9

Security

Carolina Canales-Valenzuela[1], Madalina Baltatu[2], Luciana Costa[2], Kai Habel[3], Volker Jungnickel[3], Geza Koczian[4], Felix Ngobigha[4], Michael C. Parker[4], Muhammad Shuaib Siddiqui[5], Eleni Trouva[6] and Stuart D. Walker[4]

[1] *Ericsson, Spain*
[2] *Telecom Italia, Italy*
[3] *Heinrich Hertz Institut, Germany*
[4] *University of Essex, United Kingdom*
[5] *Fundació i2CAT, Spain*
[6] *National Centre for Scientific Research "Demokritos", Greece*

9.1 Introduction

Future 5th generation (5G) technologies are anticipated to address next generation network's challenges and tackle the novel business requirements associated with different vertical sectors. This implies that 5G technologies will not only encompass new wired and wireless network technologies to support higher data rates, bandwidths, numbers of devices, etc., as elaborated in Chapter 2, but also need to be cohesively aligned from a technological as well as business standpoint with the different vertical sectors, for their optimized and efficient use of the network, for instance through customized network slices. Furthermore, convergence, automation and flexibility are expected to be intrinsic traits of any 5G system. The introduction of this multitude of complex new requirements and novel technologies immensely impacts the security landscape of 5G, and therefore the need to revisit its properties becomes essential.

Given the array of new technologies and vertical industries that 5G aims to support, as detailed in Section 2.2, it is obvious that different parts of the 5G ecosystem will be developed by different stakeholders. The interoperability of these different elements can be addressed via standard interfaces. However, a holistic approach to address the security of these elements, individually and more importantly when integrating them, is crucial, so that the required level of security can be guaranteed for a 5G system. The complexity of the security landscape of 5G systems increases exponentially due to the required tightly knitted alignment with vertical industries, and the heterogeneous nature of the integrating technologies (wired or wireless networks, virtualization, etc.). Thus, the security vision of

5G System Design: Architectural and Functional Considerations and Long Term Research, First Edition.
Edited by Patrick Marsch, Ömer Bulakçı, Olav Queseth and Mauro Boldi.

5G needs to at least comprise the security requirements associated to all of the involved technologies and vertical sectors. For example, reliability in 5G systems would go beyond availability or up-time of the network infrastructure and will also include high connectivity, virtually infinite perceived capacity and ubiquitous coverage. Furthermore, the tedious and cumbersome chores of securing legacy networks become even more complex with the inclusion of software-defined networking (SDN) and network function virtualization (NFV) technologies in 5G networks, as detailed in Section 10.2. These are just a couple of examples of how the 5G system security landscape is impacted.

The chapter is organized as follows: Section 9.2 describes the envisioned security threat landscape. Corresponding requirements for 5G security are derived in Section 9.3. Then, the main characteristics of the envisioned 5G security architecture are described in Section 9.4, before the chapter is summarized in Section 9.5.

9.2 Threat Landscape

As mentioned before, 5G enables innovative scenarios and applications making use of ultra-high speed, low-latency telecommunication networks for fixed and mobile users, and machine-to-machine communications. These scenarios, together with the introduction of the new paradigm for computing and network infrastructure which decouples the actual functionality from the underlying hardware and middleware functions, for instance through cloud computing and SDN, as detailed in Section 5.2, further reinforces the need for automated management and control of the telecommunication infrastructure. In particular, since a cloud-based paradigm that promotes that infrastructure is highly accessible and shared by multiple tenants, as for instance virtual network operators, the concept of a highly secure network gains even more relevance.

There are two different aspects that need to be taken into account in order to address the security of the upcoming 5G network: On one hand, the need to address the overall security functionalities of the network, composed of virtualized and non-virtualized functionalities, where the latter are often referred to as "traditional" or "classical" functionalities. Due to the increasingly virtualized nature of the 5G network, the effectiveness of traditional security approaches using physical network elements and devices (a.k.a. entities) is likely to diminish. When just instantiated in a cloud, the virtual network functions (VNFs) replacing any physical entities lack visibility w.r.t. changes performed over virtualized functions, service chains, and the traffic being exchanged on the virtualized network. Or putting it into different words, a holistic security approach comprising both virtualized and non-virtualized (i.e., traditional) security functions needs to be put in place.

On the other hand, there is a need to cater for an automated security management solution for the 5G network. Today we can't foresee the new and ever-changing threats that 5G networks will have to protect against, but we do have the basis to create autonomic network management solutions that shall cope with them, being fed with insights from governed real-time analytics systems on the one hand, and actuating on network resources in order to minimize or prevent the effects of the detected threats in real-time on the other hand. It is therefore of outmost importance to be able to provide robust, flexible and proactive mechanisms to detect and prevent security issues, and to be able to perform that in real-time and in an automated fashion.

Taking this into account, the infrastructure operator and the different tenants using part of the network need to maintain the overall end-to-end (E2E) security of the network with different degrees

of responsibility and different focus areas, including E2E security, physical infrastructure security, and the security of new virtualized resources (being those applications or network functions).

In detail, the 5G network assets to be secured are:

i) **User information security**, referring to the protection of E2E user data, related to human users but also machine-type communications (MTC), including the transfer of E2E control plane and user plane data;
ii) **Network element security**, related to the protection of endpoint devices and network elements, here also referred to as user equipments (UEs), including the physical security for network elements and the application SW, which in a cloud-based architecture is composed of VNFs;
iii) **Transport/interface security**, referring to the protection of communication paths.

As general examples of types of attacks, depending on the attacked 5G asset, we could mention:

- **Attacks towards UEs or network elements (NEs)**, such as infection via malware or bots infecting subscribers' devices which can generate spurious or attack traffic, create signaling storms into the network, and drain device batteries;
- **Attacks towards the different network subsystems**, such as the radio access network (RAN) and core network (CN), causing resource exhaustion, terms and conditions violations such as service level agreement (SLA) violations, or attacks on the Domain Naming System (DNS), billing, and signaling infrastructure, etc.;
- **Attacks towards end-user applications**, such as server-side malware, application-level and protocol-specific distributed denial of service (DDoS) attacks, etc.

9.3 5G Security Requirements

5G networking imposes new security requirements as it is linked to new business and trust models involving new stakeholders, new actors, and new service deployment and delivery procedures for telecommunication services, as stressed in Section 2.6. In this section, as a result of the previous analysis of threats, we draw attention to selected requirements for 5G security. These do not solely affect the design of security services deployed to offer protection and privacy, but are expected to heavily impact the overall 5G architecture design as well.

Clearly, a general requirement is that security systems in 5G should be flexible enough to accommodate the expected diversity of connected devices and systems, provide the ability to monitor their real-time status and traffic, and provide protection against the main attack vectors listed in the previous section. In the following subsections, we will dive into more details on the requirements stemming from specific aspects of the 5G system.

9.3.1 Adoption of Software-defined Networking and Virtualization Technologies

Taking into consideration the dominant trend for programmable, SDN-based infrastructures, the transition to 5G is followed by the adoption of several emerging virtualization technologies such as NFV using SDN. These paradigms, although they facilitate the orchestration and management procedures which offer increased flexibility and reduce cost for the operators, require several additional security considerations and deeply impact network security.

The SDN architecture, with the decoupling of the network control and forwarding functions, the logically centralized network control elements and the exposed interfaces that enable programmability of the network elements (switches), as detailed in Section 10.2, imports new targets for attack and exploitation at control and user planes, beside the application layer. Attackers might attempt to compromise SDN elements, controllers or switches. Especially SDN controllers are attractive targets for attacks, as such attacks may allow to compromise the SDN control plane. A compromised SDN controller enables an attacker to install new flow rules and direct traffic to flows as desired across the controlled SDN switches, altering the network design. Another common security threat towards the control plane is to perform a denial of service (DoS) attack towards the SDN controller, depleting its resources. By leveraging the security vulnerabilities in well-known southbound SDN protocols such as OpenFlow (OF), Open vSwitch Database Management Protocol (OVSDB), NETCONF, etc., attackers could divert traffic flows or simply eavesdrop the already installed flow rules for reconnaissance purposes.

Regarding NFV, although it is justifiably connected to cost reduction through the replacement of specialized hardware with virtualized services, its adoption brings significant security risks. According to the European Telecommunications Standards Institute (ETSI) NFV security (SEC) working group [1], these include risks related to network virtualization (e.g., memory leakage, interrupt isolation), traditional networking threats (e.g., flooding attacks, routing security) and those new threats that result from the combination of virtualization and networking technologies. Additionally, other security risks to consider related to the use of NFV are those specific to the software used to implement a network function. To summarize, the anticipated introduction of virtualization techniques in 5G not only introduces the risks associated to one virtualized element in itself, but also to the combination and integration of different virtualized elements from different sources.

9.3.2 Security Automation and Management

Due to the complexity of the 5G infrastructure and the increasing automation and sophistication of attacks, the automation of security functionality is essential. Along these lines, network virtualization, orchestration and analytics contribute to creating a policy-controlled network automation cycle. Additionally, virtualization technologies enable the rapid provisioning of network services, along with security services, allowing for a rapid and inexpensive creation and removal of service chains and virtualized security functions on demand.

In order to properly secure the 5G network, it must be made with out-of-the-box security management features that are highly integrated with the rest of the network elements and that are flexible enough to evolve and adapt to the network changes as soon as these happen. That is, security and security management should be architectural network principles, per network design.

Even if it might not be possible to automate all security management procedures, many of them will be highly automated, and therefore the network management and orchestration components, such as the service orchestrator, the policy manager, the inventory systems, and the existing operations support systems (OSS) and business support systems (BSS) need to provide decision support for security threat detection and mitigation. Integration and co-operation of security systems with service orchestration, policy management systems and cloud and network managers can assist the identification of security incidents and accelerate the response process towards security attacks.

Additionally, it should be possible for different tenants to manage the security features of their network slice up to a certain point, though of course the security of the whole network infrastructure is rather in the responsibility of the network infrastructure provider.

9.3.3 Slice Isolation and Protection Against Side Channel Attacks in Multi-Tenant Environments

The design of future security solutions should deal with the fact that infrastructure resources are being shared between different operators and in addition between their services and slices. In an abstract view, exchanged traffic belongs to different tenants who operate their services over a common infrastructure. Tenant isolation is vital to support self-contained and independent networks required for slice creation, to ensure the quality of service offered to tenants depending on the agreed service-level agreement (SLA), and to provide data integrity and confidentiality. Given the shared infrastructure model, which is heavily based on a shared computing environment, special security measures should be provisioned to handle possible cross-tenant side-channel attacks and prevent leakage of sensitive information amongst tenants.

The multi-tenancy model imposes new requirements to the design of 5G security systems. Security systems should provide different levels of access to services, infrastructure resources and information to tenants with different management capabilities and scope. Consequently, flexible, scalable and possibly hierarchical access schemes are required. Tenants should not be allowed to access resources or information that belongs to other tenants, without excluding special cases in which resources should be shared amongst tenants. On the other hand, infrastructure providers should be able to monitor the whole infrastructure they own, without being able to access sensitive data that reveals business information of the tenants.

9.3.4 Monitoring and Analytics for Security Purposes

Resource and service monitoring information should be used to assist security decisions in future security systems. Information that could be used for attack detection includes generic metrics from the physical and virtual network devices, notifications received from security services deployed in the network, authentication and authorization incidents and others. Traditional monitoring data such as central processing unit (CPU) utilization, random access memory (RAM) utilization, network interface bitrate and packet rate measurements, can greatly assist in detecting zero-day attacks for which identification patterns and signatures have not been defined yet. Scalable analytics engines, designed to capture, store and analyze all network activities, should be deployed in central and edge clouds. These systems could be also offered in "-as-a-Service" schemes using NFV technology and be deployed on an on-demand basis.

9.4 5G Security Architecture

9.4.1 Overall Description

Previous generation networks involved one or more security and trust domains, or more precisely, at least as many domains as the number of home networks serving each of the participants involved in the communication. This paradigm is reflected in Figure 9-1.

In 5G networks, the security and trust paradigm gets increasingly complicated. Apart from having the security and trust domains associated to the participants of the communication link, i.e. *horizontal* security and trust domains (HSTD), those HSTDs are implemented by means of virtualized and

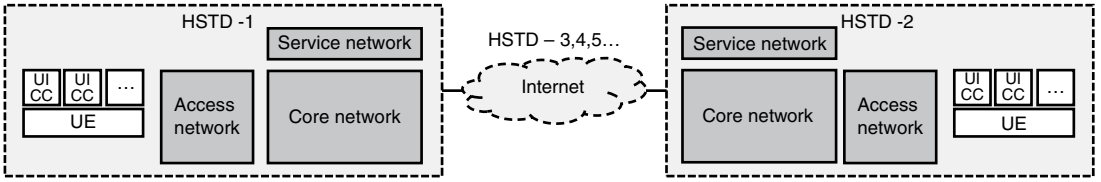

Figure 9-1. Security and trust domains in traditional networks.

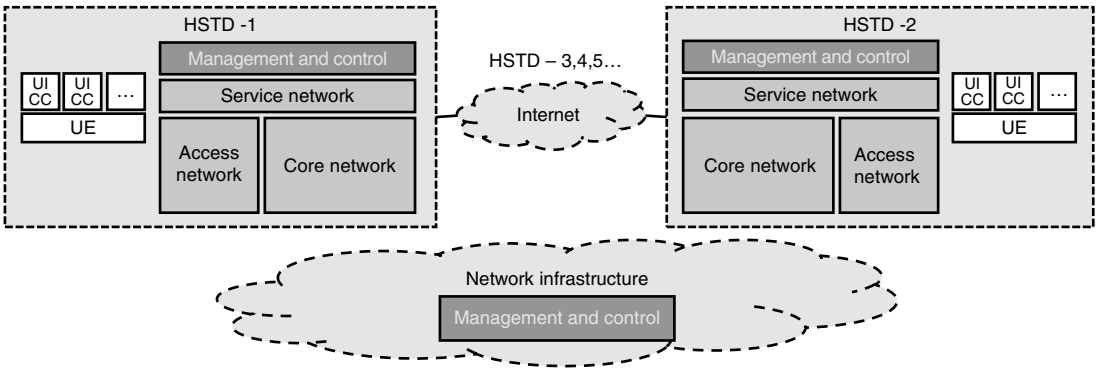

Figure 9-2. The impact of virtualization on the security and trust domains in 5G network.

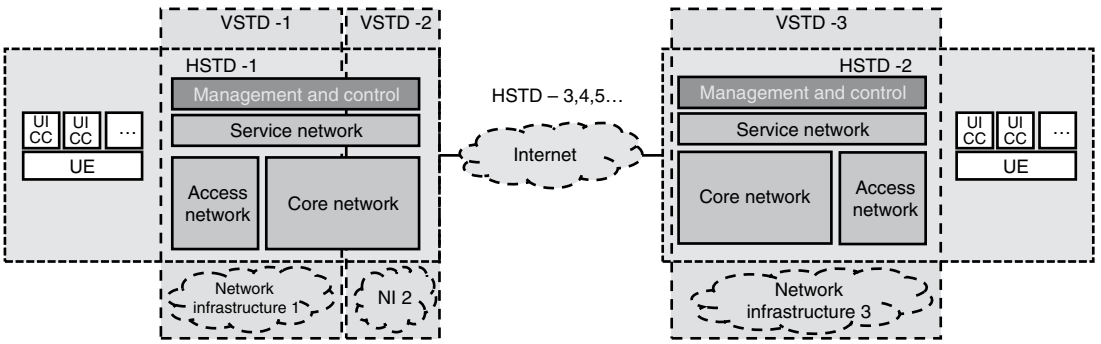

Figure 9-3. 5G security and trust domain paradigm.

non-virtualized resources which run on an infrastructure which could belong to another organization, for instance an infrastructure network provider, as reflected in Figure 9-2. For realizing a secure E2E link, obviously, joint management and control functions are needed.

We could therefore say that we have additional *vertical* security and trust domains (VSTDs) to be added to the traditional security and trust paradigm of the network. Figure 9-3 depicts the complete 5G network security and trust paradigm.

In the subsequent sections, we will focus on the security properties of the different architectural elements inside the same HSTD, taking into account that each HSTD could map to one or more

VSTDs. In particular, we here decompose each HSTD into the following elements and related security aspects: infrastructure security, physical layer security, RAN security, service-level security, and overall security management and automation.

9.4.2 Infrastructure Security

Infrastructure security (IS) can be implemented at many levels of the Open Systems Interconnect (OSI) model protocol stack in the transport network, often with a focus on the higher layers, e.g. in the form of MACSec for layer-2/data link layer, IPSec for layer-3/network layer, or transport layer security (TLS) for layer-4/transport layer. However, the layer-1/physical layer offers a complementary approach to improve communication security, particularly in the context of wireless networking, where the properties of the communications system can also be exploited, as we will see in Section 9.4.3. Thus, IS, where the transport network infrastructure inherently supports security[1] for any network functions, is becoming an important aspect of 5G security design.

Security has so far been seen as an add-on feature in the layered network design, and in particular the separation of the upper layers from the physical layer as a reliable bit pipe, and the provision of security only at those higher layers has contributed to fortifying the existing computational security techniques. Traditional solutions to mitigate the security challenges at the upper layers use various types of private and public secret keys via computation-based mechanisms, often referred to as *over-the-top* (OTT) cryptography.

While 4th generation (4G) UEs are authenticated in the mobile core, they might communicate via unsecured transport networks. The so-called Access Stratum (AS) is terminated at the enhanced Node-B (eNB) and thereby it protects control and data transport only over the air interface. In the AS, user data can be cipher-protected by the Packet Data Convergence Protocol (PDCP), where the radio control data will always be integrity-protected and can eventually be cipher-protected. The PDCP security foresees different algorithms for integrity protection, such as the Advanced Encryption Standard Counter Mode (AES-CTR), SNOW3G, and the Zu Chongzhi algorithm (ZUC). For cipher protection, it offers the same 3 schemes and also the "NULL" variant, which means that traffic is unencrypted.

In 4G, both the user plane and control plane traffic between the eNB and the Evolved Packet Core (EPC), i.e., the core network or the so-called Non-Access Stratum (NAS), is cipher- and integrity-protected. Only the encryption keys and the control traffic between eNB and Mobility Management Entity (MME) are currently protected. The NAS is realized, in practice, via a public transport network infrastructure being partly or fully shared among multiple operators and services. For UE attach procedures, the traffic is unprotected through NAS.

Now we look deeper into the methods used to further ensure secure NAS transport over the transport network infrastructure where, of course, proper protection algorithms are already available, such as TLS, IPSec and MACsec, which can be considered as potential security protocols also in 5G. These transport protocols include handshake procedures for mutual authentication and key agreement and creation.

Note that the transport network architecture is currently subject to significant changes from 4G to 5G, as stressed throughout Chapters 5, 6 and 7, one key aspect in 5G being the possibility to have

1 One example for IS is IPSec, where routing as an elementary network function inherently supports security.

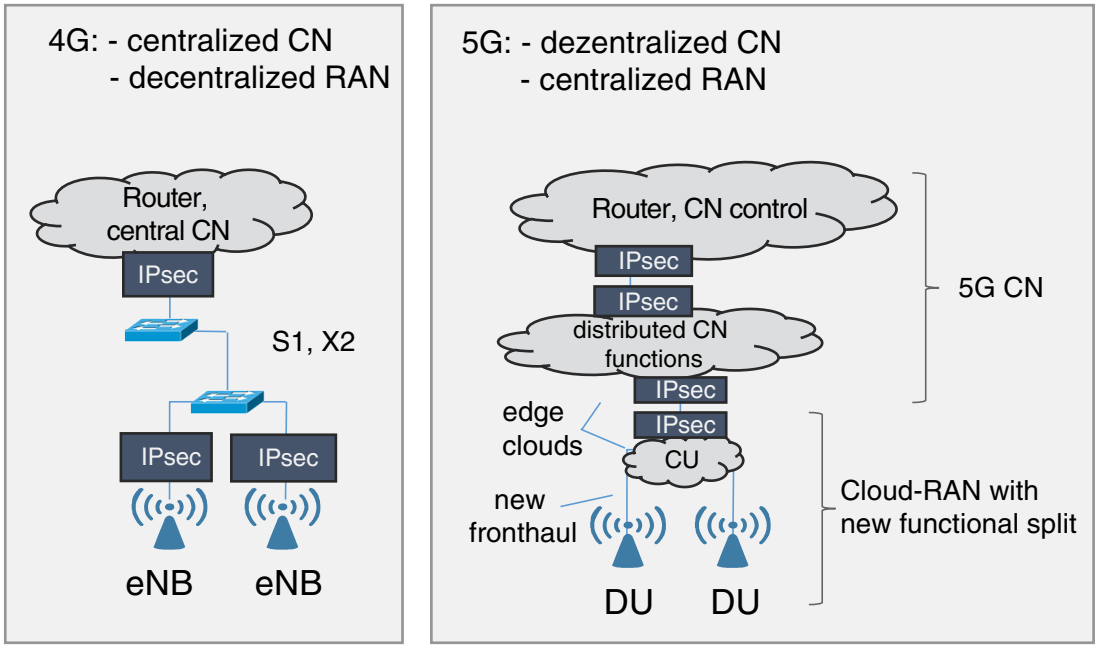

Figure 9-4. Key architectural differences between 4G and 5G, with implications on the approaches towards achieving transport infrastructure security.

more functional split options and a much more flexible assignment of network functions to physical entities, as for instance stressed in Section 6.6. This difference between 4G and 5G, and its implication on security handling, are also illustrated in Figure 9-4.

In the 4G system, depicted in the left side of Figure 9-4, the EPC is totally centralized, while the complete RAN is implemented in the form of eNBs, with an IP-based connection to the EPC. In typical 4G deployments, the complex nature of the transport network is ignored, but instead, the link between eNB and EPC is abstracted as a point-to-point (P2P) link. There is an implementation guideline to use IPSec tunnels from eNB to EPC for protecting this link, but the decision to do so is left to the mobile network operator. If not used, there is an opportunity to wiretap the user plane traffic in the transport network, for instance on unsecured microwave backhaul links.

In the 5G system [2], as shown on the right side of Figure 9-4, part of the CN functionality may be placed in edge clouds, for instance for latency-critical services, while some extent of RAN functionality may be centralized, possibly to the same edge clouds. Unlike the P2P abstraction of the 4G transport network, in 5G the complex transport network architecture, which consists of fixed core and metro networks (both typically realized as a logical ring to enhance resilience) and fixed access networks (typically realized using a logical tree), are abstracted as a hierarchical cloud network architecture with several aggregation levels at which clouds are hosted. These aggregation nodes currently evolve into data centers offering compute and storage capabilities that can be dynamically assigned to different network operators and their vertical services in 5G.

Based on a hierarchical transport network, it is possible to implement a comparatively more decentralized CN and more centralized RAN in 5G. More precisely, the 5G network will allow

CN functionality to be shifted closer to the radio, so that the future CN will be more distributed in nature. However, these distributed CN functions need to be controlled by a central CN. Starting from the top, CN clouds will host the routing into the public Internet, together with centralized CN control functions. More towards the edge, another cloud is located in which distributed CN functions can be hosted. Closer to the radio link, e.g. at the central office in the fixed access network, centralized RAN functions can be deployed in so called centralized units (CUs), whereas other RAN functionalities are implemented as distributed units (DUs) at the radio edge. From the security point of view, distributing the CN functionality and centralizing the RAN present advantages and disadvantages. On one hand, the vulnerable surface increases, while on the other hand the impact of successfully attacking one of the system sub-elements is reduced.

The functional split between centralized and decentralized RAN functions has been a matter of debate in 3GPP. One prominent higher-layer split option, the so-called 3GPP split option 2 [3], foresees the centralization of Radio Resource Control (RRC) and PDCP, while the Radio Link Control (RLC), Medium Access Control (MAC) and the physical layer (PHY) are distributed, involving a so-called F1 interface for the new fronthaul between CU and DUs. Note that this and various other RAN split options are discussed in more detail in Section 6.6.2.

The mentioned possibility in 5G to split the RAN into CUs and DUs may in fact inherently increase security for the UP traffic. In particular, if the aforementioned 3GPP split option 2 is chosen, this means that the encryption of the radio link (denoted as ciphering) that happens in the PDCP layer is placed in the CU, with the consequence that the *compound link*, i.e. from the UE over the air to the DU and then eventually via a microwave or other fixed network connection to the CU, is inherently confidentiality-protected. Because the fixed access network provides various opportunities to attack the mobile network, e.g. via unprotected microwave links in 4G, the envisioned larger degree of RAN centralization in 5G inherently closes most of the potential points of attack in the mobile network.

While the transport network can be physically deployed in various topologies (bus, star, ring), logically it is abstracted as a tree with several hierarchical aggregation nodes. The 5G network architecture foresees the availability of clouds (i.e., data centres with compute and cache capabilities) co-located with aggregation nodes where CN and RAN functions can be flexibly instantiated. If cloud resources are shared, this implies that CN and RAN network functions must be encapsulated by a secure transport infrastructure protocol, such as IPSec. Mobile network operators should aim at building secure islands inside their own cloud in each data centre, in which their own VNFs are operated. These own clouds should be isolated from the clouds of other mobie network operators or services using the same transport infrastructure. The only function that needs no further isolation is routing, which is natively safe when using IPSec. Obviously, *VNFs need security encapsulation*. Nevertheless, the cloud infrastructure is a remaining security weakness, because isolation between the tenants is virtual, and VNFs of different tenants can be physically processed in the same machine. At the low processing level, tenants are therefore not physically isolated. One way out is to only use certified cloud hardware in which interactions between the tenants can be considered to be impossible.

From a security point of view, authentication and data integrity should be inherently provided by the network infrastructure. What kind of encryption is used in CN and RAN will be a matter of future debate. Targeting an all-IP infrastructure is desired from simplified CN operation and maintenance (OAM) point of view. But the new 5G architecture forces network designers also to be aware of secure communications at layers 3, 2 and 1. Lower layer techniques may in general be simpler and have less overhead in the UP, but their OAM is more complex. A combination of most suitable techniques is needed, considering for instance physical layer security in the RAN and higher layers towards the CN.

To summarize, a combination of OTT and network-assisted E2E security techniques will be needed, both in the RAN and in the CN, to enable secure sharing of transport network infrastructures in 5G, while guaranteeing low latency and secure E2E communication.

9.4.3 Physical Layer Security

As stressed in the past chapters, 5G networking features both mobile wireless (e.g., via radio links between mobile end-users and the radio unit in the infrastructure) and fixed wireless as well as wired links (e.g., for front- and backhauling), each of which have their own security issues.

The broadcast nature of the wireless domain and the mobility of the users make them susceptible to a wide variety of security attacks such as passive (e.g., traffic analysis and eavesdropping) and active attacks (e.g., denial of service attack, resource consumption, masquerade attack, replay attack, information disclosure, and message modification [4])

Some information theory results have drawn much attention recently. Most works in this area focus on secrecy capacity, that is, the maximum rate achievable between a legitimate transmitter-receiver pair subject to the constraints on information attainable by an unauthorized receiver. Information-theoretic security is an average-information measure and it may require the knowledge of channel state information (CSI) which is not necessarily accurate. The reliability in the exchange of information between a source node "Alice" and an intended destination node "Bob", and security in terms of confidentiality and message integrity with respect to an adversary "Eve" using computational security approaches have been reported to be susceptible to attacks [5][6], and as computationally complex and intensive in a dynamic mobile environment [7][8].

From an implementation perspective, security against passive and active attacks is classified into three categories: i) channel knowledge, ii) coding and power, and iii) signal detection based techniques.

i) Channel knowledge based techniques

First, there is the channel characteristic which can be exploited to obtain secure keys. Knowledge about the radio channel is very specific for the position of the devices and appears random (i.e., uncorrelated) already at distances less than half a carrier wavelength apart. Sometimes the key is even reciprocal between a sender and receiver, for instance when using time-division duplex (TDD) at low mobility. The main idea is to derive a time-variant, symmetric key by using the channel as a random number generator. Radio channels exhibit a lot of randomness due to multipath propagation and time-variance, as well as mobility that introduces Doppler effects, as detailed in Chapter 4. Channel knowledge based security approaches can be classified into radio frequency (RF) fingerprinting, algebraic channel decomposition multiplexing (ACDM) pre-coding, randomization of MIMO precoding coefficients, and the introduction of artificial noise. In [9], a method has been proposed in which discriminatory channel estimation is performed by injecting artificial noise to the remaining space of the legitimate receiver's channel to degrade the estimation performance of the eavesdropper. An improvement to this approach is discussed in [10], which exploits the CSI feedback from the legitimate receiver at the beginning of each communication stage, and which is called a multi-stage training-based technique.

ii) Coding and power based techniques

Coding is usually considered public, as the encoder and at least one implementable decoder is described in the standard for the radio link. But if the encoder is secret, coding can be used to

improve resilience against jamming and eavesdropping. The coding approach is subdivided into the use of error correction and spread spectrum coding techniques. Information protection can also be facilitated using *power* techniques. Usual schemes here also involve the employment of directional antennas and the injection of artificial noise. A directional antenna facilitates receiving data from the direction not covered by the attacking signal.

iii) Signal detection based techniques

Any E2E security approach depends on the algorithms and methods implemented at the endpoints of a data connection, i.e. at the UE, Hence, the end-user can use an own selected algorithm for cipher and integrity protection of its own E2E user connection.

To summarize, beside other techniques from physical layer security, the randomness of the radio channel allows to generate symmetric keys in specific scenarios. This new approach offers additional security, besides the traditional encryption mechanisms in the RAN, and may be interesting for some use cases, e.g., in order to make SIM cards obsolete in billions of devices anticipated in the future wireless Internet of things (IoT).

9.4.4 5G RAN Security

The general requirement for the 5G radio network is to provide at least the same security level as the previous generation network [11]. Having this in mind, the support of traditional security functions provided by the current network, like mutual authentication and key agreement between mobiles and the network, signaling data confidentiality and integrity, user data confidentiality, security visibility and configurability, are not covered in this section. It is assumed that these functionalities will be maintained and supported also in the 5G access network. However, there are also new considerations for 5G radio security design. Most notably, trust models, as discussed in Section 9.4.1, and new aspects such as potentially misbehaving entities and devices should be catered for. This section identifies the new security functionalities which should be offered by the new RAN, considering the new security requirements listed in Section 9.3.

9.4.4.1 Protection Against a Rogue 5G RAN Node

Today, there is an implicit trust relationship between the UE and the access node, which can open the door to rogue access nodes attacks, and which the 5G security architecture should overcome. In particular, the new RAN security should be designed to provide a mechanism that allows a UE to determine the legitimacy of the access node prior to engaging in any communication with it.

It has been demonstrated in [12] that current mobile networks are still vulnerable to protocol exploits, location leaks and rogue base stations. The main issue is that RRC messages that are transferred over common radio channels are transmitted without integrity and ciphering protection. The RRC protocol in LTE includes various functions needed to set up and manage over-the-air connectivity between the eNB and the UE, as covered in Section 13.3. The eNB periodically broadcasts System Information Block (SIB) messages which carry information necessary for UEs to access the network, to perform cell selection, and other information, as covered in detail in Section 13.2.2. Such broadcast messages are neither authenticated nor encrypted. Therefore, anyone owning appropriate equipment can decode them and can exploit these messages to create a targeted denial of service attack to users or to track users' movements by setting up a rogue access point. In general, a UE always scans for eNBs around it with the best signal power. Hence, in International Mobile Subscriber

Identity (IMSI) catcher types of attacks, the rogue eNB operates with a higher power than surrounding eNBs. However, there are situations where a UE, very close to a serving eNB, does not perform scanning to save power. In this case a feature called 'absolute priority based cell reselection' [13] introduced in LTE can be exploited. Based on this feature, a UE, in the Idle state, should periodically monitor and try to connect to eNBs operating at high priority frequencies. In this way, even if the UE is close to a real eNB, a rogue eNB operating on a frequency that has the highest reselection priority would force the UE to attach to it. These priorities are defined in SIB type number 4, 5, 6, and 7 messages broadcast. Using a passive attack setup, it is possible to sniff these priorities and configure a rogue eNB accordingly for malicious purposes, for example to locate the target UE. This is possible by exploiting two other types of RRC messages. A UE, when requested by the network, sends measurement reports to the eNB in RRC protocol messages. In particular, "measurement report" and "UE information-response" messages contain serving and neighboring LTE cell identifiers with their corresponding power measurements and also similar information for GSM and 3G cells. If Global Positioning System (GPS) is supported by the device, the message can also include the GPS location of the UE, and hence also that of the subscriber. Since these messages are not protected during the RRC protocol communication, an attacker can obtain these network measurements by simply decoding the radio signals and then using them to calculate a subscriber's location. These attacks might have critical impacts on the users' privacy, and are possible because of the lack of authentication of the access nodes. The mobile devices do not have the means neither to authenticate nor to validate the messages received from the access node before the authentication phase and the NAS security activation, therefore they inherently and implicitly trust all messages coming from anything that appears to be a legitimate base station.

This issue also applies when a UE is in RRC Idle mode state, as detailed in Section 13.3, and receives services such as paging, Multimedia Broadcast/Multimedia Service (MBMS), device-to-device (D2D) services etc. For example, a UE interested in the D2D service can acquire the broadcasted system information and use the radio resources configured via the system information for the D2D discovery or communication, as detailed in Chapter 14. The lack of a mechanism to verify the authenticity of this information may mislead a UE into selecting and camping on a rogue cell, which can ultimately lead to a denial of service situation, where the UE is denied access to services such as public safety warnings, incoming emergency calls, real-time application server push services, proximity services, etc.

The 5G RAN should therefore consider and provide new security features which allow a UE to determine the authenticity of the cell before camping on it and also during Idle mode. These can probably not leverage the current mobile security architecture, which is based on a symmetric key mechanism (shared key). Such a mechanism is of course suitable and best performing for UE authentication and for deriving keys for traffic protection, but it requires the involved nodes to be identified and authenticated beforehand. Instead, solutions based on a public key architecture could be taken into consideration and evaluated in 5G to counteract the problem of rogue access nodes, since they give the possibility to digitally sign the radio broadcast messages, such as Master Information Block (MIB) and SIB packets, or the sensitive information carried within, and to verify access node authenticity before a UE camps on it. The challenge lies mostly on the management side, since network-private keys are required, while mobile devices have to be provisioned with the corresponding public keys. Such keys of course have to be periodically renewed and securely re-deployed on all elements, e.g., in the case that the confidentiality of a key is compromised. This process has a non-negligible impact on the 5G security architecture and on the practical operation of the network, and hence has to be carefully evaluated.

9.4.4.2 Security Protection of the User Plane

Current 3G and 4G networks only mandate the support of encryption for the user plane data. The encryption can be enabled or not, i.e., an operator can configure the NULL cipher algorithm and in this case no encryption will be performed, based on the countries' regulations. The support for user plane integrity is not required. The reason for this choice is mainly related to the fact that the radio layer-2 utilizes Cyclic Redundancy Checks (CRCs) in the Random Access CHannel (RACH) procedure, such that there cannot be bit-errors.

However, the lack of integrity protection exposes user plane data to man-in-the-middle attacks even when encryption is enabled. It is possible in fact to change the data en route because encryption is linear (i.e., a stream cipher). In addition, an attacker can also inject rogue data into a session with the aim to either increase subscriber bills, or to waste resources carrying the data. These risks are due likely to increase with the increasing support of IoT devices in the 5G era. Even if data integrity can be provided at the transport or application layer (in addition to encryption), there may be cases in which transport or application security conflict with performance constraints, e.g. with respect to latency or battery life, and where a bearer-level integrity provides a useful compromise [14].

This requires that the 5G RAN should be able to ensure the confidentiality as well as the authenticity and integrity for the user plane. This means that ciphering and integrity protection for the UP shall at least be supported by both the access node (i.e., the gNB in 5G) and UE in so that it can be enabled by the network based on the particular usage scenario and on the required level of protection. For example, in a scenario where latency is critical, providing bearer-level protection is better than relying on transport or application security.

The support of UP security also has several security implications based on the point where it is terminated.

Considering the choices done in the previous generation network and from a mere security point of view, it is always better to have the UP protection terminated deeper in the network rather than closer to the edge of the network. One of the main objections against terminating security in the RAN is that the security endpoint would reside in an exposed location requiring additional security measures in the security endpoint and an additional UP security gateway located at the edge of the core network to ensure security protection over backhauling. This is what has happened with 4G where the termination of UP security in the eNB exposes UP traffic to interception and injection on the interfaces between the eNB and the core network (backhauling) when additional measures like IPSec are not implemented.

With the virtualization of RAN entities, more architectural options are possible for the 5G RAN. In particular the possibility in 5G to split the RAN into CUs and DUs, as described in Section 9.4.2, allows to terminate the PDCP layer, where security is performed, in the CU, being a more secure location allowing for UP traffic protection without the need for IPSec.

However, another point that requires security considerations for UP traffic protection are the gNB internal interfaces, as for instance the F1 interface connecting the CU with the DU, as explained in Section 9.4.2., for which integrity, confidentiality and anti-replay protection shall also be provided.

To minimize the susceptibility to attacks, considering that a gNB may in many cases be located in a vulnerable location, also the security procedures internal to the gNB shall be protected to avoid that an attacker may modify the gNB's settings or software configurations via local or remote access. In addition, sensitive data is stored on the gNB like keys, user data or user identifiers, which may be

obtained by an attacker. As also done for eNBs in LTE networks, a gNB needs moreover to provide a secure environment, i.e., a logical entity within the gNB that provides a trustworthy environment for the execution of sensitive functions, such as the encryption or decryption of user data or boot processes, and the storage of sensitive data like long-term cryptographic secrets and configuration data. This secure environment shall ensure protection and secrecy of all sensitive information and operations from any unauthorized access or exposure.

9.4.4.3 Protection Against User Plane Denial of Service Attacks

The support of UP security, i.e., data ciphering and integrity, is mandatory in 5G on both the UE and network side. However, this does not ensure that it will be used in practice since its activation should depend on the particular use case considered. For example, there could be cases where it is not enabled because security is provided at an application layer, or because the traffic sent is small and spurious, and not particularly security-sensitive. This can be the case for some IoT use cases

The lack of protection of user plane over the air can expose the security of the network anyway. In the absence of integrity protection being provided by AS security, an attacker can launch DoS attacks on the user plane by flooding the path towards the network node with bogus packets. Though the bogus packets may be identified and filtered at the network node that can verify packets based on the security context of a device, the path towards the network node can still be flooded by bogus packets. This would lead to denial of service or at least throughput degradation caused by congestion for the devices whose traffic shares the same network links as that of the bogus packets.

To counteract these attacks, the possibility should be investigated that 5G network elements embed DoS detection and mitigation functions into the 5G RAN, e.g., via key security indicators and the adoption of dynamic resolution action according to the monitored security indicators. The DoS detection functions would include a set of measurable security indicators, as for example the detection or identification of an anomaly pattern of devices continuously streaming uplink data beyond a certain threshold, or some indicator related to the functions that monitor and detect performance and threshold alarms.

9.4.4.4 Protection Against Signaling DoS Attacks

In 5G, DoS and DDoS attacks originating from a very large number of connected devices, as envisioned for some use cases, will leverage and possibly also target the RAN. Each transaction or traffic flow among the IoT devices, other mobile devices or the Internet results in CP signaling. Unnecessary connection establishment and release signaling could potentially overburden the radio and core network and reduce the quality of service (QoS) experienced by other services.

Several attacks could be carried out by compromising a large number of devices in specific geographical locations. The foreseen increased use in 5G of low-cost machine-to-machine (M2M) devices, characterized by thin operating systems with limited patching capabilities, increases the likelihood that these devices can be tampered and used to coordinate DoS and DDoS attacks against the RAN. These attacks can occur by performing a large number of simultaneous network access attempts in specific geographical locations with the aim to cause a signaling plane overload by exhausting the local radio resources of the network.

Other DoS attacks towards the RAN can occur by coordinating a large number of compromised devices, in specific physical locations, which have already been granted access to the network, to transfer very short data followed by periods of inactivity. This forces a continuous allocation of radio

and network resources to support the data transfer, and afterwards their release to allow new connections. These attacks not only deplete radio and network resources needed by new devices trying to establish new network connections, but also have an impact on the signaling plane in terms of processing and computation caused by the continuous allocation and release of the needed radio and network resources.

Current networks support an overload control mechanism that is triggered when an overload situation is detected. This mechanism is not able to distinguish between malicious devices and legitimate devices. Access is instead prevented for all new connections (malicious or not) until the overload has cleared. The effect is that legitimate devices are also not able to access the network. A more sophisticated access control mechanism is required for 5G which is able to recognize via inference data transfer patterns and network access request patterns which are not conformant and selectively targeting only those devices to be disconnected from the network. The use of analytical techniques like anomaly detection should be investigated for such analysis.

9.4.5 Service-level Security

The range of end-user services provided by 5G networks will typically rely on IP and Internet technologies, so the same considerations and mechanisms used to secure traditional web services apply, including, e.g., the use of cryptography or computation-based mechanisms using various types of private and public secret keys, etc.

9.4.6 A Control and Management Framework for Automated Security

Cloud computing, software-defined networks and the fact that security attacks are becoming more automated, reinforces the need for automated management and control of the telecommunication infrastructure. In particular, since a cloud-based paradigm promotes infrastructure that is highly accessible and shared by multiple users (e.g., telco operators), the concept of a highly secure network gains even more relevance. It is therefore of utmost importance to be able to provide robust, flexible and proactive mechanisms for detecting and preventing security issues, and to be able to perform this in real-time and in an automated fashion.

Along these lines, [15] proposes a real-time and automated security framework for the 5G telecommunication network, implementing a continuous and closed loop of real-time environment inspection, analytics, policy-based decisions and actuation/enforcement via cloud and SDN orchestration procedures, as illustrated in Figure 9-5.

Today, we can address a limited number of security threats, namely those that are currently already known to the security community, but we can't foresee the new and ever-changing threats that 5G networks will have to be protected against. However, we do have the basis to create autonomic network management solutions that should cope with them, being fed with insights from governed real-time analytics systems and actuating on network resources in order to minimize or prevent the effects of the detected threats in real-time.

In order to cater for these requirements, [15] proposes a control and management plane as depicted in Figure 9-6. It closely follows the ETSI NFV architecture [16], but introduces additional entities, depicted in grey below, and described in detail in [15].

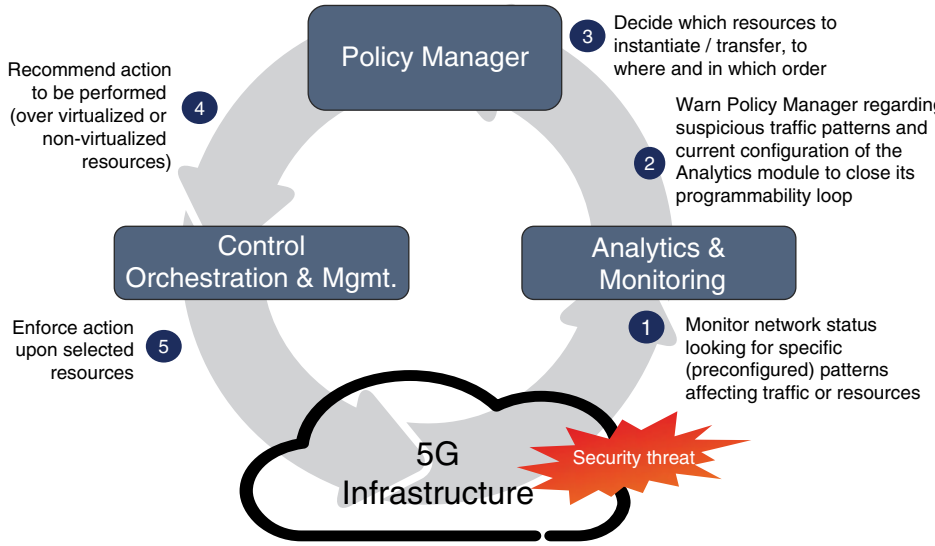

Figure 9-5. Automated 5G network security.

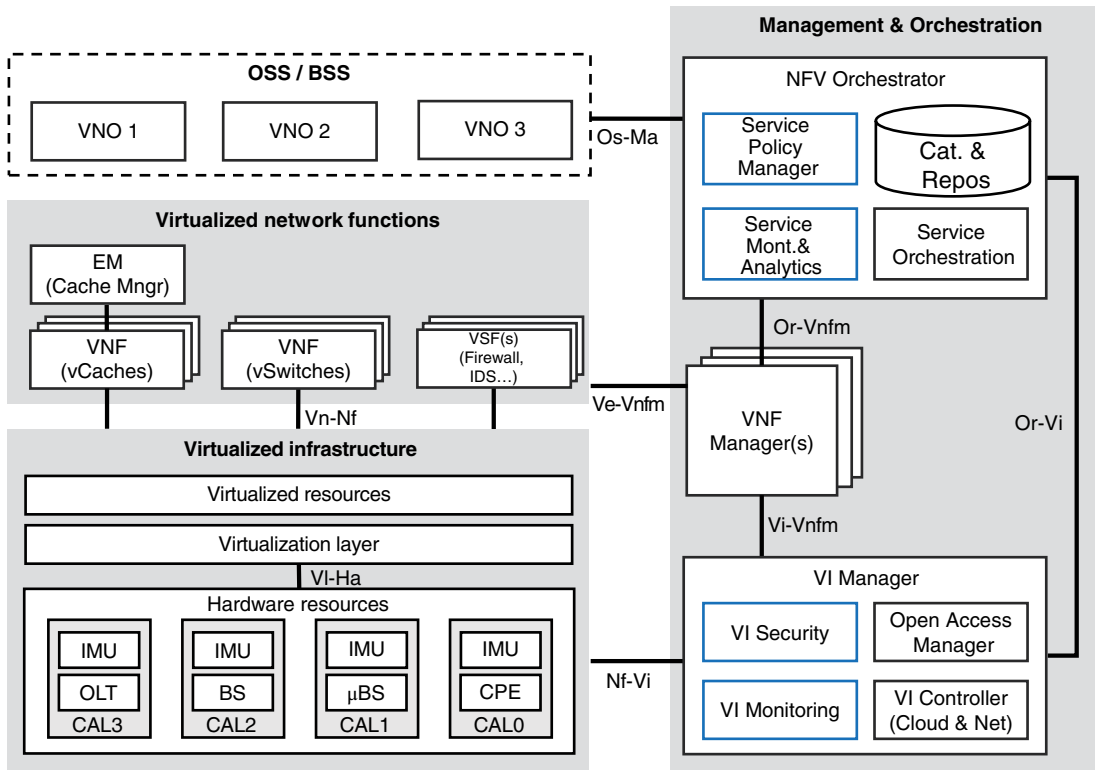

Figure 9-6. Possible control, management and orchestration architecture supporting automated security [15].

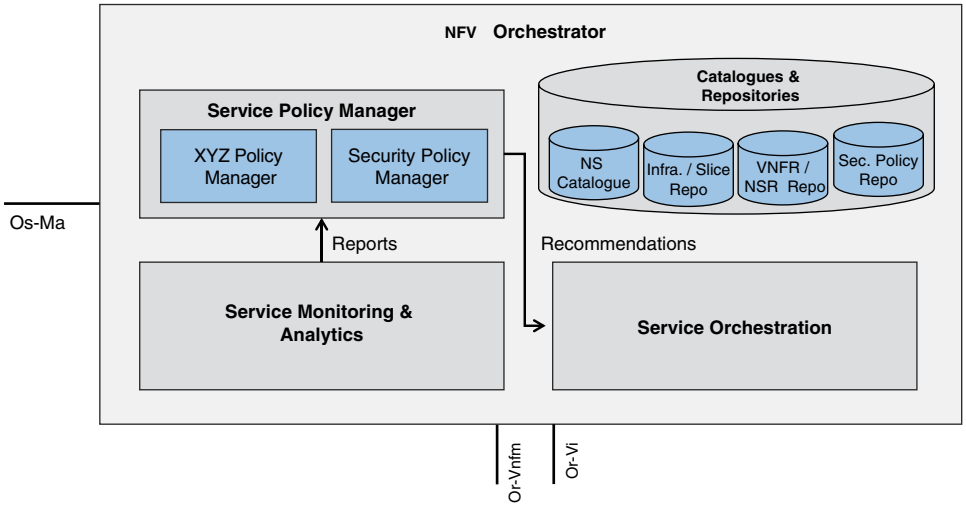

Figure 9-7. Details of NFV orchestrator.

The proposed control, management and orchestration (CMO) architecture for the 5G network consists of four groups of components: virtualized infrastructure (VI), VNFs, management and orchestration (MANO), and operations and business support systems (OSS/BSS).

Inside the MANO component, the NFV orchestrator is of special importance in order to implement the security management features of the 5G network, in particular (see also Figure 9-7):

- **Security Policy Manager (SPM):** This is in charge of making a recommendation about the best action to be taken next, taking as input events triggered by the Service Monitoring and Analytics function. The events are delivered as a result of monitoring and analyzing changes in the status of the resources. The SPM feeds recommendations into the Service Orchestration element which is in charge of enforcing these towards the network's virtualized or physical resources;
- **Service Monitoring & Analytics (SMA) component:** This is responsible for performing metrics and notifications acquisition from: i) the NFVI resources, ii) the VNFs or virtualized security functions (VSFs), and iii) the network's physical infrastructure. The NFVI resources include all physical and virtual compute, storage and network resources such as the compute resources required for the deployment VNFs. The SMA component consolidates the obtained metrics, produces events/alarms and communicates them to the Security Policy Manager. Based on these metrics, the Security Policy Manager can derive decisions and take actions in communication with the Service Orchestration component to perform changes to the network services that are already deployed, or instantiate and deploy new services;
- **Service Orchestration (Orchestrator):** The main responsibility of the Service Orchestration component is to manage the virtualized network services (NS) lifecycle procedures, according to the recommendations provided by the Security Policy Manager.

Taking these components into account, the SMA should be able to detect an attack and notify the SPM about it. The SPM has been provisioned with the appropriate instructions (i.e., policies) in

order to classify the attack and provide a best next action recommendation to the Service Orchestrator, which would finally enforce such recommendation upon the physical and virtualized resources composing the 5G network. Such action could, for instance, consist of deploying specific VSFs in order to neutralize the attack.

The goal of the previously described system is to automate the network security. However, it must be also noted that the control, management and orchestration systems are key network elements which require additional and flexible protection in itself. Policy-controlled network automation should also aim at this task.

9.5 Summary

5G networks represent both a challenge but also an opportunity from a security point of view. Among the challenges, we can mention the introduction of new actors, business models and use cases, which lead to very demanding network capabilities and security requirements. An additional challenge is the increasing number and sophistication of cyberthreats and attacks related to the new macroeconomic and geopolitical scenario.

Among the opportunities, there is the chance to design and build a network with architecturally inherent security and privacy features, donned with unprecedented capabilities for automation and self-adaptation.

References

1 ETSI GS NFV-SEC 001, "Network Functions Virtualisation; NFV Security; Problem Statement", 2014
2 5G PPP Architecture Working Group, White Paper, "View on 5G Architecture", v2.0, July 2017
3 3GPP TR 38.801, "Study on new radio access technology: Radio access architecture and interfaces", V1.1.0, March 2017
4 Y.-S. Shiu, S. Y. Chang, H.-C. Wu, S. C.-H. Huang and H.-H. Chen, "Physical layer security in wireless networks: a tutorial", IEEE Wireless Communications, vol. 18, no. 2, pp. 66–74, April 2011
5 A. Biryukov, A. Shamir and D. Wagner, "Real Time Cryptanalysis of A5/1 on a PC", International Workshop on Fast Software Encryption, Springer, 2000
6 D. Wagner, B. Schneier and J. Kelsey, "Cryptanalysis of the cellular message encryption algorithm", Annual International Cryptology Conference, Springer, 1997
7 C. Adams and S. Lloyd, "Understanding PKI: concepts, standards, and deployment considerations", Addison-Wesley Professional, 2003
8 T. Austin, "PKI: A Wiley Tech Brief", John Wiley & Sons, 2000
9 T.-H. Chang, Y.-W. P. Hong and C.-Y. Chi, "Training signal design for discriminatory channel estimation", IEEE Global Telecommunications Conference (GLOBECOM 2009), Dec. 2009.
10 I. Csiszar and J. Korner, "Broadcast channels with confidential messages", IEEE Transactions on Information Theory, vol. 24, no. 3, pp. 339–348, May 1978
11 3GPP TS 33.401, "3GPP System Architecture Evolution (SAE); Security architecture", V15.1.0, Sept. 2017
12 A. Shaik, R. Borgaonkar, N. Asokan, V. Niemi and J.-P. Seifert, "Practical attacks against privacy and availability in 4G/LTE mobile communication systems", 2015, see http://arxiv.org/abs/1510.07563

13 3GPP TS 36.133, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management", V15.0.0, Sept. 2017
14 3GPP TR 33.863, "Study on battery efficient security for very low throughput Machine Type Communication (MTC) devices", V14.2.0, June 2017
15 5G PPP CHARISMA project, Deliverable D3.2, "Initial 5G multi-provider v-security realization: Orchestration and Management", July 2016
16 ETSI GS NFV-MAN 001, "Network Functions Virtualization", v1.1.1, Dec. 2014

10

Network Management and Orchestration

Luis M. Contreras[1], Víctor López[1], Ricard Vilalta[2], Ramon Casellas[2], Raúl Muñoz[2], Wei Jiang[3], Hans Schotten[3], Jose Alcaraz-Calero[4], Qi Wang[4], Balázs Sonkoly[5] and László Toka[5]

[1] *Telefónica Global CTO Unit, Spain*
[2] *Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Spain*
[3] *German Research Center for Artificial Intelligence (DFKI), Germany*
[4] *University of the West of Scotland, United Kingdom*
[5] *Budapest University of Technology and Economics, Hungary*

10.1 Introduction

This chapter provides an insight into network management and orchestration in the 5th generation (5G), in particular highlighting how software-defined networking (SDN) and network function virtualization (NFV) will enable increased agility, scalability, and faster time-to-market of 5G communication networks.

SDN proposes the decoupling of both the control plane (CP) and user plane (UP), which are commonly integrated nowadays in the network elements (NEs), by logically centralizing the control while leaving the NEs to forward traffic and apply policies according to instructions received from the control side. This permits the network to become programmable in a way that facilitates more flexibility than traditional networks. On the other hand, NFV enables the dynamic instantiation of network functions (NFs) on top of commodity hardware, permitting the separation of the current vertical approach. This vertical approach consists of deploying integrated functional software and hardware for a given NF. Although they have emerged as separate innovative initiatives in the industry, both SDN and NFV are complementary, with the prevalent view in the industry that 'SDN enables NFV'.

Traditional telecommunications networks have been built relying on a diversity of monolithic hardware devices designed and manufactured by distinct vendors. This approach requires complex and static planning and provisioning from the perspective of the service and the network. This static and complex approach on how the network services have been conceived and deployed over the last decades has triggered a continuous process of re-architecting the network, tailoring topologies and capacity for the design and introduction of any new service in the network.

Current telecom networks require a rapid adaptation to forthcoming 5G services and demands, and if there is not an evolution of the conventional management and operation frameworks, it would

5G System Design: Architectural and Functional Considerations and Long Term Research, First Edition.
Edited by Patrick Marsch, Ömer Bulakçı, Olav Queseth and Mauro Boldi.

create difficulties to deploy the services fast enough. The carrier networks are usually multi-technology, multi-vendor and multi-layer, which translates into complex procedures for service delivery due to the different adaptations needed for the multiplicity of dimensions. In addition to that, the carrier networks are structured across regional, national and global infrastructures, motivating the need of managing and controlling a large number of physical NEs distributed over a multitude of locations. Furthermore, it is worth noting that the delivery of services implies the involvement of more than one single network domain (e.g., the access to contents not generated by the telecom operator), meaning that the interaction with other administrative domains is also critical.

Having networks built in the classical manner makes it tremendously difficult to cope with customized service creation and rapid delivery in very short times, as is expected to be required in 5G networks. A fundamental requirement identified by network operators' associations, such as Next Generation Mobile Networks (NGMN) [1], for 5G systems is to support flexible and configurable network architectures, adaptable to use cases that involve a wide range of service requirements. It is here where both network programmability and virtualization, leveraging on SDN and NFV, can solve (or at least mitigate) the complexity of the network management and orchestration needs for 5G.

The progressive introduction of both SDN and NFV into operational networks will introduce the necessary dynamicity, automation and multi-domain approach (with the different meanings of technology, network area or administration) to make the deployment of 5G services feasible. The target is to define management and orchestration mechanisms that allow for deploying logical architectures, consisting of virtual functions connected by virtual links, dynamically instantiated on top of programmable infrastructures. Undoubtedly, these new trends will change the telecom industry in many dimensions, including the operational, organizational and business ones [2] that should be carefully taken into account during the process of adoption of these new technologies.

The chapter is structured as follows. Section 10.2 introduces the main concepts of management and orchestration associated to SDN and NFV, with a review of the corresponding architecture frameworks. Section 10.3 profiles the main enablers for achieving the management and orchestration goals of 5G, through open and extensible interfaces, on one hand, and service and device models, on the other. Section 10.4 addresses the complexity derived from multi-domain and multi-technology scenarios. Section 10.5 describes the applicability of SDN to some of the scenarios foreseen in 5G, like the collapsed fronthaul and backhaul (known as Xhaul) and the transport networks. In Section 10.6, the main ideas of the role of NFV in 5G are stated. Section 10.7 provides insights about the autonomic network management capabilities in 5G. Finally, Section 10.8 summarizes the chapter.

10.2 Network Management and Orchestration Through SDN and NFV

The management and orchestration plane has an essential role in the assurance of an efficient utilization of the infrastructure while fulfilling performance and functional requirements of heterogeneous services. Forthcoming 5G networks will rely on a coordinated allocation of cloud (compute, storage and related connectivity) and networking resources. By resource, it can be considered any manageable element with a set of attributes (e.g., in terms of capacity, connectivity, and identifiers), which pertains to either a physical or virtual network (e.g., packet and optical), or to a data center (e.g., compute or storage).

For an effective control and orchestration of resources in both SDN and NFV environments, it is highly necessary to have proper levels of abstraction. The abstraction allows representing an entity in terms of selected characteristics, common to similar resources to be managed and controlled in the same manner, then hiding or summarizing characteristics irrelevant to the selection criteria. Through the abstraction of the resources, it is possible to generalize and to simplify the management of such resources breaking the initial barriers due to differences in the manufacturer, in particular aspects of the technology, or the physical realization of the resource itself.

The orchestration permits an automated arrangement and coordination of complex networking systems, resources and services. For such process, it is needed an inherent intelligence and implicitly autonomic control of all systems, resources and services.

In the case of NFV, orchestration is not formally defined, while, from the definition of the NFV orchestrator (NFVO), it can be assumed that this includes the coordination of the management of network service (NS) lifecycles, virtual network function (VNF) lifecycles, and NFV infrastructure (NFVI) resources, to ensure an optimized allocation of the necessary resources and connectivity. Similarly, for SDN, orchestration can be assumed to correspond to the coordination of a number of interrelated programmable resources, often distributed across a number of subordinate SDN platforms, for instance, per technology.

At the time of delivering a service, it will be needed to apply different levels of orchestration. On one hand, the resources that will be necessary to support a given service should be properly allocated and configured according to the needs of the service to be supported. This is known as *resource orchestration*. A resource orchestrator only deals with resource level abstraction and is not required to understand the service logic delivered by NFs, nor the topology that defines the relation among the NFs that are part of the service.

On the other hand, the *service orchestration* applies to the logic of the service as requested by the customer, identifying the functions needed to fulfill the customer request as well as the form in which these functions interrelate to provide the complete service. The service orchestrator will trigger the instantiation of the NFs in the underlying infrastructure in a dynamic way.

By the right combination of service and resource orchestration, the end-to-end (E2E) management and orchestration functionalities will be responsible for a flexible mapping of services to topologies of NFs, based on a dynamic allocation of resources to NFs and the reconfiguration of NFs according to changing service demands.

The next sub-sections generally introduce SDN and NFV frameworks in more detail.

10.2.1 Software-Defined Networking

While networks are based on distributed CP solutions, there is a huge interest around SDN orchestration mechanisms that enable not only the separation of UP and CP, but also the automation of the management and service deployment process. Current SDN approaches are mainly focused on single-domain and single-vendor scenarios (e.g., data centers). However, there is a need of SDN architectures for heterogeneous networks with different technologies (e.g., IP, MPLS, Ethernet, and optical), which are extended to cover multi-domain scenarios.

The SDN architecture, as defined by the Open Networking Foundation (ONF) in [3], is composed of an application layer, a control layer, and an infrastructure layer, as depicted in Figure 10-1. User or provider-controlled applications communicate with the SDN controller via an Application-Controller Plane Interface (A-CPI), also known as *northbound interface* (NBI). The controller is in charge of

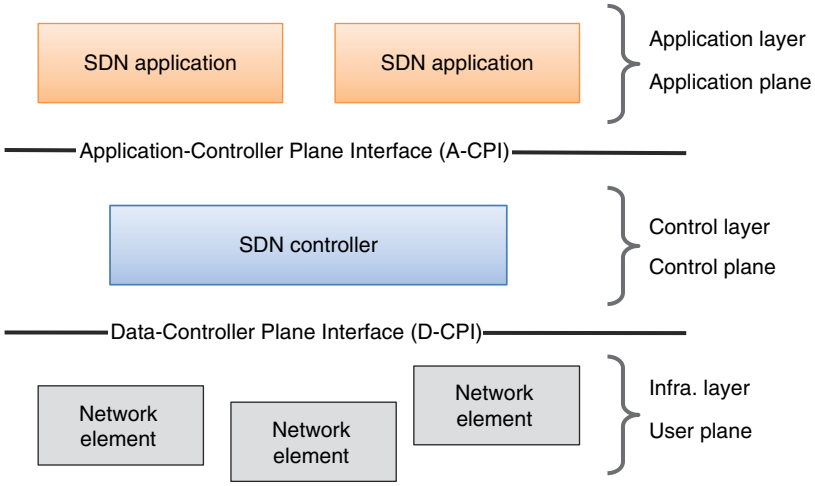

Figure 10-1. Abstract view of basic SDN components.

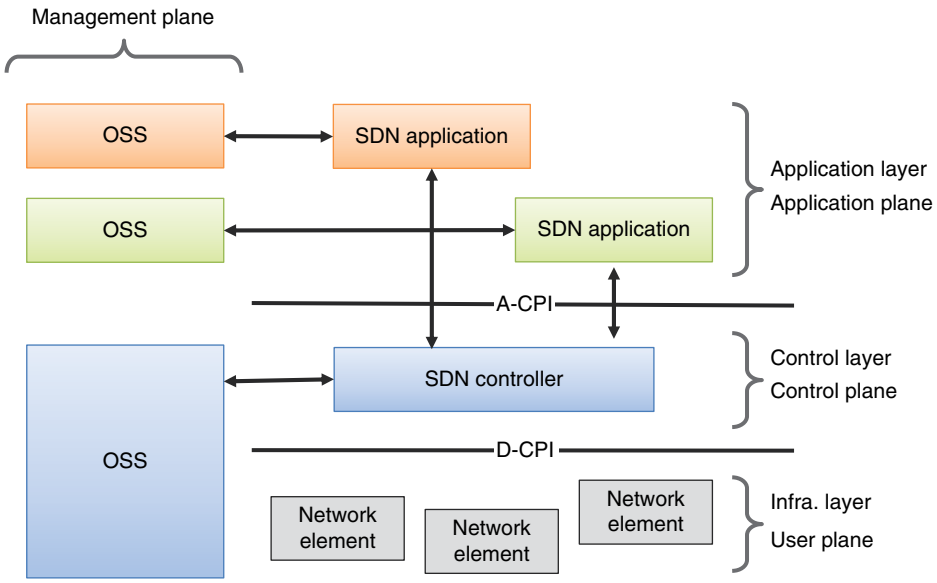

Figure 10-2. Abstract SDN architecture overview.

orchestrating the access of the applications to the physical infrastructure (i.e., the NEs), using a Data-Controller Plane Interface (D-CPI), also known as *southbound interface* (SBI).

Figure 10-2 presents a more descriptive view of a typical SDN architecture, where a management plane is also included, to carry out tasks such as registration, authentication, service discovery, equipment inventory, fault isolation, etc. In addition, Figure 10-3 shows the situation where the infrastructure owner gives away control of part of its infrastructure to a number of external entities.

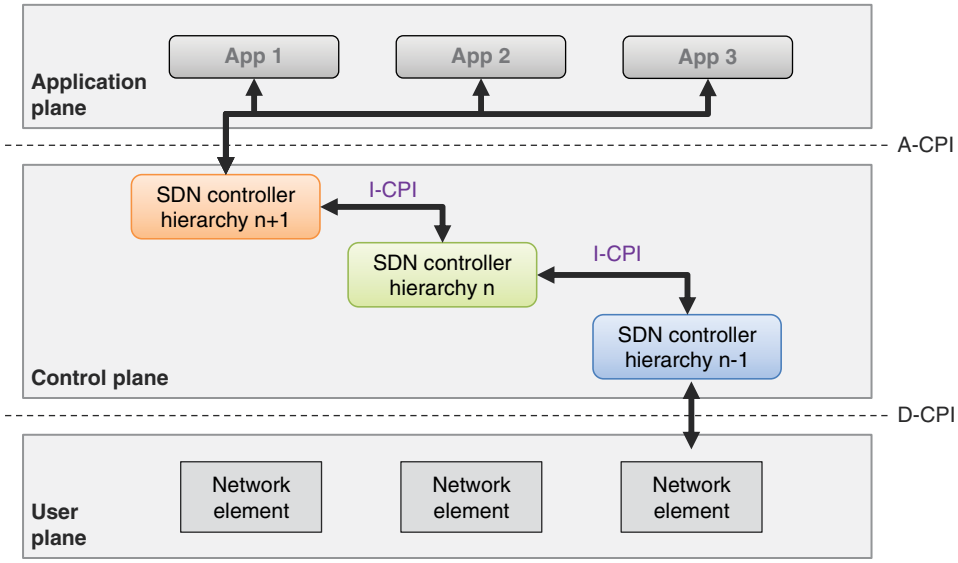

Figure 10-3. Recursive hierarchical SDN architecture.

This is relevant to scenarios where a network provider gives controlled access to equipment (or a slice of equipment through virtualization mechanisms) to some other service providers.

ONF also describes the possibilities of implementing hierarchical controllers, primarily for scalability, modularity or security reasons. Such hierarchical control structure introduces a new interface, the Intermediate-Controller Plane Interface (I-CPI), as shown in Figure 10-3. This hierarchical structure allows for recursiveness and to assure scalability, while maintaining the control of each domain in separate controllers.

In terms of functionalities, there are four main capabilities in this kind of interfaces enabling the flexible control and orchestration of different resources. Such capabilities are: (1) network topology extraction and composition, (2) connectivity service management, (3) path computation, and (4) network virtualization.

The need of network topology extraction and composition is to export the topological information with unique identifiers. Such network identifiers (such as IPv4 addresses or datapath-IDs) are required for the other functionalities. To compose the topology, it is required to export the nodes and the links in a given domain, which can be physical or virtual, as well as some parameters like the link utilization or even information about physical characteristics of the link if the operator requires the deployment of very detailed services.

The second functionality is to manage connectivity services. The operations on these services are the setup, tear down, and the modification of connections. Such services can be as basic as a point-to-point connection between two locations. Nonetheless, there are scenarios where the orchestration requires more sophistication like *(a)* exclusion or inclusion of nodes and links, *(b)* definition of the protection level, *(c)* definition of traffic-engineering (TE) parameters, like delay or bandwidth, or *(d)* definition of disjointness from another connection.

The third function is the path computation, which is fundamental as it provides the capability of properly defining an E2E service. For instance, when different controllers in a multi-domain environment are considered (e.g., in situations where multiple network segments are under a single administration, such as backhaul, metro and core networks), this permits to interact with individual controllers in each domain that are only able to share abstracted information that is local to their domain. The orchestrator with its global end-to-end view can improve end-to-end connections that individual controllers cannot configure. Without a path computation interface, the orchestrator is limited to carrying out a crank-back process that would not find proper results. This can be exploited as well when multiple technologies are considered, following a multi-layer decision approach.

Lastly, a network virtualization service allows to expose a subset of the network resources to different tenants. This advances in the direction of network slicing, where resources and capabilities of the underlying physical transport network can be offered to different users or tenants to appear as dedicated in its global network slice composition, as detailed in Chapter 8.

The ONF architecture presented here illustrates the general enablers for the objective of network programming. However, several other organizations are working on the standardization of NBIs and SBIs. In terms of maturity, there is not yet a complete solution for each model, but multiple candidate technologies for some interfaces. This is commented later on in this chapter.

10.2.2 Network Function Virtualization

European Telecommunications Standards Institute (ETSI) NFV is the most relevant standardization initiative arisen in the NF virtualization arena. It was incepted at the end of 2012 by a group of top telecommunication operators, and has rapidly grown up to incorporating other operators, network vendors, information and communications technology (ICT) vendors, and service providers. To date, the ETSI NFV industry specification group (ISG) can count on over 270 member companies. It represents a significant case of collaboration among heterogeneous and complementary kinds of expertise, in order to seek a common foundation for the multi-facet challenges related to NFV towards a solution as open and scalable as possible.

The ETSI NFV roadmap initially foresaw two major phases: The first one was completed at the end of 2014, where a number of specification documents were issued [4], covering functional specification, data models, proof-of-concept (PoC) description, etc. The second phase released a new version of the ETSI NFV specification documents. A third phase is ongoing at the time of writing, progressing the work on architectural and evolutionary aspects. The work of the ISG is further articulated into dedicated working groups (WGs). In phase 1, three WGs have been created, dealing with NFVI, Management and Orchestration (MANO), and Software Architectures (SWA). In phase 2, two additional WGs were spawned, dealing with Interfaces and Architecture (IFA) and Evolution and Ecosystem (EVE).

The currently acting specification of the ETSI NFV architecture was finalized in December 2014 [5], and its high-level picture is shown in Figure 10-4.

The ETSI NFV specification defines the functional characteristics of each module, their respective interfaces, and the underlying data model. The data model is basically made up by static and dynamic descriptors for both virtual network functions (VNFs) and network services (NSs). The latter are defined as compositions of individual VNFs, interconnected by a specified network forwarding graph, and wrapped inside a service.

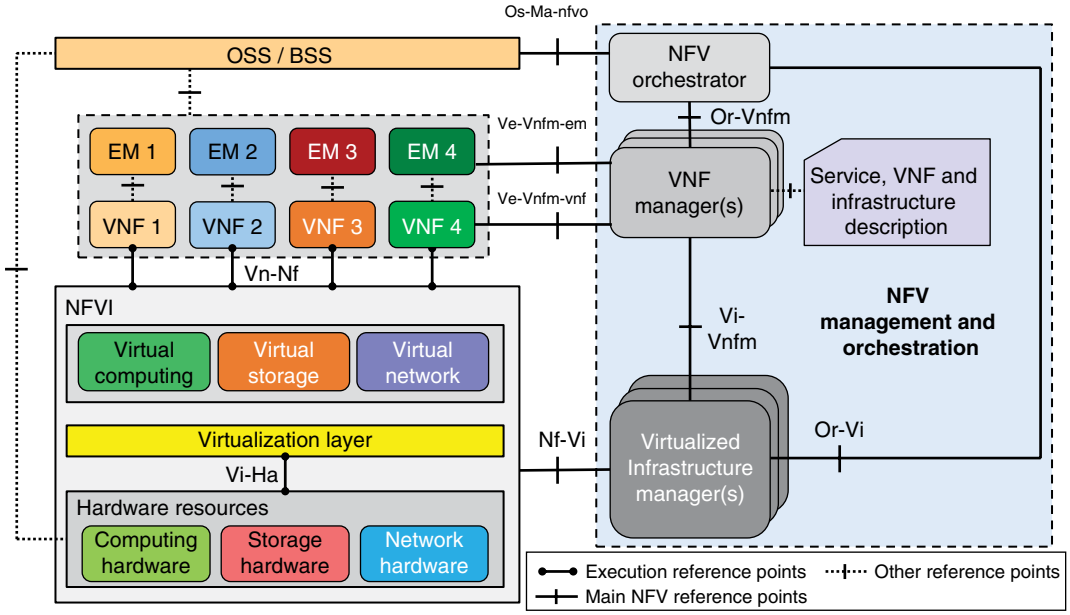

Figure 10-4. ETSI NFV architecture [5].

The ETSI NFV framework specifies the architectural characteristics common to all the VNFs. It does, though, not rule out which specific network functions can or should be virtualized, leaving this decision up to the NF provider (apart from the use cases advised for the proofs of concept).

The ETSI NFV architecture supports multi-point-of-presence (PoP) configurations, where a PoP is defined as the physical location where a NF is instantiated. A PoP can be mapped to a data center or a data center segmentation isolated from the rest of world.

A summary description of the modules in the ETSI NFV architecture is given in Table 10-1.

As it can be observed in Figure 10-4, the ETSI NFV framework assumes the existence of an outside operations support system (OSS) or business support system (BSS) layer in charge of the basic equipment and service management.

It is worth to mention that starting 2016, ETSI has launched the Open Source Mano (OSM) initiative [6]. OSM intends to develop an open-source NFV Management and Orchestration (MANO) software stack aligned with ETSI NFV specifications. This kind of open-source software initiative can facilitate the implementation of NFV architectures aligned to ETSI NFV specifications, increasing and ensuring the interoperability among NFV implementations.

10.3 Enablers of Management and Orchestration

The management and orchestration capabilities offered by SDN and NFV should be sustained by some enablers from the resource and service perspective. On one hand, there is a need for open and standard interfaces that could permit at the same time aspects like: (1) a uniform and homogeneous

Table 10-1. Components of the ETSI NFV framework.

Virtualized network function (VNF)	Virtualized instance of an NF traditionally implemented on a physical network appliance.
Element management (EM)	Component performing the typical network management functions (Fault, Configuration, Accounting, Performance and Security - FCAPS) requested by the running VNFs.
NFV infrastructure (NFVI)	Totality of hardware/software components building up the environment in which VNFs are deployed, managed and executed. Can span across several locations (physical places where NFVI-PoPs are operated). Include the network providing connectivity between such locations.
Virtualized infrastructure manager (VIM)	Provides the functionalities to control and manage the interaction of a VNF with hardware resources under its authority, as well as their virtualization. Typical examples are cloud platforms (e.g., OpenStack) and SDN controllers (e.g., OpenDaylight).
Resources	Physical resources (e.g., computing, storage, and network). Virtualization layer.
NFV orchestrator (NFVO)	Component in charge of orchestration and management of NFVI and software resources, and provisioning of network services on the NFVI.
VNF manager	Component responsible for VNF lifecycle management (e.g., instantiation, update, query, scaling, and termination). Can be 1-1 or 1-multi with VNFs.

access to the resources and services; and (2) an easy integration with supporting systems like OSS and BSS. On the other hand, a set of information and data models that could help to easily and flexible define, configure, manage and operate services and network elements in a consistent and abstract way.

10.3.1 Open and Standardized Interfaces

Through the existence of controllers allowing the programmability of the network, the operational goal is to facilitate the creation and definition of new services to be configured in the network and automatically, via OSS or directly by means of the interaction with tailored applications. The SDN controller will in this context take care of performing all the tasks needed to set up the configuration in the network (i.e., calculate the route from source to destination, check the resource availability, set up the configuration to apply in the equipment, etc.). For example, the inventory system can be better synchronized with the network, so that the provisioning can be done based on the real status of the network, avoiding any misalignment between the planning process and the deployment process.

One of the expected benefits of SDN is to speed-up the process to integrate a new vendor, or a new OSS system or application in the network. To do so, it is necessary to have standard NBI interfaces towards the OSS systems (e.g., network planning tools, inventory data bases, and configuration tools), and standard SBI interfaces towards the network element that depend only on the technology (e.g., microwave wireless transport, Metro-Ethernet/IP, or optical) and not on the vendor.

Nowadays, even for a single transport technology, the particularities per vendor implementation force a constant customization of the service constructs. This affects not only the provisioning phase,

but also the operation and maintenance of the services. Activation tools (as part of current OSS and BSS) are in some cases present, being in charge of the automated configuration of network services. However, the configuration is provided by vendor-dependent interfaces, and when a service needs to be extended by configuring different network segments, the configuration process needs to be done in each network separately, and usually by means of specific or dedicated systems. For the same reason, integrating a new vendor or new equipment (or a new release of an existing vendor or equipment) is time-consuming, and needs upgrades of the interfaces and changes in the OSS tools already deployed. It delays the introduction of new technologies, de facto blocking the transformation process towards 5G with the agility and flexibility needed by the operator. All of this renders the adoption of open and extensible interfaces, for both NBI and SBI, necessary.

Currently, there is no real progress about the definition of NBIs from the orchestrator perspective that could facilitate the smooth integration referred to before with respect to OSS and BSS. All the available NBIs are platform-dependent; in consequence, there is not a common or general approach in the industry by now. However, for the SBI there is some consensus.

For the programmability and management of the network, both NETCONF and YANG, as introduced in the following, are being recognized as future-proof options.

NETCONF [7] provides a number of powerful capabilities for a uniform configuration and management of network elements. It is transport protocol independent, meaning that it does not impose restrictions for getting access towards the devices. With NETCONF, it is possible to have a separation of the configuration data from the operational data in such a way that the administrator can set some variables from features like statistics, alarms, notifications, etc. In addition to that, thanks to the support of transactional operations, it is possible to ensure the completion of configuration tasks even on a network basis. Since NETCONF supports an automated ordering of operations, the sequential actions on the network can be defined, facilitating straightforward rollback operations if needed. NETCONF is then foreseen as the manner of managing and orchestrating multi-vendor infrastructures. However, NETCONF only defines the mechanisms to access and configure the network elements, but not the configuration information to be applied.

In this sense, YANG [8], as a data modeling language, complements NETCONF by defining the way in which the information applicable to a node can be read and written. It provides well-defined abstractions of the network resources that can be configured or manipulated by a network administrator, including both devices and services. The YANG language simplifies the configuration management as it supports capabilities like the validation of the input data, and data model elements are grouped and can be used in a transaction, etc. Nowadays, there is an intensive work in the definition of general and standard YANG models especially in the Internet Engineering Task Force (IETF), but not only. Figure 10-5 presents the evolution in the number of YANG models being proposed.

Similar to NETCONF, the RESTCONF protocol [9] provides a programmatic interface for create, read, update and delete (CRUD) operations accessing data defined in YANG based on Hypertext Transfer Protocol (HTTP) transactions, allowing Web-based applications to access the configuration data, state data, data-model-specific remote procedure call (RPC) operations, and event notifications within a networking device, in a modular and extensible manner. The purpose is then similar to the one described for NETCONF.

Regarding the orchestration of services and the management of VNF lifecycles, Topology and Orchestration Specification for Cloud Applications (TOSCA) emerges as the more solid option. ETSI NFV ISG is considering it as a description language, and recently started the specification of

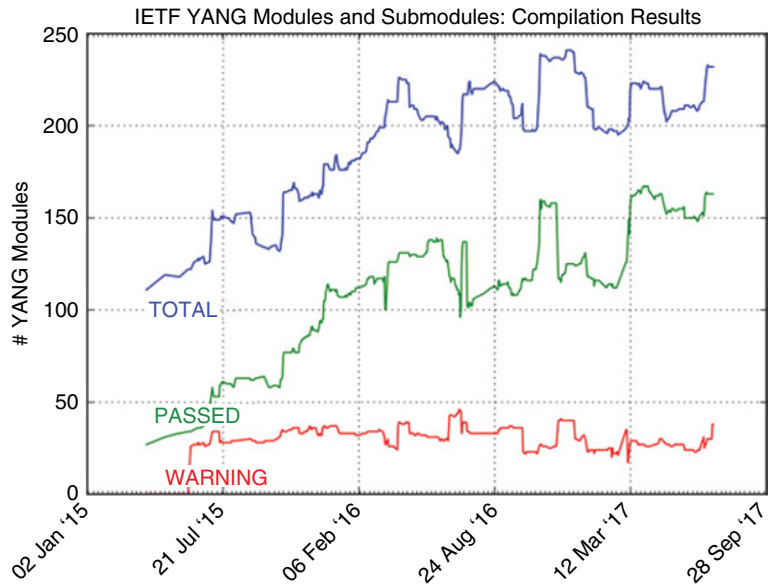

Figure 10-5. Development of YANG modules in IETF [10].

TOSCA-based descriptors [11], not yet being released at the time of writing. Nevertheless, a TOSCA template [12] is available, which is specifically designed to support describing both NS descriptors (NSDs) and VNF descriptors (VNFDs).

TOSCA is a service-oriented description language to describe a topology of cloud-based web services, their components, relationships, and the processes that manage them, all by the usage of templates. TOSCA covers the complete spectrum of service configurations, like resource requirements and VNF lifecycle management, including the definition of workflows and FCAPS management of VNFs. By this way, an orchestration engine can invoke the implementation of a given behavior when instantiating a service template.

A topology template defines the structure of a service as a set of node templates and relationships that together define the topology model as a (not necessarily connected) directed graph. Node and relationship templates specify the properties and the operations (via interfaces) available to manipulate components. The orchestrator will interpret the relationship template to derive the order in which the components of the service should be instantiated. TOSCA templates could also be used for later lifecycle management operations like scaling or software update.

From the point of view of communication method, TOSCA uses a simple REST application programming interface (API).

NETCONF/YANG and TOSCA can complement each other. Basically, the lifecycle management of the VNFs can be performed by means of TOSCA, while the VNFs can be dynamically configured at runtime by means of NETCONF/YANG. This interplay is facilitated by architectural propositions like the integrated SDN control for tenant-oriented and infrastructure-oriented actions in the framework of NFV, as described in [13]. Figure 10-6 shows the positioning of the two different levels of SDN control.

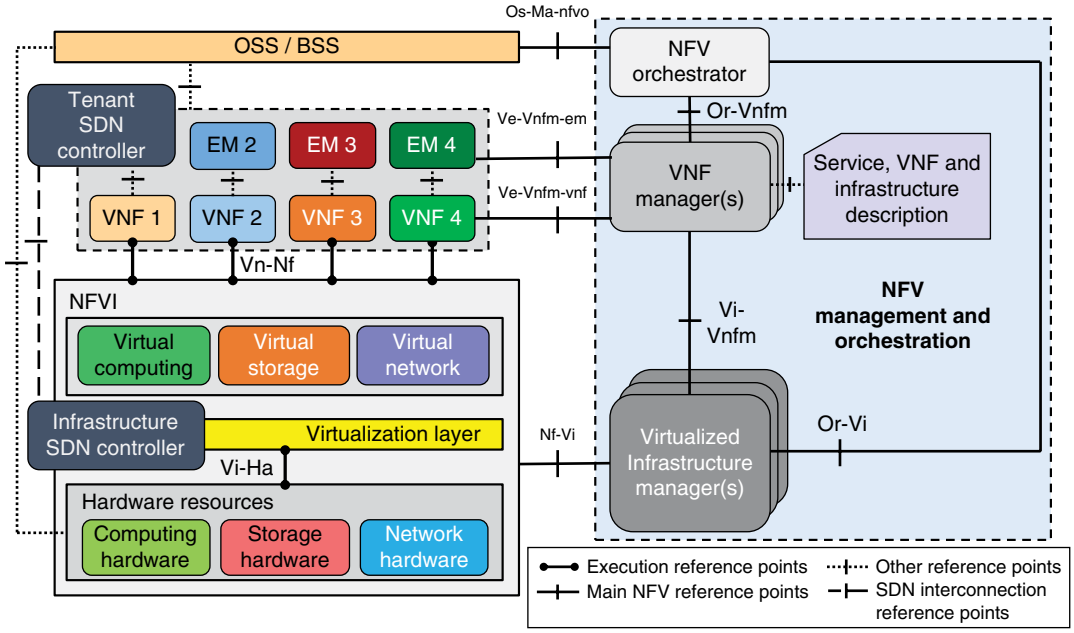

Figure 10-6. Infrastructure and tenant SDN controllers in the NFV architecture.

The SDN controller in the tenant domain can configure on-demand of NETCONF/YANG the functionality of the VNFs deployed by using TOSCA.

Furthermore, this architecture facilitates the integration of control and orchestration actions with a SDN controller at the infrastructure level for coordinating actions allowing cross-layer coordination. Both controllers manage and control their underlying resources via programmable southbound interfaces, each of them providing a different, but complementary, level of abstraction. This concept is leveraged from [14].

10.3.2 Modeling of Services and Devices

The same need for standardization as highlighted before would be also necessary for services and devices. By expressing a service to be deployed in a standard manner, it is possible to make it independent or agnostic of the actual underlying technology in which it is engineered. This provides more degrees of freedom for the decisions about how to implement a given service, and also allows for portability of such service across platforms.

Via those models, a unique entity can process all the service requests, later on triggering actions in the network for service delivery and deployment. Such an entity can be seen as a service orchestrator, which can maintain a common view across all the services deployed, instead of the legacy approach of siloed services, which renders a combined planning difficult. With such service orchestrator, dependencies can be detected in advance, allowing to improve the design by means of a coordinated usage of resources.

Similarly, the definition of common models for the same type of device simplifies the management, operation and control of the nodes in the network. A common representation of node capabilities and parametrization produce homogeneous environments removing the particularities that motivate onerous integration efforts as happens today to handle per-vendor specificities.

A generic reference about service models can be found in [15] and [16].

10.4 Orchestration in Multi-Domain and Multi-Technology Scenarios

10.4.1 Multi-Domain Scenarios

When talking about multi-domain, different meanings can be associated to the term *domain*. For instance, this can refer to different technologies, like packet, optical, microwave, etc., or different network segments. Finally, multi-domain can be understood as a multi-operator environment, with the interaction of different players for the E2E provision of a service. We use the term multi-domain for multi-operator environments and multiple administrative scenarios in this section. The importance of analyzing such scenarios was firstly raised in [17].

5G is expected to operate in highly heterogeneous environments using multiple types of access technologies, leveraging multi-layer capabilities, supporting multiple kinds of devices, and serving different types of users. The great challenge is to port these ideas to the multi-domain case, where the infrastructure (considered as network, computing, and storage resources), or even some of the necessary network functions, are provided by different players, each of them constituting a separate administrative domain.

Multi-operator orchestration requires the implementation of an E2E orchestration plane able to deal with the interaction of multiple administrative domains (i.e., different service and/or infrastructure providers) at different levels, providing both resource orchestration and service orchestration. An example would be the case of service providers offering their NFVI PoPs to host service functions of other providers, or even offering VNFs to be consumed by other service providers. However, existing interconnection approaches are insufficient to address the complexity of deploying full services across administrative domains. For instance, evolved interconnection services demanding, e.g., computing capabilities for the deployment of network building blocks as VNFs, or even inserting VNFs in the UP, cannot be satisfied with existing solutions for multi-domain environments.

An inter-provider environment imposes additional needs to be offered and served between providers like service-level agreement (SLA) negotiation and enforcement, service mapping mechanisms (in order to assign proper sliced resources to the service instance), reporting of assigned resource and service metrics, and allocation of proper control and management interfaces, to mention a few.

From the architecture perspective, an orchestration approach assuming a hierarchical top-level orchestrator playing the role of broker, with total visibility of the all providers' networks, and with the capability of orchestrating services across domains is certainly impractical, due to issues like scalability, trustiness between providers, responsibilities, etc. Instead, a peer-to-peer architecture seems to be more adequate for this kind of scenarios, as it already exists nowadays in the form of the pure interconnection for IP transit and peering.

From the point of view of SDN architecture, a primary approach to such a peer-to-peer relationship is provided by ONF in [18], which introduces an initial idea about the interaction of peer controllers,

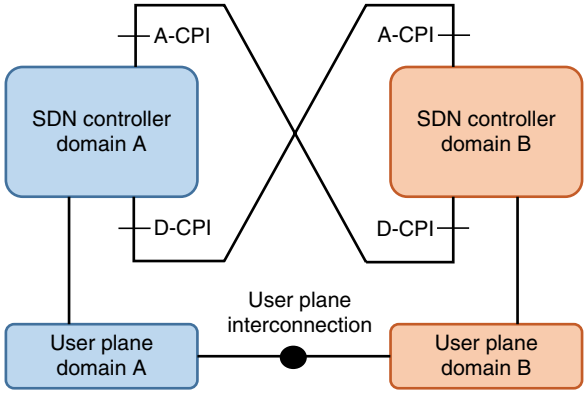

Figure 10-7. Peer controllers in the ONF architecture.

as reflected in Figure 10-7. Here, basically, each of the controllers may act as client to invoke services from the other as server, whereby A-CPI is the Application-Controller Plane Interface, and D-CPI the Data-Controller Plane Interface. The relationship among controllers is then proposed to be equivalent to a hierarchical provider/customer relationship.

For more complex orchestration scenarios, involving the provision of NFV-related services across providers, some other initiatives are in progress. To this respect, ETSI has produced a report on the description of architectural options for multi-domain [19], taking as basis for the analysis some use cases like NFVI-as-a-Service.

The Metro Ethernet Forum (MEF) Lifecycle Service Orchestration (LSO) is another initiative in the standardization arena, with a reference architecture defined in [20]. The MEF LSO architecture oversees the E2E orchestration of services where all the network domains require coordinated management and control. A shared information model for connectivity services is under definition, including the service attributes defined in MEF service specifications. Specifically, two inter-provider reference points are being proposed:

- **LSO Sonata**, which facilitates the interconnection of the BSS functions of different providers, addressing the business interactions between those providers. This includes aspects such as ordering, billing, trouble ticketing, etc.;
- **LSO Interlude**, which instead facilitates the interconnection of the OSS functions of different providers. Interlude supports control-related management interactions between two service providers and is responsible for the creation and configuration of connectivity services as permitted by service policies. It also covers notifications and queries on the operational state of services and their performance.

Co-operation between providers then takes place at the higher level, based on exchanging information, functions and control. These interfaces serve for the business-to-business and operations-to-operations relations between providers.

In addition, the 5G PPP 5G-Exchange (5GEx) project [21] has developed a multi-domain orchestration framework enabling the trading of NFs and resources in a multi-provider environment, and targeting a Slice-as-a-Service (SaaS) approach. The envisioned 5G service model is an evolution of the ETSI NFV model, proposing extensions to it. The original NFV paradigm foresees that resources used inside a service (for instance, for different VNF components) can be distributed over distinct

PoPs (i.e., physical infrastructure units, typically data centres). However, the PoPs are supposed to be under a unique administration. Furthermore, the level of control is quite limited outside the perimeter of the data centers as for instance in wide area networks (WANs). The project addressed these limitations, aiming at functionally overcoming them (i.e., enabling the integration of multiple administrative domains) and at least assessing the non-functional enablers needed to make actual business out of the technology.

5GEx has built on top of the concept of logical exchange for a global and automatic orchestration of multi-domain 5G services. A number of interfaces implement such kind of exchange for the CP perspective. This ecosystem allows the resources such as networking, connectivity, computing and storage in one provider's authority to be traded among federated providers using this exchange concept, thus enabling service provisioning on a global basis.

Figure 10-8 presents a high-level overview of the 5GEx architecture. Different providers participate in this ecosystem, each of them representing a distinct administrative domain interworking through multi-domain orchestrators (MdOs) for the provision of services in a multi-provider environment. This architecture extends the ETSI MANO NFV management and orchestration framework for facilitating the orchestration of services across multiple administrative domains. Each MdO handles the orchestration of resources and services from different providers, coordinating resource and/or service orchestration at multi-provider level, and orchestrating resources and/or services using domain orchestrators belonging to each of the multiple administrative domains.

The domain orchestrators are responsible of performing virtualization service orchestration and/or resource orchestration exploiting the abstractions exposed by the underlying resource domains that cover a variety of technologies hosting the actual resources.

There are three main interworking interfaces and APIs identified in the 5GEx architecture framework. The MdO exposes service specification APIs that allow business customers to specify their requirements for a service on business-to-customer (B2C) interface I1. The MdO interacts with other MdOs via business-to-business (B2B) interface I2 to request and orchestrate resources and services across administrative domains. Finally, the MdO interacts with domain orchestrators via interface I3 APIs to orchestrate resources and services within the same administrative domains.

Figure 10-9 presents the functional detail of the proposed architecture, showing different components identified as necessary for multi-domain service provision. In this case, all the providers are considered to contain the same components and modules, although in Figure 10-9 the complete view is only shown for the provider on the left for simplicity.

We briefly describe some of the components in the figure, particular to 5GEx.

- The **inter-provider NFVO** is the NFVO implements multi-provider service decomposition, responsible of performing the end-to-end network service orchestration. The network services operator (NSO) and resource orchestration (RO) capabilities are contained here;
- The **topology abstraction** module performs topology abstraction, elaborating the information stored in the resource repository and topology distribution modules;
- The **topology distribution** module exchanges topology information with its peer MdOs;
- The **resource repository** that keeps an abstracted view of the resources at the disposal of each one of the domains reachable by the MdO;
- The **SLA manager** is responsible for reporting on the performance of its own partial service graph (its piece of the multi-domain service);
- The **policy database** which contains policy information;

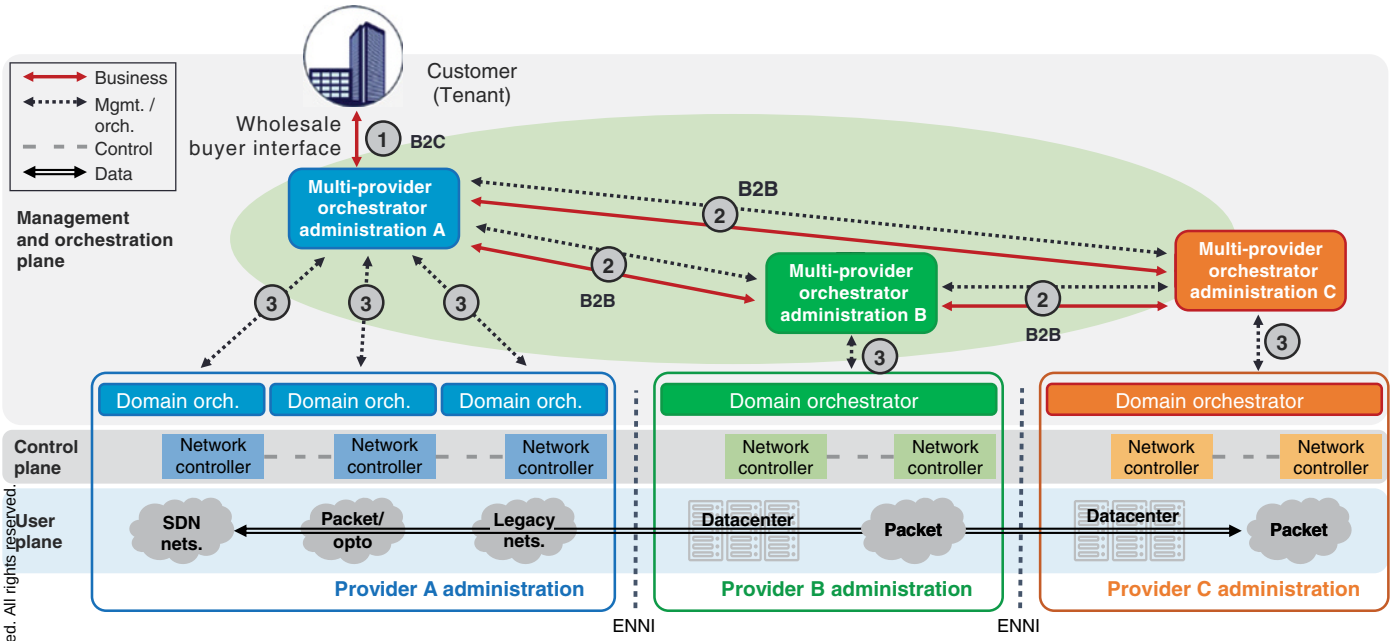

Figure 10-8. 5GEx reference architectural framework [21].

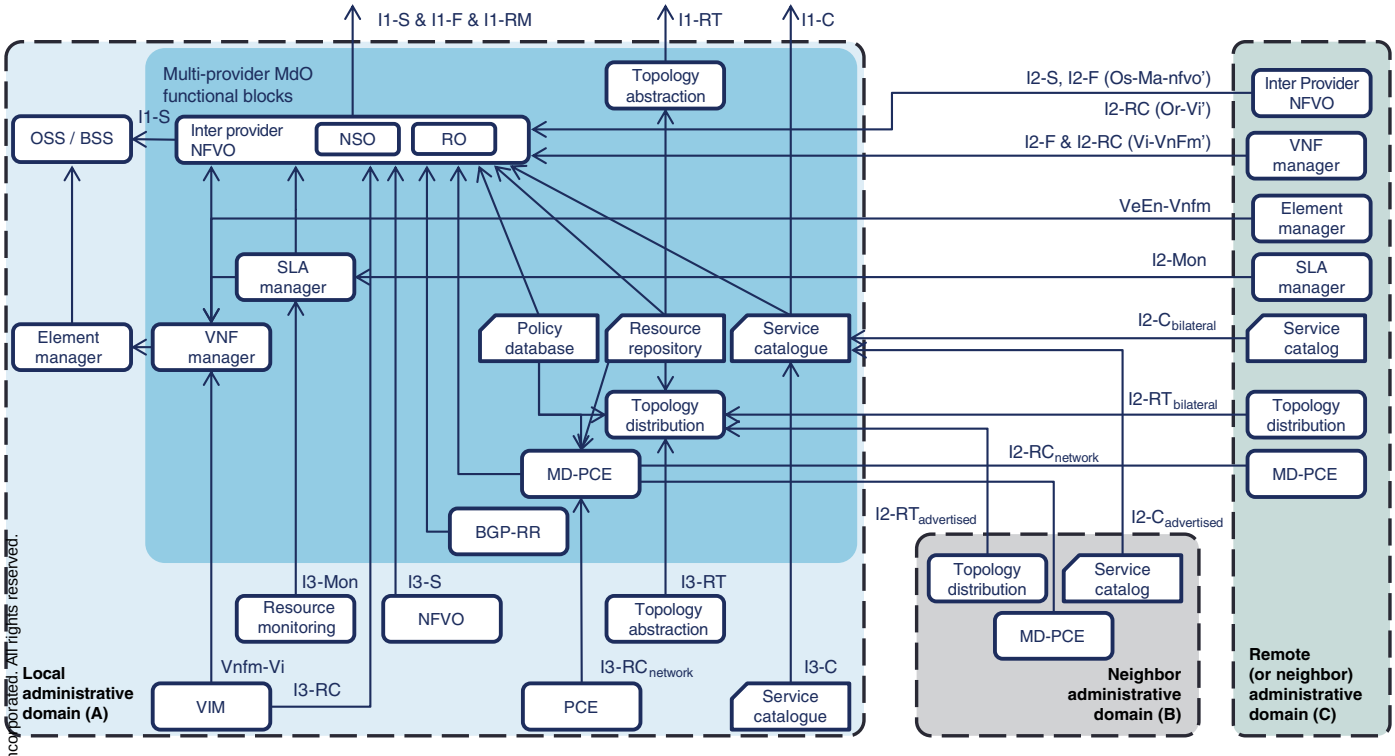

Figure 10-9. Functional architecture of 5GEx [21].

- The **resource monitoring** module dynamically instantiates monitoring probes on the resources of each technological domain involved in the implementation of a given service instance;
- The **service catalog** in charge of exposing available services to customers and to other MdOs from other providers;
- The **MD-PCE** (multi-domain path computation element) devoted to make the necessary path computations and to set up the connection between domains.

From the interfaces perspective, the functional split considered is related to service management (-S functionality), VNF lifecycle management (-F), catalogs (-C), resource topology (-RT), resource control (-RC) and monitoring (-Mon). Table 10-2 summarizes the functional needs for the mentioned interfaces as well as potential candidate solutions for their implementation. At the time of writing, the identification and specification of these interfaces is currently being defined and will be fully described in future deliverables of the project.

Figure 10-9 shows the interconnection of MdOs for three different domains. As mentioned, the left MdO is shown with full details, while the other two are skipped for simplicity. The 5GEx interfaces are presented with the corresponding functional split. The interfaces have to be considered as symmetric, since the consumer-provider role is situational in an exchange.

The left MdO is the entry point for the service request coming from the customer, through the I1 interface. Using I1-C, I1-S and I1-F, the customer (e.g., an infotainment company) will be able to request VNF instantiation and configuration, apart from expressing the way in which they are interconnected by means of a service graph.

The service will be decomposed by the NFVO of the provider A. If the service cannot be honoured by the sole use of its own resources, the NFVO will make use of resources offered by other providers in the exchange. The availability of resources from other parties is collected via I2-RT, and the availability of services offered by such parties is obtained through I2-C. Once the decision about using resources from other providers is taken, the left MdO will make use of I2-S and I2-RC for requesting and controlling the necessary resources and services. The same MdO will make use of the I3 interface for governing the own resources accordingly, in a similar manner.

In order to accomplish the negotiated SLA between the parties (i.e., both the customer and the entry provider, and the providers participating in the E2E service provision), convenient monitoring capabilities are deployed, using I1-Mon, I2-Mon and I3-Mon for the respective capabilities.

Table 10-2. Functional split of 5GEx interfaces and candidate solutions [21].

	Functional split	I1 (Customer to provider)	I2 (Inter-provider)	I3 (Intra-provider)	Candidate solutions
-S	Service management	●	●	●	TOSCA, YANG
-F	VNF lifecycle management		●	●	TOSCA
-C	Catalogs	●	●	●	Network service descriptors, TMForum
-RT	Resource topology	●	●	●	BGP-LS
-RC	Resource control		●	●	NETCONF, PCEP
-Mon	Monitoring	●	●	●	Lattice, time series data

As a reference of the different roles in the exchange, note that the provider B in Figure 10-9 (i.e., the one in the middle) participates on the E2E service only for providing UP connectivity between providers A and C.

10.4.2 Multi-Technology Scenarios

Nowadays, the automatic establishment of E2E connections is complex in a network composed of heterogeneous technological domains (that is, domains constituted by specific technologies like IP, optics, microwave, etc.). The complete process not only requires long time and high operational costs for configuration (including manual interventions), but also the adaptation to each particular technology implementation. The capability to operate and manage the network automatically and E2E is the main requirement for multi-technology scenarios. This facilitates as well the multi-vendor interworking, which is another dimension of the multi-technology issue, as already described in Section 10.2.1 in the context of the relevance of SDN. The target is to move towards a service-driven configuration management scheme that facilitates and improves the completion of configuration tasks by using global configuration procedures.

Typically, the transport network is referred to as a wide area network (WAN) in the ETSI NFV model, regardless of the complexity and diversity of the underlying infrastructure. The idea of the ETSI model is that the service orchestrator can easily interact with control capabilities that could permit the configuration and manipulations of the WAN resources to create E2E services without considering the transport domains' heterogeneity. However, this is yet far from existing capabilities and solutions.

Network operators have built their production networks based on multi-layer architectures. However, the different technologies in current transport networks are rarely jointly operated and optimized, i.e., the implications of a planning and configuration decision for different layers at the same time are typically not considered. Instead, they are usually conceived as isolated silos from a deployment and operation point of view.

This can be even more burdening across multiple domains as described before. A service deployed across domains will require actions in different networks using different technologies, inherently multiplying the intricate complexity of the E2E network provision and configuration.

A logically centralized orchestration element can have a complete and comprehensive network view independently of the technologies employed in each technological domain, and propose optimal solutions to improve the overall resource utilization. Such orchestrator, by maintaining a multi-layer view of the controlled network, can determine which resources are available to serve any connectivity request in an optimal manner, considering not only partial information (per technology domain), but the entire network resources, in a comprehensive manner. Aspects like global utilization, protection, congestion avoidance or energy saving can be optimized with such an approach. For getting the information per technology and building the multi-layer view (i.e., underlying topology, per-layer capabilities, border ports, etc.), the orchestrator could rely on lower-level controllers, for instance one per layer. In [22], an overview of the benefits obtained through a multi-layer approach is provided.

Network programmability, as enabled by SDN and already touched with relation to the radio access network (RAN) in Section 6.8, permits new ways of resource optimization by implementing sophisticated traffic engineering algorithms that go beyond the capabilities of contemporary distributed shortest path routing. Multi-layer coordination can help to rationalize the usage of technologically diverse resources. This new way of planning and operating networks requires a comprehensive view of the network resources and planning tools capable for handling this multi-layer problem.

10.5 Software-Defined Networking for 5G

5G will impose the need of a flexible network to support the diverse requirements of the distinct services and customers (i.e., verticals) on top of the providers' networks. This section introduces two particular scenarios for fronthaul, backhaul and core transport networks as examples of network segments out of the RAN also impacted by the advent of 5G. Note that SDN approaches for the RAN are covered in detail in Section 6.8.

10.5.1 Xhaul Software-Defined Networking

10.5.1.1 Introduction

The integration of fronthaul and backhaul technologies, also known as *Xhaul* and covered in more detail in Section 7.6.1, will enable the use of heterogeneous transport and technological platforms, leveraging novel and traditional technologies to increase the capacity or coverage of the future 5G networks.

The design of the Xhaul segment is driven by the detailed extracted requirements obtained from practical use cases with a clear economical target. A large number of use cases are proposed in literature, as covered in Chapter 2.

From the SDN perspective, the diversity and heterogeneity of the relevant technologies involved in the Xhaul segment means that using a single controller may not be applicable. This might be due to the need for controlling heterogeneous emerging technologies such as millimeter-wave (mmWave), while controlling a photonic mesh network. Thus, a hierarchical approach is typically proposed in order to tackle with this technological heterogeneity [23][24].

10.5.1.2 Possible Hierarchical SDN Controller Approaches for Xhaul

A possible solution to manage and control such diversity of heterogeneous technologies is to focus on a deployment model in which a SDN controller is deployed for a given technology domain (considering it as a child controller), while the whole system is orchestrated by a parent controller, relying on the main concept of network abstraction [25].

The proposed SDN architecture by ONF foresees the introduction of different levels of hierarchy, allowing for network resource abstraction and control. A level is understood as a stratum of hierarchical SDN abstraction. In the past, the need of hierarchical SDN orchestration has been justified by two purposes: a) Scaling and modularity: each successively higher level has the potential for greater abstraction and broader scope (e.g., RAN, transport, and data center network abstraction); and b) Security: each level may exist in a different trust domain, where the level interface might be used as a standard reference point for inter-domain security enforcement. The benefits of hierarchical SDN orchestration become clear in the scope of the described Xhaul with technology heterogeneity.

The Applications-Based Network Operations (ABNO) framework has been standardized by the IETF, based on standard protocols and components to efficiently provide a solution to the network orchestration of different CP technologies. An ABNO-based network orchestrator has been validated for E2E multi-layer and multi-domain provisioning across heterogeneous control domains employing dynamic domain abstraction based on virtual node aggregation [26].

Figure 10-10 shows the proposed hierarchical architecture for a future Xhaul network. It takes into account the different network segments and network technologies which are expected to be present.

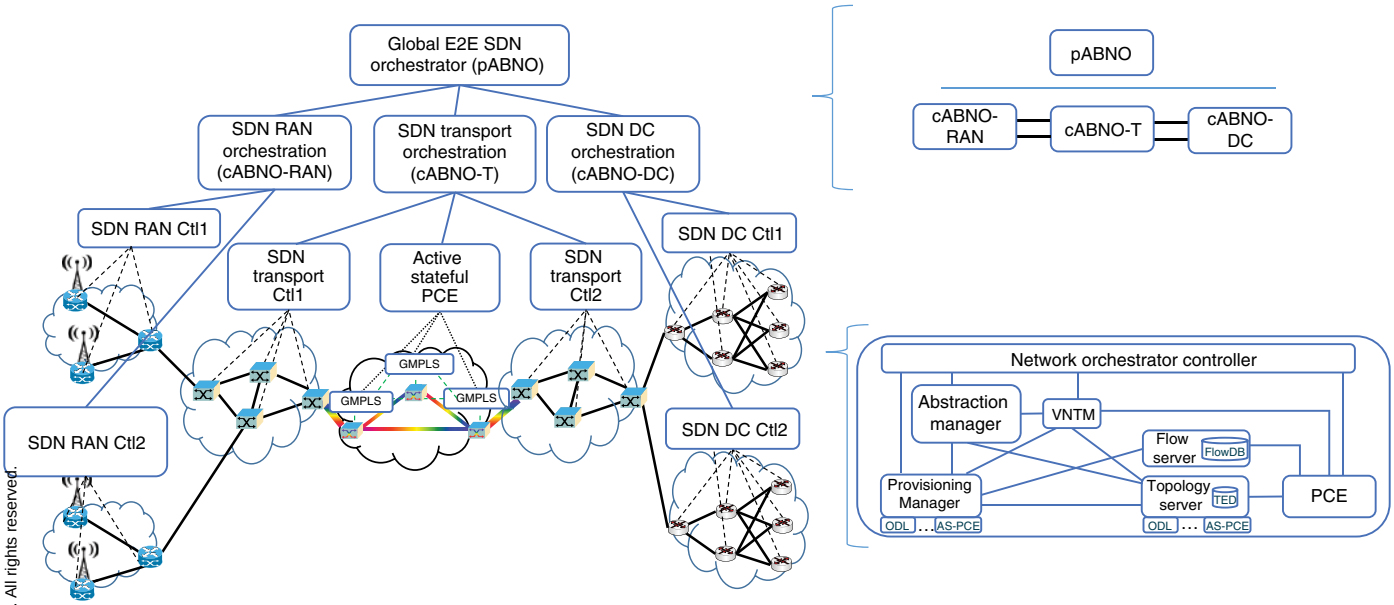

Figure 10-10. Proposed hierarchical ABNO architecture including hierarchical levels topological view and detail of hABNO architecture.

In the RAN segment, we observe several SDN-enabled controllers for wireless networks, which tackle their complexities. In a transport network, the aggregation segments and core network are taken into account. SDN-enabled Multiprotocol Label Switching - Transport Profile (MPLS-TP) can be used in the aggregation network, while a core network might use an optical SDN controller, such as an active stateful path computation element (AS-PCE) on top of an optical network. Finally, several SDN-enabled controllers are responsible for intra-data center networks, which typically run at layer-2.

Within the hierarchy, an SDN orchestrator may consider itself as the direct control entity of an information model instance that represents a suitably abstracted underlying network. It follows that, with the exception of network domain SDN controllers (which are directly related to NEs), a given SDN orchestrator might provide an abstracted network view and be present at any hierarchy level and act as parent or child SDN orchestrator. At any level of the recursive hierarchy, a resource is understood to be subject to only one controlling entity.

In the proposed architecture, several child ABNOs (cABNO) are proposed. Each cABNO is responsible for a single network segment. A recursive hierarchy could be based on technological, SDN controller type, geographical/administrative domains or network segment basis (each corresponding to a certain hierarchical level). Further, parent ABNO (pABNO) are introduced, responsible for the provisioning of E2E connections through different network segments.

For both the pABNO and the cABNO, the internal system architecture is similar, based on a set of components that are displayed in Figure 10-10 and detailed in [26]. The network orchestration controller is the component responsible for handling the workflow of all the processes involved (e.g., the provisioning of E2E connectivity services). It also exposes a NBI to offer its services to applications. For the cABNO, the NBI of the network orchestrator controller is extended to offer a REST-based interface for topology recovery and connection provisioning based on the Control Orchestration Protocol [27], which has evolved in ONF T-API and IETF TE models.

Figure 10-10 also provides the different topological views at different hierarchical levels (top hierarchical level for the pABNO, and lower hierarchical level for the different segments). The provided topological views correspond with the proposed experimental validation, where a pABNO and cABNO-T and cABNO-DC are deployed. The cABNO-T is responsible for SDN orchestration of two SDN aggregation domains and an SDN core network domain. The cABNO-DC is responsible for two intra-DC network domains.

The hierarchical SDN approach benefits single operator scenarios, where multi-layer, multi-vendor, and multi-technology SDN controllers are needed. For multi-operator scenarios, where centralized elements may be impractical, a peering model as presented in Section 10.4.1 may be the preferred option [21].

10.5.1.3 Integration with NFV Architecture

The wide adoption of NFV requires virtual computing and storage resources deployed throughout the network. Traditionally, virtual computing and storage resources have been deployed in large data centers (DCs) in the core network. Core DCs offer high computational capacity with moderate response time, meeting the requirements of centralized services with low-delay demands. However, it is also required to offer edge computing (i.e., micro-DCs and small-DCs) in different sites of the mobile network (e.g., at base stations, cell aggregation points, radio network controllers, or central offices) leveraging on low latency and high bandwidth. For example, ETSI is defining the multi-access edge computing (MEC), see Section 5.2.5, to offer applications such as video analytics, location services, mission-critical applications, augmented reality, optimized local content distribution, and data caching.

Typically, a single NFVI domain for the mobile Xhaul network is considered. The NFVI is distributed and interconnected by the Xhaul network. The VIM is commonly implemented using a cloud controller based on, e.g., OpenStack. It interfaces with the NFV reference implementations (i.e., OPNFV and OSM) using the OpenStack API. OpenStack enables to segregate the resources into availability zones for different tenants and to instantiate the creation, migration or deletion of virtual machines (VMs) and containers (CTs), related to compute services, storage of disk images (image services), and the management of the VM/CT's network interfaces and network connectivity (networking services). For example, the OpenStack compute service (named Nova) manages pools of compute nodes with many choices available for hypervisor technology (e.g., KVM, VMWare, Xen) or container technology. The OpenStack networking service (named Neutron) manages networks and IP addresses, providing flat networks or VLANs to separate traffic between hosts. Further, the Neutron service enables to configure a virtual switch such as an Open vSwitch (OVS) within a compute node (e.g., creation of new ports connecting new VMs/CTs, configuration of forwarding rules) through an SDN controller. It would allow to have a single VIM acting as global orchestrator of compute, storage and network resources. However, the current definition of the Neutron plugin does not support all the specific functionalities that would be required to control transport switches (packet or optical) external to the data center. To overcome this limitation, the ETSI NFV MANO framework has also defined the WAN infrastructure manager (WIM), as a particular VIM. In this scenario, the VIM (i.e., OpenStack cloud controller) is responsible for controlling and managing the NFVI-PoP's resources (i.e., DCs resources), whilst the WIM is used to establish connectivity between NFVI-PoP's. The WIM can be performed by a single SDN controller (e.g., OpenDaylight, ONOS, Ryu), or by an SDN orchestrator in a multi-layer (wireless, packet, optical) network with dedicated SDN controllers per technology, as explained in the previous section and further described in [28].

Additionally, each DC can be managed independently through its own cloud controller acting as a VIM. Moreover, a single cloud controller directly controlling thousands of compute nodes spread in multiple DCs does not scale. Thus, it is required to deploy a cloud orchestrator enabling to deploy federated cloud services for multiple tenants across distributed DC infrastructures. The considered cloud orchestrator may act as a parent VIM and interface with the NFVO, within a hierarchical VIM architecture. However, the cloud orchestrator should support the OpenStack API, since it has become the de-facto interface between the VIM and the reference NFVO implementations. There are two OpenStack projects aiming at developing a hierarchical OpenStack architecture. These would enable to develop a cloud orchestrator based on OpenStack (e.g., Trio2o and Tricircle) and use the OpenStack API as both the southbound interface (SBI) with the OpenStack controllers as well as the northbound interface (NBI) with the NFVO implementations. Alternatively, the NFVO should perform the orchestration of the NFV infrastructure resources (i.e., DC resources) across the multiple VIMs by directly interfacing with the multiple VIMs, instead of the cloud orchestrator.

10.5.1.4 Supporting Network Slicing over the Xhaul Infrastructure

Network slicing has emerged as a key requirement for 5G networks, although the concept itself is still not (yet) fully developed. Macroscopically and from a high-level perspective, the word slicing is understood to involve the partitioning of a single, shared infrastructure into multiple logical networks (*slices*), along with the capability of instantiating them on demand, in order to support functions that constitute operational and user-oriented services. In this setting, important characteristics of slicing are that it not only involves network resources but also computing and storage, and that

such slices are expected to be customized and optimized for a service (set) or vertical industry making use of such slice [29]. Network slicing is covered in detail in Chapter 8.

In this section, we focus on the specifics related to network management and SDN/NFV control aspects of network management. Research, development and standardization work is consequently needed, not only to define information and data models for a network slice, but also mechanisms to dynamically manage such constructs, providing multiple, highly flexible, and dedicated E2E slices (considering virtual network, cloud and function resources), while enabling different models of control, management and orchestration systems, covering all stages of slice life-cycle management. This includes the ability to deploy slices on top of the underlying infrastructure, including, where appropriate, the ability to partition network elements. The existing mechanisms to carry out this resource partitioning are multiple, and there is no formal or standard mechanism to do so.

As mentioned in Chapter 8, and from the point of view of business models, network slicing allows, e.g., mobile network operators (MNOs) to open their physical transport network infrastructure to the concurrent deployment of multiple logical and self-contained slices. In this line, slices can be created and operated by the 5G network operator or enable new business models, such as "Slice-as-a-Service" (SlaaS). As a basic, canonical example, the ETSI NFV framework, conceived around the idea and deployment model where dedicated network appliances (such as routers and firewalls) are replaced with software (i.e., guests) running on hosts, can be the basis for a slicing framework, at least for a well-scoped definition of slices. From a functional architecture perspective, the ETSI NFV framework needs to be extended to support slicing natively, by means of, e.g., a slice manager (Xhaul slice control and orchestration system) or entity that performs the book-keeping of slices and maps them to slice owners and associates them to dedicated, per-slice control and management planes.

Part of the function of such control and orchestration system is thus to ensure access rights, assign resource quotas and provide efficient means for the resource partitioning and isolation. Those functions are nonetheless assumed to be part of the network slicing lifecycle management. Support of multi-tenancy has a strong impact on the SDN and MANO functions and components. For example, at the SDN controller level, multi-tenancy requirements are related to the delivery of uniform, abstract and UP-independent views of its own logical elements, while hiding the visibility of other coexisting virtual networks, including the logical partitioning of physical resources to allocate logical and isolated network elements and the configuration of traffic forwarding compliant with per-tenant traffic separation, isolation and differentiation. At the VIM and VNF MANO level, similar considerations on virtual resource allocation and isolation are extended to computing elements and a suitable modeling of the tenant and its capabilities [30].

Related to the Slice-as-a-Service, it is commonly accepted that the tenants may need to have certain control of their sliced virtual infrastructure and resources. It is part of the actual service control model to define the degree of control over the slice [30].

In a first model, the control that each tenant (i.e., owner or operator of the allocated network slice) exerts over the allocated infrastructure is limited, scoped to a set of defined operations. For example, the tenant can retrieve, e.g., a limited or aggregated view of the virtual infrastructure topology and resource state and perform some operations, using a limited set of interfaces, allowing a limited form of control, and different from controlling or operating a physical infrastructure. For example, the actual configuration and monitoring of individual flows at the nodes may not be allowed, and only high-level operations and definitions of policies may be possible.

Alternatively, each allocated slice can be operated as a physical one, that is, each tenant is free to deploy its choice of the infrastructure operating system and control. A virtual network operator

(VNO) is able to manage and optimize the resource usage of its own virtual resources. This means that each tenant can manage its own virtual resources, implemented by deploying a per-tenant controller or per-tenant management. This approach results in a control hierarchy and recursive models, requiring adapted protocols that can be reused across the controllers' NBIs and SBIs.

10.5.2 Core Transport Networks

The evolution towards fully operational 5G networks imposes a number of challenges that are usually perceived as impacting only the access networks, although this is not actually the case. Network functions, as integral parts of the services offered to the end-users, have to be composed in a flexible manner to satisfy variable and stringent demands, including not only dynamic instantiation but also deployment and activation. In addition to that, and as a complement of it, the whole network should be programmable to accomplish such expected flexibility, allowing for interconnecting the network functions across several NFVI-PoPs and scaling the connections according to the traffic demand. The versatile consumption of resources and the distinct nature of the functions running on them can produce very variable traffic patterns on the networks, changing both the overlay service topology and the corresponding traffic demand. The location of the services is not tightly bound to a small number of nodes any more, but to distributed resources that are topologically and temporally changing. The network utilization then becomes time-varying and less predictable. In order to adapt the network to the emergence of 5G services, the provision of capacity on demand through automatic elastic connectivity services in a scalable and cost-efficient way is required. The backbone or core transport networks then become a key component for E2E 5G systems.

The transformation objectives of the core transport networks have been traditionally focused towards more affordable and cost-effective technologies, being able to cope with the huge increase in traffic experienced in the latest years, at a reduced cost per bit. 5G networks, however, present innovative requirements to be faced by the transport networks, like the need to accommodate a large number of simultaneous devices, provide transport and service resources in a flexible and dynamic manner, and reduce the provisioning time to make such flexibility functional. Specifically, 5G transport networks will have to support high traffic volumes and ultra-low latency services. The variety in service requirements and the necessity to create network slices on demand will also require an unprecedented flexibility in the transport networks, which will need to dynamically create connections between sites, network functions or even users, providing resource sharing and isolation. Key aspects on the concept of network slicing are [31]: (i) resource manageability and control, (ii) virtualization through abstraction of the underlying resources, (iii) orchestration of disparate systems, and (iv) isolation of the offered compound assets in the form of slice. 5G transport shares all those goals. Moreover, the flexibility required by 5G transport, such as the dynamic creation and reconfiguration of network slices, makes some of the requirements even more stringent.

The programmability of the transport networks will be performed through open, extensible APIs and standard interfaces that permit agile E2E service creation in a rapid and reliable way. The goal is to evolve towards E2E automated, dynamically reconfigurable and vendor-agnostic solutions based on service and device abstractions, with standard APIs able to interoperate with each other, and facilitating a smooth integration with the OSS and BSS deployed by network operators.

From a complementary angle, transport networks will also have a very relevant role in the optimization of RAN resources by enabling flexible fronthaul and backhaul systems, maximizing

the benefits provided by distributed and virtualized RAN environments, tailored to the needs of a variety of vertical customers. The support of different functional splits in the radio part, the packetized transport of such signals, and the dynamic location of the processing units will render a full programmability and dynamicity in the transport part necessary.

Network management and orchestration mechanisms at transport level are needed in order to create the programmable environment required for 5G networks. The purpose is to integrate this programmable transport infrastructure with the overall 5G orchestration system, creating, managing and operating slices for different customers.

10.6 Network Function Virtualization in 5G Environments

Virtualization is the technique which significantly reshaped the IT and the networking ecosystem in recent years. On one hand, cloud computing and related services such as Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) are the results of a successful story (and ongoing stories) from the IT field. On the other hand, networking is in the middle of a momentous revolution and important transition. The appearance of virtualization techniques for networks fundamentally redefines how telecommunications enterprises will soon operate. In the visions of 5G, the often-heard service-level keywords are cost-effectiveness and improved service offering with fast creation, fast reconfiguration and a larger geographical reach of customers. This paradigm shift is technologically triggered by NFV, i.e., the implementation of telco functions on virtual machines that can be run on general purpose computers instead of running them on expensive dedicated hardware as in the traditional way; and also by SDN, i.e., configuring network appliances with easily manageable, often centrally run controller software applications. Combined with the already mature cloud technologies, 5G services can be best implemented in service function chains (SFCs) in which basic functions are run separately, possibly in remote data centers, while network control ensures the connectivity among those, and of course among the end users, by steering traffic based on, e.g., network service headers (NSHs).

In order to enable carrier-grade network services and dynamic SFCs with strict QoS requirements, a novel UP is needed that supports high performance operations (comparable to traditional hardware-based solutions), controllable bandwidth, and delay characteristics between physical or logical ports and interfaces. Therefore, the flexible and fine-granular programming of the general purpose forwarding and processing elements is crucial. SDN is the key enabler of CP softwarization and targets a programmable UP split from the control part. Besides the activities addressing carrier-grade SDN CP platforms, such as OpenDaylight or Open Network Operating System (ONOS), significant efforts have been focused on UP solutions. For example, Intel's Data Plane Development Kit (DPDK) is a software toolkit which enables enhanced packet processing performance and high throughput on general purpose commodity servers. It is supported by the de-facto standard of software switches, i.e., Open vSwitch (OVS).

Many tools are already available for network service providers and network operators. There are open-source solutions for the orchestration of IT resources, e.g., OpenStack as a fully-fledged cloud operating system, and the building blocks, e.g., OVS and DPDK, to make the underlying networking UP programmable and efficient. However, as virtual machines (VMs) and containers (CTs) use the same hardware resources (CPU, memory) as the components responsible for networking, a low-level

resource orchestrator is also needed (besides resource orchestrators running at higher abstraction and aggregation levels), which is capable of jointly handling the requests, and of calculating, configuring and enforcing appropriate resource allocation.

In this envisioned SFC-based 5G ecosystem, multiple novel types of actors appear, as also discussed in Section 2.6: infrastructure providers that offer compute and/or network resources for service deployment, application developers who sell the code and/or the management service of VNFs from which the SFC can be built, and the customers that are, at the end of the day, the application providers to end users. The first type of actors are mostly the traditional Telcos and Internet service providers (ISPs), while the second and third types are often merged today in the form of over-the-top (OTT) solution providers.

Future 5G services, such as coordinated remote driving, remote surgery or other Tactile Internet related applications with round-trip latency requirements on the order of few ms, pose extreme requirements on the network, and call for the joint control of IT and network resources. Moreover, typical network services, realized by SFCs, span not only over multiple domains, but over multiple operators as well, as cost-effectiveness by resource sharing is envisioned, and a wide geographical reach of customers in the 5G ecosystem. As one of the most important use cases, the Factory of the Future will make an intensive usage of 5G technologies for supporting the digitization in the way conceived by the idea of Industry 4.0. A high number of connected devices, collaborative robots, augmented reality, and the integration of manufacturing, supply chain and logistics, altogether open an opportunity window to operators for monetizing the provision of virtualized infrastructures and capabilities.

The multi-provider orchestration and management of network services involves many aspects, from the resource discovery and business negotiations between operators, to the computation and monitoring of assured quality network connections among their domains, and the efficient embedding of services into the available resource set. Novel features and technical enablers are necessary for NFVO in a flexible multi-provider setup. A multi-provider NFVO handles abstract sets of compute and network resources and provisions the necessary subset to the customer in order to deploy its service within. In addition to that, it provides an integrated view of infrastructure resources to the customer, also encapsulating managed VNF capability, and ensures that the demanded service requirements are fulfilled.

With well-defined interfaces and orchestration-management mechanisms, operators can act not only as NFVI providers, but also as integrators of VNF-as-a-Service (VNFaaS) offerings from third parties. As such, operators can also act as virtualization platform providers that open interfaces for third party components, such as VNF managers (VNFMs).

10.7 Autonomic Network Management in 5G

10.7.1 Motivation

To meet the diverse and stringent KPI requirements specified in ITU-R IMT-2020, the 5G system will necessarily become more complex [32], which can be mainly characterized by the following technical features: 1) a heterogeneous network consisting of marco cells, small cells, relays, and device-to-device (D2D) links; 2) new spectrum paradigms, e.g., dynamic spectrum access, licensed-assisted access, and higher frequency at mmWave bands, as elaborated in Chapter 3; 3) cutting-edge air-interface

technologies, such as massive antenna arrays and advanced multi-carrier transmission, as detailed in Chapter 11; and 4) a novel E2E architecture for flexible and quick service provision in a cost- and energy-efficient manner, as introduced in Chapter 5.

The system's complexity imposes a high pressure on today's manual and semi-automatic network management that is already costly, vulnerable, and time-consuming. However, mobile networks' troubleshooting (related to systems failures, cyber-attacks, and performance degradations, etc.) still cannot avoid manually reconfiguring software, repairing hardware or installing new equipment. A mobile operator has to keep an operational group with a large number of network administrators, leading to a high operational expenditure (OPEX) that is currently three times that of capital expenditure (CAPEX) and keeps rising [33]. Additionally, troubleshooting cannot be performed without an interruption of the network operation, which deteriorates the end user's quality-of-experience (QoE) [34]. Without the introduction of new management paradigms, such large-scale and heterogeneous 5G networks simply become unmanageable and cannot maintain service availability.

Recently, the research community has started to explore artificial intelligence [35] in order to minimize human's intervention in managing networks to lower the OPEX and improve the system's performance. IETF has initiated a research group called Intelligence-Defined Networks to specifically study the application of machine learning technologies in networking. Moreover, the 5G PPP projects SELFNET [36] and CogNet [37] have focused on designing and implementing intelligent management for 5G mobile networks. For example, the SELFNET project has been set up to design, prototypically implement, and evaluate an autonomic management framework for 5G mobile networks. Taking advantage of new technologies, in particular SDN [38], NFV [39], self-organized networks (SON) [40], multi-access edge computing (MEC) and artificial intelligence, the framework proposed by the SELFNET project can provide the capabilities of self-healing against network failures, self-protection against distributed cyber-attacks, and self-optimization to improve network performance and end users' QoE [41]. Although the current SON techniques have a self-managing function, it is limited to static network resources. It does not fully suit 5G scenarios, such as network slicing [42] and multi-tenancy [43], where dynamic resource utilization and agile service provision are enabled by SDN and NFV technologies. Currently, existing SON can only reactively respond to detected network events, while the intelligent framework is capable of proactively performing preventive actions for predicted problems. The automatic processing in SON is usually limited to simple approaches like triggering, and some operations are still carried out manually. In addition, the self-x management mainly focuses on the RAN. An extension beyond the RAN segment to provide a self-organizing function over the E2E network is required. By reactively and more importantly proactively detecting and diagnosing differently network problems, which are currently manually addressed by network administrators, the SELFNET framework could assist network operators to simplify management and maintenance tasks, which in turn can significantly lower OPEX, improve user experience, and shorten time-to-market of new services.

In this section, a reference architecture for the autonomic management framework [36] will be introduced, including the functional blocks, their capabilities and interactions; the autonomic control loop starting from the SDN/NFV sensor and terminating at the actuators will be provided, as well as a brief exemplary loop so as to illustrate how the autonomic system may mitigate a network problem. Furthermore, several classical artificial intelligence algorithms that can be applied to implement the network intelligence are briefly shown.

10.7.2 Architecture of Autonomic Management

In addition to the software-defined and virtualized network infrastructure [44], the autonomic management framework mainly consists of: 1) SDN/NFV sensors that can collect the network metrics; 2) monitoring modules that can derive the symptoms from the collected metrics; 3) network intelligence that is in charge of diagnosing network problems and making tactical decisions; and 4) SDN/NFV actuators and an orchestrator that perform corrective and preventive actions. As shown in Figure 10-11, the potential architecture for autonomic management can be split into several layers, which are explained as follows:

- **Infrastructure layer**: All NFs managed autonomously by the framework rely on physical and virtualized resources in this layer. It encompasses physical and virtualization sublayers. The former provides an access to physical resources (networking, computing, storage, etc.), while the latter instantiates virtual infrastructures on top of the physical sublayer. It represents the NFVI as defined by the ETSI NFV terminology;
- **Data network layer**: This implies an architectural evolution towards the SDN paradigm by decoupling the CP from the UP. In this framework, the data layer represents a simple data-forwarding, which can be either a non-virtualized or virtualized NF;

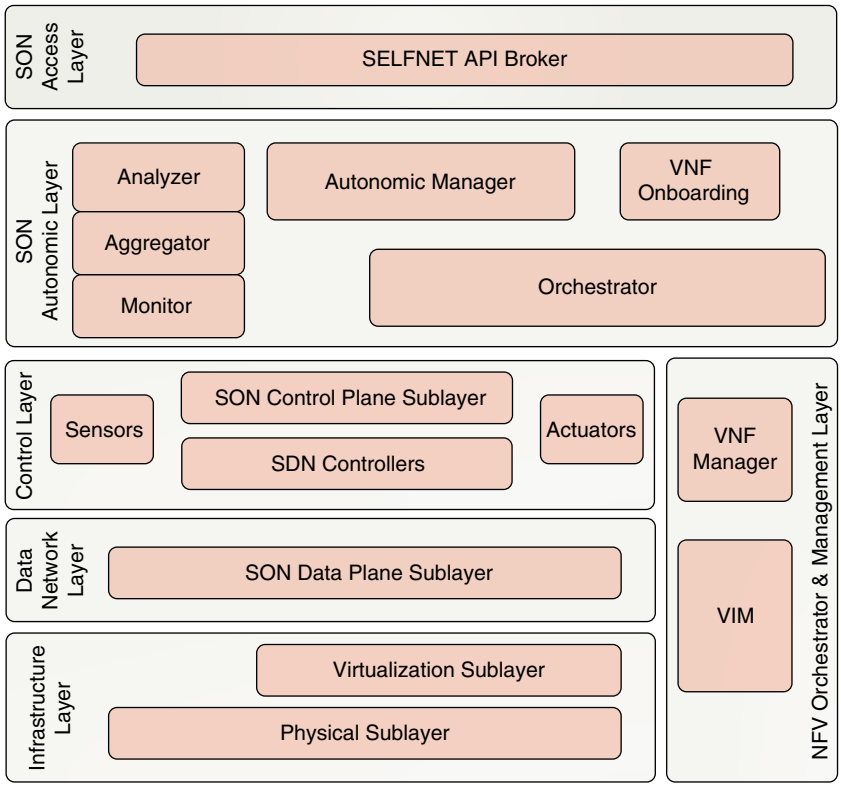

Figure 10-11. Possible architecture for autonomic management [36].

- **SON control layer**: This layer includes two internal sublayers: SDN controllers and SON CP sublayer. SDN/NFV sensors and actuators, which are capable of collecting data from the entire system and enforcing actions, respectively, are also contained. The SON control layer and data network layer have associated CPs and UPs of the network that are decoupled in the SDN paradigm;
- **SON autonomic layer**: To realize the network intelligence, this layer consists of three modules, i.e., monitor, aggregator and analyzer, autonomic manager, and orchestrator. The monitor and analyzer extract metrics related to network behavior, aggregate the collected metrics into health of network (HoN) metrics, and use these to infer the network status. The autonomic manager is in charge of diagnosing the root cause of any existing or potential network problems, and deciding which countermeasure should be conducted. Following the tactical decisions from the autonomic manager, the orchestrator coordinates the physical and virtualized resources, and manages the SDN/NFV actuators to execute the decided actions;
- **NFV orchestration and management layer**: This layer is responsible for orchestrating and managing VNFs via the VNF manager, as well as virtualized resources through VIM. It conforms to the NFV MANO specified by ETSI [5];
- **SON access layer**: This is the external interface that is exposed by the framework. Despite the fact that internal components may have specific interfaces for the particular scope of their functions, these components contribute to a general SON API, managed by the SELFNET API broker that exposes all aspects of the autonomic framework to external systems, such as BSS or OSS and administration graphical user interfaces (GUIs). The latter enable network administrators to interact with and configure the SELFNET framework and also observe the complete status of the network.

10.7.3 Autonomic Control Loop

One of the main challenging aspects of the autonomic management is the implementation of network intelligence. Apart from the underlying software-defined and virtualized network infrastructure, a closed control loop referred to as autonomic control loop, starting from the sensors and terminating at the actuators, is needed to control the processing flow. When the monitor detects or predicts a network problem, an autonomic control loop is initiated. The autonomic manager diagnoses the cause of the problem, decides on a tactic, and plans an action. Once the orchestrator receives an action request from the action enforcer (AE), it coordinates the physical and virtualized resources to enforce this action.

The autonomic manager can be regarded as the brain of the autonomic management framework and plays a vital role in the provision of network intelligence. Taking advantage of cutting-edge techniques in the field of artificial intelligence, it provides the capabilities of self-healing, self-protection and self-optimization by means of reactively and proactively dealing with detected and predicted network problems. As illustrated in Figure 10-12, the autonomic manager consists of the following functional blocks:

- A **diagnoser** is in charge of diagnosing the root cause of network problems. The monitor can derive a symptom for each detected or predicted network problem from the collected sensor data. The diagnoser processes the reported symptom to make clear its reason, and notifies the decision-maker;

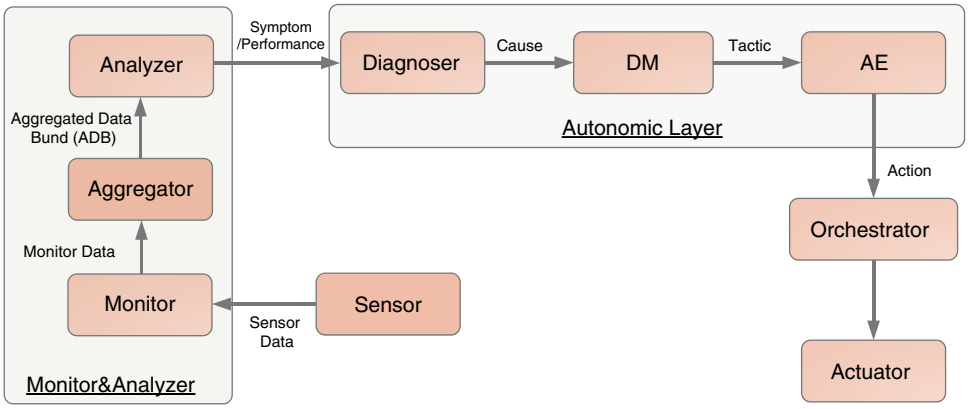

Figure 10-12. Autonomic management control loop.

- A **decision-maker (DM)** can decide a set of corrective or preventive tactics to deal with the network problems based on incoming diagnostic information. A tactic is a high-level description of a countermeasure, which needs to be transferred into an implementable action;
- An **action enforcer (AE)** is responsible for providing a consistent and coherent set of scheduled actions to be enforced in the network infrastructure. For this purpose, this module recognizes and validates these tactics by applying conflict detection and resolution in order to provide implementable actions to be enforced.

Within this control loop, the metrics collected by the sensors are processed by the monitor module first. Subsequent modules extract the required information from the previous module and provide the next-level results to the next module. The information model associated with the autonomic control loop is explained as follows:

- **Sensor data**: A range of differentiated data sources can be expected to be identified in the upcoming 5G infrastructure. All monitoring information retrieved from physical devices, UP, SDN controller, SDN/NFV sensors, VIM etc., are uniformly referred to as sensor data. The monitor is the corresponding module that is in charge of collecting sensor data from underlying infrastructures;
- **Monitor data**: The monitor regularly collects the sensor data and reports the necessary information to the aggregator. Some of the data is periodically collected, which stands for either normal or abnormal network behaviors;
- **Aggregated data bundle (ADB)**: The monitor data related to a network problem may be retrieved from a set of sensors, rather than a single one. For example, in the case of a distributed denial of service (DDoS) attack, the source and destinations are distributed. The raw information contained in monitor data should be processed to produce aggregated and correlated information, which is called *aggregated data bundle*;
- **Symptom**: A high-level health-of-network metric that may be derived from a set of correlated alarms, events, KPIs, etc., that can be evaluated to indicate the characteristics of an existing or emerging network problem, together with the additional contextual information such as metadata, is defined as a symptom;

- **Performance**: The report of achieved performance by an executed action is two-fold: i) if an action degrades performance rather than solving a problem, a roll-back mechanism will be triggered to recover the network status to the initial point before the action was performed; ii) the achieved performance, which can be regarded as the benefit or reward of action. If a large extent of operational data can be recorded, the network intelligence can be trained based on machine learning techniques;
- **Cause**: It is a description of what the reason of a network problem is or why a network problem happens or will happen. Once the diagnoser receives a symptom, it diagnoses the cause of this symptom;
- **Tactic**: After the cause of a network problem is clarified, a countermeasure that can be applied to tackle this problem needs to be decided by the decision-maker. A tactic is a high-level description of a countermeasure, which is required to be transferred into an implementable action;
- **Action**: This is an implementable version of a countermeasure with a description of how to enforce this, taking into account available physical and virtualized resources. The action provided by the AE contains more implementation details, e.g., the actuator's type, the target deployment location, and configuration information.

To close this section, let us use the following example to show the autonomic control loop and illustrate its main mechanisms. The storyline is depicted as follows: A summer concert is taking place in the city centre, where a large number of spectators gather in a small area. Some of the spectators start to share real-time videos in their social media. When the number of video users increases, especially if some of them transfer videos in ultra-high definition, the network suffers from traffic congestion and the perceived QoE deteriorates. The monitoring modules first detect this network's anomaly by means of collecting, aggregating, and analysing the sensor data. A symptom called video QoE decreasing is reported to the diagnoser. After the diagnosis, it is found that the cause of the video QoE decreasing is the increased number of video users in the zone. Then, the possible tactics, for instance, load-balancing, video coding optimizing, and admission control, are determined by the decision-maker. The AE transfers these tactics into implementable actions and notifies the orchestrator. Taking into account available resources, the action of load balancing is finally selected and executed by the orchestrator. An actuator acting as a load balancer is instantiated, configured and deployed in the local network surrounding this concert. Afterwards, the congested network is successfully recovered and the perceived QoE of end users is improved.

10.7.4 Enabling Algorithms

We will give a brief introduction about enabling intelligence algorithms. The motivation is to provide a view for the readers how to apply artificial intelligence to implement the network intelligence. Hence, only several classical algorithms are given. For further artificial intelligence technologies, such as neural networks [45], reinforcement learning [46], transfer learning [47], and deep learning [48], the reader is referred to the stated references.

10.7.4.1 Feature Selection

In practice, a large number of features (i.e., network metrics) can be extracted from the 5G infrastructure. Each feature generally needs to be periodically recorded, resulting in a huge volume of data. When the management system tackles a specific problem, e.g., traffic congestion, it is

inefficient (if not infeasible) to process all data. That is because generally only a relatively small subset of all-available features is informative, while others are either irrelevant or redundant. As a data-driven approach, the network intelligence should be built on relevant features, while discarding others, so that irrelevant and redundant features do not degrade the performance on both training speed and predictive accuracy.

Feature selection (FS) is hence one of the most important intelligence techniques and an indispensable component in machine learning and data mining. It can reduce the dimensionality of data by selecting only a subset of features to build the learning machine. A number of classical FS algorithms, such as Relief-F [49] and Fisher [50], can be directly applied to calculate the relevance of the collected features.

10.7.4.2 Classification

In the terminology of machine learning, classification is an instance of supervised learning. It is applied to identify which class a new observation belongs to on the basis of a training dataset. An example would be assigning an incoming email into SPAM or non-SPAM classes in terms of the observed features of the email (e.g., source IP address, text length, and title content). The following is a brief introduction of classification algorithms that can be used in the network intelligence:

- A **decision tree (DT)** [51] is a classical supervised learning method used for classifying. Decision rules are inferred from a training dataset and a tree-shaped diagram is built. Each node of the decision tree relies on a feature to separate the data, and each branch represents a possible decision. DT is simple, interpretable and fast, whereas it is hard to apply in a complex and non-linear case;
- A **discriminant analysis** is a classification method which assumes that different classes generate data based on different Gaussian distributions. A linear discriminant (LD) analysis [52] is to find a linear combination of features that maximize the ratio of inter-class variance to the intra-class variance in any particular dataset so as to guarantee maximal separation;
- A **support vector machine** (SVM) [53] utilizes a so-called hyperplane to separate all data points of one class from another. The number of features does not affect the computational complexity of SVM, so that it can perform well in the case of high-dimensional and continuous features. However, it is a binary classifier, and a multi-class problem can be solved only by transferring this into multiple binary problems;
- Another algorithm called **k Nearest Neighbor** (kNN) is applied for data classification following the hypothesis that close proximity in terms of inter-data distance has a similarity. The class of an unclassified observation can be decided by observing the classes of its nearest neighbors. It is among the simplest algorithms with a good predictive accuracy. But it needs high memory usage, is vulnerable to noisy data and is not easy to interpret.

10.8 Summary

This chapter has shown that the management and orchestration plane is instrumental to enable the efficient utilization of the infrastructure, while meeting the performance and functional requirements of heterogeneous services. The requirements for the forthcoming 5G networks trigger the work on a complex ecosystem where compute, storage and connectivity must be coordinated in real-time.

SDN decouples the control and user planes of the NEs to enable a central network control that can make smart decisions, while the NEs are focused on the forwarding and application of policies. Such separation enables the network to become more flexibly programmable than current networks. The programmability of SDN is required by the NFV paradigm. NFV facilitates the dynamic instantiation of VNFs on top of commodity hardware, which lets the operator separate the NFs from the hardware. Autonomics will evolve the networking technologies with the necessary support for handling its heterogeneous complexity and provide the necessary service availability and resiliency. These technologies are key enablers of the new management and orchestration technologies.

In the case of multi-operator orchestration scenarios, it is essential not only to define, but also to implement an E2E orchestration plane able to deal with the interaction of multiple administrative domains. The use of open and standard interfaces as well as the modeling of services and devices are the only way to have an ecosystem to facilitate the deployment of new paradigms in network operators. Similarly, it is the use case of multi-technology, where the scenario is a real network with legacy systems that are providing services to the end-customers.

References

1 NGMN, White Paper, "5G White Paper", Feb. 2015
2 L.M. Contreras, P. Doolan, H. Lønsethagen and D.R. López, "Operation, organization and business challenges for network operators in the context of SDN and NFV", in Elsevier Computer Networks, Vol. 92, pp. 211–217, 2015
3 ONF SDN Architecture, Issue 1, June 2014, see www.opennetworking.org
4 ETSI NFV, see http://www.etsi.org/technologies-clusters/technologies/nfv
5 ETSI GS NFV MAN 001, "Networks Functions Virtualization (NFV); Management and Orchestration", V1.1.1, Dec. 2014
6 ETSI, White Paper, "Open Source MANO", Release 2, April 2017
7 R. Enns, M. Bjorklund, J. Schoenwaelder and A. Bierman, "Network Configuration Protocol (NETCONF)", RFC 6241, IETF, June 2011
8 M. Bjorklund, "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, IETF, Oct. 2010
9 A. Bierman, M. Bjorklund and K. Watsen, "RESTCONF Protocol", RFC 8040, Jan. 2017
10 http://claise.be/IETFYANGPageCompilation.png
11 Draft ETSI GS NFV-SOL 001, "Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; NFV descriptors based on TOSCA specification", V0.4.0, Dec. 2017
12 OASIS, "TOSCA Simple Profile for Network Functions Virtualization (NFV)", v1.0, March 2016
13 ETSI GS NFV-EVE 005, "Network Functions Virtualisation (NFV); Ecosystem; Report on SDN Usage in NFV Architectural Framework", v1.1.1, Dec. 2015
14 L.M. Contreras, C.J. Bernardos, D.R. López, M. Boucadair and P. Iovanna, "Cooperating Layered Architecture for SDN", draft-irtf-sdnrg-layered-sdn-01 (work in progress), Oct. 2016
15 Q. Wu, W. Liu and A. Farrel, "Service Models Explained", May 2017
16 D. Bogdanovich, B. Claise and C. Moberg, "YANG Module Classification", May 2017
17 ONF Report TR-534, "Framework and Architecture for the Application of SDN to Carrier Networks", July 2016

18 ONF Report TR-521, "SDN Architecture – Issue 1.1", Jan. 2016
19 Draft ETSI GR NFV IFA 028, "Network Function Virtualisation (NFV); Management and Orchestration; Architecture options to support the offering of NFV MANO services across multiple administrative domains", V0.9.0, Dec. 2017
20 MEF, Service Operations Specification MEF 55, "Lifecycle Service Orchestration (LSF) Reference Architecture and Framework", March 2016
21 5G PPP 5G-Exchange project, see http://www.5gex.eu
22 V. Lopez, D. Konidis, D. Siracusa, C. Rozic, I. Tomkos and J.P. Fernandez-Palacios, "On the Benefits of Multilayer Optimization and Application Awareness", Journal of Lightwave Technology, vol. 35, no. 6, March 2017
23 5G PPP 5G-Xhaul project, see http://www.5g-xhaul-project.eu/
24 5G PPP 5G-Crosshaul project, see http://5g-crosshaul.eu/
25 M. Fiorani et al., "Abstraction models for optical 5G transport network", Journal of Optical Communications and Networking, vol. 8, no. 9, pp. 656–665, Sep. 2016
26 R. Vilalta et al., "Hierarchical SDN Orchestration for Multi-technology Multi-domain Networks with Hierarchical ABNO", European Conference on Optical Communication (ECOC 2015), Dec. 2015
27 A. Mayoral et al., "The Need of a Transport API in 5G for Global Orchestration of Cloud and Networks through a Virtualised Infrastructure Manager and Planner", JOCN Special Issue OFC 2016, 2016
28 R. Casellas, R. Muñoz, R. Vilalta and R. Martínez, "Orchestration of IT/Cloud and Networks: From Inter-DC Interconnection to SDN/NFV 5G Services", Optical Network Design and Modeling (ONDM 2016), May 2016
29 NGMN, White Paper, "Description of Network Slicing Concept", v1.0, Jan. 2016
30 Xi Li, et al., "5G-Crosshaul Network Slicing Enabling Multi-Tenancy in Mobile Transport Networks", IEEE Communications Magazine, vol. 55, no. 8, pp. 128–137, Aug. 2017
31 J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J.J. Ramos-Munoz, J. Lorca and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures and Challenges", IEEE Communications Magazine, vol. 55, no. 5, pp. 80–87, May 2017
32 J. G. Andrews et al., "What will 5G be?", IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1065–1082, June 2014
33 Aviat Networks, "Top ten pain points of operating networks", 2011
34 B. Bangerter et al., "Networks and devices for the 5G era", IEEE Communications Magazine, vol. 52, no. 2, pp. 90–96, Feb. 2014
35 A. He et al., "A survey of artificial intelligence for cognitive radios", IEEE Transactions on Vehicular Technology, vol. 59, no. 4, pp. 1578–1592, May 2010
36 5G PPP SELFNET project, see https://selfnet-5g.eu/
37 5G PPP CogNet project. see http://www.cognet.5g-ppp.eu/
38 B. A. A. Nunes et al., "A survey of software-defined networking: Past, present, and future of programmable networks", IEEE Communication Surveys, vol. 16, no. 3, pp. 1617–1634, 2014
39 R. Mijumbi et al., "Network function virtualization: State-of-the-art and research challenges", IEEE Communication Surveys, vol. 18, no. 1, pp. 236–262, 2016
40 S. Dixit et al., "On the design of self-organized cellular wireless networks", IEEE Communications Magazine, vol. 43, no. 7, pp. 86–93, Jul. 2005
41 J. P. Santos et al., "SELFNET framework self-healing capabilities for 5G mobile networks", Transactions on Emerging Telecommunications Technology, Wiley, vol. 27, no. 9, pp. 1225–1232, Sep. 2016

42 X. Zhou et al., "Network slicing as a service: enabling enterprises' own software-defined cellular networks", IEEE Communications Magazine, vol. 54, no. 7, pp. 146–153, July 2016
43 K. Samdanis et al., "From network sharing to multi-tenancy: The 5G network slice broker", IEEE Communications Magazine, vol. 54, no. 7, pp. 32–39, July 2016
44 P. Neves et al., "The SELFNET approach for autonomic management in an NFV/SDN networking paradigm", International Journal of Distributed Sensor Networks, vol. 16, no. 2, pp. 1–17, Feb. 2016
45 G. P. Zhang, "Neural networks for classification: a survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 30, no. 4, pp. 451–462, Nov. 2000
46 L. Buoniu, R. Babuka and B. D. Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 2, pp. 156–172, March 2008
47 M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey", Journal of Machine Learning Research, pp. 1633–1685, July 2009
48 Z. Fadlullah et al., "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems", IEEE Communications Surveys & Tutorials, no. 99, May 2017
49 I. Kononenko et al., "Estimating attributes: analysis and extensions of RELIEF", European Conference on Machine Learning, April 1994
50 Q. Gu, Z. Li and J. Han, "Generalized Fisher score for feature selection", Conference on Uncertainty in Artificial Intelligence, July 2011
51 S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey", Journal on Data Mining and Knowledge Discovery, vol. 2, no. 4, pp. 345–389, Dec. 1998
52 Y. Guo, T. Hastie and R. Tibshirani, "Regularized discriminant analysis and its application in microarrays", Biostatistics, vol. 1, no. 1, pp. 1–18, 2005
53 C. J. Burges, "A tutorial on support vector machines for pattern recognition", Journal on Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, Dec. 1998

Part 3

5G Functional Design

11

Antenna, PHY and MAC Design

Frank Schaich[1], Catherine Douillard[2], Charbel Abdel Nour[2], Malte Schellmann[3], Tommy Svensson[4], Hao Lin[5], Honglei Miao[6], Hua Wang[7], Jian Luo[3], Milos Tesanovic[8], Nuno Pratas[9], Sandra Roger[10] and Thorsten Wild[1]

[1] *Nokia Bell Labs, Germany*
[2] *IMT Atlantique Bretagne-Pays de la Loire, France*
[3] *Huawei German Research Center, Germany*
[4] *Chalmers University of Technology, Sweden*
[5] *Orange, France*
[6] *Intel, Germany*
[7] *Keysight Technologies, Denmark*
[8] *Samsung Electronics R&D Institute, UK*
[9] *Aalborg University, Denmark*
[10] *Universitat Politècnica de València, Spain*
With contributions from Rana Ahmed Salem, Mario Castaneda, Xitao Gong and Dinh Thuy Phan Huy.

11.1 Introduction

The 5th generation (5G) air interface (AI) constitutes the complete radio access network (RAN) protocol stack, i.e., the physical layer (PHY), Media Access Control (MAC), Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP), Radio Resource Control (RRC) and Service Data Adaptation Protocol (SDAP), and all related functionalities describing the interaction between infrastructure and device. Furthermore, it covers all services, bands, cell types, etc., expected to characterize the overall 5G system. This chapter describes the lower part of the protocol stack, namely, PHY/MAC related technologies, and highly related aspects, such as antenna design. Before heading to the detailed elaborations in subsequent sections, we start with establishing basic design criteria and assumptions. While we keep this chapter more open than the current status of discussions in 3GPP, we still relate to 3GPP where possible and reasonable.

Earlier generations of wireless mobile communications under the framework of the 3rd Generation Partnership Project (3GPP) have exclusively used transmission frequencies below 6 GHz. While this frequency range is still of high value for 5G and thus on the agenda of 3GPP, 5G will go beyond this and deploy transmission points radiating with higher frequencies up to 100 GHz. Details on the usage

5G System Design: Architectural and Functional Considerations and Long Term Research, First Edition.
Edited by Patrick Marsch, Ömer Bulakçı, Olav Queseth and Mauro Boldi.

of the different spectral regions can be found in Section 3.4. In particular, the exploitation of millimeter-wave (mmWave) frequencies, in the context of cellular systems typically associated with the wider range of 6-100 GHz, puts a number of challenges on the AI design, requiring special and dedicated mechanisms both at PHY and MAC. 5G will be required to support both frequency division (FDD) and time division duplexing (TDD) and will rely on heterogeneous deployment layouts being built upon a macro layer for providing ubiquitous coverage applying a frequency reuse of one and a small cell capacity layer (probably at mmWave frequencies in many of the deployments) for boosting the throughput in areas of high demand. The following sections will elaborate on various technical parts of the AI having those characteristics in mind and carve out respective particularities.

The support of multiple antennas both at the base station (BS) and at the device will be a fundamental corner stone of 5G. As with 4G, 5G will make use of this for enhancing both the throughput per area via spatial reuse and the coverage - both due to the virtue of beamforming gains and the use of diversity mechanisms, and thanks to coordinative means between adjacent cells exploiting the spatial selectivity of beamformed signals. Beamforming can be implemented in digital domain and/or in analog domain. Especially at mmWave frequencies, due to the large bandwidth, the large number of antennas and the lower efficiency of electronics, analog beamforming technologies need to be exploited to allow practical implementation. Currently, hybrid beamforming techniques that combine the merits of analog and digital beamforming have been developed and are widely considered as a design assumption for mmWave communications. Section 11.5 will provide detailed insights into the overall concept and will elaborate on the available options.

While each transition from one generation to the next (i.e., from 2G to 3G and from 3G to 4G) has introduced fundamentally different signal formats and mechanisms to multiplex users[1], the move from 4G to 5G is expected to be less radical when it comes to this choice. According to agreements at 3GPP RAN meetings[2], the early incarnation of 5G, referred to as New Radio (NR) phase 1, see also Section 17.2, will still rely on Cyclic-Prefix Orthogonal Frequency Division Multiplex (CP-OFDM) as 4G does, multiplexing users in time, frequency, and space. As with 4G, 5G will potentially allow for discrete Fourier transform (DFT) precoding to achieve more favorable peak to average power ratio (PAPR) conditions. Optionally, filtering or windowing functionalities can be used to further enhance specific characteristics of the signal. In Section 11.3, a more in-depth analysis of the waveform candidates and means for multiplexing user transmissions are provided. Beside candidates being in line with the current NR draft from 3GPP, we present further promising enhancements to provide the reader with a more comprehensive long-term view. This is especially relevant as later releases of 5G might still allow for the further introduction of those.

The most potent mechanisms for increasing the reliability of a single wireless transmission link are Forward Error Correction (FEC) and Hybrid Automatic Repeat reQuest (HARQ). 4G applies for the former convolutional Turbo codes (CTC) to protect the data channels, and a combination of tail-biting

1 2G: Frequency-multiplexed channels applying Time Division Multiple Access (TDMA) with Gaussian Minimum Shift Keying (GMSK), multiplexing users in time domain; 3G: Wideband Code Division Multiple Access (WCDMA), multiplexing users in code domain; 4G: Orthogonal Frequency Division Multiple Access (OFDMA) in the downlink and Single Carrier Frequency Domain Multiple Access (SC-FDMA), also known as Discrete Fourier Transform (DFT)-spread OFDM or DFT-s-OFDM, in the uplink, both making use of cyclic symbol extensions and multiplexing users in frequency and time domains.

2 In the framework of the study item 'Study on New Radio Access Technology' [1] followed by the work item 'Work Item on New Radio Access Technology' [2].

convolutional codes (TBCC) and repetition coding for the control channels. 3GPP has agreed towards the use of Low Density Parity Check (LDPC) codes for data channels and Polar codes (PC) for control channels. In the downlink (DL), 4G applies asynchronous adaptive HARQ, while the uplink (UL) is synchronous. Both variants rely on single-bit feedbacks. 5G will rely on asynchronous variants in both directions. Additionally, recent studies indicate the benefit of making the retransmissions adaptive and allowing for more sophisticated feedbacks. In Section 11.4, more details are provided on the related options that can be drawn from both FEC and HARQ, and the respective interdependencies.

The wireless channel is a shared medium, and the available spectral resources consequently have to be allocated to the respective transmission requests satisfying each user, while keeping the overall system spectrally efficient, as covered in detail in Chapter 12. 4G supports both individually scheduled access on a per transmit time interval (TTI) basis (fixed to 1 ms, at least for the earlier incarnations of 4G) and semi-persistent scheduling (SPS). Some of the new use cases being foreseen to be served by 5G and their respective requirements are implying different access types to potentially be more efficient (e.g., contention-based access and pre-emptive scheduling). In fact, 3GPP has recently agreed that NR shall support grant-free, SPS-like, Physical Uplink Shared CHannel (PUSCH) transmissions, which can be used to reduce the scheduling latency [3]. Additionally, for serving use cases with very stringent timing requirements, a more fine-grained resource allocation needs to be applied, i.e., beside the basic scheduling periodicity of 1 ms, 5G requires to support in parallel a faster scheduling process based on the time basis of, e.g., 1/4 or even 1/8 ms. In fact, 3GPP has now agreed that the time interval between scheduling request (SR) resources configured for a user equipment (UE) can be smaller than a slot (which is the basic scheduling unit). What this in practice means is an agreement by 3GPP to support shorter periodicities for transmitting SRs. Section 11.6 covers these points in more detail both for the design of control channels and data channels. Furthermore, when considering high-density deployments of mmWave small cells, reusing the same spectrum and AI for backhaul and access becomes a good option to relax backhaul and deployment cost, as detailed in Section 7.4. For this purpose, joint scheduling of backhaul and access resources across multiple cells becomes essential to allow the system to operate efficiently.

The remainder of this chapter is structured as follows. Section 11.2 covers relevant criteria for the PHY and MAC design, including considerations of harmonization, for instance between different radio access technologies (RATs). Then, Section 11.3 delves into details on waveforms, numerology and modulation schemes, followed by Section 11.4 on coding approaches and HARQ. Section 11.5 ventures in detail into antenna design, analog, digital and hybrid beamforming, Section 11.6 covers novel PHY and MAC design paradigms and specific solutions for serving and multiplexing the main service types envisioned for 5G, before Section 11.7 summarizes the chapter.

11.2 PHY and MAC Design Criteria and Harmonization

The early incarnations of 4G have focused on efficiently delivering mobile Internet services to devices, such as smart phones and tablets. For this kind of traffic, the most relevant performance indicator to improve is the throughput both per user and per area. To satisfy its customers, operators need to deliver sufficiently high and most importantly consistent data rates where and whenever needed. Hence, the main concern in designing 4G has been to maximize spectral efficiency and spatial reuse.

Especially areas not having a dominant connection to a single BS, i.e., at the cell edge, had to be treated carefully. These areas are interference-limited instead of being noise-limited and thus require special attention, e.g., by applying dedicated mechanisms, such as (further enhanced) inter-cell-interference coordination (Fe)ICIC mechanisms to coordinate the transmissions of BSs in the vicinity. For details on these mechanisms and potential improvements for 5G, the interested reader is referred to Section 12.5. 4G has additionally provided a flat network architecture and a self-organizing approach for the handover process, which has reduced latency significantly. Also, high reliability in 4G is ensured through lower-layer techniques, such as HARQ and FEC, and higher-layer techniques, such as the design of RLC with its acknowledged mode (AM) type of transfer. New and emerging use cases, such as ultra-reliable low-latency communications (URLLC), vehicular-to-anything (V2X) or industrial automation, as well as the need to simultaneously support multiple use cases, imply that the AI design needs to be substantially revisited for the 5G era.

More specifically, 5G NR is anticipated to support a much more diverse set of use cases, as outlined in Section 2.2, and with respective requirements detailed in Section 2.3. Obviously, while throughput is still of very high relevance, 5G is required to support a much wider range of requirements being related to various aspects, such as low latency, high reliability, and energy and cost efficiency at the device.

Beside the wide range of use cases to be supported, further aspects requiring special attention are the wide range of deployment types (e.g., dense urban vs. rural, pure macro-cellular vs. heterogeneous networks, and train lines) and link characteristics (e.g., a wide range of Doppler and delay spreads) as well as the wide range of spectrum below and above 6 GHz. Regarding the spectrum, the most probable mmWave frequencies would be 26 GHz, 28 GHz, 32 GHz and 40 GHz, as detailed in Section 3.4. Different frequency bands imply different bandwidths, propagation conditions and/or even regulations. Accordingly, the 5G NR should be scalable and reconfigurable to be able to properly support the different properties of these.

In a nutshell, 5G in general and the PHY and MAC layers in particular should be designed having the following criteria in mind:

- **To have a high degree of flexibility and versatility** for supporting the broad class of services with their associated broad class of key performance indicators (KPIs) and to enable efficient multi-service support (i.e., meeting the high heterogeneity of requirements) while dealing with high heterogeneity of deployment types, operating frequencies and link characteristics;
- **To be highly scalable** to efficiently support a large number of devices and a wide range of antenna system designs, for instance including different hybrid beamforming architectures, different bandwidths and carrier frequency configurations;
- **To allow for satisfactory service quality** where- and whenever needed, both related to consistent service quality (e.g., by introducing special means to improve cell edge performance, such as interference mitigation techniques) and related to the provision of capacity peaks in respective areas (e.g., by the introduction of high capacity links in crowded areas);
- **To be highly efficient** to support the requirements on energy consumption and resource utilization and to enable high spectral, energy, and cost efficiency in general;
- **To be highly robust** to hardware impairments to allow reduction of hardware costs and to enable cost-efficient operation in mmWave frequencies, where such impairments (e.g., phase noise) are in general more severe than in lower frequencies;
- **To be future-proof/forward compatible** to support easy integration of new services, functionalities and new frequencies without the need of redesigning the AI;

- **To enable tight interworking and synergies** between different RATs, such as 4G and 5G, or different 5G AI variants (AIVs), such as variants for below and above 6 GHz. Here, different levels of integration should be possible, ranging from RAN-level integration up to loose higher layer co-existence or even core network (CN) interconnection. In the following, we treat this aspect related to the lower layers of the protocol stack. Further details on the RAN-level integration of multiple RATs or AIVs can be found in Sections 6.5 and 12.4.

11.3 Waveform Design

One fundamental component of the AI is the underlying waveform, which needs to be designed to properly match to various conditions that can be expected during operation of the wireless communication system. For the most general categorization, there are two different types of waveform designs, namely single-carrier and multi-carrier. For single-carrier waveforms, a single symbol – constituted of a modulated data symbol and an appropriate pulse shape – spans the entire bandwidth B available for transmission, and symbols are transmitted consecutively at a rate of B without inserting any guard symbols or zeros in between. Transmitting these signals via delay-spread channels causes the symbols at the receiver to overlap. To compensate for this, channel estimation and calculation of the corresponding coefficients for a finite impulse response (FIR) or infinite impulse response (IIR) equalization filter is required. For channel estimation, preamble signals need to be frequently transmitted to enable capturing the channel's time variance. Based on these, the channel coefficients and appropriate equalization filter coefficients can be calculated at the receiver. The length of the preamble as well as the complexity for filter calculation and filter operations scales with the number of channel delay taps, which is one of the reasons why single-carrier waveforms are preferred for the application in channels with low delay spread. The other reason is that single-carrier waveforms cannot access the channel in a frequency-selective manner, but instead imply an averaging of the channel quality over the transmission bandwidth. Thus, deep fades that may occur in highly frequency-selective channels may severely degrade the system performance.

Multi-carrier waveforms, on the other hand, divide the available transmission bandwidth into a number of N subcarriers of equal bandwidth and thus allow transmitting independent symbols on these subcarriers in parallel. Choosing N narrowband signals instead of a single wideband signal extends the transmitted symbols in time and thus reduces vulnerabilities to delay spreads from multi-path propagation. The subcarrier bandwidth is usually chosen much smaller than the channel coherence bandwidth, so that each subcarrier signal effectively experiences a flat channel. This significantly simplifies the channel equalization process, as each subcarrier signal can be equalized by a single complex multiplication. Moreover, channel estimation is also simplified, since so-called *scattered pilot grids* can be used, where only a few of the total N subcarriers are selected to carry pilot symbols for channel estimation. From this, it becomes evident that the application of multi-carrier signals is favourable in particular for channels with high frequency selectivity, i.e., exhibiting a large delay spread. If coded transmission is applied, the effect of deep fades can further be well alleviated thanks to a frequency-selective channel access, yielding a much better performance than corresponding single-carrier systems. There is a price to pay, though, and that is related to the fluctuation of the transmit signal amplitude leading to a higher peak-to-average power ratio (PAPR) compared to single-carrier signaling, which scales with the number of subcarriers N, challenging the requirements on the power amplifier (PA).

The most prominent representative of a multi-carrier waveform is OFDM. Here, the subcarrier signals utilize the maximum bandwidth of B/N, thus attaining high spectral efficiency. Though the spectra of the subcarrier signals overlap in frequency domain, an orthogonal design of those spectra ensures an interference-free reconstruction at the receiver. The number of subcarriers for practical implementation is typically chosen as a power of 2, which allows using computationally efficient fast Fourier transform (FFT) algorithms for generating transmit signals and analyzing received signals, respectively. The spectrum of each subcarrier signal has a *sinc*-shape, which translates to the rectangular pulse in time domain. Typically, a guard interval is used for the transmission between two successive OFDM symbols in time domain, which should be larger than the channel's delay spread to protect the OFDM symbols from any inter-symbol interference. This, however, creates an additional overhead, which degrades the spectral efficiency. To keep this overhead small, the required size of the guard interval is usually decisive for the subcarrier bandwidth and thus the number of subcarriers N. Most of the OFDM systems operated today fill this guard interval with a cyclic extension of the OFDM symbol, yielding the so-called cyclic prefix of a CP-OFDM signal. However, also other measures are possible, such as zero padding yielding a zero prefix (ZP-OFDM), or using a unique word (UW-OFDM) as a prefix, which can then also beneficially be used for channel estimation purposes (see Section 11.3.1.2 and [4]). Pure ZP-OFDM is usually not favoured to be applied in practice, since it degrades the PAPR and challenges the PA due to the sudden drops of the signal to zero.

As described above, OFDM schemes as used in today's systems have indeed favourable properties for application in practice; however, they also exhibit some drawbacks. In particular, they require tight synchronization in time and frequency to maintain the signal orthogonality, and they are vulnerable to Doppler distortions in highly mobile channels. Moreover, they rely on a fixed configuration of the so-called numerology, constituted by the number of subcarriers N and the size of the guard interval, which is usually chosen as a best fit for supporting all the channel conditions that are expected during system operation. Adaptations or adjustment of the numerology during operation are not yet foreseen and are not well supported by conventional OFDM anyway, in particular if different numerologies should be supported simultaneously within the available bandwidth B to provide the system more flexibility to respond to the particular requirements of new services and use cases. The main reason for all these deficiencies is the *sinc*-shape of the subcarrier spectra, which has a poor localization of the signal power in the frequency domain due to its high side lobes. For improving the spectral containment of the subcarrier signals, OFDM can be extended by filtering components, which suppress the side lobes of the subcarrier signals. This can be done either by windowing of the time domain signal, which translates to filtering each subcarrier signal in frequency domain, or by filtering in time domain with a filter spanning a set of subcarrier signals – a so-called *sub-band*. In both cases, the filters can be designed to exhibit a steep power roll-off at the edges of a sub-band of a desired size, thus minimizing the power leaking into the adjacent band. Depending on the design constraints, successively transmitted filtered or windowed OFDM symbols may overlap, which may be accounted for at the receiver to avoid interference to arise. This becomes necessary, though, only for a larger filter length going significantly beyond the length of the prefix. While for a fixed filter length, filtered OFDM signals can attain a steeper power slope at the edge of a sub-band, windowed OFDM signals provide additional robustness against frequency errors and Doppler distortions – thanks to the fact that the signal spectra of the individual subcarriers have been uniformly modified. If properly designed, filtering and windowing do not change the orthogonality of the OFDM system, and, hence, all algorithms developed for OFDM can be reused without any alteration.

A special case of a windowed OFDM system with overlapping symbols is Filter Bank Multi-Carrier (FBMC). With FBMC, no guard interval is required, and thus the maximum spectral efficiency can be achieved. However, due to the restrictions of the Balian-Low theorem [5], which states that it is not possible to attain maximum spectrum efficiency with a well-localized pulse power in time and frequency while maintaining complex orthogonality, either the power localization needs to be compromised or the signal orthogonality needs to be relaxed. The latter can be realized with Offset Quadrature Amplitude Modulation (OQAM), where the FBMC symbols carry real-valued data on the subcarriers only, and successive symbols are transmitted at double the symbol rate $2/T = 2B/N$. A complex modulation pattern ensures that the real-valued data of symbols overlapping within a period T overlay in different dimensions of the complex signal space, so that they can be easily reconstructed at the receiver. Hence, orthogonality in OQAM-FBMC exists only in the real field and no longer in the complex field as in OFDM. As a consequence, several schemes designed for OFDM cannot be directly transferred to be used with OQAM-FBMC, but require some redesign of selected signal processing procedures. Though OQAM-FBMC received a lot of attention in research during the past years, this latter fact hampered this waveform to get commonly accepted as a mature candidate for 5G. More details on 5G waveform design can be found in [6].

In recent 3GPP discussions on the waveform to be used for NR, where the focus has been set on enhanced mobile broadband (eMBB) and URLLC services, it has been agreed that the waveform underlying NR should be based on CP-OFDM [3]. It may be extended by filtering components like filtering and windowing, but this filtering option should be transparent to the receiver. In practice, this means that the schemes should work properly even for the case that the receiver applies a simple CP-OFDM receiver without any further filtering. This transparency requirement can be supported by all extended OFDM schemes where the length of the filter tails does not go significantly beyond the length of the guard interval. Depending on the particular filter design, evaluations have shown that the tails are allowed to overlap with preceding and succeeding symbols by up to a maximum of 50% of the symbol period N/B for supporting the transparency requirement. However, it should be noted that some performance degradation will always be observed in this case: since CP-OFDM-based processing at the receiver translates to a mismatched filtering if any other filter has been used at the transmitter, the signal-to-noise ratio (SNR) at the receiver cannot be maximized, translating to an effective performance loss. The relative performance loss generally increases with the length of the filter tail. Furthermore, longer filter tails may lead to increased vulnerability to longer channel delay spreads.

To alleviate the problem of the high PAPR of multi-carrier OFDM signals, a single-carrier-like waveform based on OFDM has been introduced as DFT-spread OFDM (DFT-s-OFDM). Here, the modulated data symbols are generated in time domain with a bandwidth covering a sub-band, and this signal is then DFT-transformed and shifted to the desired frequency position. Applying the inverse FFT (IFFT) covering the entire transmission bandwidth B and adding the guard interval then creates the single-carrier-like OFDM signal. At the receiver, the signal is equalized as in OFDM and then transformed via an inverse DFT (IDFT) to obtain the data symbols. For channel estimation, full preamble (or mid-amble) signals need to be used now, meaning that a pilot signal fills the entire sub-band. This way, the simple and efficient OFDM-based processing can be maintained, while the tight envelope of single-carrier signals yielding small PAPR can be adopted. Note, however, that deep fades in the sub-band may again decrease the system performance significantly due to the implicit averaging of the channel quality over the sub-band – a feature inherent to any single-carrier transmission.

The DFT-s-OFDM scheme may also be combined with filtering and windowing, similar to its pure OFDM counterpart. Moreover, an advanced scheme with an inherent windowing operation has been proposed as zero tail (ZT) DFT-s-OFDM [7]. Here, a tail of zeros is used at the beginning and the end of the OFDM-like symbol, whose length can be adjusted to improve the resilience against inter-symbol interference in delay-spread channels.

11.3.1 Advanced Features and Design Aspects of Multi-Carrier Waveforms

In the following, we provide details on advanced features of multi-carrier waveforms and elaborate on various aspects to be accounted for in the waveform design.

11.3.1.1 Dynamic Numerology Switching

OFDM-based multi-carrier waveforms as introduced above provide three degrees of freedom for their overall design, which can be set according to the particular requirements of a desired service or use case: the numerology, consisting of the subcarrier spacing B/N and the length of the guard interval, and the filter used to attain spectral containment, which may be a time domain window or an FIR filter, or a combination of both. The spectral containment of the signal power yielded by the filtering allows partitioning the system bandwidth into separate isolated sub-bands, wherein the numerology can then be configured individually, thus supporting simultaneously different numerology configurations in the same band. This is referred to as frequency domain numerology multiplexing. Numerologies may further be changed over time following a predefined time grid, then yielding a time domain multiplexing. The time/frequency grid defining potential switching borders between different numerologies covers the total set of resources in time and frequency, and enables structuring it into so-called *tiles*, representing subsets of resources dedicated to a particular service with its unique numerology configuration [6]. Among service data, the tiles may also carry control information required for this service, yielding a self-contained structure.

The tile structure currently being discussed in 3GPP NR is based upon the length of a time slot as defined in Long-Term Evolution (LTE), constituted of 7 OFDM symbols and a constant overhead of roughly 7% for the guard interval (prefix). The subcarrier spacing (or symbol duration, respectively) is allowed to be scaled by a factor equal to an integer power of 2, i.e., 2^n with $n \in \mathrm{Z}$. Choosing these factors, while keeping the overhead for the prefix constant, yields a nested structure of the TTI time grid, i.e., consecutively transmitted frames of different TTI size will be aligned in time at regular intervals, determined by the frame with the longest TTI. This feature allows for simple and frequent switching of TTI configurations over time without creating idle times, thus facilitating the desired flexibility in the frame design to respond to different services' demands. Note that, besides improving latency, as detailed in Section 12.3.2, shorter symbol lengths also enhance the signal quality under high mobility, as inter-carrier interference is decreased and channel estimation quality is increased thanks to a denser pilot grid in time domain [6]. However, when choosing short symbol durations with a constant overhead of 7% for the prefix, it may become too short in scenarios with large delay spread, giving rise to undesired inter-symbol interference. In this case, an enhanced prefix is ready to be chosen. If the number of OFDM symbols in the TTI is reduced from 7 to 6, some additional room for the enhanced prefix is gained, yielding an overhead of 25%. This way, we can allow for using a larger prefix within a TTI without violating the TTI time grid, enabling to change the prefix length on a TTI level.

11.3.1.2 Advanced Prefix Design

As the prefix in CP-OFDM is usually discarded at the receiver, there is high motivation to make better use of this signal overhead, which triggered endeavours on advanced prefix design. One approach is to replace the prefix with a known sequence, leading to UW-OFDM. This principle is also called *known symbol padding* (KSP) OFDM in the literature. With the unique word (UW) replacing the prefix, a periodically appearing known sequence is available, which can be used as training signal that comes without any additional cost in signaling overhead. Such training can be used for various purposes, for example, to enhance channel estimation, phase noise tracking, Doppler tracking, synchronization, and to monitor the received signal power - for instance to detect blockage of the link. Furthermore, the UW can be designed to achieve lower PAPR. In [4], it has been shown that by exploiting UW for phase noise estimation, considerable gain can be achieved compared to CP-OFDM when considering low-complexity schemes.

When replacing the prefix in CP-OFDM with a UW, the circular convolution between channel impulse response and the transmitted signal is impaired. Therefore, specific demodulation schemes are needed at the receiver. Three methods have been described in [4], showing that with proper demodulation schemes and moderate complexity increase, no performance loss is observed. Further, full compatibility with CP-OFDM can be achieved regarding multiple access, multiple-input multiple-output (MIMO), and pilot usage. Practical schemes have been proposed in [4], which exploit UW to enhance channel estimation and phase noise tracking.

11.3.1.3 Mitigating Hardware Impairments

One important challenge in the design of the multi-carrier waveform is the fact that the hardware used for implementing the transceiver functionalities typically exhibits various imperfections. This gets more pronounced with higher carrier frequencies. Therefore, the waveform design and evaluation should take into account these hardware impairments. Two most typical hardware impairments are oscillator phase noise and nonlinear characteristics of the PA. The following paragraphs will provide more details on this.

Phase Noise

Free-running oscillator and phase-locked loop (PLL) based oscillators are the most common implementations assumed in the literature [8]. The phase noise of the PLL-based oscillator consists of three main noise sources: the reference oscillator, the phase-frequency detector along with the loop filter, and the voltage controlled oscillator (VCO). Each of these noise sources includes both white noise (thermal noise) and colored noise (flicker noise). The detailed modeling of phase noise can be found in Table 4-2 of [9]. The detrimental effect of phase noise increases as a function of carrier frequency. Phase noise will cause common phase errors (CPEs) and inter-carrier interference (ICI), resulting in an increased error vector magnitude (EVM) of the desired signal. CPE refers to a common phase rotation of all sub-carriers, which can be compensated quite easily in frequency domain. The actual phase rotation requiring compensation is estimated with the help of pilot subcarriers. ICI may be modeled as additive noise (not always Gaussian) and is usually hard to be compensated. It requires denser pilots for phase noise and channel tracking, and estimation and compensation can be computationally intensive. The most straightforward method to mitigate the effect of phase noise is the use of a larger subcarrier spacing, though this may increase the vulnerability to frequency-selective channels if the overhead of the prefix is kept constant, as discussed earlier in Section 11.3.1.1.

Non-linear Characteristics of the PA

When digitally modulated signals go through a PA having non-linear characteristics, spectral regrowth appears, which in turn causes adjacent channel interference. The power series model or the polynomial model is widely used in the literature for the modeling of memoryless nonlinear PAs [10], which is given by:

$$y(t) = \sum_{k=0}^{K} c_{2k+1} \left| x(t) \right|^{2k} x(t) \tag{1}$$

where K is the non-linear order, $y(t)$ is the output signal, $x(t)$ is the input signal, and c_{2k+1} is the $(2k+1)$th complex-valued polynomial coefficient. The coefficients c_{2k+1} can be calculated by using least squares estimation (LSE). More recently, there has been growing interest in modeling nonlinear PAs with memory effects, for instance based on a memory-polynomial model or Volterra series [4].

In order to have high PA efficiency, it is required that the input signals have low PAPR. One may also employ linearization techniques, e.g., pre-distortion, to compensate for PA nonlinearity. However, such technique comes with major baseband complexity and may not be effective for large bandwidth signals and/or hybrid beamforming architectures. It remains an open question to which extent such techniques can be effective for large bandwidth signals, as for instance envisioned in the context of mmWave, and how much additional complexity is required.

11.3.1.4 PAPR Reduction Techniques

One of the main drawbacks of OFDM as compared with single-carrier waveforms is the high PAPR, which requires the PA to operate linearly in a very wide range. As explained in the introduction of Section 11.3, DFT-s-OFDM is a means to create a single-carrier-like signal based on OFDM, though its performance suffers from deep fades, which may occur in particular if the transmit signal covers a broader bandwidth. Hence, PAPR reduction techniques for conventional OFDM are worthwhile to be explored. It is noteworthy that a low PAPR is not only important in the UL, but also in the DL, in particular at very high frequencies (e.g., mmWave bands) due to the need for low-cost BSs. Various simple and effective PAPR reduction techniques for OFDM (e.g., amplitude clipping, exponential companding, and constrained clipping) have been proposed in the literature, see [4] for further details and corresponding references.

Figure 11-1 (top) shows the complementary cumulative distribution function (CCDF) of the PAPR for DFT-s-OFDM as well as for OFDM with and without various PAPR reduction schemes. Through applying appropriate PAPR reduction techniques, it is shown that OFDM can achieve a similar PAPR performance like DFT-s-OFDM.

Figure 11-1 (bottom) shows the EVM performance for DFT-s-OFDM and OFDM signals with and without the different PAPR reduction schemes at different power back-off settings. Here, the EVM is measured at the receiver, where the signal distortions are arising from the noise, from distortion being introduced by the PAPR reduction techniques, and from the nonlinear characteristics of the PA. With high power back-off, the signal is less distorted by the nonlinear PA; thus, the EVM performance of DFT-s-OFDM and OFDM is almost the same, while OFDM with the PAPR reduction schemes shows degraded EVM performance due to the distortion introduced by those schemes. With low power back-off, the nonlinear effects of the PA are introducing additional signal distortions, where the severity of the distortion depends on the PAPR of the input

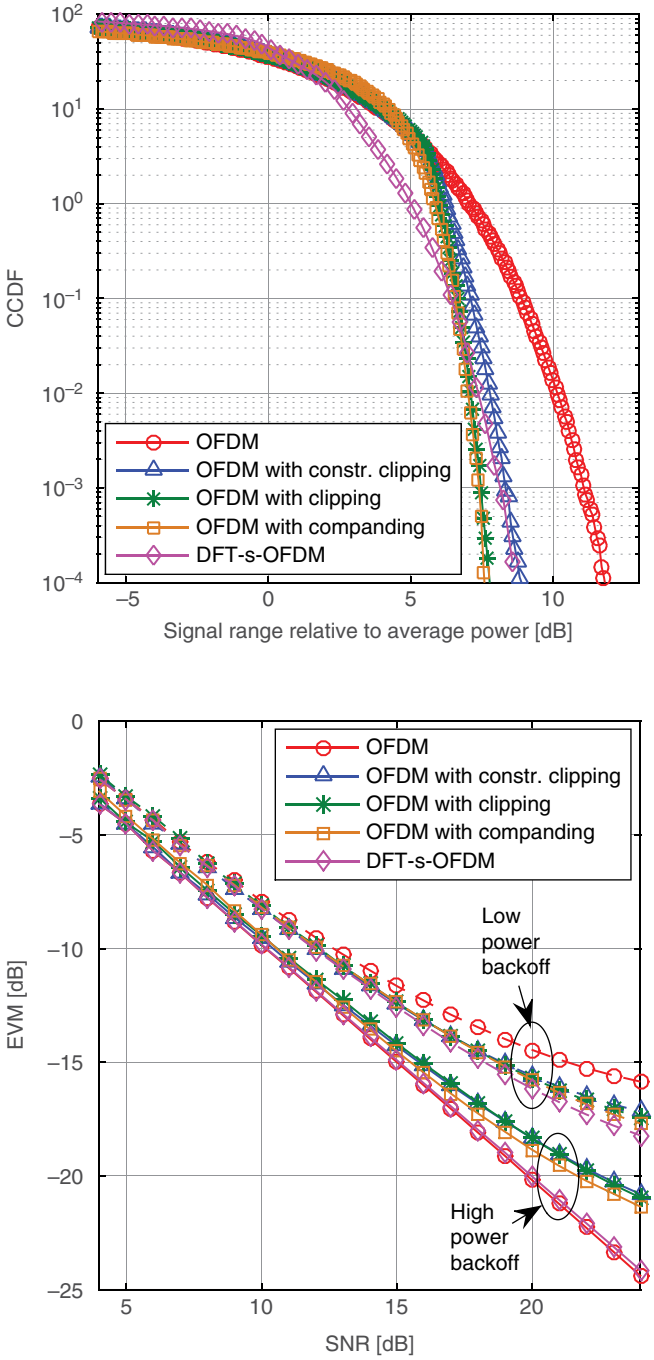

Figure 11-1. CCDF of the PAPR for DFT-s-OFDM and OFDM signals with/without PAPR reduction schemes (top) and corresponding EVM performance (bottom).

signal. As DFT-s-OFDM and OFDM with PAPR reduction have lower PAPR than OFDM, OFDM performs the worst among all schemes, while OFDM with PAPR reduction schemes achieves similar EVM performance as compared to DFT-s-OFDM. It is assumed here that the low power back-off is insufficient to provide the required "headroom" to accommodate the OFDM signal, but sufficient to accommodate the OFDM signal with PAPR reduction. Furthermore, it is shown that the degradation of EVM in OFDM with PAPR reduction is more pronounced at high SNRs than for low SNRs, where the additive channel noise is the dominant adverse factor.

11.3.2 Comparison of Waveform Candidates for 5G

Several windowed and filtered multi-carrier schemes have been investigated in recent years, targeting on finding the most suitable candidates for 5G. These have been evaluated and compared with respect to their capability to address the particular requirements of the future 5G system as well as to their behaviour under given hardware impairments, which get particularly pronounced in the context of higher frequencies, as discussed in the previous section. In this section, a summary of the most important findings from this evaluation is presented. A detailed description of all multi-carrier schemes as well as further details on their evaluation and comparison can be found in [4][6]. The following multi-carrier waveform candidates have been in the focus of those investigations:

- **Conventional (non-filtered)**: CP-OFDM; or single-carrier-like variant DFT-s-OFDM;
- **Subcarrier-wise filtered**: Windowed (W)-OFDM, pulse-shaped (P)-OFDM, UW-OFDM, flexibly configured (FC)-OFDM, and OQAM-FBMC;
- **Sub-band-wise filtered**: Universal-filtered (UF)-OFDM and block-filtered (BF)-OFDM.

As detailed at the beginning of Section 11.3, one favorable feature of the novel waveforms is the spectral containment of the signal power, which facilitates easy coexistence of different radio configurations in the same frequency band and allows for asynchronous UL access. To highlight how well the different waveform candidates can realize this feature, two appropriate scenarios have been selected for evaluation and direct comparison [6]:

- **Scenario 1 (asynchronous UL access)**: Here, 3 UEs are assigned to adjacent sub-bands, each spanning 48 subcarriers, corresponding to a total of 720 kHz for 15 kHz subcarrier spacing. The UE in the center sub-band is the one being evaluated. The receiver is synchronous with this UE, while the two UEs in the adjacent sub-bands are misaligned in their timing with a constant time offset relative to the center UE, going beyond the size of the prefix. A guard band of configurable size is used between the adjacent sub-bands of different UEs.
- **Scenario 2 (UL synchronous transmission with mixed numerology)**: Here, 2 UEs are assigned to adjacent sub-bands of equal size (720 kHz), whereas the UE located aside the evaluated UE uses double the subcarrier spacing (30 kHz) of the evaluated UE (15 kHz). A guard band of configurable size is used between the adjacent sub-bands of different UEs.

Performance comparisons in terms of the Turbo-coded block error rate (BLER) versus the effective SNR are shown in Figure 11-2 (top) for scenario 1 and in Figure 11-2 (bottom) for scenario 2. The effective SNR reflects the total average power spent per subcarrier signal, including the power contained in the prefix overhead, if applicable. Standard system settings for an LTE system operating at 4 GHz with 10 MHz bandwidth have been applied, if not stated otherwise. In Figure 11-2 (top), 16-QAM modulation has been used, and the guard band between the frequency sub-bands of different users

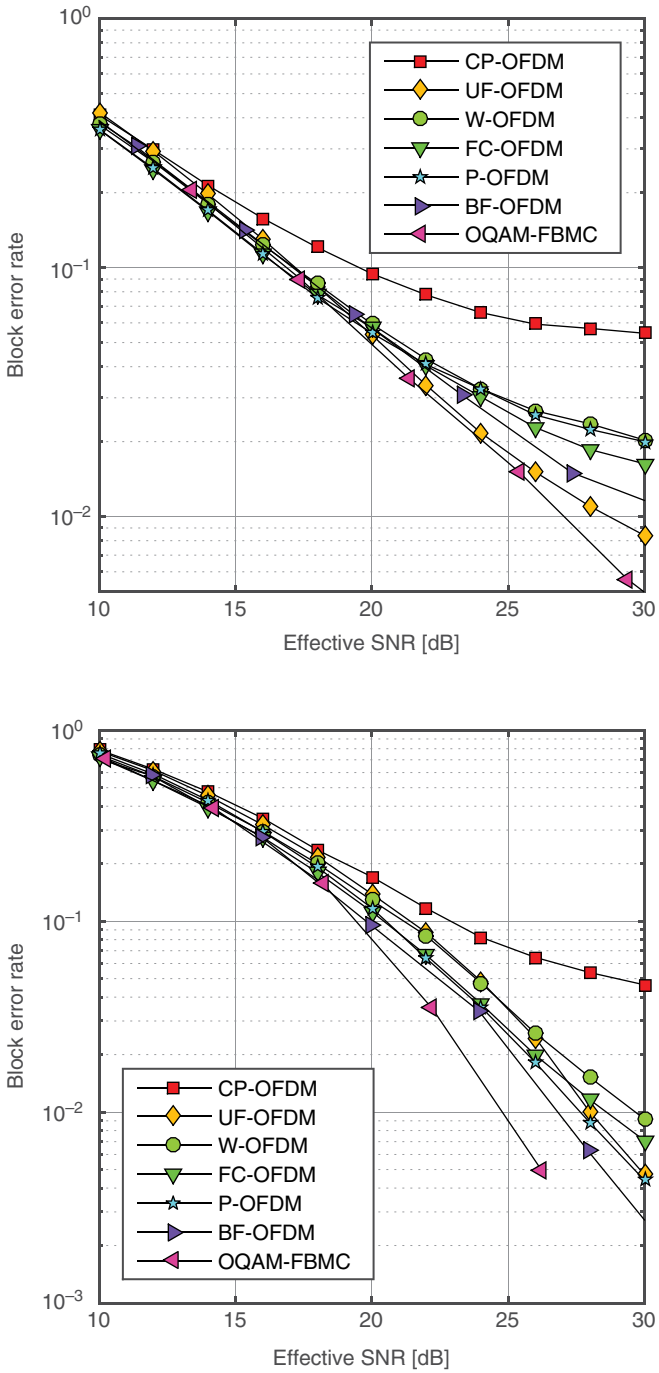

Figure 11-2. Performance comparison of the waveform candidates for asynchronous UL access (scenario 1, top) and mixed numerology coexistence (scenario 2, bottom).

is set to 30 kHz. The adjacent users are misaligned in their timing by a number of samples amounting to 1/8 of the FFT block size. From the figure, it can be observed that all evaluated waveforms attain much better BLER performance than conventional CP-OFDM. The best performance is attained by the OQAM-FBMC scheme, which is not fully OFDM-compatible, though. The sub-band filtered schemes get closest to this superior performance, while the windowed schemes follow at some distance. It is worth mentioning that with increasing the guard band size, the performance of all modified OFDM schemes converges towards that of OQAM-FBMC, while the performance of W-OFDM is kept at some distance.

In Figure 11-2 (bottom), 64-QAM modulation has been used, and the guard band between the frequency sub-bands of different users is set to 60 kHz. We observe that OQAM-FBMC again attains the best performance, followed at some distance by the modified OFDM schemes, which do not differ too much from each other, but all significantly outperform conventional CP-OFDM. Again, W-OFDM performance is kept at some distance. For further details on these waveform comparisons, the interested reader is referred to [6].

Effects of Hardware Impairments in the Context of mmWave Transmission

Non-ideal properties of the hardware implementing mmWave transceiver components cause impairments, as addressed in Section 11.3.1.3. Further impairments include in-phase and quadrature phase (I/Q) imbalance, sampling jitter and sampling frequency offset, carrier frequency offset, etc. These imperfections are present in every hardware implementation, but their impact in the mmWave frequency range is larger than in sub-6 GHz bands, because these hardware components are operated closer to the overall physical limits and therefore closer to the limit of their capabilities. In particular, the PA has lower efficiency at mmWave frequencies, so that increased power consumption for a given transmit power target is expected. Therefore, it is important to have low-PAPR waveforms, as stressed earlier in Section 11.3.1.4.

Based on those mmWave-related challenges, a number of KPIs have been selected for the evaluation of the waveform candidates, including: spectral efficiency, PAPR, phase noise robustness, robustness to frequency/time selective channels, MIMO compatibility, time localization, out-of-band emissions (with and without PA), complexity and flexibility. A summary of prominent results is presented here, while detailed evaluation results can be found in [4].

Figure 11-3 shows the power spectral density (PSD) of different waveforms with and without hardware impairments. Without any hardware impairments (top figure), it is shown that indeed very low out-of-band (OOB) emissions can be achieved with FBMC, W/P-OFDM and UF-OFDM due to the filtering/windowing operations, as compared to CP-OFDM and DFT-s-OFDM. When phase noise is included (bottom figure), the sharp spectrum roll-off provided by FBMC-OQAM, W/P-OFDM, and UF-OFDM is significantly reduced, but is still much lower than that of CP-OFDM and DFT-s-OFDM. When a nonlinear PA is further added, it is observed that the sharp spectrum roll-off promised by these waveforms is unlikely to be achieved, due to the fact that PA non-linearity leads to spectral regrowth. However, for low power transmission, i.e., with a relatively high power back-off, OOB advantages over OFDM can still be maintained.

Figure 11-4 shows the EVM performance of different waveforms under different hardware impairments. It is observed that there is no significant difference in the EVM performance among the various candidate waveforms. It is generally known that multi-carrier waveforms are sensitive to phase noise. However, with phase noise compensation and sufficiently large subcarrier spacing, the multi-carrier

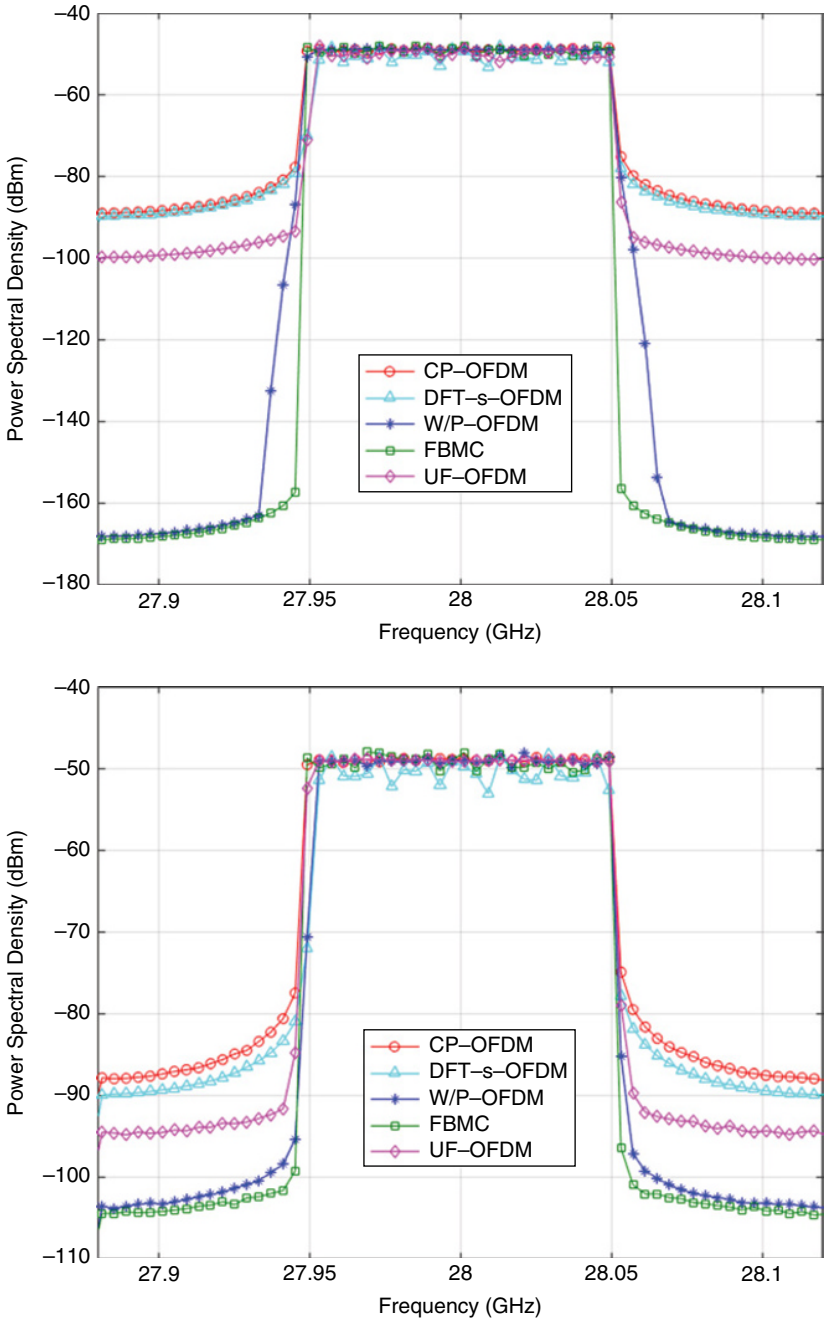

Figure 11-3. PSD of different waveforms without any hardware impairments (top) and with phase noise (bottom).

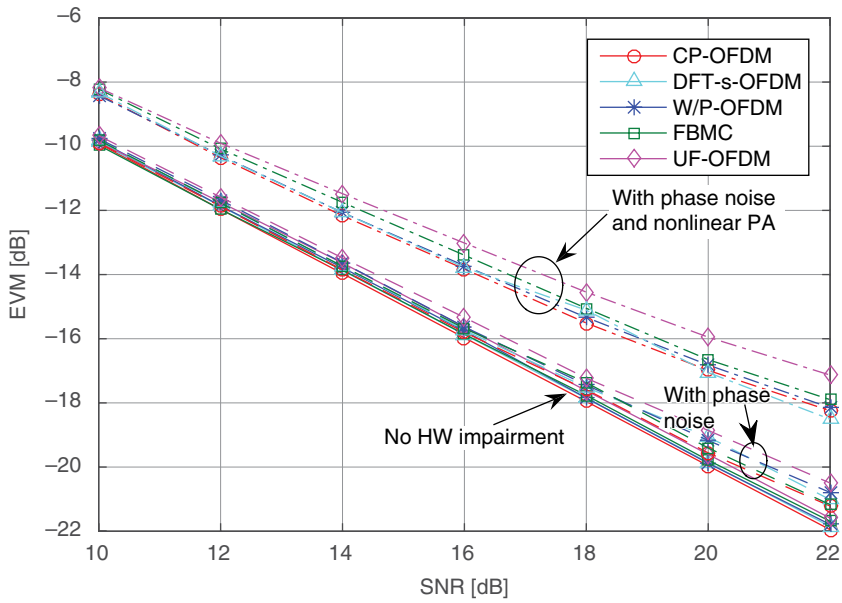

Figure 11-4. EVM performance of different waveforms with hardware impairments.

waveforms can be made robust against phase noise effects. Besides, scattered pilot-based channel estimation was used, which also compensates for the CPE caused by phase noise.

The waveform comparisons presented before have revealed that the waveforms extended by filtering or windowing provide significant gains compared to conventional CP-OFDM in scenarios reflecting novel applications envisaged for 5G. Looking at their performance under hardware impairments, it is observed that the studied waveforms have similar EVM performance as conventional CP-OFDM. The new waveforms exhibit improved OOB performance as compared to CP-OFDM, but the improvement gets smaller when hardware impairments, such as phase noise and nonlinear PAs, are taken into account. For nonlinear PAs with high power transmission, i.e. small power back-off, the OOB advantage finally vanishes, and similar OOB emissions are observed for all waveforms.

11.3.3 Co-existence Aspects

The 5G waveform should be capable to coexist with the CP-OFDM waveform, which is motivated by the following two reasons: First, during the early deployment of 5G systems, the existing 4G bands will not be re-farmed for 5G usage immediately. However, to allow for a gradual and effective penetration of the 5G system, some of the 4G bands, especially under-used UL carriers, could be shared with the 5G RAT. This kind of sharing could be semi-static or dynamic. Thus, with 4G and 5G RATs sharing the same band, the respective waveforms used in 4G and 5G, respectively, need to

co-exist. Second, 5G is envisioned to support various kinds of services beyond conventional (e)MBB traffic, as outlined in Section 2.2. At the moment of writing this section, 3GPP NR has decided to adopt a CP-OFDM based waveform for eMBB services [3], while other waveforms are not precluded for other services, such as mMTC or V2X. This suggests that even within the 5G system, possible other waveforms need to coexist with CP-OFDM. For facilitating this coexistence, two options can be considered:

1) The co-existing 5G waveform should be designed as orthogonal (or quasi-orthogonal) as possible with respect to the 4G waveform to effectively minimize the mutual interference. Naturally, this will limit the degrees of freedom in the overall waveform design. Some possible solutions are: The 5G waveform can be based on subcarrier-filtered waveforms on top of CP-OFDM, with the window length being limited to the length of the prefix, or, alternatively, the 5G waveform can be based on sub-band filtering on top of CP-OFDM, where the filter length has similar restrictions.
2) Another option is to release the requirement of the 5G waveform to be fully orthogonal to CP-OFDM. In this case, the inter-system interference needs to be controlled by applying a frequency guard band, combined with a power back-off at the edge of the sub-band used by any 5G service. This will allow for more ambitious 5G waveform designs. The inter-sub-band interference from the "4G sub-band" to the "5G sub-band" can be handled by the 5G receiver thanks to spectral confinement techniques, whereas the interference from 5G to 4G is mitigated by proper guard band dimensioning and by using a power back-off at the edge of the sub-band used by 5G.

Another aspect constraining 4G-5G co-existence resides in the basic frame design of LTE. The Physical Downlink Control CHannel (PDCCH) occupies one up to three successive OFDM symbols across the entire band. Furthermore, LTE applies cell-specific reference symbols (CRS), which are scattered across the bandwidth and which are not allowed to be muted. To care for this, the 5G signals would have to be muted at the respective positions in the time-frequency grid. The introduction of mini-slots in 5G NR supports to solve this issue. In the UL, the respective frame design decisions in 4G are less restrictive, as both the control channel and the reference symbol placements are frequency-localized.

11.3.4 General Framework for Multi-Carrier Waveform Generation

Multi-waveform harmonization was not critical for the design of LTE, because the main target use case has been MBB, traditionally well-served by the use of CP-OFDM. However, in 5G, the use cases as well as the respective requirements are much more diverse, motivating the use of different waveform alternatives with different features, as already discussed in previous sections. In particular, one of the major motivations for waveform harmonization is to enable the support of different waveforms for different services with minimized implementation effort. This objective can be attained by a modular structure based on the reuse of hardware components.

In [11], a general framework based on the mathematical tool known as Gabor systems [5] was introduced, which allows to represent different multi-carrier waveforms under the same system model by selecting the appropriate prototype filter, subcarrier spacing and symbol spacing in time.

In fact, the system model is useful to represent conventional CP-OFDM, W-OFDM, P-OFDM and ZT-DFTs-OFDM. Furthermore, the framework is also valid to represent waveforms of the FBMC family, such as FBMC-QAM and FBMC-OQAM. As a result, this general framework properly facilitates a harmonized hardware implementation capable to generate multiple waveforms.

Harmonized Implementation Concept for Multiple Waveforms

The block diagram of a harmonized transmitter capable to implement the generic multi-carrier waveform following the general framework is presented in Figure 11-5. The diagram corresponds to an implementation based on poly-phase filtering, carried out in time domain through a poly-phase network (PPN) [12]. As further elaborated next, by selectively enabling or disabling particular blocks, the harmonized implementation is able to generate each of the mentioned waveform variants.

Note that the blocks necessary to generate the CP-OFDM signal are shown in white, whereas the extra blocks required for the generation of any of the other waveforms are highlighted in grey. The special blocks included in the harmonized implementation, which require specific configuration for some waveforms, are detailed next:

- **DFT spreading**: This block is intended to perform the spreading operations necessary for the generation of ZT-DFT-s-OFDM;
- **OQAM pre-processing**: This set of blocks contains the necessary preparative multiplexing steps required for FBMC-OQAM, that is, complex-to-real number conversion of QAM complex symbols, up-sampling, and time staggering;
- **PPN**: The task of this block is to perform the convolution of the discrete signals with a filter implemented through a PPN.

The harmonized block diagram illustrates the usefulness of the proposed implementation to provide flexible adaptation to a particular communication scenario, i.e. to allow for a dynamic waveform selection, and, at the same time, to reduce implementation costs. Regarding the OFDM variants, all of them will leave aside the OQAM pre-processing and will include the prefix addition, except for ZT-DFT-s-OFDM, which leaves also the prefix aside. Only ZT-DFT-s-OFDM makes use of the DFT spreading block, though. With respect to the PPN block, its inclusion will actually depend on the specific variant: Plain CP-OFDM will not require this block, whereas W-OFDM and P-OFDM will need it for windowing. Concerning the FBMC transmitters, they must enable all the blocks in Figure 11-5 except those for involving the prefix and DFT spreading. The operations in charge of OQAM generation are only necessary for the transmission of FBMC-OQAM. Further

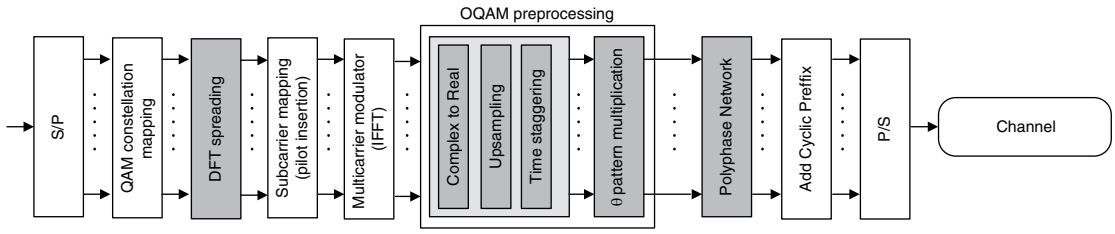

Figure 11-5. Harmonized transmitter for multi-carrier waveform generation.

implementation aspects and complexity evaluations of other waveforms and of the harmonized transceiver are covered in detail in Section 16.3.3, where the typical savings provided by the harmonized transmitter compared to the independent implementation of all constituting waveforms are shown to be in the range between 60–75%.

11.4 Coding Approaches and HARQ

11.4.1 Coding Requirements

An unprecedented variety of new applications and services are foreseen to be introduced in the future 5G communications systems, as detailed in Section 2.2. This results in challenges and constraints for the envisioned usage scenarios, such as very high user data rates for eMBB services, stringent reliability and latency constraints for URLLC, or the transmission of short packet messages with sporadic traffic for mMTC. Therefore, a special emphasis has to be placed on the design of FEC solutions able to efficiently support the underlying constraints. In this regard, three main KPIs provided in 3GPP NR [13] can be clearly identified as of high importance for FEC choice and design:

- For eMBB, the target for peak data rate should be 20 Gbps for DL and 10 Gbps for UL;
- For URLLC, the target for user plane latency should be 0.5 ms for UL, and 0.5 ms for DL;
- For URLLC, the target for reliability should be 10^{-5} of packet error rate (PER) with keeping the time constraint of 1 ms. This reliability performance shall be supported together with user experienced data rate on the order of 300 Mbps.

Unfortunately, the FEC coding and modulation components of LTE and LTE-Advanced (LTE-A) are not optimal in this respect, as they were not designed to meet such requirements. Actually, the following weak points have been identified [14] for the Turbo codes (TCs) employed in LTE:

A known issue related to TCs resides in their poor performance at low error rates when transmitting data with coding rates higher than 1/3. This is due to the so-called *error floor*, which can be observed when a TC is punctured with the rate matching mechanism. A detrimental resulting effect is the frequent resort to HARQ retransmissions. Consequently, the LTE FEC code cannot simultaneously meet the reliability and latency constraints of URLLC usage services. Moreover, the error floor issue of TCs is also not compatible with the requested increased user data rate for eMBB services, since high coding rates are then required.

Using convolutional component codes, conventional TCs have not been originally designed for encoding short blocks. So, the LTE/LTE-A FEC code does not provide capacity-approaching performance for the transmission of short data packets. In particular, it is called for using tail bits for trellis termination. On one hand, this results in a non-negligible bandwidth efficiency reduction for short blocks. On the other hand, this type of trellis termination introduces low-weight truncated codewords and does not ensure the same protection for all data bits, since tail bits are not encoded twice (i.e., Turbo encoded) as the rest of the data. Therefore, the LTE TCs need to be improved to be able to efficiently cope with the sporadic traffic of short messages, as typical for mMTC services.

The target peak rate of 20 Gbps for eMBB represents a major challenge for any family of FEC codes. This is particularly true when taking into account the required flexibility, on the order of what is

specified in LTE in terms of supported packet sizes and coding rates. Regarding TCs, due to the recursive nature of the underlying convolutional codes, their decoding structure is serial by nature, and they are known to have difficulties in achieving extremely high data rates at a reasonable implementation cost.

Accordingly, coding solutions able to answer favorably to the identified constraints of the different usage scenarios of 5G have been identified and evaluated in the FEC selection process of 3GPP.

11.4.2 Coding Candidates

Defined around 10 years ago, the LTE TC is somewhat dated, delivering performance far from what can be achieved from best FEC code designs nowadays. For example, an enhanced TC family was designed to target the requirements of the different scenarios of 5G [15][16]. On the other hand, taking the envisioned peak data rates into account, LDPC codes [17] may represent a better choice as a coding solution. Furthermore, Polar codes (PCs) [18] concatenated with an outer error detecting code, such as a cyclic redundancy check (CRC) code, have recently emerged as strong coding candidates for short block sizes. In addition to these capacity-approaching coding solutions, convolutional codes and block codes, such as Reed Muller or Bose–Chaudhuri–Hocquenghem codes (BCH) can be of interest for the particular case of extremely short packet sizes, for instance less than 40 bits.

In the rest of this section, a special focus is put on describing the latest advances regarding the three main families of codes represented by TCs, LDPC codes, and concatenated PCs.

11.4.2.1 Enhanced TCs

The enhanced TC (eTC) family [15][16] was designed to address the drawbacks of the existing TC solution in LTE, when targeting the requirements of the different scenarios of 5G. The encoder structure is a parallel concatenation [19] of two 8-state recursive systematic convolutional encoders, as shown in Figure 11-6. Each component code is a modified version of the LTE TC component code with an additional parity symbol W, resulting in a TC with a mother coding rate equal to $R = 1/5$. The generator polynomials for C1 and C2 are $(1,(1+D+D^3)/(1+D^2+D^3)$ and $(1+D+D^2+D^3)/(1+D^2+D^3))$, respectively.

Tail-biting, also called circular encoding, is introduced. It ensures that, when encoding a message of length K, the initial and the final states are identical for each component encoder C1 and C2. Tail-biting is the best-known termination method for TCs, since it avoids the transmission of tail bits. Thus, there is no rate loss and the spectral efficiency of the transmission is not reduced. Moreover, with tail-biting, all the information bits are protected in the same way by the TC, and the circular property prevents the occurrence of low-weight truncated codewords. Therefore, tail-biting termination helps to lower the error floor.

Rate adaptation is performed via the application of a periodic puncturing pattern of length Q. The information block size K is assumed to be a multiple of Q. Typical values for Q are 4, 8, 16 or 32, but others values are possible, provided that Q is a divisor of K. The selection of the puncturing patterns is performed on the basis of a joint analysis of the Hamming distance spectrum of the punctured component convolutional code and of the mutual information exchange between the two component encoders [15]. Incremental puncturing patterns have been designed, enabling inherent HARQ support via incremental redundancy.

The interleaver has an important impact on the performance of TCs in the low error rate region. For implementation-friendly designs, algebraic interleaving is adopted in most standards. Amongst

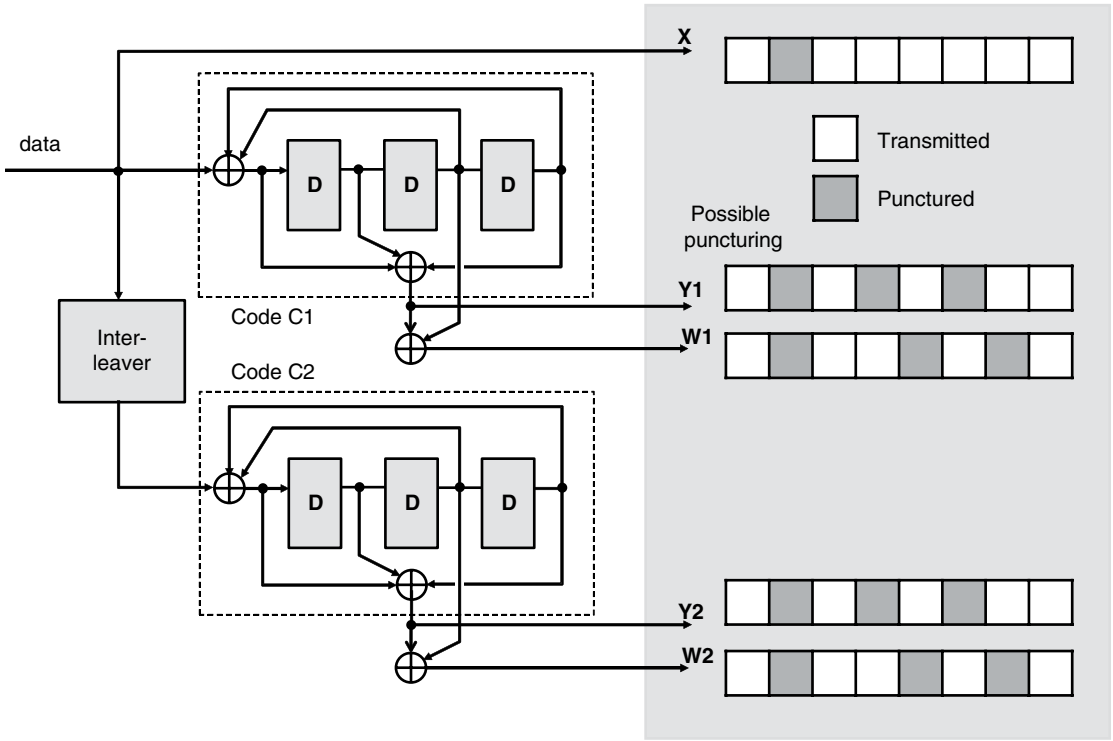

Figure 11-6. Turbo encoder with periodic puncturing.

the different existing models, it was shown in [20] that the *almost regular permutation* encompasses most of the popular algebraic interleavers, including the quadratic permutation polynomial (QPP) model of the LTE TC. Therefore, almost regular permutation interleavers can achieve at least the same minimum Hamming distances as QPP interleavers and were adopted for the proposed eTC. The corresponding interleaving function is given by the following equation:

$$\Pi(i) = (Pi + S(i \bmod Q)) \bmod K \tag{2}$$

where i denotes the address of the data symbol after interleaving, and $\Pi(i)$ represents its corresponding address before interleaving. P is a positive integer relatively prime to K. S is a vector containing Q integer values. The values of parameters P and $S(i)$, $i = 0 \cdots Q-1$, are chosen to support the different block sizes and coding rates. Their selection procedure follows the steps described in [16] to generate a so-called *protograph-based* interleaver design.

The eTCs are decoded using an iterative process that exchanges probabilistic extrinsic information between two component convolutional decoders, each applying a variant of the Bahl, Cocke, Jelinek and Raviv (BCJR) algorithm in the logarithmic domain, commonly named *scaled Max-Log MAP (maximum a posteriori) algorithm* [21]. After several decoding iterations, for instance 6 to 8, the final binary decision is provided.

11.4.2.2 LDPC Codes

LDPC codes, first proposed in [17], were re-discovered in the mid-90s by McKay and Neal [22]. These are block codes with sparse parity check matrices (PCM). The sparsity property is enforced to reduce correlation in decoding. Bipartite Tanner graphs are used to define the connections between the variable nodes associated to code bits and the check nodes associated to the parity-check equations. Thanks to their inherent parallel structure, LDPC codes present advantages in terms of achieved decoding throughput, making them strong candidates to comply with the peak data rate requirements of 5G.

A specific family of LDPC codes named multi-edge (ME)-LDPC [23] codes was proposed by several parties at 3GPP as a coding solution for eMBB data in 5G. The ME-LDPC family can be seen as a generalization of the irregular LDPC ensemble framework with larger degrees of freedom for code design and flexibility. This family is characterized by the presence of several edge types in the Tanner graph, as opposed to the standard irregular LDPC code ensemble. The construction consists of a base Tanner graph with several edge types and includes punctured nodes, called *state nodes*, for enhanced performance. Systematic ME-LDPC codes are obtained by the introduction of degree-two parity variable nodes via an accumulate chain.

A base ME Tanner graph is designed for each desired coding rate. Afterwards, the final PCM is obtained by lifting the base graph to the desired codeword size. Targeting hardware-friendly designs, used lifting is generally a cyclic copy obtained through circulant matrices. This makes the ME-LDPC code a quasi-cyclic code, therefore greatly simplifying the encoding operation as well as the code description.

A Tanner base graph example of a ME-LDPC code with size 24 before lifting is shown in Figure 11-7. The circles correspond to the variable nodes of the base graph, and the squares are the parity-check nodes. The T-shaped dongles on the top of each variable node represent the transmitted bits. The variable node with no dongle represents a punctured state node, intentionally introduced to improve performance. The degree-two parity node accumulate chain is shown on the right with dashed edges. For a codeword size N, the lifting size Z satisfies $N = 24 \times Z$. To obtain a cyclic lifting, each edge has to be associated with an integer from the cyclic group modulo Z.

The design of the base Tanner graph for each target coding rate has an important impact on the overall performance of the code. Parameters such as the girth, i.e., the minimum cycle length in the Tanner graph, also play an important role in resulting performance. ME-LDPC codes are generally designed using the density evolution technique [24].

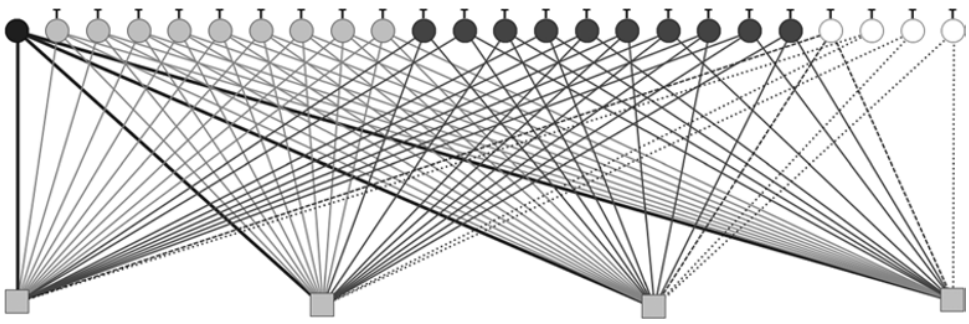

Figure 11-7. Tanner graph example of an ME-LDPC code with a base graph size of 24 nodes.

HARQ support via incremental redundancy (IR) can be implemented by designing the highest-rate code and then extending this to lower rates with the addition of extra parity bits. However, due to the incremental redundancy constraint, the base graphs of the higher-rate codes have to be subgraphs of the base graphs of the lower-rate codes. This generally results in a low-rate base graph different from the one obtained directly by density evolution, with poorer performance.

The decoding of ME-LDPC codes requires iterative processing and exchange of extrinsic probabilistic information between variable nodes and check nodes, based on a low-complexity variant of the belief propagation principle in the logarithmic domain, called *scaled* or *offset min-sum algorithm* [25]. After several iterations, for instance 10 to 25 layered iterations, the final binary decision is provided. Since the lift size Z is generally larger than the number of columns in the base graph, a decoding hardware capable of processing Z liftings in one clock cycle would allow a high decoding throughput, which is particularly appealing for 5G. Note that the complexity of LDPC decoders is also discussed in further detail in Section 16.3.4.

11.4.2.3 Polar Codes

A PC can be viewed as a recursive concatenation of a base short block code designed to transform the encountered transmission channel into a set of virtual channels with variable levels of reliability. The first description introducing the idea of channel polarization and the framework of PC design was explicitly provided in [18].

The capacity C of a binary input symmetric discrete memoryless transmission channel T satisfies $0 \le C(T) \le 1$. The problem of designing an error correcting code for the two extreme values is simple to solve. The target of polarization is to transform N channel uses of T into N virtual channels with capacity 0 or 1, as shown in Figure 11-8. Polarization is performed via the application of XOR operations as described in [18]. It can be achieved for quasi-infinite packet sizes. However, for finite-length packets, a non-uniform reliability distribution is obtained for the different virtual channels.

The simplest application of this principle advocates the design of a base matrix respecting the polarization constraints that is recursively used until polarization is achieved over the codeword length N.

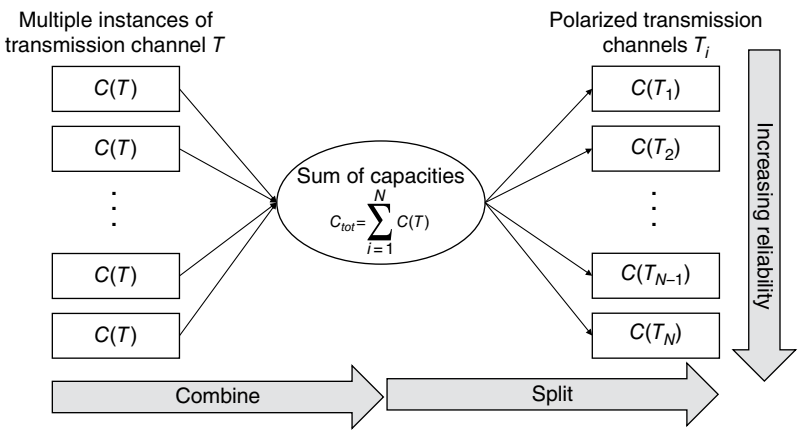

Figure 11-8. Channel polarization.

The right lower dashed rectangle in Figure 11-9 represents the encoder of a size-4 Polar codeword. The upper dashed rectangle represents the code extension to a codeword size of 8 bits. Rate compatibility is achieved by transmitting a subset of bits u_i enjoying the highest reliability levels after polarization. Bits that are not transmitted, called frozen bits, correspond to the least reliable channels after polarization.

The corresponding code shows capacity-achieving performance for quasi-infinite packet sizes with a successive interference decoder [18]. However, for most packet sizes used in practical applications, PCs decoded with this type of decoder show poor performance, far from the best achieved by TC and LDPC codes.

At a later stage, a list-based decoder was proposed in [26]. It classifies codewords in increasing reliability order. Then, a concatenation with an outer error detection code, typically based on a CRC, was introduced to eliminate the least reliable codewords from the list [27]. The resulting concatenated structure is able to bridge the existing performance gap with TC and LDPC codes. However, there still exist practical drawbacks in terms of implementation efficiency and parallelization.

During the standardization process of 5G, a novel concatenated structure, called parity-check PCs, was proposed [3]. This structure divides the codeword into independent sub-codewords

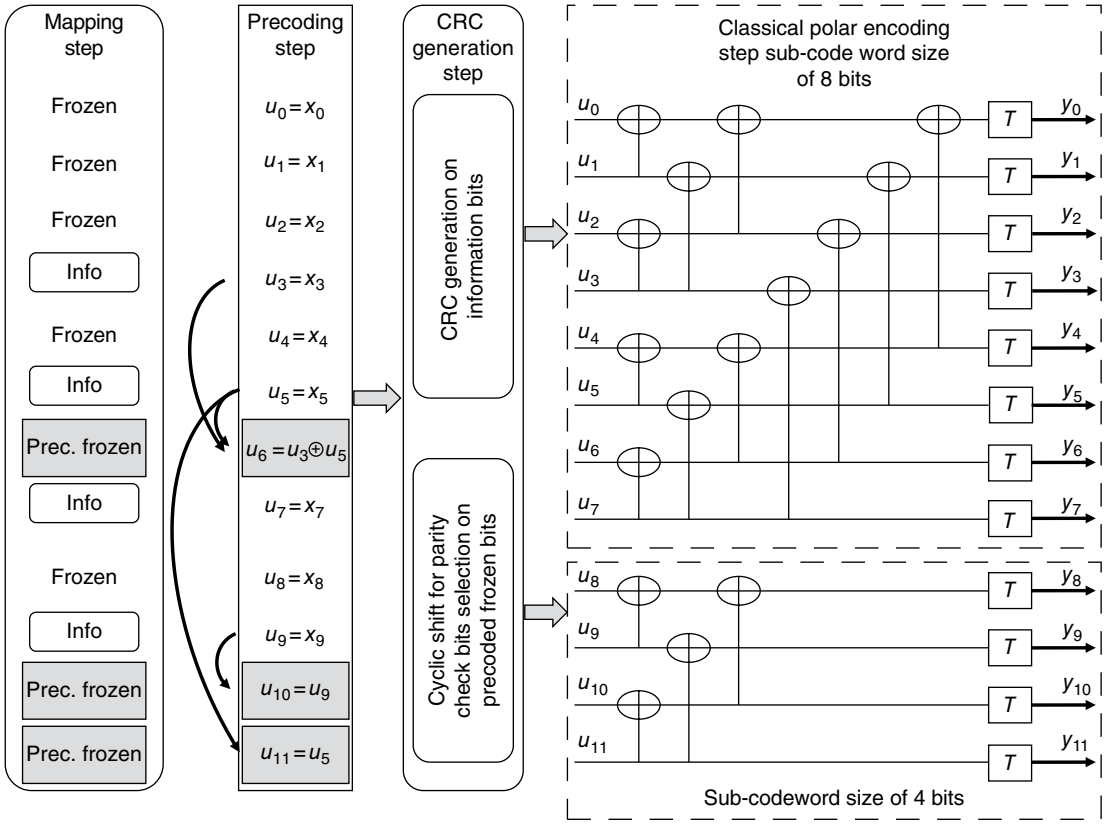

Figure 11-9. Parity-check Polar encoder including classical CRC Polar encoder.

(potentially of different sizes) linked solely by a precoding step. This latter consists of an XOR or a repetition operation that generates precoded frozen bits (u_6, u_{10} and u_{11}), as shown in Figure 11-9. The precoding step is followed by the generation of cyclic parity-check bits, separately on the precoded frozen bits and on the information bits, used for error detection. Two families of parity-check bits are generated from precoded frozen bits: self-parity-check bits computed from bits of the same sub-codeword (such as, u_6 and u_{10}), and cross-parity-check bits computed from bits of adjacent sub-codewords (such as, u_{11}). Finally, classical Polar encoding is performed.

For each sub-codeword, the corresponding decoder unrolls and sorts its list of candidate codewords in increasing reliability order. Then, adjacent sub-codewords exchange information on cross-parity-check bits. These bits can now be decoded as frozen bits by each individual list decoder, pruning the list of candidates of the corresponding sub-codeword. The resulting decoder, called *chained-list* decoder enjoys lower latency and can achieve higher throughputs since sub-codewords can be decoded in parallel, making it particularly appealing for the peak data rates of 5G. Moreover, a well-constructed PC with a proper parity-check function over some well-chosen frozen bits increases its minimum Hamming distance, thereby improving the error correction performance. In fact, appending CRC bits as in [27] can be seen as a special case of the parity-check PC.

HARQ support is achieved through shortening via incremental frozen bits selection. Shortening is applied on the sub-codewords, such that the resulting total size is strictly larger than the size of the codeword for the lowest supported coding rate.

11.4.3 General Summary and Comparison

Strong and weak points of each of the three different families of codes considered for 5G are identified in this section. They are based on observed tendencies regarding performance, implementation complexity, flexibility and maturity, and hence characterizing these families.

It is extremely difficult to draw conclusions from sole performance comparisons, since they should be performed at equal complexity. Moreover, computational complexity is far from being an accurate representative of final hardware complexity and power efficiency. It is indeed widely acknowledged that memory resources and memory accesses have a large impact on chip area and power consumption, respectively. Therefore, fair comparisons require the availability of hardware designs for all three coding families, and comparable assumptions related to quantization, number of decoding iterations, achieved throughput, technology, etc. Nevertheless, since performance results comparing these families have shown that the observed gaps for large block sizes (i.e., larger than $K = 1024$ bits and code rates ranging from $R = 1/3$ to $R = 8/9$) were within the limits (≤ 0.3 dB) related to the assumptions made for reduced complexity hardware designs, it can be concluded that all three families of codes offer a satisfactory level of performance for these cases. A similar conclusion was reached during the selection process of 3GPP for 5G [28].

LDPC codes are considered as widely implemented in commercial hardware supporting several Gbps throughput with attractive area and energy efficiency, but with a flexibility support which is far below the requirements for the eMBB services of 5G. Actually, the area efficiency of LDPC decoders reduces when decreasing the coding rate, and their complexity rises when the flexibility is increased. Moreover, despite the ability of achieving large parallel decoding degrees, some of this parallelism may not be exploited for all code block lengths and code rates, resulting in a penalizing impact on energy and area efficiency. In addition, for the shortest block sizes of some 5G scenarios, penalizing short cycles cannot be avoided. This leads to poor performance of this family of codes for

short block sizes. On another note, as previously mentioned, IR-HARQ support by ME-LDPC codes entails performance penalties for low rate ME-LDPC codes, compared to the best-known low-rate ME-LDPC codes and to TCs. This is due to the design constraints forcing high-rate base graphs to be embedded into low-rate base graphs. To conclude, hardware implementation with attractive area and energy efficiency is considered challenging when simultaneously targeting the peak data rate and flexibility requirements of eMBB services of 5G.

PCs are considered implementable, although there are currently no commercial implementations available, and, in relation to the eMBB service of 5G, there are some concerns linked to the maturity and the availability of decoding hardware. In addition, most existing work in the literature is related to successive interference cancellation decoders and not to list-based decoders that are required to enable the excellent performance of this family of codes. The implementation complexity of list-based decoding increases with the list size, especially for large block sizes. Moreover, the area efficiency reduces for shorter block lengths and lower coding rates. A list-4 decoder is largely agreed as implementable for all codeword sizes. However, in practice, most existing simulations results considered list-8 decoders that could be argued implementable only for short block sizes. Besides, IR-HARQ support for PCs entails performance penalties because the best positions of frozen bits are not necessarily incremental when lowering the coding rate. To conclude, decoding hardware can now achieve acceptable latency, performance and flexibility for PCs, but there are still some concerns about the feasible area efficiency and energy efficiency, and about the maturity of the technology.

TCs are widely implemented in commercial hardware, supporting IR-HARQ and the flexibility constraint required for 5G, but not at the high data rates or low latencies needed for the eMBB usage scenarios. In fact, TCs meet the flexibility requirements of 5G with the most attractive area and energy efficiency except at very high throughputs. With TCs, for a given code structure, the area and energy efficiency is constant when varying the coding rate, via puncturing and HARQ. Another advantage resides in the fact that the decoding complexity increases linearly with the information block size for a given mother code rate.

Due to the peak data rate requirements of 5G, in addition to a stand-alone TC solution, a combination of TC (for flexibility) and LDPC (for high throughputs) codes was proposed [29]. It considers designing a Turbo decoder capable of decoding both LTE and, at least, lower information block sizes ($K \leq 6144$ bits) of the eMBB scenario of 5G. For the high throughput case of 5G, a LDPC code can be designed with a limited flexibility or equivalently for a few combinations of code rates ($R > 1/2$) and block sizes ($K > 6144$ *bits*). This proposal has the benefit of combining the advantages of each family (TC and LDPC) of codes without bearing the burden of their drawbacks. Indeed, it was shown in [29] that this combination of codes could answer favorably all the requirements of 5G, especially in terms of complexity.

As a general conclusion, we can state that each family of codes presents its challenges when trying to satisfy simultaneously all the requirements of 5G. Therefore, it is quite difficult to clearly identify an all-around favourite without performing a joint thorough analysis of performance, complexity and latency taking into account real implementations. Due to timing constraints, the framework for such a comparison was not agreed in 3GPP, and individual technical contributions were used as a basis for the selection process. Finally, regardless of potential technical drawbacks, a compromise was found that led to the adoption of LDPC codes for eMBB data channels and PCs for control channels [3].

The choice of the coding solution for mMTC and possibly URLLC scenarios remains an open issue. While simulation conditions for these two scenarios are quite different from those of eMBB, they are partly

related to the comparisons performed for short block sizes. In fact, block sizes lower than 1024 bits were considered for coding rates from $R = 2/3$ down to $R = 1/12$. Error rates of 10^{-4} to 10^{-5} of PER are targeted.

Taking into account the performance results provided in [30] comparing PCs and eTCs, we can clearly identify these two families of codes as strong candidates for URLLC and mMTC from the performance point of view, with a slight edge for eTCs showing improved performance for these targeted low rates [31]. LDPC codes cumulate two main drawbacks: the first lies in the large performance penalty (more than 1.0 dB for short block sizes and low rates) in some cases, the second is the fact that decoding complexity increases by orders of magnitude when decreasing the coding rate, compared to the two other families of codes.

From the complexity point of view, TCs present an advantage for such low rates as $R = 1/12$ since the decoding complexity of this family of codes scales linearly with the information block size K and not with the codeword size N. This is not the case for both Polar and LDPC decoders. To identify the best technical choice, in-depth complexity comparisons, going beyond simple computational complexity, would have to be carried out at comparable performance. The framework for such comparisons should be clearly set and agreed between the parties proposing these coding solutions.

Finally, the selection process for 5G coding solutions has launched a wave of new proposals for the three families of codes. Current studies are focusing on improving the decoding efficiency of Turbo decoders when targeting high throughput scenarios. A large number of studies are also focusing on the design and implementation efficiency of PCs and related decoders. Therefore, significant improvements are being made, and a thorough investigation taking into consideration performance and hardware complexity between these two strong candidates represented by TCs and PCs should be performed before a final selection.

11.4.4 Hybrid Automatic Repeat reQuest (HARQ)

HARQ is a tool being applied in many communication systems to both increase the reliability of a single transmission and the overall spectral efficiency. The cost to pay is a higher device complexity (e.g., a buffer is needed), processing effort, and latency. With a system applying HARQ, any data transmission is to be acknowledged by the respective recipient, i.e., in the DL by the device, and in the UL by the BS. Based on a given decision criterion (e.g., a CRC check), the integrity of a received packet is checked. If this check is not passed, a respective feedback message "negative acknowledged" (NACK) is sent to request a retransmission, or otherwise the successful reception is acknowledged (ACK). In the former case, the original message or another redundancy version is subsequently transmitted. The receiver combines both the original message and the retransmission(s) to eventually detect the original message. So, HARQ is a tool to increase the reliability by making use of time and (depending on the variant) frequency diversity, energy, and (again depending on the variant) coding gain. Obviously, there is a multitude of design choices possible for implementing this mechanism. In the following we introduce these, highlight the design choices of 4G, and present potential improvements for 5G NR.

The core functionality of HARQ is to enable the system to potentially retransmit a given packet if the reception has not been successful so far. For the actual implementation various aspects have to be decided for:

How to Check if a Transmission has been Successfully Received?

The receiver needs to be able to check the integrity of each received packet. The most reliable and complex variant is to base this decision on the CRC. The CRC check relies on the received bits after

FEC and hard decision. Less complex but also less reliable is to base this decision on the statistical characteristics of the soft-symbols before the demapper or of the soft-bits before or after the decoder. The potential costs to pay are reduced reliabilities (if the check wrongly indicates a successful reception) or reduced system throughput (if the check wrongly indicates a non-successful reception). The potential merits are less processing burden and reduced processing time, which can contribute to reducing the overall latency.

How to Design the Feedback Carrying the Outcome of this Check (Single-bit vs. Multi-bit Feedback)?

Data transmission both in 4G and in 5G NR is based on so-called transport blocks. A transport block comprises all bits being transmitted within a given transmission opportunity. For efficiency reasons, a transport block may be segmented into code blocks, each being individually encoded. The variant with the least overhead is to spend a single-bit feedback referring to the whole transport block. If only parts of the transport block are corrupted (e.g., due to localized interference), resources are wasted, as the complete transport block has to be unnecessarily retransmitted. With allowing single code blocks to be (non-)acknowledged, this can be avoided. The cost to pay is the increased overhead (i.e., one bit per code block instead of a single bit for the complete transport block).

The Timing between the (re-)Transmissions and the Related Feedback (Synchronous vs. Asynchronous HARQ) and the Overall Number of (re-)Transmissions Possible

HARQ can be implemented both in a synchronous and in an asynchronous manner. The former has a fixed timing between the (re-)transmissions and the related ACK/NACK feedbacks. The 4G UL applies this variant with 4 ms, or 4 TTIs, being the time span between (re-)transmissions in the UL and ACK/NACK feedback in the DL, and allows for up to 4 retransmissions. In contrast to this, 4G applies asynchronous HARQ for DL transmissions. In this case, the BS has more degrees of freedom related to setting up the timing, as the retransmissions are treated alike scheduled transmissions. The latter, however, implies more overhead, as the BS has to communicate its decisions. The advantage of this is a higher flexibility. As outlined several times, 5G NR is foreseen to require a very high degree of flexibility to be able to serve the various use cases, and, for instance, meet the stringent latency requirements for URLLC services. Consequently, non-elastic mechanisms, such as synchronous HARQ, should be avoided in 5G, as already agreed in 3GPP [3].

How to Configure the Retransmission both w.r.t. Resources to be used and the Transmission Parameters (Adaptive vs. Non-adaptive HARQ)?

Again two variants are possible. With non-adaptive HARQ, the retransmission requires to use the very same resources (naturally, in a respectively later sub-frame) and configuration (e.g., selected modulation and coding scheme). As above, the merit of this variant is a lower overhead, since the configuration is implicitly known at the cost of fewer degrees of freedoms. In the 4G UL, non-adaptive HARQ is applied, while the 4G DL applies adaptive HARQ. 5G NR is foreseen to use in both directions the adaptive variant to avoid troublesome restrictions.

In a nutshell, compared to 4G, 5G NR requires to move the implementation of HARQ towards a higher flexibility at the cost of higher processing effort and overhead. In this section, we have treated

all available options in a rather abstract manner. For further details, the interested reader is for instance referred to [6].

11.5 Antenna Design, Analog, Digital and Hybrid Beamforming

As highlighted in the previous sections, there is a need to more densely reuse the spatial domain in wireless communications. This needs to be done both at each network node and by densifying the network. More spatial efficiency at the node level can be obtained by using many antennas at the transmitter and/or receiver. This approach is in general referred to as MIMO in the literature. Depending on the channel properties and scenario, MIMO systems can be configured for spatial (transmit and or receive) diversity, beamforming, multiplexing, and spatial multiple access. The maximum diversity gain that can be achieved in a MIMO system equals the number of independent paths between antenna pairs, and the maximum number of spatial streams that can be supported equals the minimum of the number of transmit and receive antenna elements in the system. However, note that full diversity and spatial multiplexing cannot be obtained simultaneously [32].

In the diversity mode, sufficiently separated antenna elements are used to make the link more robust (i.e., lower the outage) by transmitting and/or combining different redundancy versions of the signal by taking advantage of more-or-less independently fading propagation paths between the transmit and receive antenna elements. Receive diversity is sometimes called single-input-multiple-output (SIMO), and transmit diversity is sometimes called multiple-input-single-output (MISO) in the literature, when the transmitter/receiver has only one transmit/receive chain, respectively.

In the beamforming mode, the complex baseband weights of the transmitted and/or the received signals at each antenna element are chosen to adjust and to shape the transmit and/or receive beams in order to increase the signal-to-interference-plus-noise ratio (SINR) of the link, and to avoid harmful interference on other links. Transmit beamforming can be done either in a user-agnostic manner or in a per-user channel-aware manner. The latter case is normally denoted as precoding in the literature (sometimes also the per-user power allocation is included in the notion of precoding), and requires knowledge of the users' channels at the transmitter.

In spatial multiplexing, multiple signals are sent as independent streams to obtain throughput gains whenever there is sufficient multipath propagation in the environment, and provided that channel state information (CSI) at the transmitter side can be obtained. The multi-user case of spatial multiplexing is called spatial multiple access, in which the streams can be aimed for different users.

As highlighted above, the more transmit and receive antennas can be used, the higher the potential for spatial diversity, beamforming, spatial multiplexing and spatial multiple access systems is, provided the spatial channels are sufficiently uncorrelated. In particular, with spatial multiplexing and multiple access, the capacity increases and the required transmit power decreases with the number of antennas. For this reason, intensive research has been carried out on so-called massive or large MIMO [33], addressing both theoretical and practical challenges. From the theoretical side, asymptotic results show that in rich multipath fading environments so-called *channel hardening* [34] appears, essentially eliminating the fading, and with suitable precoding orthogonal spatial channels for the users can be obtained. However, to practically implement large MIMO systems, there are challenges related to, in particular, channel state acquisition, hardware impairments and signal processing complexity. In FDD systems, CSI needs to be obtained by reference or pilot symbols in UL

and DL, whereas, with transceiver chains in TDD that are calibrated to compensate for the different Rx/Tx hardware (HW) properties, it is possible to take advantage of the reciprocity of the propagation channel, and thus probing signals need to be sent in only one direction. For this reason, TDD is seen as the most practical possibility to implement large MIMO systems. A drawback with any half duplexing system, however, is the so-called half-duplex loss, since the transceivers take turn in transmitting and receiving. This loss would be avoided if full duplex (FD) at the same frequency would be possible to implement, cf. Section 16.2.4. With FD, there is the potential to double the spectral efficiency and increase the resource allocation flexibility. Perhaps even more important for some 5G use cases, FD has the potential to reduce the access delay by a factor of two in the system, since all time slots can be made available for transmission. Still, reference symbol design for large MIMO systems is a challenge in order to obtain sufficient CSI knowledge at the transmitter side with a reasonable overhead. In particular, intensive research has been carried out on the so-called pilot contamination problem [35], since orthogonal reference symbols would impose too much overhead on the system.

More spatial efficiency at the network level can be obtained by a dense deployment of infrastructure nodes. This is especially important at mmWave frequencies, due to challenging propagation conditions, such as a higher path loss, penetration losses, and shadowing, as detailed in Chapter 4. In addition, the hardware capabilities are worse with weaker output power and noisier oscillators, causing phase noise to be a design issue, as detailed in Section 11.3.1.3. On the other hand, the small wavelength at higher frequencies can be exploited to pack a large number of antennas in a small area, allowing for large beamforming gains at reasonable form factor, which allow to overcome the weaker output power. At the receiver, sufficiently large antenna area could be achieved creating a larger effective aperture, thus enabling large enough antenna array gain, and also enabling receive beamforming to be implemented for additional directivity gains. It should be noted that mmWave channels are typically wideband and spatially confined causing sparse multipath, and, hence, the massive MIMO focus for mmWave frequencies is rather on beamforming, while spatial multiplexing is less relevant for these frequencies, unless the antennas are spatially distributed.

In the following, an overview of the multi-antenna support in 3GPP NR is given, followed by a more detailed discussion on the so-called hybrid-beamforming architecture, which is a promising approach for large antenna arrays in particular at mmWave carrier frequencies. Then, a short discussion is devoted to an alternative MIMO architecture that has received increased attention to address the complexity, energy consumption, and cost: digital beamforming with finite precision digital to analog converters (DACs). Finally, although currently not supported by NR, the potential for so-called *massive multiple-input massive multiple-output* (MMIMMO) is shown to be promising to boost the spectral efficiency in specific scenarios. It might be feasible from a complexity point of view using a novel spatial multiplexing scheme.

11.5.1 Multi-Antenna Scheme Overview of 3GPP NR

In the 3GPP LTE standard, there has been support of various MIMO schemes from the start, i.e., since Release 8. In Release 13, the eNodeB can be configured in ten different so-called transmission modes (TMs), implementing transmit diversity, beamforming and spatial multiplexing. The TM can be selected per UE based on the channel properties, UE and eNodeB capabilities. The CSI feedback can contain the so-called rank indicator (RI) describing the rank of the MIMO channel, i.e., the number of sufficiently spatially separable channels, the precoding matrix indicator (PMI) describing the

preferred precoder from a set of pre-defined codebooks, and a channel quality indicator (CQI) used for adaptive transmission and multi-user scheduling. Up to four spatial layers are supported, but only two codewords can be transmitted simultaneously to a given UE. Due to the potential of dense spatial reuse, in 3GPP NR, a beam-oriented approach is adopted, and there is inherent support of distributed cooperative transmission and reception schemes, as well as more advanced MIMO schemes. Below, the main novelties of NR are summarized.

11.5.1.1 Beam Management

In 3GPP NR [1], beam management is defined as a set of layer-1 (L1) and layer-2 (L2)[3] procedures to acquire and maintain a set of transmit-receive points (TRPs) and/or UE beams that can be used for DL and UL transmission and reception. Specifically, at least the following aspects are addressed by beam management procedures:

- **Beam determination**: For TRP(s) or UEs to select their own Tx/Rx beam(s);
- **Beam measurement**: For TRP(s) or UEs to measure characteristics of received beamformed signals;
- **Beam reporting**: For UEs to report information of beamformed signal(s) based on beam measurements;
- **Beam sweeping**: Operation of covering a spatial area, with beams transmitted and/or received during a time interval in a predetermined way.

According to [1], the following DL L1/L2 beam management procedures are supported within one or multiple TRPs:

- **DL beam alignment**: Utilized to enable UE measurements on different TRP Tx beams to support selection of TRP Tx beams and UE Rx beam(s);
- **DL beam refinement**: Utilized to enable UE measurement on different TRP Tx beams to possibly change inter- or intra-TRP Tx beam(s);
- **UE receive beam refinement**: Utilized to enable UE measurements on the same TRP Tx beam to change the UE Rx beam in the case that a UE uses beamforming.

At least network-triggered aperiodic beam reporting is supported under the above three beam management related operations. UE measurements based on reference signals (RSs) for beam management (at least CSI-RS) are composed of K beams, and UEs report measurement results for N selected Tx beams, where N is not necessarily a fixed number. Note that the procedure based on RS for mobility purposes, such as synchronization signal blocks, is not precluded. Reporting information at least includes measurement quantities for N beam(s) and information indicating N DL Tx beam(s), if $N < K$.

NR also supports the following beam reporting considering L groups, where $L \geq 1$ and each group refers to an Rx beam set or a UE antenna group. For each group l, the UE reports at least the following information:

- Information indicating group at least for some cases;
- Measurement quantities for $N \cdot l$ beam (s);
- Information indicating $N \cdot l$ DL Tx beam(s), when applicable.

3 L1 refers to the PHY sublayer, while L2 refers to MAC/RLC/PDCP sublayers.

NR supports that the UE can trigger a mechanism to recover from beam failure. A beam failure event occurs when the quality of beam pair link(s) of an associated control channel falls low enough (i.e., involving comparison with a threshold and time-out of an associated timer). The mechanism to recover from beam failure is triggered when beam failure occurs.

11.5.1.2 MIMO Schemes

For NR, the number of codewords per Physical Downlink Shared Channel (PDSCH) assignment per UE is 1 codeword for 1 to 4-layer transmission and 2 codewords for 5 to 8-layer transmission.

DL demodulation reference signal (DMRS) based spatial multiplexing is supported. At least 8 orthogonal DL DMRS ports are supported for single user MIMO (SU-MIMO), and a maximum of 12 orthogonal DL DMRS ports are supported for multi-user MIMO (MU-MIMO). At least the following DMRS based DL MIMO transmissions are supported for data in NR:

- **Scheme 1**: Closed-loop transmission where data and DMRS are transmitted with the same precoding matrix;
- **Scheme 2**: Open loop and semi-open loop transmissions, where data and DMRS may or may not be restricted to be transmitted with the same precoding matrix.

For the DL data, at least a precoding resource block group (PRG) size for physical resource block (PRB) bundling equal to a specified value is supported. A configurable PRG size is also supported for data DMRS. DL transmission scheme(s) achieving diversity gain at least for some control information transmission are supported.

11.5.1.3 CSI Measurement and Reporting

For NR, DL CSI measurements with up to 32 antenna ports are supported. At least for CSI acquisition, NR supports CSI-RS and SRS. NR further supports aperiodic, semi-persistent, and periodic CSI reporting. The periodic CSI reporting can be configured by a higher layer, above PHY. Higher-layer configuration includes at least reporting periodicity and timing offset. By semi-persistent CSI reporting, configuration of CSI reporting can be activated or de-activated.

CSI reporting with two types of spatial information feedback is supported.

Type I CSI feedback is the normal CSI feedback scheme. As in 3GPP LTE, it consists of codebook based PMI feedback with normal spatial resolution. The PMI codebook has at least two stages, where the first stage comprises of beam groups and vectors. Type I feedback supports at least the following (DL) CSI reporting parameters:

- Resource selection indicator (i.e., reference signal sequence or beam);
- RI;
- PMI;
- Channel quality feedback.

There is support for multi-panel scenarios by having a co-phasing factor across antenna panels.

Type II CSI feedback is an enhanced CSI feedback scheme, enabling explicit feedback and/or codebook-based feedback with higher spatial resolution. At least one scheme must be supported from the following Category 1, 2, and/or 3 for Type II CSI:

- **Category 1**: Precoder feedback based on a linear combination of dual-stage codebooks. Specifically, stage one consists of a set of L orthogonal beams taken from a set of 2-dimensional DFT beams,

and the beam selection is wideband. The L beams with common stage one precoder are combined in stage two, which supports subband reporting of phase quantization of beam combining coefficients;

- **Category 2**: Covariance matrix feedback. The feedback of the channel covariance matrix is long-term and wideband. A quantized/compressed version of the covariance matrix is reported by the UE. Specifically, the quantization/compression is based on a set of M orthogonal basis vectors, where M is the number of supported simultaneous beam pair links, and the maximum value of M may depend at least on UE capability. The reporting can include indicators of the M basis vectors along with a set of coefficients;
- **Category 3**: Hybrid CSI feedback. A type II Category 1 or 2 CSI codebooks can be used in conjunction with 3GPP LTE Class-B CSI feedback using beamformed CSI-RS to reduce the RS overhead and improve coverage. The LTE Class B CSI feedback can be based on either Type I or Type II CSI codebook.

11.5.2 Hybrid Beamforming

Due to the large number of antennas in the transmit and receive arrays required to enable mmWave communication, equipping each antenna with a separate radio frequency (RF) transceiver chain along with a high-resolution converter, as done with smaller arrays at lower frequencies, would result in a high complexity, cost, and power consumption. This is mainly due to the implementation of RF components at mmWave frequencies, as well as the expected large bandwidths, which impose the requirement on the DACs at the transmitter and analog to digital converters (ADCs) at the receiver to operate at a high sampling rate. Thus, equipping one converter per antenna in a large antenna array translates inevitably into a high power consumption and cost. For this reason, analog beamforming with a single RF chain has been adopted in early standards, such as in IEEE 802.11ad. However, since this architecture offers limited signal processing capability, a hybrid beamforming architecture [36] using a reduced number of RF chains, and subsequently converters, has attracted substantial attention as a promising solution in particular for mmWave scenarios, and is depicted in Figure 11-10.

With hybrid beamforming, the number of RF chains N^{RF} is smaller than the number of antennas in the array, e.g., ${N_{\mathrm{tx}}}^{\mathrm{RF}} \leq N_{\mathrm{tx}}$ at the transmitter. By splitting the beamforming operation between the analog RF domain and the digital baseband, this architecture provides a reduction of complexity and power consumption [36], at the expense, however, of reduced degrees of freedom for the baseband digital processing, and a consequently reduced possible number of streams $N_{\mathrm{s}} \leq {N_{\mathrm{tx}}}^{\mathrm{RF}}$. Compared to

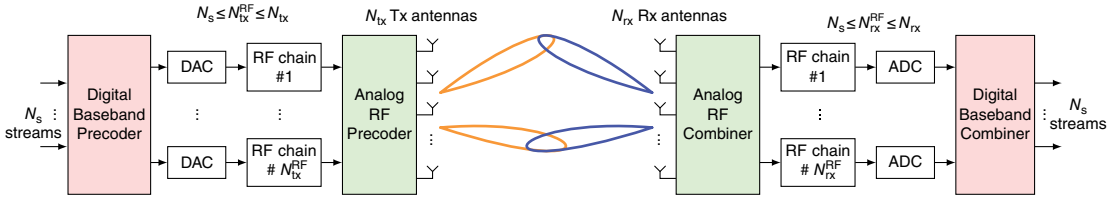

Figure 11-10. Hybrid beamforming with analogue RF beamforming.

a fully digital system, the hybrid beamforming architecture poses different challenges for the CSI acquisition and beamforming design due to the constraints on the analog processing and the need of directional transmission at mmWaves.

For example, the analog processing is frequency flat, which implies that the analog beamforming matrix is fixed for all subcarriers in a multi-carrier system, whereas the digital beamformer can be adapted for each subcarrier. For the wideband hybrid beamforming design, however, the fact that the spatial characteristics of the channel are frequency-invariant can be exploited. Similarly, for hybrid beamforming in a multi-user scenario, the design of the analog beamforming needs to consider that it is common for all users. Furthermore, the analog processing can be implemented via a network of phase shifters, RF switches or with a lens antenna array, and can be implemented at different stages including RF, intermediate frequency and baseband. In case of phase shifters, the entries of the analog beamformer are constrained to have unit modulus. An open question is, however, how the limited number of RF chains should be connected. A partially connected architecture has recently been proposed [37], where the output of each RF chain is connected only to a subset of the transmit antennas. This approach reduces the required number of phase shifters as well as the losses, thereby facilitating the implementation of hybrid beamforming at the expense of reduced design flexibility. If each RF chain (or set of RF chains) is connected to a distinct set of antennas, the architecture is based on subarrays, where each subarray is basically connected to its own transceiver. The partially connected architecture is applicable at both the transmitter and receiver side, and the performance can be rather close to a fully connected hybrid beamforming architecture. For further discussions about such a hybrid beamforming architecture, see also Section 2.3 in [38].

11.5.3 Digital Beamforming with Finite DACs

Hybrid beamforming is currently the most promising approach to tackle the power consumption and complexity bottleneck mmWave transceivers are facing. An alternative approach to address these aspects that is currently attracting increased attention is digital beamforming with low resolution DACs [39], as also discussed in detail in Section 16.2.3.2. Reducing the precision of the converters enables to reduce the power consumption, which scales roughly exponentially with the number of resolution bits. This enables to have a large antenna array with many active elements at a reduced power consumption and cost. Despite the increased signal processing capabilities compared to analog beamforming, the nonlinearity introduced by the quantization leads to limited capacity at high SNR, and imposes certain challenges on the channel estimation and data detection [39]. Still, investigations show promising performance even with 1-bit DACs in multi-user MIMO DLs, as shown in Section 2.4 in [38].

11.5.4 Massive Multiple-Input Massive Multiple-Output

The hybrid beamforming architecture has good potential to be used for access in particular in mmWave bands, but it can also be used for wireless relaying and backhauling, see for instance Section 7.4. However, in rather static high-throughput wireless backhaul scenarios, massive and symmetric MIMO might be feasible. In such a case, arrays with hundreds of antenna elements at both the transmitter and the receiver sides can be used to multiplex hundreds of data streams in the spatial domain, as illustrated in Figure 11-11. In theory, such MMIMMO could deliver spectral

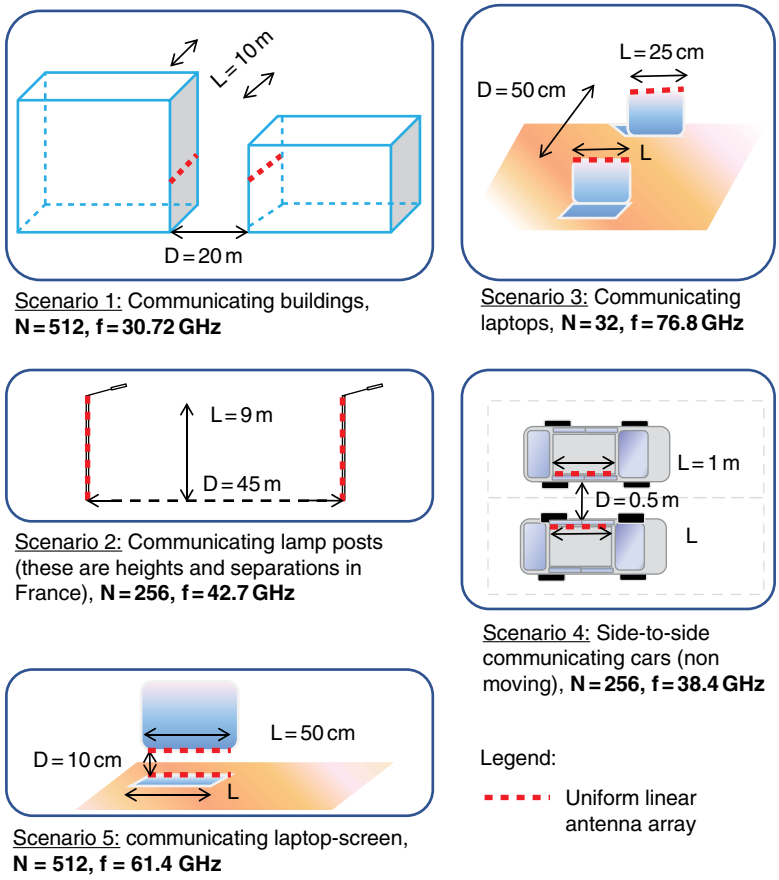

Figure 11-11. Practical examples of deployment scenarios of MMIMMO systems, see Section 3.4 in [9].

efficiencies of hundreds of bps/Hz, and therefore, provide multi-Gbps throughput, which are essential for the backhaul of future wireless communication systems.

So far, MMIMMO has been regarded as infeasible due to the complexity; thus, it has yet not been in focus for 3GPP NR. However, in [40] it has been recently shown that hundreds of data streams can be spatially multiplexed through a short range and line-of-sight MMIMMO propagation channel thanks to a new low-complexity spatial multiplexing scheme called block DFT based spatial multiplexing with maximum ratio transmission (B-DFT-SM-MRT). The block-based approach is beneficial to control that the spatial subchannels have similar properties, and maximum ratio transmission (MRT) is used to mitigate the effect of scattering and to deal with cases where the uniform linear arrays are not perfectly parallel. Its performance in real and existing environments was assessed using accurate ray-tracing tools and antenna models. In the best simulated scenario, depicted in Figure 11-12, 1.6 kbps/Hz of spectral efficiency is attained, corresponding to 80% of singular value decomposition (SVD) performance, with a transmitter and a receiver that are 200 and 10,000 times less complex, respectively.

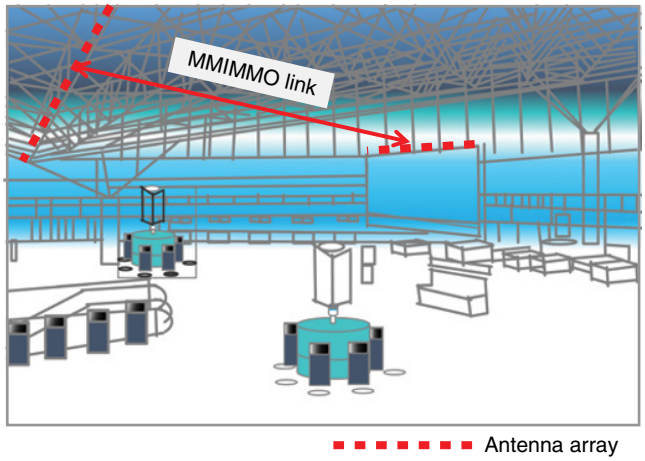

Figure 11-12. Helsinki airport simulated deployment scenario.

11.6 PHY/MAC Design for Multi-Service Support

As outlined in the introduction, 5G NR is to obey to various design criteria for making the system ready for the foreseen ecosystem of devices and services benefitting from having wireless access. This requires means to configure single connections according to the respective needs of the connected device or service. The overall system should be able to do so concurrently for any given combination of devices and services requesting access at a given point in time, enabling efficient *multi-service support*. Furthermore, its basic functionalities, such as access protocols and reference symbols, need to be able to scale with reasonable effort and resource consumption, following the respective network conditions, e.g., the number of devices requesting access, or the number of antenna ports or beams that are available or active. Beyond Release 15, 5G should be able to add new functionalities and use cases with low efforts and without requiring major redesign of the initial versions, commonly referred to as the notion of *forward compatibility*. Finally, its design should allow for tight interworking (or even coexistence) with other access technologies. Naturally, while obeying to all those design criteria, all design choices have to keep track of resource, energy and cost efficiency, and allow the system to be robust against, e.g., harsh channel conditions.

While any component of the communication system needs to care for the points given above, the design of the radio frames and related functionalities is pivotal for the overall communication system in general and the AI in particular in achieving the aforementioned targets. Before heading to specific parts of the frame design and the status of the discussions in 3GPP, some fundamental aspects are outlined.

11.6.1 Fundamental Frame Design Considerations

5G requires to support a reasonable set of options with respect to supported bandwidth to cover all relevant deployment scenarios and spectral bands being available, while keeping the overall number reasonably low. For each of these options, the respective sampling rate, number of subcarriers covering the bandwidth and the supported subcarrier spacing should be integer multiples or fractions of a given baseline, to keep system complexity and testing efforts at a reasonable level. Optimally, this baseline is

aligned with 4G (i.e., 15 kHz) to both allow for efficient multi-RAT implementations and to ease interworking and multi-connectivity among 4G and 5G at low and high bands. [3] is summarizing the status of the discussions in 3GPP following similar lines as given above: For below 6 GHz, the smallest bandwidth to be supported is 5 or 10 MHz, while in regions above 6 GHz and up to 52.6 GHz, the smallest bandwidth is 40 or 80 MHz. To avoid excessively large FFTs, the subcarrier scales with increasing carrier frequency. The baseline is 15 kHz, and further options are $15\text{kHz}*2^n$ (with n being an integer).

For schemes dealing with concurrent transmissions of multiple cells (e.g., inter-cell interference coordination), it is advantageous to have their transmissions to be time-aligned on symbol level. For energy reasons, the amount of always-on signal components should be minimized, and the actual repetition rate (e.g., for synchronization signals in DL) should be configurable. The on-demand principle should be applied as far as possible to increase energy efficiency especially in low-load scenarios. As an example, system information blocks (SIBs) shall be transmitted on-demand only, while master information blocks (MIBs) would always be transmitted.

Special care has to be taken for the new requirements being brought up by the new use cases. Hence, the overall frame design requires to obey to the following rules:

- Low-end devices, as typically envisioned for mMTC services, require to be able to detect the DL signal in a sub-sampled manner for energy efficiency reasons. This calls for applying narrow-band implementations in general (e.g., DL control channel) and eventually the introduction of complementary narrow-band signals (e.g., DL synchronization signals);
- To achieve very low latencies for URLLC services, very short TTIs and very quick scheduling processes have to be enabled. To avoid inefficient implementations, these options need to be introduced in a complimentary manner;
- Efficient support of various antenna configurations by implementing scalable RS designs is needed to increase the spatial reuse of spectral resources and for improving coverage, e.g., for eMBB services.

The frame requires various control channels (both for carrying the relevant global system configurations and for maintaining/configuring the various device connections and transmissions) and data channels (potentially different formats for different use cases). Additionally, one needs to account for DL synchronization symbols, UL sounding, UL random access opportunities, and RSs for various means, such as beam selection and channel estimation. Both TDD and FDD configurations have to be accounted for, and different areas of the spectrum require at least partially varying treatments. Finally, to allow for energy-efficient devices, pipelined processing should be enabled. Figure 11-13 depicts the basic sub-frame configurations being foreseen taking those points into account.

In the figure, options (c), (d), and (e) depict bidirectional sub-frames containing both DL and UL transmissions. They are the basic building blocks to enable efficient TDD, while (a) and (b) are exclusively carrying either DL or UL transmissions. Control (including DMRS that are not depicted here) is separated from the data block for enabling pipelined processing at the device and is placed at the borders, i.e., *frontloaded* in the case of the DL, and at the end of the sub-frame for the UL.

Having laid out the fundamental design paradigms to follow, the next sections provide some further details to the single building blocks of the signaling frames of 5G NR. First, the required frame elements for enabling initial access are introduced, followed by a discussion of the design of the control channels and the different data channels variants that are foreseen. The section is then completed by covering further access variants beyond unicast, namely, device-to-device (D2D) and broadcast/multicast.

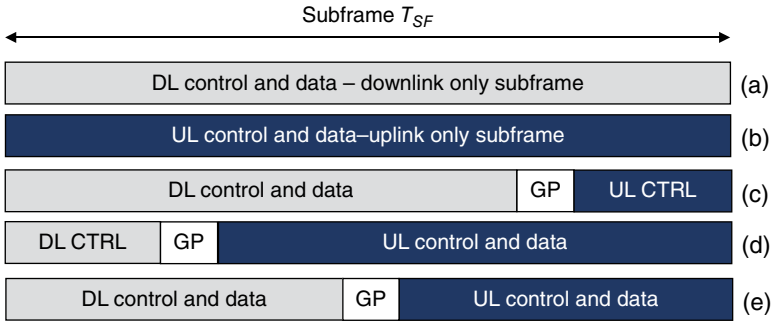

Figure 11-13. Sub-frame design variants. GP: Guard Period.

11.6.2 Initial Access

This section is closely related to Section 13.2. Before being able to transmit or receive data, a device needs to perform various steps when powered up. It has to:

1) identify close-by cells or transmission points (cell discovery);
2) align its transmission parameters (time and frequency) to the reference;
3) read the system configuration; and
4) perform the registration process, i.e., perform the random access procedure and message exchange for setting up the device configuration, e.g., related to authentication and encryption.

The BS has to regularly transmit synchronization signals for the device to be able to perform steps 1 and 2. Similarly as with 4G, 5G DL frames will regularly carry primary synchronization signals (PSS) and secondary synchronization signals (SSS). PSS and SSS are spanning 12 physical resource blocks (PRBs), each spanning 12 subcarriers, within a single multi-carrier symbol carrying selected sequences[4] being used to identify the respective transmission node. One of the key targets for the design is to enhance one-shot detection by improving the respective sequence characteristics (e.g., by applying a longer sequence and by avoiding time/frequency ambiguities). With achieving this target, the periodicity of the synchronization signals can be reduced without significantly increasing latency. By correlating the received signals with the known sequences, the device is able to perform DL synchronization (both on frame and symbol level) and to identify the transmission point. Once this has been done, the device is able to locate and decode the Physical Broadcast CHannel (PBCH). This channel carries the relevant system configuration (e.g., location and frequency of the Physical Random Access CHannel, PRACH) for the device to be able to continue the access procedure, and it spans 24 PRBs in frequency direction over two successive symbols in time. At this point, the device is aware of the available network and how to access it. The next step is concerned with setting up the physical (involving, e.g., power

4 The actual sequence design is defined in [41].

control) and logical connection (involving, e.g., authentication and encryption). This includes various measurements and message exchanges. In Chapter 13, more details are given.

The frequency of transmitting the set of signals given above (i.e., PSS, SSS, PBCH in DL, PRACH in UL) has an impact on the control plane latency (i.e., the time it takes to register to the network) and the signaling overhead (i.e., the more often those signals are transmitted, the more spectral resources are required).

It is foreseen to add high-capacity transmission points operating at mmWave bands, both in a standalone and a non-standalone manner. For the latter, a close-by transmission point from the coverage layer (<6GHz) acts as supporting node for the steps given above, e.g., for all actions related to the messaging. The cell discovery and the alignment of its transmission parameters though, must still be done directly with the mmWave transmission points for both deployment cases. As discussed in Section 11.5, mmWave requires directional and beam-based transmission to overcome the large free-space loss. For cell discovery, the broadcast of synchronization and control signals via beam-based transmissions is a challenge. In [4], different beam sweeping strategies, including time division, frequency division, code division, and spatial division, are compared systematically with respect to cell discovery latency and signaling overhead. It was found out that time division achieves the lowest latency at the price of high signaling overhead, while spatial division allows much lower signaling overhead and provides flexibility to achieve a trade-off between latency and signaling overhead. An auxiliary transceiver based scheme has been proposed to further reduce the signaling overhead and avoid interruption of data transmission (due to hybrid transceiver constraints) during the broadcast of synchronization and control signals.

11.6.3 Control Channel Design

For controlling the system in general and the transmission of data in particular, various physical control message exchanges between the network and the connected devices are required. Some messages carry system-wide settings and are of relevance for all devices, while some messages characterize the connection setup between a specific device (or a sub-group of devices) and the BS and thus are of less relevance for other devices not being part of this sub-group. The former needs to follow similar means as, e.g., applied in 4G for the PBCH design. As any device needs to be aware of the system configuration, the respective transmission needs to be of broadcast type and thus needs to be configured having the weakest possible link in mind to achieve a given minimal reliability. Details related to the PBCH design have already been given in the prior section.

5G will be a packet-switched network, as 3G and 4G have been. Hence, one needs to employ structures for controlling the flow of packets from and to different sources and recipients[5]. In DL, for example, the Physical Downlink Control CHannel (PDCCH) is carrying the instructions (e.g., scheduling grants, resource configurations, and HARQ feedback) from the BS to the connected devices. The first releases of 4G have physically structured the PDCCH in a broadcast manner: The control messages, also known as DL control information elements (DCIs), are multiplexed into a single structure and mapped to the beginning of each TTI, meaning that data and control are

5 A more comprehensive treatment of dynamic scheduling can be found in Chapter 12.

multiplexed in a time division multiple access (TDMA) manner. In Release 11, 3GPP has introduced an enhanced PDCCH (EPDCCH) allowing to dedicate single PRBs for the transmission of control messages in an frequency domain multiple access (FDMA) manner.

Discussions in 3GPP indicate that NR will follow a different structure [3]. Instead of separating control and the respective data transmission, so-called *in-resource messaging* is implemented, also referred to as "staying in the box". The main reasons for doing so are:

- "No race to the bottom" (i.e., no need to configure the resources for the weakest possible link);
- Data and control can share reference symbols;
- Control can make use of rank 1 precoding if respective CSI is available;
- Blanking of frequency resources is improved (e.g., for inter-cell interference coordination);
- More degrees of freedom are available for designing the DCIs.

For the actual realisation of the control channels, 3GPP has defined so-called resource element groups (REGs) as the basic control channel building block. These comprise 12 consecutive resource elements (REs) within a single OFDM symbol. Moreover, like in 4G, so-called control channel entities (CCEs), consisting of a number of REGs, are defined as the smallest unit of a scheduled PDCCH transmission. Both localized and distributed variants are foreseen, the latter being able to exploit frequency diversity. As such, each NR-PDCCH is transmitted by using one or several CCEs depending on the respective channel quality. The number of CCEs employed is called aggregation level (AL). Currently, similar to LTE, several ALs, namely 1, 2, 4, 8 or even 16 and 32 are considered. The higher the required coding gain is, the higher the aggregation level has to be chosen.

The transmission of physical control messages is typically required to be very robust to avoid packet loss. The control channel coverage in 5G NR is intended to be at least as good as in 4G. The main building block used to match the transmission to the respective link quality is the already mentioned aggregation level. The higher the aggregation level, the more copies of the control message are transmitted and thus the overall transmission becomes more robust. In addition, if the BS is aware of the channel state, rank 1 precoding may be employed to improve the link quality for the transmission of device specific DCIs. Otherwise, one needs to make use of diversity mechanisms. Two variants, namely space-frequency block codes (SFBC) and per-RE precoder cycling, have been extensively discussed as potential options. Finally, frequency hopping may be applied for the sake of frequency diversity. As coding scheme, PCs are selected, as covered in Section 11.4.2.3.

To enable channel estimation for coherent demodulation, some REs in the REG are dedicated for carrying DMRS. It is envisioned that REGs with configurable DMRS patterns can provide additional trade-offs between channel estimation performance and achieved coding rate for the control channel transmission.

11.6.4 Data Channel Design

While the previous section has covered control plane related aspects, the focus shall now be on the relevant PHY/MAC design choices for the user plane or data channels. The support of a multitude of different use cases is much more emphasized in 5G NR than in earlier generations. To allow the system to be efficient, various means have to be provided for devices to transmit and receive the data.

Most likely, the bulk of the connections accessing the system will still exhibit the following characteristics:

- The overall amount of data to be transmitted is far bigger than a single transmission opportunity is able to accommodate (i.e., a single transmission request requires several transmission opportunities);
- The energy consumption related to the transmission and reception of the data is small compared to other parts of the device (e.g., display);
- The system is able to regularly collect rather concrete context information, such as channel quality.

The listed aspects are mostly, but not exclusively, connected to (e)MBB services. Beside this, 5G NR will include new types of services introducing a new set of relevant aspects either related to the needs of the respective service or the respective device(s):

- In the area of mMTC, some use cases imply the need for transmitting tiny amounts of data (requiring only few transmission opportunities) in a rather sporadic and unpredictable manner. The related devices are typically constrained with respect to cost and energy;
- In the area of URLLC, the allowed transmission latency is required to be very short and the reception of the packet has to be extremely reliable. Both sporadic and regular/periodic transmissions are possible.

Obviously, one needs to make use of various tailored access mechanisms for being able to meet those partly contradicting targets. While these are presented later in this section, initially an overview is given on the most recent progress in 3GPP on the fundamental scheduling concepts and how they are being optimized for multi-numerology operation, followed by a summary of the open issues.

PHY layer processing is not aware of abstract concepts like service-categorization, e.g., URLLC or eMBB. Instead, the differentiation is done via selecting different DCI depending on the current transmission. While initially the research and standards community was heading towards reserving specific types of PHY resources exclusively for some specific services (e.g., so-called mini-slots for URLLC traffic), the prevailing trend in 3GPP now goes towards enhancements of PHY control signaling to indicate the length of a data assignment making the overall system more flexible.

In LTE, if a UE is not in discontinuous reception (DRX) mode, it needs to continuously monitor PDCCH meaning every 1 ms. A potential game-changer on this topic introduced by 3GPP for 5G NR is to allow for more differentiated options. 3GPP has, for instance, agreed that the minimum PDCCH monitoring period may go down to a single symbol, and the consecutive data duration is to be indicated. All that matters is the PDCCH monitoring period itself, especially since data can be scheduled over any number of successive symbols indicated by a respective DCI.

The scheduling delay represents a significant portion of the overall delay occurring in the radio network. In fact, reducing UL scheduling delay has been singled out in 3GPP as a means for achieving the latency required for URLLC applications and for future-proofing the system design. Ideally, URLLC transmissions should get an UL grant right away, i.e., for the very first PUSCH transmission, allocating the service resources with appropriate size and physical layer numerology according to underlying QoS requirements of the data buffered in the UE.

To alleviate the above issues, there are in principle three possible approaches worth exploring:

- Contention-based/grant-free UL data transmission, i.e. without prior scheduling request;
- Enhancements to scheduling requests and buffer status reporting (BSR) mechanisms;
- Semi-persistent scheduling (SPS).

As will be shown, the above are not mutually exclusive. In fact, support for all three is possible and will enable the reduction of the UL data scheduling latency, although optimizing for all three approaches simultaneously within a single system design may not be possible, meaning that compromises will be needed.

11.6.4.1 Contention-based/Grant-free Access

3GPP has agreed to support grant-free, SPS-like, PUSCH transmissions [3]. More specifically, an UL transmission scheme without a grant shall be supported.

Contention-based access in previous generations of cellular systems has been exclusively used as the basis for the initial connection of a device to the network, i.e., where a device would switch from RRC Idle state to RRC Connected, as detailed in Section 13.3. With the introduction of mMTC and URRLC services in 5G, the potential use of this type of access has been extended; namely, by the possibility of a device connecting to the network with minimal signaling overhead and latency. This is especially useful for sporadic small packet transmissions.

Contention-based access protocols are of three types, as depicted in Figure 11-14: (a) multi-stage, (b) two-stage, and (c) one-stage. These can be interpreted very differently, and each can contain several variants.

The one-stage access protocol (c) means that both the access notification and data delivery are done in a single transaction, i.e., using one or several consecutive packets in a single transmission. A two-stage access protocol (b) allows the UE to separate the access notification stage from its data delivery stage, e.g., by allowing for an intermediate feedback message. A multi-stage access protocol (a), for which the current LTE connection establishment protocol is a prime example, is composed of at least three phases, namely, the access, connection establishment phase (including authentication and security), and finally the data phase.

In [42], several proposals have been put forward that realize two-stage and one-stage accesses. In particular, the signaling associated with the connection establishment (i.e., mostly the establishment of mutual authentication and security) is assumed to be reused from a previous session, where the

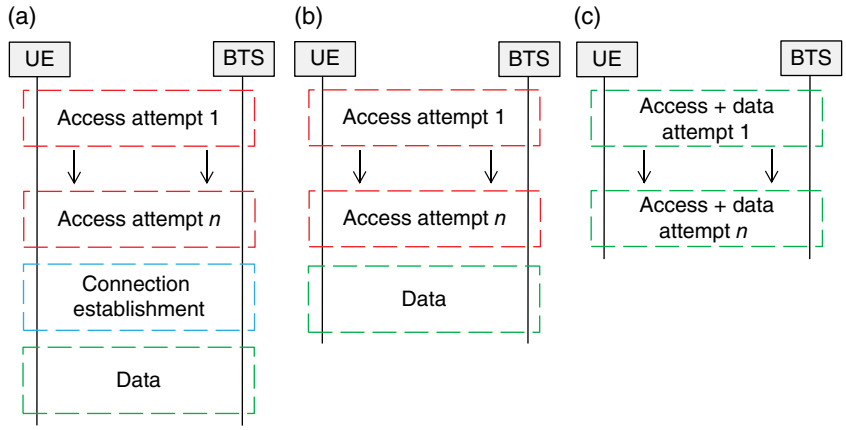

Figure 11-14. High-level description of the three access protocols types considered: (a) Multi-stage access protocol with an access, connection establishment and data phase; (b) Two-stages access protocol with access and data phases; and (c) One-stage access with combined access and data phase.

multi-stage access protocol has been carried out. This reuse of the connection context is achieved through the RRC Extant state introduced in [42]; a new additional state complementing RRC Idle and RRC Connected, as detailed further in Section 13.3. 3GPP has recently applied a similar state named RRC Inactive [43].

In both the one-stage and the two-stage access schemes, contention may occur only in the initial step of the protocol. Designing this type of access is a joint effort between the PHY design (to improve detection and to cope with collisions via advanced signaling processing) and the higher layers (to keep the number of arrivals to the network within the capabilities of the PHY).

A tailored design of the access scheme for mMTC and URLLC services is needed due to the short packet length and the need for supporting a massive number of simultaneously active devices attempting to access the network.

The main PHY functionality requirements for this can be summarized as follows:

- DL synchronization at the device;
- Signature/preamble generation at the device;
- Signature/preamble detection, e.g., via correlation (or compressive sensing) at the BS;
- Repetition and retransmission scheme, eventually based on frequency hopping;
- (Non-orthogonal) multiple access, e.g., through spatial separation;
- Autonomous link adaptation and power control.

The PHY preamble design uses false-alarm probability and missed detection as key metrics. The LTE PRACH is based on Zadoff-Chu sequences [44]. 5G has the following options to adjust the original design of LTE to its needs:

- A higher number of cyclic shifts in at least a part of the PRACH allocations, though reducing the distance between the preambles;
- Simultaneous usage of different root sequences. However, cyclic shifts generated from different root sequences are not orthogonal; hence, residual cross-correlations may increase false alarm rates.
- Usage of m-sequences instead of Zadoff-Chu sequences for higher PRACH capacity, though at the price of a higher PAPR;
- Orthogonal CDMA codes providing much higher capacity, though being more sensitive to weak synchronization;
- Reshaping the arrival distribution of access attempts. Ideally, the access attempts are equally distributed over time, which may be attained by dividing the PRACH opportunities into N slots, and each UE accesses the slot number derived from a modulo N operation of its ID. This solution is limited to delay tolerant use cases and devices with relaxed energy consumption requirement.

Obviously, adopting one or more of the above listed options alleviates the support of higher loads in massive access scenarios compared to the LTE preamble design, as long as the respective downsides can be tolerated.

11.6.4.2 Enhancements to the Design of Scheduling Requests and Buffer State Reporting

In 4G, a SR is used to inform the network that the respective UE has data to transmit, but it does not have enough resources available to transmit the BSR itself, which carries information on buffer status. The buffer status is crucial for the BS scheduler and is typically unknown until the BSR is received. Most SR enhancements proposed recently focus on reducing the delay between initial SR and the first UL data transmission.

Assume a device requiring eMBB data to be transmitted right after having terminated an URLLC session. Without having differentiated SRs, the BS would not be aware of the altered connection requirements, for instance related to latency and reliability, and would thus not be able set up the subsequent data connection as required. This leads either to a connection not meeting the requirements (e.g., if these are more stringent than those of the prior transmission), or the other way round, i.e., a connection overshooting the actual requirements, resulting in inefficient resource usage. Hence, to avoid this, the use of a different SR settings or configurations are required. With allowing for differentiated SRs, the BS is able to allocate the appropriate amount of UL resources, perform link adaptation matched to the use case, and initiate the respective access method. In essence, SR and BSR functionalities are somewhat merged to reduce the latency. The basic idea is to include more details about the BSR into the SR, where some proposals target only to indicate whether the BSR is long or short, while others go further and include even more details, such as, the type of the service the data is originating from.

11.6.4.3 Semi-persistent Scheduling and Grant-free Scheduling

SPS is used in LTE as a scheduling technique with minimal overhead being suitable for traffic with periodic characteristics. SPS is configured (but not activated) via RRC messages with signaling its periodicity. The SPS is then activated via PDCCH-messaging, enabling to re-tune parameters on a faster basis and with less control signaling overhead. More specifically, a UE being configured for SPS-like transmissions waits for a DCI scrambled in relation to a special kind of cell radio network temporary identifier (SPS-C-RNTI), and once received, the UE starts transmitting data with a pre-configured periodicity, as set via RRC signaling.

As with any dedicated resource scheduling scheme, SPS has some inherent inefficiencies. Furthermore, the empty transmissions as needed for implicit release are raising further concerns. Some enhancements to SPS to support 5G are centred around the following items:

- **Configuration/activation split**: Which parameters are configured via RRC signalling, and which are signalled via PDCCH?
- **Sharing of SPS resources**: Should it be allowed, and what sort of collision resolution would need to be introduced?
- **SPS periodicity**: Is support for extreme (i.e., sub-ms range) values needed and justified, in the light of potential usage for URLLC?

As discussed in Section 11.6.4.1, 3GPP is additionally working on standardizing grant-free/contention-based transmissions. It has recently been agreed that UL data transmission without UL grant can be configured by the network to be carried out after semi-static resource configuration in RRC without PHY signaling [3]. If the network configures the system accordingly though, PHY signaling for the activation and deactivation and/or modification of parameters for UL data transmission without UL grant can be applied. Work is ongoing to find a harmonized MAC design for UL SPS and grant-free access based on configured PHY signaling.

11.6.4.4 Pre-emptive Scheduling

As discussed throughout this chapter, 5G will need to support shorter TTI lengths in order to enable lower latencies. On the other side, as long as the latency requirement is not very strict, as for instance for eMBB services, and the traffic volumes to be transmitted are not small, longer TTI durations typically result in higher performance due to lower control overhead. So it is neither reasonable to fix

transmission times to be very short (below 1 ms), nor to be long in any case and allow for specific selection instead. Naturally, when allowing transmissions with varying time bases to access the system, special treatments have to be accounted for as outlined next. Let us consider mixed traffic scenarios with both eMBB and URLLC traffic being present. For the DL, in case the BS has already scheduled and indicated eMBB traffic with, e.g., 1 ms TTI duration or longer, and an urgent URLLC packet arrives without respective resources being available, parts of the eMBB DL traffic may be punctured, and the URLLC data symbols inserted instead. Equivalently, in UL the URLLC packet may be allowed to be superimposed to the running eMBB transmission. This solution efficiently embeds URLLC into the frame with avoiding excessive resource reservation [42].

The disadvantage of applying puncturing/superposition as described above, is the negative impact on the eMBB transmission. If the eMBB victim is not aware of the puncturing or superposition, the inserted URLLC symbols will degrade the symbol detection performance significantly. Hence, a first option to lower this effect, is to inform the victim device via control signaling about the particular part of its transmission being punctured. Further solutions for improvement are to use: (i) code-block based HARQ re-transmissions (where only the punctured part is retransmitted) or even to (ii) retransmit the punctured code block immediately in the next scheduling opportunity (i.e., without NACK feedback). For further details, the interested reader is referred to Section 12.3.3.

11.6.4.5 Device-to-Device and Broadcast/Multicast

So far, we have handled unicast transmissions between device and network. For specific use cases, though, it is more efficient to introduce novel access paradigms, related to device-to-device (D2D) and broadcast/multicast (BMS) transmission. The former allows devices to directly communicate without the network being part of the data exchange, though, for controlling the connections, the network may still act in a supporting manner. When payloads are of relevance for a group of devices, BMS is typically more efficient than relying on multiple unicast connections. In the following, we provide some high-level viewpoints. For a deep-dive, the interested reader is referred to the wide range of available publications, e.g., given in [42].

D2D connections can be used for various means:

- Network offloading with the help of content caching within dedicated devices, e.g., [45];
- Coverage extension with the help of data relaying;
- Direct interactions between the respective devices, e.g., between different road users.

The actual transmission can be 'underlaid' or 'overlaid', i.e., either the D2D connections have dedicated resources available, or they use superposition. Before being able to perform the data transfer, though, the related devices need to be aware of the available links and their respective quality. For this, proximity discovery is performed. Various approaches have been discussed in the literature to perform this step. Recently, both strategies with network support [46] and without network support [47] have been analyzed. For the latter, the achievable mean discovery time (i.e., the time until all nodes have been made aware of all potential partners) has been investigated depending on the number of active nodes and with or without the application of FD, as covered in detail in Section 16.2.4. Dedicated discovery messages have been designed, and the transmission probability for each node has been optimized to minimize the mean discovery time. The use of FD during the discovery phase helps to further reduce the mean discovery time. The given references provide further details on specifics of the concepts and performance results. A more thorough treatment of D2D can be found in Chapter 14.

Although being part of early releases of 4G already, services relying on broadcast/multicast are not yet widely applied in 4G networks. Different applications can benefit from this kind of access such as software updates of sensor networks, or multimedia streams to a group of people, e.g., video feeds during concerts or sports events. Recent studies have investigated means to enhance broadcast/multicast transmission:

- Non-orthogonal transmission schemes for stream multiplexing, including both beam-based and multi-level coding based variants;
- The introduction of complementary unicast based feedback schemes.

The former approach allows to increase the spatial reuse by means of multi-antenna precoding in order to superimpose several multi-cast streams on top of a broadcast transmission. A possible usage scenario for such a technique is the transmission of area specific video feeds in a stadium (e.g., the camera feed capturing the area of the game field being far away from the respective multicast group) in addition to a broadcast transmission being of interest for all spectators. An alternative to this beamforming-based approach is to rely on multi-level coding, multiplexing different streams in code domain [42].

A common bottleneck of broadcast systems is the need to configure the transmission according to the weakest possible link. 5G naturally will allow for unicast UL transmissions, which may be used to improve the DL broad-/multicast connections in various ways, such as:

- transmitting channel quality indicators providing context that the system can use to allow for a more efficient initial broadcast transmission, and
- enabling the system to introduce a HARQ mechanism to follow-up the initial broadcast stream by subsequent unicast retransmissions.

For details and performance results, the interested reader is referred to [42].

11.7 Summary and Outlook

5G will more prominently than any of the earlier generations invite new and existing players to improve their products, systems and services by allowing those to obtain access to a wireless communication system. This wide range of new use cases and device classes accessing the network most often requires special treatment of those to ensure proper and efficient functioning. The air interface (AI) is one of the fundamental building blocks to address this. In particular, the AI should be designed in a way to provide more flexibility while avoiding being excessively complicated to build and test. The two possible design extremes are either having a single AI as a "one-fits-all solution", where fixed configurations support each and every single requirement at any point in time, or allowing for an excessive number of solutions tailored to each single family of use cases, where the service provider requires to operate a respective number of RATs concurrently. This chapter has explored various design options within these stated extremes, though ultimately centering on a single multi-service AI supporting some flexible adaptations to allow various kinds of services and device classes to be served according to their specific needs, and also the support of different transmission frequencies with very different transmission characteristics.

The chapter has in particular treated the wide topic of waveform design for 5G NR - a key technology defining many features of the overall AI - and compared a rich set of enhancements improving

conventional CP-OFDM in various aspects. Similarly, various coding options to be used in various settings have been identified, e.g., related to the size of the packets being transmitted, and efficient schemes to implement HARQ both for UL and DL have been discussed. While for bands below 6 GHz fully digital beamforming may be applied in the context of massive MIMO, systems operating at higher frequencies might have to rely on hybrid variants adding an analog component, as detailed in the chapter, along with the application of massive MIMO to an example scenario. Finally, various design principles for multiplexing the different service types within a single transmission band have been presented, both taking data and control signaling into account.

References

1 3GPP RP-170379, "Study on New Radio (NR) Access Technology", March 2017
2 3GPP RP-170855, "Work Item on New Radio (NR) Access Technology", March 2017
3 3GPP TR 38.802, "Study on new radio access technology physical layer aspects", V14.1.0, June 2017
4 5G PPP mmMAGIC project, Deliverable D4.2, "Final radio interface concepts for mm-wave mobile communications", June 2017
5 H. G. Feichtinger and T. Strohmer, "Gabor Analysis and Algorithms: Theory and Applications", Springer, 1998
6 5G PPP FANTASTIC-5G project, Deliverable D3.2, "Final results for the flexible 5G air interface link solution", May 2017
7 G. Berardinelli, F. M. L. Tavares, T. B. Sørensen, P. Mogensen, and K. Pajukoski, "On the potential of zero-tail DFT-spread-OFDM in 5G networks", IEEE Vehicular Technology Conference (VTC Fall 2014), Sept. 2014
8 D. Petrovic, W. Rave, and G. Fettweis, "Effects of phase noise on OFDM systems with and without PLL: characterization and compensation", IEEE Transactions on Communications, vol. 55, no. 8, pp. 1607–1616, Oct. 2007
9 5G PPP mmMAGIC project, Deliverable D5.1, "Initial multi-node and antenna transmitter and receiver architectures and schemes", Mar. 2016
10 S. C. Cripps, "Advanced Techniques in RF Power Amplifer Design", Artech House, 2002
11 5G PPP METIS II project, Deliverable D4.2, "Final air interface harmonization and user plane design", Apr. 2017
12 P. Siohan, C. Siclet, N. Lacaille, "Analysis and design of OFDM/OQAM systems based on filterbank theory", IEEE Transactions on Signal Processing, vol. 50, no. 5, pp 1170–1183, Aug. 2002
13 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies", V14.2.0, Dec. 2016
14 5G PPP FANTASTIC 5G project, Deliverable D3.1, "Preliminary Results for Multi-Service Support in Link Solution Adaptation", May 2016
15 R. Garzon-Bohorquez, C. Abdel Nour and C. Douillard, "Improving Turbo Codes for 5G with parity puncture-constrained interleavers", Int. Symp. on Turbo Codes and Iterative Information Processing (ISTC 2016), Sept. 2016
16 R. Garzon-Bohorquez, C. Abdel Nour and C. Douillard, "Protograph-Based Interleavers for Punctured Turbo Codes", IEEE Transactions on Communications, Dec. 2017
17 R. G. Gallager, "Low Density Parity-Check Codes", MIT Press, Cambridge, 1963

18 E Arıkan, "Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels", IEEE Transactions on Information Theory, vol. 55, no. 7, pp. 3051–3073, July 2009

19 C. Berrou and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding: Turbo-Codes", IEEE Transactions on Communications, vol. 44, no. 10, pp. 1261–1271, Oct. 1996

20 R. Garzon-Bohorquez, C. Abdel Nour and C. Douillard, "On the Equivalence of Interleavers for Turbo Codes", IEEE Wireless Communications Letters, vol. 4, no. 1, pp. 58–61, Feb. 2015

21 J. Vogt and A. Finger, "Improving the max-log-MAP turbo decoder", Electronics Letters, vol. 36, no. 23, pp. 1937–1939, Nov. 2000

22 D. J. C. MacKay and R. M. Neal, "Near Shannon Limit Performance of Low Density Parity Check Codes", Electronics Letters, vol. 32, no. 18, pp. 1645–1646, Aug. 1996, Reprinted Electronics Letters, vol. 33, no. 6, pp. 457–458, Mar. 1997

23 T. Richardson and R. Urbanke, "Multi-Edge Type LDPC Codes", 2004, see http://citeseerx.ist.psu.edu/index

24 T. J. Richardson, M. A. Shokrollahi and R. L. Urbanke "Design of Capacity-Approaching Irregular Low-Density Parity-Check Codes", IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 619–637, Feb. 2001

25 C. Jones, S. Dolinar, K. Andrews, D. Divsalar, Y. Zhang and W. Ryan, "Functions and Architectures for LDPC Decoding", IEEE Information Theory Workshop, Sept. 2007

26 I. Tal and A. Vardy, "List Decoding of Polar Codes", IEEE Transactions on Information Theory, vol. 61, no. 5, pp. 2213–2226, May 2015

27 B. Li, H. Shen and D. Tse, "An Adaptive Successive Cancellation List Decoder for Polar Codes with Cyclic Redundancy Check", IEEE Communications Letters, vol. 16, no. 12, pp. 2044–2047, Dec. 2012

28 3GPP RAN1 meeting #86bis, Nokia, "Chairman's notes of AI 8.1.3 on channel coding and modulation for NR", Oct. 2016

29 3GPP RAN1 meeting #86bis, AccelerComm, "Complementary turbo and LDPC codes for NR, motivated by a survey of over 100 ASICs", Oct. 2016

30 3GPP RAN1 meeting #87, AccelerComm, Ericsson, Orange, IMT, LG Electronics, NEC, "WF on channel codes for NR short block length eMBB data", Nov. 2016

31 3GPP RAN1 meeting #88, R1-1702856, AccelerComm, "Enhanced turbo codes for URLLC", Feb. 2017

32 L. Zheng and D. N. C. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels", IEEE Transactions on Information Theory, vol. 49, no. 5, pp. 1073–1096, May 2003

33 T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas", IEEE Transactions on Wireless Communications, vol. 9, no. 11, pp. 3590–3600, Nov. 2010

34 M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling", IEEE Transactions on Information Theory, vol. 50, no. 9, pp. 1893–1909, 2004

35 O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo and S. Z. Iliya, "A Comprehensive Survey of Pilot Contamination in Massive MIMO—5G System", IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 905–923, Q2 2016

36 A. Alkhateeb, O. Ayach, G. Leus and R. W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems", IEEE Journal on Selected Topics on Signal Processing, vol. 8, no. 5, pp.831–46, Oct. 2014

37 S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G", IEEE Communication Magazine, vol. 53, no. 1, pp. 186–194, Jan. 2015

38 5G PPP mmMAGIC project, Deliverable D5.2, "Final multi-node and multi-antenna transmitter and receiver architectures and schemes", June 2017

39 J. Mo and R. W. Heath, "Capacity Analysis of One-Bit Quantized MIMO Systems with Transmitter Channel State Information", IEEE Transactions on Signal Processing, vol. 63, no. 20, pp. 5498–5512, Oct. 2015

40 D.-T. Phan-Huy, P. Ratajczak, R. D'Errico, A. Clemente, J. Järveläinen, D. Kong, K. Haneda, B. Bulut, A. Karttunen, M. Beach, E. Mellios, M. Castaneda, M. Hunukumbure and T. Svensson, "Massive Multiple Input Massive Multiple Output for 5G Wireless Backhauling", IEEE Globecom'2017 ET5GB workshop, Singapore, Dec 2017

41 3GPP TR 38.211, "Physical channels and modulation", V1.0.0, (2017–09)

42 5G PPP FANTASTIC-5G project, Deliverable D4.2, "Final results for the flexible 5G air interface multi-node/multi-antenna solution", May 2017

43 3GPP TR 38.804, "Study on new radio access technology radio interface protocol aspects", V14.0.0, Mar. 2017

44 S. Sesia, I. Toufik and M. Baker, "LTE The UMTS Long Term Evolution: From Theory to Practice", John Wiley & Sons Ltd., 2009

45 A. Masucci, S. E. Elayoubi and B. Sayrac, "Flow level analysis of the offloading capacity of D2D communications", IEEE Wireless Communications and Networking Conference (WCNC 2016), Apr. 2016

46 N. K. Pratas and P. Popovski, "Network-Assisted Device-to-Device (D2D) Direct Proximity Discovery with Underlay Communication", IEEE Global Conference on Communications (GLOBECOM 2015), Dec. 2015

47 M. G. Sarret, G. Berardinelli, N. H. Mahmood, B. Soret and P. Mogensen, "Can full duplex reduce the discovery time in D2D communication?", International Symposium on Wireless Communication Systems (ISWCS 2016), Sept. 2016

12

Traffic Steering and Resource Management

Ömer Bulakçı[1], Klaus Pedersen[2], David Gutierrez Estevez[3], Athul Prasad[4], Fernando Sanchez Moya[5], Jan Christoffersson[6], Yang Yang[7], Emmanouil Pateromichelakis[1], Paul Arnold[8], Tommy Svensson[9], Tao Chen[10], Honglei Miao[7], Martin Kurras[11], Samer Bazzi[1], Stavroula Vassaki[12], Evangelos Kosmatos[12], Kwang Taik Kim[13], Giorgio Calochira[14], Jakob Belschner[8], Sergio Barberis[14] and Taylan Şahin[1,15]

[1] *Huawei German Research Center, Germany*
[2] *Nokia Bell Labs, Denmark*
[3] *Samsung Electronics R&D Institute, UK*
[4] *Nokia Bell Labs, Finland*
[5] *Nokia, Poland*
[6] *Ericsson, Sweden*
[7] *Intel, Germany*
[8] *Deutsche Telekom, Germany*
[9] *Chalmers University of Technology, Sweden*
[10] *VTT, Finland*
[11] *Fraunhofer Heinrich Hertz Institute, Germany*
[12] *WINGS ICT Solutions, Greece*
[13] *Samsung Electronics, Republic of Korea*
[14] *Telecom Italia, Italy*
[15] *Technische Universität Berlin, Germany*
With contributions from Chao Fang and Behrooz Makki.

12.1 Motivation and Role of Resource Management in 5G

One of the main differentiating factors of the 5th generation (5G) mobile and wireless network compared to previous generations is the support of a wide range of services associated with a diverse set of requirements, which are typically grouped under the main service types enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications (mMTC), as highlighted in Section 2.2. 5G also aims at enabling new business opportunities by addressing the requirements of vertical industries. Accordingly, the resulting 5G system shall be flexible to cope with such diversity in an efficient way. On this basis, resource management (RM) plays a critical role to fulfil the service and slice requirements, such as quality of service (QoS) requirements, while efficiently mapping service flows to appropriate resources in order to optimally adapt to the current traffic and dynamic radio-environment conditions. To this

5G System Design: Architectural and Functional Considerations and Long Term Research, First Edition.
Edited by Patrick Marsch, Ömer Bulakçı, Olav Queseth and Mauro Boldi.

end, the resource landscape is extended beyond conventional radio resource management (RRM). In addition to licensed radio bands, the extended realm of resources includes the native use of unlicensed bands, which shall be adaptive and coupled with the changing radio topology, energy, hardware and software resources, such as computational and storage resources, as well as backhaul (BH) and fronthaul (FH) resources.

A fast resource allocation can typically be performed on a radio frame unit basis (e.g., a radio subframe of one millisecond), which is controlled by a scheduler. The scheduler can allocate time and frequency resources considering, e.g., the link qualities and transmission power constraints. A much slower resource allocation can be achieved through balancing the load among different cells and radio access technologies (RATs), which is handled by traffic steering mechanisms and hard handovers with execution times spanning several hundred milliseconds. Nevertheless, the 5G system will include various paradigm changes that will re-shape the RM methods towards more agile solutions.

As also highlighted in Chapter 11, it is envisioned that a flexible frame structure will be employed in 5G, such that physical layer (PHY) numerologies are optimized for specific frequency bands, such as sub-6 GHz, millimeter-wave (mmWave), and for one or more target use cases [1] [2] [3]. Thus, the developed RM schemes need to be built upon this additional degree of flexibility, where it is required to introduce new functionalities, such as dynamic multi-service scheduling (see Section 12.3), considering QoS requirements and classifications (see Section 12.2).

In the 5G era, a tight interworking between novel 5G air interfaces (AIs) and enhanced Long-Term Evolution (eLTE) is targeted, where the integration is done on the radio access network (RAN) level, as detailed in Section 6.5. Along with forms of multi-connectivity where a user equipment (UE) connects to multiple access nodes at the same time, and a new QoS architecture, as introduced in Section 5.3.3 and detailed in Section 12.2, functions that are traditionally slow are envisioned to be operated on a faster time scale. This enables a fast routing of service flows to appropriate access nodes and AIs, as presented in Section 12.4 under dynamic traffic steering. Furthermore, a proactive analysis can be performed to determine new mmWave links to be utilized by traffic steering.

The application of full frequency reuse is an effective way of increasing the network capacity considering the scarcity of available spectrum. This, however, results in inter-cell interference (ICI) which needs to be mitigated to ensure the targeted high capacity and wide coverage. In addition, novel communication modes and new deployment options foreseen in 5G lead to new interference challenges to be overcome. Therefore, interference mitigation schemes, as presented in Section 12.5, are essential and should take into account different deployment options, such as fixed and dynamic radio topologies, flexible duplexing schemes, such as dynamic time division duplex (TDD), and high frequency band operation, while exploiting new interference-resistive designs.

Network slicing is seen as a key enabler for new 5G businesses (e.g., related to vertical industries) that require one or more services with associated service-level agreements (SLAs), as detailed in Chapter 8. In particular, the scope of QoS fulfilment is enhanced toward SLA fulfillment, where a real-time SLA monitoring is crucial to avoid SLA violations. Multi-slice RM in Section 12.6 introduces schemes that are needed to respond to multiple SLAs, where different slices share the same RAN infrastructure.

As detailed in Section 15.3, energy efficiency is one of the design goals of 5G to attain a sustainable system. Hence, active-mode operation needs to be optimized in the 5G network with the help of QoS and channel quality awareness, as presented in Section 12.7. The energy saving gains of the moderated network are obtained due to the unique design of 5G which enables access nodes to be in sleep mode longer in case of no traffic.

Eventually, functional extensions and changes in the device measurement context are needed to enable the aforementioned new functionalities tailored to different 5G use cases, while considering device performance, as covered in Section 12.8. Finally, the chapter is summarized in Section 12.9.

12.2 Service Classification: A First Step Towards Efficient RM

The envisioned 5G services have different requirements in terms of QoS, both regarding delay and throughput as well as reliability and availability aspects. The simultaneous provisioning of these services using a common infrastructure is an issue that should be addressed by the 5G network to have a functional and efficient operation. Toward this direction, in this section, we present briefly the QoS framework for 5G networks, focusing on service classification mechanisms.

12.2.1 QoS Mechanisms in 5G Networks

A first step to be able to allocate efficiently the network resources to the heterogeneous services is the accurate identification of the service types and the corresponding requirements. This knowledge could be provided by the higher layers as in High Speed Packet Access (HSPA) and LTE, where sets of QoS parameters are available for RRM functionalities, such as admission control and packet scheduling decisions. A similar approach to the one followed in LTE, where different bearers are set up within the Evolved Packet System (EPS) to support multiple QoS requirements, also holds for 5G networks. As introduced in Section 5.3.3 and described in detail in [4] [5], the 5G QoS model supports a QoS flow based framework where a QoS flow ID (QFI) is used to identify a QoS flow in the 5G system. The QFI is carried in an encapsulation header without any changes to the end-to-end (E2E) packet header. It can be applied to Protocol Data Units (PDUs) with different types of payload, i.e., Internet Protocol (IP) packets, non-IP PDUs, and Ethernet frames. The QFI will be unique within a PDU session. In the Next-Generation Radio Access Network (NG-RAN), the data radio bearer (DRB) defines the packet treatment on the radio interface (Uu). Particularly, in the downlink (DL), the NG-RAN maps QoS Flows to DRBs based on NG-U (or N3 interface) marking (QFI) and the associated QoS profiles. In the uplink (UL), the UE marks UL packets with the QFI for the purposes of marking forwarded packets to the core network (CN). Specifically, in the UL, the NG-RAN may control the mapping of QoS Flows to DRB in two different ways: Either using reflective mapping, where the UE monitors the QoS flow ID(s) of the DL packets and applies the same mapping in the UL, or using explicit configuration where the NG-RAN may configure by Radio Resource Control (RRC) an UL "QoS Flow to DRB mapping". However, sometimes an incoming UL packet matches neither an RRC configured nor a reflective QoS Flow ID to DRB mapping. In this case, the UE will map that packet to the default DRB of the PDU session, even though this may not correspond to the adequate QoS requirements of the flow. To address this issue, the use of novel service classification techniques should be considered, in which the traffic flows are monitored to extract more detailed service classification information and to identify the service type providing this information as input to RRM functionalities.

Toward this direction, this section focuses on service classification techniques based on machine learning (ML). The proposed classification methods reside in the area of statistical-based classification techniques and are implemented exploiting several flow-level measurements (e.g., traffic volume, packet length, inter-packet arrival time, and so forth) to characterize the traffic of different services.

Other methods of traffic classification, like payload-based classification, need to analyze the packet payload or use deep packet inspection technologies. On the contrary, statistical-based classification techniques are usually very lightweight, as they do not need access to packet payload and can also leverage information from flow-level monitors. It should be noted that these techniques may be used both as independent mechanisms as well as additional mechanisms that will support the already established techniques in order to effectively assign flows to the appropriate bearers and, therefore, fulfill the QoS requirements of the flows. The goal is to further increase the effectiveness of packet scheduling algorithms and other RRM functionalities by assigning the appropriate QoS to each service type or even to different flows of the same service type.

12.2.2 A Survey of Traffic Classification Mechanisms

During the last years, several studies that focus on application and service discrimination based on traffic classification learning techniques have been proposed. Both supervised as well as unsupervised ML mechanisms have been considered [6], as also depicted in Figure 12-1. Regarding the unsupervised techniques, principal component analysis (PCA) based mechanisms [7] and clustering algorithms like K-Means, density-based spatial clustering of applications with noise (DBSCAN) and Autoclass [6] have been investigated. These mechanisms group flows that have similar patterns into a set of disjoint clusters. Their major advantage is that they automatically discover the classes via the identification of specific patterns in the dataset without requiring a training phase like the supervised mechanisms. However, the resulting clusters do not necessarily map 1:1 to services and, even in this case, the clusters still need to be labeled in order to be mapped to the corresponding services.

As far as the supervised ML techniques are concerned, there are various classification schemes that have been proposed for the traffic classification problem like Naïve Bayes, decision trees, random forests and others [8]. A main approach focuses on Bayesian classification techniques [9] in which flow parameters are used to train the Naïve Bayes classifier and create a group of services.

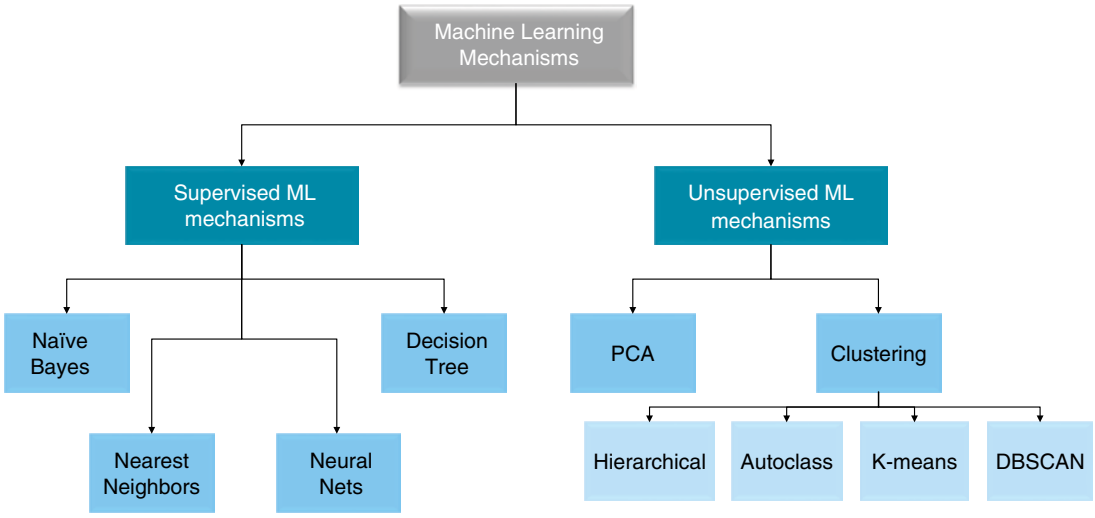

Figure 12-1. Overview of service classification techniques.

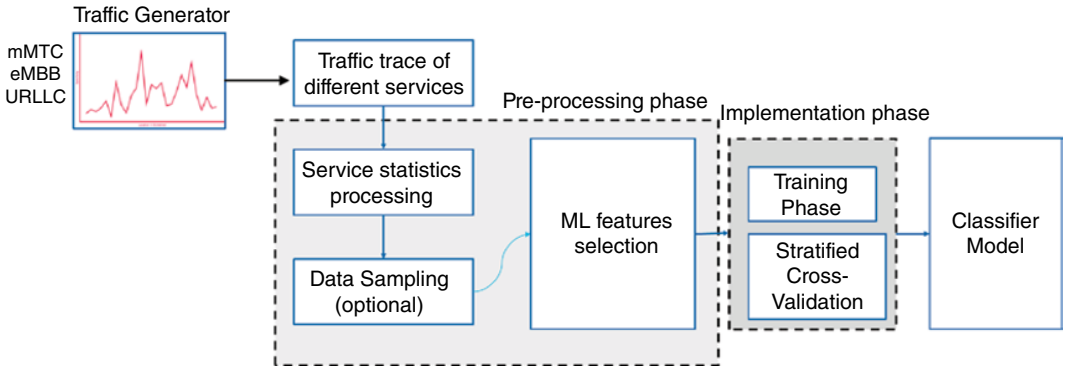

Figure 12-2. Example mechanism for the service classification process.

Then, when new flows arrive, they are assigned to the class for which the class membership probability is maximized. Also, decision tree algorithms [10] represent a completely orthogonal approach to the classification problem, using a tree structure to map the observation input to a classification outcome.

Finally, another concept, which has also been employed for the traffic classification process, is the concept of artificial neural networks (ANNs) that consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs, inspired by the way biological nervous systems works. Based on this approach, a multilayer perception classification network can be used for assigning probabilities to flows [11]. A set of flow features can be used as input to the first layer of the network, while the output classifies flows into a set of traffic classes by calculating the probability density function of class membership.

12.2.3 ML-based Service Classification Approach

In this section, the problem of service classification is investigated through the example of supervised ML mechanisms. The different steps included in the classification mechanism are presented in Figure 12-2. The first step of the process refers to the collection of traces from different services, whereas the next step focuses on the statistical processing of these traces to separate them in flows. It should be noted that a flow is considered a series of packet transmissions that have the same source and destination, and for which the inter-arrival time is below a specific threshold.

After this processing, a set of features is generated for each flow, including inter-arrival time statistics, packet size statistics and other flow characteristics like the total number of packets, source, destination as well as flow direction. Subsequently, some feature engineering tasks (e.g., data imputation tasks and data cleaning) are performed, completing the pre-processing steps. These tasks include the selection of the most representative features, the transformation of categorical features into numerical values, the normalization of features' values, and other tasks, such as the replacement of missing values, in order to guarantee high data quality. Finally, the implementation of the ML mechanism follows, including two main phases, namely a training phase and a cross-validation phase. In this case, stratification is applied in order to randomly sample the flows' dataset in such a way that each service type is properly represented in both training and testing datasets.

To evaluate the performance of the classification mechanisms and select the most adequate mechanism, several metrics can be used in the train/test sets, like the so-called accuracy metric, the precision, the recall and the F1-score, all of which will now be explained. A very useful tool that illustrates the relationship between the different evaluation metrics and provides a holistic view of each algorithm's performance is the confusion matrix, where the horizontal axis represents the predicted class (i.e., the outcome of the algorithm), and the vertical axis represents the true class. The confusion matrix formulation includes information for the false positives (*FP*), true positives (*TP*), false negatives (*FN*) and true negatives (*TNeg*). In this context, *accuracy* is defined as the percentage of correct predictions to the total number of predictions, i.e. $(TP + TNeg)/(TP + FP + TNeg + FN)$, whereas *precision* represents the percentage of the instances that were correctly predicted as belonging in a class among all the instances that where classified as belonging in this class, as given by $TP/(TP + FP)$. *Recall* is defined as the percentage of the instances of a specific class that were correctly classified as belonging to this class, given by $TP/(TP + FN)$, and *F1-score* is defined as the harmonic mean of the precision and recall.

It should be highlighted that the investigation of a single metric, like accuracy, is not always enough to choose the best mechanism for a classification problem, as the misclassification of a specific class instance may have more significance than the correct classification of others. For this reason, other evaluation metrics have also to be applied to make the most appropriate choice depending on the problem's characteristics.

12.2.4 Numerical Evaluation of Service Classification Schemes

In this subsection, a numerical evaluation of the service classification schemes based on simulation results is provided. In particular, for the simulation scenario, the three main service types URLLC, eMBB and mMTC are considered. The different traces for each service have been generated using specific traffic models, where URLLC and mMTC traces are generated according to the models presented in 802.16p [12], and eMBB traffic is generated based on [13], assuming video streaming traffic (i.e., YouTube). For the classification mechanism, a splitting of 70%-30% for training and testing sets has been considered. For the training set, the label 'service type' of each flow is considered as known; whereas, for the testing set, this label is considered as unknown, and each flow is labeled using the classifier model. The outcome of the proposed mechanism is a classifier model that can be applied to unknown flows to recognize them in an accurate way.

In the simulation scenario, the performance of various ML mechanisms has been investigated including base classifiers as well as ensemble-based classifiers. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm to improve generalizability and robustness over a single classifier. To compare the different ML mechanisms, the accuracy metric of each algorithm is presented in Table 12-1, where a Dump classifier that classifies all the flows as type 0 (mMTC service) is also considered. From this table, it can be seen that Decision Tree and Random Forest algorithms lead to the highest accuracy values, outperforming the other ML algorithms.

To provide a more complete view of each classifier's performance, the corresponding confusion matrices are also illustrated in Figure 12-3. Considering that it is desired to eliminate the possibility that a URLLC service is misclassified as another service type, the optimal model should have high values of recall and high accuracy values for the case of mMTC and eMBB services. The confusion matrix shows that the Decision Tree and the Random Forest algorithms result in extremely good

Table 12-1. Accuracy score for each classification mechanism.

Classifier	Accuracy
Dump Classifier	0.51
Naïve Bayes	0.82
Support Vector Machine (SVM)	0.93
Decision Tree	0.99
K Nearest Neighbour Classifier	0.97
Logistic Regression	0.89
Random Forest	0.99
Adaboost Classifier	0.98

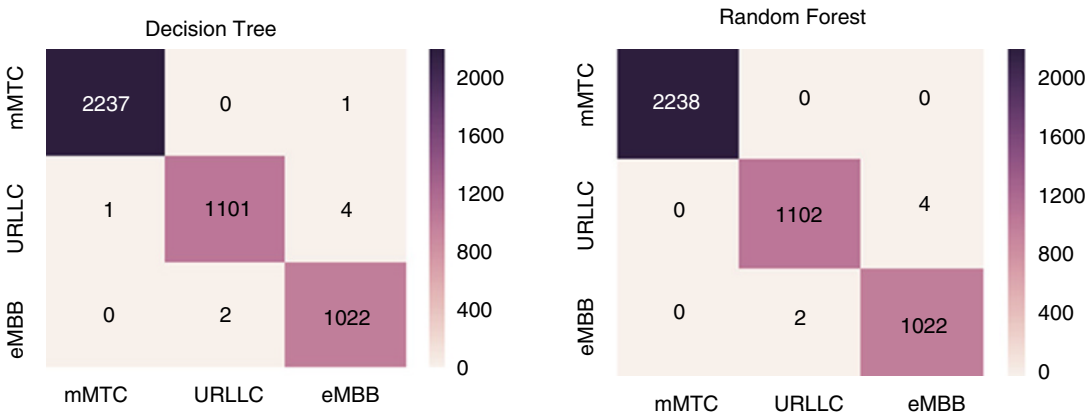

Figure 12-3. Performance evaluation results considering the indicative scenario.

results, as they miss-classify only a few flows, resulting also in high values of recall and precision. Therefore, these two classification mechanisms can be selected for further consideration for the problem of service classification.

12.3 Dynamic Multi-Service Scheduling

Allocation of radio transmission resources is one of the most essential RRM tasks. A major challenge is to balance these resources among different transmission types, e.g., unicast, multicast, and broadcast, as well as scheduled and non-scheduled UL access. Here, we first focus on the network-controlled resource allocation for scheduled unicast transmissions between the network infrastructure and UEs, as handled by the Medium Access Control (MAC) layer dynamic scheduler. In addition to

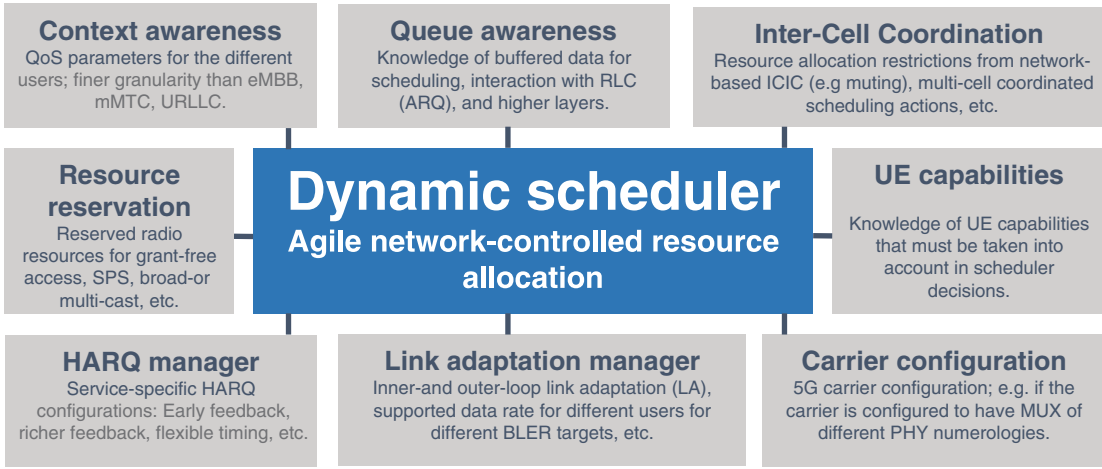

Figure 12-4. High-level illustration of the main interfaces to the dynamic scheduler for unicast transmissions.

dynamic scheduling, also semi-persistent scheduling (SPS) mechanisms and various forms of grant-free scheduling are of relevance for 5G, especially for the UL direction.

The main interfaces to the dynamic unicast packet scheduler are illustrated in Figure 12-4 [14]. Note that this is an example where diverse services are multiplexed on a shared carrier, or set of carriers. For scenarios with a stronger slice separation on MAC and PHY layers, e.g., slices with dedicated spectrum, the same principle as depicted in the figure could be applied in parallel per slice. For an overview on the possible extents of slice multiplexing on different layers in the RAN, please refer to Section 8.2.3.

The dynamic scheduler needs context information for the users under its control to fulfill the users' QoS and quality of experience (QoE) requirements, as well as knowledge of the buffered data amounts for each of the users. This queue awareness functionality serves as an interface with higher layers, e.g., Transmission Control Protocol (TCP) and other RM functionalities, such as multi-AI/multi-slice traffic steering as presented in the following Section 12.4 [15]. In particular, TCP-awareness can help to avoid re-buffering events for streaming and queuing delays for latency-critical traffic. Furthermore, the dynamic scheduler needs to know which radio resources are available for unicast transmissions, and which resources are reserved for other purposes, such as SPS, broadcast, and non-scheduled access. Knowledge of UE categories and related constraints is obviously also needed by the scheduler. UE categories basically express terminal capabilities, such as the maximum supported data rate and multiple-input multiple-output (MIMO) capability which influence on how much data can be scheduled to the UEs, and which formats are supported. As 5G will support cell carriers with a mixture of different PHY numerologies, information about carrier configuration is also needed by the scheduler. The link adaptation manager essentially provides information on the supported data rate for the different users. It also facilitates to operate different users with a service-specific first transmission block error rate (BLER); for example, using an initial BLER target of less than 1% for latency-critical traffic with tight reliability constraints, while using 10% - 20% BLER for best effort eMBB traffic.

As the scheduler also needs to dynamically allocate transmission resources for hybrid automatic repeat request (HARQ) transmissions, it interfaces the HARQ manager as well. Notice here that in

5G, asynchronous HARQ is assumed for both link directions (i.e., UL and DL), giving the scheduler freedom to decide at which point in time it wants to schedule HARQ retransmissions [16]. Similarly, other HARQ enhancements are considered, such as a flexible timing of acknowledged (ACK)/ negative acknowledged (NACK) messages and number of HARQ processes [17], options for early HARQ feedback to reduce latency [18], as well as richer HARQ feedback, allowing the scheduler to optimize the usage of radio resources for HARQ retransmissions [19].

Finally, the scheduler is also subject to constraints from multi-cell cooperation, e.g., the network-based ICI coordination (ICIC) schemes that are further described in Section 12.5.

12.3.1 Scheduling Formats and Flexible Timing

In order to support the diverse service requirements handled by the previously described dynamic scheduler in the 5G New Radio (NR) system, especially variable transmission time interval (TTI) sizes, a flexible resource frame structure and timing are required.

The flexible frame structure is realized by a scalable numerology [3], due to parametrized subcarrier spacing of $2^m \cdot 15$ kHz, where m is an element of the set of integer values [0, 1, 2, 3, 4, 5]. This scalable subcarrier spacing allows flexibility of resource units in the frequency domain. It also corresponds to flexible orthogonal frequency division multiplexing (OFDM) symbol lengths in time domain. Similar to LTE, multiple symbols are combined into a slot. The number of OFDM symbols is given by the reference numerology such that the slot time duration is $1/2^m$ ms.

Consequently, with the same control plane (CP) overhead, the number of OFDM symbols is independent of the subcarrier spacing scaled by m. It is further agreed that 5G NR should also support the LTE normal CP duration for each subcarrier spacing. As an example, $m = 0$ and $m = 2$ correspond to 15 kHz and 60 kHz subcarrier spacing, respectively, resulting in a slot duration of 1 ms and 0.25 ms. Similar to the time domain, a minimum scheduling unit for scheduling comprises also multiple subcarriers in the frequency domain, called a physical resource block (PRB) in LTE. Additional to the flexibility by this scalable numerology, further degrees of freedom are provided by the option to schedule slots or mini-slots. Specifically, a slot is defined as 14 OFDM symbols, and the smallest mini-slot is 1 OFDM symbol [4] [16]. Note that, on the other hand, also larger scheduling units can be assigned, spanning multiple slots or mini-slots.

However, to support the diverse traffic and latency requirements mentioned in the previous section, also flexible timing of the scheduled resources is essential. An illustrative example for the envisioned 5G flexible timing is provided in Figure 12-5, where a TDD scenario with three slots for DL followed by two for UL transmissions is shown. Further signaling, such as reference signals, are omitted for visibility.

The flexible timing is applied by the following main design principles. First, DL control and data channel transmissions are multiplexed on a per-user basis and control channels are front loaded, which means that they appear in the start of the transmission before the corresponding DL data transmission, as detailed in [14], Chapter 2. Second, the NR PHY design should allow devices with different bandwidth capabilities to access the same carrier regardless of the NR carrier bandwidth [3]. Third, for an UE scheduled with a resource spanning multiple slots, the corresponding DL control channel signaling is transmitted only once in order to utilize benefits of reduced control channel overhead of longer TTIs. For example, user #3 in Figure 12-5 is scheduled to a resource unit spanning three slots with a single DL grant in slot #1.

While the benefits of flexible timing and scheduling are clear, such a framework requires additional signaling, e.g., transmitted in the DL control information (DCI), which contains among other things

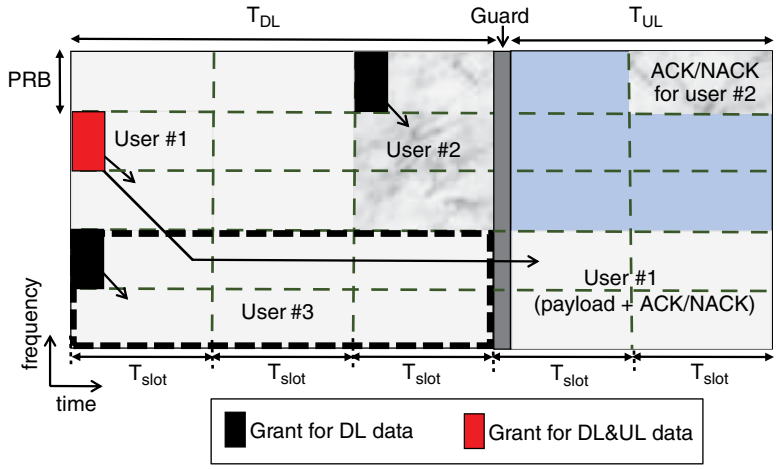

Figure 12-5. Basic illustration of possible scheduling of users on a slot resolution for a TDD scenario, where T_{slot} is the time duration of a slot and UL/DL indicates UL/DL.

resource assignment for UL or DL in an LTE system. Among other fields in the DCI, this includes the timing between a DL assignment and data transmission, the timing between the UL assignment and data transmission, and the timing between DL data reception and acknowledgement. Optionally, a DL grant can also include a future UL transmission allocation, as given in Figure 12-5 by the grant for user #1, and the timing has to be part of the DCI.

These degrees of freedom in scheduling and timing, supported by the flexible numerology and scheduling framework, allow a centralized implementation in one physical location, as seen in Figure 12-4, while RF frontends can be distributed over multiple base station (BS) sites.

12.3.2 Benefits of Scheduling with Variable TTI Size

It is well known from the existing literature that there are fundamental trade-offs between scheduling users to maximize their spectral efficiency, coverage, latency or reliability [20]. This calls for flexible scheduler functionality that allows scheduling each user in accordance with its desired optimization target. One option for this is to design the 5G system to support scheduling with different TTI sizes [21], which is the foundation of the scheme detailed in the following, based on a flexible frame structure that dynamically allows variable TTI size configuration per user and per scheduling instance.

In this way, the following scheduling decisions are possible:

- Use short TTI for low-latency communications users to optimize their latency, at the expense of increased control overhead and lower channel coding gains;
- eMBB users can be scheduled with longer TTIs and wider frequency allocations to cope with the high data rate demands;
- mMTC users can benefit from narrow bandwidth allocation and long TTIs, which are attractive characteristics from cost and coverage perspectives.

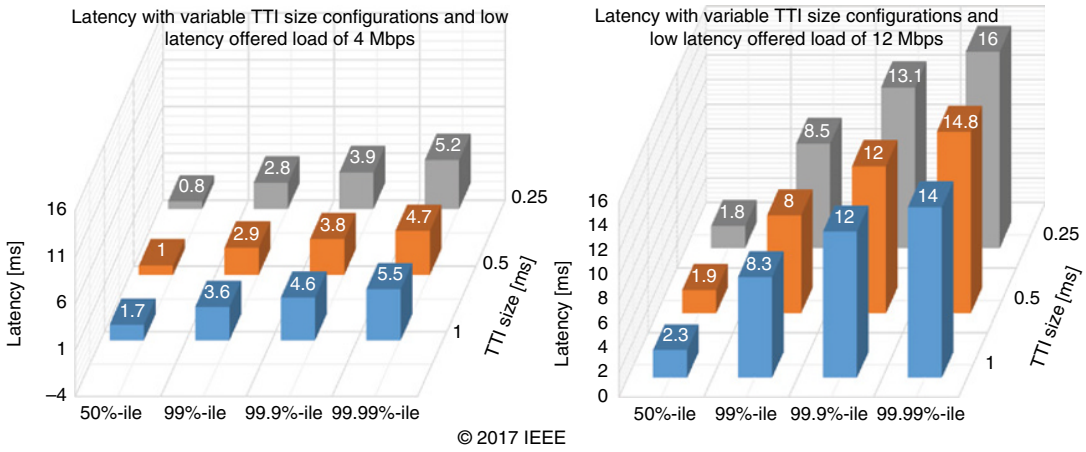

Figure 12-6. Latency values from packet latency cumulative distribution function (CDF) with variable TTI configurations and offered loads for a mix of eMBB and low-latency traffic © 2017 IEEE [25].

In addition, the possibility to set the TTI size per scheduling instance enables optimizing eMBB services that make use of TCP. A short TTI duration can be used in the first transmissions to reduce the round-trip time of the flow control mechanism in the slow start phase of TCP, and later longer TTIs can be configured to maximize the spectral efficiency, when steady operation is reached.

System-level performance evaluations have been carried out to compare the performance of several TTI size configurations, in order to estimate the most suitable TTI size that should be dynamically chosen per UE depending on service requirements, traffic type, radio channel quality and system load. The evaluation is performed in a 3GPP Urban Macro scenario with 7 BSs, each having 3 sectors, 500 m inter-site distance (ISD) and using 10 MHz bandwidth [22]. In-resource control channel (CCH) scheduling grants with link adaptation are assumed, which allows to model different degrees of CCH overhead (i.e., aggregation levels or number of resource elements dedicated to CCH) depending on the UE radio conditions [21] [23].

The packet latency, i.e., the MAC layer one-way UP latency (including scheduling delay) achieved with different TTI sizes and system loads with a mix of eMBB and low-latency services is shown in Figure 12-6. The eMBB traffic is modelled with a single user full buffer download, whereas higher priority low-latency traffic follows a Poisson arrival process with 1 kB payload and varying total cell offered load. More details can be found in [24] [25]. As depicted in Figure 12-6, using a short TTI at low system loads is in general a more attractive solution to achieve low-latency communications. However, looking at the tail values (99.9%-ile and above), a 0.5 ms TTI size offers better latency than the 0.25 ms TTI, even for low loads.

As the load increases, longer TTI configurations with lower relative CCH overhead, and consequently higher spectral efficiency, provide better performance, as these better cope with the non-negligible queuing delay. The 1-ms TTI configuration is beneficial from a latency point of view for high loads and above the 99.9%-ile, due to queuing delay. As the offered load increases, or as UEs with the worst channel conditions are considered, the queuing delay becomes the most dominant component of the total latency; therefore, it is beneficial to increase the spectral efficiency of the transmissions, by using a longer TTI, in order to reduce the experienced delay

in the queue. The observed trends are relevant for URLLC use cases requiring latency guarantees of a few milliseconds and reliability levels up to 99.999%.

The results presented above and detailed in [24] [25], as well as related studies performed in [23] (focused on TCP performance with variable TTI size configuration), indicate that the optimum TTI size varies depending on multiple factors. Therefore, it is beneficial to be able to dynamically adjust the TTI size per user's service requirements and scheduling instance, rather than operating the system with a fixed TTI.

12.3.3 Punctured/Preemptive Scheduling

Punctured scheduling, as pictured in Figure 12-7, is proposed to enable multi-service scheduling with diverse requirements [14]. A single block in Figure 12-7 is a scheduling unit in time domain from UE #2 perspective, which can be a mini-slot of one or several OFDM symbols. The symbol duration depends on the subcarrier spacing of the system. Here, UE #1, receiving eMBB traffic, is scheduled by the BS for transmission on the DL shared radio channel. The former is facilitated by the BS sending a scheduling grant, transmitted on a PHY control channel, followed by the actual transmission of the transport block. During the scheduled transmission time of the transport block for UE #1, the DL shared channel for this transmission is in principle monopolized by the UE. As illustrated in Figure 12-7, it may happen that URLLC data for UE #2 arrives at the BS shortly after the transmission towards UE #1 has started. In order to avoid having to wait for the completion of the transport block transmission to UE #1, it is instead propose to immediately transmit the URLLC data to UE #2 by puncturing (i.e., over-writing) part of the ongoing transmission to UE #1. The advantage of this solution is that the latency of the data to UE #2 is minimized, and there is no need to a priori reserve radio resources for transmission of URLLC that may potentially come.

The drawback is in the performance of the transmission towards UE #1. Depending on how large a fraction of the resources for the transmission towards UE #1 is punctured, UE #1 may still be able to correctly receive the data thanks to efficient forward error correction (FEC).

In Figure 12-8, the simulated performance of three transceiver approaches with different extents of knowledge about the puncturing at both transmit and receive sides is shown [26]. In the first approach, assuming prior knowledge about puncturing at both transmitter and receiver sides, rate

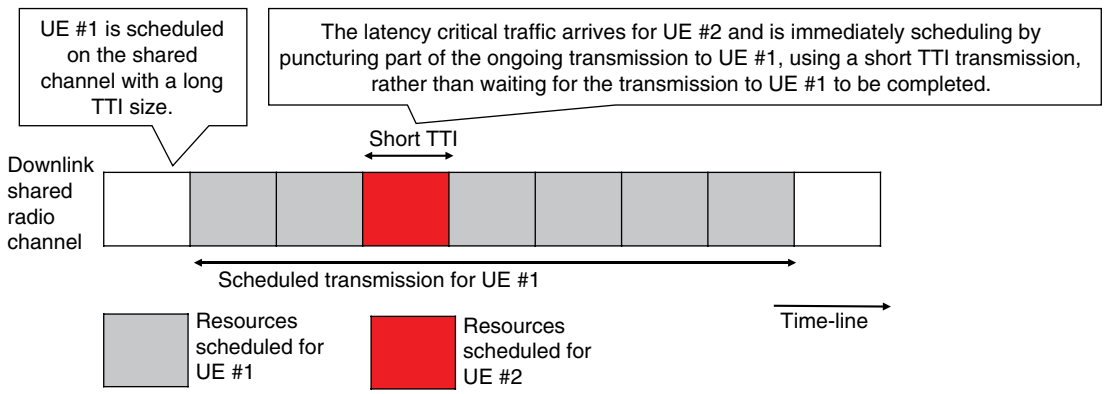

Figure 12-7. Sketch of the basic principles of punctured scheduling on the DL shared data channel.

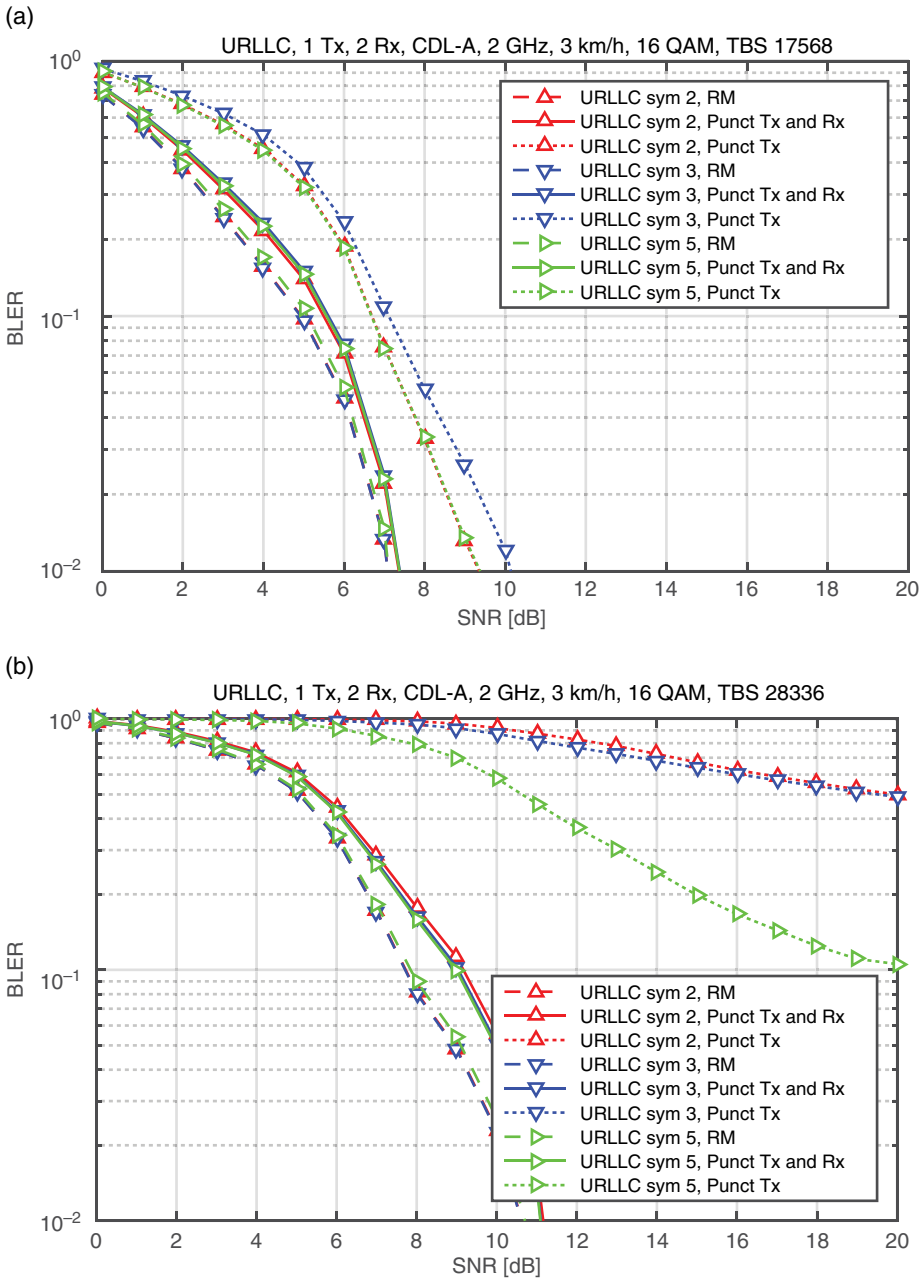

Figure 12-8. Three transceivers: Ideal rate matching, puncturing at transmitter (Tx) only, and puncturing at both Tx and receiver (Rx) with (a) transport block size (TBS): 17568 bits, (b) TBS: 28366 bits [26]. QAM: Quadrature Amplitude Modulation.