EMERGING TRENDS IN WIRELESS INFRASTRUCTURE

White paper | Version 01.00 | Dr. Nishith Tripathi and Dr. Jeffrey Reed

ROHDE&SCHWARZ

Make ideas real

CONTENTS

INTRODUCTION

5G, the fifth-generation cellular technology, is being deployed around the world. 5G is a transformational technology designed to provide significant flexibility and support a variety of use cases. This paper describes the emerging infrastructure trends of wireless networks for 4G, 5G, and beyond 5G. These trends provide even more opportunities to service providers for network deployments, network customization, and network optimization. These key trends include (i) spectrum trends, (ii) densification & coverage extension methods, (iii) virtualization and cloudification, and (iv) network customization and intelligence.

In emerging spectrum trends, higher integration of sub-6 GHz/sub-7 GHz frequencies and millimeter-wave frequencies, increased availability of unlicensed spectrum, and spectrum sharing opportunities are discussed.

In the area of densification and coverage extension, trends such as small cells with beamforming, Integrated Access and backhaul, and special infrastructure enhancements in support of Vehicle-to-Everything communications are observed in addition to traditional areas such as Fixed Wireless Access and Distributed Antenna System.

In the area of virtualization and cloudification, the technologies such as Network Functions Virtualization, Software Defined Networking, and orchestration are expected to provide scalability and reduce costs of deployment. Furthermore, the disaggregation of the gNB into gNB-Central Unit and gNB-Distributed Unit and further disaggregation of the gNB-Central Unit provide scalability and additional opportunities for implementing virtualization.

In the area of network customization and intelligence, trends such as Network Slicing, Multi-access Edge Computing, Non-Public Network, Non-Terrestrial Network, Self-Optimizing Network enhancements for 5G, and Open-Radio Access Network are observed.

1 5G NETWORK IN A NUTSHELL

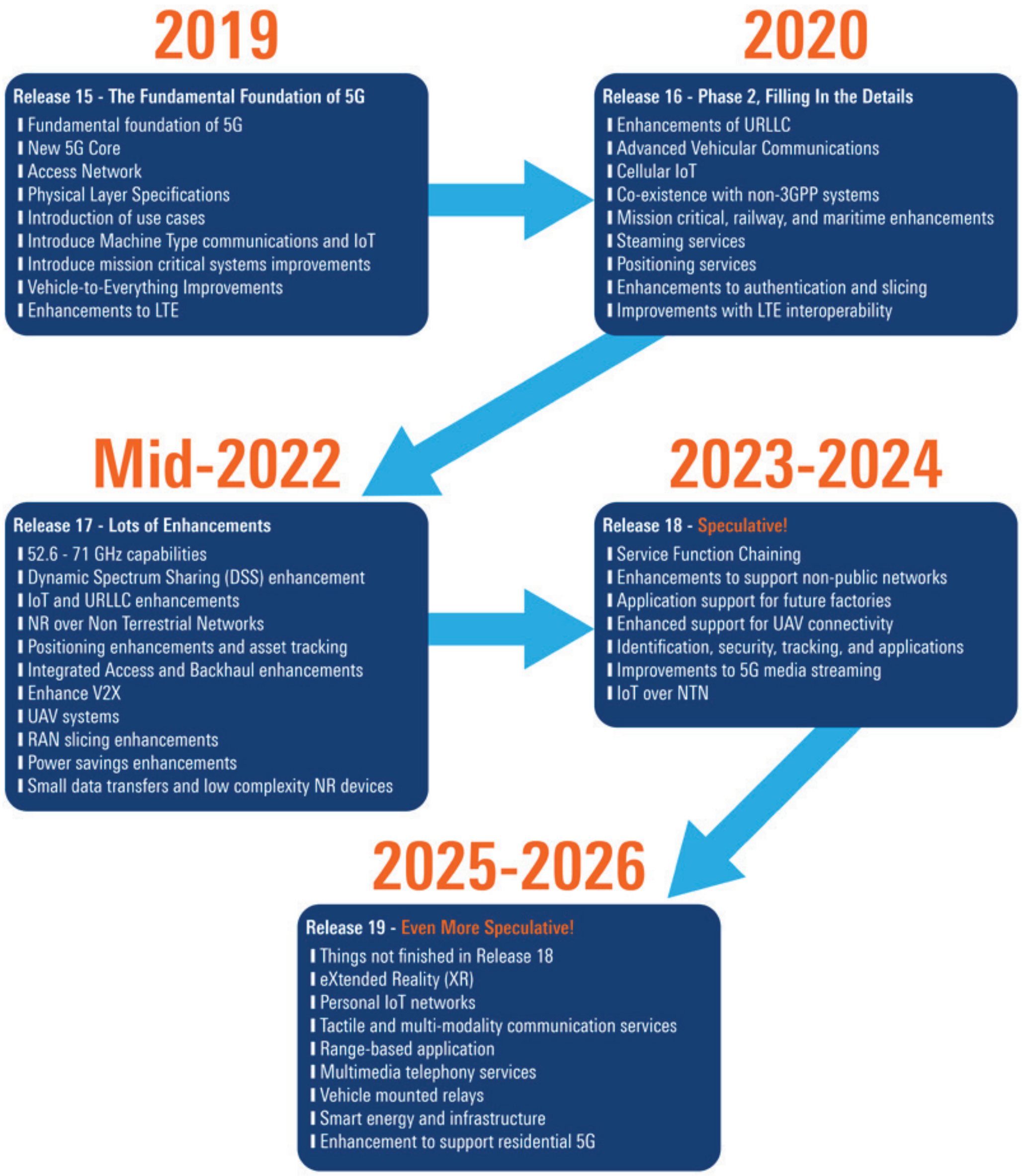
The Third Generation Partnership Project (3GPP) released specifications for the fifth-generation (5G) cellular technology in Release 15 in 2019. 5G is a transformational technology that is expected to revolutionize the way we live and work and transform numerous industries and various facets of world economies. Section 1.1 provides a glimpse of evolving 5G specifications beyond Release 15. The end-to-end architecture of the 5G network consisting of the radio network, the packet-switched core network, and the services network is illustrated in Section 1.2. 5G supports various configurations or architecture options for connecting LTE-based radio network and 5G New Radio (NR)-based radio network to the LTE's Evolved Packet Core (EPC) and 5G Next-Generation Core (NGC). Section 1.3 explains the two most popular architecture options, Standalone NR with the NGC and Non-Standalone NR with the EPC. The emerging trends in the wireless infrastructure that utilizes the Release 15-based logical network architecture as the baseline and facilitate and broaden the deployment of 5G are mentioned in Section 1.4 and elaborated in Sections 2 to Section 5 of the paper.

1.1 5G: An Evolving Standard

5G is an evolving standard. The initial version of 5G was accomplished with the freezing of Release 15 late-drop in summer 2019. Enhanced features originally were introduced in Release 15 and new features were incorporated in Release 16 in 2020. More enhancements of features defined in prior releases and new features are part of Release 17 that is currently in progress. Freezing of Release 17 is currently anticipated to be around mid-year 2022, and any further timeline beyond Release 17 is speculative.

The initial specifications of 5G, Release 15, provided the fundamental architecture of 5G and addressed the use cases that include the famous “magic triangle” of enhanced Mobile Broadband (eMBB), ultra-reliable and low-latency communications (URLLC), and massive machine-type communications (mMTC). Release 16 and Release 17 have focused on enhancing those use cases, especially URLLC and mMTC. By the end of 2021 or early 2022, priority goals for Release 18 should be available in the 2023-2024 timeframe. A summary of selected features defined or refined in various releases is shown in Figure 1.1. While Release 15 provided the fundamental architecture of 5G, subsequent releases have been simultaneously focusing on “three” directions:

Figure 1. Evolution of 5G: Features for Various 3GPP Releases



2019

Release 15 - The Fundamental Foundation of 5G

- Fundamental foundation of 5G
- New 5G Core
- Access Network
- Physical Layer Specifications
- Introduction of use cases
- Introduce Machine Type communications and IoT
- Introduce mission critical systems improvements
- Vehicle-to-Everything Improvements
- Enhancements to LTE

2020

Release 16 - Phase 2, Filling In the Details

- Enhancements of URLLC
- Advanced Vehicular Communications
- Cellular IoT
- Co-existence with non-3GPP systems
- Mission critical, railway, and maritime enhancements
- Steaming services
- Positioning services
- Enhancements to authentication and slicing
- Improvements with LTE interoperability

Mid-2022

Release 17 - Lots of Enhancements

- 52.6 - 71 GHz capabilities
- Dynamic Spectrum Sharing (DSS) enhancement
- IoT and URLLC enhancements
- NR over Non Terrestrial Networks
- Positioning enhancements and asset tracking
- Integrated Access and Backhaul enhancements
- Enhance V2X
- UAV systems
- RAN slicing enhancements
- Power savings enhancements
- Small data transfers and low complexity NR devices

2023-2024

Release 18 - Speculative!

- Service Function Chaining
- Enhancements to support non-public networks
- Application support for future factories
- Enhanced support for UAV connectivity
- Identification, security, tracking, and applications
- Improvements to 5G media streaming
- IoT over NTN

2025-2026

Release 19 - Even More Speculative!

- Things not finished in Release 18
- eXtended Reality (XR)
- Personal IoT networks
- Tactile and multi-modality communication services
- Range-based application
- Multimedia telephony services
- Vehicle mounted relays
- Smart energy and infrastructure
- Enhancement to support residential 5G

[Ref: https://www.3gpp.org/specifications/releases]

Release 16 was a major release and necessary to broaden the overall system specifications of Release 15. Release 16 brought extensions to V2X communications (i.e., 5G NR-based direct device-to-device communications or sidelink communications) to extend automated and remote driving, Industrial Internet of Things (IIoT), enhancements to URLLC, numerous energy efficiency changes, Integrated Access and Backhaul (IAB) (that brings a relay function to 5G), coexistence support for non-3GPP systems such as improvements in wireline and wireless systems, mission-critical public warning systems, improvement in voice, multimedia, and streaming services, NR-Unlicensed (NR-U), and 5G positioning or location-based services (LBS).
[Ref: https://www.3gpp.org/release-16]

Release 16 included many studies in key areas that may guide future releases, such as Release 17 and 18. These include (Non-Terrestrial Networks) NTNs, coverage and positioning enhancements, NR and slicing QoE work, the addition of new frequency ranges, NR reduced-capacity devices, enhanced support of non-public networks (NPNs), support for unmanned aerial systems, support for edge computing in 5GC, proximity-based services in 5GS, network automation for 5G and for access traffic steering, switch and splitting (ATSSS), among others.

Release 17 features already in the pipeline include new work and/or enhancements for URLLC for NR-based IIoT, NR-based NTN, MIMO, integrated access and backhaul (IAB), MBS positioning, NR multicast, and broadcast services, RAN slicing for NR, NR sidelink, multi-RAT dual-connectivity (MR-DC), support for multi-SIM devices for LTE/NR, and NR small data transmissions in an inactive state and multimedia priority service. [Ref: https://www.3gpp.org/release-17]

The schedule for Release 18 may not be known until late 2021 or early 2022, and hence it might be late 2023 or even into 2024 before Release 18 is frozen or finalized. Work items possible for that release may include enhancements of multimedia telephony services, vehicle-mounted relays, smart energy and infrastructure, and enhancements to support residential 5G. [Ref: 7] Release 17 study items include support for eXtended Reality (XR) and IoT over NTN, and these study items may set the stage for Release 18 work items.

One of the more interesting items anticipated for Release 18 is a study on the performance requirements for AI/ML model and data distribution, distributed/federated learning, model transfer and training requirements, operations splitting, and characterization of use cases such as image recognition, video improvements, robotic control, speech recognition, automotive networks, and "flocking." [Ref: Technical Report 3GPP TR 22.874 v 1.0.0, March 2021 found at https://itectec.com/archive/3gpp-specification-tr-22-874/] Release 18 may start in 2022 and finish in 2024. Products with Release 18 features may be available starting in 2026-2027, especially if it takes more than one release to perfect the specification around those features.

Other study items possible for Release 18 include personal IoT networks, support for tactile and multi-modality communication services, range-based applications, smart energy and infrastructure, enhancements of multimedia telephony services, improved timing resilience, and railway communications. Given that these are study items, the transition of some of these studies to work items may begin in Release 19. If release development intervals continue past trends, Release 19 may start in 2024 or 2025.

Trends with Releases

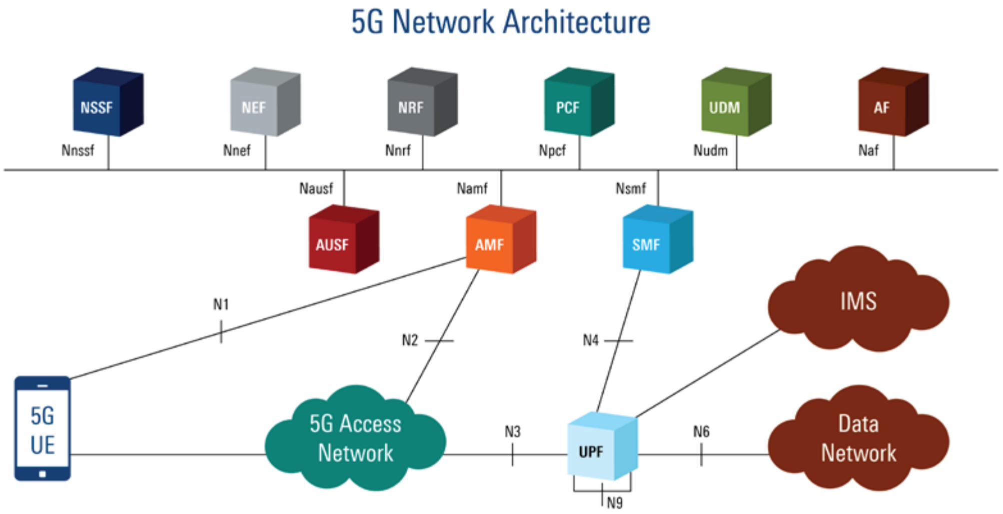
One aspect that has been consistent in the evolution of the 5G standard is the focus on developing standards around application use cases. The flexibility built into 5G was envisioned from the start and considered essential to support the evolving applications and use cases of wireless communications. To some extent, predictions can be made about future steps in 3GPP's processes based on studies being performed to set the stage for working groups. However, one thing that isn't as predictable is consumer acceptance of the various services and the appearance of unexpected services that 5G technology enables. One may expect that the 5G standardization process to morph as the success or failure of expected and unexpected applications becomes apparent. New generations of wireless systems tend to appear every ten years. Some research efforts are underway toward 6G. Hence, 6G may become a reality in the early 2030s. However, given the flexibility of 5G, it has the potential to progress with incremental improvements for a long time. There are still many improvements and applications that can be addressed by evolving the 5G standard.

New generations of wireless systems tend to appear every ten years. Some research efforts are underway toward 6G. Hence, 6G may become a reality in the early 2030s. However, given the flexibility of 5G, it has the potential to progress with incremental improvements for a long time. There are still many improvements and applications that can be addressed by evolving the 5G standard.

1.2 End-to-End 5G Network

3GPP has defined a logical network architecture of 5G. Figure 2 illustrates a simplified service-based representation of the 5G system architecture [REF: 3GPP, TS23.501, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144]

Figure 2. 5G System Architecture: A Service-Based Representation.



The 5G system includes the User Equipment (UE), the (Radio) Access Network (AN/RAN), and the 5G Core (5GC) or NGC entities. In addition, the 5G system defines Network Functions (NFs) that perform specific functions. Examples of NFs include the UE, the next-generation Node B (gNB) in the RAN, and the Access and Mobility Management Function (AMF). Compared to an LTE network, the 5G network includes more entities or NFs. While LTE defines point-to-point interfaces, 5G follows the paradigms of modularity and self-containment for network functions to foster reuse and extensibility of system functionality. Furthermore, the 5G network is virtualization-friendly, where the NFs can be implemented using a virtualized network infrastructure or “cloud infrastructure” (see Section 4 for details). In the service-based representation of the 5G system, an NF offers its services to other NFs using a service-based interface. For example, the AMF offers its services to other authorized NFs via the serviced-based interface $N_{amf.}$

The 5G AN includes the base stations called gNBs. The gNB communicates with 5G UEs via the NR-based Uu interface. The gNB has a User Plane (UP) interface with the User Plane Function (UPF). The UPF provides access to Data Networks (DNs) such as the Internet. For example, Internet Protocol (IP) packets carrying a video from a server travel through the routers in the Internet arrive at the UPF. The UPF forwards such IP packets to the gNB, and, the gNB utilizes an NR-based radio protocol stack to deliver the packets to the UE.

The gNB has a Control Plane (CP) interface with the AMF via the N2 Reference Point. The UE and the AMF can exchange Non-Access Stratum (NAS) signaling messages with the help of Radio Resource Control (RRC) signaling on the radio interface and the Next Generation Application Protocol (NGAP) signaling on the N2 interface. Such NAS signaling facilitates the management of security, sessions, and network slices. The Authentication Server Function (AUSF) utilizes the assistance of the AMF to carry our mutual authentication with the UE. The Session Management Function (SMF) anchors and manages the users' sessions and allocates IP addresses to UEs. The Unified Data Management (UDM) stores information about the subscribers, such as subscribers' identities. The Application Function (AF) interacts with the policy control framework to facilitate the implementation of Quality of Service (QoS) in the 5G network. While the IP Multimedia Subsystem (IMS) is external to the 5G System, the role of the Proxy-Call Session Control Function (P-CSCF) of the IMS network is also performed by the Application Function (AF). More specifically, the AF extracts QoS requirements of a multimedia session from Session Initiation Protocol (SIP) signaling messages between two communication endpoints (e.g., two UEs) and provides such QoS needs to the Policy Control Function (PCF). The PCF translates the overall QoS needs into 5G-specific QoS and provides the 5G QoS rules to the SMF. The SMF, the AMF, and, the gNB work with each other to set up a (5G) QoS Flow that meets the QoS requirements.

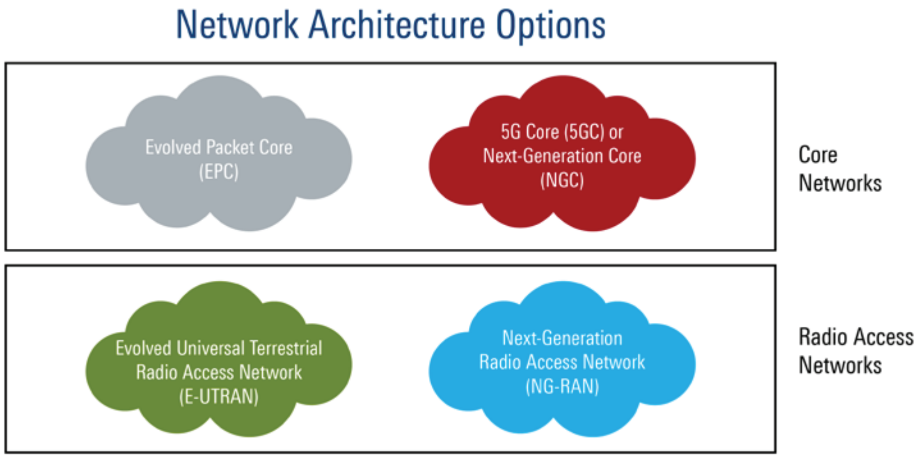
The Network Repository Function (NRF) enables an NF to store its information (e.g., availability) so that other NFs can find it when needed. The Network Exposure Function (NEF) enables entities external to the 5G network to securely access a 3GPP-defined NF and make use of the 5G network. The Network Slice Selection Function (NSSF) enables the selection of a suitable network slice for a given UE when needed. A concept of a Network Slice can be used to meet requirements of different customers of the wireless service provider (e.g., two different enterprises with different sets of requirements on availability or reliability) and diverse QoS requirements of services (e.g., enhanced Mobile Broadband Service (eMBB) service vs. Ultra-Reliable Low-Latency Communications (URLLC) service).

While 3GPP allows implementation of the NFs using dedicated physical pieces of equipment similar to typical third-generation (3G) and initial 4G LTE deployments, virtualized NFs are expected to become more prevalent in the coming months and years.

1.3 Radio-Core Connectivity [Overview of options including SA and NSA NR]

5G defines various architecture options, where LTE-based RANs and NR-based RANs connect to LTE's EPC and 5G's NGC. Figure 3 lists the architecture options for such connectivity, which 3GPP has discussed. All these options are not currently fully supported; more options are expected to be available in different releases of the 5G specifications. The existence of multiple radio-core connectivity options provides flexibility to service providers.

Figure 3. 5G Architecture Options: Radio-Core Connectivity Choices.


Network Architecture Options
Evolved Packet Core (EPC)
5G Core (5GC) or Next-Generation Core (NGC)
Core Networks
Evolved Universal Terrestrial Radio Access Network (E-UTRAN)
Next-Generation Radio Access Network (NG-RAN)
Radio Access Networks

Option 1: E-UTRA with EPC (Baseline)
Option 2: Standalone NR with NGC
Options 3, 3a, 3x: Non-Standalone NR with EPC
Options 4, 4a: Non-Standalone E-UTRA with NGC
Option 5: Standalone E-UTRA with NGC
Option 6: Standalone NR with EPC
Options 7, 7a, 7x: Non-Standalone NR with NGC
Options 8, 8a: Non-Standalone E-UTRA with EPC

The baseline option is Option 1, where the LTE-based RAN called Evolved Universal Terrestrial Radio Access Network (E-UTRAN) is connected to the EPC. The two most popular deployment options are Non-Standalone (NSA) NR with the EPC and Standalone (SA) NR with the 5GC.

Most 5G service operators started offering 5G using the NSA NR with the EPC, which is formally called E-UTRA-NR Dual Connectivity, where the LTE eNB acts as the Master Node (MN) and the NR gNB acts as the Secondary Node (SN). In EN-DC, the LTE eNB (as the MN) provides the UE with a Control Plane connection to the EPC so that NAS signaling can be exchanged between the UE and the MME in the EPC. The NR gNB as the Secondary Node provides additional radio resources. Since the NR gNB needs the help of the LTE eNB (due to the NAS signaling connection passing through the eNB), this option is referred to as the Non-Standalone NR option. In one of the EN-DC flavors ("Option 3x"), the UE can simultaneously obtain user traffic from an LTE eNB and an NR gNB, and the packets can be forwarded on the Xn interface between the eNB and the gNB. For example, the NR gNB can send some packets directly to the UE using the NR air interface and forward some packets to the eNB so that the eNB can send some packets to the UE using the LTE air interface. Since the EN-DC architecture utilizes the EPC and not the 5GC, and 5G can be deployed quickly without waiting for the 5GC. Furthermore, LTE is available as default technology and 5G is utilized whenever the 5G coverage overlaps with the LTE coverage. A key benefit of the EN-DC approach is faster time-to-market.

In the SA NR with the NGC (“Option 2”) architecture, the NR gNB does not rely on an LTE eNB for connecting the UE to the core network; the gNB itself is connected to the 5GC. Hence, such an option is called the Standalone NR option. Since this option utilizes the 5GC in addition to the 5G NR radio interface, full 5G capabilities can be exploited. For example, in the SA NR with the NGC architecture, 5G QoS can be provided and features such as network slices can be fully and properly implemented.

After 5G Phase 1 (i.e., Release 15), 3GPP has begun to specify the support for other Multi-Radio Dual Connectivity (MR-DC) architectures such as Next Generation- RAN (NG-RAN) E-UTRA-NR Dual Connectivity (NGEN-DC), NR-E-UTRA Dual Connectivity (NE-DC), and NR-NR Dual Connectivity (NR-DC) [TS37.340]. All these MR-DC architectures utilize the 5GC as the core network. In MR-DC, two independent schedulers operate at the MN and the SN, and the MN and the SN may be connected with each other via a non-ideal (or ideal) backhaul. A non-ideal backhaul implies that the timings at the MN and the SN may not be aligned (i.e., there could be some delay between the MN transmission and the SN transmission).

In the case of NGEN-DC, the UE is connected to a next-generation eNB (ng-eNB) that acts as the MN and a gNB that acts as an SN. An ng-eNB is an LTE eNB that communicates with the UE using the LTE air interface and is connected to the 5GC using the NG N2 interface. In the NE-DC scenario, the UE is connected to an NR gNB that acts as the MN and an eNB that acts as an SN. The NR-DC utilizes a gNB as the MN and another gNB as the SN.

1.4 Emerging Trends: Overview and Motivation

As shown in Section 1.3, 5G offers deployment flexibility to service providers. Figure 4 gives examples of emerging trends in the wireless infrastructure that provide even more opportunities to service providers for network deployments, network customization, and network optimization. These key trends include (i) spectrum trends (ii) densification & coverage extension methods, (iii) virtualization and cloudification, and (iv) network customization and intelligence.

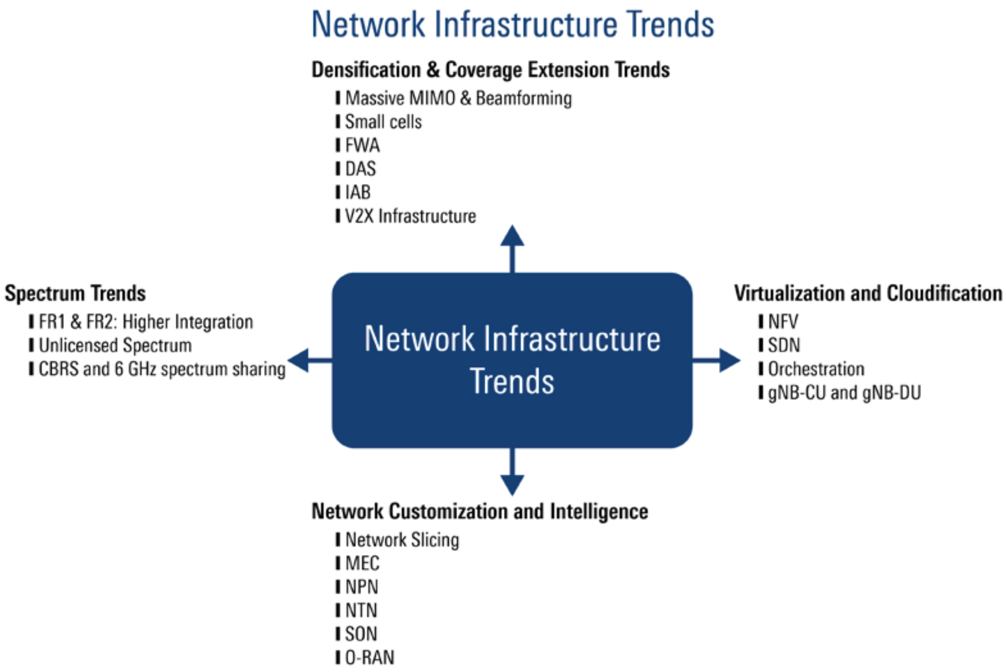
In emerging spectrum trends, higher integration of sub-6 GHz/sub-7 GHz Frequency Range 1 (FR1) and millimeter wave (mmW) Frequency Range 2 (FR2), increased availability of unlicensed spectrum, and spectrum sharing opportunities (e.g., Citizen Broadband Radio Service (CBRS) in the U.S. and 6 GHz) are observed.

In the area of densification and coverage extension, trends such as small cells with mmW beamforming, Integrated Access and backhaul (IAB), and special infrastructure enhancements in support of Vehicle-to-Everything (V2X) communications (e.g., Road Side Units (RSUs)) are observed in addition to traditional areas such as Fixed Wireless Access (FWA) and Distributed Antenna System (DAS).

In the area of virtualization and cloudification, the technologies such as Network Functions Virtualization (NFV), Software-Defined Networking (SDN), and orchestration are expected to provide scalability and reduce costs of deployment. Furthermore, the disaggregation of the gNB into gNB-Central Unit (gNB-CU) and gNB-Distributed Unit (gNB-DU) and further disaggregation of the gNB-CU into gNB-CU-CP and gNB-CU-UP provide scalability and additional opportunities for implementing virtualization.

In the area of network customization and intelligence, trends such as Network Slicing, Multi-access Edge Computing (MEC), Non-Public Network (NPN), Non-Terrestrial Network (NTN), Self-Optimizing Network (SON) enhancements for 5G, and Open-RAN (O-RAN) with intelligent RAN control are observed.

Figure 4. Emerging Trends in the Wireless Infrastructure.


Network Infrastructure Trends
Densification & Coverage Extension Trends
Massive MIMO & Beamforming
Small cells
FWA
DAS
IAB
V2X Infrastructure
Spectrum Trends
FR1 & FR2: Higher Integration
Unlicensed Spectrum
CBRS and 6 GHz spectrum sharing
Network Infrastructure Trends
Virtualization and Cloudification
NFV
SDN
Orchestration
gNB-CU and gNB-DU
Network Customization and Intelligence
Network Slicing
MEC
NPN
NTN
SON
O-RAN

2 SPECTRUM TRENDS

Spectrum provides precious radio interface resources for wireless communications. As newer generations of wireless technologies are defined by standards bodies such as 3GPP and IEEE, additional frequency bands are frequently introduced. For example, initial 4G LTE deployments in the U.S. began using 700 MHz spectrum that was previously not used in the first three generations of cellular technologies. Similarly, 28 GHz millimeter wave (mmW) spectrum was first used in 5G deployments. Such mmW spectrum was not used in the first four generations of cellular technologies. Furthermore, spectrum is also re-farmed, whereby the older technologies are replaced by newer technologies in a given frequency band. For example, 850 MHz and 1900 spectrum previously used by 2G and 3G technologies are re-farmed so that 5G can make use of such spectrum. Section 2.1 introduces three different types of spectrum (i.e., licensed, unlicensed, and shared) and the frequency ranges defined by 3GPP for 5G. A huge amount of unlicensed spectrum is becoming available for cellular and Wi-Fi technologies, which is the focus of Section 2.2. Spectrum sharing is an important trend that is gaining momentum. Section 2.3 shows how spectrum sharing can be implemented using an example of the 3.5 GHz spectrum in the U.S.

2.1 3GPP FR1 and FR2

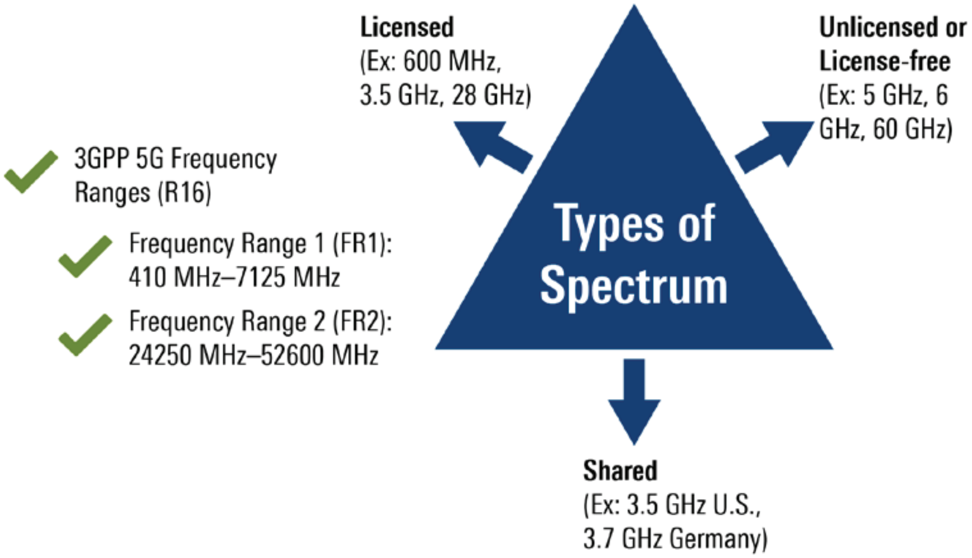
As shown in Figure 2.X, there are three types of spectrum: licensed, unlicensed or license-free, and shared. While licensed spectrum has been widely utilized since the birth of cellular technologies and unlicensed spectrum has been widely utilized in Industrial, Scientific, and Medical (ISM) applications and Wi-Fi technologies, shared spectrum has gained popularity in the recent years.

Licensed spectrum is typically auctioned by regulatory agencies in specific countries (e.g., Federal Communications Commission CC in the U.S.). The main benefit of the licensed spectrum to a service provider is that the service provider obtains exclusive access to the spectrum and hence can manage interference and performance of its network (e.g., user QoS, network capacity, and network throughput) in a predictable manner. Examples of new licensed spectrum bands used in 5G include Band n71 (600 MHz FDD band in the U.S.; used by T-Mobile; 663 MHz to 698 MHz Uplink and 617 MHz to 652 MHz Downlink), Band n261 (28 GHz TDD band in the U.S., South America, and Asia; 27500 MHz to 28350 MHz; used by AT&T and Verizon Wireless), and Band n78 (3.5 GHz in Europe and Asia, and South America; 3.3 GHz to 3.8 GHz).

The main benefit of unlicensed spectrum is that the spectrum is available for use without any charges and facilitates wireless access to the Internet when mobility and stringent QoS are not major considerations (e.g., homes and enterprises). An example of unlicensed spectrum band used in 5G include Band n46 (global 5 GHz TDD band; 5150 MHz – 5925 MHz).

The main benefit of shared spectrum is that the spectrum that is typically intended for a primary entity can be used by other secondary entities based on centralized coordination of spectrum usage. The spectrum usage coordination eliminates the unpredictability of interference that exists in unlicensed spectrum. The shared spectrum approach maximizes the utilization of the spectrum and expands the wireless ecosystem. Examples of shared spectrum bands used in 5G include Band n48 in the U.S. (3.5 GHz TDD CBRS band; 3550 MHz to 3700 MHz).

Figure 2.1. Types of Spectrum and 3GPP Frequency Ranges for 5G.



3GPP has defined two frequency ranges, Frequency Range 1 (FR1) and Frequency Range 2 (FR2) for 5G. FR1 covers the frequencies from 410 MHz to 7.125 GHz, and FR2 covers the frequencies from 24.250 GHz to 52.6 GHz. A given NR frequency band belongs to one of these ranges [Ref: 3GPP, TS 38.104, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3202]. For example, the NR operating band n71 (i.e., 600 MHz FDD band) belongs to FR1 and the NR operating band n261 (i.e., 28 GHz TDD band) belongs to FR2.

FR1 and FR2 can be deployed in a variety of ways. For example, a frequency band in FR1 can provide basic coverage and a frequency band in FR2 can be deployed in small cells for very high throughput. An architecture such as New Radio- Dual Connectivity (NR-DC) can be deployed with one NR cell as the Primary Cell providing connectivity toward a 5G Core and another NR cell as Primary Secondary Cell (PSCell). The PCell and its SCells can use FR1 for coverage and PSCell and its SCells can provide additional radio resources for enhanced throughput. In other examples, Carrier Aggregation and Dual Connectivity can use FR1 and FR2 spectrum.

2.2 Unlicensed Spectrum

The unlicensed spectrum has been used in a non-cellular network for decades. For example, 2.4 GHz unlicensed spectrum has been used by ISM applications, WiFi, Bluetooth, and IoT technologies. 3GPP initially introduced a 5 GHz frequency band to make use of unlicensed spectrum in conjunction with licensed spectrum as part of the Licensed-Assisted Access (LAA) in Release 13. Since the unlicensed spectrum is shared between LAA and WiFi, such spectrum can also be viewed as shared spectrum. In LAA, LTE-based radio interface is used in both the licensed spectrum and the unlicensed spectrum. The licensed spectrum provides an anchor carrier frequency for reliable Radio Resource Control (RRC) signaling and user traffic transfer. The unlicensed spectrum provides one or more carrier frequencies in the 5 GHz frequency band for user traffic transfer. While Release 13 utilizes Carrier Aggregation (CA) in the downlink to combine licensed and unlicensed carrier frequencies for enhanced throughput in the downlink, Release 14 supports CA in the uplink to combine licensed and unlicensed carrier frequencies for enhanced throughput in the uplink. Such CA is part of the enhanced LAA (eLAA) feature in Release 14.

5G also supports the 5 GHz unlicensed spectrum by defining band n46. A Release 16 feature called New Radio- Unlicensed (NR-U) makes use of the unlicensed spectrum using both a standalone option and an anchor (i.e., licensed-assisted) option. In the standalone option, NR-U only utilizes the unlicensed spectrum; no help from the licensed spectrum is needed. The standalone option is suitable for private networks, which are formally called Non-Public Networks or NPNs in 3GPP. Of course, the NPN can also make use of licensed spectrum. A large amount of unlicensed spectrum around 60 GHz is expected to be defined by 3GPP soon. The availability of 60 GHz unlicensed spectrum can significantly increase the utility of NR-Unlicensed when local wireless access is desirable. NR-U will share the 60 GHz unlicensed spectrum with the IEEE 802.11ad technology called WiGig. Due to its high frequency, the 60 GHz spectrum is more suitable for short-range communications, where the transmitter and the receiver are separated by few tens of meters.

2.3 Spectrum Sharing and Associated Architecture Enhancements

When people think about cellular networks, they tend to think of a traditional model where service providers purchase spectrum so that they have exclusive rights to utilize that spectrum. Spectrum is a commodity very much like real estate; it tends to increase in value in the long run, and it is the backbone for commercial enterprises. And like real estate where they “just don’t make new land,” they don’t make new spectrum either! Instead, we make more efficient use of both land and spectrum. The highly prized bands are below 6 GHz, where propagation characteristics have smaller propagation losses (so-called “beachfront property”), and practically every Hertz of spectrum is assigned to some purpose. Getting exclusive access to existing spectrum/land to support new services/buildings requires dislocating a current user of spectrum/land to make room for new services/buildings.

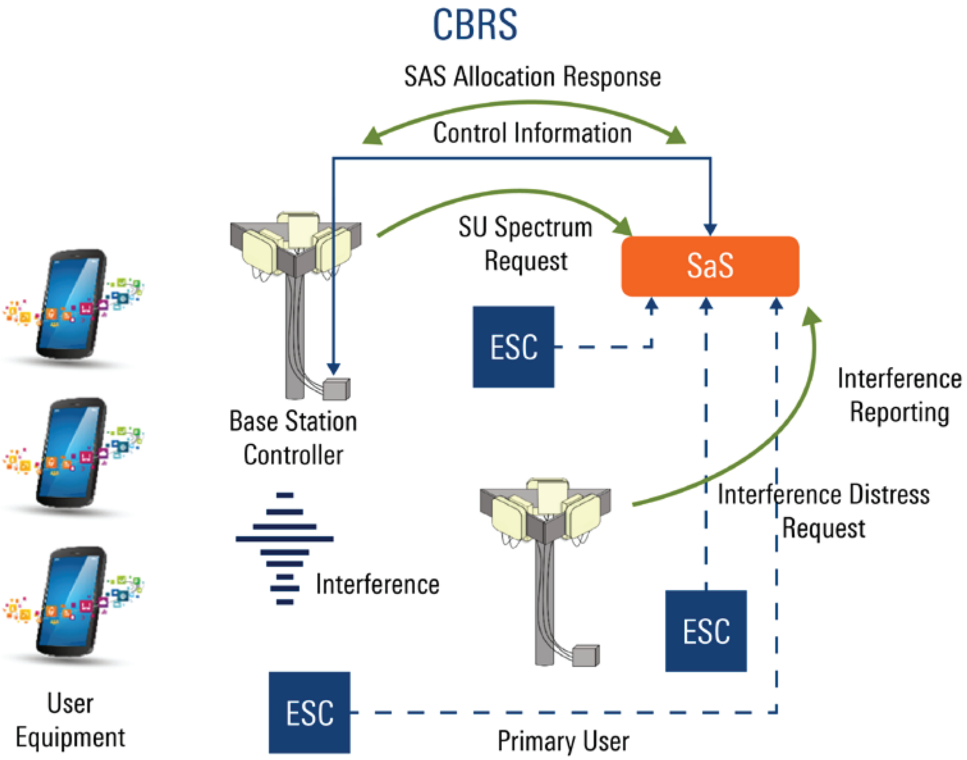
As time has progressed, the ability to transition spectrum to a different exclusive use is becoming much more difficult. The legacy users of spectrum who were easy to move to a different band or transport mechanism have typically been moved. Hence, new models have arisen on how to better leverage spectrum that is encumbered with legacy users.

One model is unlicensed, license-free, or license-exempt access, such as the spectrum used in WiFi. Analogous to real estate, license-free bands represent areas of public commons, such as a recreational park. Anyone can use this area, but they must live with others swho may be in the park at the same time. Likewise, license-free spectrum users must contend with the interference that other users in that band may be causing to their communications.

A more modern approach is spectrum sharing – non-exclusive use of the spectrum. With better-defined access rights, shared-spectrum avoids interference through processes that harmonize co-existence with other users. The real estate analogy is making a reservation at a park instead of just showing up and hoping for space. Spectrum sharing approaches started with “white space” sharing of TV bands with fixed wireless communication systems used to support rural broadband [Ref: FCC, “White Space,” found at https://www.fcc.gov/general/white-space]. A database approach is used to allocate spectrum to wireless users of these TV bands by checking if their transmission could cause interference to television stations in the region. If the wireless user location and frequency are not expected to cause interference, that transmission is allowed. White space communications were the first generation of spectrum sharing techniques. This type of sharing is found in many countries, including the United States. In other countries, spectrum sharing may be allowed using another process. For example, instead of an automated approach that uses a database to avoid conflicts, this process is done manually by a regulatory agency to provide long-term regional spectrum access and avoid interference situations.

In the United States, the next generation of spectrum sharing systems is the Citizens Broadband Radio Service (CBRS), a spectrum sharing regime used in the 3.55-3.7 GHz region. Here, the spectrum is shared using three tiers of users: incumbent users, priority access license (PAL) users, and general authorized access (GAA) users. Unlike "white space," this approach does allow for the mobility of users. Incumbent users have the highest or the first level of rights to use the spectrum. PAL users obtained the 2nd level of rights from an auction held in 2020, giving them rights over some defined interval (3-10 years) and some regions. GAA users can use the spectrum for free if the spectrum is not utilized by legacy or PAL users. The resolution of the conflict between users occurs at the database, the Spectrum Access System (SAS), with the help of RF Environment Sensing Capability (ESC) monitoring a region feeding spectral usage to the database and information about other legacy users provided by the FCC. For a PAL or GAA user, permission for the access point or base station to use the spectrum and the spectrum's allocation is provided by the database. Such an approach allows incumbent users such as radar systems to dynamically operate as needed and without a significant probability of interference. In the case of CBRS, incumbent users are typically high-power military radar systems.

Figure 2.2. CBRS System Operations



The CBRS model is powerful because it allows new business models to form; companies can acquire the CBRS spectrum to support enterprise networks to support smart cities, industrial plants, and hospitals with a level of certainty for spectrum availability. Some companies may acquire the shared spectrum to serve as a neutral host, augmenting traditional service providers' spectrum to satisfy a surge in demand to the providers' networks or even support new service providers' entry. The system supports a competitive environment for engaging new service providers such as cable TV companies. CBRS can be used as a lower-cost alternative to a distributed antenna system (DAS) to fill in a building's coverage gaps. To carry the real estate analogy along, the CBRS spectrum is the equivalent of leasing real estate instead of buying, adding a new approach to managing the financial costs of spectrum that had not been possible before. In the future, we may even see the equivalent of real estate investment trust (REIT) to give diversification to spectrum investments.

More recently, the FCC has adopted a hybrid of the unlicensed and "white space" approach to managing the U-NII-5 (5.925-6.425 GHz) and U-NII-7 (6.525-6.875 GHz) bands to operate outdoors or indoors with similar power levels as permitted for unlicensed portions of the 5 GHz band using an automated frequency coordination (AFC) system. This AFC system, protecting incumbent fixed microwave radio and radio astronomy observatories, follows the pattern of the "white space" approach, creating exclusion zones to protect the incumbents while providing frequency and power ranges for operations.

No doubt that other bands will be transitioned in the future using spectrum sharing as the final or intermediate solution to complete spectrum transition. Each transitioned band has its own characteristics, implying that each band will have to have unique ways to perform spectrum sharing. Nevertheless, the database approach to spectrum management seems to be the way of the future until research reveals more autonomous and safer ways to share spectrum.

3 DENSIFICATION & COVERAGE EXTENSION TRENDS

When a new generation of cellular technology is launched, services associated with such technology may initially be offered in limited geographic areas and then expanded gradually over a multi-year period to large areas including nationwide and continentwide coverage. Unless a new operator is carrying out “greenfield deployment,” an existing infrastructure is reused to the extent possible. For example, the facilities (e.g., towers and buildings) where the radio equipment of a previous generation cellular technology is residing are upgraded to accommodate the newer generation cellular technology. Deployment of 5G on top of 4G LTE has often occurred with such approach. There are special coverage extension or enhancement needs such as in-building coverage in large venues (e.g., large exhibition halls and convention centers). Similarly, there are special capacity needs in densely populated areas, requiring densification of the network (i.e., relatively more cells per unit area) to increase the available capacity. While traditional coverage extension and densification means such as Distributed Antenna Systems, Fixed Wireless Access, and small cells can be utilized with 5G, newer infrastructure trends are also emerging. These newer trends include enhanced antenna technologies, Integrated Access and Backhaul (IAB), and Vehicle-to-Everything (V2X) communications infrastructure.

Section 3.1 summarizes how antenna technologies can help with capacity needs and coverage needs. The traditional methods of small cells, FWA, and DAS are briefly explained in Sections 3.2, 3.3, and 3.4, respectively. Section 3.5 describes how IAB can facilitate deployments in situations where the physical fiber network is not widely available for the backhaul connectivity between the radio network and the core network. Section 3.6 illustrates the infrastructure enhancements needed for V2X communications.

3.1 Enhancements in Antenna Technologies

Antenna technologies can help address coverage needs as well as capacity needs. Transmit diversity and receive diversity antenna techniques improve coverage reliability. For example, the transmit diversity scheme of Space Frequency Block Coding (SFBC) in LTE improves the coverage reliability by exploiting space diversity (i.e., the use of multiple transmit antennas) and frequency diversity (i.e., the use of different frequencies or subcarriers). In contrast, the use of multiple receive antennas results in receive diversity, where multiple propagation paths between a given transmit antenna and the receiver are created and the receiver can overcome even a deep fade on a given propagation path. Furthermore, spatial multiplexing techniques such as Multiple Input Multiple Output (MIMO) increase capacity or throughput by enabling the reuse of the same time and frequency radio resources on multiple spatial layers. For example, compared to the single transmit antenna and single receive antenna case, throughput can theoretically double when two transmit antennas and two receive antennas are used as part of a (2x2) Single User- MIMO (SU-MIMO) technique.

While passive antennas have been quite common prior to 5G, active antennas have been gaining further popularity with 5G. In the traditional passive antenna system, a given transmit antenna consists of an array or set of passive antenna elements separated by a small space in the antenna closure called the radome. In contrast, active antennas refer to the generalized concept of having an antenna or even antenna array directly connected to the RF frontend that includes active components (e.g., power amplifiers). In general, characteristics of the overall antenna radiation pattern can be modified by adjusting the amplitude and the phase of the RF signals associated with active antenna elements. High-gain beamforming can be achieved in the millimeter-wave (mmW) deployment because many antenna elements (e.g., hundreds or even more than a thousand) can be housed in a small antenna radome. Since the mmW spectrum has huge propagation path losses and hence may have difficulty providing good in-building coverage, the high-gain massive MIMO beamforming can be exploited to overcome a huge deficit in the propagation path loss. High-gain beamforming is a key contributor toward making the commercial mmW deployments a reality.

5G does not have a separate diversity technique such as SFBC like LTE but has high-performance beamforming for enhanced reliability compared to SFBC. 5G defines flexible precoding in support of various antenna techniques. Furthermore, precoding can be applied to the Reference Signals that accompany the shared channel carrying user-specific signaling messages and traffic. 5G supports SU-MIMO, where time-frequency resources are reused on the transmit-receive paths between the gNB and a given UE. Additionally, 5G supports Multi-User MIMO (MU-MIMO), where time-frequency resources are reused between the gNB and different UEs (e.g., the same radio resources are given to two different UEs in the cell).

Release 15 supports relatively wide beams for Synchronization Signal/Physical Broadcast Channel Blocks (SSBs). Furthermore, Release 15 supports flexible Channel State Information- Reference Signal (CSI-RS). For example, beams for the CSI-RS can be narrower than the beams for the SSBs for more focused information transfer between the gNB and the UE in a beam, resulting in beam-specific reference signal. Different numbers of SSB beams are supported in different frequency ranges. For example, in Release 15, a maximum of four SSB beams are supported below 1 GHz and a maximum of 64 beams are supported in the mmW spectrum. Release 15 supports up to 8 spatial multiplexing layers for MIMO for 5G NR.

5G supports flexible transceiver architectures for analog beamforming, digital beamforming, and hybrid beamforming. Advanced signal processing techniques are utilized to enhance the performance of 5G radio interface. Release 15 defines a flexible codebook with the support for up to 32 (logical) antenna ports. Type I codebook supports standard-resolution MIMO operations and Type II codebook supports high-resolution MU-MIMO operations. A higher-resolution codebook enables finer beamforming at the expense of increased complexity.

3GPP introduced several MIMO enhancements in 5G after Release 15 [5G_Americas]. For example, efficiency is increased by reducing the overhead in the frequency-domain. Type II codebook for MU-MIMO is extended to support a rank greater than two. Multi-Transmission Reception Point (TRP)/Multi-Panel transmission enhancements are introduced to support ideal backhaul and non-ideal backhaul and inter-cell and intra-cell multi-TRP transmission. Multi-beam operations are enhanced for FR2 to reduce latency and overhead and to support beam failure recovery in an SCell. Additionally, full-power uplink transmissions with multiple power amplifiers are supported. Furthermore, new low-Peak-to-Average Power Ratio (PAPR) reference signals are defined for the downlink and the uplink.

3.2 Small Cells

Small cells enable reuse of the same spectrum more times per unit area compared to large cells, significantly increasing the available capacity per unit area. Furthermore, since the UE power is limited, coverage reliability improves as well.

When higher frequency bands are used, larger propagation path losses naturally lead to relatively smaller cells. The use of mmW spectrum works in 5G results in smaller cells for a given amount of fixed transmit power. Additionally, since the mmW spectrum has a significantly larger amount of spectrum available, more radio resources are available at the mmW spectrum compared to lower frequency bands. Hence, higher capacity and throughput can be achieved at higher frequency bands; although cell coverage shrinks, requiring more cells to contiguously cover a given geographic area.

Small cells are deployed for sub-6 GHz spectrum per operator choice. In general, relatively higher frequencies make it easy to contain the RF energy in a given small cell.

The use of unlicensed spectrum also naturally leads to small cells, because such spectrum imposes relatively lower power limits. For example, while 30 dBm power limit is often used in the 5 GHz unlicensed spectrum, traditional cellular networks in the licensed spectrum would often exceed 55 or 60 dBm power. The use of lower power limits restricts the cells to be small in the unlicensed spectrum deployments.

3.3 Fixed Wireless Access

Fixed Wireless Access (FWA) systems have been around for quite some time. As their name reveals, they provide Internet access in a wireless fashion to residential homes and enterprises. The superior performance of 5G has caused a resurgence of FWA systems. Indeed, some of the initial 5G deployments involved fixed (and not mobile!) broadband services. The flexible and high-performance beamforming in 5G can be exploited to focus energy toward buildings for enhanced coverage and toward the areas where mobile traffic is. The superior performance of 5G enables a service provider to support ultra-high-speed broadband connectivity to customers, potentially at a lower cost.

5G-based FWA systems can be used by a cellular service operator to compete with cable and DSL operators that run fixed networks. Furthermore, 5G can offer high-speed mobile broadband services where broadband services are not currently available such as unserved areas in developed or developing countries, where installation of a cable/fiber network is not economically viable. While a fixed network requires a wireline connection to each premise, an existing cellular network infrastructure can be enhanced to support fixed broadband users. 5G enables faster deployment of fixed broadband services. Furthermore, 5G radio networks and/or 5G Core can be shared between fixed and mobile broadband users for efficient operations.

An FWA system may support different types of a Customer Premise Equipment (CPE). The CPE acts like a UE or smartphone and connects in a wireless fashion to a 5G gNB using the 5G NR radio interface. The CPE typically supports Wi-Fi to enable various devices in a home or an enterprise building to access the Internet. The CPE may be an indoor device, in which case there can be a large building penetration loss depending upon the structure of the building. The CPE may have a built-in outdoor directional antenna (e.g., with a 10 to 14 dB gain) that avoids any building penetration loss. The CPE antenna can potentially be oriented toward a 5G gNB to optimize the CPE-gNB radio link. An FWA system utilizes a CPE management system that allows the operator to log in to devices, configure them, and check their status remotely.

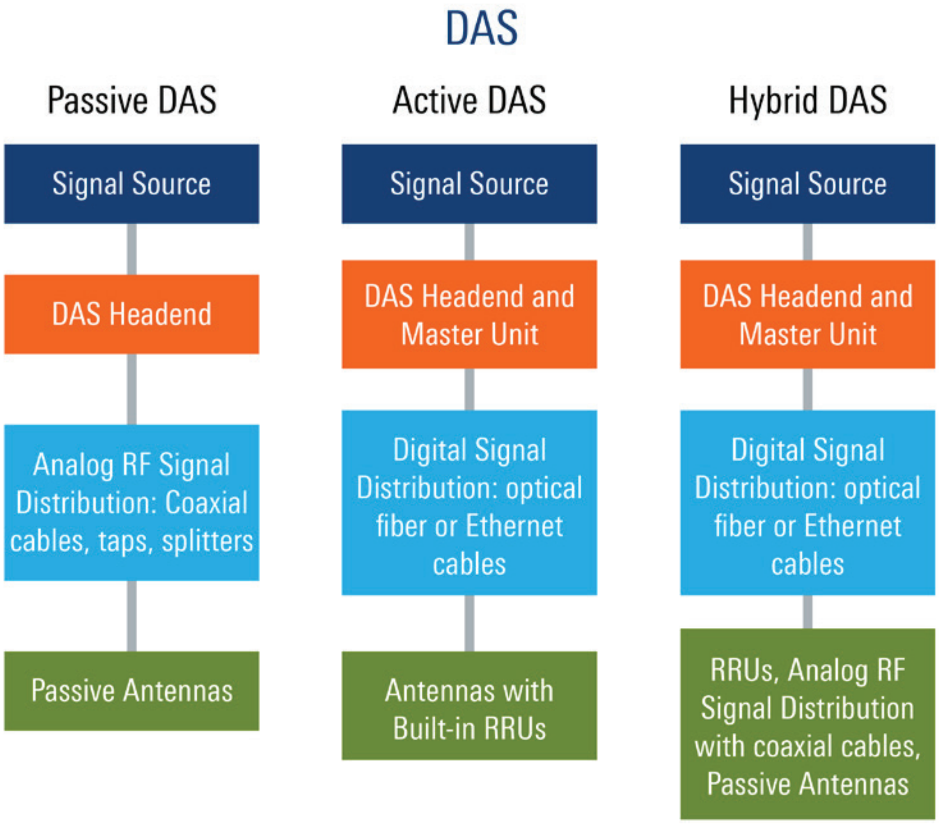
The choice of spectrum to be used between the CPE and the gNB is one of the important design considerations for the FWA systems. A lower frequency band provides better coverage due to smaller propagation path loss but limits the achievable throughput due to the smaller amount of available spectrum. In contrast, a higher frequency band such as the mmW spectrum provides smaller coverage due to larger propagation path loss but very high throughput due to larger amount of available spectrum.

3.4 Distributed Antenna System (DAS)

A Distributed Antenna System (DAS) has been widely used to extend coverage to large venues such as stadiums, convention centers, and multi-storied buildings. The signals between the traditional macro gNB and users in such large venues may be too weak for reliable communications. Hence, one possible solution is to exploit a DAS. Another alternative to cover large venues is to deploy small cells. For a given coverage area, a DAS can make use of a single backhaul connection or fewer backhaul connections between the radio network and the core network, while small cells would need a backhaul connection per cell or gNB depending on the small cell configuration. The small cell approach usually provides a higher capacity or throughput per unit area than a DAS in the target service area.

Figure 3.1 illustrates three major types of a DAS- passive, active, and hybrid [DAS_Waveform].

Figure 3.1. Major Types of Distributed Antenna Systems (adapted from [DAS_Waveform]).


DAS
Passive DAS
Active DAS
Hybrid DAS
Signal Source
Signal Source
Signal Source
DAS Headend
DAS Headend and Master Unit
DAS Headend and Master Unit
Analog RF Signal Distribution: Coaxial cables, taps, splitters
Digital Signal Distribution: optical fiber or Ethernet cables
Digital Signal Distribution: optical fiber or Ethernet cables
Passive Antennas
Antennas with Built-in RRUs
RRUs, Analog RF Signal Distribution with coaxial cables, Passive Antennas

For the network-to-the device link information transfer, a DAS utilizes a signal source to obtain a cellular signal. The DAS processes such signal in a DAS headend and possibly a Master Unit and evenutally distributes an analog or digital signal to antennas using a distribution network. In the device-to-the-network information transfer, antennas receive the uplink signals from devices and these uplink signals eventually reach the DAS headend before the DAS headend supplies the signal to the cellular network.

Let's contrast passive DAS, active DAS, and hybrid DAS for the network-to-the-device information transfer. The cellular signal is obtained by the DAS headend from a signal source. A donor antenna oriented toward the traditional gNB (e.g., the gNB of the microcellular network or microcellular network) can serve as a signal source. The signal processed by the donor antenna is "off-air" signal. The donor antenna receives the cellular RF signal from the gNB and transmits the cellular RF signal toward the gNB. This donor antenna approach works well when the signal at the donor antenna is strong. However, the decoupling between RX and TX at the donor antenna is a major challenge in this approach. A gNB can be deployed at the venue itself. In such a case, the gNB is the signal source. Since the gNB on the premise needs to be connected to the service provider's core network via backhaul, this approach requires a longer deployment time. Additionally, space, cooling, and power requirements increase. A yet another possibility for the signal source is a small cell. Small cells by themselves may be too expensive to cover a large venue. Hence, few small cells with a DAS could be an attractive solution that provides a good tradeoff among coverage, capacity, and cost.

In a passive DAS, for the network-to-the device information transfer, the DAS headend amplifies the downlink analog RF signal and provides the amplified signal to passive antennas via a distribution network. This distribution network consists of coaxial cables, taps, and splitters. At the end of the distribution are passive antennas that simply transmit RF signals to UEs. For example, different floors of a multi-storied building may have different passive antennas covering different floors. In an active DAS, for the network-to-the device information transfer, the DAS headend amplifies the downlink analog RF signal and provides the amplified signal to a Master Unit. The Master Unit converts the RF signal into a digital signal. The digital signal is distributed by the distribution network made up of optical cables or Ethernet cables. The antennas with built-in Remote Units (RRUs) process the digital signals such as by performing digital-to-analog conversion (DAC) and power amplification. The amplified RF signals are transmitted over the air by the antennas. The active DAS has a unique advantage compared to the passive DAS; it results in a smaller cable loss and less overall attenuation.

In a hybrid DAS, characteristics of both passive DAS and active DAS are combined. For the network-to-the-device information transfer, the DAS headend amplifies the downlink analog RF signal and provides the amplified signal to a Master Unit. The Master Unit converts the RF signal into a digital signal. The distribution network utilizes optical cables or Ethernet cables to send digital signals to RRUs. The RRUs perform digital-to-analog conversion (DAC) and amplifies the analog RF signal. The amplified RF signal is distributed to passive antennas using another distribution network consisting of coaxial cables.

With the advent of 5G, an additional DAS architecture becomes possible. A gNB-CU can be located in one location at the venue and gNB-DUs can be scattered across the venue to efficiently create small cells. These small cells can be integrated with a DAS for achieving the desired tradeoff among coverage, capacity, and costs.

3.5 Integrated Access and Backhaul (IAB)

Integrated Access and Backhaul (IAB) is a mechanism for base stations to do self-backhauling using a wireless link at the same frequency or at a frequency different from that used to communicate with the UEs via the wireless access link [Tripathi19]. An IAB base station or IAB-Donor gNB, can service UEs in this region and act as a backhaul to other gNBs or IAB-Donor gNBs, as shown in Figure 3.2. Implementation of IAB is trickier than one might expect because of interference potential with the backhaul. The sharing of resources between the backhaul link and wireless access link can constrain data throughput and latency. This potential for backhaul interference is mitigated through complex scheduling and beamforming.

The initial study leading to IAB came out with Release 15, and the physical layer of IAB was specified in Release 16; higher layer protocols were part of Release 16. Refinements to IAB are expected in Release 17 and beyond.

IAB offers many benefits, but most importantly, it can reduce cost by eliminating the need for fiber backhaul (or connectivity to the radio node, fronthaul). It can also function with the same Operations and Management system as the rest of the 5G network since the radio heads serve both the access link and the backhaul link. Cost reductions are particularly impactful for extending range and coverage in rural areas and for mitigating the obstructions in non-line-of-sight cases that occur at higher frequencies. This improved coverage is also very valuable for deploying public safety systems that need universal coverage. Resilience against the loss of a link is essential too, and IAB allows for routing to be changed when a link is lost. Another benefit is the dynamic deployment of cells, where vehicle-based base stations (so-called “Cells on Wheels”) can be deployed on the fly for coverage extensions.

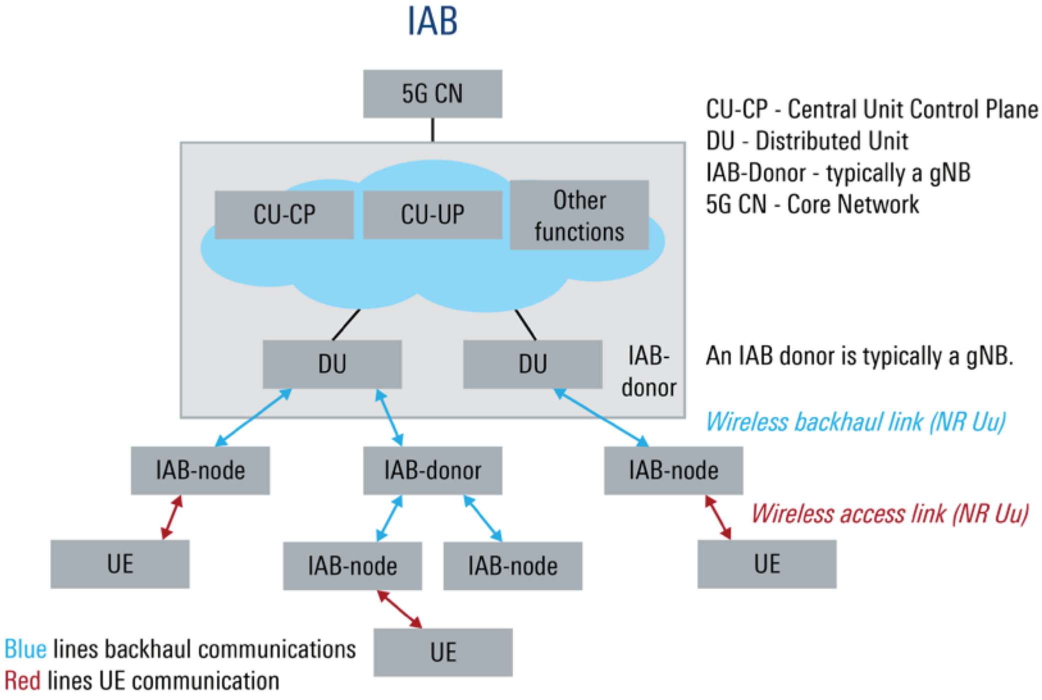
A relay approach similar to IAB was introduced for LTE in Release 10, but it never gained much popularity. The reasons are multi-fold, but the key problem was that the LTE relay function used the same valuable spectrum as that was used to communicate with the UEs. Second, small cells were not as popular in 4G as had been anticipated, and the need for fiber-less backhaul never materialized to the extent that had been anticipated.

However, for 5G, small cells are expected to be the norm, especially for mmW systems, and backhaul for those numerous small cells needed because of propagation limitations is problematic. Second, 5G can leverage mmW spectrum, and that spectrum has enough bandwidth to support access to the users as well as reach other gNBs (i.e., IAB-nodes). Beamforming, an inherent feature of 5G, effectively provides additional capacity for supporting both the backhaul link and the access link.

Furthermore, rural broadband is becoming more important because extensive wireless deployment is anticipated for improving rural broadband infrastructure. IAB is a cost-effective way to provide this coverage.

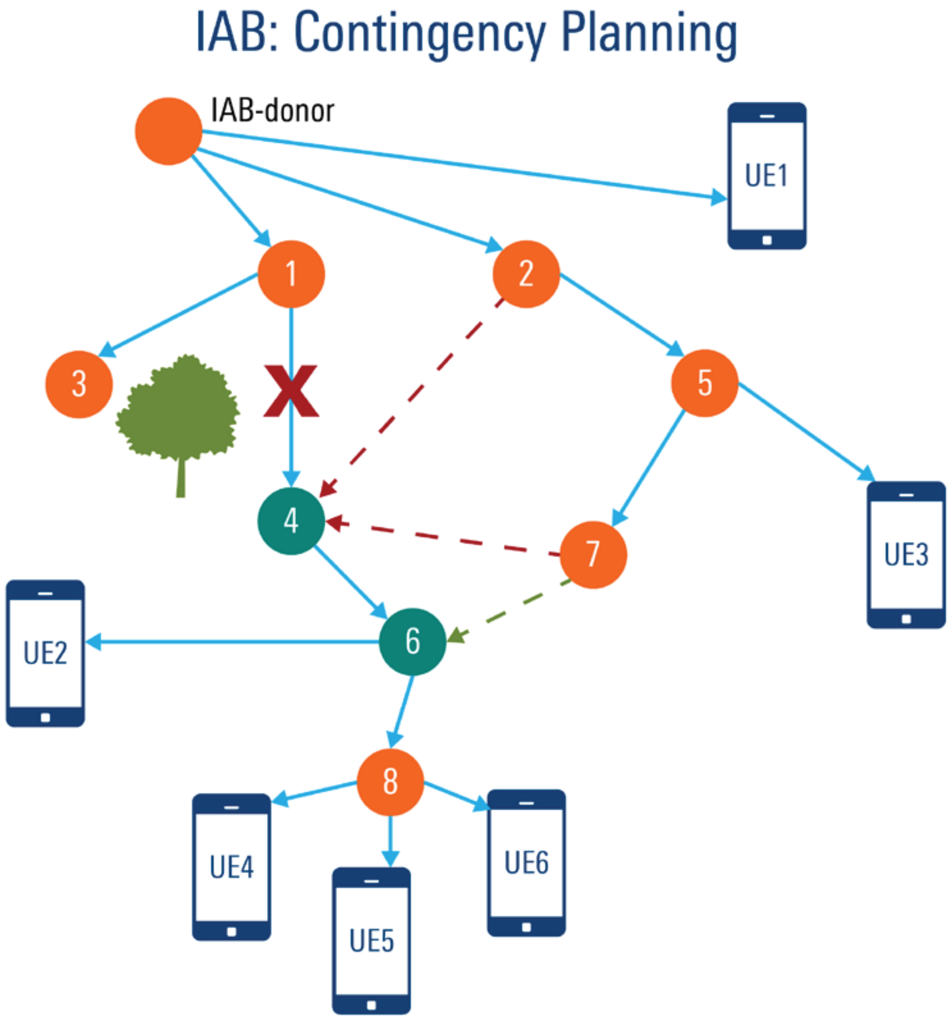
The architecture for an IAB network is shown in Figure 3.2 below. The IAB-doner gNB can act as a relay or as an access point to UEs. One can chain IAB-donor nodes together. Contingency routing can help with link disruption by selecting a different IAB-doner node. QoS management also helps on the backhaul links helps with reliability. Interference measurements are made to interference between the backhaul link and access link, and network synchronization also helps reduce interference for time-division duplex (TDD) systems. There is flexibility in the allocation of the spectrum between the backhaul and the access link to handle variations in traffic.

Figure 3.2. Architecture of an IAB-enabled network.


IAB
5G CN
CU-CP - Central Unit Control Plane
DU - Distributed Unit
IAB-Donor - typically a gNB
5G CN - Core Network
CU-CP
CU-UP
Other functions
DU
DU
IAB-donor
An IAB donor is typically a gNB.
Wireless backhaul link (NR Uu)
IAB-node
IAB-donor
IAB-node
Wireless access link (NR Uu)
UE
IAB-node
IAB-node
UE
UE
Blue lines backhaul communications
Red lines UE communication

Multi-hop routing capability for IAB is illustrated in Figure 3.3. Here the link between nodes 1-to-4 is disrupted. The links 2-to-4 or 7-to-4, along with 7-to-6, could be used to replace links 1-to-4. The selected replacement link can depend on many factors, including maintaining QoS features such as latency or throughput.

Figure 3.3: Contingency Planning in IAB: Rapid Link Replacement.



Release 17 will focus on making IAB more robust, spectrally efficient, resilient to backhaul/access link interference, and better multi-hop latency and end-to-end performance. There are also expected to be improvements in scheduling and fairness, congestion control, load balancing, and support for dual connectivity.

Beyond Release 17, there is still plenty of room for improvement in topology optimization, network coding, adaptive routing for a mesh-based network, mobile IAB-doner nodes, and the use of intelligent surfaces. Given the anticipated demand for IAB, likely, the features of IAB will continue to improve. IAB serves as an example that even though technology is introduced by 3PP, such as LTE relay function, it may take several releases before the approach becomes viable or desirable for commercial deployment.

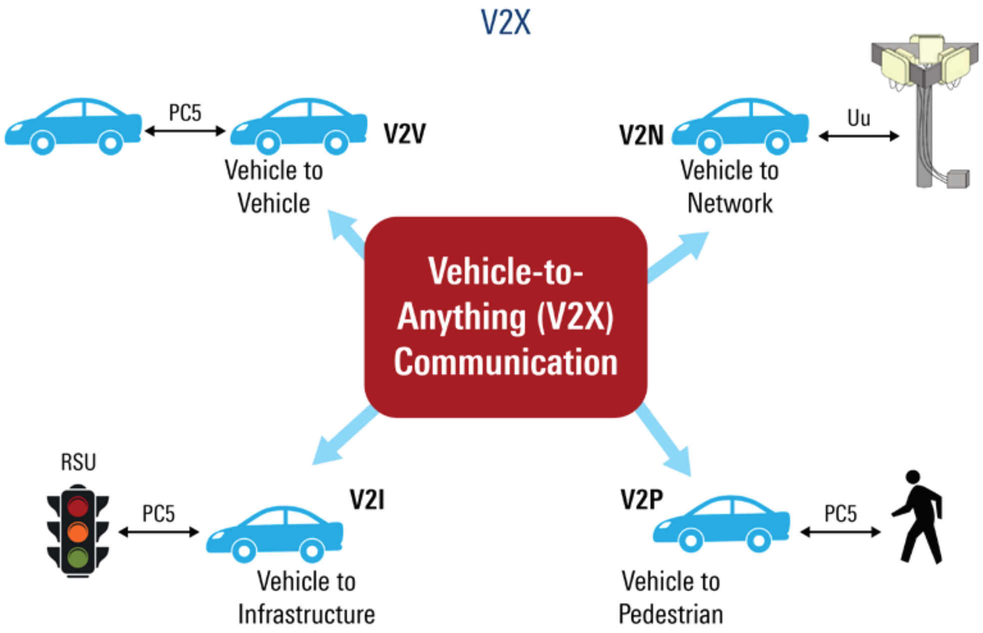
3.6 Infrastructure Enhancements Supporting V2X Communications

Connected vehicles with automated driving represent one of the emerging technology trends in modern times. Automated driving brings advantages such as increased safety, higher fuel efficiency, and reduced traffic congestion. The Intelligent Transport System (ITS) band around 5.9 GHz is widely available globally, and some countries (e.g., China) have mandated V2X services. In the U.S., the FCC had initially designated a 75 MHz spectrum around 5.9 GHz for Dynamic Short Rage Communications (DSRC) to support communications for improved traffic safety. Some systems were designed to operate in such DSRC spectrum using the IEEE 802.11p-based physical and MAC layers. However, in November 2020, the FCC repurposed the spectrum to allow 45 MHz for Wi-Fi and the remaining 30 MHz for auto safety where technologies such as Cellular- Vehicle-to-Anything (C-V2X) can be deployed. C-V2X provides a 360° Non-Line of Sight (NLOS) vision that goes well beyond a driver's vision of sensors residing on a given vehicle. C-V2X provides information about the weather, nearby accidents, road conditions road construction warnings, the arrival of emergency vehicles, and the activities of other drivers/vehicles. C-V2X complements sensor technologies that use radar, camera, and Light Detection And Ranging (LiDAR).

3GPP defined LTE-based C-V2X communications in support of automated driving in Release 14. Release 14 provides a foundation for C-V2X technologies, where cellular technology such as LTE is adapted to meet the requirements of V2X communications. For example, LTE enhanced the Device-to-Device (D2D) Communications feature originally defined in Release 12 to support the C-V2X technology. Example enhancements included improved signal design to facilitate better synchronization and channel estimation, the transmission of control and data in the same subframe, and efficient resource allocation [V2X_Qualcomm]. Release 15 introduced enhancements such as transmit diversity and the QoS of 10 ms Packet Delay Budget. Release 16 supports 5G NR-based C-V2X, and Release 17 continues to enhance 5G NR-based C-V2X (also known as NR-V2X). Note that LTE-V2X and NR-VTX are complementary; they do not substitute each other. NR-V2X builds on LTE-V2X and enables additional uses cases that benefit from reduced latency and higher throughput of 5G NR.

Figure 3.4 illustrates four types of communications supported by C-V2X: Vehicle-to-Vehicle (V2V), Vehicle-to-Pedestrian (V2P), Vehicle-to-Infrastructure (V2I), and Vehicle-to-Network (V2N).

Figure 3.4: C-V2X Communication.



The three types of communications- V2V, V2P, and V2I- utilize the PC5 interface between two UEs. In contrast, the V2N communication utilizes the Uu interface between the UE and the network. The radio link used on the PC5 interface is called the sidelink. The V2V communication is intended for safety and the Advanced Driver Assistance System (ADAS). The V2P communication ensures the pedestrian's safety. The V2I communication enables the vehicle and the roadside infrastructure such as the traffic signals, parking spaces, and toll collection systems to communicate with each other. The V2N communication provides over-the-top cloud-based services.

The V2I communication aspect of V2X communications involves deploying new communications infrastructure such as Road Side Units (RSUs) with which On-Board Units (OBUs) of the vehicles communicate. The RSU would need to be connected to the backend of the Intelligent Transport System (ITS) infrastructure. The RSUs would need to support the PC5 interface defined by the 3GPP. The RSU can also act as a small cell and provide NR-Uu-based access to UEs. Furthermore, while lower layers such as Layer 1 and Layer 2 would make use of 3GPP-defined LTE and 5G specifications, upper layers make use of standards such as the IEEE 1609 family of standards (e.g., to provide higher-layer security).

Due to its superior performance capabilities, such as peak data rates on the order of Gbps and latencies of just few milliseconds, 5G is expected to enable a new set of use cases for automated driving.

4 VIRTUALIZATION AND CLOUDIFICATION

Various Network Functions or Network Elements of a 5G network can be implemented using tightly-coupled dedicated hardware and software. For example, similar to 3G networks and early LTE networks, a physical piece of equipment can act as a base station such as the 5G Network Function of the gNB. However, the 5G network is defined by 3GPP such that the implementation of the network can benefit from virtualization technologies. Section 4.1 explains the motivation behind virtualization. Key virtualization technologies such as Network Function Virtualization (NFV), Software-Defined Networking (SDN), and orchestration are discussed in Section 4.2. Finally, Section 4.3 illustrates the RAN architecture enhancements such as gNB-Central Unit (gNB-CU) and gNB-Distributed Unit (gNB-DU).

4.1 Motivation for Virtualization

Virtualization makes use of suitable software to implement the Network Functions using a cloud infrastructure. Examples of cloud infrastructures include Amazon AWS (Amazon Web Services), Microsoft Azure, and Google's Cloud Platform, Alibaba Cloud, and IBM. While service providers could create their own cloud infrastructure, they often team up with providers of the cloud infrastructure. For example, AT&T has collaborated with Microsoft to implement the edge computing aspect of 5G using the Microsoft's Azure cloud infrastructure. Verizon has collaborated with Amazon to implement the edge computing aspect of 5G using the Amazon's AWS cloud infrastructure.

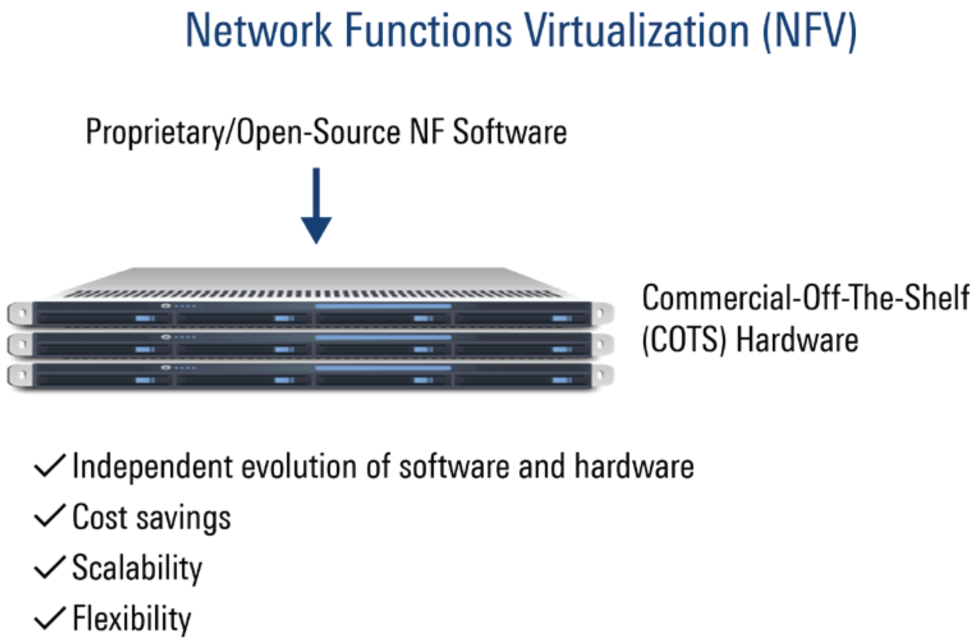
Cloud infrastructure is a physical infrastructure that includes computing, storage, and networking resources. Compute resources involve processors that carry out desired processing (e.g., software for a given 5G Network Function such as the AMF). Storage resources include memory units that store information such as information about wireless subscribers and their user traffic sessions based on the needs of a given Network Function (NF). Finally, networking devices enable the information such as a user's IP packets to travel from one NF to another NF (e.g., from the gNB to the UPF).

Virtualization offers benefits such as cost savings, scalability, and flexibility. A service provider can purchase different virtualized Network Functions (i.e., software packages for different NFs) to minimize the total cost of ownership. It is much faster to scale up a virtualized network by adding more resources of the cloud infrastructure to deploy more NFs compared to scaling up a network that has NFs implemented on dedicated physical platforms. The service provider also has more flexibility in mixing and matching NFs from different vendors based on factors such as cost and performance.

4.2 Fundamental Virtualization Technologies

Key fundamental technologies related to virtualization include Network Function Virtualization (NFV), Software-Defined Networking (SDN), and orchestration [R1].
Figure 4.A shows the key idea behind NFV.

Figure 4.A. NFV: Concept and Benefits.


Network Functions Virtualization (NFV)
Proprietary/Open-Source NF Software
Commercial-Off-The-Shelf (COTS) Hardware
✓ Independent evolution of software and hardware
✓ Cost savings
✓ Scalability
✓ Flexibility

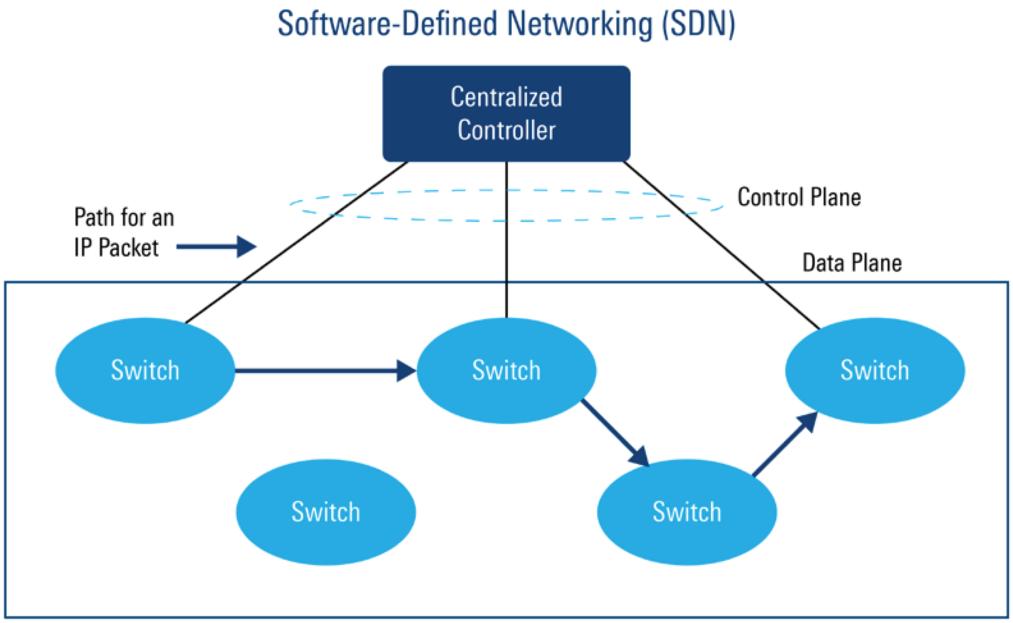
A Physical Network Function (PNF) utilizes proprietary software that is tightly integrated with specific hardware. The only functions performed by such hardware are the functions of a specific Network Function. For example, the AMF implemented as a PNF can only perform functions of the AMF. NFV deploys the NF software on generic Commercial-Off-The-Shelf (COTS) hardware. Such NF software may be open-source software or proprietary software of a vendor. A Virtualized Network Function (VNF) does not have tight coupling between software and hardware. The same software can be deployed on any other hardware. Additionally, one piece of hardware, such as a processor, can implement multiple NFs.

NFV has several advantages. Since hardware and software are decoupled, both can evolve independently from each other. Hence, if a better processor (e.g., a higher-performance processor) becomes available, the existing processor can be replaced. Similarly, if a new software version is available, it can be deployed using the existing hardware. Since a service provider can make use of COTS hardware, the operator gains an economies-of-scale advantage. The network capacity can be scaled up or down by adding or removing the cloud infrastructure resources based on the needs. The separation of hardware and software also provides significant flexibility. Furthermore, NFV also provides implementation flexibility; one processor can support multiple NFs when needed and different processors can be used for different NFs based on performance requirements of NFs and capabilities of processors.

A Software-Defined Networking (SDN) framework can identify optimal routes for IP packets and forward IP packets cost-effectively. A typical IP router implements both a Control Plane (CP) and a Data Plane (DP) (equivalent to the term "User Plane," which is a term widely used in wireless communications). The Control Plane enables the router to exchange signaling messages with neighboring routers to create a routing table. Protocols such as Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) are used for CP communications between IP routers to determine optimal routes toward destinations. The User or Data Plane of an IP router makes use of the Routing Table to forward an incoming packet to a suitable next hop router so that the overall path for such IP packet is optimal. A typical IP router can be quite complex because it may implement a large set of features. The SDN separates the CP and the DP functions of IP routers for enhanced performance and reduced costs.

Figure 4.B illustrates how SDN can be utilized in a virtualized network.

Figure 4.B. Software-Defined Network: Overall Framework.



Two main components of SDN are an SDN controller and an SDN switch. The SDN controller centralizes intelligence of the CP of traditional IP routers. The SDN controller determines optimal routing paths and provides such paths to SDN switches. Since the SDN controller is centralized, it has a more comprehensive view of the IP network and makes better routing decisions. For example, when one part of the network is congested, the SDN controller can promptly determine new routing paths and inform SDN switches about new optimal routing paths. The SDN switches are low-complexity packet forwarding Data Plane devices.

SDN can be used to connect one NF of a 5G network to other NFs in a dynamic and automatic manner. Moreover, as new NFs are added to a 5G network, the SDN framework makes it easy to establish connectivity among those NFs.

The main benefits of SDN are low cost, enhanced routing performance, and scalability. Since SDN switches are low-complexity packet forwarding devices, they are much less expensive than full-fledged IP routers. The centralization of intelligence in the SDN controller provides a more comprehensive view of the network traffic and hence the overall routing performance improves relative to decentralized IP routers. Since SDN dynamically determines optimal routing paths, manual configurations are minimized, improving the scalability of the connectivity solution.

Orchestration is a framework to launch and maintain services in a highly automated manner [R1]. Orchestration can make use of multiple orchestrators. An orchestrator may be single-tenant or multi-tenant. A single-tenant orchestrator serves a single tenant (e.g., an enterprise customer), and, hence, multiple single-tenant orchestrators are needed to serve different tenants. In contrast, a multi-tenant orchestrator can serve multiple tenants.

When offering a new service such as the 5G eMBB service to customers requires multiple steps to be executed. These steps include equipment procurement, equipment configuration, testing, service launch, performance monitoring, troubleshooting, and optimization. The traditional approach that executes all these steps is slow and prone to errors due to manual configurations and inter-department/inter-organization transitions during the service deployment. Service orchestration automates the workflow of various processes to instantiate new services and maintain deployed services.

Service orchestration makes use of NFV orchestration and SDN orchestration. NFV orchestration automates the resource management of the Virtualized Network Functions (VNFs) and the NFV Infrastructure (NFVI) (which is the cloud infrastructure). For example, NFV Orchestrator uses the help of the VNF Manager (VNFM) to manage VNFs and the Virtualized Infrastructure Manager (VIM) to manage the NFVI. OpenStack is an example platform to implement VIM. SDN orchestration automates network connectivity and network monitoring. For example, an SDN orchestrator facilitates the connectivity between the AMF and the SMF in a 5GC. Service orchestration can be implemented using Open Network Automation Platform (ONAP), an overall framework for modeling, orchestration, and analysis of services. Benefits of orchestration include service velocity, simplified equipment interoperability and integration, and reduced CapEx and OpEx.

4.3 RAN Evolution- Aggregation or Disaggregation?

In 3G networks, the Radio Access Network (RAN) was centralized, with a Base Station Controller or a Radio Network Controller controlling hundreds of Base Stations. 4G LTE introduced a flat or distributed RAN architecture with only eNBs (i.e., LTE base stations) in the RAN; a centralized controller is absent. An LTE eNB in initial LTE deployments utilized (i) a Base Band Unit (BBU) that implements all the layers of the radio protocol stack except the lower physical layer and (ii) a Radio Unit (RU) that implemented the lower physical layer. Such eNB configuration is available in recent deployments as well. The RU consists of components such as RF filters and High-Power Amplifier (HPA) and connects to the antenna. The BBU and the RU can be housed in the same cabinet or can be separated by a short distance such as few meters. The BBU-RU interface utilizes an optical fiber and is managed by a protocol called Common Public Radio Interface (CPRI).

The BBU-RU interface carries a digital baseband signal. LTE (as well as 5G) utilizes OFDMA in the downlink. The baseband OFDMA signal can be represented as a series of complex-valued samples. For example, the baseband OFDMA signal consists of 2048 complex-valued samples in the OFDMA symbol period (1/subcarrier spacing) when the channel bandwidth is 20 MHz. In the case of 15 kHz subcarrier spacing and 20 MHz channel bandwidth in LTE, during the OFDMA symbol period of (1/15 kHz= 66.6 µs), there are 2048 complex-valued samples in the baseband signal. Each complex-valued sample has an in-phase component (so-called "I-channel") and a quadrature-phase component (so-called "Q-Channel"). The in-phase component of the sample is quantized using a certain number of bits such as 15 bits, and, the quadrature-phase component of the sample is also quantized similarly. TheSC-OFD MA symbol is eventually represented in the uplink by I-channel bits of all samples and Q-channel bits of all samples during the symbol period. Additionally, some CPRI overhead is also carried on the BBU-RU interface. A protocol such as CPRI allows the transfer of control signaling between the BBU and the RU to support of the transport of bits.

The distributed LTE RAN later evolved to Centralized RAN (C-RAN), where a pool of BBUs is centralized in a local data center or C-RAN Hub close to the cell sites. As a result, the Radio Units are few miles away from the BBUs. Such RUs are called Remote Radio Units (RRUs) or Remote Radio heads (RRHs) because they are separated from their BBUs by a relatively long distances.

C-RAN yields benefits such as cost savings, reduced real-estate needs, fewer routers, and reduced electricity usage. Since a pool of BBUs can be used to control numerous RRUs and since 1:1 mapping between the RRU and the BBU is not needed, cost savings are realized. RAN deployments are facilitated due to the reduced footprint at cell sites. Since BBUs are no longer at cell sites, fewer IP routers are needed to connect BBUs to the core network via the backhaul. Centralization of BBUs in a local Data Center or C-RAN Hub reduces the air conditioning requirements at individual cell sites, leading to reduction in utility bills. Certain advanced features such as Coordinated Multi-Point (CoMP) and dual connectivity perform better due to low-latency communications made feasible by the proximity of BBUs in a Data Center. The end-to-end latency can be further reduced by exploiting edge computing, where an Application Server is placed in or close to a Data Center.

When RAN starts using the cloud infrastructure it becomes a Virtualized RAN (VRAN). VRAN derives benefits of virtualization, because selected BBU functions can be implemented by using the cloud infrastructure that can be shared between the RAN and the core network.
In 5G, distributed gNBs, Centralized RAN, and VRAN can be utilized like LTE. Furthermore, 5G gNB can be disaggregated as illustrated in Figure 4.C.

3GPP allows the gNB to be divided into two logical components, gNB-Central Unit (gNB-CU) and gNB-Distributed Unit (gNB-DU). The gNB-CU implements the upper layers of the radio protocol stack for the Control Plane and the User Plane, while the gNB-DU implements the lower layers of the radio protocol stack for the Control Plane and the User Plane. More specifically, the gNB-CU implements Radio Resource Control (RRC), Service Data Adaptation Protocols (SDAP), and Packet Data Convergence Protocol (PDCP). The gNB-DU implements Radio Link Control (RLC), Medium Access Control (MAC), and Physical (PHY) layer. From the 3GPP perspective, the gNB-DU includes the entire PHY layer. However, the PHY layer, in an implementation-specific manner, can be further divided into Upper PHY and Lower PHY such that Upper PHY carries out baseband processing (which would be typically carried out by a BBU) and lower PHY implements RF processing (which would typically be carried by an RU/RRU). The gNB-CU can interface with other gNBs and eNBs and the core network such as the 5GC.

The gNB-CU and the gNB-DU have an F1 interface between them. The Control Plane protocol on the F1 interface is F1 Application Protocol (F1AP), while the User Plane protocol is GTP-U or GPRS Tunneling Protocol, where GPRS is General Packet Radio Service. The gNB-CU can be further disaggregated or decomposed into gNB-CU-CP and gNB-CU-UP. The interface between gNB-CU-CP and gNB-CU-UP is called the E1 interface and is managed by the E1 Application Protocol (E1AP).

Benefits of disaggregation of the gNB include flexibility, cost savings, reduced transport bandwidth requirements, and low latency. In an example implementation, a service provider can mix and match gNB-CUs and gNB-DUs (and Lower PHY components such as RUs/RRUs) from different vendors. A gNB-CU can control multiple (e.g., tens or hundreds) of gNB-DUs. The use of virtualization further reduces costs. The interface between the gNB-CU and the gNB-DU is also called mid-haul. Since raw data is sent on the F1 interface instead of quantized samples of a baseband signal, the transport bandwidth requirements are reduced. Furthermore, since retransmissions at the RLC and MAC/PHY layers occur between the user devices and the gNB-DU, the average latency can be reduced.

Figure 4.C. Disaggregated gNB in 5G.

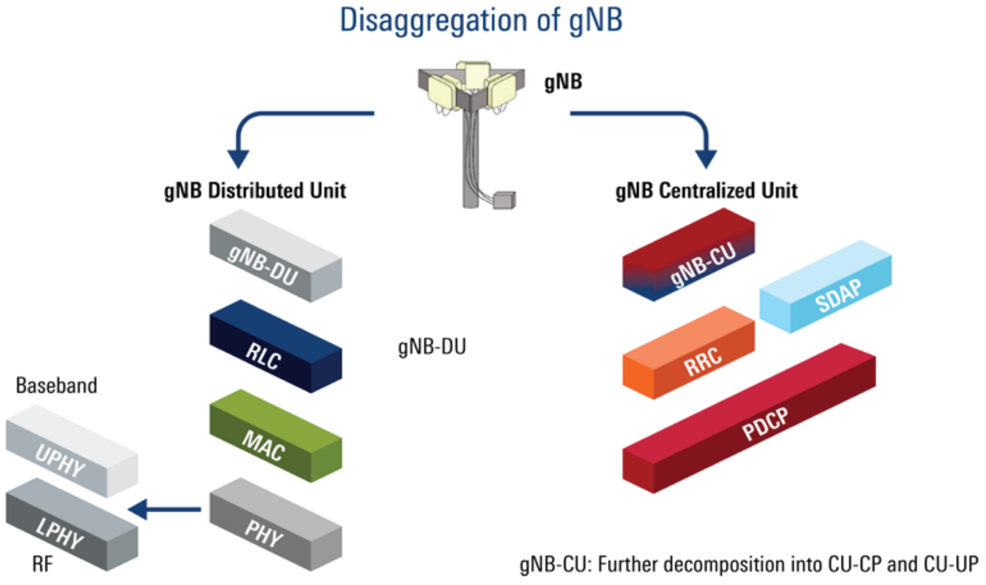
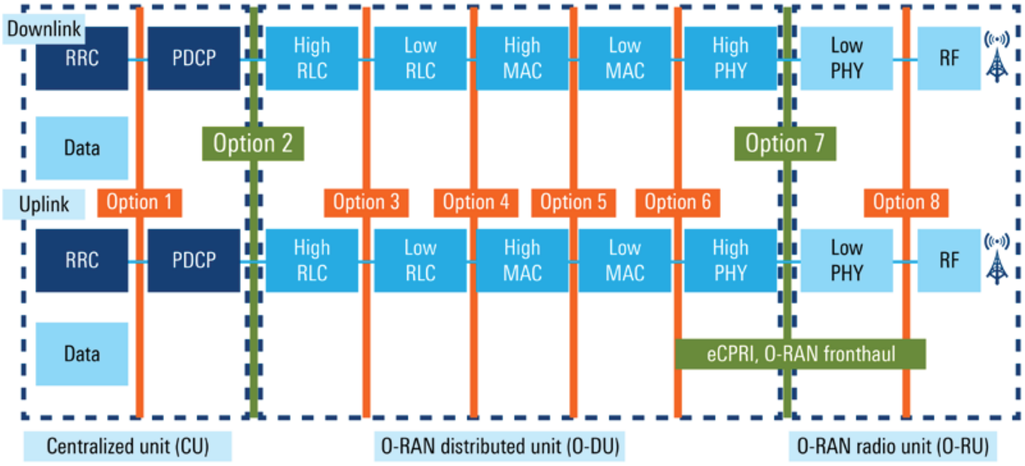


Figure 4.D illustrates various split options that specify functional splits at Layer 1 and Layer 2 [Ref: 3GPP, TR 38.801, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3056]. A given gNB can be implemented with one of such splits. Note that 3GPP already supports Option 2. Furthermore, observe that O-RAN has already defined functionalities beyond the 3GPP specifications in support of the split at the physical layer, where O-RAN-defined fronthaul is used to carry the baseband signal. A popular method to carry a baseband signal between the baseband processor and the RF processor is to use the protocol called enhanced Common Public Radio Interface (eCPRI). Note that the utility of a given function split becomes quite low if the associated interface is not standardized. The lack of standardization of a given interface complicates interworking among the equipment from different vendors.

Figure 4.D. Example \Functional Splits for the gNB Implementation.



5 NETWORK CUSTOMIZATION AND INTELLIGENCE

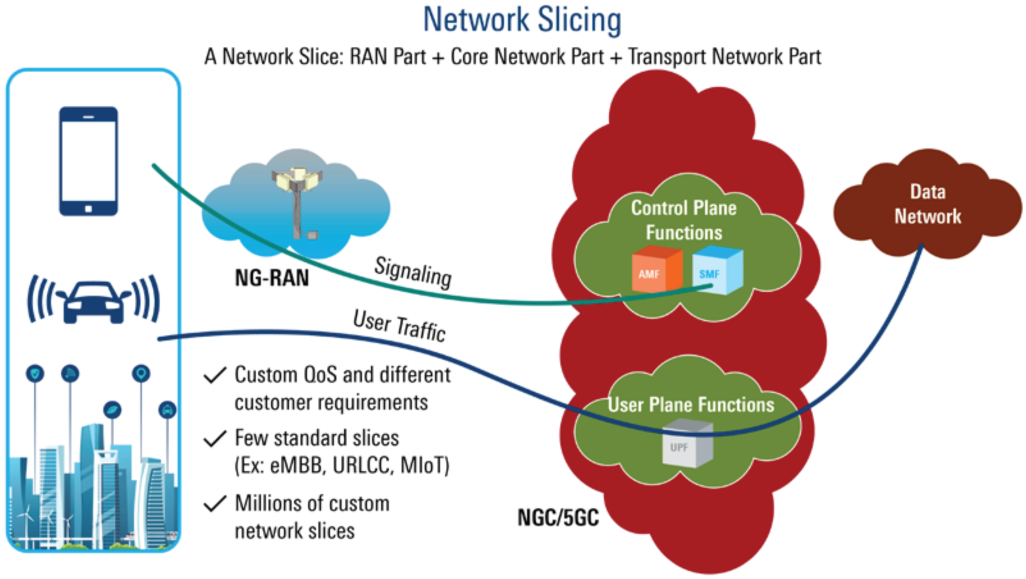
5G is expected to revolutionize numerous industries and transform the way we live. The ability of 5G to customize the network for different industries, use cases, and deployment scenarios is quite important. Hence, 5G is designed to be quite flexible. Furthermore, 5G is quite complex. The power of Artificial Intelligence (AI) and Machine Learning (ML) can be harnessed to optimize a complex 5G network. This section discusses various technologies that can be used to customize a 5G network and make the 5G network more intelligent.

Section 5.1 explains Network Slicing that enables the network to customize the network to meet QoS requirements of diverse services and service requirements of different customers or industries. The concept of Multi-access Edge Computing (MEC) can be exploited to reduce the overall end-to-end delay between the device and a server and reduce the amount of traffic in the network. Section 5.2 introduces MEC. Section 5.3 narrates how a Non-Public Network (NPN) can expand the utility of 5G by enabling use cases such as industrial Internet of Things (IIoT) and Enterprise networks. The goals of service ubiquity, service mobility, and service scalability can be facilitated by using a Non-Terrestrial Network (NTN). Section 5.4 describes characteristics of an NTN. Automation and intelligence can be utilized in a Self-Organizing Network (SON) to optimize various operations. Section 5.6 gives examples of SON applications in 5G. The emerging trend of Open-RAN (O-RAN) that provides additional flexibility and cost savings to service providers and makes use of AI techniques for advanced RAN performance is described in Section 5.7.

5.1 Network Slicing

3GPP introduced the concept of Network Slicing in 5G to enable the service provider to customize the network to support a wide variety of services and meet diverse customer requirements. Network Slicing involves creating different logical networks or network slices in the same physical network to meet a target set of requirements. For example, one network slice can be created to support enhanced Mobile Broadband (eMBB) services, while another network slice can be created to support Ultra-Reliable Low Latency Communications (URLLC). Yet another network slice can be created to support Massive IoT (MIoT) services. Indeed, 3GPP has defined three standard network slices in Release 15- eMBB, URLLC, and MIoT. In the case of the eMBB slice, high data rates are important, and the network can allocate a large number of resources to achieve high data rates. In the case of the URLCC network slice, reliability and latency are essential. Hence, the network can utilize a higher degree of redundancy, faster retransmissions, and suitable resource reservation to increase reliability and reduce latency. For a MIoT slice, the support for a large number of devices with (typically) delay-tolerant and low-rate services is needed.
Figure 5.1 illustrates the concept of a Network Slice.

Figure 5.1. Network Slicing: The Concept.


Network Slicing
A Network Slice: RAN Part + Core Network Part + Transport Network Part
NG-RAN
Signaling
User Traffic
Control Plane Functions
AMF
SMF
Data Network
User Plane Functions
UPF
NGC/5GC
✓ Custom QoS and different customer requirements
✓ Few standard slices (Ex: eMBB, URLLC, MIoT)
✓ Millions of custom network slices

From the 3GPP and functional perspectives, a Network Slice formally includes a RAN part and a core network part. In practice, a suitable transport network part is also needed to meet the Network Slice requirements. As shown in Figure 5.1, both signaling and user traffic are considered for the network slice. For example, for an eMBB network slice, which is widely used in 5G Phase 1 commercial deployments, suitable Control Plane Functions such as the AMF and the SMF are chosen for signaling. Some AMFs may be optimized for eMBB and other AMFs may be optimized for URLLC.

Similarly, some SMFs may be optimized for eMBB, and other SMFs may be optimized for URLLC. For the user traffic, some UPFs may be optimized for eMBB and other UPFs may be optimized for URLLC. In particular, the concept of MEC may be utilized to select a UPF that is close to the device or the gNB to reduce the end-to-end latency for a given network slice.

Virtualization also plays a vital role in ensuring an optimal implementation of Network Slicing. For example, different enterprise customers may have different requirements for a given network slice. Two self-driving car manufacturers would like to get a network slice for their customers. However, they may have different service or performance requirements (e.g., X% availability vs. Y% availability). The service provider would then need to use different computing, networking, and storage resources to meet the specific requirements of these two car manufacturers. Suitable network slicing management functions are needed to carry out life cycle management of network slices.

To enable service providers to create a wide variety of network slices to meet specific service and customer requirements, 3GPP supports built-in flexibility for the network slicing framework. A given network slice is specified by an 8-bit Slice/Service Type (SST) and a 24-bit Slice Differentiator (SD). SST corresponds to a set of features and services. 3GPP defined three standardized or standard slices of eMBB (with SST=1), URLLC (with SST=2), and MIoT (with SST=3) in Release 15. A Vehicle-to-Everything (V2X) slice with SST=4 was introduced in Release 16 [Ref: 3GPP, TR 38.801, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3056]. Standardized SSTs facilitate global interoperability. The service provider can also define non-standard SSTs. The SD value allows differentiation among network slices that have the same SST but different characteristics. For example, two self-driving car manufacturers may be allocated the same SST=4 for V2X services but different SDs for differentiated services. Since the SST consists of 24 bits, the service provider can define about 16 million slices for each SST!

5.2 Edge Computing or Multi-access Edge Computing (MEC)

In the traditional method of providing services to the UEs, the servers are located far away from the UEs in a relatively centralized location. More and more traffic is aggregated in the transport network as the distance from the UE to the server increases. For example, there is relatively less traffic in the local area's transport network surrounding a gNB than the transport network near a relatively centralized UPF, because traffic from multiple gNBs from local geographic areas gets aggregated toward the centrally-located UPF. For example, there is less traffic at the city level but more traffic at the state level in the transport network, because traffic from multiple cities gets aggregated for a state before such traffic reaches a UPF located near central servers. Central locations of servers lead to a long end-to-end latency between the UE and the server and require more bandwidth (or equivalently, increase the load) in the transport network.

Edge computing locates services (e.g., Application servers) close to the UEs. Edge computing, in different standards bodies and at different times, has also been referred to as Mobile Edge Computing and Multi-access Edge Computing (MEC). For a given UE, edge computing involves choosing a UPF near the gNB that is serving the UE. Such selection of the "local" UPF, as opposed to a "central" or "far-away" UPF, enables steering of the traffic toward the local Data Network, where suitable Application Servers are hosted. Such traffic steering may be based on the UE's subscription data, UE location, Application Function (AF) information, policy, or other related traffic rules [Ref: 3GPP, TS 23.501, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144]. Furthermore, the 5GC may expose network information and capabilities to an Edge Computing Application Function.

Edge computing reduces the end-to-end latency for user traffic due to relatively short distances between UEs and Application Servers. Furthermore, since the user traffic is routed toward a local Data Network instead of a remote or "central" Data Network, less traffic is aggregated, reducing the transport bandwidth requirements. Virtualization technologies such as NFV, SDN, and the cloud infrastructure or NFV Infrastructure facilitate the realization of Edge Computing.

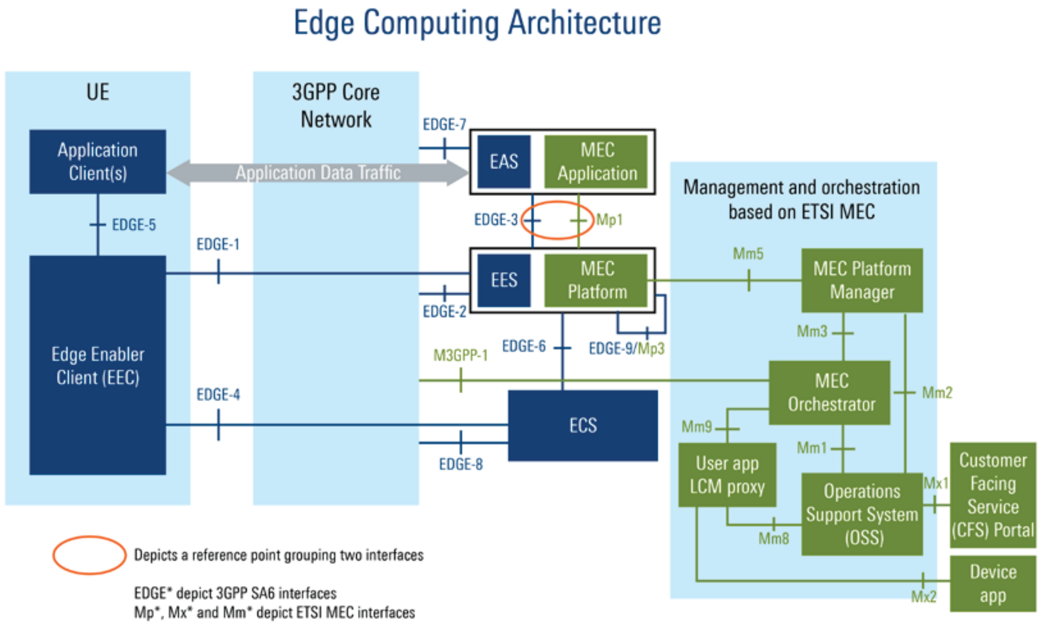
Edge Computing is an evolving concept in the context of 5G. 3GPP aims to provide native support of Edge Computing in 3GPP networks in Release 17. Edge Computing would enable 5G to effectively support use cases such as virtual and augmented reality, IoT, Industrial IoT, autonomous driving, real-time multiplayer gaming, and split computing[Ref: 3GPP, found at https://www.3gpp.org/news-events/2152-edge_sa6]. A harmonized 3GPP-ETSI architecture for Edge Computing is shown in Figure 5.2, which combines 3GPP-defined Edge Computing architecture called EDGEAPP and the ETSI-defined Multi-access Edge Computing (MEC) architecture [Ref: ETSI, found at https://www.etsi.org/images/files/ETSIWhitePapers/ETSI_wp36_Harmonizing-standards-for-edge-computing.pdf].

Application Clients (ACs) on the UE exchange application data traffic with Edge Application Servers (EASs). The UEs may use the Domain Name System (DNS), the MEC Application, or the EEC to perform the discovery of the EAS. ACs can be edge-aware or edge-unaware, with edge-aware ACs directly communicating with the network architecture. For example, in the case of the edge-aware AC, the Edge Enabler Client (EEC) helps the ACs in the UE with EAS discovery by interacting with the Edge Enabler Server (EES). The EES enables the discovery of the EASs. The Edge Configuration Server (ECS) specifies configurations to the EEC on the UE to connect with an EAS.

The Customer-Facing Portal (CFS) enables enterprise customers to place orders for MEC applications. The Operations Support System (OSS) grants or rejects requests for instantiation and termination of MEC applications. The MEC orchestrator automates the overall offering of the MEC applications and services by interfacing with the OSS, the MEC Platform Manager, and the User Application Life Cycle Management (LCM) Proxy, and the 3GPP core network. The User Application LCM Proxy authorizes requests from device/UE applications for instantiation and termination of applications and provides the state of these applications to the device applications. The MEC Platform Manager performs LCM of MEC applications, manages the MEC platform, and processes fault reports and performance measurements. The MEC Platform allows MEC applications to discover, consume, or offer services and hosts MEC services such as location and bandwidth manager.

The edge services are exposed to the ACs by the ECS and the EES via the EEC in the UE. The address of the ECS is provided to the EEC by the Mobile Network Operator (MNO) or the Edge Computing Service Provider. The platforms in the network such as the 3GPP EES or the ETSI MEC Platform perform functions such as application authorization, application service registration, application service discovery, and context transfer. The platforms also expose Application Programming Interfaces (APIs) towards edge cloud applications provided by the MEC application or the EAS residing in the local data network.

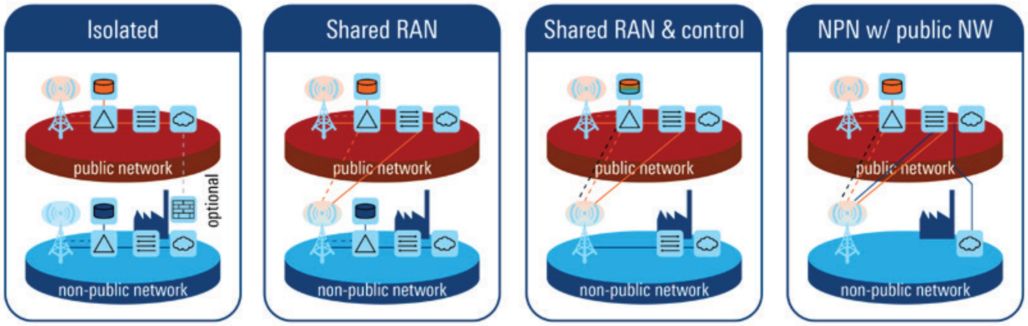
Figure 5.2. Harmonized 3GPP-ETSI architecture for Edge Computing.


Edge Computing Architecture
UE
Application Client(s)
EDGE-5
Edge Enabler Client (EEC)
3GPP Core Network
Application Data Traffic
EDGE-1
EDGE-4
EDGE-7
EAS
MEC Application
EDGE-3
Mp1
EES
MEC Platform
EDGE-2
M3GPP-1
EDGE-6
EDGE-9/Mp3
ECS
EDGE-8
Management and orchestration based on ETSI MEC
Mm5
MEC Platform Manager
Mm3
MEC Orchestrator
Mm2
Mm9
User app LCM proxy
Mm1
Operations Support System (OSS)
Mm8
Mx1
Customer Facing Service (CFS) Portal
Device app
Mx2
Depicts a reference point grouping two interfaces
EDGE* depict 3GPP SA6 interfaces
Mp*, Mx* and Mm* depict ETSI MEC interfaces

5.3 Non-Public Network (NPN)

3GPP introduced the concept of a Non-Public Network (NPN) in support of private networks in Release 16. Release 17 aims for enhanced support of an NPN. There are two types of NPNs—Standalone NPN (SNPN) and Public Network Integrated NPN (PNI-NPN) [Ref: 3GPP, TS 23.501, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144]. The SNPN is operated by an NPN operator and offers services to UEs without relying upon network functions provided by a PLMN. In contrast, the Public Network Integrated NPN (PNI-NPN) is an NPN deployed with the support of a PLMN [REF: 3GPP, TS23.501, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144]. An NPN and a PLMN can share the same Next Generation Radio Access Network (NG-RAN) through the concept of network sharing. The NPN may use licensed spectrum, unlicensed spectrum, or shared spectrum based on the spectrum used by the NG-RAN.

Figure: Different deployment scenarios for Non-Public Networks (NPN) in industrial settings



5G-based private 5G networks are emerging to address critical wireless communications requirements in market segments such as enterprises, public safety, and industrial manufacturing. Private networks are deployed for private use by a government, company, or group of companies. In enterprises, 5G-based private networks can offer enhanced capabilities and user experience. Older public safety technologies can be migrated to more advanced 4G LTE and 5G technologies for advanced services and on a larger scale. Private networks can be exploited in relevant Industry 4.0 applications to facilitate the digital revolution in various industries.

Compared to existing private network technologies such as Wi-Fi, Bluetooth, and TETRA, 3GPP-based LTE and 5G technologies offer significant advantages in meeting private network requirements. Key private network requirements include availability, reliability, interworking, Quality of Service, and security. [Ref: Ericsson, https://www.ericsson.com/4af9b6/assets/local/reports-papers/white-papers/criticalcapabilities5g.pdf]. High availability implies that the end-user can always use the service. A network can ensure high availability by minimizing or eliminating downtime, scheduling maintenance intelligently, and ensuring a suitable degree of redundancy. High reliability means that a given amount of traffic can be transmitted within a target duration with a high probability of success. Robust network coverage, high network capacity, and effective mobility management can help achieve high reliability. Interworking with public networks can be quite important for certain private networks. For example, in the case of a private public safety network, ambulances would need service continuity while moving between a private network and a public network. QoS can be quantified by metrics such as throughput, latency, delay jitter, and packet drop rate. If QoS requirements are too strict for an unlicensed spectrum for a given private network, licensed spectrum can be used. End-to-end security is of paramount importance to protect information, infrastructure, and people from threats. While LTE is quite secure, 5G offers enhanced security capabilities compared to LTE.

Key characteristics of the SNPN and the PNI-APN within the scope of Release 16 are briefly mentioned below.

SNPN in Release 16

An SNPN utilizes the network architecture shown in Figure 1.2 of Section 1 and may also utilize an additional node of N3 Interworking Function (N3IWF) such that the UE can access the 5GC via a non-3GPP network using such N3IWF. In case of such non-3GPP access, the UE-AMF communications and the UE-UPF communications pass through the N3IWF. Such N3IWF-based access to the 5GC is typically used in case of untrusted non-3GPP access such as WiFi access. In Release 17, direct access to the SNPN is supported for 3GPP access only.

An SNPN is identified by the combination of a PLMN ID and a Network identifier (NID). The PLMN ID consists of two parts- Mobile Country Code (MCC) and Mobile Network Code. The PLMN ID used for SNPNs does not need to be unique. Indeed, PLMN IDs are reserved for use by private networks (e.g., with MCC set to 999). A PLMN operator can use its own PLMN IDs for SNPN(s) along with NID(s). The NID can be determined in two ways. In the self-assignment model, the SNPN operator chooses the NIDs individually at the time but uses a different numbering space than NIDs determined the coordinated assignment model. In the coordinated assignment model, the NID may be determined such that the NID itself is globally unique independent of the PLMN ID or the combination of the NID and the PLMN ID is globally unique. NG-RAN nodes (e.g., gNBs) broadcast one or multiple PLMN IDs and a list of NIDs per PLMN ID.

An SNPN-enabled UE is configured with a subscriber identifier called Subscription Permanent Identifier (SUPI), just like a typical 5G UE. Additionally, such UE is configured with credentials for each subscribed SNPN. Furthermore, if an SNPN-enabled UE is configured with the address of an N3IWF, the country identifier corresponding to such N3IWF is also specified. An SNPN-enabled UE supports the SNPN access mode, and the UE only selects and registers with SNPNs while operating in the SNPN access mode.

The UE may access PLMN services via an SNPN or SNPN services via a PLMN. To access PLMN services via an SNPN, the UE in SNPN access mode first needs to register successfully with an SNPN. The UE then performs another registration via the SNPN User Plane with a PLMN using the credentials of such PLMN. The SNPN takes the role of “untrusted non-3GPP access” of the PLMN. To access SNPN services via a PLMN, the UE first registers successfully with a PLMN over 3GPP access. The UE then performs another registration with an SNPN using the credentials of such SNPN via the PLMN User Plane, with the PLMN taking the role of “Untrusted non-3GPP access” of the SNPN.

In Release 17, the SNPN has certain limitations. For example, interworking with EPS, emergency services, roaming between SNPNs, and certain types of handover such as handover between two SNPNs, handover between SNPN and PLMN, and handover between SNPN and PNI NPN, and Cellular CIoT 5GS optimizations are not supported in SNPNs.

PNI-NPN in Release 16

The PNI-NPN is the NPN that is made available via PLMNs through means such as dedicated DNNs or one or more Network Slice instances allocated for the NPN. The UE needs to have a subscription for the PLMN to access the PNI-NPN.

The PNI-NPN may utilize the Closed Access Group (CAG) mechanism to apply access control, where a UE's access to the PNI-NPN is controlled on a per-cell basis. A CAG is a group of subscribers who are permitted to access one or more CAG cells associated with the CAG. A CAG is identified by a CAG Identifier that is unique in a PLMN. A CAG cell broadcasts one or multiple CAG Identifiers per PLMN to enable CAG-capable UEs to access such CAG cells. As part of mobility restrictions, the CAG-capable UE may be pre-configured, configured, or reconfigured with the CAG information such as an Allowed CAG list. This list consists of CAG Identifiers that the UE can access and an optionally CAG-only indicator that informs the UE whether the UE can only access the 5G System (5GS) via CAG cells. Emergency Services are supported in CAG cells for CAG-capable UEs. Furthermore, emergency services for non-CAG UEs may be supported based on the operator policy.

Beyond Release 16: eNPN

Example NPN enhancements targeted for Release 17 include (i) support for the SNPN along with credentials owned by an entity separate from the SNPN (e.g., broadcasting of information to enable SNPN selection, support for associated cell selection/reselection and connected mode mobility, and related changes on the network interfaces) and (ii) support for UE onboarding and provisioning for an NPN, (e.g., broadcast of parameters relevant to the UE onboarding, associated cell selection/reselection, cell access control and the connected mode mobility support, and related changes on the network interfaces), and (iii) support for IMS voice and emergency services for SNPN (e.g., broadcasting of relevant parameters).

5.4 Non-Terrestrial Network (NTN)

3GPP introduces a 5G NR-based Non-Terrestrial Network (NTN) in Release 17. 3GPP has two work items related to the NTN: (i) NR-based NTN that extends the applicability of 5G from the Terrestrial Network (TN) to the NTN, especially supporting the eMBB use case and (iii) Narrowband-Internet of Things (NB-IoT)-based NTN that utilizes the NB-IoT air interface for IoT.

An NTN is a network that utilizes an airborne vehicle or a spaceborne vehicle for transmission and carries out at least some RF processing [Ref: 3GPP, TR 38.821, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3525]. Examples of RF processing include power amplification, filtering, and frequency conversion. Unmanned Aerial Vehicles (UAVs) or drones are not considered to be part of the 3GPP NTN. The NTN may make use of NTN platforms such as different types of satellites (e.g., the Low Earth Orbit (LEO), Medium Earth Orbit (MEO), and Geostationary Earth Orbit (GEO) satellites) and High Altitude Platform Station (HAPS). A stationary aircraft in the stratosphere is an example of HAPS. 3GPP also supports air-to-ground (ATG) systems as part of the NTN feature.

The NTN use cases can be classified into one or more of the three main categories- service ubiquity, service scalability, and service continuity [Ref: 3GPP, TR 38.822, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3372].

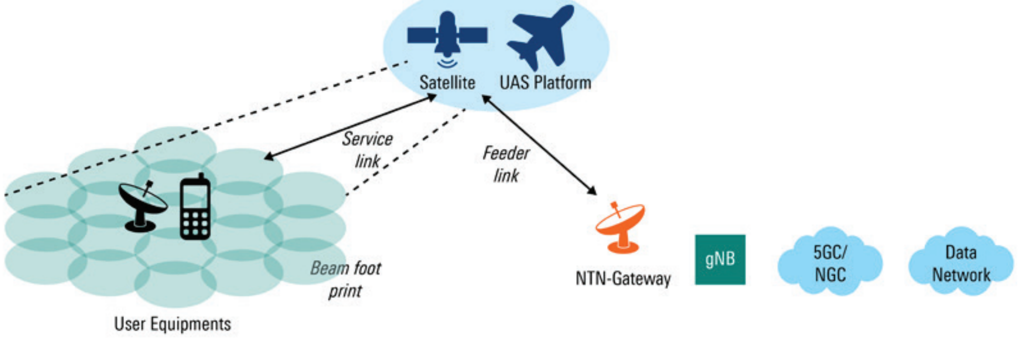
The service uniquity use cases aim to provide the user NTN services anywhere, especially in unserved or underserved areas where terrestrial networks have challenges providing communications services. For example, a terrestrial network (TN) may not be available in a given geographic area due to economic rationales or partial or complete destruction of the terrestrial network infrastructure due to disasters such as earthquakes and floods. Example uses cases for the service uniquity category include (i) Internet of Things (IoT) applications for agriculture, metering and control of critical infrastructures such as pipelines, and asset tracking, (ii) public safety and associated emergency networks, and (iii) broadband access in rural homes.

The service scalability use cases aim to provide efficient broadcast transmissions. An NTN typically has much broader coverage compared to a TN. For example, a typical TN cell has a cell diameter of few kilometers, but an NTN cell has a diameter of hundreds of kilometers. Indeed, one NTN cell's coverage may correspond to the combined coverage of thousands of TN cells. Hence, an NTN can be highly efficient in multicasting or broadcasting a given content over a large geographic area. Furthermore, traffic can be off-loaded from a TN to an NTN by multicasting or broadcasting non-time-sensitive data in non-busy hours. Example service scalability use cases include distribution of rich or very rich TV content in media encoding formats such as 3D and Ultra High Definition (UHD). Users can enjoy such rich multimedia content without causing a considerable radio resource drain on the network.

The service continuity use cases provide continuity of services between a TN and an NTN so that the users do not lose services or users obtain better services (e.g., higher data rates and lower latencies). For example, users currently using services in a TN may lose services when they go outside the TN coverage. Such a situation arises when users embark on land platforms such as trains, airborne platforms such as a commercial aircraft or a private jet, or maritime platforms such as a cruise ship. Similarly, when being serviced by a NTN enters a geographic area with TN coverage the user can switch to the TN and likely gain access to enhanced services due to a relatively larger amount of radio resources per user in a TN compared to an NTN.

3GPP may eventually support multiple NTN architectures. Figure 5.3 illustrates an example architecture for an NTN for Release 17 [Ref: 3GPP, TR 38.821, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3525].

Figure 5.3. Simplified NTN Architecture with a Transparent Payload.



The link between the UE and the NTN platform (i.e., a satellite or HAPS) is called the service link or the access link, and the link between the NTN platform and the NTN-Gateway (NTN-GW) is called the feeder link. The service link utilizes the NR-based radio interface. The UE may be a handheld device such as a smartphone or a Very Small Aperture Terminal (VSAT) device with an external antenna (e.g., a dish antenna). A VSAT device may be mounted on a fixed structure or a moving structure. Figure 5.3 utilizes a transparent payload, where the satellite or HAPS acts as an enhanced relay or repeater and performs RF processing such as frequency conversion between the service link and the feeder link, filtering, and power amplification. The gNB carries out the NR radio protocol stack processing on the ground in case of such a transparent NTN payload.

The NTN poses several new challenges compared to a TN. The propagation delays could be quite long and different for GEOs and LEOs. The UE-gNB propagation delay could be hundreds of milliseconds for GEO satellites and tens of milliseconds for LEO satellites. Furthermore, since LEO satellites move relative to a point on the Earth's surface, the UE-gNB propagation delay varies from one instant to another. A given NTN platform generates different types of beams. For example, GEO satellites and HAPS create Earth-fixed beams, where a beam illuminates a fixed area of Earth at all times. LEO satellites, depending upon their capability and configuration, create quasi-Earth-fixed beams or Earth-moving beams. For example, in the case of quasi-Earth-fixed beams, a LEO satellite utilizes its beam steering capability and illuminates one fixed area on Earth during one period and another fixed area on Earth during another period. In contrast, in the case of Earth-moving beams, a LEO satellite utilizes a fixed beam that illuminates one area on Earth at one instant and a different area on Earth at the very next instant as the satellite with a fixed beam pattern moves in its orbit.

Another challenge in the NTN is the cell size. An NTN cell has a diameter of hundreds of kilometers, covering a geographic area equivalent to hundreds or thousands of TN cells. Depending upon the user density, per-user radio resources may be limited compared to a TN. Quasi-Earth-fixed beams or Earth-moving beams mean that the cell identities on the NR-based radio interface as seen by a stationary UE would be different during different periods. Such moving cells would cause frequent and massive handovers. Additionally, non-GNSS satellites such as LEO satellites cause much larger Doppler shifts compared to a TN. The LEO satellite speed is approximately 7 km/s, high-speed trains move at the speed of up to 0.17 km/s (corresponding to 375 miles per hour), and aircraft speed is 0.27 km/s (corresponding to 600 miles per hour).

3GPP is considering solutions to the challenges mentioned above in Release 17. Timing and frequency pre-compensation can be used to address long and time-varying propagation delays and large Doppler shifts. Timers at Layer 2 of the protocol stack such as those related to the random access procedure, Hybrid Automatic Repeat Request (HARQ) retransmissions, and Radio Link Control (RLC) retransmissions would need to be adjusted to address the long and time-varying delays. The NTN platform mobility needs enhanced management of Tracking Areas or Registration Areas, cell selection, cell reselection, handover, and uplink scheduling. In particular, frequent and massive handover would require enhancements to overcome the challenges of the NTN platform mobility and large cells. Relaxed QoS, especially for GEO satellites, may also be needed.

5.5 Self-Organizing Network (SON)

A Self-Organizing Network (SON) is a network that can configure and optimize parameters to enhance various network operations and to recover automatically from failures. Hence, SON can be viewed as an important step in the direction of network automation and network intelligence. Indeed, SON functions can utilize AI for enhanced network performance. 3GPP began defining SON functions in Release 8 for LTE. A comprehensive set of SON functions are available for LTE. 3GPP has been extending SON to 5G. The actual SON functions or algorithms that modify relevant parameters are implementation-specific and hence not standardized by 3GPP. 3GPP defines suitable measurements and signaling mechanisms for reporting of measurements and configurations so that SON algorithms can perform their tasks.

A SON process may be open-loop or closed-loop [Ref: 3GPP, TR 28.861, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3556]. An open-loop SON process is semi-autonomous and has pre-defined stop points for human intervention. In contrast, a closed-loop SON process is fully autonomous and involves human intervention only when exceptions occur. A transition between the open-loop SON process and the closed-loop SON process is also supported.

A SON function may be implemented using a centralized SON (C-SON) architecture, a distributed SON (D-SON) architecture, or a hybrid SON (H-SON) architecture that combines centralized and distributed architectures [Ref: 3GPP, TR 28.861, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3556]. An example of the C-SON architecture is an Operations, Administration, and Maintenance (OAM) system implementing a SON function. The OAM system at a centralized location obtains required inputs and decides updated parameters relevant to the SON function. An example of the D-SON architecture is an eNB or an entity near an eNB implementing a SON function. In the case of H-SON, part of the SON function is implemented at a central entity such as the OAM system and part of the SON function is implemented at a distributed entity such as an eNB or an entity close to an eNB.

SON functions can be divided into one of the three categories: self-configuration, self-optimization, and self-healing. Self-configuration is applicable to a pre-operational state of the eNB between (i) the time the eNB is powered up and has backbone connectivity and (ii) the time the RF transmitter of the eNB is switched on. In contrast, self-optimization (and self-healing) is applicable to the operational state that exists after the RF interface has been switched on. During self-optimization, a SON function utilizes UE and eNB measurements and performance measurements to auto-tune the network.

As of Release 16, support for the following functionalities is available for LTE SON.

- Self-configuration Features: Dynamic configuration of the S1-MME interface, Dynamic configuration of the X2 interface, Automatic Neighbor Relation (ANR) Function, Physical Cell Identifier (PCI) Selection, Transport Network Layer (TNL) address discovery, and Dynamic configuration of the Xw-C interface (Xw-C is an interface between the eNB and a WLAN Termination when LTE-WLAN Aggregation is used).
- Self-optimization Features: Mobility Load Balancing (MLB), Mobility Robustness Optimization (MRO), Random Access Channel (RACH) Optimization, Energy Saving, and Radio Link Failure (RLF) reports.

As of Release 16, support for the following functionalities is available for NR SON.

- Self-configuration Features: Dynamic configuration of the NG-C interface, Dynamic configuration of the Xn interface, ANR, and Xn-C TNL address discovery.
- Self-optimization Features: MLB, MRO, RACH Optimization, and UE History Information from the UE.
- Energy Saving.

The feature of Minimization of Drive Tests (MDT) can be viewed as a companion feature of SON or one of the SON functions. While are some commonalities between SON and MDT, they can be independently utilized. 3GPP is aiming to address the following features in support of SON and MDT in Release 17 [Ref: 3GPP, RP-201281, found at https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_88e/Docs].

- Data collection for SON features such as Capacity and Coverage Optimization (CCO), inter-system inter-RAT energy-saving, inter-system load balancing, 2-step RACH optimization, mobility enhancement optimization, PCI selection, energy efficiency, successful handovers reports, UE history information in EN-DC, load balancing enhancement, MRO for Secondary Node (SN) change failure, and RACH Optimization enhancements.
- Data collection for MDT features for use cases such as 2-step RACH optimization and MDT for MR-DC.
- NR-U related SON/MDT optimization.

Selected SON functions of PCI Selection, ANR, RACH Optimization, MRO, MLB, Energy Saving, MDT, and CCO are briefly described below [Ref: 3GPP, TS 36.300, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2430] [Ref: 3GPP, TS 38.300, found at https://portal.3gpp.org/desktopmodules/Specifications/

SpecificationDetails.aspx?specificationId=3191].

PCI Selection.

This feature involves the automatic selection of a PCI for a given cell. There are 505 PCIs in LTE and 1008 PCIs in NR. A PCI on a given carrier frequency in a given geographic area is reused in a different geographic area on the same carrier frequency. Hence, the goal of PCI selection is to minimize any PCI collision or PCI confusion. For example, when different PCIs are used in neighboring cells, interference is reduced, and confusion is eliminated. In the centralized PCI assignment approach, the OAM signals a specific PCI value, and the eNB selects such value. In the distributed PCI assignment approach, the OAM signals a list of PCIs and the eNB can choose a PCI after removing PCIs that are reported by UEs and neighboring eNBs and observed using a downlink receiver at the eNB.

Automatic Neighbor Relation (ANR).

Manual neighbor list configuration for a given serving cell may introduce human errors, adversely affecting cell reselection and handover. The ANR function manages the conceptual Neighbor Cell Relation Table (NCRT). The Neighbor Detection Function of the ANR function helps find new neighbors and adds them to the NCRT. Similarly, the Neighbor Removal Function of the ANR function removes outdated Neighbor Cell Relation (NCRs). Assume that the eNB has implemented the ANR function. An existing Neighbor Relation from a source cell to a target cell means that eNB controlling the source cell knows the ECGI/CGI and PCI of the target cell and has the relevant entry in the NRCT. Each NCR has three attributes: “NoRemove,” “NoHO,” and “NoX2.” If “NoRemove” is checked, the eNB shall not remove the Neighbor Cell Relation from the NRT. If “NoHO” is checked, the NCR is not used by the eNB for handover. Finally, if “NoX2” is checked, the X2 interface is not used to initiate procedures towards the eNB controlling the target cell. Note that NCRs are unidirectional cell-to-cell relations, while an X2 link is bidirectional. The ANR can be used for intra-frequency neighbors, inter-frequency neighbors, and inter-Radio Access Technology (RAT) neighbors.

Mobility Load Balancing (MLB).

The MLBLMS aims to distribute load or traffic evenly among cells or alleviate congestion in cells by moving traffic from one cell to another via optimization of mobility parameters or handover actions. MLB can be carried out within a RAT or between RATs. MLB involves obtaining load reports, load balancing based handovers, and adapting handover and/or reselection configuration. Examples of the load information in the UL and the DL include radio resource usage (e.g., GBR and non-GBR PRB usage), hardware load indicator (e.g., low, mid, high, and overload), TNL load indicator (e.g., low, mid, high, and overload), Cell Capacity Class value (e.g., a relative capacity indicator), capacity value (e.g., available capacity for load balancing as a percentage of total cell capacity). Based on the load reports, the source cell may initiate load balancing handovers. The target cell performs admission control for such load balancing handover requests. A source cell may request a change in handover and/or reselection parameters at a target cell. The source cell keeps the target cell informed about any new mobility settings. Automatic changes in handover and/or cell reselection parameters need to be within the range specified by the OAM system.

Mobility Robustness Optimization (MRO).

MRO aims to detect and rectify problems such as connection failure due to intra-RAT or inter-RAT mobility, unnecessary handover to another RAT, and inter-RAT/-system ping-pong (e.g., between

LTE and NR). MRO detects Radio Link Failures (RLFs) occurring due to "Too Early Handover," "Too Late Handovers," or "Handover to Wrong Cell." In the case of "Too Early Handover," an RLF occurs soon after a successful handover from a source cell to a target cell or a handover failure occurs during the handover procedure itself, and, the UE attempts to re-establish the radio link connection in the source cell. In the case of "Too Late Handovers," an RLF occurs after the UE has stayed for a long time in the source cell, and, the UE attempts to re-establish the radio link connection in a different cell. In the case of "Handover to Wrong Cell," an RLF occurs shortly after a successful handover from a source cell to a target cell or a handover failure occurs during the handover procedure itself, and, the UE attempts to re-establish the radio link connection in a cell that is neither the source cell nor the target cell. The MRO function processes handover and RLF reports to detect such scenarios and modifies handover parameters (e.g., event thresholds and hysteresis values) to optimize the handover performance.

Random Access Channel (RACH) Optimization.

This SON function aims to improve the performance of the random access (RA) procedure. It processes the RA reports from UEs and the RA parameters exchanged between the eNBs/gNBs and adapts the RA parameters. Examples of RA information for NR useful for RACH optimization include contention detection indication per RACH attempt, indexes of the Synchronization Signal/Physical Broadcast Channel Blocks (SSBs) and the number of RACH preambles sent on each tried SSB, and indication whether the selected SSB is above or below a signal threshold per RACH attempt. After processing such RA information, this SON function tries to modify the RACH parameters. Examples of RACH parameters that can potentially be tuned include the RACH configuration (i.e., time0frequency resources of the Physical Random Access Channel (PRACH), PRACH preamble distribution (e.g., distribution among dedicated RA preambles, group A RA preambles, and group B RA preambles, PRACH backoff parameter value, and PRACH transmission power control parameters (e.g., target received power level at the gNB and power ramping step size).

Minimization of Drive Tests (MDT).

The goal of the MDT feature is to enable network optimization without carrying out extensive traditional drive testing. Traditional drive testing consumes a significant amount of time and financial resources. Furthermore, drive testing may not reflect where users are accessing the network. The MDT involves the configuration of selected UEs during finite periods (and potentially for a certain amount of reporting). These UEs and the network make measurements in target geographic areas. The UE measurements and the network measurements can be correlated and processed to optimize various operations such as handover and random access. There are two modes for the MDT measurements: immediate MDT and logged MDT [Ref: 3GPP, TS 37.320, found at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specification-Id=2602]. Additionally, some measurements, such as accessibility measurements do not belong to either of these two modes. The immediate MDT mode is applicable to UEs in RRC_CONNECTED state and the reporting of measurements occurs (almost) at the time of measurements. In contrast, the logged MDT mode is applicable to UEs in RRC_IDLE and RRC_INACTIVE states and the measurements are reported at a future instant. MDT can be UE-specific or non-UE-specific (e.g., at the cell level). Examples of immediate mode MDT measurements include RSRP, RSRQ, data volumes, throughout, packet delays, and packet error rates. Examples of logged mode MDT measurements include RSRP and RSRQ.

Energy Saving.

The goal of this SON function is to reduce Operational Expenditure (OpEx) through energy savings. In a commercial network deployment, some cells provide the basic coverage so that users can access the network anywhere. However, the network often utilizes capacity booster cells that provide additional capacity to enhance user experience (e.g., higher data rates). When the capacity or resource requirements in a given geographic area are low due to low traffic (i.e., user data activity), some capacity booster cells can be switched off to save energy and reduce utility bills. When needed, such booster cells can be re-activated. The energy savings SON function considers the load to turn off a booster cell. An NG-RAN node (e.g., gNB) may initiate handovers to move the users from the capacity booster cell to other cells, such as the cells on different frequencies providing the basic coverage. The gNB controlling the capacity booster cell being switched off informs the neighboring gNBs about such a switch off using the Xn interface. A gNB controlling non-capacity booster cells may request a re-activation of a capacity booster cell using the cell activation procedure.

Capacity and Coverage Optimization (CCO).

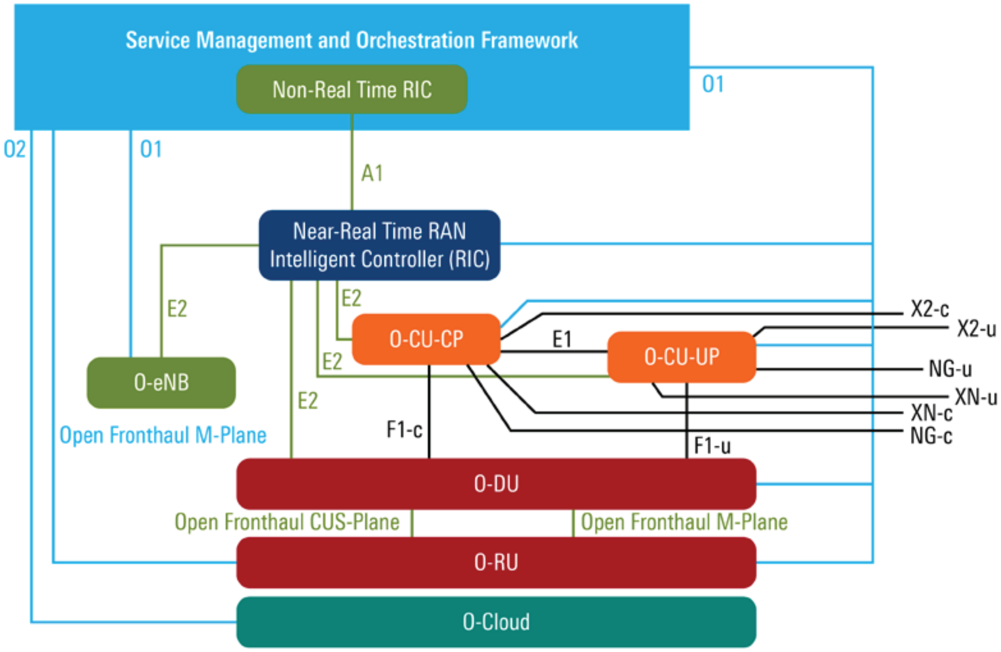
The goal of the CCO function is to serve the maximum number of users at required levels of service in each geographic area. From the perspective of the radio network, such optimization requires a suitable configuration of radio resources so that the radio resources are utilized efficiently, and radio resource allocations are adapted to reflect the prevailing radio channel conditions and service requirements of the traffic of connected users. The CCO function aims to minimize manual intervention and automate by an optimization process that uses UE, network measurements, and configuration status. For example, the CCO function tries to detect coverage holes and identify capacity improvement opportunities automatically to optimize network performance while reducing OpEx. Furthermore, where needed, a tradeoff between capacity and coverage is made by the CCO function. The overall centralized CCO consists of 3 stages: (i) Monitoring (where the UE and the network make continuous measurements with time-space granularity), (ii) Detailed improvement analysis (an optional stage where a fine-grained detection tool such as MDT with periodic measurements is used after the monitoring stage has detected an opportunity of improvement but additional analysis is needed to take any specific action), and (iii) Improvement action (where changes in the cell configuration are made).

5.6 O-RAN with Intelligent RAN Control

The Open-Radio Access Network (O-RAN) Alliance has been working on the RAN architecture and relevant use cases to evolve the RAN by making it more open and smarter than currently deployed networks to realize benefits such as reduced costs, enhanced performance, and increased agility. The O-RAN Alliance aims to facilitate the building of a cost-effective and agile RAN by exploiting the concepts such as open interfaces, open hardware, and open source code, and intelligence “to meet the requirements for increasingly complex, denser and richer networks through deep learning techniques and embedded intelligence in every layer of the RAN architecture” [Ref: O-RAN Alliance, “O-RAN: Towards an Open and Smart RAN,” found at https://static1.squarespace.com/static/5ad774cce74940d7115044b0/t/5bc79b371905f4197055e8c6/1539808057078/O-RAN+WP+FInal+181017.pdf]. The O-RAN Alliance has defined an O-RAN architecture that utilizes AI/ML to make the network intelligent while using open and standardized interfaces in a multi-vendor network environment.

Figure 5.4 illustrates the O-RAN architecture.

Figure 5.4. The O-RAN Architecture [O-RAN_2].



Note that the 3GPP defines a logical network architecture, while O-RAN adds extensions to the 3GPP-defined network. Hence, the O-RAN architecture can be viewed as a network implementation approach that builds on the top of the base 3GPP-defined network architecture. For example, 3GPP defines gNB-CU-CP, gNB-CU-UP, and gNB-DU and O-RAN makes use of these logical network functions as the baseline and adds O-RAN-specific functionalities to define O-CU-CP, O-CU-UP, O-DU, and O-RU. O-DU and O-RU together can be viewed as O-RAN-compliant gNB-DU. O-RAN supports the double-split: low-layer split and high-layer split. To enhance the 3GPP-defined LTE eNB, the O-RAN architecture includes O-eNB. For example, O-CU-CP can be viewed as O-RAN specifications-compliant gNB-CU-CP and O-CU-UP can be viewed as O-RAN specifications-compliant gNB-CU-UP when 5G NR-based gNB is used in the RAN. Note that 3GPP supports disaggregation of the gNB as explained in Section 4.3, where the gNB is disaggregated into gNB-CU and gNB-DU, and gNB-CU can be further disaggregated into gNB-CU-CP and gNB-CU-UP. While 3GPP does not formally divide the gNB-DU into a baseband portion and an RF portion, a commercial implementation can make such distinction such that O-DU can implement the baseband processing of the gNB-DU, and O-RU can implement the RF processing.

In Figure 5.4, the Service Management and Orchestration (SMO) framework contains the Non-Real Time (RT) RAN Intelligent Controller (RIC). The Non-RT RIC supports intelligent RAN optimization in non-real-time, implying a reaction time of greater than one second per O-RAN. The Non-RT RIC utilizes data analytics and AI/ML techniques and provides policy-based guidance to Near-RT RIC. The Non-RT RIC can make use of the SMO services such as data collection and provisioning services of O-RAN nodes (e.g., O-CU and O-DU) to obtain necessary inputs for its processing. The Near-RT RIC performs near real-time control and optimization of O-RAN nodes and relevant resources using the E2 interface. The Near-RT RIC operates with control loops with typical reaction times ranging from 10 ms to 1 s. Examples of primitives used by the Near-RT RIC to control O-RAN nodes include monitor, suspend/stop, override, and control. Furthermore, the Near-RT RIC hosts xApps on the E2 interface to collect near real-time RAN information from O-RAN nodes under the guidance of the Non-RT RIC. For example, the Non-RT RIC and the Near-RT RIC interact with each other to optimize algorithms or functions such as load balancing, mobility management, multi-connection control, QoS management, and network energy saving. Note that some of these are SON functions discussed earlier.

In Figure 5.4, O-Cloud is an O-RAN cloudification and orchestration platform. The O-Cloud is intended to automate deployment and providing of O-RAN. The O-Cloud is a cloud computing platform that consists of physical infrastructure nodes hosting suitable O-RAN functions in support of management and orchestration functions. The O-Cloud builds upon the ETSI-defined Network Functions Virtualization (NFV) architecture to facilitate flexible deployment of virtualized O-RAN network elements using the cloud computing technologies such as Virtual Machines (VMs) and containers.

The O-RAN Alliance has defined several O-RAN use cases [O_RAN2], intended to be gradually supported by O-RAN specifications. Examples of O-RAN use cases include low-cost RAN white-box hardware, RAN sharing, handover management for V2X, Quality of Experience (QoE) optimization, traffic steering, massive MIMO optimization, and RAN Slice Service Level Agreement (SLA) assurance.

6 SUMMARY

Key infrastructure trends include spectrum trends, densification & coverage extension methods, virtualization and cloudification, and network customization and intelligence.

In emerging spectrum trends, higher integration of FR1 and FR2 increases the availability of unlicensed spectrum and spectrum sharing opportunities. Increased integration of FR1 and FR2 provides deployment flexibility, enhanced coverage (due to lower frequencies of FR1), and higher data rates (due to larger channel bandwidths of FR2). A large amount of unlicensed spectrum can be exploited on an opportunistic basis to increase overall capacity and throughput. Spectrum sharing with suitable prioritization maximizes the use of available spectrum while minimizing interference through coordination.

In the area of densification and coverage extension, trends such as small cells with beamforming, IAB, and special infrastructure enhancements in support of V2X communications are seen. The use of millimeter-wave spectrum along with beamforming enables high-performance small cells. IAB obviates the need for the backhaul by connecting some gNBs to the core network through donor gNBs. RSUs plan an important role in the V2X infrastructure.

In the area of virtualization and cloudification, trends such as NFV, SDN, and orchestration are observed. NFV utilizes COTS hardware and enables the independent evolution of hardware and software. SDN centralizes intelligence for routing and makes use of simple packet forwarding devices. Orchestration automates service offerings. The disaggregation of the gNB into the gNB-CU and the gNB-DU and further disaggregation of the gNB-CU into the gNB-CU-CP and the gNB-CU-UP are seen. Such disaggregation provides scalability.

In the area of network customization and intelligence, trends such as Network Slicing, MEC, NPN, NTN, SON enhancements for 5G, and ORAN are observed. Network Slicing provides custom QoS and enables a service provider to meet varying customer requirements by creating different logical networks using the same physical network. MEC reduces the end-to-end latency and the transport network bandwidth requirements by locating Application Servers closer to the users. The NPN makes it easy to deploy private networks for customized services. SON, originally defined for LTE, is being enhanced to support 5G features. ORAN aims to further standardize intra-network interfaces and make the RAN more intelligent than traditional RAN be using RICs.

The emerging infrastructure trends discussed in the paper are expected to continue in the coming years as deployments of 5G and 5G-Advanced become widespread around the globe.

Rohde & Schwarz
The Rohde & Schwarz electronics group offers innovative solutions in the following business fields: test and measurement, broadcast and media, secure communications, cybersecurity, monitoring and network testing. Founded more than 80 years ago, the independent company which is headquartered in Munich, Germany, has an extensive sales and service network with locations in more than 70 countries.

www.rohde-schwarz.com

Rohde & Schwarz customer support
www.rohde-schwarz.com/support

R&S® is a registered trademark of Rohde & Schwarz GmbH & Co. KG
Trade names are trademarks of the owners
Version 01.00 | March 2021
White Paper | 5G Voice over New Radio (VoNR)
Data without tolerance limits is not binding | Subject to change