea Telecom
LAA	License Assisted Access
LADN	Local Area Data Network
LAN	Local Area Network
LBT	Listen-Before-Talk
LCP	Logical Channel Prioritization
LEA	Law Enforcement Agency
LI	Lawful Interception
LLS	Low Layer Split
LOL	Laugh Out Loud!
LPWAN	Low-Power, Wide-Area Network
LS	Location Service
LTE	Long-Term Evolution
LTE-M	Long-Term Evolution Machine Type Communication
MAC	Medium Access Control
MANO	Management and Orchestration
Max|TE|	Maximum Absolute Time Error
MB	Megabyte
MBB	Mobile Broadband
MBMS	Multimedia Broadcast Multicast Service
Mbps	Megabits Per Second
MCPTT	Mission-Critical-Push-To-Talk

MDBV	Maximum Data Burst Volume
MEC	Multi-Access Edge Computing
ME-LDPC	Multi-Edge Low-Density Parity-Check Code
MEO	MEC Orchestrator
MEP	Multi-Access Edge Platform
MEPM	MEC Platform Manager
MFBR	Maximum Flow Bit Rate
MGCF	Media Gateway Control Function
MGW	Media Gateway
MIMO	Multiple-Input Multiple-Output
MIoT	Massive Internet of Things
MMIMO	Massive Multiple-Input Multiple-Output
MISO	Multiple Input, Single Output
ML	Machine Learning
MME	Mobility Management Entity
MMI	Man-Machine Interface
MHz	Mega Hertz
mMTC	Massive Machine Type Communication
mmWave	Millimeter Wave
MO	Mobile Originated
MOCN	Multi-Operation Core Network
MORAN	Multi-Operation Radio Access Network
MPLS	Multiprotocol Label Switching
MSC	Mobile Switching Center
MT	Mobile Terminated
MTC	Machine Type Communication
MTIE	Maximum Time Interval Error
MU-MIMO	Multi-User Multiple-Input Multiple-Output
MVNO	Mobile Virtual Network Operators
N1	Interface between the 5G UE and the AMF
N6	Interface between the 5G UPF and the data network
NAS	Non-Access Stratum

NB-IoT	Narrowband Internet of Things
NEF	Network Exposure Function
NEM	Network Equipment Manufacturer
NF	Network Function
NFCI	Network Function Cloud Infrastructure
NFProfile	Network Function Profile
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
NG	Next Generation
NGC	Next Generation Core
NGMN	Next Generation Mobile Networks
NG-RAN	Next Generation Radio Access Network
NIDD	Non-IP Data Communication
NOC	Network Operations Center
NOMA	Non-Orthogonal Multiple Access
NR	New Radio
NRF	Network Repository Functions
NSA	Non-Standalone
NSI	Network Slicing Instance
NSSAI	Network Slice Selection Assistance Information
NSSF	Network Slice Selection Function
NTP	Network Timing Protocol
OAI	OpenAPI Initiative
OAS	OpenAPI Specification
OC	Ordinary Clock
OCP	Open Compute Project
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
ONAP	Open Network Automation Platform
OpEx	Operational Expenditure
O-RAN	Open-Radio Access Network
OSI	Open Systems Interconnection
OSM	Open Source MANO
OSS	Operations Support System
OTN	Optical Transport Networking
OTT	Over The Top
OTTP	Over The Top Players
PA	Power Amplifier
PAPR	Peak to Average Power Ratio
PBCH	Physical Broadcast Channel
PC	Personal Computer
PCF	Policy Control Function
PCH	Paging Channel
PCI	Physical Cell Identity
PCRF	Policy and Charging Rules Function
PCS	Direct interface between UEs
PDB	Packet Delay Budget
PDC	Personal Digital Cellular
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDH	Plesiochronous Digital Hierarchies
PDN	Packet Data Network
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Units
PDV	Packet Delay Variation
PEC-M	Packet-Based Equipment Clock Master
PEC-S-F	Packet-Based Equipment Clock Slave Frequency
PER	Packet Error Rate
P-GW	Packet Gateway
PHY	Physical Layer
PLMN	Public Land Mobile Network
PMI	Precoding Matrix Indicator
PNF	Physical Network Functions
PNT	Packet Networking Timing

PNT-F	Packet Networking Timing Function
PON	Passive Optical Network
PRC	Primary Rate Clock
PRTC	Primary Reference Telecom Clock
PS	Packet Switched
PSM	Power Saving Mode
PSS	Primary Synchronization Signal
PSTN	Public Switched Telephone Network
PTP	Precision Timing Protocol
PTT	Push-To-Talk
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QFI	QoS Flow ID
QL	Quality Level
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QUIC	Quick UDP Internet Connection
RAB	Radio Access Bearer
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RCSe	Rich Communications Services Enhanced
RD	Retained Data
RDI	Remote Defect Indication
RE	Resource Element
REC	Radio Equipment Control
REST	Representational State Transfer
RF	Radio Frequency
RFC	Request For Comment
RFI	Requests-For-Information
RI	Rank Indicator
RIC	RAN Intelligent Controller
RLC	Radio Link Control
RNC	Radio Network Controller
RNI	Radio Network Information
RoE	Radio over Ethernet
RoHC	Robust Header Compression
RPU	Revenue Per Unit
RQA	Reflective QoS Attribute
RRC	Radio Resource Control
RRH	Remote Radio Head
RRU	Remote Radio Unit
RTT	Round Trip Time
RU	Radio Unit
S1	Interface between the e-UTRAN and the S-GW
SA	Standalone
SBA	Service-Based Architecture
SBI	Service-Based Interfaces
SCEF	Service Capability Exposure Function
SCS	Service Capabilities Server
SD	Slice Differentiator
SDAP	Service Data Application Protocol
SDH	Synchronous Digital Hierarchy
SDI	Software Defined Infrastructure
SDN	Software Defined Networking
SDO	Standards Development Organization
SDU	Service Data Unit
SFI	Slot Format Indication
SGi	Interface to the internet from the PDN
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SIMO	Single Input, Multiple Output

SISO	Single Input, Single Output
SLA	Service Level Agreement
SLF	Subscriber Location Function
SME	Session Management Entity
SMF	Session Management Function
SMS	Short Message Service
SNR	Signal-to-Noise Ratio
S-NSSAI	Single Network Slice Selection Assistance Information
SON	Self-Organizing Networks
SONET	Synchronous Optical Network
SR	Scheduling Request
SRB	Signaling Radio Bearer
SRE	Site Reliability Engineering
SRS	Sounding Reference Signal
SRVCC	Single-Radio Voice Call Continuity
SSB	Synchronization Signal Block
SSC	Session and Service Continuity
SSM	Synchronization Status Messages
SSS	Secondary Synchronization Signal
SST	Slice/Service Type
STA	WLAN Station
S-TMSI	Serving Temporary Mobile Subscriber Identity
STTD	Space Time Transmit Diversity
SU-MIMO	Single-User Multiple-Input Multiple-Output
SUT	System Under Test
SyncE	Synchronous Ethernet
TA	Timing Advance
TAI	Tracking Area Identity
TAE	Time Alignment Error
T-BC	Telecom Boundary Clock
TB	Transport Block
TBF	Temporary Block Flow

TC	Transparent Clock
TCE	Trace Collection Entity
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TDEV	Time Deviation
TDF	Traffic Detection Function
TDMA	Time Division Multiple Access
TE	Time Error
TEID	Tunnel Endpoint Identifier
TETRA	Terrestrial Trunked Radio
T-GM	Telecom Grandmasters
TIP	Telecom Infra Project
T-GM	Telecom Grandmaster
TM	Tele Management
TR	Technical Report
TS	Technical Specification
TSN	Timing Sensitive Network
T-TC	Telecom Transparent Clock
T-TSC	Telecom Time Slave Clock
TTI	Transmission Time Interval
UCI	Uplink Control Information
UDM	Unified Data Management
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UltraHD	Ultra-High Density
UMTS	Universal Mobile Telecommunication System
UP	User Plane
UPF	User Plane Function
URA	UTRAN Registration Area
URLLC	Ultra-Reliable Low Latency Communications
UTC	Coordinated Universal Time

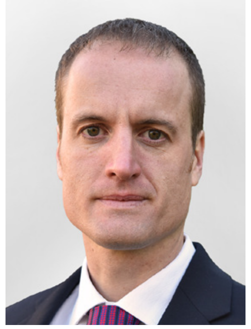
UTRAN	UMTS Terrestrial Radio Access Network
V2N	Vehicle to Network
V2X	Vehicle to Everything
vCPE	Virtual Customer Premises Equipment
vDNS	Virtual Domain Name Service
vEPC	Virtual Evolved Packet Core
vFW	Virtual Firewall
VGCS	Voice Group Call Service
VHS	Video Home System
VIM	Virtualization Infrastructure Manager
VLAN	Virtual Local Area Network
VM	Virtual Machine
VNF	Virtual Network Function
VNFI	Virtual Network Function Infrastructure
VNFM	Virtual Network Function Manager
VoIP	Voice Over IP
VoLTE	Voice Over LTE
VPLMN	Visited Public Land Mobile Network VR Virtual Reality
WAN	Wide-Area Network
WCDMA	Wideband Code Division Multiple Access
WDM	Wavelength Division Multiplexing
WFA	Wi-Fi Alliance
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Local Area Network
YAML	YAML Ain't Markup Language
ZSM	Zero-Touch Network and Service Management

About the Authors

Jonathan Brooksby

Jonathan Brooksby is a twenty-five-year veteran of the communications industry. He currently leads the VIAVI market segment team responsible for combined CORE and RAN Assurance data solutions focusing on mobile service providers worldwide. Jonathan's expertise in telephony, mobile, mobile broadband, LTE, and 5G comes from developing, supporting, marketing, and consulting on major data networking and mobile telecommunications projects.

He is a seasoned spokesperson on many networking issues for a variety of international conferences and forums. Jonathan is a graduate of Heriot-Watt University in Scotland with a master's degree in electrical and electronic engineering.



Walter Featherstone, PhD

Dr. Featherstone served as a Principal Research Engineer at VIAVI Solutions. Prior to joining VIAVI, he worked for Motorola Mobility during Google's ownership, Nokia Networks, and Motorola Solutions. He holds a PhD in Electronic and Electrical Engineering from the University of Leeds, specializing in Super Resolution Direction Finding. Walter is passionate about new fields of research, emerging technologies, and innovation. He has thirteen granted patents. Walter is an active participant in ETSI ISG MEC, where he's been elected to lead the Deployment and Ecosystem Development working group (WG DECODE). The WG focuses on enabling implementation of systems that take advantage of MEC-defined frameworks and provide/consume services using MEC-standardized APIs. Walter is a chartered member of the Institute of Engineering and Technology (IET) and Chairman of the IET's Swindon branch organizing events, presentations, and support for technology related festivals.

Per Kangru

Per Kangru works in the VIAVI Business Development and Technology office, based in Sweden. Per has twenty years of experience supporting leading Telecom and IT companies in their transformation projects. The current work is focused on the evolution to 5G with a major focus on the Automation of the Network and Service Management. Per with VIAVI was one of the founding members of the ETSI ZSM ISG.

Previously Per was involved in building the global success of the 4G (LTE) Technology supporting global operators and vendors in the development, deployment, and optimization of their networks and services. Per actively participates in several industry bodies such as LSTI, TIA/Quest, NGMN, and TM-Forum.

Per is a co-author of a bestselling book on LTE published by Wiley. Per has several patents granted in the area of Telecom Test & Measurement and Assurance.

Eng Wei Koo

Eng Wei Koo served in the VIAVI CTO office and contributed to developing 5G strategy, products, and architecture roadmaps for the VIAVI end-to-end lab verification, field test, and optimization 5G solution portfolio. He represented VIAVI on global organizations which define standards and solutions for 5G, Open RAN, and autonomous driving, and participated in international technology conferences and workshops on advanced and emerging wireless technologies.

Eng Wei has filed several patents and has authored several publications and book chapters in telecommunications.

Chris Murphy, PhD

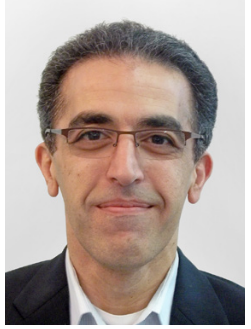
Dr. Murphy has nearly twenty years of commercial experience in telecommunications covering network performance measurement, optimization, and SON, particularly in cellular RAN including LTE, UMTS, CDMA, and 5G. His focus has been on the modelling, simulation, and optimization of next-generation technologies to build a revenue stream early in the lifecycle of each new generation of telecommunication systems. He has contributed to various industry and standardization bodies including 3GPP, NGMN, ATIS, and the WiMAX Forum.

Chris joined VIAVI through the acquisition of Arieso. Prior to that he worked for Motorola and Nokia developing new technical capabilities for the service portfolio.

Chris holds a degree in mathematics and computing and a PhD in the calibration of oceanic remote sensing satellites for improved models of climate change. He has authored various journal articles, conference papers, and book chapters, and has filed eighteen patent applications.

Howard Thomas, PhD

Dr. Thomas is a Telecommunications Consultant with over thirty years of experience in the mobile radio business including contributions on mobile radio propagation research, the creation, design, and 3GPP standardization of features and systems, and research and design of self-organizing-network algorithms. In parallel with this, he has been involved in the creation, management, and enforcement of intellectual property rights, including authoring over sixty patents and acting as an expert witness to the Chancery Division of the High Court in London.



Reza Vaez-Ghaemi, PhD

Dr. Vaez-Ghaemi has held positions in research, engineering, marketing, and management in the telecommunications industry in Germany and the United States. At VIAVI he is responsible for research in emerging technologies and development of product roadmaps for Carrier Ethernet, Mobile Backhaul, Fronthaul, and Optical Transport networks. Dr. Vaez-Ghaemi received his BS, MS, and PhD in electronics and electrical engineering from Technical University of Berlin (Germany).

Sameh Yamany, PhD

Dr. Yamany, Chief Technology Officer (CTO) of VIAVI Solutions, leads long-term technology vision, and is responsible for applied research programs, industry thought leadership, and advanced technology incubations.

Prior to VIAVI, Dr. Yamany was CEO and President of Trendium, a wireless monitoring, probing and assurance company acquired by JDSU (now VIAVI). Trendium products were later integrated into the VIAVI market-leading NITRO™ Mobile platform. His industry experience also includes leading the vision and development of the Iris™ wireless and wireline monitoring and troubleshooting suite at Techtronix Communications.

Dr. Yamany worked as an assistant professor at Old Dominion University in Virginia and has a Doctorate in computer science and engineering (CSE) from the Speed School of Engineering at University of Louisville, Kentucky, a Master of Science and a Bachelor of Science in systems and biomedical engineering from Cairo University, Egypt.

He authored and co-authored several patents, numerous scientific journal papers, conference and industry related publications, and book chapters in artificial intelligence, telecommunications, systems, and biomedical engineering.

Additional Contributors

Paul Gowans and Kashif Hussain

Managing Editor

Laurie Rerko

Table of Contents

Index

Page numbers in *italic* denote figures, page numbers in **bold** denote tables.

2G architecture 6-9, 211
3G architecture 10-12
3GPP Enhanced Coverage 213-214
3GPP standardization 25, 27, 100, 146, 210
  5G New Radio (5G NR) 83, 87-88
  architecture 32-33
4G architecture 13-14, 16, 92-93
5G architecture blueprint 33-35
5G Core (5GC) *131*, *136*, *139*, *142*
5G interfaces, Cellular Internet of Things (CIoT) 216, *217*
5G New Radio (5G NR) 42-46, *48*, *49*, 62-64
  see also massive MIMO
  3GPP standardization 83, 87-88
  beamforming support 76-82
  Code Block Group (CBG) based re-transmission 61-62
  configurable 5G NR MAC layer 55-56
  configuring PDCCH monitoring 58-59
  configuring PDSCH 57-58
  configuring PUCCH duration 59-60
  energy saving support 82-83, 258
  Enhanced Mobile Broadband (eMBB) 84-86
  massive Machine Type Communications (mMTC) 84-86
  Multiple-Input Multiple-Output (MIMO) support 76-82
  network slicing 89-90
  standards evolution 83-87
  Ultra-Reliable Low Latency Communications (URLLC) 24, *25*, 63-64, 83-84
5G NGC and network slicing 145-149
  5G Core network slices topology 146, *147*
  5G network slice identifiers 148
  5G network slice selection 149
  Single Network Slice Selection Assistance Information (S-NSSAI) 148-149
5G technologies life cycle *252*
5G transport networks 111-122
  emerging 5G synchronization requirements 120, **121**
  Enhanced Mobile Broadband (eMBB) 111
  Evolved Common Public Radio Interface (eCPRI) 113-118
  fronthaul bridged network *118*
  fronthaul challenges and functional split options 111-112
  Fronthaul Gateways (FHGW) 121-122
  functional split options 111-113
  Higher Layer Split (HLS) 113-118
  Lower Layer Split (LLS) 113-118
  Massive Machine Type Communications (mMTC) 111
  radio over Ethernet (RoE) 121-122
  service requirements categories 111
  timing accuracy categories (eCPRI transport requirements) **121**
  Timing Sensitive Network (TSN) 118-119
  Ultra-Reliable Low Latency Communications (URLLC) 111
5GS QoS architecture and models 141-145
  5G QoS identifier and characteristics 144, **145**
  5G QoS model 142, *143*
  5G QoS profile 143-144
  PDU connectivity service and PDU session 141
  reflective QoS 145

AI *see* artificial intelligence
Amazon Web Services (AWS), cloud environment 189
anomalies detection, machine learning (ML) 275-278
APTS *see* assisted partial timing support
architecture
  2G architecture 6-9, 211
  3G architecture 10-12
  3GPP standards for architecture 32-33
  4G architecture 13-14, 16, 92-93
  5G architecture blueprint 33-35
ARQ *see* automatic repeat reQuest
artificial intelligence (AI) 269-280
  see also machine learning (ML)
  automation and optimization 262
  definitions 270
  cf. machine learning (ML) 270
aspirational capabilities of 5G 39
assisted partial timing support (APTS) 108, *109*
automatic repeat reQuest (ARQ) 6, 54
  see also Hybrid Automatic Repeat reQuest
automation, zero-touch networks 199-202
automation and optimization 250-280
  5G service enablers 254
  Artificial intelligence (AI) 262
  benefits 254-260
  business drivers 250-254
  capital expenditure (CapEx) 257-258
  centralization 264-265
  cost 253
  cyber security and network security 258-259
  disaggregation 260-261
  dynamic nature of the networks 252
  energy consumption reduction 258
  frictionless service creation 255-256
  intent-driven optimization 259-260
  machine learning (ML) 262
  market in connectivity 255-256
  network flexibility and complexity 261
  network infrastructure 253
  network programmability 261-262
  network slicing 254-255
  network tuning for optimization 265-269
  operational expenditure (OpEx) 256-257
  optimization timescales 263
  regulatory requirements 259

  - Service-Based Architecture (SBA) 262
  - service creation 255-256
  - spatial extent of optimization 264-265
  - stakeholders 253-254
  - technology enablers 260-262
  - user equipment (UE) 264
  - virtualization 262
- AWS *see* Amazon Web Services

- backhaul networks 92-93, *94*
- bandwidth management service, Multi-Access Edge Computing (MEC) 161
- bandwidth parts (BWP) 46, 52
- bandwidth/latency/capacity trade-off 64
- base station *4*
  - protocol layers *5*
- base station controller (BSC) 7, *8*
- base transceiver station (BTS) 7, *8*
- baseband unit (BBU) 93-97, 112, 113
- battery life 22, 23, 58-59, 83, 84
  - Cellular Internet of Things (CIoT) 211-214, 243-244
  - extended Discontinuous Reception Cycle (eDRX) 212
  - Power Saving Mode (PSM) 213
- BBU *see* baseband unit
- beam configuration, network tuning for optimization 265-266
- beamforming 65, 74-82
  - characteristic 74-75
  - digital beamforming 75-76
  - hybrid analog/digital beamforming 75-76
  - cf. Multiple-Input Multiple-Output (MIMO) **81-82**
  - support 76-82
- blank slots 53
- BSC *see* base station controller
- BTS *see* base transceiver station
- business drivers
  - automation and optimization 250-254
  - Cellular Internet of Things (CIoT) 205-210
- BWP *see* bandwidth parts

- capacity/bandwidth/latency trade-off 64
- CapEx *see* capital expenditure
- CAPIF *see* Common API Framework
- capital expenditure (CapEx) 16, 17-18, 19-20, 30-31, 34-35
  - automation and optimization 257-258
  - Open Compute Project (OCP) 179-180
  - reduced or deferred CapEx 257-258
- cash flow, mobile operator 20, 34-35
- Cat M1, Cellular Internet of Things (CIoT) 210, 213, 218
- Cat M1 CIoT Using S1-U Data Transfer or User Plane CIoT EPC Optimization [3] 218-221
- CBG *see* Code Block Group based re-transmission
- CDMA *see* code division multiple access
- Cellular Internet of Things (CIoT) 204-247
  - 2G architecture 211
  - 5G interfaces 216, *217*
  - battery life 211-214, 243-244
  - business drivers 205-210
  - Cat M1 210, 213, 218
  - Cat M1 CIoT Using S1-U Data Transfer or User Plane CIoT EPC Optimization [3] 218-221
  - CIoT ecosystem 206-209
  - CIoT interfaces 215-217
  - CIoT transfer methods 217-242
  - coverage extension 213-214
  - device types 210-211
  - EC-GSM-IoT 211
  - end-to-end IoT service provider *208*
  - end-to-end troubleshooting 243
  - extended Discontinuous Reception Cycle (eDRX) 212
  - implications 242-243
  - indoor deep coverage device placement 244-245
  - IoT coexistence with other services 242-243
  - low-power, wide-area network (LPWAN) 211-212
  - LTE interfaces 215-216
  - management philosophies 243-245
  - managing battery life 243-244
  - markets 205-206
  - Narrowband IoT (NB-IoT) 210, 211, 213
  - NB-IoT Switching Between IP-Based Tunnel and Non-IP Data 242
  - NB-IoT Using CIoT EPS Optimization Data Transfer Using NAS PDU and IP-Based Tunnels [3, 5] 222-235
  - NB-IoT Using Non-IP Data Communication (NIDD) [4, 5] 235-241
  - NB-IoT Using SMS 242
  - network slicing 245-247
  - new infrastructure 242
  - new quality metrics 243
  - opportunities 205-210
  - Power Saving Mode (PSM) 213
  - radio access network (RAN) 211
  - radio efficiencies 242
  - radio quality 242
  - Revenue Per Unit (RPU) 209-210
  - service level agreements (SLAs) 245-247
  - signaling storms and security 244
  - uses 204-210, *205*
- Central Office Re-architected as a Datacenter (CORD) 192-193
- centralization, automation and optimization 264-265
- channel coding 53
- Channel-State Information Reference Signal (CSI-RS) 76, 77, 78, *79*, *80*
  - CSI-RS Resource Indicator (CRI) 78, *79*, *80*
- cloud environment
  - Amazon Web Services (AWS) 189
  - private and public cloud environments 188-189
  - virtualization 188-189

CN *see* core network
coarse wavelength division multiplexing (CWDM) 95
Code Block Group (CBG) based re-transmission 61-62
code division multiple access (CDMA) 6, 10
  wideband CDMA (WCDMA) 10, 12
Common API Framework (CAPIF), Multi-Access Edge Computing (MEC) 168-170
Common Public Radio Interface (CPRI)
  bandwidth inefficiency 111-112
  CPRI bandwidth as a function of bandwidth and antenna ports **112**
  fronthaul networks 97-99
  high level CPRI interface *97*
  hyperframe structure *99*
  options and line rates **98**
competitive advantage, virtualization 176-177
configurable 5G NR MAC layer 55-56
configuring PDCCH monitoring, 5G NR 58-59
configuring PDSCH, 5G NR 57-58
configuring PUCCH duration, 5G NR 59-60
conformance testing, Multi-Access Edge Computing (MEC) 170
Control and User Plane Separation (CUPS) 26-27
control plane (CP) 5
CORD *see* Central Office Re-architected as a Datacenter
core network (CN) 3-5
coverage extension, Cellular Internet of Things (CIoT) 213-214
CP *see* control plane
CPRI *see* Common Public Radio Interface
CRI *see* CSI-RS Resource Indicator
CSI-RS *see* Channel-State Information Reference Signal
CSI-RS Resource Indicator (CRI) 78, *79*, *80*
CUPS *see* Control and User Plane Separation
CWDM *see* coarse wavelength division multiplexing
cyber security and network security, automation and optimization 258-259

DAS *see* distributed antenna system
delay spread 43-46
demodulation reference signal (DMRS), Physical Downlink Shared Channel (PDSCH) 57-58
dense wavelength division multiplexing (DWDM) 95
DevOps 194
digital beamforming 75-76
disaggregation 28-31
  automation and optimization 260-261
distributed antenna system (DAS) 92
DMRS *see* demodulation reference signal
drivers for 5G 18, 20-21
  use-case based service drivers, 5G 22-25
DWDM *see* dense wavelength division multiplexing

EC-GSM-IoT, Cellular Internet of Things (CIoT) 211
eCPRI *see* Evolved Common Public Radio Interface
EDGE *see* enhanced data rates for GSM evolution
edge computing 152-173
  see also Multi-Access Edge Computing (MEC)
  application function 165
  edge computing in 5GS 165-167
  historical context 152-153
  Local Area Data Network (LADN) 166
  QoS and charging 166
  routing and traffic steering 165
  Service-Based Architecture (SBA) 165
  Session and Service Continuity (SSC) 165-166
  standardization 153-156
eDRX *see* extended Discontinuous Reception Cycle
eMBB *see* Enhanced Mobile Broadband
energy consumption reduction, automation and optimization 258
energy efficiency/latency/reliability trade-off 53-64
energy saving support, 5G New Radio (5G NR) 82-83, 258
enhanced data rates for GSM evolution (EDGE) 7
Enhanced Mobile Broadband (eMBB) 23, 24, *25*, 39, 42-43, 50, 64, 65, **82**, 111
  evolution 86-87
  network slicing 123, 147
enhanced multiplexing with wavelength conversion *95*
EPC *see* Evolved Packet Core
ETSI ISG MEC 170-173
ETSI NFV, virtualization 179
ETSI ZSM 200-202
Evolved Common Public Radio Interface (eCPRI) 113-118
  CoS High Latency Requirements for split options E, ID, IID, and IU (eCPRI Transport Networks) **119**
  eCPRI transport requirements (timing accuracy categories) **121**
  message types **117**
  requirements for Class 1 (CPRI) traffic (IEEE 802.1cm) **119**
  requirements for split options E, ID, IID, and IU (eCPRI Transport Networks) **119**
Evolved Packet Core (EPC), vs. Next Generation Core (NGC) 129-130
extended Discontinuous Reception Cycle (eDRX) 58-59, 83
  Cellular Internet of Things (CIoT) 212

FHGW *see* Fronthaul Gateways
flexibility 20-21, 22-24, 63
  network flexibility and complexity 261
flexible subcarrier spacing and numerologies 45-46
frame structure and slots 47-52
frequency bands for 5G 39-42
frequency synchronization standards, synchronization networks 102-107
frictionless service creation 255-256
Fronthaul Gateways (FHGW) 121-122
fronthaul networks 93-99
  Common Public Radio Interface (CPRI) 97-99
  optical layer technology 94-97

General Packet Radio Services (GPRS) 7
global navigation satellite system (GNSS), machine learning (ML) 278-279
global system for mobile communications (GSM) 6-9
gNB and UE processing time capability 56-57
GNSS *see* global navigation satellite system
GPRS *see* General Packet Radio Services
GSM *see* global system for mobile communications
guard bands 45
  spectral efficiency 45

handover parameters, network tuning for optimization 267
HARQ *see* Hybrid Automatic Repeat reQuest
Higher Layer Split (HLS) 113-118
historical context xii-xx, 38
  4G architecture 92-93
  1980s and early 1990s *xiv*, 6-9
  1990s to early 2000s *xiv*, 10-12
  2010s *xiv*, 13-18
  2017 18-21
  edge computing 152-153
  radio access network (RAN) 92-93
  spectrum reuse in wireless access 65
  virtualization 178-180
HLS *see* Higher Layer Split
hybrid analog/digital beamforming 75-76
Hybrid Automatic Repeat reQuest (HARQ) 6, 53-64
  benefits 54-55
  review of NR HARQ v LTE HARQ (with some examples) 62-64
  trade-off 55
  versions 55
hybrid operational models 198-199

idle and inactive state operation and access parameters, network tuning for optimization 268
IEEE 1588v2 standard 100-102
independence of RAN from CN 25-26
indoor deep coverage device placement, Cellular Internet of Things (CIoT) 244-245
integrated antenna systems 93-94
intent-based optimization
  machine learning (ML) 280
  network slicing 280
intent-driven optimization, automation and optimization 259-260
intent-driven orchestration 196-197
Internet of Things (IoT) 9
  see also Cellular Internet of Things (CIoT)
interoperability testing, Multi-Access Edge Computing (MEC) 170-171
IoT *see* Internet of Things

LADN *see* Local Area Data Network
latency/capacity/bandwidth trade-off 64
latency/reliability/energy efficiency trade-off 53-64
layer management, network tuning for optimization 268
life cycle, 5G technologies life cycle *252*
LLS *see* Lower Layer Split
Local Area Data Network (LADN), edge computing 166
location service, Multi-Access Edge Computing (MEC) 161
long-term evolution (LTE) 9, 13-16
  LTE interfaces, CIoT 215-216
Lower Layer Split (LLS) 113-118
low-power, wide-area network (LPWAN), Cellular Internet of Things (CIoT) 211-212
LPWAN *see* low-power, wide-area network
LTE *see* long-term evolution

MAC *see* medium access control
machine learning (ML) 269-280
  see also artificial intelligence (AI)
  algorithms 271-273
  analytics to deliver value from ML models 274-275
  anomalies detection 275-278
  applications 275-279
  cf. artificial intelligence (AI) 270
  automation and optimization 262
  global navigation satellite system (GNSS) 278-279
  models for wireless communications 273-275
  network integrity 278
  network slicing 280
  neural networks 271-273
  non-parametric models 271-273
  parametric models 271-273
  predictive analytics 274-275
  radio interface performance 277
  signal-to-noise ratio (SNR) 277
machine type communications (MTC) 9
management philosophies, Cellular Internet of Things (CIoT) 243-245
mapping to antennas 6
market in connectivity, automation and optimization 255-256
massive Machine Type Communications (mMTC) 23, 24, *25*, 50, 111
  evolution 84-86
  use cases 84-86
massive MIMO (M-MIMO) 40, 42, 65, 72-74
  see also 5G New Radio (5G NR)
MEC *see* Multi-Access Edge Computing
medium access control (MAC) 6, 53-57, 60
  balancing reliability, latency and power efficiency 55-56
MIMO *see* Multiple-Input Multiple-Output
MISO *see* Multiple Input Single Output
ML *see* machine learning
M-MIMO *see* massive MIMO
mMTC *see* massive Machine Type Communications
mmWave 40-42, 45-46, 72, 75-76, 81, 261
mobile core history 126-127
  LTE EPC components 126-127
mobile switching center (MSC) 7, *8*

Table of Contents

model and intent-driven orchestration 196-197
models for wireless communications, machine learning (ML) 273-275
modulation 6, 42-43, 47
MSC *see* mobile switching center
MTC *see* machine type communications
Multi-Access Edge Computing (MEC) 153-173
  see also edge computing
  5GS common API framework 168-170
  bandwidth management service 161
  Common API Framework (CAPIF) 168-170
  conformance testing 170
  control and user plane separation (CUPS) 164
  ETSI ISG MEC 170-173
  interoperability testing 170-171
  location service 161
  MEC application enablement framework 156-158
  MEC deployment in 3GPP 5GS 166-168
  MEC deployment integrated in the 5G network 166-168
  MEC deployment options 161-164
  MEC impact on operational systems 170-173
  MEC service producing and consuming applications 158, *159*
  MEC services 158-161
  network information services 159-160
  network resource function (NRF) 166
  Policy Control Function (PCF) 165, 166-167
  S1 based deployment 161-162, *164*
  Service-Based Architecture (SBA) 166
  SGi based deployment 163-164
  standardization 153-156
  UE identity service 161
  V2X service 161
  virtualization 153-155
Multiple Input Single Output (MISO) 68-69
Multiple-Input Multiple-Output (MIMO) 65-82
  cf. beamforming **81-82**
  history of spectrum reuse in wireless access 65
  massive MIMO (M-MIMO) 40, 42, 65, 72-74
  Multi-User Multiple-Input Multiple-Output (MU-MIMO) 71, 72, **81**
  Single User-Multiple Input Single Output (SU-MIMO) 70-71, 72
  support 76-82
Multi-User Multiple-Input Multiple-Output (MU-MIMO) 71, 72, **81**

Narrowband IoT (NB-IoT)
  Cellular Internet of Things (CIoT) 210, 211, 213, 222-242
  NB-IoT Switching Between IP-Based Tunnel and Non-IP Data 242
  NB-IoT Using CIoT EPS Optimization Data Transfer Using NAS PDU and IP-Based Tunnels [3, 5] 222-235
  NB-IoT Using Non-IP Data Communication (NIDD) [4, 5] 235-241
  NB-IoT Using SMS 242
NAS *see* non-access-stratum
NB-IoT *see* Narrowband IoT
neighbor lists, network tuning for optimization 266
NetOps 194
network flexibility and complexity, automation and optimization 261
Network Function Virtualization Infrastructure (NFVI) 182-184
  evaluating NFVIs 183-184
Network Function Virtualization (NFV) 31
network information services, Multi-Access Edge Computing (MEC) 159-160
network integrity, machine learning (ML) 278
network programmability, automation and optimization 261-262
Network Repository Functions (NRF) 134-140, 166, *167*
network resilience 194-196
network resource function (NRF), Multi-Access Edge Computing (MEC) 166
network security and cyber security, automation and optimization 258-259
network slicing 31-32, 50
  5G Core network slices topology 146, *147*
  5G network slice identifiers 148
  5G network slice selection 149
  5G New Radio (5G NR) 89-90
  5G NGC 145-149
  5G RAN 122-124
  automation and optimization 254-255
  Cellular Internet of Things (CIoT) 245-247
  CIoT SLA management 245-247
  coexistence 254-255
  eMBB and uRLLC applications (Control Plane stack) *123*
  Enhanced Mobile Broadband (eMBB) 123, 147
  intent-based optimization 280
  machine learning (ML) 280
  rich service ecosystems 254-255
  Single Network Slice Selection Assistance Information (S-NSSAI) 148-149
network tuning for optimization 265-269
  beam configuration 265-266
  handover parameters 267
  idle and inactive state operation and access parameters 268
  layer management 268
  neighbor lists 266
  physical cell identity 267
  physical layer parameters 267-268
  placement and configuration of network functions 268
  scheduler configuration 268
  transport configuration 269
  user equipment (UE) 265-267
neural networks, machine learning (ML) 271-273
Next Generation Core (NGC) 127-130
  vs. Evolved Packet Core (EPC) 129-130
NFV *see* Network Function Virtualization
NFVI *see* Network Function Virtualization Infrastructure
NGC *see* Next Generation Core
non-access-stratum (NAS) 5
non-parametric models, machine learning (ML) 271-273
NRF *see* Network Repository Functions; network resource function
numerologies 45-50
  flexible subcarrier spacing and numerologies 45-46

OCP *see* Open Compute Project
OFDM *see* Orthogonal Frequency Division Multiplexing
ONAP *see* Open Network Automation Platform
Open Compute Project (OCP), capital expenditure (CapEx) 179-180
open interfaces 30-31
Open Network Automation Platform (ONAP) 193-194
Open Source MANO (OSM) 192
open source projects
  Open Network Automation Platform (ONAP) 193-194
  standardization 190-194
  virtualization 190-194
Open-Radio Access Network (O-RAN) 30-31, *131*, *132*, 260
operational considerations and patterns
  DevOps 194
  hybrid operational models 198-199
  model and intent-driven orchestration 196-197
  NetOps 194
  radically different concepts 194-196
  reactive vs. proactive assurance 197-198
  resilience 194-196
operational expenditure (OpEx) 19-20, 30-31, 34-35
  automation and optimization 256-257
  operational complexity management 256-257
optical layer technology, fronthaul networks 94-97
optimization *see* automation and optimization
O-RAN *see* Open-Radio Access Network
orchestration 27-28
Orthogonal Frequency Division Multiplexing (OFDM) 15, 38
  waveforms 42-45
OSM *see* Open Source MANO
outages, network resilience 194-196

packet data convergence protocol (PDCP) 6
parametric models, machine learning (ML) 271-273
partial timing support (PTS) 108, *109*
passive optical network (PON) 96
passive wavelength division multiplexing (DWDM) 95
PCF *see* Policy Control Function
PDCCH *see* Physical Downlink Control Channel
PDCP *see* packet data convergence protocol
PDSCH *see* Physical Downlink Shared Channel
phase noise 45, 46, 86, 267
physical (PHY) functions 6
physical cell identity, network tuning for optimization 267
Physical Downlink Control Channel (PDCCH), configuring PDCCH monitoring to trade-off latency, reliability, power efficiency *56*, 58-59
Physical Downlink Shared Channel (PDSCH)
  configuring PDSCH duration to trade-off latency, reliability, power efficiency 59-60
  configuring PDSCH to trade-off latency, reliability and power efficiency *56*, 57-58
  demodulation reference signal (DMRS) 57-58
  time domain interleaving 58
physical layer parameters, network tuning for optimization 267-268
Physical Uplink Control Channel (PUCCH), configuring PUCCH duration to trade-off latency, reliability, power efficiency *56*, 59
Physical Uplink Shared Channel (PUSCH) 59
  Code Block Group-based re-transmission 61-62
  configuring PUSCH duration to trade-off latency, reliability, power efficiency 60-62
  grant-free operation for UL transmission 61
  pre-emption of UL and DL resources 61
Policy Control Function (PCF), Multi-Access Edge Computing (MEC) 165, 166-167
PON *see* passive optical network
power outages, network resilience 194-196
Power Saving Mode (PSM), Cellular Internet of Things (CIoT) 213
precision timing protocol (PTP) 100-102
predictive analytics, machine learning (ML) 274-275
primary reference time clocks (PRTC) *103*, 108, 109
primary synchronization signal (PSS) 76, *79*, *80*
proactive vs. reactive assurance 197-198
PRTC *see* primary reference time clocks
PSM *see* Power Saving Mode
PSS *see* primary synchronization signal
PSTN *see* Public Switched Telephone Network
PTS *see* partial timing support
Public Switched Telephone Network (PSTN) 7, *8*
PUCCH *see* Physical Uplink Control Channel
PUSCH *see* Physical Uplink Shared Channel
PWDM *see* passive wavelength division multiplexing

quality-of-service (QoS) 5-6
  see also 5GS QoS architecture and models

radio access network (RAN) 3-6
  5G RAN 122-124
  Cellular Internet of Things (CIoT) 211
  historical context 92-93
radio interface performance, machine learning (ML) 277
radio link control (RLC) 6
Radio over Ethernet (RoE) 121-122
radio resource control (RRC) 5
RAN *see* radio access network
rate matching 6
reactive vs. proactive assurance 197-198
receive diversity, Single Input Multiple Output (SIMO) transmission 67-68
regulatory requirements, automation and optimization 259
reliability/energy efficiency/latency trade-off 53-64
Reliance Jio, virtualization 176-177
remote radio head (RRH)/remote radio unit (RRU) 15, 93-94, 112, 113
remote radio unit (RRU) 93-97, 112
resilience, network 194-196
resource element (RE) mapping 6
revenue
  mobile operators 2, *3*, 20, 34-35, 177-178
  revenue models shift 177-178

Revenue Per Unit (RPU), Cellular Internet of Things (CIoT) 209-210
rich service ecosystems, network slicing 254-255
RLC *see* radio link control
RoE *see* Radio over Ethernet
routing and traffic steering, edge computing 165
RPU *see* Revenue Per Unit
RRC *see* radio resource control
RRH *see* remote radio head (RRH)/remote radio unit
RRU *see* remote radio unit

SBA *see* Service-Based Architecture
scheduler configuration, network tuning for optimization 268
SDAP *see* service data adaptation protocol
SDI *see* Software Defined Infrastructure
SDN *see* software defined networks
secondary synchronization signal (SSS) 76, *79*, *80*
self-organizing networks (SON) 250-251
service creation 255-256
service data adaptation protocol (SDAP) 5-6
service level agreements (SLAs), Cellular Internet of Things (CIoT) 245-247
Service-Based Architecture (SBA) 27, *28*, 130-140
  edge computing 165
  Multi-Access Edge Computing (MEC) 166
  virtualization 262
Session and Service Continuity (SSC), edge computing 165-166
signaling storms and security, Cellular Internet of Things (CIoT) 244
signal-to-noise ratio (SNR), machine learning (ML) 277
SIMO *see* Single Input Multiple Output
simple wavelength multiplexing *95*
Single Input Multiple Output (SIMO) 67-68
  receive diversity 67-68
Single Input, Single Output (SISO) 66, *67*
Single Network Slice Selection Assistance Information (S-NSSAI) 148-149
Single-User Multiple-Input Multiple-Output (SU-MIMO) 70-71, 72
SISO *see* Single Input, Single Output
SLAs *see* service level agreements
SNR *see* signal-to-noise ratio
S-NSSAI *see* Single Network Slice Selection Assistance Information
Software Defined Infrastructure (SDI), Virtualized Infrastructure Managers (VIM) 187-188
software defined networks (SDN), virtualization 179, 180-181, 193
SON *see* self-organizing networks
spectral efficiency, guard bands 45
spectrum reuse in wireless access 65
SSC *see* Session and Service Continuity
SSS *see* secondary synchronization signal
standardization 29
  3GPP standardization 25, 27, 83, 100, 146, 210
  edge computing 153-156
  Multi-Access Edge Computing (MEC) 153-156
  open source projects 190-194
  virtualization 190-194

subcarriers 44-50
  flexible subcarrier spacing and numerologies 45-46
subscriber numbers 2, *3*
subscribers *3*
SU-MIMO *see* Single-User Multiple-Input Multiple-Output
symbols 42-43
synchronization networks 100-110
  assisted partial timing support (APTS) 108, *109*
  frequency synchronization standards 102-107
  G.8263 testing procedure for noise generation 105
  G.8265.1 PTP message rates 106-107
  G.8271.1 time synchronization deployment cases 107, *108*
  general packet network timing architecture *106*
  IEEE 1588v2 standard 100-102
  overview of ITU-T synchronization standards **103**
  partial timing support (PTS) 108, *109*
  precision timing protocol (PTP) 100-102
  primary reference time clocks (PRTC) *103*, 108, 109
  reference points for network limits G.8261.1 *104*
  synchronization requirements and technologies 100-102
  Synchronous Ethernet (SyncE) 101, *103*
  time and phase synchronization standards 107-110
Synchronous Ethernet (SyncE) 101, *103*

TDMA *see* Time Division Multiple Access
terminology 3-6
Time Division Multiple Access (TDMA) 6
time domain interleaving, Physical Downlink Shared Channel (PDSCH) 58
Timing Sensitive Network (TSN) 118-119
transport configuration, network tuning for optimization 269
transport networks, 5G *see* 5G transport networks
TSN *see* Timing Sensitive Network

UE *see* user equipment
Ultra-Reliable Low Latency Communications (URLLC) 24, *25*, 63-64, 111
  evolution 83-84
UMTS *see* universal mobile telecommunication system
UMTS Terrestrial Radio Access Network (UTRAN) 10
universal mobile telecommunication system (UMTS) 10, 12
UP *see* user plane
URLLC *see* Ultra-Reliable Low Latency Communications
usage scenarios, key capabilities in *25*
use-case based service drivers, 5G 22-25
user equipment (UE) 3-5
  automation and optimization 264
  beam configuration 265-266
  gNB and UE processing time capability 56-57
  idle and inactive state operation and access parameters 268
  network tuning for optimization 265-267
  spatial extent of optimization 264
  UE identity service, Multi-Access Edge Computing (MEC) 161

user plane (UP) 5
UTRAN *see* UMTS Terrestrial Radio Access Network

V2X service, Multi-Access Edge Computing (MEC) 161
VIMs *see* Virtualized Infrastructure Managers
Virtual Machines (VM) 185-186
virtualization 27-28, 176-202
  automation and optimization 262
  competitive advantage 176-177
  ETSI NFV 179
  historical context 178-180
  hybrid operational models 198-199
  key technology options 181-189
  model and intent-driven orchestration 196-197
  Multi-Access Edge Computing (MEC) 153-155
  Network Function Virtualization Infrastructure (NFVI) 182-184
  open source projects 190-194
  operational considerations and patterns 194-199
  possibilities 176
  private and public cloud environments 188-189
  reactive vs. proactive assurance 197-198
  Reliance Jio 176-177
  revenue models shift 177-178
  Service-Based Architecture (SBA) 262
  software defined networks (SDN) 179, 180-181, 193
  standardization 190-194
  value 180-181
  Virtualized Infrastructure Managers (VIM) 184-188
  zero-touch networks 199-202
virtualization and telecommunications lifecycle 16-18
Virtualized Infrastructure Managers (VIM) 184-188
  containers 186-187
  Software Defined Infrastructure (SDI) 187-188
  Virtual Machines (VM) 185-186
VM *see* Virtual Machines

waveforms, OFDM waveforms 42-45
wavelength division multiplexing (WDM) 94-96
WCDMA *see* wideband CDMA
WDM *see* wavelength division multiplexing
wideband CDMA (WCDMA) 10, 12
worldwide interoperability for microwave access (WIMAX) 13

zero-touch networks 199-202
  ETSI ZSM 200-202

Technology vendors, network equipment and device manufacturers, and service providers worldwide have begun offering 5G products and services – yet 5G technical standards are still being finalized. This book explores the revolutionary 5G architecture and describes how each segment of the 5G network is redesigned to provide the promised characteristics and benefits, as well as offer new use cases and applications that define the sixth technological evolution era.

To help business executives and network professionals understand the path to 5G implementation and adoption, a panel of industry experts has collaborated on this indispensable reference, breaking this complex technology into distinct categories containing the critical details necessary for multiple audiences. The authors draw upon their experiences in the development and deployment of all previous generations of wireless technology, as well as their collaborations with every major network equipment manufacturer and service provider worldwide. They are recognized contributors to all of the principal industry working groups including 3GPP, 5GAA, ETSI, IEEE and ITU among many others.

Authors

Jonathan Brooksby

Walter Featherston, PhD

Per Kangru

Eng Wei Koo

Chris Murphy, PhD

Howard Thomas, PhD

Reza Vaez-Ghaemi, PhD

Sameh Yamany, PhD

VIAVI Solutions

Contact Us **+1 844 GO VIAVI**
(+1 844 468 4284)

To reach the VIAVI office nearest you,
visit viavisolutions.com/contact

understanding-5g-2nd-edition-
bk-fop-nse-ae 30187507 900 0121