Introduction and Background

1 Introduction to 5G

The tremendous growth in the number and variety of connected devices and the substantial increase in user/network traffic volume and types as well as the performance limitations of 4G technologies have motivated industry efforts and investments toward defining, developing, and deploying systems for the fifth generation (5G) of mobile networks. The 5th generation of mobile broadband wireless networks have been designed to meet the challenging system and service requirements of the existing and the emerging applications in 2020 and beyond. 5G is a fast, secure, and reliable connected ecosystem comprising humans and machines, which enables seamless mobility, efficient connectivity, increased connection density, increased industrial productivity, automation, and sustainability. The future connected societies are characterized by the significant growth in connectivity and traffic density, network densification, and the broad range of new use cases and applications. As a result, there is a continuous need to push the performance envelope of the wireless systems to the new limits in order to satisfy the demands for greater network capacity, higher user throughput, more efficient spectrum utilization, wider bandwidths, lower latency, lower power consumption, more reliability, increased connection density, and higher mobility through virtualized and software-defined network (SDN) architectures. While extending the performance envelope of mobile networks, 5G includes intrinsic flexibility and configurability to optimize the network usage and services, accommodating a wide range of use cases, and business models. The 5G network architecture encompasses modular network functions that can be deployed, configured, and scaled on demand, in order to accommodate various use cases in a smart and cost-efficient manner.

The scope of 5G is best understood by scrutinizing the main usage models targeted by this megatrend namely enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications (mMTC). The high-throughput mobile broadband use case is primarily based on the evolution and enhancement of LTE technology. Evolutionary nature of mobile broadband use case implies coexistence with the existing radio access networks, but this consideration is only limited to sub-6 GHz RF spectrum.

Any mobile broadband deployments above 6 GHz are expected to use new baseband and radio technologies that improve spectral efficiency, throughput, latency, and other key performance metrics. The mission-critical machine type communication and mMTC use cases are revolutionary in nature. Technology and design choices for the latter use cases are unaffected by the burden of legacy support or backward compatibility (see Fig. 1). The Internet of Things (IoT), automotive, smart grid, traffic safety, emergency response services are some of the examples of low-latency machine type communications. Tactile Internet is an interesting aspect of this use case that is expected to create new services.

As shown in references [11,14], peak and average data rates as well as transport delay are the leading performance indicators of the high-throughput mobile broadband scenario. Interactive gaming, augmented/virtual reality, and immersive entertainment are notable consumer mobile broadband services that are in nascent state and hold significant promise of new business opportunities and user experience. The eMBB delivers over 20 Gbps downlink peak data rates with expected user experienced data rates of approximately 100 Mbps anywhere anytime. This would require substantial increase in network capacity in 2020 and beyond. Network operators are gearing up toward meeting such a dramatic increase in capacity by a combination of new spectrum utilization, spectral efficiency improvements, and ultra-dense network deployments.

The key challenge in the design and deployment of 5G systems is not only about the development of a new radio interface(s) but also about the coordinated operation in a highly heterogeneous environment characterized by the existence of multi-RAT systems, multi-layer

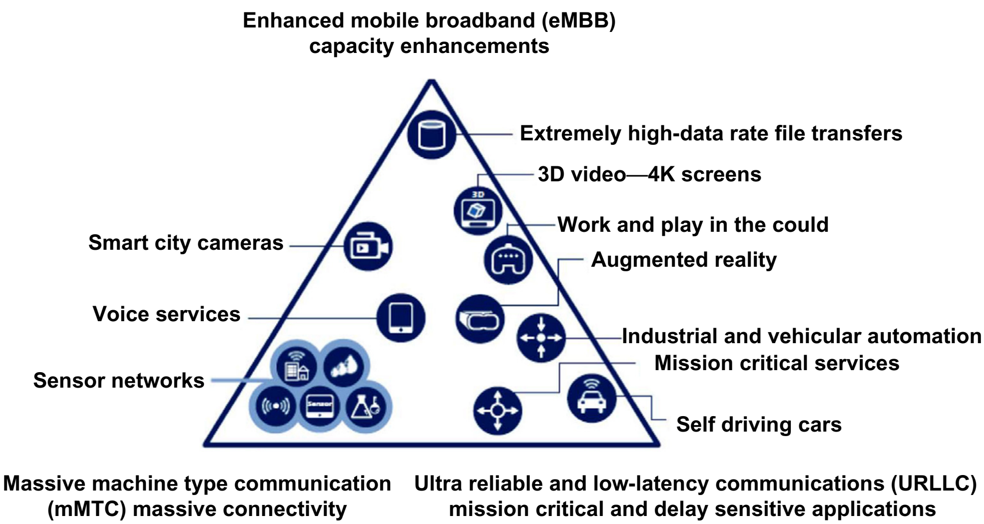

Figure 1
5G use case categories [1].

networks, multi-mode devices, and diverse user interactions. Under these conditions, there is a fundamental need for 5G to achieve seamless mobility and consistent user experience across time and space. New business models and economic incentives with fundamental shift in cost, energy, and operational efficiency are expected to make 5G feasible and sustainable. 5G also is going to offer value proposition to vertical businesses and consumers through the definition and exposure of capabilities that enhance today's overall service delivery.

In addition to supporting the evolution of the established mobile broadband usage models, 5G is going to support numerous emerging applications with widely different requirements and service attributes; that is, from delay-tolerant applications to ultra-low-latency applications, from high-speed scenarios in trains/airplanes to nomadic/stationary scenarios in home or office, and from best effort applications to ultra-reliable applications such as in healthcare and public safety. Furthermore, content and information will be delivered across a wide range of devices (e.g., smartphones, wearables) and through heterogeneous environments. Fulfilling such stringent requirements will require larger number of base stations/access nodes in the form of a heterogeneous network architecture, which can be realized as small cells and high-capacity cloud-RAN or virtual RAN architectures serving a large number of remote radio heads. Content caching and processing at the cloud edge in 5G networks will overcome bottlenecks in the network transport for delay-sensitive applications.

5G use cases and business models require allocation of additional spectrum for mobile broadband and flexible spectrum management capabilities. The new networks are employing a set of new frequency bands (licensed, unlicensed, and shared) to supplement the existing wireless frequencies, enabling larger bandwidth operation and substantially improved capacity. 5G inevitably requires some of the sub-6 GHz spectrum to be repurposed for deployment of the new technologies. Existing cellular bands can be augmented with new spectrum allocations above 6 GHz in order to create extremely wider operating bandwidths. Utilizing RF spectrum above 6 GHz for mobile broadband, low-latency mission-critical machine type communication, and low-energy massive machine type communication requires development of new radio interfaces or access technologies as well as substantial innovations in analog RF-integrated circuits and semiconductor technologies. In addition, carrier aggregation techniques are used to combine segments of spectrum that are not co-located within the same band to further improve peak data rates. 5G systems are required to support up to 400 MHz of contiguous bandwidth and to support a total bandwidth of up to 3.2 GHz via carrier aggregation beyond 6 GHz.

Advanced driver assistance systems (ADAS) and autonomous vehicles are emerging trends in the automotive space. Together, they bring a number of benefits, including improved safety, reducing collision risks and road congestions, improved fuel economy, and higher productivity for the drivers. 5G wireless technologies supporting high-speed, low-latency vehicle-to-vehicle and vehicle-to-infrastructure communications are key enablers of ADAS and autonomous vehicles. In addition, automotive industry is demanding richer infotainment options, which are adding more traffic bottlenecks to wireless networks. eCall is an initiative by the transportation safety regulatory bodies to provide immediate assistance to people involved in accidents. It is a subsystem of telematics systems being installed in vehicles. Implementation of eCall requires significant investment in infrastructure. In the case of a crash, it contacts the closest emergency center and forwards the GPS location of the crash site to the emergency center operators. eCall can be activated manually as well as automatically. It can be activated automatically, if the vehicle is involved in an accident and the airbags are inflated.

As in the previous generations, improvements in radio access technologies, in the form of either evolutionary or revolutionary, have continued to be the core focus of the research and development in 5G. However, it should not be forgotten that the proper design of network architecture and functionality along with ease of programmability/configurability have been equally important to 5G to ultimately achieve the goal of an integrated communication platform serving multiple vertical markets and diverse consumer applications. Network function virtualization (NFV) has been an underlying method to enable network operators to create network slices per end-user application or service requirements with guaranteed performance and quality corresponding to service level agreements. Both cloud and edge computing components will be needed in these network slices to address varying performance and latency requirements. Each network slice will rely on enhanced LTE and/or NR access technologies to provide reliable wireless link between the user terminals and the network.

The comprehensive research and studies conducted on 5G in the past few years and the ongoing standards development efforts are revealing a number of key technologies that have helped achieve the ambitious system and service requirements as follows (see Fig. 2):

- New spectrum: Use of large blocks of spectrum in higher frequency bands and heterogeneous carrier aggregation (particularly above 6 GHz and up to 100 GHz) in addition to unlicensed spectrum have made wider system bandwidths (up to 3.2 GHz) feasible, leading to higher peak data rates and network capacity.
- New waveforms and multiple access schemes: OFDM-based LTE air-interface may not be suitable for some use cases, and therefore a number of new waveform candidates and multiple access schemes have been studied. However, the complexity and practical limitations of some of those new candidates have convinced the 3GPP community to adopt OFDM with adaptive numerology which would allow configurable frame structures and radio resource allocations depending on use case, available spectrum, and bandwidth without affecting backward compatibility with legacy LTE systems.
- Massive MIMO: As the extension of multi-user MIMO concept to large number of antennas at the base station, it has offered a promising solution to significantly increase the user throughput and network capacity by allowing beamformed data and control transmission and interference management. The effect of considerably increased path

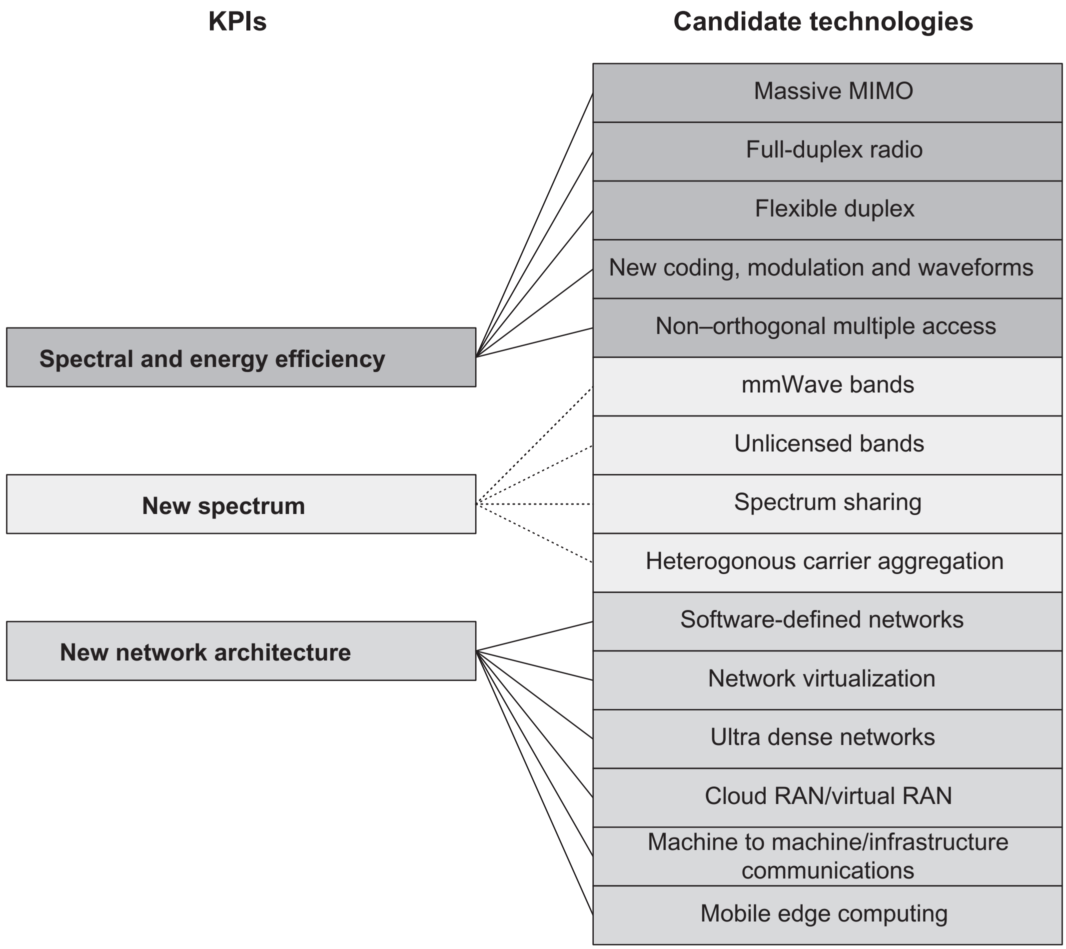

Figure 2
Mapping KPIs to candidate 5G technologies [2].

loss in very high frequencies is compensated by higher antenna gains (directivity), which is made possible by increasing number of antennas at the base station using adaptive beamforming and active antenna schemes.

- New network architecture: NFV and cloud-based radio access and core network architecture have been promising approaches to attain flexibility and versatility in 5G network design, where the objective is to run today's network functions, which are typically implemented on dedicated hardware, as virtualized software functions on general-purpose hardware comprising servers, storage, and hardware accelerators. This is further extended to the radio network by splitting base stations into remote radio units and baseband processing units (connected via optical fiber links, high-speed Ethernet cables, or wireless fronthaul) and pooling baseband functions in the virtualized

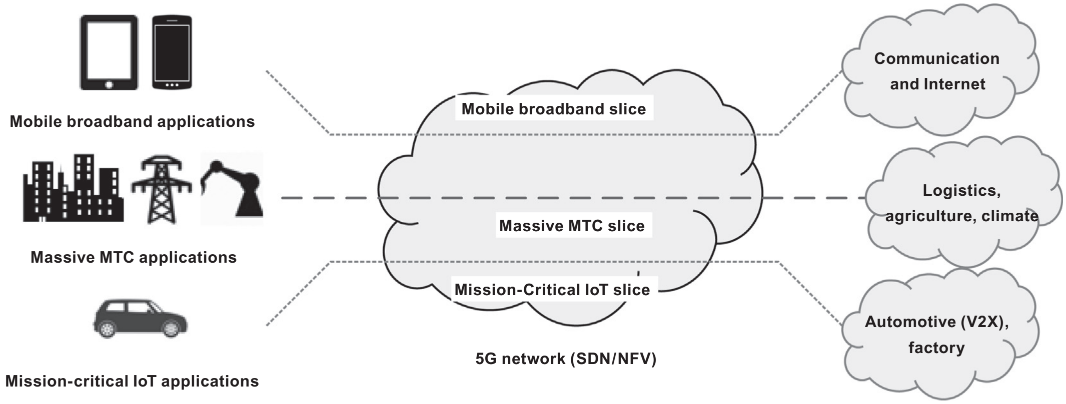

Figure 3
Example network slices accommodating different use cases [43].

environment to serve a large number of remote radio units and to form different network slices along with core network functions.

- Software-defined networking: Separation of control-plane and user-plane is the first step toward centralized network control, which allows network administrators to programmatically initialize, control, change, and manage network behavior dynamically via open interfaces and abstraction of lower level functionality.
- Multi-connectivity: Decoupling of downlink and uplink control/data paths would allow control of user devices on a macro-cell level, where user terminal data packets can be independently sent or received via one or several small cells in a heterogeneous network layout while the control signaling is exchanged with a macro-cell node.

In this chapter, we will study the 5G use cases and prominent deployment scenarios. We will further define the performance metrics and review architectural, system, and service requirements for 5G systems. The standards development activities and timelines within 3GPP and ITU-R will be discussed, and later the key aspects of spectrum allocations and regulatory issues concerning 5G systems will be described. The chapter will be concluded with a discussion about the future of wireless systems and services beyond 5G era (see Fig. 3).

2 Use Cases and Deployment Scenarios

The breakthroughs in 5G network design are expected to sustain the extensive growth and enhancement of mobile Internet and the IoT services. The use of 5G technologies in the IoT and vertical industries are providing new business opportunities for network operators. In addition, expanded and enhanced mobile Internet services have been further improving consumer experience and satisfaction. To adequately support the enhancement and expansion of mobile Internet and IoT, 5G networks have to become the primary means of network access for person-to-person, person-to-machine, and machine-to-machine connectivity. This means that 5G will need to be widely deployed and adapted to the diversity of service requirements and their unique characteristics. The fulfillment of the requirements for a service-oriented network in order to provide better user experience in a flexible and efficient manner is nontrivial. Wireless networks will need to match the performance offered by the advances in fixed networking in terms of delivering a high level of quality of service (QoS), reliability, and security. Based on the requirements, two major challenges are needed to be addressed for the design of the 5G systems, as follows:

- The system should be capable of flexible and efficient use of all available spectrum types from low-frequency bands to high-frequency bands, licensed bands, unlicensed bands, and shared bands.
- The system should be adaptable in order to provide efficient support for the diverse set of service characteristics, massive connectivity, and very high capacity. Flexible network design is required to improve spectral efficiency, to increase connection density, and to reduce latency. These requirements and challenges have also impacts on the design of 5G air interface and the network architecture.

The foundation of the 5G radio access networks has been laid on two major concepts: (1) software-defined networking and (2) network function virtualization. An elastic (adaptable) air interface would allow optimizing the network to support various applications and deployment scenarios. The 5G networks further encompass self-organization and inter-cell coordination algorithms that utilize the features, protocols, and interfaces to overcome the limitations of the cell-based network topology (realizing virtual cell concept). The 5G network architectures are typically heterogeneous in order to enable the cooperation between low-frequency wide-area and high-frequency small-area network topologies. There is a consensus in the industry that higher frequency bands are the complementary bands to deploy 5G, whereas low-frequency bands (i.e., sub-6 GHz) are still the primary bands of the 5G spectrum. High frequency also enables unified access and backhaul since the same radio resources are shared. 5G is expected to use a unified air interface and hierarchical scheduling for both radio access and backhaul which enables flexible backhauling and low-cost ultra-dense network (UDN) architectures.[1] Future radio access may also employ bands with different levels of access regulation including exclusively licensed, non-exclusively licensed

[1] UDN is the most promising solution to address high traffic rates in wireless networks. It increases network capacity, improves link efficiency, and spectrum utilization while reducing power consumption. In fact, densification of infrastructure has already been used in current cellular networks, where the minimum distance between base stations has been drastically reduced. Nonetheless, UDN faces many technical challenges, including difficulties in mobility management and backhauling (including self-backhaul). Advanced interference and mobility management schemes need to be developed at the physical and network layers to support UDN. UDNs further require comprehensive characterization in terms of cost, power consumption, and spectrum utilization efficiency [4].

(shared), and unlicensed bands. The 5G system supports both licensed and unlicensed spectrum with a flexible and unified air interface framework [35].

The ever increasing user traffic in mobile networks necessitates increasing the amount of spectrum that may be utilized by the 5G systems. High-frequency bands in the centimeter wave and millimeter wave range are being used due to their potential for supporting wider channel bandwidths and eventually the capability to deliver ultra-high data rates. Some new spectrum below 6 GHz was allocated for mobile communication at the World Radio Conference (WRC)[2] 2015, and the bands above 6 GHz are expected to be allocated at WRC 2019 [40].

The 5G systems are expected to support diverse use cases and applications that will expand beyond the current systems. The wide range of capabilities would be tightly coupled with different use cases and applications for International Mobile Telecommunications (IMT)-2020 initiative. The main use cases of IMT for 2020 and beyond systems include the following [1]:

- *Enhanced Mobile Broadband:* This use case addresses the user-centric applications for wireless access to multimedia content, services, and data. The demand for mobile broadband services has continuously increased in the past decade, leading the industry to significantly enhance the mobile broadband capabilities. The eMBB use case encompasses new application areas (with their rigorous requirements) in addition to the existing applications that strive for improved performance and increasingly seamless user experience. This usage scenario covers a range of cases including wide-area coverage and hotspots, which have different requirements. For the hotspot case, that is, for an area with high user density, very high traffic capacity is needed, while the requirement for mobility is relaxed and user data rate is higher than that of wide-area coverage. For the wide-area coverage case, seamless coverage and medium to high mobility are desired with improved user data rates compared to the existing data rates. However, the data rate requirement may be less stringent compared to the hotspot scenario.
- *Ultra-reliable and Low-Latency Communications:* This use case has stringent requirements for capabilities such as throughput, latency, and availability. Some examples include wireless control of industrial manufacturing or production processes, remote medical surgery, distribution automation in a smart grid, and transportation safety.
- *Massive Machine Type Communications:* This use case is characterized by very large number of connected devices (e.g., a sensor network) typically transmitting relatively small payloads containing non-delay-sensitive data. These devices are required to be low cost and consume very low power.

[2] WRC, http://www.itu.int/en/ITU-R/conferences/wrc.

A large number of deployment scenarios in each category of eMBB, mMTC, and URLLC are foreseen to emerge in the next few years, which require more flexibility, programmability and configurability in the design, development and deployment of 5G access and core networks. These design features help curtail increasing operators' CAPEX and OPEX for future network upgrades. In the following sections we study various use cases and the deployment scenarios supported by 5G systems.

2.1 Use Cases

In an information society, users will be provided with wide range of applications and services, ranging from infotainment services, safe and efficient transportation, as well as new industrial and professional applications. These services and applications include very high data rates and very dense distribution of devices, with stringent requirements on the end-to-end performance and user experience. The new type of challenges that arise from the emerging application areas include very low latency, very low power, very low cost, and massive number of devices. In many cases, one of the key challenges is related to support of seamless mobility (i.e., transparent to application layer) over a wide range of speeds. Mobile broadband is the most prominent use case today, and it is expected to continue to be one of the key use cases driving the requirements for the 5G systems. The challenge goes far beyond providing the users with basic mobile Internet access and covers support of rich interactive work, media, and entertainment applications in the cloud or reality augmentations (both centralized and distributed frameworks).

5G has strived to enable new usage models for mobile networks in various vertical markets, ranging from wearable electronic devices to autonomous driving. However, the approach for spectrum usage for vertical markets seems to be still open to different solutions and the most important drivers appear to be market demand, time to market, and efficient spectrum utilization. The requirements for 5G are derived from the use cases that consider how the 5G system will provide services to the end user and what services are provided. The use cases are grouped into broad categories in order to contrast the differences in their requirements. A number of use cases are related to scenarios where the user selects from a number of network options, depending on service, performance, cost, or user preferences. When the service transfers from one network configuration to another, one of the important considerations is the degree of network interoperability.

One of the 5G promises was to provide a unified experience for users across multiple devices. From an operator's point of view, 5G also aims to unify multiple access types into a single core network as well as requirements for interoperability for one device operating on different access technologies; this also introduces requirements for service continuity between devices. 5G systems provide the user with the ability to transfer existing service sessions from one device to another device with a minimal perceptible service interruption to the user.

The modern cities of this era greatly depend on a number of critical infrastructure to function properly: electricity, water, sewer, gas, etc. Critical infrastructure monitoring is a cumbersome undertaking, often requiring service levels achievable only by dedicated wireline connectivity. For instance, in order to detect a fault in high-voltage transmission lines and be able to take corrective actions to prevent cascading failures, the required communication latency is beyond what current wireless networks can achieve. Similarly, structural monitoring requires the provisioning of a large number of low-data-rate battery-powered wireless sensors, and today's wireless networks are not optimized to support this deployment model, both in terms of battery life and cost efficiency. Also, with the massive migration of the world's population toward urban environments, there is an increasing demand for cities to modernize their infrastructure and services. From water and power management to buildings and transportation, city planners will rely on new scalable, interconnected services that do not require cities to overhaul private and public infrastructure and are built for the future. 5G systems are enabling smart cities around the world to build long-term connectivity strategies that help improve livability and sustainability.

The massive number of 5G connected devices (e.g., sensors and actuators) collects and processes data to monitor critical parameters and optimize performance based on environmental conditions. Sensors also enable service providers to detect when hidden pipes and cables need repair or when an unauthorized access occurs. 5G systems are designed to support reliable low-latency communications among densely deployed devices that are subject to power constraints with wide range of data rate requirements.

Video and audio streaming, video calls, social networking, and multimedia messaging are just some of the popular communications and entertainment applications used on today's wireless networks. In addition, new applications are emerging such as real-time multi-user gaming, virtual/augmented reality (VR/AR), 3D multi-site telepresence, ultra-high-resolution video streaming (e.g., 4K and 8K video), and photo and video sharing. These applications will require significant increase in data rate, capacity, and very low transport latency that is not supported by today's wireless networks. 5G systems provide solutions that enable continued evolution of communications and entertainment applications. There has been a great interest and early work around in the industry to enable smart, connected devices and realization of the IoT. This use case is about enabling growth through reducing the cost of connectivity of smart devices to support massive deployments, enabling carrier networks to provide connectivity for both cellular and non-cellular IoT devices, and enabling the efficient use of spectrum for device-to-device (D2D) communications.

A wide range of new and diverse use cases will need to be supported by the 3GPP ecosystem. This needs to be done at the same time that the industry continues to support the traditional mobile broadband use cases. The new use cases are expected to come with a wide range of requirements on the network operation. For example, there will be different requirements on functionalities such as charging, policy control, security, mobility, speed, and availability. Some use cases such as mobile broadband may require application-specific charging and policy control while other use cases can efficiently be managed with simpler charging or policies. Some use cases also have significant differences in performance requirements, for example, power consumption and complexity, meaning that it will not be acceptable to simply provide applications with the superset of all requirements.

In order to manage the multitude of service segments and verticals in a sustainable manner, there is a need to isolate the often intrinsically different segments from each other. For example, a scenario where a large number of consumer premises' electricity meters are malfunctioning in the network should not negatively impact the services provided to eMBB users or the healthcare and safety applications. In addition, with new verticals supported by the 3GPP community, there will also be a need for independent management and orchestration (MANO) of segments, as well as providing analytics and service exposure functionality that is customized to each vertical's or segment's need. The functionality should not be restricted to providing isolation between different segments but also to allow an operator to deploy multiple instances of the same network partition. Fig. 3 provides a high-level illustration of the concept. A slice is composed of a collection of logical network functions that support the communication service requirements of particular use case(s). Some slices will be very rich in functionality, while other slices will be very minimalist, but the network slices are not arranged in any form of hierarchy. A given network slice will simply contain the functions required for a given application or class of service(s). Devices can be directed to the appropriate network slice in a way that fulfills operator and user needs, for example, based on subscription or device type. The network slicing primarily targets a partition of the core network, but it is not excluded that the RAN may need specific functionality to support multiple slices or even partitioning of resources for different network slices.

Migration toward network slicing for 5G services necessitates careful consideration of the security requirements, isolation between slices, and the specific individual needs of each operator, whether it is the hosting platform provider or network slice tenant, in ensuring the integrity of the data processing and data stored in the network slices [45].

The presence of a very large number of IoT devices on the cellular network may require radically new technology and solutions that the current 4G networks may not be able to provide. A fundamental feature needed in order to support massive IoT deployments is scalability on the device and the infrastructure sides. Current LTE networks are designed in a manner that makes such scalability technically and financially prohibitive. 5G systems are designed to provide connectivity to an extremely large number of low-cost, low-power, low-complexity devices under mMTC usage scenario. In general, IoT devices require to operate in low-power mode during communication while maintaining a high level of reliability and coverage.

Mobility-on-demand provides a selection of options, which may be dynamically assigned to a device or application according to the device and application context, or statically configured for specialized devices and applications.

Mobility-on-demand consists of two components: (1) managing mobility of active devices and (2) tracking and reaching out to devices that support a power-saving idle mode. At the same time, the requirements on mobility support may also vary based on the applications and services. While some services require the network mobility events to be transparent to the application layer in order to avoid interruption in service delivery, other applications may have specific means to ensure service continuity. The act of concealing mobility events may include some aspects such as minimizing interruption time and packet loss or maintaining the same IP address during intra-RAT or inter-RAT cell changes.

An important part of improving user experience is to collect and analyze some information from individual users in order to understand their service priorities and perception of the network service quality. In order to allow such data collection, an end-user opt-in and control element should be provided to users to specify which aspects of service performance are the most important to them and to set their preferences for data collection in order to maintain a balance between user privacy and service quality optimization. The ecosystem around user experience monitoring should include third-party entities such as data brokers and content providers, which is further considered as part of the user control. Content providers also need to give consent to collection of data on the usage of their applications and content.

The typical usage model for the development of previous generations of cellular technologies has been centered on the telephony and wireless multimedia services with human interfaces. This type of service will continue to be important in the 5G timeframe, but cellular technology is now required to efficiently support a range of other emerging applications. The IoT is widely mentioned as an umbrella for a range of applications that include remote healthcare, smart grid, smart city, intelligent transport, and smart cars. 5G systems are required to support a wide range of IoT applications including geographically and/or power constrained, low-data devices and/or sophisticated devices with requirements for large amounts of real-time data. The IoT is an area where direct communication between devices, rather than communication via the network, may be desirable.

Augmented and virtual reality services are rapidly growing markets driven by advances in device capabilities, consumer excitement about new user experiences, and a range of practical applications. These services are expected to be demanding in their requirements for high bandwidth (particularly for virtual reality) and low latency (particularly for augmented reality). Meeting users demand for these services is expected to be an important driver for 5G.

Another emerging application area for cellular systems is in public safety and mission-critical communications for first responders and other users. 5G technology should maintain and extend the mission-critical communication technology being developed as part of LTE. This area will also have specific requirements for IoT devices and wearable technology. Users' growing reliance on mobile connectivity continues to present technical and economic challenges for the provision of adequate coverage and capacity for these critical applications.

2.2 Deployment Scenarios

The study of 5G deployment scenarios is the first step toward identifying service continuity and performance requirements. In this section, we discuss the prominent scenarios considered for 5G systems deployment. While these scenarios are similar to the use cases, they are primarily focused on the deployment issues rather than user experience. As shown in Fig. 4, the initial deployments of 5G systems will likely be predominantly in the urban areas with high user density, low mobility, and extremely high-capacity demand. Over time, network services will be expanded to include suburban and rural areas with lower user density but higher mobility and wider coverage requirements. Note that as illustrated in the figure, the urban coverage consists of macrocells and small cells where macrocells may operate at sub-6 GHz frequencies and small cells may operate at above 6 GHz frequencies (e.g., mmWave bands). As the coverage is expanded to suburban and rural areas, macrocells will likely use higher transmit powers at sub-6 GHz bands in order to provide larger coverage and extended range. As a result, multi-connectivity will play an important role in early deployments of 5G systems particularly in urban and suburban areas.

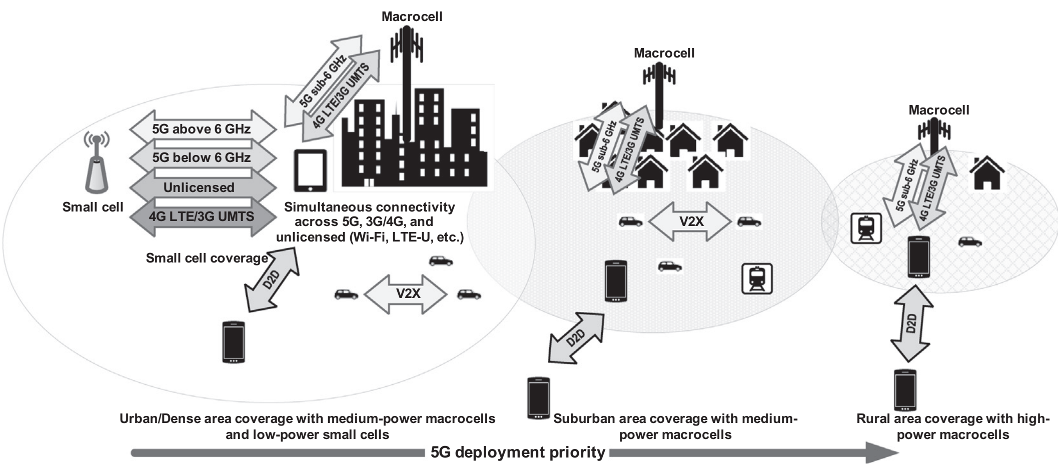

Figure 4
Multi-connectivity across frequency bands and access technologies [45].

In order to understand 5G deployment scenarios, we start our analysis with the study of high-level RAN architectures and network topologies supported by 5G. The design principles of 5G RAN architecture are based on disaggregation of the base stations and separation of the control plane (CP) and user plane (UP) entities. The architecture further supports connectivity between 3GPP LTE and 3GPP NR access and core network entities. The RAN architecture scenarios shown in Fig. 5 support co-located deployment of LTE eNBs and NR gNBs connected to either LTE core network (EPC) or 5G core (5GC) network in a non-standalone (NSA) mode as well as standalone (SA) deployment of NR disaggregated gNBs connected to 5GC. In the early deployments of 5G systems and before 5GC equipment is available, the NSA scenarios are going to be predominant and later SA deployments of 5G systems will prevail [16].

In the deployment scenarios shown in Fig. 5, gNB-CU and gNB-DU denote the central unit (CU) and distributed unit (DU) of a logical gNB, respectively, which may or may not be co-located. The logical interfaces F1 and Xn are new network reference points that comprise control plane and user plane paths between the CU and the DU(s) that will be discussed in more detail in Chapter 1. There are two gNB deployment options when considering C-RAN architectures, as shown in Fig. 6 [16,17]:

- *Collapsed gNB deployment:* In this scenario, all RAN protocols and functions are co-located within the same site. This deployment option corresponds to the current deployments of LTE systems and as such it ensures backward compatibility with the existing LTE deployments.
- *Disaggregated gNB deployment:* In this scenario, RAN protocols and functions are distributed across different sites; that is, in a C-RAN architecture, the DU and the CU may be physically apart. The CU may be further divided into CP and UP entities. The DU hosts the RLC, MAC, and PHY protocols, the CU-CP entity hosts the

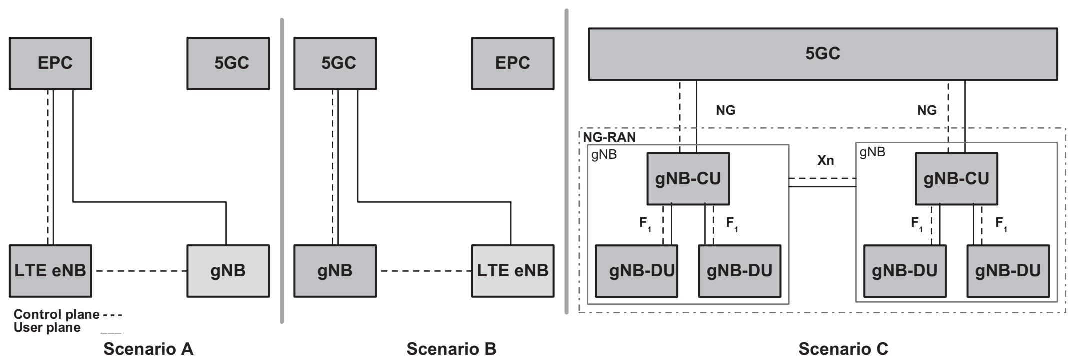

Figure 5
Different deployment scenarios of NR and LTE [16].

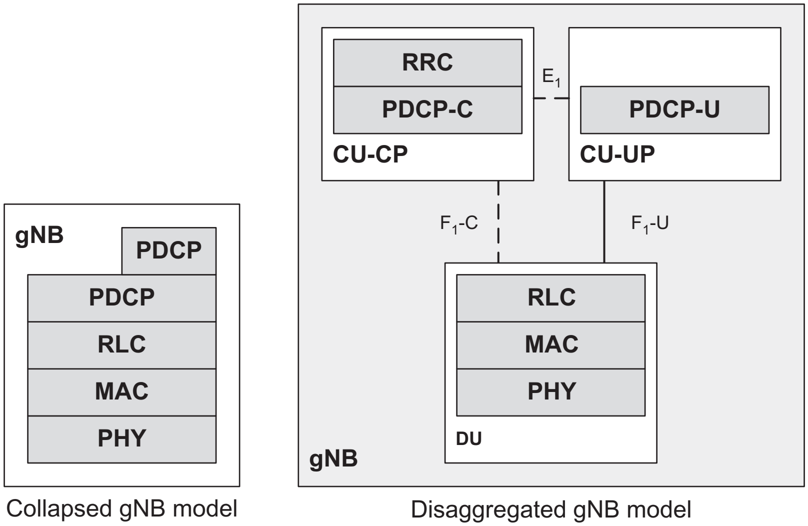

Figure 6
Different deployment options for gNB (NR base station) [16].

PDCP-C and radio resource control (RRC) protocols whereas the CU-UP entity hosts the PDCP-U (and SDAP) protocols. In the disaggregated gNB deployment, the separation of control-plane CU-CP and user-plane CU-UP entities offers the possibility of optimizing the location of different RAN functions based on the desired topology and performance. For example, the CU-CP can be placed in a location close to the DU entities. It could also be co-located with the DU, thus providing shorter latency for the critical CP procedures such as connection (re)establishment, handover, and state transition. On the other hand, the CU-UP can be centralized in a regional or national data center, thus favoring centralized implementations and providing a central termination point for the UP traffic in dual connectivity and tight interworking scenarios. An additional CU-UP can also be placed closer (or co-located) with the DU to provide a local termination point for the UP traffic for applications that require very low latency (e.g., URLLC traffic).

In the disaggregated gNB architecture shown in Fig. 6, it is necessary to coordinate the CU-CP and CU-UP entities. Some of the functions that may require a control-plane interface between CU-CP and CU-UP include CU-CP set-up, modification, and configuration of the data radio bearers (DRBs) in the CU-UP and when CU-CP entity configures the security keys in the CU-UP entity for RAN-level security activation and configuration. A new open interface between CU-CP and CU-UP has been defined to enable these functions, which is denoted as E_1. The interface E_1 is a control-plane interface, and it does not require a UP part because CU-CP and CU-UP do not exchange UP traffic.

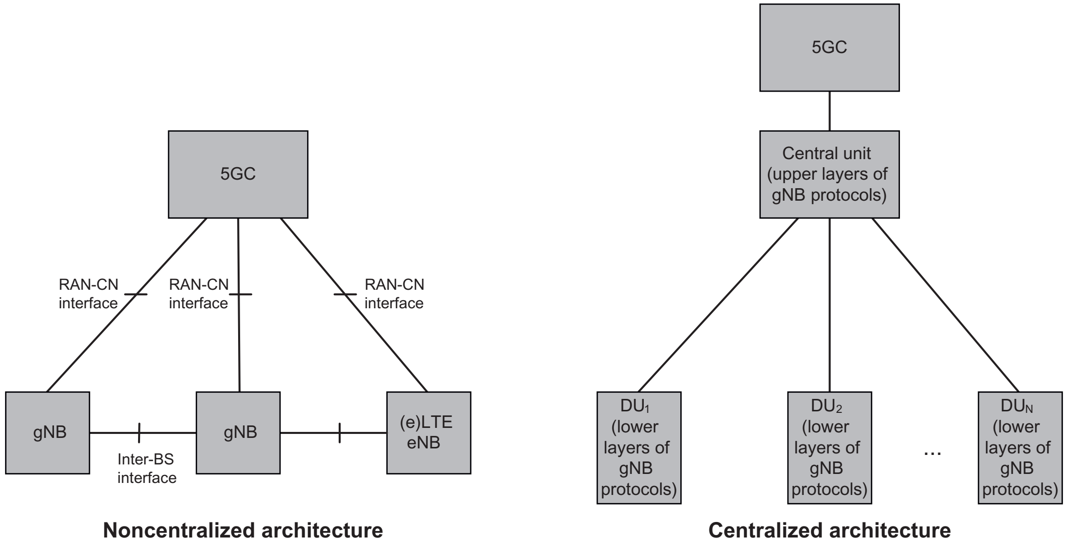

Figure 7
Illustration of the non-centralized and centralized deployment scenarios [16,17].

The new 5G RAN architecture, as shown in Fig. 7, supports the following topologies in order to enable diverse use cases and deployment scenarios [16]:

- *Non-centralized RAN architecture:* In this scenario, full user-plane and control-plane radio protocol stacks are implemented at the gNB, for example, in a macrocell deployment or an indoor (public or enterprise) hotspot environment. The gNBs can be connected to each other using any transport mechanism (e.g., Ethernet, optical fiber). However, it is assumed that the gNB is able to connect to other gNBs or eLTE eNBs[3] through standard interfaces defined by 3GPP.
- *Centralized RAN architecture:* In this scenario, the upper layers of the radio protocol stack (PDCP and RRC layers) are implemented in the CU which is typically located in the edge cloud. Different functional splits between the CU and the DU(s) are also possible, depending on the transport layer configuration. High-performance transport mechanisms, for example, optical transport networks (OTN),[4] between the CU and the lower layers of the stack located at the DUs would enable advanced coordinated multipoint (CoMP) transmission/reception schemes and inter-cell scheduling optimization, which could be useful in high-capacity scenarios, or scenarios where inter-cell

[3] The eLTE eNB is the evolution of LTE (i.e., LTE Rel-15 and later) eNB that supports connectivity to EPC and 5GC.
[4] OTN is a set of optical network elements connected via optical fiber links that are able to provide transport, multiplexing, switching, and management functions as well as supervision and survivability of optical channels carrying users' signals.

coordination is desired. The higher layers of NR radio protocol stack can be moved to the CU, if a low performance transport mechanism is used between the CU and the DU (s), because in this case the requirements on the transport layer in terms of bandwidth, delay, synchronization, and jitter are relatively relaxed.

- *Co-sited deployment with LTE:* In this scenario, the NR functions are co-located with LTE counterparts either as part of the same base station or as multiple base stations at the same site. Co-located deployment can be applicable in all NR deployment scenarios such as urban macro. As shown in Fig. 8, in this scenario it is desirable to fully utilize all spectrum resources assigned to both radio access technologies (RATs) using load balancing or connectivity via multiple RATs (e.g., utilizing lower frequencies as wide-area coverage layer for the cell-edge users).
- *Shared RAN deployment:* The NR supports shared RAN deployments, implying multiple hosted core operators. The shared RAN topology can cover large geographical areas as in the case of national or regional network deployments. The shared RAN architecture can also be heterogeneous; that is, limited to a few smaller areas, as in the case of shared in-building RANs. A shared RAN should be able to efficiently interoperate with a non-shared RAN. Each core operator may have its own non-shared RAN serving areas adjacent to the shared RAN. The mobility between the non-shared RAN and the shared RAN is supported in a manner similar to that of LTE. The shared RAN may operate either in the shared spectrum or in the spectrum of each host operator (see Fig. 8).

The cell layouts corresponding to the prominent deployment scenarios for NR are shown in Fig. 9. In the homogeneous deployment scenario, all cells (macro or small cell only) provide the same coverage, whereas in the heterogeneous deployment scenarios, cells of different sizes have overlapped coverage, that is, a combination of macro and small cells where typically macrocells provide wide-area coverage and are used as anchor points and small-

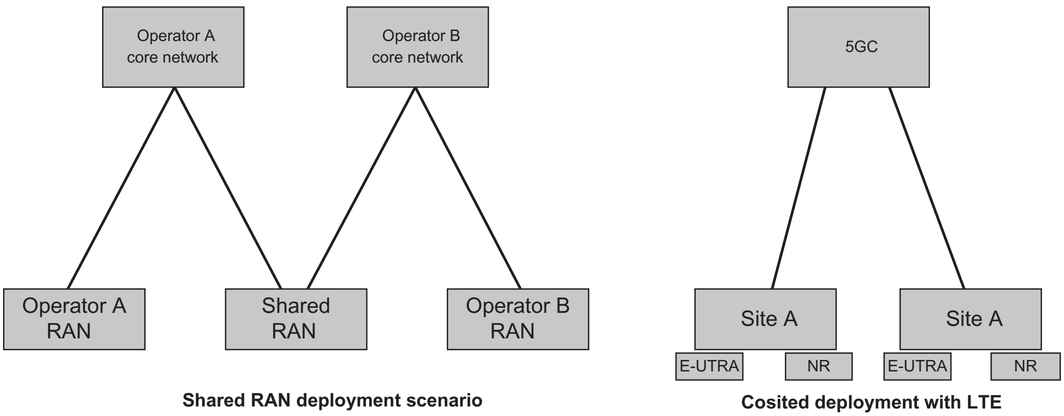

Figure 8
Cosited and shared RAN deployment scenarios [16].

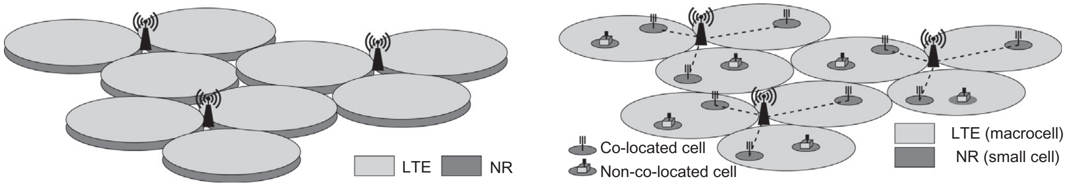

Figure 9
Co-deployment scenarios of NR and LTE [15,16].

cell access points provide high-throughput small coverage areas within macrocell coverage. In these deployment scenarios, depending on the cell layout and the location of the access nodes, both NR and LTE coverage may coexist in the same geographical area. One of the scenarios in Fig. 9 illustrates both LTE and NR cells that are overlaid and co-located, both providing similar coverage area. In this case, both LTE and NR cells are either macro or small cells. Fig. 9 further depicts a scenario where LTE and NR cells are overlaid, but not necessarily co-located, where each providing different coverage area. In this case, LTE serves macrocells and NR serves small cells. The opposite scenario is also possible which may not be practically useful for initial deployments of 5G networks. According to 3GPP definition, a co-located cell refers to a small cell together with a macro cell for which the eNBs are installed at the same location, whereas a non-co-located cell refers to a small cell alongside a macro cell for which the respective eNBs are installed at different locations.

The distributed base station model leads way to the centralized and cooperative C-RAN architectural concept where all or part of baseband functions are performed in a CU. Centralized signal processing greatly reduces the number of site equipment needed to cover the same areas served by a distributed network of base stations. Cooperative radio transmission combined with distributed antenna scheme provided by RRHs provides higher spectral efficiency and interference coordination. Real-time cloud infrastructure based on open platform and base station virtualization enables processing, aggregation, and dynamic resource allocation, reducing the power consumption and increasing the infrastructure utilization rate. The processing resources on CU can be managed and allocated dynamically, resulting in more efficient utilization of radio and network resources and improved energy efficiency (EE). Moreover, network performance is also improved significantly because collaborative techniques can be supported in C-RAN. The concept of C-RAN has been evolving in the past few years as the technology matures. For example, CU/DU functional split and next-generation fronthaul interface (NGFI) have been introduced in C-RAN to better meet the 5G requirements, for example, higher frequency, larger bandwidth, increased number of antennas, and lower latency. [23,24].

It is observed that in a typical 5G C-RAN architecture, the baseband unit often comprises a CU and one or more DUs. The principle of CU/DU functional split lies in real-time processing requirements of different functions. As shown in Fig. 10, the CU functions mainly

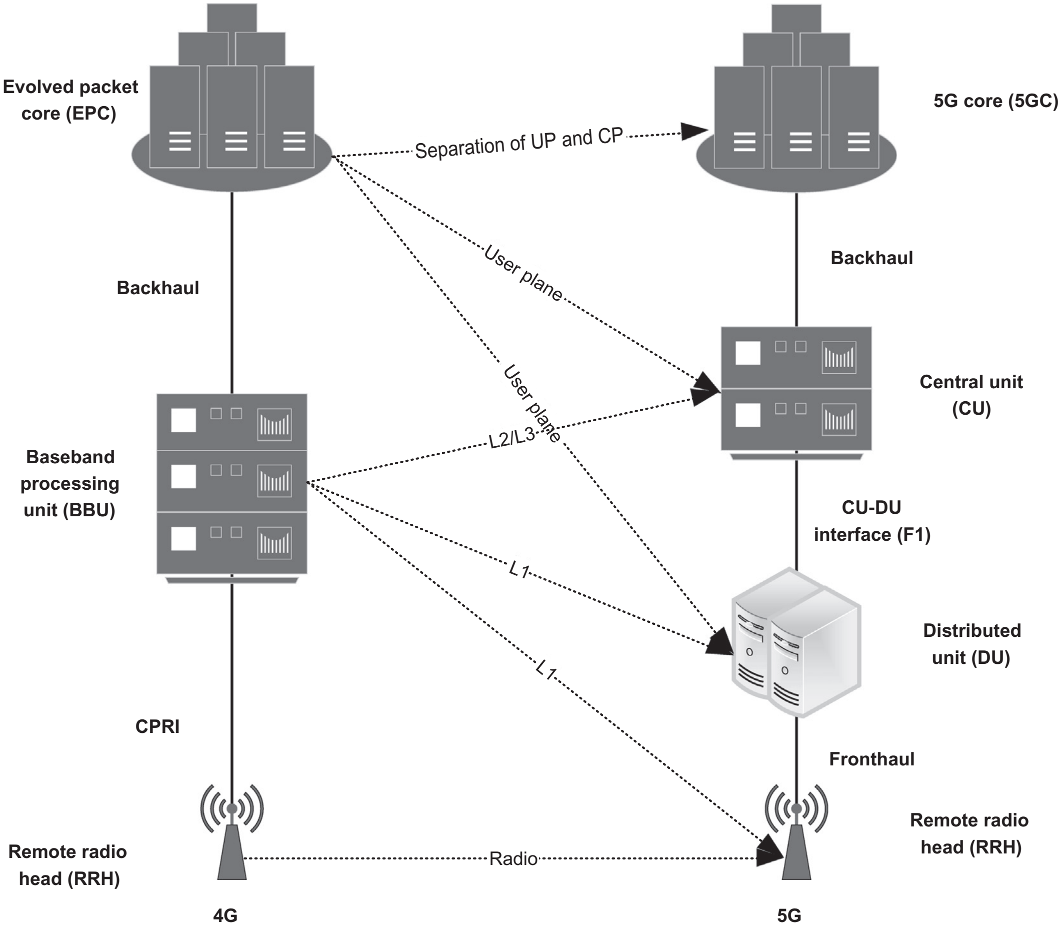

Figure 10
Evolution of single-node 4G BBU to separated CU/DU architecture in 5G [23].

include non-real-time higher layer RAN protocol processing as well as core network functions that have been moved to the edge of the network to enable mobile edge computing (MEC) services and tactile Internet. Accordingly, a DU is mainly responsible for physical layer processing and real-time processing of some or all layer 2 protocols. In order to minimize the transport capacity and latency requirements between RRU and DU(s), part of physical layer processing can be moved from DU to RRU(s). From hardware implementation point of view, the CU equipment can be developed based on general purpose processors or server platforms. The DU hardware can be implemented using customized SoCs and/or hybrid platforms, which enable real-time processing of computationally intensive functions. With NFV infrastructure, system resources (including CU and DU) can be orchestrated and managed flexibly via MANO entity, SDN controller, and traditional operating and maintenance center, which can support operators' requirements for fast service rollout.

In order to address the transport challenges among CU, DU, and RRU in 5G wireless systems, IEEE 1914 working group[5] and Common Public Radio Interface (CPRI) Forum[6] have independently specified new fronthaul interfaces.

- *IEEE P1914.1* is developing the standard for packet-based fronthaul transport networks. It is defining an architecture for the transport of mobile fronthaul traffic (e.g., Ethernet-based), including user data traffic as well as management and control-plane traffic. The group is further defining the requirements for the fronthaul networks, including data rates, timing and synchronization, and QoS. The standard also analyzes functional partitioning schemes between the remote radio units and the baseband processing units with the goal of improving fronthaul link efficiency and interoperability on the transport level, and facilitating the realization of cooperative radio functions such as massive MIMO, and CoMP transmission and reception.
- *IEEE P1914.3* standard specifies the encapsulation of the digitized complex-valued in-phase and quadrature radio signal components, and (vendor-) specific control information channels/flows into an encapsulated Ethernet frame payload field. The header format for both structure-aware and structure-agnostic encapsulation of existing digitized radio transport formats have been specified. The structure-aware encapsulation is assumed to have detailed knowledge of the encapsulated digitized radio transport format content, whereas the structure-agnostic encapsulation is only a container for the encapsulated digitized radio transport frames. The standard further defines a structure-aware mapper for CPRI frames and payloads to/from Ethernet encapsulated frames. It must be noted that the structure-agnostic encapsulation is not restricted to CPRI.
- *CPRI Forum* has developed a new specification (eCPRI) for fronthaul interface which includes increased efficiency in order to meet the foreseeable requirements of 5G mobile networks. Note that the widely used CPRI (radio over fiber) specifications were also developed by this special interest group in the past. The eCPRI specification is based on new functional partitioning of the cellular base station functions within the physical layer. The new split points enable significant reduction of the required fronthaul link capacity. The required bandwidth can scale flexibly according to the user-plane traffic transport over the Ethernet. The use of Ethernet opens the possibility to carry eCPRI traffic and other traffic simultaneously in the same switched network. The new interface is a real-time traffic interface enabling use of sophisticated coordination algorithms and guaranteeing the best possible radio performance. The eCPRI interface is meant to be future proof, allowing new feature introduction by software updates

[5] IEEE 1914 Working Group, Next Generation Fronthaul Interface, http://sites.ieee.org/sagroups-1914/.
[6] CPRI, http://www.cpri.info/.

in the radio network. In addition to the new eCPRI specification, the special interest group continues to further develop the existing CPRI specifications in order to maintain a competitive option for all deployments with dedicated fiber connections in the fronthaul including 5G wireless systems.

As shown in Fig. 11, an NGFI switch network connects CU and one or more DU(s) entities. The use of NGFI interface allows flexible configuration and deployment of CU and DU entities in different deployment scenarios. In case of ideal fronthaul, the deployment of DU can also be centralized, which can support physical layer cooperation across various transmission/reception nodes. In case of non-ideal fronthaul, the DU(s) can be deployed in a distributed manner. Therefore, a C-RAN architecture based on NGFI interface supports not only the centralized DU deployments, but also enables distributed DU deployments.

Mobile networks have traditionally been optimized for voice and data services. However, in 5G era, they have to serve a variety of devices with different characteristics and service/performance requirements. As mentioned earlier, some of the most common use cases for 5G include mobile broadband, massive IoT, and mission-critical IoT, and they all require

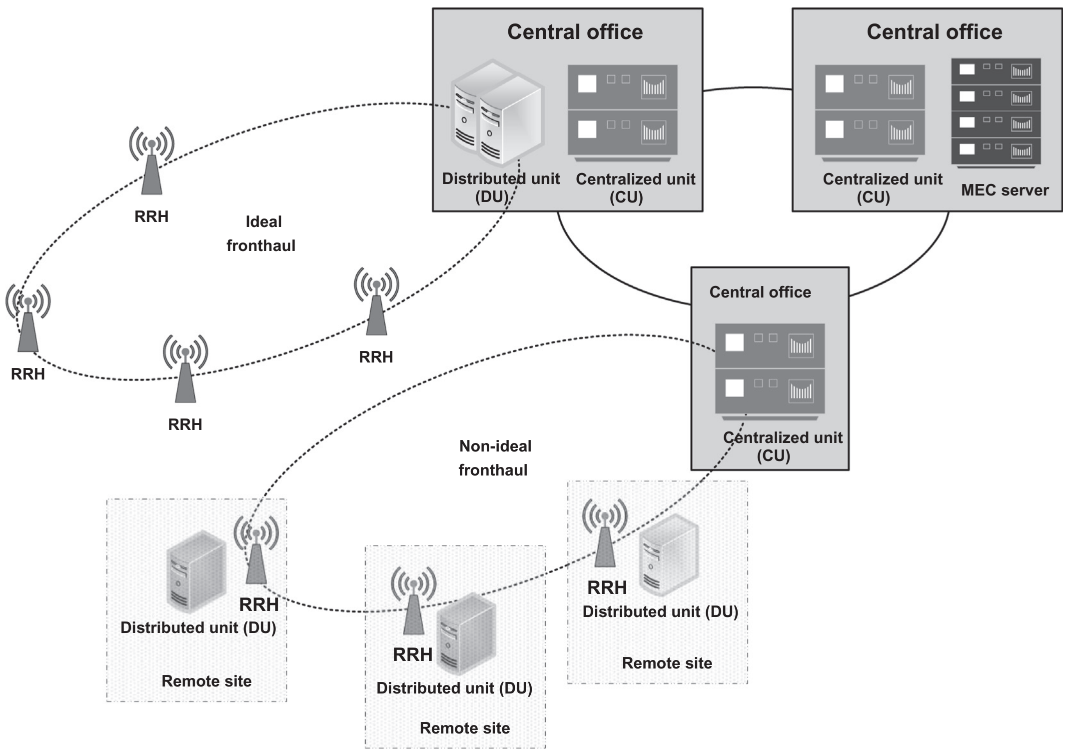

Figure 11
CU/DU-based C-RAN architecture [23–26].

different types of features and networks in terms of mobility, charging, security, policy control, latency, reliability, etc. For instance, a massive IoT service that connects stationary sensors measuring temperature, humidity, precipitation, etc. to mobile networks does not require features such as handover or location update, which have been critical in serving voice services. A mission-critical IoT service such as autonomous driving or remote-controlled robots requires a substantially low end-to-end latency in the order of a fraction of millisecond, whereas some massive MTC services are delay-tolerant.

The objective of network slicing is to transform the network from using a rather static one-network-fits-all approach to a dynamic model where multiple logical networks can be created on top of the same physical infrastructure to fulfill the diverse needs of different use cases. Network slicing is a concept that enables the operators to create service-customized (logical) networks in order to provide optimized solutions for different services which have diverse requirements in terms of functionality, performance, and isolation. The fundamental assumption in this solution is that slicing of an operator's network should be transparent to the UEs at the radio interface level. A network slice is a logical network that includes a set of network functions, which may be virtual network functions (VNFs) or physical network functions, and corresponding resources, such as compute, storage, and networking resources. A slice can also be seen as a unique profile for an application, defined as a set of services within the network to support a given use case, traffic type, or a customer. A slice may serve a particular purpose or a specific category of services (e.g., a specific use case or a specific traffic type) or even individual customers, in which case it may be created on-demand using a network-as-a-service (NaaS)[7] approach. Each network slice exists within an SDN overlay. Within the slice, a collection of network functions are chained together according to the requirements of the service using the slice. Each slice sharing the same underlying physical infrastructure is a logically separated and isolated system that can be designed with different network architecture, hardware/software and network provisioning. In some cases, network slices can share functional components.

To implement network slices, the NFV is a prerequisite. The main idea of NFV is to implement network functions in software (i.e., packet core and selective radio access functions) mapped to virtual machines (VMs)[8] that are run on commercial server platforms as opposed to processing over dedicated network equipment. In this case, the access network works as

7 NaaS is a business model for delivering network services virtually over the Internet on a pay-per-use or subscription basis.

8 A VM is an operating system or application environment that is installed on software, which emulates dedicated hardware. The end users have the same experience on a virtual machine as they would have on dedicated hardware. Virtual machines do not require specialized, hypervisor-specific hardware. Virtualization does, however, require more bandwidth, storage, and processing capacity than a traditional server or desktop if the physical hardware is going to host multiple running virtual machines. VMs can easily be moved, copied, and reassigned between host servers to optimize hardware resource utilization.

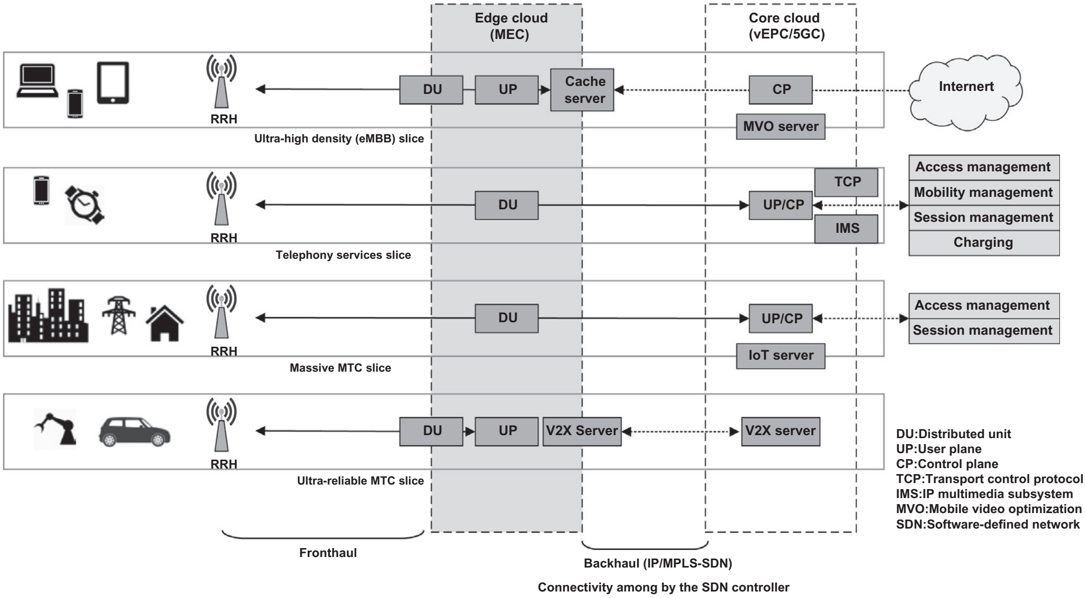

Figure 12
Illustration of network slicing concept [43].

edge cloud, whereas the core network operates as core cloud. Connectivity among VMs located on the edge and core clouds are provisioned using SDN. Network slices are automatically configured and created for each service (i.e., voice service slice, massive IoT slice, mission-critical IoT slice). Fig. 12 shows how applications dedicated for each service can be hypothetically virtualized and implemented in each slice. In the example shown in the figure slices that have been configured for the specific services contain the following components [43]:

- *Mobile broadband slice:* All virtualized DU, 5GC user-plane, and cache server in the edge cloud, and virtualized 5GC control-plane as well as mobile video optimization (MVO)[9] server in the core cloud.
- *Voice services slice:* 5GC user-plane and control-plane functions with full mobility support features (e.g., access management, session management, mobility management, and charging), and IP multimedia subsystem and TCP optimization VNFs running in a core cloud.

[9] MVO covers the use of technologies and solutions that enable mobile network operators (MNOs) to optimize the delivery of video content through a host of video optimization techniques, prioritization policies, user-based customization, and intelligent management of overall traffic on their data networks. MVO aims at maximizing the quality of video services and enhancing end user experience, while enabling MNOs to monetize the mobile video services, both operators' own and third-party over-the-top services.

- *Massive MTC slice:* Light-duty 5GC without mobility management functions in the core cloud.
- *Mission-critical IoT slice:* 5GC user-plane and the associated servers (e.g., V2X server) are all located in the edge cloud to minimize transport delay.

Dedicated slices are created for services with different requirements and virtualized network functions are assigned to different locations in each slice (i.e., edge cloud or core cloud) depending on the services. Some network functions, such as charging, policy control, might be essential in one slice but unnecessary in other slices. Operators can customize network slices in the most cost-effective manner.

In the RAN side the slicing can be realized based on grouping of physical radio resources or logical radio resources that are abstracted from physical radio resources. RAN slicing can be implemented by mapping a slice identifier to a set of configuration rules applied to the RAN control-plane and user-plane functions. Some network functions, such as mobility management, can be common to several slices. There will also be common control functions that coordinate RAN resource usage among the slices. Radio slices in the RAN may share radio resources (time, frequency, code, power, and space) and the corresponding communication hardware such as digital baseband processing components or analog radio components. The sharing may be done in a dynamic or static manner depending on the configuration rules of the network slice. In the case of dynamic sharing, each slice obtains resources based on its demand and priority using either scheduling (that is, the slice requests resources from a centralized scheduler, which allocates resources based on the overall traffic load or the priority of the slice) or contention. In the case of static sharing, a slice is pre-configured to operate with a dedicated resource throughout its operation time. Static sharing enables guaranteed resource allocation to the slice, whereas dynamic resource sharing allows overall resource usage optimization. Slice-specific configuration rules for the control-plane functions adapt the RAN control-plane functions for each slice. This is because slices may not require all control-plane functions. There is also a need for network slice-specific admission control, so that the system can meet the initial access requirements of various network slices.

3 Key Performance Indicators, Architectural, System, and Service Requirements

Unlike previous generations of 3GPP systems that attempted to provide a general-purpose framework, the 5G systems are expected to provide customized and optimized support for a variety of different services, traffic patterns, and end-user demographics. Several industry-led 5G white papers and in particular the one from Next-generation Mobile Network[10] Alliance [14,25,32,35] described a multi-purpose 5G system capable of simultaneously

[10] Next-Generation Mobile Network Alliance, http://www.ngmn.org/.

supporting various combinations of requirements such as reliability, latency, throughput, positioning, and availability to enable different use cases and deployment scenarios. Such revolutionary systems would be achievable with introduction of new technologies, both in access network and the core network, with configurable, flexible, and scalable assignment of network resources. In addition to increased flexibility and optimization, a 5G system needs to support stringent key performance indicators (KPIs) for latency, reliability, throughput, etc. Enhancements in the radio access network contribute to meeting these KPIs as do core network improvements such as network slicing, in-network caching, and hosting services closer to the end nodes [21]. Flexible network operations are the cornerstone of the 5G systems and the new features that allow this flexibility include NFV, network slicing, SDN, network scalability, and seamless mobility. Other network operational requirements address the indispensable control and data plane resource allocation efficiencies as well as network configurations that optimize service delivery by adopting optimal routing strategies between end users and application servers. Enhanced charging and security mechanisms manage new types of UEs connecting to the network in different ways.

The enhancements in mobile broadband access mechanisms aim to meet a number of new KPIs, which pertain to high data rates, high user density, high user mobility, highly variable data rates, flexible deployment options, and improved coverage. High data rates are driven by the increasing use of data for services such as streaming (e.g., video, music, and user generated content), interactive services (e.g., augmented reality), and some IoT use cases. The availability of these services impose stringent requirements on user experienced data rates and latency in order to meet certain QoS requirements. In addition, increased coverage in densely populated areas such as sport arenas, urban areas, and transportation hubs has become an essential demand for pedestrians and users in vehicles. New KPIs related to traffic and connection density enable transport of high volume of user traffic per area (traffic density) and transport of data for high number of connections (UE density or connection density). Many UEs are expected to support a variety of services which exchange either a very large (e.g., streaming video) or very small (e.g., sensor data burst) amount of data. The 5G systems are required to manage this service variability in a resource efficient manner. All of these cases introduce new deployment requirements for indoor and outdoor, local area connectivity, high user density, wide-area connectivity, and UEs travelling at high-speed scenarios.

Another aspect of 5G KPIs includes requirements for various combinations of latency and reliability, as well as higher accuracy for positioning. These KPIs are driven by support of commercial and public safety services. On the commercial side, industrial control, industrial automation, control of aerial vehicles, and the augmented/virtual reality are examples of such services. Services such as aerial vehicle control will require more precise positioning information that includes altitude, speed, and direction, in addition to planar coordinates. Support for massive connectivity brings many new requirements in addition to those for the enhanced KPIs. The expansion of connected things paradigm introduces need for significant improvements in resource efficiency in all system components (e.g., UEs, IoT devices, access network, and core network). The 5G system also targets enhancement of its capabilities to meet the KPIs that the emerging V2X applications require. For these advanced applications, the requirements, such as data rate, reliability, latency, communication range, and speed, become more stringent. These applications have specific technical requirements that need to be addressed through sophisticated design of the 5G radio interface and the use of appropriate frequency bands in order to ensure consistent satisfaction of their requirements.

3.1 Definition of the Performance Metrics

In order to quantify how certain technical solutions would affect end-user experience or what would be the 5G system performance in a desired use case, specific performance metrics are needed for evaluation. This section provides definitions of 5G main characteristics and KPIs, similar to the ones defined in references [1,14]. The following section provides the definitions of the performance metrics that have been used to measure and benchmark the performance of 5G systems. These metrics are related to the KPIs that describe the main 5G use cases and certain deployment scenarios.

Peak Data Rate: Described in bits/s, this is the maximum (theoretically) achievable data rate under error-free conditions at the physical layer. It denotes the received data bits allocated to a single user terminal, when all assignable radio resources for the corresponding link direction are utilized excluding the overhead resources which include radio resources that are used for physical layer synchronization, reference signals, guard bands, and guard times. The peak data rate is defined for a single mobile station. For a single frequency band, it is related to the peak spectral efficiency in that band. Let W denote the channel bandwidth and SE_p denote the peak spectral efficiency in that band. Then the user peak data rate R_p is given by

$$R_p = SE_p W$$

The peak spectral efficiency and available bandwidth may have different values in different frequency ranges. If the bandwidth is aggregated across multiple bands, the peak data rate will be summed over those bands. Therefore, if bandwidth is aggregated across N bands, then the total peak data rate is defined as follows:

$$R_p = \sum_{n=1}^{N} SE_p^n W_n$$

where W_n and SE_p^n represent the nth component bandwidth and its peak spectral efficiency, respectively.

Average Spectral Efficiency: Let $R_n\,(T)$ denote the number of correctly received bits by the nth user in the downlink or from the nth user in the uplink. It is assumed that there are N users (or devices) and N_{TRP} transmission/reception points (TRPs)[11] in the system within the coverage area. Furthermore, let W denote the channel bandwidth and T the time interval over which the data bits are received. The average spectral efficiency may be estimated by running system-level simulations over N_{drops} (number of drops), where each drop results in certain rate denoted as $R^{(1)}(T),\ldots,R^{(N_{drops})}(T)$, subsequently

$$SE_{avg} = \frac{1}{N_{drops}TWN_{TRP}}\sum_{n=1}^{N_{drops}} R^{(n)}(T) = \frac{1}{N_{drops}TWN_{TRP}}\sum_{n=1}^{N_{drops}}\sum_{t=1}^{N} R_i^{(n)}(T)$$

where SE_{avg} is the estimated average spectral efficiency that approaches the actual average value with increasing number of drops, and $R_i^{(n)}(T)$ is the total number of correctly received bits by the ith user in the nth drop. The average spectral efficiency is evaluated through system-level simulations using the evaluation configuration parameters of Indoor Hotspot-eMBB, Dense Urban-eMBB, and Rural-eMBB test environments defined in references [9,10] while the temporal and spectral resources consumed by layers 1 and 2 overhead are taken into consideration. Examples of layer 1 overhead include the time/frequency resources used by synchronization signals, guard band and DC subcarriers, guard intervals or switching times, reference signals, and cyclic prefixes. Example layer 2 overheads include time/frequency resources utilized to carry common control channels, HARQ ACK/NACK signaling, CSI/CQI feedback, random access channel, payload headers, and CRC. Power allocation and/or boosting should also be accounted for in modeling resource allocation for control channels.

5th Percentile User Spectral Efficiency: This is the 5% point of the cumulative distribution function (CDF)[12] of the normalized user throughput, calculated at all possible user locations. Let's assume that the ith user in the nth drop can correctly decode $R_i^{(n)}(T)$ accumulated bits in the time interval $[1,T]$. For the non-scheduled duration of user i, zero bits are accumulated. During this total time, user i receives accumulated service time of $T_i \leq T$, where the service time is the time duration between the first packet arrival and when the

[11] TRP is defined as physical node, located at a specific geographical location within a cell, with an antenna array that consists of one or more antenna elements that are available for the network operation.

[12] The distribution function $D(x)$, also called the CDF or cumulative distribution function, describes the probability that a random variable X takes on a value less than or equal to a number x. The distribution function is sometimes also denoted as $F_X(x)$. The distribution function is therefore related to a continuous probability density function $p(x)$ by $D(x) = P(X \leq x) = \int_{-\infty}^{x} p(\tau)d\tau$. As a result, $p(x)$ (when it exists) is simply the derivative of the distribution function $p(x) = \partial D(x)/\partial x$. Similarly, the distribution function is related to a discrete probability $p(x)$ by $D(x) = P(X \leq x) = \sum_{X \leq x} p(x)$. There exist distributions that are neither continuous nor discrete.

last packet of the burst is correctly decoded. In the case of full buffer traffic model $T_i = T$. Therefore, the rate normalized by time and bandwidth of user i in drop n is given as

$$r_i^{(n)} = \frac{1}{T_i W} R_i^{(n)}(T)$$

By running the system-level simulations N_{drops} times, $N_{drops}\ N$ values of $r_i^{(n)}$ are obtained of which the lowest 5% point of the CDF is used to estimate the fifth percentile user spectral efficiency. The fifth percentile user spectral efficiency is evaluated by system-level simulations using the evaluation configuration parameters of Indoor Hotspot-eMBB, Dense Urban-eMBB, and Rural-eMBB test environments. It should be noted that the fifth percentile user spectral efficiency is evaluated by system-level simulations only using a single-layer layout configuration even if the test environment comprises multi-layer layout configuration. Furthermore, the fifth percentile user spectral efficiency is evaluated using identical simulation assumptions as the average spectral efficiency for that test environment taking into account the layers 1 and 2 overhead. Note that for time division duplex (TDD) systems, the effective bandwidth is used which is the operating bandwidth normalized by the uplink/downlink ratio.

User Experienced Data Rate: This is the 5% point of the CDF of the user throughput. The user throughput during active state is defined as the number of correctly received bits, that is, the number of bits contained in the service data units (SDUs) delivered to network layer over a certain period of time. In the case of single frequency band and a single TRP layer, the user experienced data rate can be derived from the fifth percentile user spectral efficiency. Let W and SE_{user} denote the system bandwidth and the fifth percentile user spectral efficiency, respectively, then the user experienced data rate denoted by R_{user} is given as $R_{user} = SE_{user} W$. In the case where the system bandwidth is aggregated across multiple bands (one or more TRP layers), the user experienced data rate will be summed over the bands.

The user experienced data rate is evaluated with non-full-buffer traffic assumption. For non-full-buffer traffic, user experienced data rate is fifth percentile of the user throughput. The user throughput during active time is defined as the size of a burst divided by the time between the arrival of the first packet of a burst and the reception of the last packet of the burst. It must be noted that the 5% user spectrum efficiency depends on the number of active users sharing the channel and for a fixed transmit power; it may vary with the system bandwidth. A dense network architecture would increase the 5% user spectrum efficiency and results in fewer users sharing the channel, that is, 5% user spectrum efficiency may vary with the site density. Support of very large system bandwidths in 3GPP 5G systems would further increase the user experienced data rates. The target 5% user spectrum efficiency is three times the IMT-Advanced minimum requirements [8].

Area Traffic Capacity: This metric represents the system throughput over a unit of geographic area and can be evaluated using either a full-buffer traffic model or a non-full-buffer traffic model. The area traffic capacity is described in bits/s/m^2. Both the user experienced data rate and the area traffic capacity must be evaluated at the same time using the same traffic model. The throughput is defined as the number of correctly received bits, that is, the number of bits contained in the SDUs delivered to the network layer over a certain period of time. It can be derived for a particular deployment scenario which utilizes one frequency band and one TRP layer based on the achievable average spectral efficiency, TRP density, and system bandwidth. Let W and $d_{TRP} = N_{TRP}/S$ denote the system bandwidth and the TRP density, respectively, where S is the coverage area. The area traffic capacity C_{area} is proportional to the average spectral efficiency SE_{avg} and can be expressed as follows:

$$C_{area} = SE_{avg} d_{TRP} W$$

In case the system bandwidth is aggregated across multiple frequency bands the area traffic capacity will be accumulated over those bands.

In other words, the area traffic capacity is a measure of traffic a network can deliver over a unit area. It depends on site density, bandwidth, and spectrum efficiency. In the case of full-buffer traffic and a single-layer single-band system, the previous equation may be expressed as

$$\begin{aligned}&\textit{area traffic capacity}\ (\mathrm{bps}/m^2) = \\ &\textit{site density}\ (\mathrm{site}/m^2) \times \textit{bandwidth}\ (\mathrm{Hz}) \times \textit{spectrum efficiency}\ (\mathrm{bps/Hz/site})\end{aligned}$$

Several techniques have been utilized in 3GPP NR to achieve high spectrum efficiency and thereby to improve area traffic capacity that will be discussed in the next chapters. To this end, spectrum efficiency improvements in the order of three times of IMT-advanced are targeted [8]. Furthermore, 3GPP has strived to develop standards in order to support large system bandwidths in the excess of 1 GHz (aggregated bandwidth) in mmWave spectrum bands. The available bandwidth and site density, which both have direct impact on the available area capacity, vary in different parts of the world and for different operators.

Connection Density: This metric is defined as the total number of devices served by a network within a geographical area where the service satisfies the requirements of a specific QoS using a limited bandwidth and a limited number of TRPs. The target QoS is set such that the delivery of a payload of a certain size within a certain time and with a certain success probability can be guaranteed. This requirement was defined for the purpose of evaluation in the mMTC use case. More specifically, the connection density represents the total number of terminals that can achieve a target QoS per unit geographical area, where the target QoS is to ensure that the system packet drop rate remains less than 1% at a given packet inter-arrival time and payload size. Connection density is proportional to the connection efficiency, the channel bandwidth and TRP density. Let W denote the channel bandwidth and d_{TRP} the TRP density (i.e., number of TRPs per square kilometer). Then connection density $d_{connection}$ can be expressed as follows:

$$d_{connection} = \alpha d_{TRP} W$$

The connection efficiency α is evaluated with a small packet model. Considering the number of users/devices N in the coverage area with average inter-packet time interval T_{packet}, the average number of packets arrival in 1 second denoted as $N_{packet} = N/T_{packet}$. The system must guarantee that the percentage of packets in outage is less than 1%, where a packet is defined to have experienced outage, if the packet fails to be delivered to the destination within a permissible packet delay bound of 4 milliseconds. The packet delay is measured from the time when a packet arrives at the transmitter buffer to the time when it is successfully decoded at the receiver.

According to the IMT-2020 evaluation methodology [10], in order to compute the connection efficiency, one must run system-level simulations by assuming certain number of devices N, packet arrival rate N_{packet}, number of TRPs N_{TRP} and system bandwidth W. One then generates user traffic packets according to the small packet model specified in reference [10] with packet arrival rate N_{packet}. The system-level simulations are run again to obtain the packet outage rate. The number of devices will be iterated until the number of devices satisfying the packet outage rate of 1% is reached, which is denoted as $N_{capacity}$. The connection efficiency is then calculated as $\alpha = N_{capacity} W/N_{TRP}$.

User-Plane Latency: The one-way user-plane latency is the defined as delay introduced by the radio access network during transmission of an IP packet from the source to the destination. The user-plane latency also known as transport delay is the transmission time between an SDU packet being available at the IP layer in the source UE/BS and the availability of this packet at IP layer in the destination BS/UE depending on the link direction. The user-plane packet delay includes delay introduced by associated protocols and control signaling in either uplink or downlink direction assuming the user terminal is in the connected mode and the network operates in unloaded condition. This requirement is defined for the purpose of evaluation in the eMBB and URLLC use cases. The breakdown of the user-plane latency components is as follows [10]:

1. UE/BS (transmitter) processing delay (typically one subframe)
2. Frame alignment (typically half of subframe duration)
3. Number of transmission time intervals (TTIs)[13] used for data packet transmission, which typically includes UE scheduling request and access grant reception times.

[13] The transmission time interval (TTI) is defined as the duration of the transmission of the physical layer encoded packets over the radio air-interface. In other words, the TTI refers to the length of an independently decodable transmission on the air-link. The data on a transport channel is organized into transport blocks. In each TTI, at most one transport block of variable size is transmitted over the radio air-interface to or received from a terminal in the absence of spatial multiplexing. In the case of spatial multiplexing, up to two transport blocks per TTI are transmitted.

The number of TTIs used for each packet transmission depends on channel quality, allocated frequency resource, and the use of multi-connectivity. In case of user-plane multi-connectivity, this delay component should be derived with respect to different multi-connectivity configurations, that is, whether different data streams are transmitted over different links or multiple links are simply used for redundant data transmission. In 5G, both transmitter and receiver can be user devices considering D2D communication. In D2D communications, the user terminal may need some time for downlink/uplink synchronization.

4. HARQ retransmission time which is typically calculated assuming 10%–30% error rate for transmission over the air prior to HARQ. Both control-plane and user-plane multi-connectivity impact this delay component.
5. BS/UE (receiver) processing delay (typically two to four subframes)

Control-Plane Latency: The control-plane latency is the time for the UE to transition from the IDLE state to the CONNECTED state. In other words, the time that it takes to establish the control-plane and data bearer for data transmission over the access link excludes downlink paging delay and backhaul signaling delay. The control-plane latency includes random access procedure duration, uplink synchronization time, connection establishment and HARQ retransmission interval, and data bearer establishment inclusive of HARQ retransmissions. More specifically, the breakdown of control-plane latency components is as follows [10,39]:

1. UE wakeup time: Wakeup time may significantly depend on the implementation. 3GPP NR has introduced an intermediate state referred to as INACTIVE state in addition to the LTE IDLE and CONNECTED states, for the purpose of control-plane latency reduction and improving device energy consumption. The new NR UE state provides a broadly configurable discontinuous reception (DRX) and thus contributes to different control-plane latencies for different traffic patterns and power saving modes. Since UE can be configured by the network with different DRX cycles in different conditions, this delay component is better characterized through simulation.
2. Downlink scanning and synchronization + broadcast channel acquisition: This step may further require taking into consideration the beam tracking procedures in the terminal side. On the other hand, 5G has introduced different forms of multi-connectivity which would allow skipping this step, for example, the broadcast information for the idle link can be delivered over one of the active links where UE is able to receive. With different configuration of multi-connectivity, broadcast information for the idle link might be delivered in different ways. In the case of control-plane/user-plane decoupling, detection of cells providing the user-plane needs to be taken into account. Note that the periodicity of certain common signals/channels for access may vary over different links.
3. Random access procedure: The delay due to the use of random access (physical channel) preamble for transmission of small payloads needs to be considered. Furthermore,

in case of mMTC traffic, the delay due to possible collision of random access preambles and their retransmissions must be taken into consideration.

4. Uplink synchronization: During the NR feasibility studies, it was observed that some candidate waveforms may relax the requirements for uplink synchronization by allowing asynchronous uplink access. In certain cases, the delay due to capability negotiation procedure and the associated HARQ retransmissions can also avoided, if the UE capability information is already available at the base station. In the case of control-plane/user-plane decoupling in two or more cells, the information about the capabilities of user-plane and control-plane needs to be obtained prior to the uplink access.
5. RRC connection establishment/reconfiguration + HARQ retransmissions: In some new 5G multi-connectivity scenarios, this step is considered as complete when the RRC connection is established with the target radio access network. When data bearers are split/aggregated at the core network, the RRC connection setup is required across all of those radio links.
6. Authorization and authentication/key exchange + HARQ retransmissions: Security information may already be available in the intermediate state introduced by 5G NR. Nevertheless, it is important to consider whether the security context has been discarded during the transition between the UE states.
7. Registration with the anchor gNB + HARQ retransmissions: In the case of user-plane/control-plane split, the UE often registers with the cell that provides the control plane. In that case, when user-plane and control-plane belong to different radio access networks, the UE may also register with both cells. In the case of control-plane establishment in multi-connectivity, the UE may register with multiple cells which provide control-plane functionalities. If the radio access network does not require registration, this component can be omitted, assuming that the context can be retrieved from the previous session.

Mobility: Mobility is the maximum UE speed at which certain QoS can be achieved or sustained. There are four mobility classes defined in IMT-2020 as follows:

- Stationary: 0 km/h
- Pedestrian: 0−10 km/h
- Vehicular: 10−120 km/h
- High-speed vehicular: 120−500 km/h

The upper range of high-speed vehicular mobility class is mainly considered for high-speed trains. A mobility class is supported, if the minimum required data channel link-level spectral efficiency on the uplink can be achieved. This assumes the user is moving at the maximum speed in that mobility class in each of the test environments. The mobility class compliance is evaluated under Indoor Hotspot-eMBB, Dense Urban-eMBB, and Rural-eMBB test environments. In order to verify satisfaction of the mobility requirement, one should run system-level simulations, identical to those for calculating the average spectral efficiency, at the maximum speed of each mobility class, starting with link-level simulations and an appropriate link-to-system mapping (also known as physical layer abstraction), in order to collect overall statistics for uplink SINR values (i.e., SINR distribution or geometry), and to calculate the CDF corresponding to those values for each test environment. The CDF for the test environment(s) is used to find the respective 50th percentile of SINR values. A new link-level simulation in the uplink is run for the selected test environment(s) under either NLoS or LoS propagation conditions with the purpose of obtaining link-level data rate and the residual packet error rate as a function of SINR. The link-level simulation uses air interface configuration(s) supported by the NR while taking into consideration the HARQ retransmissions, channel estimation, and phase noise impairments. The link-level spectral efficiency values (link data rate normalized by channel bandwidth) obtained from the previous step are compared using the set of SINR values obtained earlier for selected test environments. The mobility requirement is satisfied, if the value of the spectral efficiency is larger than or equal to the corresponding threshold value and if the residual decoded packet error rate is less than 1%, for all selected test environments. For the selected test environment, it is sufficient if one of the spectral efficiency values (of either NLoS or LoS propagation model) satisfy the minimum requirement. A similar methodology can be used for the assessment of the mobility requirement in the downlink.

Reliability: Reliability refers to the capability of transmitting certain amount of user traffic within a predetermined time duration with high probability of success. In other words, reliability is the success probability of transmitting a layer 2/3 packet within a permissible time, which is the time it takes to deliver the small payload from the source layer 2 SDU ingress point to the destination layer 2/3 SDU egress point under certain channel conditions. This metric was defined for the purpose of system evaluation in the URLLC usage scenario. The minimum requirement for the reliability is a success probability in the order of $1-10^{-5}$ for transmitting a layer 2 protocol data unit of 32 bytes (i.e., 20 bytes application data + protocol overhead) within 1 ms in channel quality corresponding to the edge of the coverage (cell-edge conditions) for the Urban Macro-URLLC test environment. Note that target coverage and reliability requirement depend on the deployment scenario and type of operation (e.g., V2X with certain average inter-vehicle speed and distance).

In order to verify the compliance, one must run system-level simulations, identical to those for average spectral efficiencies. Using link-level simulations and an appropriate link-to-system mapping for the desired mobility class and the test environment(s), overall statistics for (downlink and/or uplink) SINR values are obtained and the CDF of SINR distribution is calculated for each test environment. Using a new downlink and/or uplink link-level simulation, the residual packet error rate within the maximum delay time as a function of SINR is calculated. The satisfaction of reliability requirement is ensured, if the fraction of messages that are correctly delivered within the permissible delay bound is larger than or equal to the required success probability.

Energy Efficiency: Network energy efficiency is the capability of a network to minimize the energy consumption for radio access (and/or core) network operations without affecting the performance of the network. Device energy efficiency (RAN aspects) is the capability of the network to minimize the power consumed by a user device communication subsystem (modem). Network and the device energy efficiencies are related in the sense that they mutually affect efficient data transmission in a loaded network. The energy consumption of a device can be reduced when there is no data transmission to or from the device. Efficient data transmission in a loaded network is demonstrated by the average spectral efficiency. The device energy consumption can be estimated based on the ratio of sleep and wake cycles of the device. The sleep ratio is the fraction of unoccupied time slots (for the network) or sleeping time (for the device) in a period of time corresponding to the cycle of the control signaling (for the network) or the periods of DRX (for the device) when no user data transfer takes place. Furthermore, the sleep duration, that is, the continuous periods of time with no transmission (for network and device) and reception (for the device), should be sufficiently long.

Network energy efficiency (both quantitative and qualitative aspects) is considered as a basic principle in the NR system design. This is the capability to minimize energy consumption for radio access while providing the highest possible area traffic capacity. The target is to design the system with the ability to efficiently deliver data, and the ability to provide sufficiently granular transmission cycles, so that transmission can be discontinued when there are no data to transmit. It further includes the ability to provide the operator with sufficient flexibility to adapt sleep cycles of base stations depending on the load, services, and coverage area. In order to quantitatively express and compare this KPI, the following network energy efficiency metric in bits per Joule is defined. This definition would allow quantitative comparison of different solutions or energy-saving mechanisms on the basis of efficiency merits, when their impact is not obvious from qualitative analysis. In order to evaluate the overall improvements achieved in 3GPP NR system, the network energy efficiency metric is defined as follows:

$$EE_{global} = \sum\nolimits_{scenario_k} b_k EE_{scenario_k}$$
$$EE_{scenario_l} = \sum\nolimits_{load\ level_l} \frac{a_l V_l}{EC_l}$$

In the above equation b_k refers to the weights of each deployment scenario where the network energy efficiency is evaluated. Furthermore, V_l, EC_l, a_l denote the traffic per second served by a base station (in bits/s); the power consumed by a base station to serve V_l in Watt = Joule/s, and the weight for each traffic load level, respectively. In other words

$$EE_{scenario_k} = \sum_{load\ level_k} a_k \frac{V_k(\text{traffic per second served by a base station})}{EC_k(\text{power consumed by a base station to serve} V_k)} \quad \text{Bits/Joules}$$

The UE battery life can be measured as the total time that a UE can operate with one charge. For mMTC, the UE battery life in extreme coverage is based on the activity of the mobile-originated data transfer consisting of 200 bytes of uplink traffic per day followed by 20 bytes of downlink traffic at maximum coupling loss of 164 dB (a coverage criterion), assuming that the UE battery has a power-rating of 5 Watts-hour. The desired UE battery life for mMTC use cases is between 10 and 15 years. This metric is analytically evaluated. The UE energy efficiency is an indication of how a UE can sustain high mobile broadband data rate while minimizing the UE modem energy consumption.

Coverage: Downlink or uplink coverage in a cellular system is measured by the maximum coupling loss between the transmit antenna of the transmitter and the receive antenna of the receiver. More specifically, the maximum coupling loss in the uplink or downlink between the user device and the base station antenna connectors is defined for a reference data rate of 160 bps, where the data rate is measured at the egress/ingress point of the MAC layers in the uplink and downlink. The target for 3GPP NR coverage is maximum coupling loss of 164 dB. Link budget and/or link-level analysis is used as the evaluation methodology to derive the parameters necessary for calculation of the coverage. In practice, the coupling loss is defined as the total long-term channel loss over the link between the UE antenna ports, and the base station antenna ports and includes antenna gains, path loss, shadowing, and body loss. The maximum coupling loss is the maximum value of the coupling loss at which a minimum SINR can be achieved and a particular service can be delivered. It is defined in the downlink/uplink direction as follows:

$$MaxCL_{UL} = \max(Tx\ \ power_{UL}) - BS\ \ Sensitivity$$
$$MaxCL_{DL} = \max(Tx\ \ power_{DL}) - UE\ \ Sensitivity$$

The parameter *MaxCL* denotes maximum coupling loss and is calculated using link budget analysis and based on the data obtained though link-level simulations. Other parameters including max (*Tx power*) and *BS*/*UE Sensitivity* depend on implementation and the deployment scenario [14].

For extreme coverage scenario and for a basic mobile broadband service characterized by a downlink data rate of 2 Mbps and an uplink data rate of 60 kbps for stationary users, the target maximum coupling loss is 140 dB. In this case and for most users, a downlink data rate of 384 kbps is acceptable. For a basic eMBB service characterized by a downlink data rate of 1 Mbps and an uplink data rate of 30 kbps for stationary users, the target maximum coupling loss is 143 dB. In this case, the downlink and uplink control channels must satisfy the performance requirements at this coupling loss.

Handover Interruption Time: This metric refers to the shortest time interval during which a user terminal cannot exchange user-plane packets with any base station when traversing boundary of the two or more cells in a cellular network. The handover interruption time includes the time required to execute any radio access network procedure, RRC signaling protocol, or other message exchanges between the mobile station and the radio access network. This benchmark only applies to eMBB and URLLC usage scenarios where effectively a seamless handover interruption time (zero second) is required.

3.2 Test Environments

In the course of development of technical standards, satisfaction of the service and system requirements by candidate technologies are often verified under certain channel propagation conditions, cell layouts, and the technical parameter set pertaining to the deployment scenario under consideration. A test environment is thereby an approximate representation of a practical deployment scenario with predefined channel propagation model, topology, and system configuration. Appropriate channel models are often developed and used in the evaluations of the candidate technologies to allow realistic modeling of the propagation conditions for the radio transmissions in different test environments. The channel models need to cover all required test environments and usage scenarios where the 5G networks are going to be deployed.

Several test environments related to the main use cases of 5G (i.e., eMBB, mMTC, URLLC) have been defined where the candidate 5G radio access technologies will be evaluated, which include the following [32]:

- *Indoor Hotspot-eMBB* test environment represents an isolated indoor environment with very high user density, which is typically found at office buildings and/or in shopping malls populated by stationary and pedestrian users. This scenario is characterized by small coverage areas per site or TRP, high user throughput, high user density, and consistent user experience in an indoor environment. The typical radial coverage in this environment is less than 10 m.
- *Dense Urban-eMBB* test environment exemplifies an urban environment with high user density and large amount of user traffic where pedestrians and users in vehicles are covered in an outdoor and outdoor-to-indoor setting. This interference-limited scenario mainly focuses on overlaying macro-TRPs with or without micro-TRPs, which is typically found in city centers and dense urban areas. The radial coverage in this environment is typically less than 100 m.
- *Rural-eMBB* test environment represents a rural environment with larger and continuous wide-area coverage, supporting pedestrian, vehicular, and high-speed vehicular users. The main feature of this scenario is unspotted wide-area coverage supporting high-speed vehicles. This scenario is noise-limited and/or interference-limited and uses a setup comprising macro-TRPs. The radial coverage in this environment is typically less than 2500 m.

- *Urban Macro-mMTC* test environment characterizes an urban macro environment with blanket coverage over a large cell serving a large number of connected (machine-type) devices. The key characteristics of this interference-limited scenario are continuous and ubiquitous coverage in urban areas comprising macro-TRPs. The typical coverage in this environment is less than 250 m (in sub-GHz frequency bands an inter-site distance of 1732 m is assumed).
- *Urban Macro-URLLC* test environment symbolizes an urban macro environment targeting URLLC services.
- *Extreme Long Distance* deployment scenario is defined to provision services for very large areas with low density of users that could be humans or machines (e.g., distribution of sensors over a large geographical area). The main features of this scenario include macro-cells with very large coverage area supporting basic data rates and voice services, with low-to-moderate user throughput and low user density.

The mapping of the IMT-2020 test environments to the usage scenarios is shown in Table 1. This mapping is based on the similarities between the key features of each test environment and those of the use case. It must be noted that a test environment represents a certain system/geometrical configuration to evaluate the operation of a candidate radio access technology in a particular usage scenario and to verify satisfaction of the requirements for that usage scenario. A test environment typically signifies a practical deployment scenario with limited parameter set that are selected to model the realistic behavior of the radio access technology as accurate as possible. Therefore, the choice of parameters and their values is very important to the accuracy of the modeling of deployment scenarios. In the past two decades, 3GPP and ITU-R technical working groups have made significant efforts to define the test environments and their associated parameter sets to facilitate characterization of the radio access technologies.

The selected parameter sets corresponding to the test environments are shown in Table 2. The complete list of parameters of each test environment is provided in reference [10]. The configurations shown in Table 2 are chosen to ensure fulfillment of the performance requirements in the associated test environment while taking into consideration the services and the practical limitations (e.g., form factor, hardware, and installation) of the gNB/TRP and/or the terminals operating in that environment. Furthermore, DL + UL bandwidth in this table refers to symmetric bandwidth allocations between DL and UL in frequency

Table 1: Mapping of test environments and usage scenarios [10].

Usage Scenarios	eMBB	mMTC	URLLC		
Test environments	Indoor Hotspot-eMBB	Dense Urban-eMBB	Rural-eMBB	Urban macro-mMTC	Urban macro-URLLC

Table 2: Selected parameters of the test environments [10,14].

Deployment Scenario	Frequency Band (GHz)	Maximum System Bandwidth (MHz)	Maximum Number of Tx/Rx Antennas (TRP)	Maximum Number of Tx/Rx Antennas (UE)	Mobility (km/h)
Indoor hotspot	4	200 (DL + UL)	256Tx × 256Rx	8Tx × 8Rx	3 (Indoor)
	30	1000 (DL + UL)	256Tx × 256Rx	32Tx × 32R	
	70		1024Tx × 1024Rx	x64Tx × 64Rx	
Dense urban	4	200 (DL + UL)	256Tx × 256Rx	8Tx × 8Rx	3 (Indoor) 30 (Outdoor)
	30	1000 (DL + UL)	256Tx × 256Rx	32Tx × 32Rx	
Rural	0.7	20 (DL + UL)	64Tx × 64Rx	4Tx × 4Rx	3 (Indoor) 120 (Outdoor)
	4	200 (DL + UL)	256Tx × 256Rx	8Tx × 8Rx	
Urban macro	2 and 4	200 (DL + UL)	256Tx × 256Rx	8Tx × 8Rx	3 (Indoor) 30 (Outdoor)
	30	1000 (DL + UL)	256Tx × 256Rx	32Tx × 32Rx	
High speed	4	200 (DL + UL)	256Tx × 256Rx	8Tx × 8Rx	500 (Outdoor)
	30, 70	1000 (DL + UL)	256Tx × 256Rx	32Tx × 32Rx	
Extreme long distance coverage	0.7	40 (DL + UL)	N/A	N/A	160 (Outdoor)
	3				
Urban macro-mMTC	0.7 (2.1)	10 (50)	64Tx × 4Rx	2Tx × 2Rx	3 (Indoor/Outdoor)

division duplex (FDD) systems and the aggregated system bandwidth used for either DL or UL via switching in time-domain in TDD systems.

3.3 High-Level Architectural Requirements

The desire to support different UE types, services, and technologies in 5G has been driving the development of the NR standards in 3GPP. The key objective of the 5G systems is to support new deployment scenarios and to address the requirements of diverse market segments. As we mentioned earlier, 5G systems have certain key characteristics such as support for myriad of radio access technologies, scalable and customizable network architecture, rigorous KPIs, flexibility and programmability, and resource efficiency (both on the user and control planes) as well as seamless mobility in densely populated heterogeneous environments and support for real-time and non-real-time multimedia services and applications with improved quality of experience (QoE).[14] Thus the architecture of the next-generation

[14] QoE is an important measure of the end-to-end performance at the service level from the user's perspective and an important metric for the design of systems. QoE is related to but differs from QoS, which embodies the notion that hardware and software characteristics can be measured, improved, and guaranteed. In contrast, QoE expresses user satisfaction both objectively and subjectively. It is often used in information technology and consumer electronics. QoE, while not always numerically quantifiable, is the most significant single factor in the real-world evaluation of the user experience. Major factors that affect QoE include cost, reliability, efficiency, privacy, security, interface user-friendliness, and user confidence.

networks is required to address some key design principles in order to meet the above-mentioned requirements. The following is a list of key architectural focus areas and their associated requirements that were identified by system architecture working group in 3GPP in the early stages of 5G standards development [13,21]:

1. Support of the 3GPP NR access, evolution of LTE, legacy 3GPP access, and non-3GPP access such as Wi-Fi networks is a key requirement for the new architecture. As part of non-3GPP access types, trusted and untrusted WLAN access support is necessary for heterogeneous networks. Satellite radio access network needs to be supported in order to enable coverage of rural and remote areas using satellite access nodes. The 5G system supports most of the existing LTE evolved packet system services in addition to the new services that includes seamless handover between the NR and the legacy 3GPP radio access networks. Interworking between the new radio and LTE is required considering the necessity for efficient inter-RAT mobility and aggregation/distribution of data flows via dual connectivity between LTE and the new radio in NSA scenarios. This requirement applies to both collocated and non-collocated site deployments.
2. Unified authentication framework for different access systems is required to simplify access to the next-generation networks via any of their constituent radio access networks. The authentication protocol is responsible for the validation of the user identity that is presented to the network when a UE requests to receive service(s) from the next-generation network.
3. Support of several simultaneous connections of a UE via multiple access technologies served by a single 5GC network is another important requirement. For UEs that can be simultaneously connected to both 3GPP access and non-3GPP access, the next-generation system should be able to exploit the availability of multiple radio links in a way that improves the user experience, optimizes the traffic distribution across various access links, and enables the provision of new high-data-rate services. The new RAN architecture is also required to support operator-controlled side-link (D2D and V2X) operation in both in-coverage and out-of-coverage scenarios. The UE can connect to the network directly (direct network connection), connect via another UE acting as a relay (indirect network connection), or connect using both types of connections. These user terminals can be anything from a wearable, monitoring human biometrics, to non-wearable devices that communicate in a personal area network such as a set of home appliances (e.g., smart thermostat and entry key) or the electronics in an office environment (e.g., smart printers). The relay UE can access the network using 3GPP or non-3GPP access (e.g., WLAN access, fixed broadband access). 3GPP and non-3GPP radio technologies and fixed broadband technologies over licensed bands or unlicensed bands can be supported as connectivity options between the remote UE and the relay UE.
4. Separation of control- and user-plane functions is a key step toward enabling SDN-controlled networks. This feature is also an enabler for the network slicing and the NFV

infrastructure. The next-generation network is built on NFV and software-defined networking to reduce the total cost of ownership and to improve operational efficiency, energy efficiency, and simplicity and flexibility of the network for offering new services.

5. Support of network sharing and network slicing is yet another important requirement. Network slicing allows the operators to provide several customized network services over the same physical network. For example, there can be different requirements on each network slice functionality (e.g., priority, charging, policy control, security, and mobility), differences in performance requirements (e.g., latency, mobility, availability, reliability, and data rates), or each network slice can serve only specific users (e.g., multimedia priority service users, public safety users, enterprise customers, and mobile virtual network operators). A network slice can provide the functionality of a complete network including radio access and core network functions. Each physical network can support one or more network slices [21]. It must be possible to verify if the UE is allowed to access a specific network slice. The management of life cycle of network slice and network function instances is also a key aspect in the overall framework for network slicing in 3GPP.
6. Energy efficiency is a critical issue in 5G systems. The potential to deploy systems in areas without a reliable energy source requires new methods of managing energy consumption not only in the UEs but also through the entire 5G network. The 5G access network is required to support an energy-saving mode which can be activated/deactivated either manually or automatically and the service can be restricted to a group of users (e.g., public safety user, emergency callers). When in energy-saving mode, the inactive UEs transmit power may be reduced or turned off or their latency and jitter constraints may be relaxed with no impact on active users or applications in the network. Small form factor UEs also typically have a small battery, and this not only imposes restrictions on power optimization but also on how the energy is consumed.
7. The unambiguous definition of the functional split between the next-generation core and the access network(s) is an important consideration for the support of heterogeneous access types where new interfaces between the next-generation core and the new 3GPP RAN need to be specified.
8. Mobility management is a key feature of 5G to support UEs with different mobility conditions which include stationary UEs during their entire usable life (e.g., sensors embedded in an infrastructure), stationary UEs during active periods, but nomadic between activations (e.g., fixed access), limited mobility within a constrained area (e.g., robots in a factory), and fully mobile UEs. Some applications require the network to guarantee seamless mobility, transparent to the application layer, in order to avoid service interruption, and to ensure service continuity. The increasing multimedia broadband data traffic necessitates offloading of IP traffic from the 5G network to traditional IP routing networks via an IP anchor node close to the network edge. As the UE

traverses the network, changing the IP anchor node may be needed in order to avoid traffic congestion in highly loaded networks, to reduce end-to-end latency, and to provide better user experience. The flexible nature of the 5G systems support different mobility management options that minimize signaling overhead and optimize network access for different categories of UEs.

3.4 System Performance Requirements

The groundbreaking improvements initiated by 5G are achievable with the introduction of new technologies, both in the access and the core networks, which include flexible and scalable allocation of network resources to various services and applications. In addition to increased flexibility and optimization, a 5G system must support stringent criteria for latency, reliability, throughput, etc. Enhancements in the air-interface contribute to meeting these KPIs as do improvements in the core network, such as network slicing, in-network caching, and hosting services closer to the network edges [21]. 5G systems further support new and emerging business models. Drivers for the 5G KPIs include services such as unmanned aerial vehicle (UAV) remote control, VR/AR,[15] and industrial automation. Increasing network flexibility would allow support of self-contained enterprise networks, which are installed and maintained by network operators while being managed by the enterprise. Enhanced connection modes and improved security enable support of massive IoT use cases, where those are expected to include numerous UEs of different categories sending and receiving data over the 5G network. The essential capabilities for providing this level of flexibility include network slicing, network capability exposure, scalability, and diverse mobility. The requirements for other network operations are meant to address the necessary control and user-plane resource allocation and utilization efficiencies, as well as network configurations that optimize service delivery by minimizing routing between end users and application servers.

Unlike previous generations of wireless systems, the main objective of 5G has been to develop a unified system that can be exclusively configured and optimized for different use cases. Increasing user expectations and diversified and challenging service requirements necessitate a coherent approach to technology development and deployment. In the previous generations of wireless technology, functionality was primarily provided by network equipment, and devices were primarily used to access network services. However, in the case of 5G, it is recognized that the boundary between functionality provided by the network and

[15] VR is the term used to describe a three-dimensional, computer-generated environment which can be explored and interacted with by a person. That person becomes part of this virtual world or is immersed within this environment and whilst there, is able to manipulate objects or perform a series of actions. AR is the integration of digital information with the user's environment in real time. Unlike virtual reality, which creates a totally artificial environment, augmented reality uses the existing environment and overlays new information on top of it.

functionality provided by the device is subtle. Virtualization of network functions, the role of applications, and the ubiquitous availability of increasingly sophisticated smart devices will only accelerate this trend. Nonetheless, there are some cases where the desired functionality is expected to be exclusively provided in the network [45].

The eMBB use cases require a number of new KPIs related to higher data rates, user density, user mobility, and coverage. Higher data rates are driven by the increasing use of data for services such as streaming (e.g., video, music, and user generated content), interactive services, and some IoT applications. These services have stringent requirements for user experienced data rates and end-to-end latency in order to guarantee good user experience. In addition, increased coverage in densely populated areas such as sports arenas, urban areas, and transportation hubs is essential for nomadic and mobile users. The new KPIs on traffic and connection density would enable transport of large amount of user traffic per area (traffic density) and transport of data for a large number of connections (connection density). A large group of UEs are expected to support variety of services which exchange either a very large (e.g., streaming video) or very small (e.g., data burst) amount of data. The 5G systems are designed to manage this variability in an extremely resource and energy efficient manner. All of these scenarios introduce new requirements for indoor and outdoor deployments, local area connectivity, high user density, wide-area connectivity, and UEs travelling at high speeds. Another aspect of 5G KPIs includes requirements for various combinations of latency and reliability, as well as higher position determination accuracy. These KPIs are further driven by support for commercial and public safety services. Support for mMTC presents several new requirements in addition to those for the eMBB use cases. The proliferation of the connected things introduces a need for significant improvements in resource efficiency in all system components (e.g., UEs, IoT devices, radio access network, core network). The 5G systems are intended to extend their capabilities to meet KPIs required by emerging mission-critical applications. For these advanced applications, the requirements, such as data rate, reliability, latency, communication range, and speed, are made more stringent. Fig. 13 illustrates the latency and throughput requirements of the prominent 5G applications (see also Table 3).

3.5 Service Requirements

The 5G systems are required to support the existing LTE network services in addition to the newly introduced services, which means that the existing EPS services can be accessed via the new 5G access technologies. The following is a list of important service requirements for 5G networks:

- *Multimedia Broadcast/Multicast Service:* The proliferation of video streaming services, software delivery over wireless network, group communications, and multicast/broadcast IoT applications have created a demand for flexible and dynamic allocation of

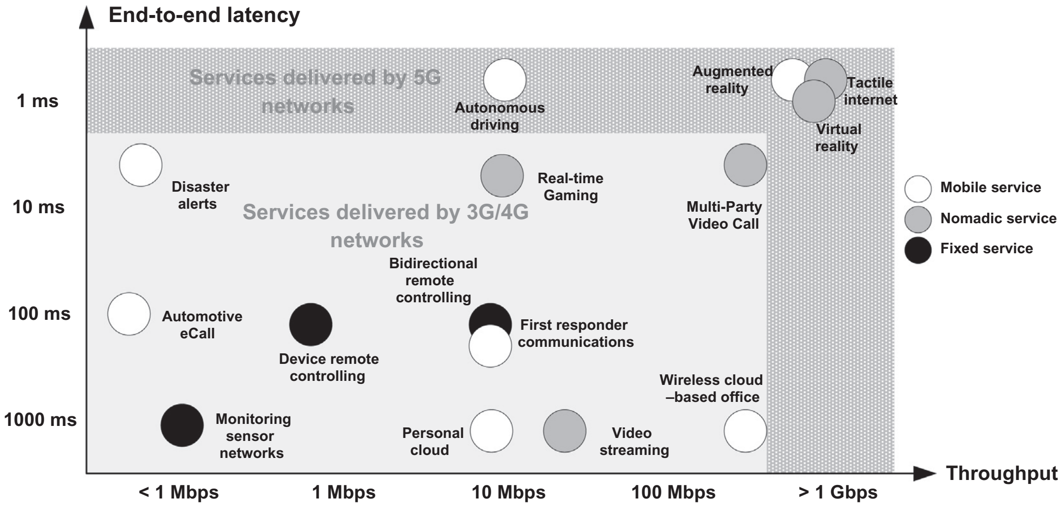

Figure 13
Performance requirements of some 5G applications [44].

radio resources between unicast and multicast services within the network as well as support for a standalone deployment of multicast/broadcast network. Moreover, enabling such a service over a network for a wide range of intersite distances between the radio base stations will enable a more efficient and effective delivery system for real-time and streaming multicast/broadcast content over wide geographical areas as well as in specific geographical areas spanning a limited number of base stations. A flexible multicast/broadcast service will allow the 5G system to efficiently deliver such services.

- *Location/Positioning Service:* The NR systems enable and improve, if necessary, advanced positioning techniques such as RAN-assisted (Cell-ID, E-Cell ID, OTDOA, and UTDOA)[16] and non-RAN-assisted (GNSS, Bluetooth, WLAN, terrestrial beacon systems (TBS),[17] and sensor-based location determination) that have been specified and used in LTE systems. In addition, support of D2D-based positioning techniques has also been considered in NR. The NR positioning schemes exploit high bandwidth, massive antenna systems, new network architecture, and functionalities as well as deployment of massive number of devices to support indoor and outdoor location-based services. Higher precision positioning is characterized by ambitious system requirements for

[16] Cell-ID, E-Cell ID, OTDOA, UTDOA, and GNSS denote cell identifier, enhanced cell identifier, observed time difference of arrival, uplink time difference of arrival, and global navigation satellite system and are commonly used methods for positioning in 3GPP systems.

[17] TBS consist of a network of terrestrial beacons broadcasting signals for positioning purposes. They may use dedicated, unshared spectrum or spectrum shared with other users, including FDD and TDD licensed spectrum.

Table 3: 3GPP and ITU-R service and performance requirements [11,14].

Service/Performance Metric	Category Usage Scenario	Test Environment	Link Direction	ITU-R Requirements	3GPP Requirements
Peak data rate (Gbps)	eMBB	N/A	Downlink	20	20
			Uplink	10	10
Peak spectral efficiency (bits/s/Hz)	eMBB	N/A	Downlink	30	30
			Uplink	15	15
User experienced data rate (Mbps)	eMBB	Dense Urban-eMBB	Downlink	100	3 × IMT-Advanced cell-edge spectral efficiency at 1 GHz Bandwidth
			Uplink	50	3 × IMT-Advanced cell-edge spectral efficiency at 1 GHz bandwidth
5th Percentile user spectral efficiency (bits/s/Hz)	eMBB	Indoor Hotspot-eMBB	Downlink	0.3	3 × IMT-Advanced cell-edge spectral efficiency in InH[a]
			Uplink	0.21	3 × IMT-Advanced Cell-edge Spectral Efficiency in InH
	eMBB	Urban macro	Downlink	N/A	3 × IMT-Advanced cell-edge spectral efficiency in UMa
			Uplink	N/A	3 × IMT-Advanced cell-edge spectral efficiency in UMa
	eMBB	Dense Urban-eMBB	Downlink	0.225	3 × IMT-Advanced cell-edge spectral efficiency in UMi
			Uplink	0.15	3 × IMT-Advanced cell-edge spectral efficiency in UMi
	eMBB	Rural-eMBB	Downlink	0.12	3 × IMT-Advanced cell-edge spectral efficiency in RMa
			Uplink	0.045	3 × IMT-Advanced cell-edge spectral efficiency in RMa
Average spectral efficiency (bits/s/Hz/TRP)	eMBB	Indoor Hotspot-eMBB	Downlink	9	3 × IMT-Advanced cell spectral efficiency in InH
			Uplink	6.75	3 × IMT-Advanced cell spectral efficiency in InH
	eMBB	Dense Urban-eMBB	Downlink	7.8	3 × IMT-Advanced cell spectral efficiency in UMi
			Uplink	5.4	3 × IMT-Advanced cell spectral efficiency in UMi
	eMBB	Rural-eMBB	Downlink	3.3	3 × IMT-Advanced cell spectral efficiency in RMa
			Uplink	1.6	3 × IMT-Advanced cell spectral efficiency in RMa

(*Continued*)

Table 3: (Continued)

Service/Performance Metric	Category Usage Scenario	Test Environment	Link Direction	ITU-R Requirements	3GPP Requirements
Area traffic capacity (Mbit/s/m^2)	eMBB	Indoor Hotspot-eMBB	Downlink	10	No explicit requirement
User-plane latency (ms)	eMBB	N/A	Uplink and downlink	4	4 *The system user plane should support RTT of 600, 180, and 50 ms in the case of GEO, MEO, LEO satellite systems, respectively*
	URLLC	N/A	Uplink and downlink	1	0.5
Control-plane latency (ms)	eMBB	N/A	Uplink and downlink	20	10 *The system control plane should support RTT of 600, 180, and 50 ms in the case of GEO, MEO, and LEO satellite systems, respectively.*
	URLLC	N/A	Uplink and downlink	20	10
Connection density (devices/km^2)	mMTC	Urban Macro-mMTC	Uplink	1000,000	1000,000
Energy efficiency	eMBB	N/A	N/A	Capability to support a high sleep ratio and long sleep duration	No explicit requirement
Reliability	URLLC	Urban Macro-URLLC	Uplink or downlink	$1-10^{-5}$ success probability of transmitting a layer 2 PDU of size 32 B within 1 ms in channel quality of coverage edge	$1-10^{-5}$ for a packet of 32 bytes with a user-plane latency of 1 ms *For eV2X cases (direct communication via side-link or when the packet is relayed via gNB), the requirement is $1-10^{-5}$ for packet size 300 bytes and user-plane latency 3-10 ms.*
Mobility classes	eMBB	Indoor Hotspot-eMBB	Uplink	Stationary, pedestrian	No explicit requirement
	eMBB	Dense Urban-eMBB	Uplink	Stationary, pedestrian, vehicular (up to 30 km/h)	No explicit requirement
	eMBB	Rural-eMBB	Uplink	Pedestrian, vehicular, high-speed vehicular	No explicit requirement
Mobility traffic channel link	eMBB	Indoor Hotspot-eMBB	Uplink	1.5 (10 km/h)	No explicit requirement
Data rates (bits/s/Hz)	eMBB	Dense Urban-eMBB	Uplink	1.12 (30 km/h)	No explicit requirement
	eMBB	Rural-eMBB	Uplink	0.8 (120 km/h) 0.45 (500 km/h)	No explicit requirement No explicit requirement
Handover interruption time (ms)	eMBB and URLLC	N/A	N/A	0	0

(*Continued*)

Table 3: (Continued)

Service/Performance Metric	Category Usage Scenario	Test Environment	Link Direction	ITU-R Requirements	3GPP Requirements
Bandwidth (MHz/GHz) and scalability	N/A	N/A	N/A	At least 100 MHz Up to 1 GHz Support of multiple different bandwidth values	No explicit requirement No explicit requirement No explicit requirement

[a]The IMT-Advanced minimum requirements for 5th percentile user spectral efficiencies in InH, UMa, UMi, and RMa test environments can be found in reference [8].

positioning accuracy. One use case where higher precision positioning capability is critically needed is collision avoidance in a busy street where each vehicle must be aware of its own position, the positions of neighboring vehicles, and their expected driving paths and movements, in order to reduce the risk of collisions. In another use case on the factory floor, it is important to locate moving objects such as machinery or parts to be assembled with sufficiently high precision. Depending on the use case, the 5G system is required to support the use of 3GPP and non-3GPP technologies to achieve higher accuracy indoor/outdoor positioning. The corresponding positioning information must be acquired with proper timing, to be reliable, and to be available (e.g., it is possible to unambiguously determine the position). The 5G UEs must be able to share positioning information between each other or with a cloud-based controller if the location information cannot be processed or used locally. Table 4 summarizes the positioning requirements for the latter use case. Another aspect of 5G KPIs includes requirements for various combinations of latency and reliability, as well as higher accuracy for positioning in order to support mission-critical services such as public safety, emergency communications, and public warning/emergency alert systems. These KPIs are driven by support for both commercial and non-commercial public safety services. On the commercial side, industrial control, industrial automation, UAV control, and AR exemplify those services. Services such as UAV control will require more precise positioning information that includes altitude, speed, and direction, in addition to planar coordinates.

- *Service Continuity and Reliability:* Communication service availability is defined as percentage of time where the end-to-end communication service is delivered according to an agreed QoS, divided by the amount of time the system is expected to deliver the an end-to-end service according to the specification in a specific area. Note that the end point in an end-to-end service is assumed to be the communication service interface. The communication service is considered unavailable, if it does not meet the pertinent QoS requirements. The system is considered unavailable in case an expected message is not received within a specified time, which, at minimum, is the sum of end-to-end

Table 4: Performance requirements for higher precision positioning service [21].

Service	Position Acquisition Time (ms)	Survival Time (s)	Availability (%)	Dimension of Service Area (m)	Position Accuracy (m)
Mobile objects on factory floor	500	1	99.99	500 × 500 × 30	0.5

latency, jitter, and the survival time.[18] Reliability expressed in percent is the amount of transmitted network layer packets successfully delivered to a given node within the time constraint required by the targeted service, divided by the total number of sent network layer packets. Communication service availability and reliability are well-defined terms that have used not only within 3GPP but also in vertical industries. Communication service availability addresses the availability of a communication service, in vertical applications in accordance to IEC 61907,[19] whereas reliability is related to the availability of the communication network (see Fig. 14).

As depicted in Fig. 14, reliability covers the communication-related aspects between two nodes which are often the end nodes, while communication service availability addresses the communication-related aspects between two communication service interfaces. In other words, the gap between the two concepts is the communication interface. This might seem to be a small difference, but this difference can lead to conditions where reliability and communication service availability have different values.

- *Priority, QoS, and Policy Control:* The 5G networks are expected to support a wide range of commercial services and regional/national regulatory services with appropriate access prioritization. Some of these services share common QoS characteristics such as latency and packet loss rate but may have different priority requirements. For example, UAV control and air traffic control may have stringent latency and reliability requirements but not necessarily the same priority requirements. In another example voice-based services for multimedia priority service (MPS)[20] and emergency services share

[18] Survival time is the time that an application consuming a communication service may continue without an anticipated message.

[19] IEC 61907 2009 provides guidance on dependability engineering of communication networks. It establishes a generic framework for network dependability performance, provides a process for network dependability implementation, and presents criteria and methodology for network technology designs, performance evaluation, security consideration, and quality of service measurement to achieve network dependability performance objectives (see https://webstore.iec.ch/publication/).

[20] MPS, supported by 3GPP, is a set of services and features which create the ability to deliver calls or complete sessions of a high-priority nature from mobile-to-mobile networks, mobile-to-fixed networks, to fixed-to-mobile networks. MPS provides broadband IP-based multimedia services (IMS-based and non-IMS-based) over wireless networks in support of voice, video, and data services. Network support for MPS will require end-to-end priority treatment in call/session origination/termination including the non-access stratum (NAS) and access stratum (AS) signaling establishment procedures at originating/terminating network side as well as resource allocation in the core and radio networks for bearers. The MPS will also require end-to-end priority treatment in case of roaming if supported by the visiting network and if the roaming user is authorized to receive priority service.

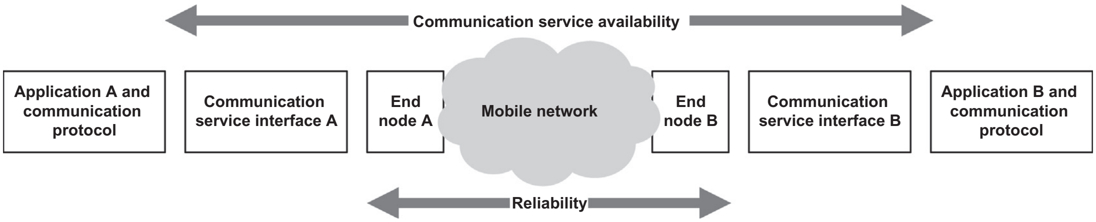

Figure 14
Illustration of the relationship between communication service availability and reliability [21].

common QoS characteristics as applicable for normal public voice communications, yet they may have different priority requirements. The 5G network supports mechanisms that allow decoupling of the priority of a particular communication from the associated QoS characteristics such as latency and reliability to allow flexibility and configurability to support different priority services in operators' networks while adhering to their network policies and the corresponding national/regional regulations. The network needs to support flexible means to make priority decisions based on the state of the network (e.g., during disaster events and network congestion) recognizing that the service priorities may change during a crisis. The priority of any service may be different for each user depending on the operational requirements and regional/national regulations. The 5G systems provide flexible means to prioritize and enforce prioritization among the services (e.g., MPS, emergency, medical, public safety) and among the users of those services. The traffic prioritization may be enforced by adjusting resource utilization or preempting lower priority traffic. The networks are typically capable of providing the required QoS (e.g., reliability, latency, and bandwidth) for a service and are able to prioritize resources when necessary in order to meet the service requirements. The existing QoS and policy frameworks manage latency and improve reliability by traffic engineering. In order to support 5G service requirements, it is necessary for the 5G network to offer QoS and policy control for reliable communication with latency required for a service and enable the resource adaptations as necessary. Also, as 5G network is expected to operate in a heterogeneous environment with multiple access technologies, multiple types of UE, etc., it should support a harmonized QoS and policy framework that can be applied to multiple access networks. Unlike the QoS control in EPS which only covers RAN and core network, the 5G network end-to-end QoS (e.g., RAN, backhaul, core network, network-to-network inter-connect) is required to achieve the 5G user experience (e.g., ultra-low latency, ultra-high bandwidth).

- *V2X Services:* Platooning enables the vehicles to dynamically form a group travelling together. All the vehicles in the platoon receive periodic data from the leading vehicle, in order to carry on platoon operations. This information allows the distance between vehicles to become extremely small, that is, the inter-vehicle distance when translated to time can be very small and in the order of fraction of a second. Platooning applications may allow the participating vehicles to be autonomously driven. Advanced driving enables semi-automated or fully automated driving where longer inter-vehicle distance is assumed. Each vehicle and/or road side unit (RSU)[21] shares data obtained from its local sensors with vehicles in proximity, thus allowing vehicles to coordinate their courses and/or maneuvers. In addition, each vehicle shares its driving plan with vehicles in proximity. The advantage of this feature is safer driving, collision avoidance, and improved traffic efficiency. Extended sensors enable the exchange of raw or processed data gathered through local sensors or live video data among vehicles, road site units, UEs of pedestrians, and V2X application servers. The vehicles can enhance the perception of their environment beyond what their own sensors can detect and have a more holistic view of the local conditions. Remote driving enables a remote driver or a V2X application to operate a remote vehicle for those passengers who cannot drive themselves or a remote vehicle located in dangerous environments. For a case where variation is limited and routes are predictable, such as public transportation, driving based on cloud computing can be used.

Intelligent transport systems (ITS) embrace a wide range of applications that are intended to increase travel safety, minimize environmental impact, improve traffic management, and maximize the benefits of transportation to both commercial users and the general public. Over the years, the emphasis on intelligent vehicle research has turned into cooperative systems in which the traffic participants (vehicles, bicycles, pedestrians, etc.) communicate with each other and/or with the infrastructure. In this context, a cooperative ITS scheme is a technology that allows vehicles to be wirelessly connected to each other, to the infrastructure and other parts of the transport network. In addition to what drivers can immediately see around them, and what vehicle sensors can detect, all parts of the transport system will increasingly be able to share information to improve decision-making in the system. Thus this technology can improve road safety through avoiding collisions and further can assist the transportation network by reducing congestion and improving traffic flows. Once the basic technology is in place as a platform, subsequent applications can be developed. Cooperative ITS can greatly increase the quality and reliability of information available about vehicles, their location and the driving environment. In the future, cars will know the location of road works and the switching phases of traffic lights ahead, and they will be able to react accordingly. This will make travel safer and more convenient. On-board driver assistance, coupled with two-way communication between vehicles and between cars and road infrastructure, can help drivers to better control their vehicle and hence have positive effects in terms of safety and traffic efficiency. RSUs play an important role in this

[21] RSU is a communication/computing device located on the roadside that provides connectivity support to passing vehicles.

technology. Vehicles can also function as sensors reporting weather and road conditions including incidents. In this case, cars can be used as information sources for high-quality information services. RSUs are connected to the traffic control center (TCC) for management and control purposes. The traffic light information and the traffic information obtained from the TCC can be broadcast via the RSUs to the vehicles and the RSUs can further collect vehicle probe data for the TCC. For reliable distribution of data, low-latency and high-capacity connections between RSUs and the TCC are required. This type of application is made possible by setting stringent end-to-end latency requirements for the communication service between RSU and TCC since relayed data needs to be processed in the TCC and the results need to be forwarded to the neighboring RSUs. It must be noted that the availability of the communication service has to be very high in order to compete with existing wired technology and in order to justify the costly deployment and maintenance of RSUs.

- *Security:* IoT introduces new UEs with different characteristics, including IoT devices with no user interface (e.g., embedded sensors), long life spans during which an IoT device may change ownership several times (e.g., consumer goods), and non-provisioned IoT devices (e.g., consumer goods). These applications necessitate secure mechanisms to dynamically establish or refresh credentials and subscriptions. New access technologies, including licensed and unlicensed, 3GPP and non-3GPP, are driving new efforts for creating access independent security mechanisms that are seamlessly available while the IoT device is active. A high level of 5G security is essential for mission-critical communications, for example, in industrial automation, V2X services, industrial IoT, and the smart grid. Expansion into enterprise, vehicular, and public safety markets will drive the efforts toward increased end-user privacy protection. 5G security is attempting to address all of these new requirements while continuing to provide connection/session security consistent with 3GPP legacy systems.

4 ITU-R IMT-2020 Standardization Activities

ITU has a long and reputable history in developing radio interface standards for mobile communications in the form of recommendations and reports. The development of standards for IMT systems, which started in late 1990's, encompassing IMT-2000 and IMT-Advanced [3], helped globalized the 3G and 4G networks and this trend will continue to evolve in to 5G with publication of IMT-2020 recommendations (see Fig. 15). In early 2012, ITU-R[22] initiated a path finding program to develop vision for IMT systems in 2020 and beyond, setting the stage for widespread 5G research activities that subsequently ensued around the world.

[22] ITU toward IMT for 2020 and beyond at http://www.itu.int/.

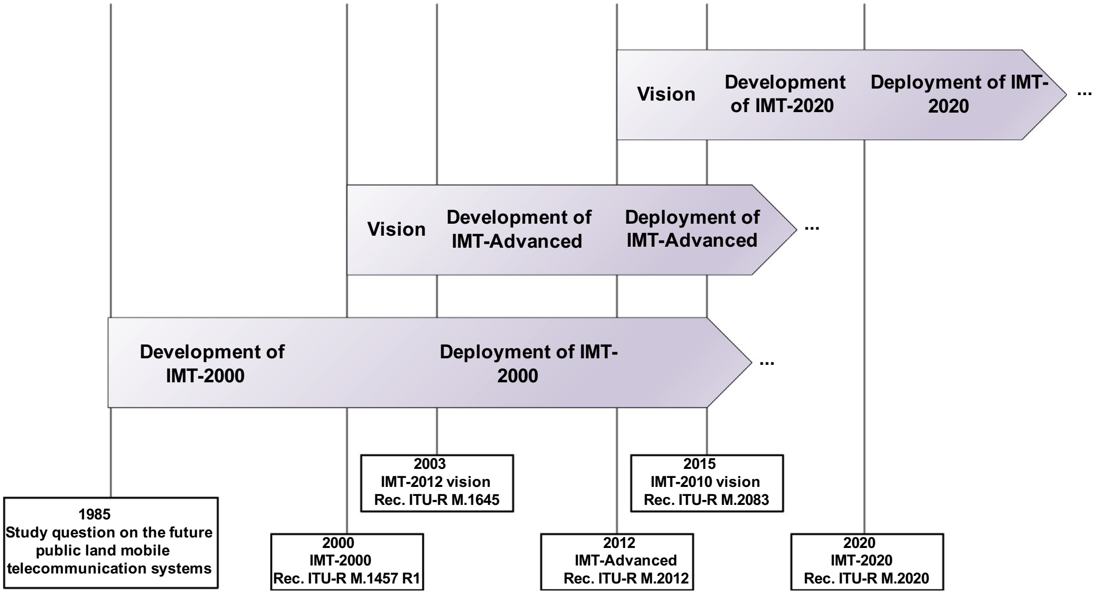

Figure 15
Overview of timelines for IMT standards development and systems deployment [1].

The role of the ITU-R[23] is to ensure efficient and economical use of the worldwide spectrum by all wireless services, including maritime and satellite communications. ITU-R further allocates and regulates the radio spectrum for global deployments of wireless and radio telecommunication technologies. The vision documents published by ITU-R (e.g., recommendations ITU-R M.1645 [2] and ITU-R M.2083 [1]) have been used by standards development organizations as guideline to develop technologies that materialize those visions.

Within ITU-R, there are several study groups that develop the technical bases for decisions taken at world radio-communication conferences and develop global standards in the form of recommendations, reports, and handbooks concerning radio telecommunication topics. The ITU-R members are typically from administrations, the telecommunications industry as a whole, and academic/research organizations throughout the world, who participate in the work of the study groups on topics such as efficient management and use of the spectrum/orbit resources, radio systems characteristics and performance, spectrum monitoring, and emergency radio communications for public safety. Working Party 5D (WP 5D)[24] is part of ITU-R Study Group 5 which is responsible for the overall radio system aspects of IMT systems. WP 5D has already finalized its views on the IMT-2020 timeline and submission

23 ITU-R, https://www.itu.int/en/ITU-R.
24 WP 5D—IMT Systems, https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d.

process as well as the minimum requirements and evaluation methodology for the IMT-2020 systems.

The WRC charter is to constantly review and revise the worldwide spectrum regulations. The WRC decisions can impact 5G in two important aspects: (1) it can designate certain bands for mobile services, and (2) it can allocate specific spectrum for IMT systems deployments. The IMT-2020 is the ITU initiative to standardize and specify 5G technologies that are included in the IMT related recommendations. WRC-15 took a significant step toward 5G spectrum at low, medium, and high frequencies by identifying bands above 6 GHz for sharing studies prior to WRC-19. Some of the frequency bands identified in 2015 are already being utilized for the deployment of 4G networks in many regions of the world. As 5G standards and technologies continue to mature, bands already in use for 4G will be reallocated to 5G systems as network operators migrate their networks from 4G to 5G. Availability of spectrum is crucial for testing and early deployments of 5G systems before 2020; therefore both higher and lower frequencies are presently needed to meet the requirements of field trials of relevant 5G use cases that are being conducted by major network operators [29,30].

According to ITU-R vision, IMT-2020 systems are mobile systems which include the new capabilities of IMT systems that go beyond those of IMT-Advanced (see Fig. 16). The IMT-2020 systems support low to high-mobility applications and a wide range of data rates in accordance with user and service requirements in various deployment scenarios. IMT-2020 systems also have capabilities for high-quality multimedia applications within a wide range of services and platforms, providing a significant improvement in performance and quality of service. A broad range of capabilities, tightly coupled with intended use cases and applications for IMT-2020 have been laid out in ITU-R vision in order to support emerging new usage scenarios and applications for 2020+ [1].

While the minimum technical requirements and corresponding evaluation criteria defined by ITU-R based on the envisioned capabilities for IMT-2020 systems could have been met by certain enhancements to the existing IMT systems, true fulfillment of the IMT-2020 system and service requirements necessitated development of groundbreaking technology components and functionalities, and development of a set of new radio access technologies. Nevertheless, IMT-2020 systems continue to interwork with and complement the existing IMT systems and their enhancements [1–12].

IMT-2020 can be considered from multiple aspects including the users, manufacturers, application developers, network operators, and service and content providers as the usage scenarios of IMT-2020 will continue to expand. Therefore, it is expected that the technologies for IMT-2020 can be applied in a broader range of usage scenarios and can support a wider range of environments, diverse service capabilities, and technology options compared to IMT-Advanced and IMT-2000 systems. The task of ITU-R WP 5D is to determine a set

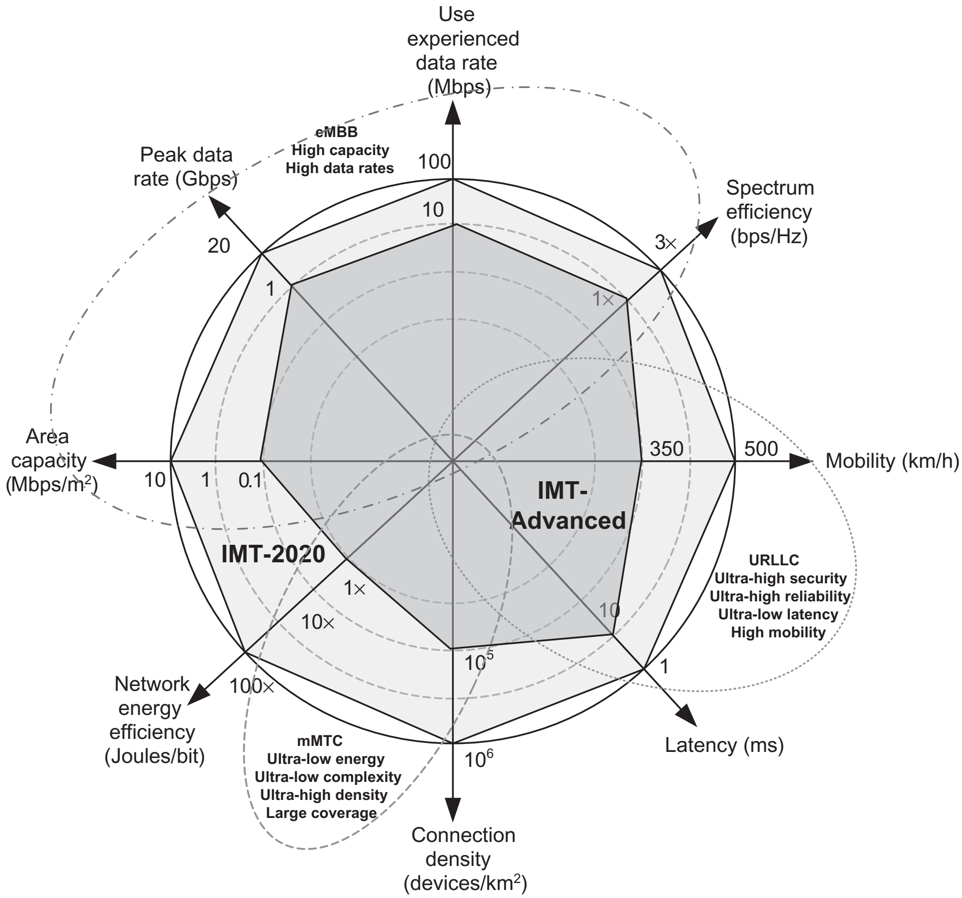

Figure 16
Comparison of IMT-advanced and IMT-2020 features and requirements [1].

of representations of IMT-2020 systems that can satisfy the requirements related to technical performance for IMT-2020 radio interfaces in the aforementioned test environments. Similar to IMT-Advanced process, the proponents could propose a single radio interface technology or a set of radio interface technologies to fulfill the requirements.

The technical characteristics of IMT-2020 systems chosen for evaluation have been explained in detail in Report ITU-R M.[IMT-2020.SUBMISSION], including service and spectrum aspects as well as the requirements related to technical performance, which are based on Report ITU-R M.[IMT-2020.TECH PERF REQ]. These requirements are summarized in Table 3, together with the high-level assessment methods which include system-level and link-level simulations, analytical, and inspection by reviewing the functionalities and supported configurations of the candidates. The process defined by ITU-R for standardization of IMT-2020 systems is illustrated in Fig. 17. According to the IMT-2020 process,

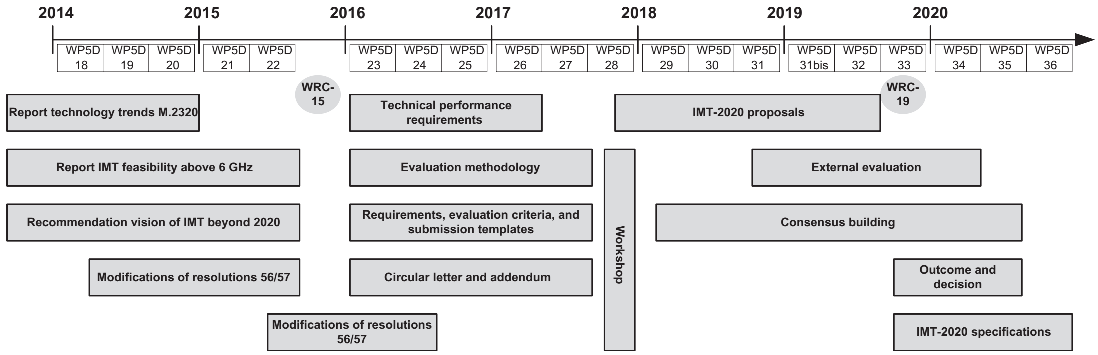

Figure 17
Detailed timeline and process for IMT-2020 in ITU-R [1].

the deadline for submission of candidates is 3Q 2019, followed by independent evaluations by external groups, which ultimately results in IMT-2020 specifications in 3Q 2020 [1].

5 3GPP 5G Standardization Activities

3GPP[25] initiated the study of 5G NR in September 2015 by inviting all 3GPP membership to share their 5G visions during a 2-day workshop. A preliminary study on 5G candidate technologies was conducted as part of 3GPP Rel-14 that was concluded in March 2017 [15]. The 3GPP services and requirements working group has contributed to the study phase for 5G service requirements, known as SMARTER [21]. In late 2015, the 3GPP system architecture working group approved a technical report on the study of architecture for next-generation systems [13] which laid out 3GPP vision for the design of 5G architecture, support of new RAT(s), the evolved LTE, as well as interworking with non-3GPP access networks. The latter study investigated high-level architectural requirements and defined a common terminology for the next-generation networks. At the same time, based on the outcome of the RAN feasibility studies, 3GPP began to develop 5G specifications in two phases. 3GPP Rel-15 is associated with the first phase of 5G specifications that define basic features of 5G systems. 3GPP Rel-16 will specify the second phase of 5G specifications defining additional features of 5G that combined with those of the first phase will satisfy the IMT-2020 requirements. The two phases of 5G specifications development in 3GPP are scheduled to end in September 2018 (already completed in March 2019) and March 2020, respectively [34].

While the prospect of the next-generation networks has inspired extensive and non-backward compatible changes in the air interface design (e.g., channel coding schemes, use of

[25] 3GPP, http://www.3gpp.org/.

non-orthogonal multiple access) and the transport layer [e.g., information centric networking (ICN)[26]], it has further stimulated the development of numerous incremental improvements in the current networks. Many of these technologies are already in the pipeline for development, standardization, and deployment. In fact, what 3GPP calls LTE-Advanced Pro, that is, revisions of LTE technology that build upon 3GPP Rel-8 through 14, along with the first drop of the NR air interface features with Rel-15 in late 2018, will form the basis of 5G in its initial deployments around 2020. Full satisfaction of IMT-2020 requirements is not anticipated until the completion of 3GPP Rel-16 in late 2019. The momentum is clearly and strongly behind what is being provided with many enhancements to LTE in a sequence of new releases. LTE has already been significantly improved since its inception with 3GPP Rel-8 and with further improvements in LTE-Advanced, as initially standardized in 3GPP Rel-10. In the LTE track of 5G in 3GPP, enhancements will continue to enable the standard to support as many 5G requirements and use cases as possible.

In order to increase the performance of the existing systems by an order of magnitude, it would require a combination of new spectrum, improved spectrum efficiency and increased network density. In particular, carrier aggregation which was introduced in 3GPP Rel-10 has been continuously enhanced in conjunction with availability of new spectrum. Among various additional capabilities in 3GPP Rel-11 and Rel-12, the performance of mobile broadband data has been increased to peak data rates of 4 Gbps in the downlink and 1.5 Gbps in the uplink through aggregation of multiple bands. Enhanced MIMO schemes and higher order modulations (i.e., 256QAM and 1024QAM) have further increased the cell spectral efficiency in the latest releases of LTE. For the time being, the evolution of mobile broadband will remain the most viable path because the demand for such services is proven to be strong and profitable for the operators and device/equipment vendors. Commercial implementation of other capabilities introduced by 3GPP in the past releases of LTE has not been as strong as anticipated. Deployment of HetNets with various features coordinating radio resources among large and small cells has been slow. Challenges include the significant difficulties and cost of upgrades and deployment of large number of small cells in urban areas and providing backhaul connection via optical fiber for user data and signaling.

The process of making LTE as of part of 5G standards package requires various enhancements and new features in LTE Rel-14 and Rel-15. The most significant ones are

[26] ICN is an approach to evolve the Internet infrastructure away from a host-centric paradigm based on perpetual connectivity and the end-to-end principle, to a network architecture in which the focus is on content or data [38]. In other words, ICN is an approach to evolve the Internet infrastructure to directly support information distribution by introducing uniquely named data as a core Internet principle. Data becomes independent from location, application, storage, and means of transportation, enabling or enhancing a number of desirable features, such as security, user mobility, multicast, and in- network caching. Mechanisms for realizing these benefits are the subject of ongoing research in Internet Engineering Task Force (IETF) and elsewhere. Current research challenges in ICN include naming, security, routing, system scalability, mobility management, wireless networking, transport services, in-network caching, and network management.

enhancements to user data rates and system capacity with full-dimension MIMO (FD-MIMO), improved support for unlicensed operations [47], and latency reduction in both control and user planes. The enhancements in LTE Rel-14 and Rel-15 are also intended to provide better support for use cases such as massive MTC, mission-critical communications, and intelligent transportation systems. The MIMO enhancement in 3GPP makes it possible to dynamically adapt the transmission both vertically and horizontally by steering a two-dimensional antenna array. The concept of FD-MIMO in the current LTE releases builds on the channel state information feedback mechanisms introduced in LTE Rel-13, in which precoding matrix codebooks support two-dimensional port layouts with up to 16 antenna ports. To enhance both non-precoded and beamformed CSI-RS operation, LTE Rel-14 introduced several new features, including hybrid non-precoded/beamformed CSI mode with optimized feedback; aperiodic triggering of CSI-RS measurements and support for up to 32 antenna ports. Other 3GPP work items include uplink enhancements such as higher order modulation and coding (e.g., 256QAM in the uplink) in baseline LTE systems. While it is theoretically possible to increase uplink peak data rates with improvements such as this in small-cell deployments, commercially available mobile devices are almost invariably use modulation orders limited to 16QAM in the uplink due to practical constraints. The ability to combine unlicensed spectrum with licensed spectrum is a highly attractive opportunity for network operators. This can be achieved by LTE−WLAN link aggregation and/or by carrier aggregation of LTE in both licensed and unlicensed bands through licensed-assisted access. These features are included in 3GPP Rel-13 onward.

There has been significant interest expressed by some mobile operators and broadcasters for enhanced multimedia broadcast multicast service (eMBMS) in the past few years; nevertheless, the use of eMBMS in general has been only nascent and the standardization work is ongoing with further improvements in 3GPP Rel-14.

Reduction in end-to-end latency is required for various applications, including those envisaged for 5G. However, somewhat lower latency than current levels is desirable for more conventional mobile broadband use cases. LTE control-plane latency figures vary quite widely among different networks with figures up to 75 ms quite common. This relatively high latency can result from the routing of data packets through many network elements before they reach a network gateway to the Internet. From that point onward, there will be additional packet-routing delays which are beyond operator's control. 3GPP has been studying the possibility of reducing the LTE TTI of 1 ms and to further identify additional latency reductions at layers 2 and 3 of the LTE protocol stack.

While LTE evolution is mainly focused on eMBB aspects, 3GPP is also trying to fulfill some of the 5G use cases with the specific performance requirements in the LTE path. For example, a downscaled UE Category 0 was introduced in LTE Rel-12 for IoT applications. This and the subsequent LTE-based NB-IoT standardization could satisfy some of the 5G requirements for mMTC use cases.

The scope of NR phase 1 standardized in 3GPP Rel-15 included both NSA and SA operations. In NSA mode, 5G NR uses LTE as an anchor in the control plane. The SA mode allows 5G NR to work independently with full control plane capability. While focusing on eMBB use case, it also provides support for some URLLC services. The LTE and NR share much more than the common 3GPP release schedule shown in Fig. 18. It is inevitable that much of what is being developed for LTE will also significantly contribute to 5G, and there will be significant commonalities and interdependencies, if not compatibilities between the two. For example, 5G might be dependent on low-band LTE for coverage and control channel signaling where such spectrum is not directly available for 5G. However, the transition to 5G also has opened up the possibility of introducing an entirely new air interface and creating network architecture and designs that are not necessarily backward compatible with LTE. There are significant tradeoffs in NR implementation and deployments. For example, dual connectivity of the UE can improve performance, but this requires compromise on

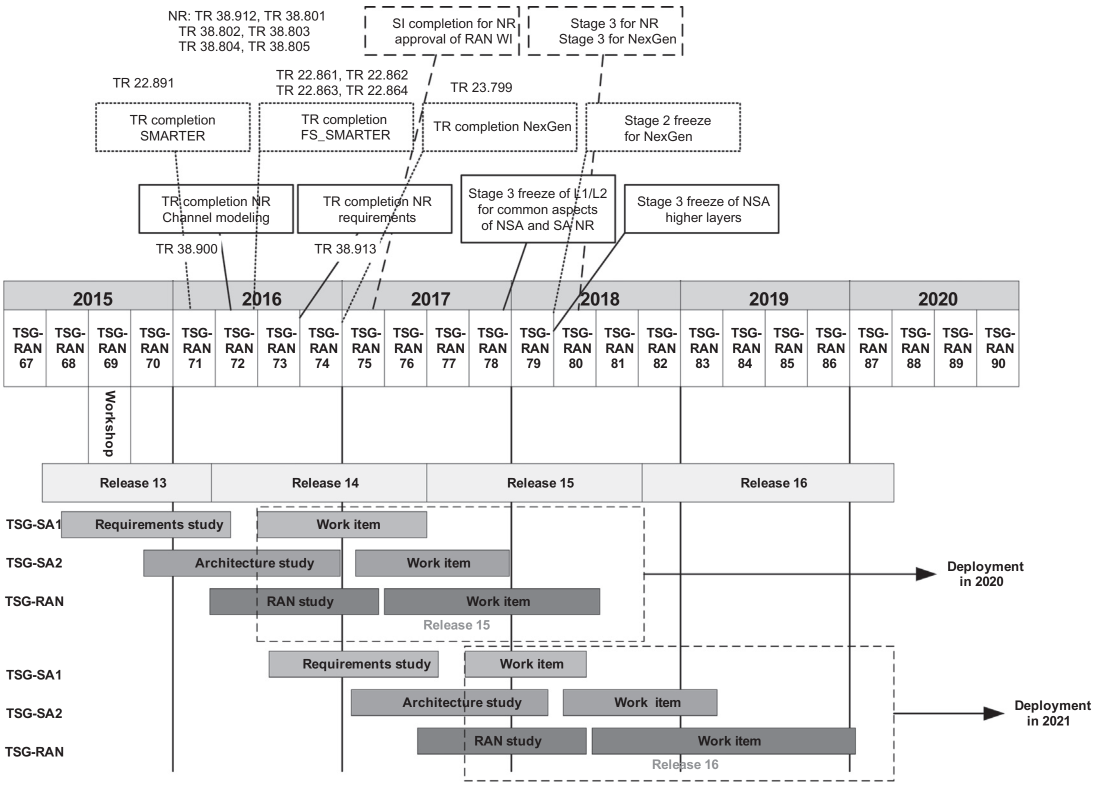

Figure 18

Detailed timeline and process of 5G development in 3GPP (3GPP Work Plan, http://www.3gpp.org/specifications/work-plan).

other aspects. The robustness of LTE control mechanisms (such as mobility, resources management, and scheduling) at lower frequencies makes it very attractive as an anchor technology for the different access techniques used in 5G.

6 Spectrum and Regulations

Spectrum is the key asset for a communications service provider. The available spectrum has a major impact on how a network's maximum capacity and coverage are defined. From VR/AR and autonomous cars to the industrial automation and smart cities, 5G is expected to enable a wide range of new use cases and applications and ultimately define a new framework for future telecommunication systems. Although the mobile industry, academia, and international standards developing organizations have been actively involved in developing the technologies that are central to 5G systems, the success of the 5G services will essentially be heavily reliant on the work of regional regulatory bodies and how quickly they can allocate the right amount and type of spectrum to those services. Significant amount of new and harmonized spectrum are needed to ensure that 5G systems/services can meet the future expectations and deliver the full range of expected capabilities. In this section, frequency bands with ongoing activities and/or potential for 5G are discussed, noting that not all of these bands are included in the studies being carried out within ITU-R in the lead up to WRC-19.

5G systems rely on the spectrum within three frequency ranges in order to deliver extensive coverage and throughput and to support all use cases. Those frequency regions include sub-1 GHz, 1−6 GHz, and above 6 GHz. The sub-1 GHz region will support widespread coverage across urban, suburban, and rural areas and will enable IoT services. The large wavelengths below 1 GHz; however, limit the option to use 5G features such as massive MIMO. That's because handheld devices are typically not large enough to accommodate more than two sub-1 GHz antennas, while size, weight, wind resistance, and visual impact considerations limit the number of deployable antenna elements at base stations. Frequency range 1−6 GHz offers a good tradeoff between coverage and capacity with spectrum within the 3.3−3.8 GHz range expected to form the basis for many initial 5G deployments under 6 GHz [29]. Macrocell network coverage is uplink limited because the 23 dBm maximum output power of mobile devices is much lower than the base station power which can be in excess of 46 dBm. Therefore, a 3500 MHz uplink falls short of LTE 2100 MHz or LTE 1800 MHz coverage. The frequency range above 6 GHz is needed in order to enable extremely high data rates required for some 5G applications where the focus will be on frequency bands above 24 GHz which include growing interest in the 24 GHz and/or 28 GHz bands. There is also some interest in exploring bands in the 6−24 GHz frequency region; however, this range has not been included in WRC-19 sharing studies [30,37].

One of the main objectives of the telecommunication industry is to connect everything and everywhere by more comprehensively and intelligently integrating LTE (in licensed and unlicensed bands), Wi-Fi, and cellular IoT technologies, together with 5G radio interface. This will allow mobile networks to dynamically allocate resources to support the varying needs of an extremely diverse set of network connections ranging from factory automation, sensor networks to self-driving and connected vehicles as well as smartphones and traditional voice services. The increase in the capacity of the 5G radio network needs to be supported by faster and wider bandwidth backhaul that incorporates fast Ethernet, optical fiber, or microwave/mmWave wired/wireless point-to-point links. The use of satellite networks may also provide backhaul for 5G systems despite their limited ability to satisfy 5G's stringent latency and bandwidth requirements.

A central component in the evolution of all mobile technology generations in the past two decades has been the use of continuously increasing bandwidths to support higher peak data rates and larger amounts of user traffic. Ultra-fast 5G services require large amounts of spectrum, compelling the governments and regulators to actively look at significantly higher frequencies that have been traditionally used in mobile services for spectrum allocation. This includes spectrum above 24 GHz where wider bandwidths are more readily available. Without making these higher frequency bands accessible to 5G, it may not be possible to materialize significantly higher data rates and to support rapidly growing mobile data traffic particularly in crowded urban areas.

5G services will struggle to reach beyond urban centers and deep inside the buildings without the sub-1 GHz spectrum. There is existing mobile spectrum in this range which should be refarmed and used in the future. For example, the European Commission has already expressed a desire for the use of 700 MHz band for 5G services in Europe. The United States Federal Communications Commission (FCC)[27] has indicated that the 600 MHz band could be used to drive 5G services in the United States. ITU-R is also considering identifying additional spectrum for mobile broadband from 470 to 694/8 MHz in 2023 which would be well-timed for the large-scale rollout of 5G services, especially if countries prepare to use it quickly after international agreement is reached [30]. The 3.4−3.6 GHz range is practically globally harmonized, which can drive the economies of scale needed for low-cost devices. A number of countries are exploring whether a portion of other bands could be used such as 3.8−4.2 GHz and spectrum in the 4−5 GHz range, in particular 4.8−4.99 GHz region. There are also numerous other mobile bands in the 1−6 GHz range that are currently used for 3G and 4G services which should be gradually repurposed for 5G use. Due to its favorable properties such as radio wave propagation and available bandwidth, the bands in the range of 3300−4200 and 4400−4990 MHz will be the primary spectrum under 6 GHz for the introduction of 5G. Parts of the band between 3300−4200 MHz and 4400−4990 MHz are being considered for the first field trials and

[27] United States FCC, https://www.fcc.gov/.

introduction of 5G services in a number of countries and regions in the world, including Europe (3400–3800 MHz), China (3300–3600, 4400–4500, 4800–4990 MHz), Japan (3600–4200, 4400–4900 MHz), Korea (3400–3700 MHz), and the United States (3100–3550, 3700–4200 MHz).

The above 6 GHz spectrum comprises a combination of licensed and unlicensed mobile bands. There are several mmWave bands including 24.25–27.5, 31.8–33.4, 37–43.5, 45.5–50.2, 50.4–52.6, 66–76, and 81–86 GHz which have been identified for sharing studies by WRC-15 that must be ratified by WRC-19 for worldwide allocation. However, some countries are also investigating other mobile bands above 6 GHz for 5G services, which are not being considered at WRC-19. The 28 GHz band is of particular interest as it has been permitted for 5G use in the United States and is being closely examined by Japan and Korea. This would complement the 24 GHz band, which is being studied at WRC-19 and is supported in the European Union, because the same equipment could easily support both bands thus helping to lower device costs. Fig. 19 illustrates the frequency ranges that are being studied for specification at WRC-19 in different regions of the world [40].

The smaller coverage areas of higher frequencies, when employed terrestrially, provide more frequency reuse opportunities with limited inter-cell interference issues. As such, 5G services in urban areas may be able to occupy the same bands as other wireless services (e.g., satellite and fixed links) which operate in different geographical areas (e.g., rural)

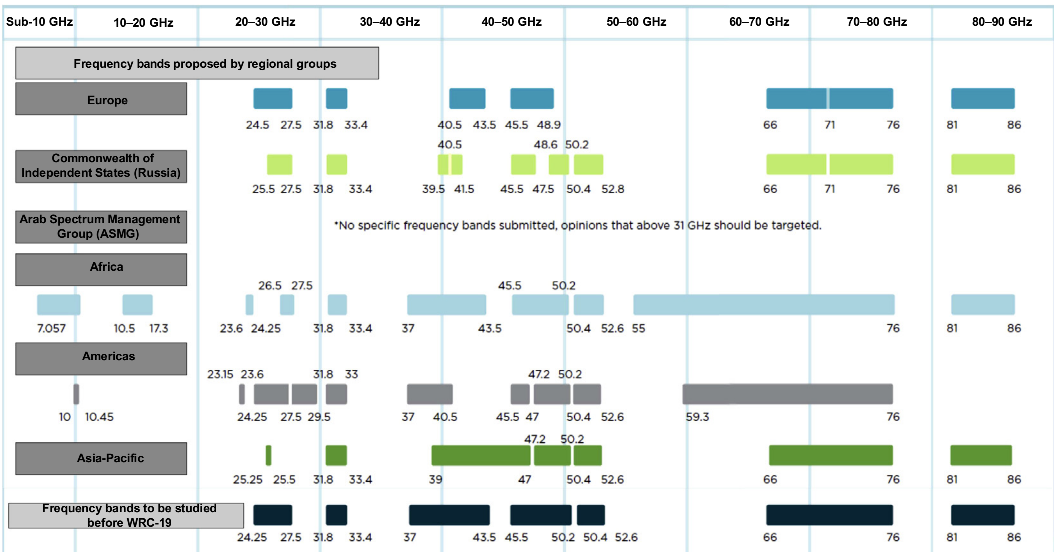

Figure 19
Frequency bands under study for ratification in WRC-19 [40].

when suitable interference mitigation techniques such as beamforming are in place. The potential for spectrum sharing will be explored through sharing studies being conducted by ITU-R, in preparation for WRC-19. Adopting viable sharing methods is particularly important as many of the bands above 24 GHz that are being considered for 5G access are or will be used for 4G/5G mobile backhaul especially in rural areas where fiber links might not be available or practical [30].

As we mentioned earlier, the wide bandwidth requirements for some 5G applications have necessitated comprehensive study of frequency bands above 6 GHz for utilization in various 5G deployment scenarios and use cases. There are ongoing global activities to identify and specify suitable spectrum including frequency bands that can be used in as many countries as possible to enable global roaming and economies of scale. While not very successful for 4G and previous generations, various efforts around the world are underway to achieve spectrum harmonization in 5G networks. In addition to setting different requirements on the network features, new applications will drive a wide variety of deployment scenarios. The different physical characteristics of various spectrum bands (e.g., range, path loss, penetration into structures, propagation) make some applications more suitable in certain frequency bands [46]. While sub-1 GHz spectrum has very good propagation properties that make it the most suitable option for large area coverage, it has limited capacity due to lack of available spectrum and component design considerations. The sub-6 GHz spectrum provides a type of coverage that is more suitable for urban deployment, with increased capacity. The above 6 GHz spectrum is more limited in coverage but can provide very high capacity due to potentially very large bandwidth available at these frequencies.

In late 2015, WRC-15 agreed that ITU-R should conduct sharing and compatibility studies for a number of frequency bands between 24.25 and 86 GHz prior to WRC-19 in late 2019. Some of these frequency bands enable wide contiguous bandwidths, which would allow satisfaction of the stringent requirements of bandwidth-demanding applications. Although the band 27.5–29.5 GHz was not selected for ITU-R studies, the United States and Korea have continued to make progress with trials in this band for 5G which could, when combined with the band 24.25–27.5 GHz (which was selected at WRC-15 for ITU-R studies), provide a good solution for a global implementation of 5G systems. The following frequency bands in sub-6 GHz range have been identified for IMT in ITU-R radio regulations since 1992 [6]:

- 450–470 MHz
- 470–698 MHz
- 694/698–960 MHz
- 1427–1518 MHz
- 1710–2025 MHz
- 2110–2200 MHz

- 2300−2400 MHz
- 2500−2690 MHz
- 3300−3400 MHz
- 3400−3600 MHz
- 3600−3700 MHz
- 4800−4990 MHz

As shown in Table 5, there are two frequency regions which are playing an important role in 5G NR deployments. In the 3−6 GHz band, there is generally globally available spectrum in the range of 3.3−3.8, 3.8−4.2, and 4.4−4.9 GHz. These bands, which are all based on TDD or unpaired spectrum, generally have wider bandwidth than their 4G predecessors. They will be particularly important in user equipment implementations that will use LTE as the anchor (or primary) and the NR as a secondary access scheme (i.e., NSA deployments). 5G systems can also use sub-1 GHz FDD bands to provide wide-area coverage, including deep indoor penetration. The low-band spectrum can take advantage of the new 700 MHz allocation in Europe or 600 MHz allocation in the United States, as well as the refarming of 850/900 MHz released by minimizing the use of legacy 2G/3G spectrum. The aggregation of the different spectrum bands from sub-1 GHz to millimeter waves provides the best combination of coverage, capacity, and user data rates. In addition to the licensed bands in the 3−6 GHz range, there is the possibility of using supplemental unlicensed bands to create even wider operating bandwidth. The 3GPP NR will utilize mmWave spectrum to achieve wider bandwidths and higher data rates. Industry consensus is building around the use of mmWave spectrum for fixed wireless applications due to challenges of mobility management in very high frequencies. Applying mmWave technology to mobile devices represents a significant technological challenge for the near future. Small cells are going to be a critical part of the 5G network deployments. We referred to this trend earlier as network densification. The use of mmWave does leverage large antenna arrays and directional beams in the downlink over a short range, enabling design of high-capacity small cells.

Spectrum harmonization remains an important issue for the development of 5G and even more important for higher frequencies in order to support the development of a new ecosystem as well as the deployment of massive MIMO systems. The United States has adopted new rules to enable rapid development and deployment of 5G technologies and services in licensed spectrum in 28 and 39 GHz. China is also planning to deploy commercial 5G networks to meet the demands for the extremely high peak data rates in the frequency ranges of 26 and 42 GHz. In Europe, the spectrum around 26 GHz has been identified as 5G pioneer band and work is well underway in order to harmonize the frequency bands in Europe for 5G before WRC-19 through adoption of a harmonized band and to promote this band for worldwide use. The use of higher frequency bands for early and commercial deployments of 5G can be summarized as follows [30,40,46]:

Table 5: Candidate spectrum for 5G new radio [48].

Region	Sub-6 GHz Spectrum							mmWave Spectrum			
	New Band		Existing Band		3GPP Band	Bandwidth (MHz)	Total Bandwidth (MHz)	F_{Low} (GHz)	F_{High} (GHz)	Bandwidth (GHz)	Total Bandwidth (GHz)
	F_{Low} (MHz)	F_{High} (MHz)	F_{Low} (MHz)	F_{High} (MHz)							
Korea	3400	3700				300	300	26.50	29.50	3.00	3.00
European	3400	3800	2570	2620	38	50	450	24.25	27.35	3.10	7.70
Union			3400	3800	42 + 43	400		31.80	33.40	1.60	
								40.50	43.50	3.00	
Japan	3600	4200	2496	2690	41	194	1494	27.50	29.50	2.00	2.00
	4400	4900	3400	3600	42	800					
						500					
United	–	–	2496	2690	41	194	344	27.50	28.35	0.85	10.85
States			3550	3700	48	150		37.00	38.60	1.60	
								38.60	40.00	1.40	
								64.00	71.00	7.00	
China	3300	3600	2300	2400	40	100	790	TBD			
	4400	4500	2555	2655	41B	100					
	4800	4990	3400	3600	42	300					
						100					
						190					

- United States: 27.5−28.35 and 37−40 GHz pre-commercial deployments in 2018
- Korea: 26.5−29.5 GHz trials in 2018 and commercial deployments in 2019
- Japan: 27.5−28.28 GHz trials planned from 2017 and potentially commercial deployments in 2020
- China: 24.25−27.5 and 37−43.5 GHz for deployments in 2020 and beyond
- Sweden: 26.5−27.5 GHz awarding trial licenses for use in 2018 and onward
- European Union: 24.25−27.5 GHz for commercial deployments from 2020

Note that the range 24.25−27.5 GHz (26 GHz) is overlapping with the band 26.5−29.5 GHz (28 GHz), which suggests that countries supporting 26 GHz may also benefit from early ecosystem development for the 28 GHz band in other regions. As shown in Table 5, in addition to the bands above 6 GHz, the spectrum in 600, 700, 800, 900, 1500, 2100, 2300, and 2600 MHz may be of particular interest for both legacy and new applications and are key to deliver necessary 5G broadband coverage for IoT, industry automation, and mission-critical usage scenarios.

Extensive research is currently being conducted on the use of much higher frequency bands for 5G systems than are currently used for existing mobile services. The frequencies as high as 300 GHz, according to some reports, have the potential of supporting extremely wideband 5G applications. However, due to their radio propagation characteristics, these bands are likely to be utilized primarily in small cells and may have only limited practical usage outside urban areas. To ensure 5G services provide good coverage that extends beyond small urban hotspots, it will be important to ensure that the suitable spectrum is available in sub-1 GHz region. Significant cooperation among mobile industry leaders, standards developing organizations, and national/regional regulatory bodies is required to pave the way for large-scale commercial 5G deployments in 2020 and beyond. Progressive refarming of existing mobile bands should be possible and permitted to accommodate future 5G services and to ensure spectrum is used as efficiently as possible. The success of refarming is best illustrated by the 1800 MHz band which is now the most common LTE band worldwide and was previously used for 2G services.

A secondary effect of densification of the network with high-data-rate links is the need to improve backhaul transport. Today, optical fiber is the preferred backhaul option. However, as networks grow to become denser with the deployment of more small cells, licensed fixed or point-to-multipoint (PMP) mmWave wireless backhaul may emerge as the most flexible and viable solution. For example, an operator (leveraging guaranteed QoS with licensed PMP) with a 10 Gbps + hub could aggregate backhaul traffic from multiple base stations. The economies of this scenario improve as more base stations are added progressively through network densification. Millimeter wave communications technologies in the 60 GHz and 70−80 GHz frequency range for high capacity in the last mile and pre-aggregation backhaul have been explored in the literature [27,48]. Table 6 provides the existing

Table 6: Mapping of RF components and technologies to spectrum bands [48].

TDD bands	700 MHz–3 GHz	3–6 GHz B42/B43/B46	24.25–29.50 GHz	37–71 GHz
FDD bands	700 MHz–3 GHz	3–6 GHz	-	-
Front-end technology	*Sub-6 GHz frequencies*		*mmWave frequencies*	
	FEMiD[a]/PAMiD[b]/DRx[c]	FEMiD/PAMiD/DRx	8Tx × 8Rx antenna complete front-end	8Tx × 8Rx antenna complete front-end
Power amplifier	III-V[d]/SiGe[e]/Bulk CMOS[f]	III-V/SiGe/Bulk CMOS	InP/SiGe BiCMOS/Advanced SOI	InP/GaN/SiGe BiCMOS/Advanced SOI
Low-noise amplifier	III-V/SiGe/SOI[g] CMOS	III-V/SiGe/SOI CMOS	Advanced SOI/GaN	SiGe BiCMOS/Advanced SOI
RF switching	SOI CMOS	SOI CMOS	Advanced SOI	Advanced SOI
Filtering	Acoustic/IPD/Ceramic	Acoustic/IPD/Ceramic	IPD/Ceramic	IPD
Antenna integration	N/A	N/A	Yes	Yes
Signal generation method	N/A	N/A	Advanced SOI/SiGe BiCMOS	Advanced SOI/SiGe BiCMOS

DRx, Diversity receive module; *FEMiD*, front-end module with integrated duplexers; *IPD*, integrated passive device; *PAMiD*, power amplifier module integrated duplexer; *SOI*; silicon on insulator.

[a]FEMiD.

[b]PAMiD is a multimode multiband RF front-end module that integrates multiplexers, filters, antenna switches, and low-noise amplifiers allowing highly integrated transmit and receive chains.

[c]DRx combines switches, filters, and low-noise amplifiers into a module, providing a highly integrated solution for implementing high-order diversity receive paths for advanced 4G/5G architectures.

[d]A substance that can act as an electrical conductor or insulator depending on chemical alterations or external conditions. Examples include silicon, germanium, and gallium arsenide. It is called III-V materials since semiconductor elements are in groups III and V of the periodic table of chemical elements.

[e]Silicon Germanium process

[f]Bulk CMOS refers to a chip manufactured on a standard silicon wafer.

[g]SOI technology refers to the use of a layered silicon–insulator–silicon substrate in place of conventional silicon substrates in semiconductor manufacturing, especially microelectronics, to reduce parasitic device capacitance, thereby improving performance.

technology and spectrum, as well as the planned 5G NR spectrum. The table further shows the mapping between those applications and the technologies needed to implement power amplifiers, low-noise amplifiers, RF switches, RF filters, and integrated antennas to enable 5G transceivers to be implemented. One important conclusion is that the entire new spectrum for 5G NR, whether sub-6 GHz or greater than 6 GHz, is mainly TDD bands; thus frequency duplexers are not required in order to implement a front-end solution. Filtering, as needed, is accomplished using band-pass filters. In the 3–6 GHz region, filtering can be accomplished using acoustic, integrated passive devices (IPDs)[28] or ceramic technologies.

[28] IPDs or integrated passive components are attracting an increasing interest due to constant needs of handheld wireless devices to further decrease in size and cost and increase in functionality. Many functional blocks such as impedance matching circuits, harmonic filters, couplers and RF baluns, and power combiner/divider can be realized by IPDs technology. IPDs are generally fabricated using standard wafer fabrication technologies such as thin film and photolithography processing. IPDs can be designed as flip chip mountable or wire bondable components and the substrates for IPDs usually are thin film substrates like silicon, alumina or glass.

For the 5G mmWave fixed wireless applications, requirements for massive MIMO and beamforming mean that transmit and receive functions will most likely be in distributed array formats. As a result, there will be multiple transmit/receive chains to accomplish transceiver functionality for fixed/mobile wireless devices. Filter technology for mmWave 5G is likely to be based on transmission line and waveguide cavity technology. micro-electromechanical systems cavity resonators are also an attractive choice to avoid manual-tuning filters and leveraging silicon wafer-based manufacturing approaches. Multi-pole filters suitable for operation in frequencies of 20–100 GHz have been demonstrated in the literature [27]. User equipment PA technology for 5G sub-6 GHz will leverage traditional GaAs heterojunction bipolar transistor (HBT) technology,[29] but it will require some improvement or innovation in dealing with wide bandwidth signals, especially in biasing. The current skills in IEEE 802.11ac and IEEE 802.11ax implementations (up to 160 MHz bandwidth with extremely challenging error vector magnitude requirements) can be leveraged for development of 5G wideband transceivers. Above 6 GHz, the experience in system-in-package technology will be leveraged to create antenna front-ends on organic laminates.

7 Future Outlook

Users' growing reliance on mobile connectivity continues to present technical and economic challenges for the industry and in particular network operators for provisioning adequate coverage and capacity. Increasing utilization of VR/AR services are examples of changes in user behavior that can generate significantly greater performance demands on the wireless networks. In this chapter we have shown various scenarios for how the 5G technologies can enhance coverage and capacity beyond what can be achieved by legacy networks. Some key aspects of this trend include intrinsic capabilities of 5G systems to provide service in various indoor/outdoor environments particularly in dense urban areas, flexible use, and aggregation of available spectrum and access technologies, including Wi-Fi, LTE-U, and other technologies in the unlicensed bands, use of spectrum above 6 GHz to provide high-capacity short-range connectivity in small cells, and flexible network architectures and models of network operation and coverage.

The industry recognizes that sustaining excellent coverage and capacity is not a transitory event but a continuous process. It must be possible to efficiently reconfigure 5G systems to

[29] HBT is a type of bipolar junction transistor (BJT) which uses different semiconductor materials for the emitter and base regions, creating a heterojunction. The effect is to limit the injection of holes from the base into the emitter region, since the potential barrier in the valence band is higher than in the conduction band. Unlike BJT technology, this allows a high doping density to be used in the base, reducing the base resistance while maintaining gain. The HBT improves on the BJT in that it can process signals of very high frequencies, up to several hundred GHz. It is commonly used in modern ultra-fast electronic circuits, RF systems, and in applications requiring high-power efficiency, such as RF power amplifiers in cellular phones.

account for special events, including planned events such as sports games and unplanned events such as natural disasters. For smart devices, coverage requirements include global roaming and support for services on moving platforms such as cars, planes, high-speed trains, and mass transit. 5G will continue to improve the user and operational experience in these areas.

Leveraging recent advances in storage/memory technologies, communication/connectivity, computing, big-data analytics, artificial intelligence (AI), machine learning (ML), machine vision, and other related areas will enable the fruition of immersive technologies such as augmented and virtual reality. The use of virtual reality is expected to go beyond early adopters such as interactive gaming to enhancing cyber-physical and social experiences such as conversing with family, acquaintances, business meeting, or the disabled individuals. The AR/VR technologies would hypothetically allow walking around a street where everyone is talking in a different language, and interacting with those people in the same language in a fully immersive experience. The growing number of drones, robots, and other self-driving vehicles taking cameras to places that humans could never imagine reaching would allow us to collect new content from attractive points of view around the globe. Ultimately, AR/VR will provide the most personal experience with the closest screen, providing the most connected, most immersive experience witnessed thus far.

The concept of content caching has been recently investigated in great detail, where the idea is to cache essential content at the edge of the network (e.g., a base station), devices, or other intermediate locations. One distinguishes between reactive and proactive caching. While the former serves end users when they request contents, the latter is proactive and anticipates users' requests. Proactive caching depends on the availability of fine-grained spatial−temporal traffic predictions. Miscellaneous information such as user's location, mobility patterns, and social behavior and connections can be further exploited to optimize the services provided to the end user. Storage will play a crucial role in AR/VR where for instance upon the arrival of a task query, the network/server needs to promptly decide whether to store the object if the same request will come in the near future or instead recompute the query from scratch if the arrival rate of the queries will be sparse in the future. Content/media location and delivery will also be important in terms of storing different qualities of the same content at various network locations.

Migrating computationally intensive tasks/processes from AR/VR devices to more resourceful cloud/fog servers is necessary in order to relax some additional computational requirements for the low-cost devices and to extend the battery life of such devices. For this purpose, MEC is used to enable the client on-demand access to the cloud/fog resources (e.g., infrastructures, platforms, and software). While the current solutions allocate cloud-based radio and computing resources in a centralized manner, for optimal AR/VR performance, both radio access and computational resources must be brought closer to the users by binding the availability of small-cell base stations with proximity access to computing/storage/memory resources. Furthermore, the network infrastructure must enable a fully distributed cloud-based immersive experience where intensive computations occur on very powerful servers that are in the cloud/network edge while sharing the sensor data that are being delivered by end-user devices at the client side.

Leveraging short-range communication such as D2D and edge proximity services among co-located AR/VR users can help alleviate network congestion. The idea is to extract, appropriately combine and share relevant contextual information among AR/VR users in terms of views and camera feeds. In the context of self- driving vehicles equipped with ultra-high-definition (UHD) cameras capturing their surrounding area, the task for the vehicle is not only to recognize objects/faces in real time, but also to decide which of the objects should be included in the map and to share the information with neighboring vehicles for richer and more context-aware maps.

The introduction of UHD cameras (e.g., 8K cameras with 360-degree panoramic video recorders) has enriched new video and media experiences. User interaction with media content currently falls at two extreme ends of the spectrum. On the one hand, there are lean-back experiences such as movies and television where consumers are passive and are led through a story by content authors/producers. On the other hand, there are lean-forward experiences in the form of games in which the user is highly engaged and drives the action through an environment created by content authors/producers. The next generation of interactive media where the narrative can be driven by authors/producers will be personalized dynamically to the situation and preferences of end users.

The use of context-aware communication has already been promoted as a means of optimizing complex networks. In order to improve user's connected and immersive experience, the users' personal habits, emotions, and other behavioral and psychological aspects must be factored in their service delivery and optimization. This involves predicting users' disengagement and preventing it by dynamically shifting the content to better match individual's preferences, emotional state, and situation. Since a large amount of users' data in the network can be considered for the big-data processing, ML tools can be exploited to analyze the contextual user information and to react accordingly. Of particular importance is the fact that deep learning models have been recently on the rise in ML applications, due to their human-like behavior in training and good performance in feature extraction.

Previous generations of mobile networks enabled voice, data, video, and other impactful services. In comparison, 5G is going to change our society by opening up the telecommunication ecosystem to vertical industries. 5G will help vertical industries to achieve the IoT vision of ubiquitously connected, seamlessly mobile, highly reliable, ultra-low-latency services for massive number of devices. Service-guaranteed network slicing is one of the essential features for 5G to achieve this vision. Prominent operators, vendors, and vertical industries have to come together in order to establish a common understanding of service-guaranteed network slicing in terms of the vision, requirements, end-to-end solution, key enabling technologies, and the impacts on vertical industries.

The 3GPP standard releases are driven by users and operator demands, with smaller releases launched every few years to enhance features and performance of the status quo. Therefore, one might reasonably expect a beyond 5G release to arrive at some point in the future. There is some debate about what systems beyond 5G would entail and whether indeed it is relevant to consider the term, as industry/user requirements will change greatly in the next 10−20 years. A high-level answer is that systems beyond 5G will explore and include relevant technologies that will be left out from 5G, due to being late, experimental in status or simply outside the defined scope for 5G. Future applications and technologies will be integrated/incorporated when they achieve maturity. For systems beyond 5G, one proposal is to integrate terrestrial wireless with satellite systems, for ubiquitous always-on broadband global network coverage. Many of the cellular devices currently connected are machines rather than people, and with the rise of smart homes, smart buildings, and smart cities, 5G and systems beyond 5G will encounter increased demands for machine-to-machine communications, including robotic interactions and autonomous drone delivery and transport systems. Other trends predicted for systems beyond 5G include ultra-dense cellular networks, reconfigurable and fully programmable hardware, and extended use of mmWave spectrum for user access, enhanced optical-wireless interface and photonics, networked visible light communication (VLC),[30] intelligent networking, and technologies to enable full immersive experience for users [28].

The user demand for greater global coverage, higher network capacities, seamless and always-on connectivity, new Internet services and applications continue to grow, and systems beyond 5G will be expected to deliver even more features and services. One of the potential drivers for systems beyond 5G is the growing prospects of software-defined radio and software-defined networking, which means that the underlying technologies for the systems beyond 5G will be easier to upgrade. This reduces the expensive and disruptive upgrades of previous mobile standards which generally involve replacement of physical infrastructure. Inter-vendor interoperability is also an increasing trend, with disaggregation fueled by open-source/white-box development of systems/technologies.

[30] Networked VLC for ultra-dense wireless networking is based on VLC, which promises quantum step improvements in area spectral efficiency while exploiting existing infrastructures by piggy-backing high speed data communication on existing lighting infrastructures. Networked VLC is also referred to as Li-Fi. The visible light spectrum is unlicensed and 10000 times wider than the range of radio frequencies between 0 and 30 GHz. The use of the visible light spectrum for data communication is enabled by inexpensive and off the-shelf available light emitting diodes which also form the basis for next generation energy efficient lighting. Individual LEDs can be modulated at very high speeds of 3.5 Gbps at approximately 2 m distance have been demonstrated using micro-LEDs with a total optical output power of 5 mW [36].

In the early stage of network slicing deployment there will be only a few network slice instances (NSIs)[31] and the deployment may occur in a semi-automatic mode. As the number of NSIs increases and scenarios, such as, dynamic instantiation of NSIs or run-time adaptation of the deployed NSI emerge, more advanced technologies will be needed to support network slicing and its further evolution [31]. Management functions will become real-time, implying that the difference between management and control will gradually disappear. Some management functions will be tightly integrated with the NSIs as well as the network infrastructure. In current networks, technical domains are normally coordinated via centralized network management system. In 5G, performing real-time cross-domain coordination through distributed lower layer such as control plane would be possible, with potentially unified control logic of different domains. Advanced automation and AI algorithms can be applied in a unified, holistic network manner, which would be scalable and flexible, which might then result in run-time deployment and adaptation of NSIs. The 5G networks are not only envisioned as a support for IoT, but also as means to give rise to an unprecedented scale of emerging industries. IoT requires support for a diverse range of service types, such as e-health, Internet of Vehicles, smart households, industrial control, and environment monitoring. These services will drive the rapid growth of IoT and facilitate numerous devices to connect to the network, leading to the notion of the Internet of Everything especially for vertical industries.

From ITU-R perspective, the IMT systems serve as a communication tool for people and a facilitator which enable the development of other industry sectors, such as medical science, transportation, and education. IMT systems will continue to contribute to the following aspects [1,4,5]:

- Wireless Infrastructure: Broadband wireless connectivity will attain the same level of importance as access to electricity. IMT will continue to play an important role in this context as it will act as one of the key pillars to enable mobile service delivery and information exchange. In the future, private and professional users will be provided with a wide variety of applications and services, ranging from infotainment services to new industrial and professional applications.

[31] NSI is defined as a set of network functions and resources to run network functions, forming a complete instantiated logical network to meet certain network characteristics required by the service instance(s). A network slice instance may be fully or partly, logically and/or physically, isolated from another network slice instance. The resources consist of physical and logical resources. A network slice instance may be composed of subnetwork instances, which as a special case may be shared by multiple network slice instances. The network slice instance is defined by a network slice blueprint. Instance-specific policies and configurations are required when creating a network slice instance. Network characteristics examples are ultra-low-latency, ultra-high reliability, etc. [25,26].

- Integrated Information and Communication Technologies (ICT): The development of future IMT systems is expected to promote the emergence of an integrated ICT industry which constitute a driver for economies around the globe. Some possible areas include the accumulation, aggregation and analysis of big data as well as delivering customized network services for enterprise and social network groups on wireless networks.
- Service Availability and Affordability: IMT systems will continue to help closing the gaps caused by an increasing digital divide. Affordable, sustainable, and easy-to-deploy mobile and wireless communication systems can support this objective while effectively saving energy and maximizing efficiency. IMT systems will enable sharing of any type of content anytime, anywhere through any device. Users will generate more content and share the content without being limited by time and location.
- Energy Efficiency: IMT enables energy efficiency across a range of industries by supporting machine-to-machine communication and solutions such as smart grid, teleconferencing and telepresence, smart logistics, and transportation.
- Education and Culture: IMT systems can change the method of education by providing easy access to digital textbooks or cloud-based storage of knowledge on the Internet, advancing applications such as e-learning, e-health, and e-commerce. IMT systems will support people to create collaborative art works or remotely/virtually participate in group performances or activities.

It is an inevitable fact that future spectrum is going to be highly fragmented, ranging from frequencies below 6 GHz to the mmWave band and THz frequency region. New hardware design and manufacturing paradigms are key to the exploitation of such a fragmented spectrum and in particular for the mmWave part of the spectrum, essential for achieving the data rates required for systems beyond 5G. Electronically steered high-gain 3D antennas and high levels of RF/DSP integration are going to be essential. With unprecedented aggregate data rates in the backbone network, optical fiber connections and wireless backhaul at E and W-bands and beyond will be required. It is an immense challenge to realize hardware (i.e., transceivers, filters, power amplifiers, low-noise amplifiers, mixers, and antennas) at frequencies from 28 to 300 GHz and beyond, with the low manufacturing costs demanded by network operators. Silicon RFIC technology is the key enabler, potentially even up to 1 THz, but considerable advances in RF design techniques are required in order to realize complete beyond 5G subsystems. Furthermore, while mmWave spectrum is the key to providing the expected 10 Gbps + data rates, it should not be forgotten that the massive IoT presents another set of RF design challenges, such as ultra-low-power operation, energy harvesting, and sensing in ultra-low-cost technologies such as printed and wearable electronics [48].

References

ITU-R Specifications[32]

[1] Recommendation ITU-R M.2083-0, IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond, September 2015.
[2] Recommendation ITU-R M.1645—Framework and Overall Objectives of the Future Development of IMT 2000 and Systems Beyond IMT 2000, June 2003.
[3] Recommendation ITU-R M.2012—Detailed Specifications of the Terrestrial Radio Interfaces of International Mobile Telecommunications Advanced (IMT-Advanced), January 2012.
[4] Report ITU-R M.2320—Future Technology Trends of Terrestrial IMT Systems, November 2014.
[5] Report ITU-R M.2370—IMT Traffic Estimates for the Years 2020 to 2030, July 2015.
[6] Report ITU-R M.2376—Technical Feasibility of IMT in Bands Above 6 GHz, July 2015.
[7] Report ITU-R M.2133, Requirements, Evaluation Criteria and Submission Templates for the Development of IMT-Advanced, December 2008.
[8] Report ITU-R M.2134—Requirements Related to Technical Performance for IMT Advanced Radio Interface(s), December 2008.
[9] Report ITU-R M.2135-1, Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced, December 2009.
[10] Report ITU-R M. [IMT-2020.EVAL], Guidelines for Evaluation of Radio Interface Technologies for IMT-2020, June 2017.
[11] Report ITU-R M. [IMT-2020.TECH PERF REQ], Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s), March 2017.
[12] FG IMT-2020: Report on Standards Gap Analysis, Focus Group on IMT-2020, IMT-O-016, October 2015.

3GPP Specifications[33]

[13] 3GPP TR 23.799: Study on Architecture for Next Generation System (Release 14), December 2016.
[14] 3GPP TR 38.913: Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14), March 2017.
[15] 3GPP TR 38.912: Study on New Radio (NR) Access Technology (Release 14), March 2017.
[16] 3GPP TR 38.801: Study on New Radio Access Technology: Radio Access Architecture and Interfaces (Release 14), March 2017.
[17] 3GPP TS 38.300: NR; NR and NG-RAN Overall Description, Stage 2 (Release 15), April 2019.
[18] 3GPP TR 38.802: Study on New Radio Access Technology, Physical Layer Aspects (Release 14), March 2017.
[19] 3GPP TR 38.803: Study on New Radio Access Technology, RF and Co-existence Aspects (Release 14), March 2017.
[20] 3GPP TR 38.804: Study on New Radio Access Technology, Radio Interface Protocol Aspects (Release 14), March 2017.
[21] 3GPP TS 22.261: Service Requirements for the 5G system, Stage 1 (Release 16), June 2017.
[22] 3GPP TR 38.900: Study on Channel Model for Frequency Spectrum Above 6 GHz (Release 14), December 2016.

Articles, Books, White Papers, and Application Notes

[23] China Mobile Research Institute, C-RAN: The Road Towards Green RAN (Version 3.0), June 2014.
[24] China Mobile Research Institute, NGFI: Next Generation Fronthaul Interface (Version 1.0), June 2015.

[32] ITU-R specifications can be accessed at the following URL: http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/.

[33] 3GPP specifications can be accessed at the following URL: http://www.3gpp.org/ftp/Specs/archive/.

[25] NGMN Alliance, NGMN 5G White Paper v1.0, February 2015.
[26] NGMN, Further Study on Critical C-RAN Technologies, April 2015.
[27] D. Bojic et al., Advanced wireless and optical technologies for small cell mobile backhaul and dynamic software-defined management, IEEE Commun. Mag., 51 (9) (2013).
[28] M. Dohler, et al., Beyond 5G Challenges: White Paper, The University of Sheffield, March 2015.
[29] Global mobile Suppliers Association (GSA), The Future of IMT in the 3300-4200 MHz Frequency Range, June 2017.
[30] GSM Association, 5G Spectrum, Public Policy Position, November 2016.
[31] 5G Service-Guaranteed Network Slicing White Paper, China Mobile Communications Corporation, Huawei Technologies Co., Ltd., Deutsche Telekom AG, Volkswagen, February 2017.
[32] Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society-II (METIS II): Deliverable D1.1, Refined Scenarios and Requirements, Consolidated Use Cases, and Qualitative Techno-Economic Feasibility Assessment, January 2016.
[33] Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society (METIS), Deliverable D6.1, Simulation Guidelines, October 2013.
[34] K. Mallinson, The path to 5G: as much evolution as revolution, 3GPP. <http://www.3gpp.org/news-events/3gpp-news/1774-5g_wiseharbour>, May 2016.
[35] 5G Forum, Republic of Korea, 5G vision, requirements, and enabling technologies (V.1.0). <https://www.5gforum.org>, March 2015.
[36] IEEE Std 802.15.7-2011 – IEEE Standard for Local and Metropolitan Area Networks—Part 15.7: Short-Range Wireless Optical Communication Using Visible Light, September 2011.
[37] Evolving LTE to fit the 5G future, Ericsson Technology Review, January 2017.
[38] Wikipedia, Information-centric networking. <https://en.wikipedia.org/wiki/Information-centric_networking>.
[39] 5G-PPP Use Cases and Performance Evaluation Models. <http://www.5g-ppp.eu/>, April 2016.
[40] NTT DoCoMo Technical Journal, ITU Radio-Communication Assembly 2015 (RA-15) Report – Future Mobile Phone Technologies Standardization – 2015 ITU World Radio-communication Conference (WRC-15) Report – Standardization of Mobile Phone Spectrum Vol. 18 No. 1, July 2016.
[41] NMC Consulting Group (NETMANIAS), Timeline of 5G Standardization in ITU-R and 3GPP, January 2017.
[42] NMC Consulting Group (NETMANIAS), Network Architecture Evolution from 4G to 5G, December 2015.
[43] NMC Consulting Group (NETMANIAS), E2E Network Slicing – Key 5G Technology: What Is It? Why Do We Need It? How Do We Implement It?, November 2015.
[44] Ixia, Test Considerations for 5G, White Paper, July 2017.
[45] 5G Reimagined: A North American Perspective, Alliance for Telecommunications Industry Solutions (ATIS), February 2017.
[46] 5G Americas, White Paper on 5G Spectrum Recommendations, April 2017.
[47] Qualcomm Technologies, Inc. LTE-U/LAA, MuLTEfire™ and Wi-Fi; Making Best Use of Unlicensed Spectrum, September 2015.
[48] Skyworks Solutions, Inc., White Paper, 5G in Perspective: a Pragmatic Guide to What's Next, March 2017.