NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

MSc THESIS

Study, Measurements and Characterisation of a 5G system using a Mobile Network Operator Testbed

Ukiwo J. Anya

Supervisor: **Dimitris Syvridis,** Professor
Co-supervisor: **George Agapiou,** Dr.

ATHENS

SEPTEMBER 2020

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Study, Measurements and Characterisation of a 5G system using a Mobile Network Operator Testbed

Ukiwo J. Anya

Επιβλέποντες: **Dimitris Syvridis,** Professor
George Agapiou, Dr.

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2020

MSc THESIS

Study, Measurements and Characterisation of a 5G system using a Mobile Network Operator Testbed

Ukiwo J. Anya

S.N.: 1199001

SUPERVISOR: **Dimitris Syvridis,** Professor
CO-SUPERVISOR: **George Agapiou,** Dr.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Study, Measurements and Characterisation of a 5G system using a Mobile Network Operator Testbed

Ukiwo J. Anya

A.M.: 1199001

Επιβλέποντες: **Dimitris Syvridis,** Professor
George Agapiou, Dr.

ABSTRACT

The goals for 5G are aggressive. It promises to deliver enhanced end-user experience by offering new applications and services through gigabit speeds, and significantly improved performance and reliability. The enhanced mobile broadband (eMBB) 5G use case, for instance, targets peak data rates as high as 20 Gbps in the downlink (DL) and 10 Gbps in the uplink (UL).

While there are different ways to improve data rates, spectrum is at the core of enabling higher mobile broadband data rates. 5G New Radio (NR) specifies new frequency bands below 6 GHz and also extends into mmWave frequencies where more contiguous bandwidth is available for sending lots of data. However, at mmWave frequencies, signals are more susceptible to impairments. Hence, extra consideration is needed to determine test approaches that provide the precision required to accurately evaluate 5G components and devices.

Therefore, the aim of the thesis is to provide a deep dive into 5G technology, explore its testing and validation, and thereafter present the OTE (Hellenic Telecommunications Organisation) 5G testbed, including measurement results obtained and its characterisation based on key performance indicators (KPIs).

SUBJECT AREA: Wireless Communications

KEYWORDS: 5G, mmWave, massive MIMO, Over-the-Air Testing, Testbed

In loving memory of my mother, Mrs Ure Ukiwo (née Kalu Agbai), who passed away on 28 August 2020

ACKNOWLEDGEMENTS

I wish to specially thank my supervisor and co-supervisor, Prof. Dimitris Syvridis and Dr. George Agapiou, for assisting me immensely in realising this thesis.

Thanks also to Prof. Stylianos Sygletos and Prof. Sergei Turitsyn for their great contributions in setting up and running the SMARTNET programme.

Finally, I wish to thank all the administrative staff of the programme, most especially Zorina, who were always available to help whenever needed.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This Master Thesis has been accomplished within the framework of the European Funded Project: SMART Telecom and Sensing Networks (SMARTNET) - Erasmus+ Programme coordinated by Aston University, and with the participation of Télécom SudParis (TSP) and National and Kapodistrian University of Athens.

The goal of SMARTNET is to provide training in the inter-disciplinary fields of photonic and 5G wireless technologies for data communication, sensing and big data processing.

It has been my motivation right from the start of my enrolment in the programme to write a final thesis focused on 5G, which has been billed as an innovative technology that will radically transform the way we live, work, and play.

Therefore, this thesis with the title “Study, Measurements and Characterisation of a 5G system using a Mobile Network Operator (MNO) Testbed” is my attempt at giving a tangible outcome to my motivation.

1. INTRODUCTION

1.1 Introduction to 5G

The 5th Generation (5G) of mobile networks has arrived and promises to deliver a unifying connectivity fabric that will take on a much bigger role than previous generations. For more than three decades, mobile networks have evolved to interconnect people in new and better ways. Although this evolution of new features and technologies is continuous, roughly every 10 years a new generation of mobile technologies is introduced that delivers a big leap in performance, efficiency, and capability. And now, the world sits at the dawn of the fifth generation of mobile — 5G, a unifying connectivity fabric for a broad range of industries that will bring significant economic and societal benefits [1].

It is a brand-new network that will not only interconnect people, but also interconnect and control machines, objects, and devices — a platform for innovations that can enable new services, empower new user experiences, and connect new industries. 5G will considerably enhance and lower cost for mobile broadband services available today on 3G and 4G networks, while also bringing new kinds of services to life, such as enabling mission-critical control through ultra-reliable, low-latency links and connecting the massive Internet of Things [1].

While the first four generations of mobile networks connected people by delivering better voice and faster data services, it is anticipated that 5G will do and connect much more. 5G is a platform for innovations that will redefine a broad range of industries by connecting virtually everyone and everything, from workers and patients to machines and crops, supporting the connectivity needs across an array of world-changing use cases. 5G will bring together people's worlds to achieve new levels of efficiency that will benefit the entire society. The Figure 1 below shows some example applications that 5G will bring to life [1].

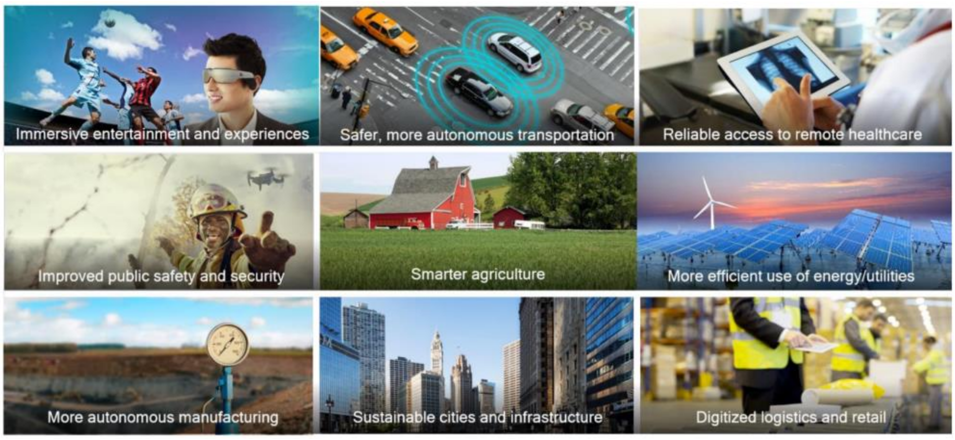

Figure 1: 5G value creation

Unlike previous generations of mobile networks, the fifth generation (5G) technology is expected to profoundly transform the role that telecommunications technology plays in society. 5G is also expected to enable further economic growth and widespread digitalisation of a hyper-connected society, where not only are all people are connected to the network whenever needed, but also many other devices/things thereby creating a society with everything connected (i.e. Internet of Everything). It is not a coincidence that governments around the world (especially in the most advanced economies such as China, EU, Japan, Korea, and USA) have accelerated the introduction of 5G technology in their respective markets [2].

1.2 Mobile Networks and the Evolution to 5G

The 1st generation cellular networks were basic analogue systems designed for voice communications. A move to early data services and improved spectral efficiency was achieved in 2G systems using digital modulations and time division or code division multiple access. 3G introduced high-speed Internet access, highly improved video, and audio streaming capabilities using technologies such as Wideband Code Division Multiple Access (W-CDMA) and High-Speed Packet Access (HSPA). HSPA is a combination of two mobile telephony protocols, High-Speed Downlink Packet Access (HSDPA) and High-Speed Uplink Packet Access (HSUPA), which extends and improves the performance of existing 3G mobile telecommunication networks utilising WCDMA protocols. An improved 3rd Generation Partnership Project (3GPP) standard, HSPA+ (or Evolved HSPA), was released in late 2008 with subsequent worldwide utilisation beginning in 2010. HSPA has been deployed in over 150 countries by more than 350 communications service providers (CSP) on multiple frequency bands and is now the most extensively sold radio technology worldwide, although LTE is closing the gap quickly [3].

The International Mobile Telecommunications-Advanced (IMT-Advanced) standard includes capabilities outstripping those of IMT-2000 (3G) mobile communication. The International Telecommunication Union (ITU) refers to IMT-Advanced as a 4G mobile communications technology. LTE radio access technology has been developed by the 3GPP to offer a fully 4G-capable mobile broadband platform. LTE is an orthogonal frequency-division multiplexing (OFDM)-based radio access technology that supports a scalable transmission bandwidth up to 20 MHz and advanced multi-antenna transmission. As a key technology in supporting high data rates in 4G systems, Multiple-Input Multiple-Output (MIMO) enables multi-stream transmission for high spectrum efficiency, improved link quality, and adaptation of radiation patterns for signal gain and interference mitigation via adaptive beamforming using antenna arrays [4]–[6]. The combination of HSPA and LTE increased the peak mobile data rates of the two systems, with data rates exceeding 100 Mbps, and also allowed for optimal dynamic load balancing between the two technologies [3].

As the demand for capacity in mobile broadband communications increased dramatically every year, wireless carriers prepared to support up to a thousand-fold increase in total mobile traffic by 2020, which required researchers to seek greater capacity and to find a new wireless spectrum beyond the 4G standard [3].

Then came 5G which is the next generation of mobile standards defined by the ITU. IMT-2020 (5G) is a name for the systems, components, and related elements that support enhanced capabilities beyond those offered by IMT-2000 (3G) and IMT-Advanced (4G) systems [7].

5G promises to deliver enhanced end-user experience by offering new applications and services through gigabit speeds, and significantly improved performance and reliability. It will build on the successes of 2G, 3G and 4G mobile networks, which have transformed societies, supporting new services and new business models. It will also provide an opportunity for mobile network operators (MNOs) to move beyond providing connectivity services, to developing rich solutions and services for consumers and industry across a range of sectors –and at affordable cost. 5G is an opportunity to implement wired and wireless converged networks and offers opportunities in integrating network management systems [7].

Commercial 5G networks are expected to be deployed from 2020, as shown in Table 1, as 5G standards are finalised. By 2025, the GSM Association (GSMA) expects 5G connections to reach 1.1 billion, around 12 per cent of total mobile connections. It also forecasts overall operator revenues to grow at a Compound Annual Growth Rate (CAGR) of 2.5 per cent, to reach USD 1.3 trillion by 2025 [7].

Table 1: Evolution of mobile networks

	1G	2G	3G	4G	5G
Approximate deployment date	1980s	1990s	2000s	2010s	2020s
Theoretical download speed	2kbit/s	384kbit/s	56Mbit/s	1Gbit/s	10Gbit/s
Latency	N/A	629 ms	212 ms	60-98 ms	< 1 ms

5G is also expected to increase data rates dramatically and reduce latency compared to 3G and 4G. This significant latency reduction is expected to be below 1 ms, suited to mission-critical services where data are time-sensitive. Its high-speed capability means it can deliver a range of high-speed broadband services and offer an alternative to last-mile access such as fibre-to-the-home (FTTH) or copper connections [7].

As fifth generation (5G) is developed and implemented, the main differences compared to 4G are the use of much greater spectrum allocations at untapped mm-wave frequency bands, highly directional beamforming antennas at both the mobile device and base station, longer battery life, lower outage probability, much higher bit rates in larger portions of the coverage area, lower infrastructure costs, and higher aggregate capacity for many simultaneous users in both licensed and unlicensed spectrum. The backbone networks of 5G will move from copper and fibre to mmWave wireless connections, allowing rapid deployment and mesh-like connectivity with cooperation between base stations [3].

Essentially, 5G is a paradigm shift that includes very high carrier frequencies with massive bandwidths, extreme base station and device densities, and unprecedented numbers of antennas [8].

In a nutshell, the main target of wireless systems, from 1G to 4G, is the quest for increasingly higher wireless speeds to support the transition from voice-centric to multimedia-centric traffic. As wireless speeds have approached those of their wired counterparts, the mission of 5G is different and much more complex, which is to support the explosive evolution of ICT and Internet. In terms of functions, 5G systems will support communications, computing, control, and content delivery (4C). In terms of applications, a broad range of new applications and services for 5G are emerging, such as real-time online gaming, virtual reality (VR) and ultra-high-definition (UHD) video streaming, which require unprecedented high access speed and low latency [9].

1.3 Why is 5G Needed?

Initial interest and discussions about a possible 5G standard evolved into a full-fledged conversation that has captured the attention and imagination of researchers and engineers around the world. As the 4G LTE system reached maturity, where only incremental improvements and small amounts of new spectrum can be expected, researchers began to ponder “what’s next?” [10]

However, this was not a mere intellectual exercise. According to the annual visual network index (VNI) reports released by Cisco, quantitative evidence shows that the wireless data explosion is real and will continue. Driven largely by smartphones, tablets, and video streaming, VNI report forecasts that an incremental approach will not come close to meeting the demands that networks will face by 2020 [11]. In addition to the sheer volume of data, the number of devices and the data rates will continue to grow exponentially [8].

According to a 2019 Ericsson Mobility Report, mobile data traffic grew 82 percent year on year in the first quarter of 2019. The report predicts mobile traffic will rise at a compound annual growth rate of 30 percent between 2018 and 2024. It forecasts a total of 8.8 billion mobile subscriptions by 2024, including 1.9 billion for 5G enhanced mobile broadband [12].

Moreover, with the introduction of a myriad of smart hand-held devices, user demands for mobile broadband are undergoing an unprecedented rise. The drastic growth of bandwidth-hungry applications such as video streaming and multimedia file sharing are already pushing the limits of current cellular systems. Some envisioned media-rich mobile applications such as telepresence and three-dimensional (3D) holography will require data rates simply not possible with fourth generation (4G) networks [13]. Therefore, researchers and engineers have worked to meet these intense demands via innovative new technologies [8].

The popular sentiment is that 5G will be an innovation engine, bringing disruptive change across industries and society, and creating new use cases, new revenue streams, and new business models for industries and consumers. With 5G, industries will have connectivity that is customised for their requirements and the agility to move quickly to meet customer needs [14].

In addition to the high demand for more network capacity, there are a number of other factors that make 5G interesting, including the potentially disruptive move to mmWave spectrum, new market-driven ways of allocating and re-allocating bandwidth, a virtualisation of the core network that will progressively spread to the edges, the possibility of an “Internet of Things” comprised of billions of devices, and the increasing integration of past and current cellular and WiFi standards to provide an pervasive high-rate, low-latency experience for network users [8].

And even though, many 5G conversations focus on higher data rates, 5G will offer other significant benefits, including higher reliability for mission-critical communications, lower latencies that enable new gaming and business applications, and massive IoT connectivity that can support tens of thousands of sensors per cell. The 5G standard has been purposefully designed to provide higher flexibility allowing service providers to make a variety of trade-offs between speed, capacity, energy consumption (e.g., battery life), and cost [15].

5G is set to become an integral part of the world we live in and the way our world works. With 5G:

- Peak speeds will reach and exceed 1Gbps.
- Mobile networks will manage traffic more efficiently than with 4G. This means network capacity will increase, so that users will enjoy higher and more consistent average speeds-even in crowded scenarios or in areas with less-than-ideal coverage.
- Latency[1] will decrease and it will continue to do so overtime as 5G devices evolve. As the time between performing an action (such as moving a character in an

[1] Latency is the round-trip time it takes for a packet to go to and from the application server, measured in milliseconds.

online game) and getting a response will be reduced, user experience will be greatly improved.

- More devices will be able to connect to a 5G cell site, supporting the expected explosion in the number of devices as part of the Internet of Things (IoT) [16].

It is a springboard for service providers to bring a range of new services and service improvements to market, as presented in Figure 2 [15].

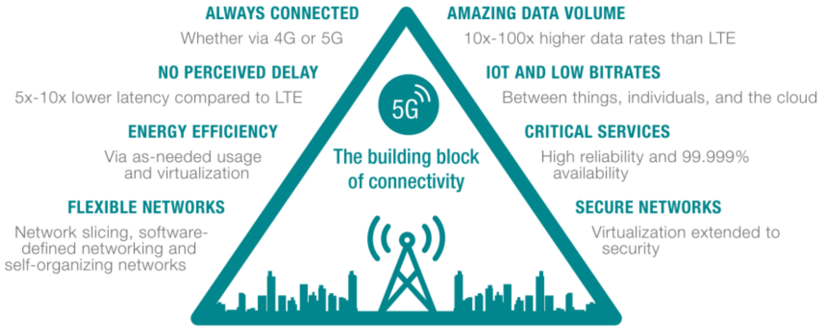

Figure 2: 5G opportunities

1.4 Thesis Outline and Contribution

The remainder of this document is structured as follows:

Chapter 2 examines 5G development standardisation and the timeline, the envisioned usage scenarios and key performance indicators (KPIs).

Chapter 3 is an in-depth look at 5G New Radio (NR), the new air interface or radio access technology developed for 5G.

Chapter 4 explains the various enabling technologies for 5G and why they are needed.

Chapter 5 describes the architecture of the 5G Next Generation System (NGS) and its components.

Chapter 6 provides an overview of 5G deployment options and some new features that have been incorporated.[examples of how policy makers are starting to work through the issues associated with deploying 5G networks]

Chapter 7 explores the testing and validation of 5G focusing on what is tested, how and why. It, thereafter, concludes with a presentation of the OTE 5G testbed.

Chapter 8, the final chapter, presents measurement results obtained from the testbed and its characterisation, and then makes some recommendations for future work.

2. 5G DEVELOPMENT

2.1 Standardisation and Timeline

The International Telecommunications Union (ITU) allocates global radio spectrum and satellite orbits, developing the technical standards that ensure that networks and technologies can seamlessly interconnect, and strive to improve access to information and communications technologies to underserved communities worldwide. Since the advent of 3G mobile communication, each generation of wireless communication has been associated with a specific International Mobile Telecommunications (IMT) technology within the ITU Radiocommunication Sector (ITU-R) [17].

Starting from year 2000, 3G wireless access corresponded to IMT-2000, the global standard for 3G that has opened the way to enabling innovative applications and services (e.g. multimedia entertainment, infotainment and location-based services, among others), while 4G wireless access corresponds to IMT-Advanced [17].

ITU's current family of standards for International Mobile Telecommunication systems (IMT-2000 and IMT-Advanced), serve as a global framework and reference for today's 3G and 4G mobile systems. As an evolution of that, ITU set the stage for the 5th Generation of mobile wireless broadband technologies, known as "IMT-2020" [18].

In 2015, ITU published Recommendation ITU-R M.2083: "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond" on the vision of the 5G mobile broadband connected society and future IMT [19]. This Recommendation defines the framework and overall objectives of the future development of International Mobile Telecommunications (IMT) for 2020 and beyond. It includes a wide variety of capabilities associated with foreseen usage scenarios [18].

The standards for 5G are approved by the ITU-R. Its Working Party 5D (WP 5D) prepared evaluation criteria [20] followed by submissions of proposals and evaluation of candidate technologies. This process was completed in late 2019, leading to the first certified 5G standards. ITU Telecommunication Standardisation Sector (ITU-T) completed a study into networking innovations required to support the development of 5G systems through a Focus Group on IMT-2020 as part of Study Group 13. This study took a system-wide view of 5G architectures and encompassed Proofs-of-Concept (PoC) [21].

The ITU worked with operators, network equipment manufacturers and standards organisations to define the International Mobile Telecommunications 2020 (IMT-2020) vision [22].

The 5G vision defines an advanced mobile broadband communication system that can support an ultra-connected society [23]. The high speeds and low latency promised by 5G will propel societies into a new age of smart cities and the Internet of Things (IoT) [7].

3rd Generation Partnership Project (3GPP), which unites seven telecommunications standard organisations, develops the protocols for mobile telecommunications. 2G, 3G and 4G networks are based the work done by 3GPP and 5G will be too [16].

3GPP uses a system of parallel "Releases" which provide developers with a stable platform for the implementation of features at a given point and then allow for the addition of new functionality in subsequent Releases. Figure 3 shows the timeline for the 3GPP Releases [17].

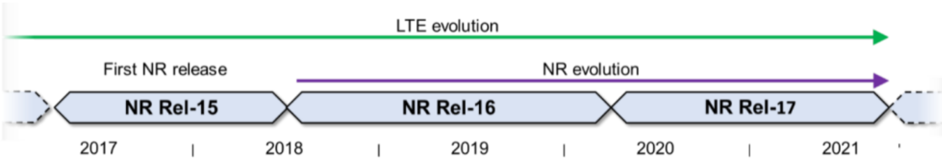

Figure 3: 3GPP timeline for Releases 15, 16, and 17

Standardisation of 5G technology is broken into two phases. The goal of this phased standardisation approach is to complete initial specifications to allow deployments in the 2020 timeframe. Phase 1 will be completed by September 2018 in Release 15. Phase 2 will incorporate more functions to extend the capabilities of 5G to progressively support more services, scenarios, and higher frequency bands. Phase 2 will be completed around the end of 2019 in Release 16 [21].

By using a phased approach (Figure 4), ITU and 3GPP enable widespread 5G commercialisation from 2020. Field trials have already taken place in select cities around the world. The first commercial networks rolled out in late 2018 [23].

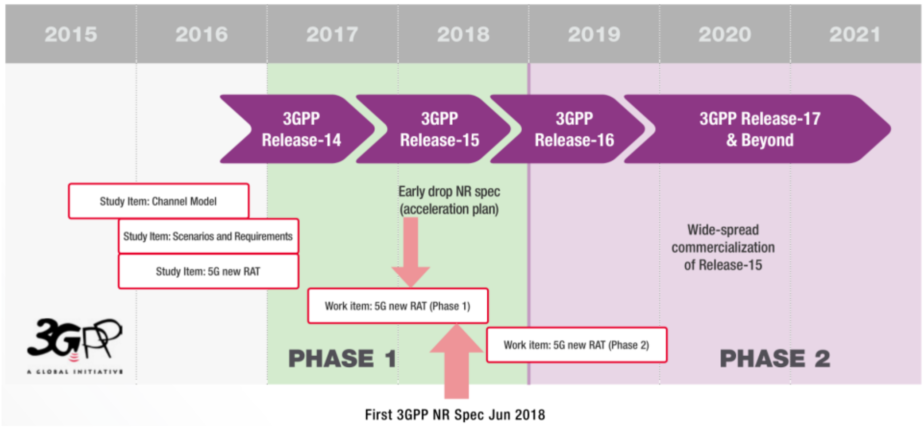

Figure 4: 5G release phases

Figure 5 illustrates the ITU timeline for IMT-2020 and the corresponding 3GPP submissions. After the finalisation of the IMT-2020 requirements in mid-2017, ITU-R opened for the submission of IMT-2020 proposals. 3GPP made an initial submission to ITU in December 2017 and a final submission in June 2019 [17].

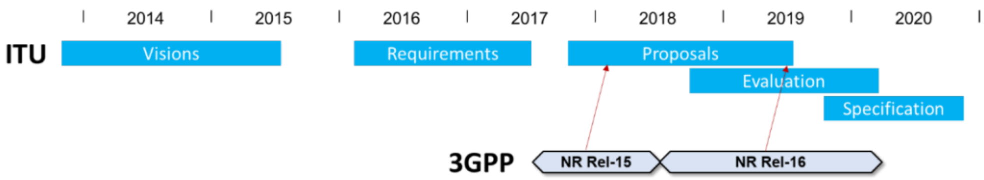

Figure 5: Timeline for IMT-2020 (5G) in ITU and 3GPP submission

So far, major 3GPP milestones are as follows:

- December 2017: Non-Standalone New Radio (Release 15)
- June 2018: Standalone 5G New Radio and New Core (Release 15)
- December 2018: Further RAN – Core Network deployment options (Release 15).
- December 2019: Further evolution of 5G New Radio (Release 16) [16].

Overall, the mobile wireless industry continues to make great strides in research, development, standardisation, and deployment of 5G technologies. The evolution and revolution in wireless continues with new standardised technical features at 3GPP as the mobile wireless industry connects more people and things [17].

2.2 Usage Scenarios

The 5G vision set forth by IMT-2020 is amazing. Three defined use cases represent the foundation for the 5G specifications: enhanced mobile broadband (eMBB) to support extreme data rates, ultra-reliable low-latency communications (URLLC) for near instant communications, and massive machine-type communications (mMTC) for massive IoT interconnects (Figure 6). [24].

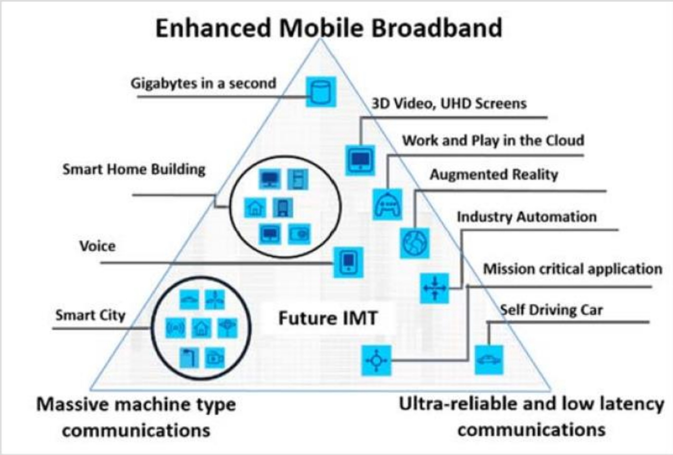

Figure 6: 5G usage scenarios

Enhanced Mobile Broadband: The enhanced mobile broadband usage scenario will come with new application areas and requirements, in addition to existing mobile broadband applications, for improved performance and an increasingly seamless user experience [19].

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 [19].

Massive machine type communications: This use case is characterised by a very large number of connected devices typically transmitting a relatively low volume of non-delay-sensitive data. Devices are required to be low cost and have a very long battery life [19].

2.3 Key Performance Indicators

The Key Performance Indicators (KPIs) of 5G (IMT-2020) compared to 4G is depicted in Figure 7 below.

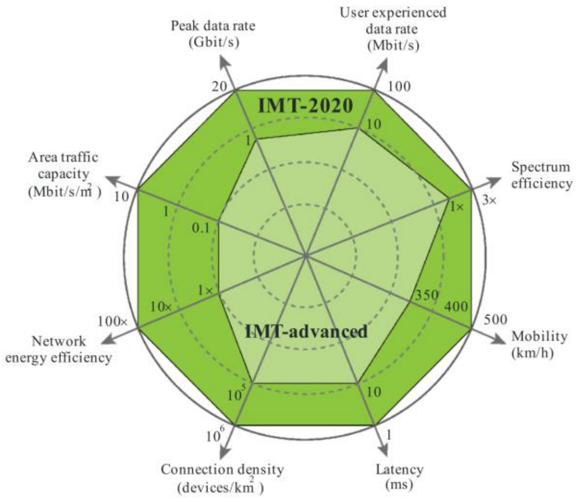

Figure 7: Enhancement of KPIs from 4G to 5G

Firstly, its peak data rate for eMBB is expected to reach 10 Gbps. However, under certain conditions and scenarios IMT-2020 would support up to 20 Gbps peak data rate. 5G would support different user experienced data rates covering a variety of environments for eMBB. For wide area coverage cases, e.g. in urban and sub-urban areas, a user-experienced data rate of 100 Mbps is expected to be enabled. In hotspot cases, the user-experienced data rate is expected to reach higher values (e.g. 1 Gbps indoor) [7].

Spectrum efficiency is expected to be three times higher compared to 4G for eMBB. The achievable increase in efficiency from 4G will vary and could be higher in some scenarios. 5G is expected to support 10 Mbps/m^2 area traffic capacity, for example in hot spots [7].

The energy consumption for the radio access network (RAN) of 5G should not be higher than those of earlier mobile network generations, even as it delivers enhanced capabilities. The network energy efficiency should therefore be improved by a factor at least as great as the envisaged traffic capacity increase of 5G relative to 4G for eMBB [7].

5G would be able to provide 1 ms over-the-air latency, capable of supporting services with very low latency requirements. It is also expected to enable high mobility up to 500 km/h with acceptable Quality of Service (QoS). This is envisioned for high-speed trains [7].

Finally, 5G is expected to support a connection density of up to 10^6/km^2, for example in mMTC scenarios [7].

Whilst all key capabilities may to some extent be important for most use cases, the relevance of certain key capabilities may be significantly different, depending on the usage scenario. The importance of each key capability for eMBB, URLLC and mMTC is illustrated in Figure 8. This is done using an indicative scaling in three steps as “high”, “medium” and “low” [19].

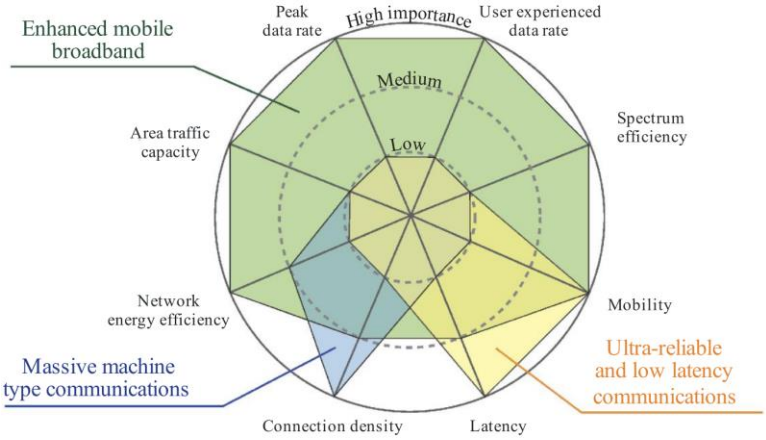

Figure 8: The importance of the KPIs in different usage scenarios

In the eMBB scenario, user experienced data rate, area traffic capacity, peak data rate, mobility, energy efficiency and spectrum efficiency all have high importance, but mobility and the user experienced data rate would not have equal importance simultaneously in all use cases. For example, in hotspots, a higher user experienced data rate, but a lower mobility, would be required than in wide area coverage case [19].

In some URLLC scenarios, low latency is of highest importance while high data rates could be less important, e.g. to enable safety critical applications like transportation [19].

In the mMTC scenario, high connection density is needed to support huge number of devices in the network that may transmit only occasionally, at low bit rate and with zero or very low mobility. A low-cost device with long operational lifetime is vital for this usage scenario [19].

3. 5G NEW RADIO (NR)

3.1 The new Radio Access Technology

A Radio Access Technology or (RAT) is the underlying physical connection method for a radio-based communication network. Many modern mobile phones support several RATs in one device such as Bluetooth, Wi-Fi, GSM, UMTS, LTE or 5G NR [17].

Approximately once every 10 years, wireless communications standards have marched forward, advancing through 2G, 3G, 4G and now moving into 5G [22].

5G NR is the global standard for a unified, more capable 5G wireless air interface. It will deliver significantly faster and more responsive mobile broadband experiences and extend mobile technology to connect and redefine a multitude of new industries [17].

It specifies new features that require the development of new technologies to meet the aggressive goals of the IMT-2020 vision by adding new operating bands with advanced ways to package and transmit signals. mmWave operating bands, wider modulation bandwidths, scalable numerologies and new initial access procedures introduce many changes to understand and implement in new infrastructure and mobile designs. Chipsets and devices will operate at higher frequencies. Devices and base stations will use new technologies to make connections and networks will evolve to handle more data, more users, and different levels of service. Initially, 5G NR will operate alongside 4G LTE to deliver enhanced services for users [22].

3.2 5G NR Specifications

3GPP RAN working groups define the 5G NR specifications. The workgroup outputs are public: all documents, meeting reports and published specifications are available on the 3GPP website.[2] The 5G NR specifications appear in the 38.xxx series documents. The 5G NR RAN study items and specifications define the functions, requirements, and interfaces of the networks. RAN study items are followed by work items that are followed by the release of specifications [22].

The radio interface between the user equipment (UE) and the network is comprised of layers 1, 2 and 3 of the communications stack, commonly known as the physical layer, the data link layer, and the network layer. The physical layer, defined in TS 38.200, represents the interface to the “real world” and includes the hardware and software to control this linkage. The physical layer provides a transport channel and specifies how information is transferred over the radio interface. Layers 2 and 3, defined in the TS 38.300 series (see Figure 9), work in conjunction with the physical layer. The data link layer, also known as the medium access control (MAC) layer, enables data transfer between the different networks. The MAC layer provides different logical channels to the radio link control (RLC) in the network layer. Layer 3, the radio resource control (RRC) layer, connects with the nodes in the network so that the UE can move seamlessly throughout the network [22].

[2] 3GPP Specifications Groups, www.3gpp.org/specifications-groups

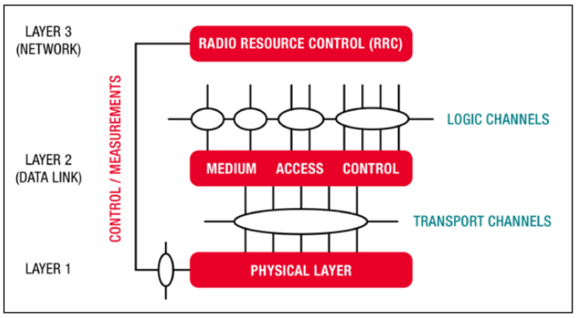

Figure 9: Radio interface protocol architecture (3GPP document TS 38.201)

The RAN working groups are responsible for developing the 5G NR specifications in certain areas, such as the 5G NR physical layer. TR represents a technical report while TS represents a technical specification. The 5G NR RAN working groups and technical specifications are [22]:

- **RAN1 (radio layer 1, TS 38.201–38.215)** is responsible for the physical layer (layer 1) of the UE and the data transport to the radio interface protocol architecture (layers 2 and 3). It includes specifications of the physical channel structures, mapping of the transport channels into physical channels, multiplexing, modulation, and channel coding, as well as the physical layer procedures, such as cell search, power control and beam management.
- **RAN2 (radio layers 2 and 3, TS 38.300–TS 38.331)** is responsible for the radio interface architecture and protocols. This includes interfaces between the 5G NR and the 5G core network. It covers the network interfaces, the physical layer, and connections to MAC, RLC and the packet data convergence protocol (PDCP). RAN2 is also responsible for the RRC protocol, the strategies of radio resource management (RRM) and the services provided by the physical layer to the upper layers.
- **RAN3 (radio network, TS 38.401–38.474)** is responsible for the overall architecture and the protocol specifications. TS 38.2xx and TS 38.3xx in RAN1 and RAN2 define the radio interface protocols, and RAN3 defines the next-generation interface protocols.
- **RAN4 (radio performance and protocol, TS 38.101–38.307)** is responsible for the radio frequency (RF) aspects of communications and the development of the minimum requirements for 5G NR transmission and reception, as well as the parameters for channel demodulation. RAN4 also provides test procedures for base station conformance and specifications for electromagnetic compatibility (EMC), radio link, cell selection/reselection and performance supporting RRM.
- **RAN5 (mobile terminal conformance tests, TS 38.508–38.533)** is responsible for the specifications of conformance testing at the radio interface for the UE, based on the specification defined in RAN4 for signalling and protocol test cases. RAN5 has the responsibility for RF and signalling subgroups, including RF conformance and inter-RAT procedures.

Consequently, the 5G NR specifications can be characterised in terms of the following:

3.2.1 Frequency Bands

5G NR is designed to support a wide range of frequency bands including support of mmWave frequency bands (see Table 2). Radio propagation in mmWaves is characterised by higher attenuation and is impacted more by blockage due to obstacles and foliage. To counteract these effects, NR uses beamforming to concentrate the RF energy in the direction of the UE. With beamforming, the user's mobility and time varying environment require the UE to constantly look for new beams and dynamically change beams [18].

Table 2: Definition of frequency ranges in 5G NR Release 15

Frequency Range Designation	Corresponding Frequency Range
FR1	450 MHz – 6000 MHz
FR2	24.25 GHz – 52.6 GHz

Previous generation networks primarily operated in licensed spectrum bands below 3 GHz, 5G will bring the next level of convergence with support for licensed, shared, and unlicensed spectrum. Moreover, 5G will expand spectrum usage to low-bands below 1 GHz, mid-bands between 1 GHz and 6 GHz, and high-bands above 24 GHz, loosely known as mmWave, which will open up vast amount of bandwidths for extreme data rates and capacity that were previously not usable for wide-area mobile communications [1].

3.5 GHz has the broadest support for 5G globally and has been identified in Europe as the primary 5G band to bring necessary capacity for 5G services. mmWave spectrum (in frequencies >24 GHz) is also identified for 5G for ultra-high capacity and innovative new services. 26 GHz band is a pioneer band identified by China and Europe for 5G in mmWave. The US, Korea and Japan have identified 28 GHz as their primary 5G mmWave band [16].

The 5G New Radio (NR) specification adds new operating bands in both sub-6-GHz and mmWave frequencies to extend the available spectrum. 5G NR identifies almost 10 GHz of new spectrum from sub-6 GHz to mmWave frequencies [24]:

- **FR1 (Frequency Range 1)**: 400 MHz to 6 GHz adds 1.5 GHz of new spectrum in frequency bands 3.3 – 4.2 GHz, 3.3 – 3.8 GHz and 4.4 – 5 GHz [24]
- **FR2 (Frequency Range 2)**: 24.25 to 52.6 GHz adds 8.25 GHz of new spectrum in frequency bands: 26.5 – 29.5 GHz, 24.25 – 27.5 GHz and 37 – 40 GHz [24].

In the FR1 (sub-6 GHz) frequency bands, massive MIMO (multiple-input, multiple output) technologies will be deployed to boost cell capacity and realise the throughput envisioned by eMBB use cases. Massive MIMO will use a much greater number of antenna elements on the base station directed to multiple UE simultaneously in the same frequency and time resource. Whereas in the FR2 (mmWave) frequencies, carrier aggregation is used to combine multiple component carriers. This additional spectrum is essential to enabling 5G's promise of extreme data rates of 20 Gbps in the downlink (DL) and 10 Gbps in the uplink (UL) [24].

While mmWave frequencies enable more bandwidth, they are also more susceptible to signal propagation issues that are not a problem at sub-6 GHz frequencies. Signal propagation issues such as increased path loss, delay spread, and blockage make it more difficult to establish and maintain a wireless communications link between a mobile device and base station. At mmWave frequencies, the placement of a user's hand on a mobile device or the orientation of their body can significantly degrade the radio link performance. Signals also experience more attenuation going through different types of materials, which can limit outdoor-to-indoor coverage [24].

To overcome these signal propagation issues, 5G radio systems will utilise multiple antennas on base stations and mobile devices to implement spatial diversity and beam steering techniques to reliably direct narrow beams in a specific direction [24].

These spectrum bands earmarked for deployment of 5G can also be sub-divided in three categories in terms of their coverage capacity: sub-1 GHz, 1-6 GHz and above 6 GHz (Figure 10) [2].

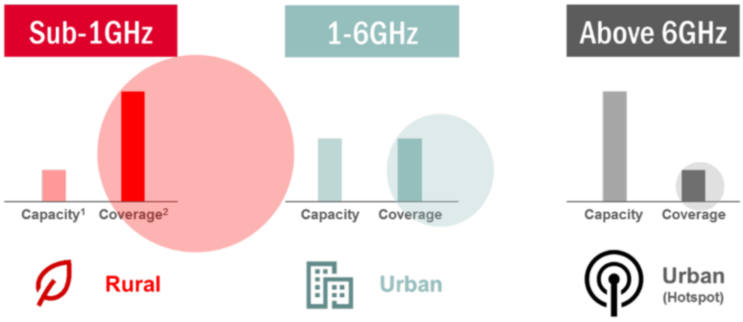

Figure 10: Capacity and coverage of spectrum categories

Sub-1 GHz bands are suitable to support IoT services and extend mobile broadband coverage from urban to suburban and rural areas. This is because the propagation properties of the signal at these frequencies enable 5G to create very large coverage areas and deep in-building penetration [2].

The 1-6 GHz bands offer a reasonable mixture of coverage and capacity for 5G services. There is a reasonable amount of existing mobile broadband spectrum identified with this range which could be used for initial 5G deployments [2].

Spectrum bands above 6 GHz provide significant capacity thanks to the very large bandwidth that can be allocated to mobile communications and thus enable eMBB applications. The downside of using high spectrum (mmWave) bands is the highly reduced coverage size of each cell and its susceptibility to blocking [2].

3.2.2 Flexible Waveforms

5G NR relies on variants of OFDM to meet the extreme requirements of 5G. In 5G NR, Cyclic Prefix Orthogonal Frequency Division Multiplex (CP-OFDM) is the modulation format (or waveform) in the downlink (DL) and uplink (UL) [25].

CP-OFDM use is well-known for DL transmissions but is new for UL transmissions in mobile. Having the same waveform in both UL and DL enables easier device-to-device communication in future releases. Discrete Fourier Transform spread OFDM (DFT-s-OFDM) is an optional waveform in the UL. DFT-s-OFDM uses a single transmission, which is helpful in limited-power scenarios [25].

Discussing each of them in more detail:

CP-OFDM (Downlink and Uplink): Because orthogonal frequency division multiplexing (OFDM) implementations lend themselves well to Time Division Duplex (TDD) operation and delay-sensitive applications, and because they have demonstrated successful commercial implementation by efficiently processing ever-larger bandwidth signals, CP-OFDM became the preferred choice for 5G NR. The strong benefits of CP-OFDM that make it a great fit for 5G implementation are [26]:

- **High Spectral Efficiency** — This important feature of OFDM access helps meet the extreme data rate needs, especially for backhaul links. Also, in future cases like vehicular communication in dense urban environments, high spectral efficiency will help address capacity constraints when many users broadcast periodically and asynchronously.

- **MIMO Compatibility** — Both base stations and mobile devices will take advantage of MIMO technology to implement spatial and frequency multiplexing with Single-User MIMO and Multiuser MIMO (MU-MIMO). MIMO deployments also overcome high propagation losses and extend coverage range with beamforming.
- **Phase Noise Resistance** — As the frequency of operation (and with it the oscillator phase noise) increases, an OFDM system can minimise inter-symbol interference due to phase noise by applying larger OFDM subcarrier spacing (SCS).
- **Transceiver Simplicity** — OFDM transceivers offer lower implementation complexity compared to other waveforms that designers considered for 5G deployments. Having worked with OFDM designs for several years, the wireless industry knows that their well-understood operation and wide commercial deployment can enable 5G devices with powerful OFDM baseband processing at lower prices.
- **Channel Time- and Frequency-Selectivity Resistance** — With the right selection of SCS and frequency of operation, an OFDM system can finish a transmission between devices in an interval shorter than the channel coherence time and enable high-mobility (high-speed) and high-data-rate scenarios while minimising the effects of time selectivity. Also, as Figure 11 shows, with channel estimation and equalisation techniques, OFDM waveforms prove to be resilient against frequency-selective channels.

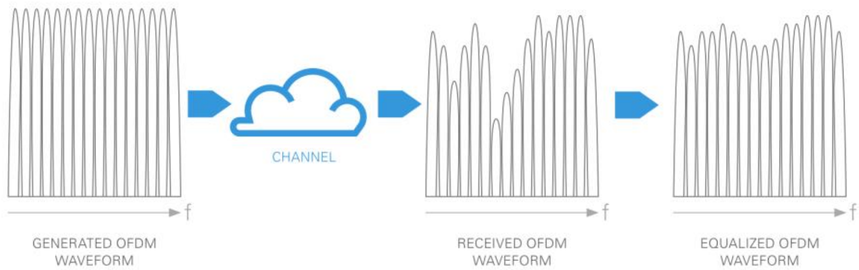

Figure 11: An OFDM Waveform’s frequency-selectivity resistance

- **Timing Error and Inter-symbol Interference Resistance** — Because of the CP, a receiver can tolerate synchronisation errors better and prevent the previous OFDM symbol from spreading into the currently received OFDM symbol. Figure 12 shows two subsequent OFDM symbols, each with a dedicated CP. The CP at the beginning of each OFDM symbol contains a copy of the end of the OFDM symbol. When the receiver demodulates the signal, it operates on the symbol after the CP (FFT window). This mechanism prevents inter-symbol interference between adjacent OFDM symbols.

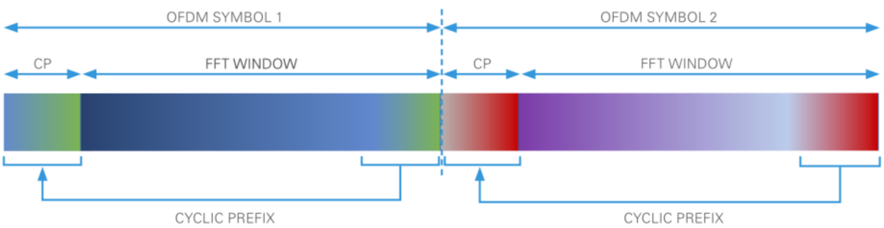

Figure 12: A cyclic prefix separates OFDM symbols

With CP-OFDM, UE support QPSK, 16-QAM, 64-QAM and 256-QAM modulation schemes [26].

DFT-s-OFDM (Higher Efficiency Uplink): One of the main drawbacks of OFDM waveforms is their high PAPR. As a result, RF output power amplifiers on transmitters lose efficiency and cannot minimise high-order, nonlinear effects well. For UE such as smartphones, preserving battery life and being energy efficient is important. The RF power amplifier that transmits the signal to the base station consumes the most power within the mobile device, so system designers required a type of waveform that promotes high-efficiency amplifier operation while meeting the spectral demands of 5G applications. Although single-carrier waveforms have very low PAPR and more efficient power amplifier operation, they do not offer high spectral efficiency and dynamic spectrum utilisation, their compatibility with MIMO systems is lower, and they are susceptible to frequency-selective channels [26].

For uplink, NR allows UE to use CP-OFDM or the DFT-s-OFDM hybrid format waveform. In DFT-s-OFDM, the transmitter modulates all subcarriers with the same data. The right side of Figure 13 shows that the first group of subcarriers (all red) takes the same amount of bandwidth as the OFDM symbol on the left. The DFT-S-OFDM modulator maps the same data to all subcarriers but for a shorter duration. It then maps the next data symbol (green) to all subcarriers for another short interval. By the end of the equivalent OFDM symbol time, the transmitter sends the same amount of data as it sends with an OFDM waveform by mapping the data symbols to all subcarriers simultaneously but with shorter transmission intervals. This DFT-s-OFDM waveform combines a lower PAPR with the multipath interference resilience and flexible subcarrier frequency allocation that OFDM provides [26].

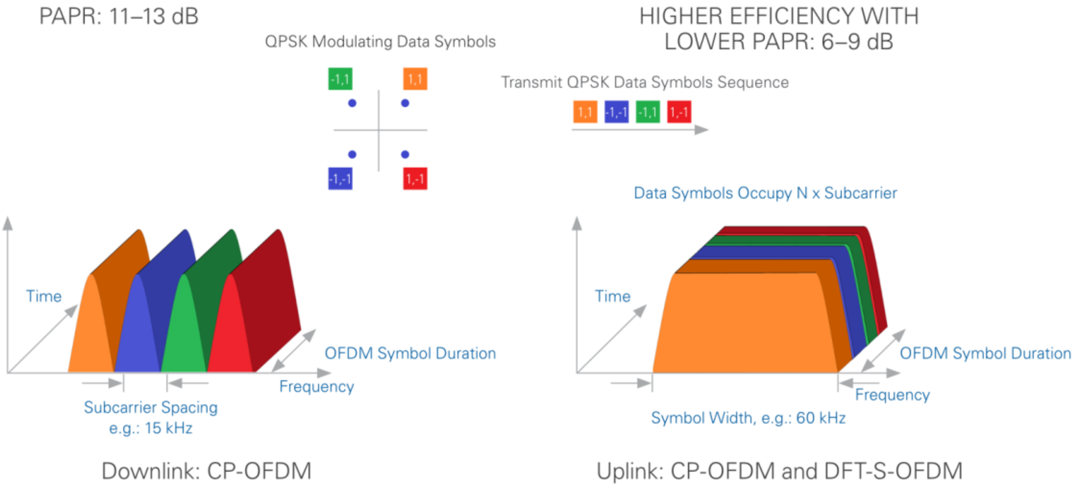

Figure 13: OFDM vs. DFT-S-OFDM

With DFT-S-OFDM, UE support Pi/2-BPSK, 16-QAM, 64-QAM and 256-QAM modulation schemes [26].

3.2.3 Scalable Numerology

Numerology refers to the physical layer subcarrier spacing (SCS) and symbol length. [26]. 4G LTE supports carrier bandwidths up to 20 MHz with a fixed OFDM numerology — 15 kHz spacing between OFDM subcarriers. 5G NR, on the other hand, uses a scalable OFDM numerology to support diverse frequency bands and deployment models [1]. In essence, the subcarrier spacings are no longer fixed to 15 kHz [25].

5G NR numerology scales according to the following formula [27],[28]:

$$SCS = 15 \cdot 2^{\mu}\, kHz, \mu \in \{0, 1, 2, 3, 4\}$$

The standard specifies that the smallest allocatable frequency unit consists of 12 subcarriers, defined as a physical resource block (PRB). Consequently, the smaller the SCS, the narrower the PRB, as shown in Figure 14 [26].

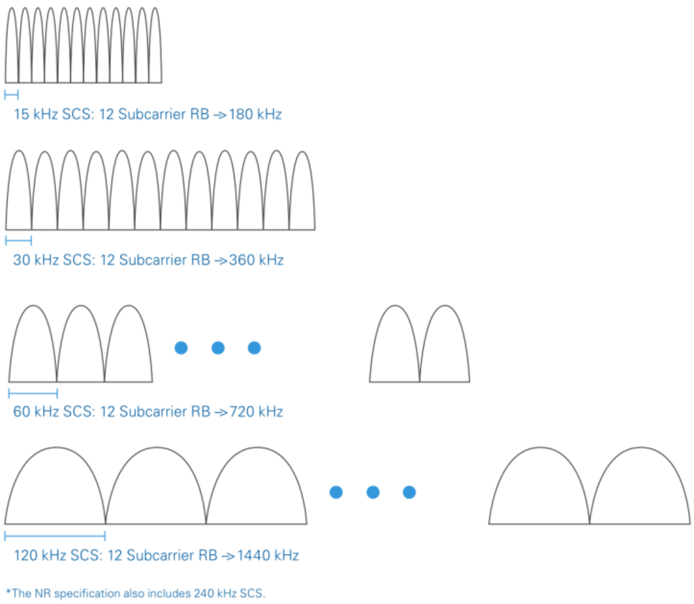

Figure 14: Flexible subcarrier spacing (SCS)

Lower frequency bands use 15, 30, and 60 kHz SCS, and higher frequency bands use 60, 120, and 240 kHz SCS. Scalable numerology enables scalable slot duration to optimise for different service levels in throughput, latency, or reliability [25].

Larger SCS at higher frequencies makes the waveform more robust against integrated phase noise, which is an issue in mmWave designs [25]. As the carrier frequency increases, so does the system phase noise. For instance, in the carrier phase noise plot of Figure 15, the difference in phase noise between a carrier at 1 GHz and 28 GHz is about 20 dB. This phase noise increase makes it difficult for a receiver to demodulate OFDM waveforms with their characteristic narrow, fixed SCS and symbol duration at mmWave frequencies [26].

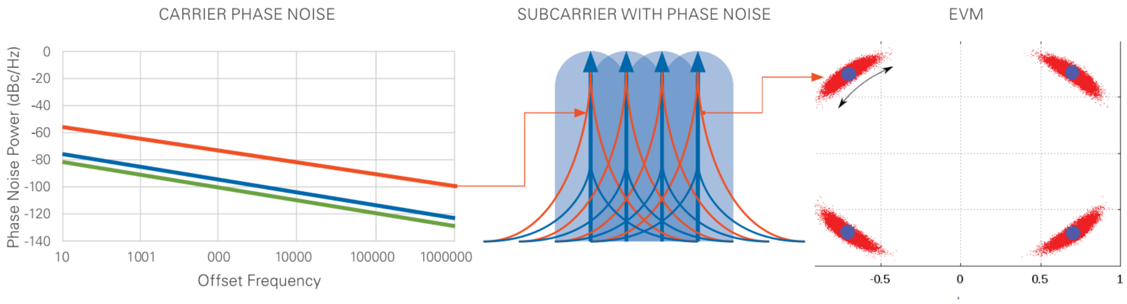

Figure 15: Effect of phase noise on error vector magnitude (EVM)

Likewise, a scalable numerology tackles the issue of Doppler shift. Doppler shift increases with carrier frequency, as shown in Figure 16. For example, a UE traveling at a speed of 60 km/h using a carrier frequency of 28 GHz sees a Doppler shift of close to 1500 Hz, or 10 percent of a 15 kHz SCS. Because the channel coherence time, or the time when the system can assume that the radio channel remains constant, is approximately inversely proportional to the Doppler shift, it decreases as mobility increases. Therefore, at higher carrier frequencies and higher speeds, the system has less time to measure the channel and finish a single slot transmission, hence the need for scalable numerology in 5G NR [26].

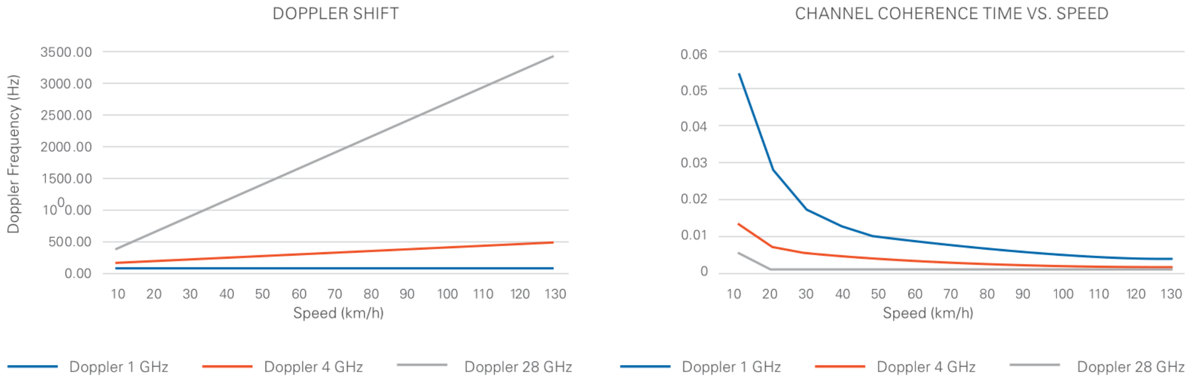

Figure 16: Doppler shift and channel coherence

Hence, phase noise and Doppler shift define the requirements for SCS to meet specific error vector magnitude (EVM) criteria. That means using narrow SCS causes higher EVM because of phase noise. Also, when SCS is small, the system performance can suffer because of Doppler shift in high-mobility scenarios. On the other hand, selecting a large SCS results in excessive channel bandwidth. Furthermore, given that SCS is inversely proportional to the OFDM symbol duration, the OFDM symbol and CP length shortens as SCS increases and makes the system more susceptible to delay spread. Therefore, SCS should be as small as possible while providing enough performance in the presence of phase noise and Doppler shift for a desired channel bandwidth [26].

Also, SCS has an inversely proportional relationship with OFDM symbol duration. 5G NR slots have 14 OFDM symbols per slot. The duration of the slots scales down as SCS increases which implies that slot configurations scale with SCS. Figure 17 shows the conventional 5G NR slot configurations [26].

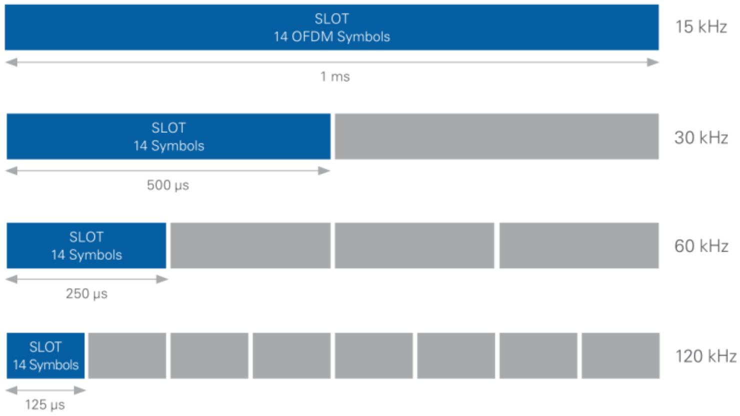

Figure 17: TDD slot-based scheduling

The frame structure (Figure 18) numbers the slots and groups them into subframes of 1 ms duration. Ten 1 ms subframes form a complete 5G NR frame. The number of slots within a frame also varies with the choice of numerology, for instance:

- Using 15 kHz of SCS results in a single 1 ms slot within the subframe, amounting to 10 slots per frame
- Whereas 30 kHz of SCS results in a subframe with two 500 µs slots within the subframe, amounting to 20 slots per frame [26].

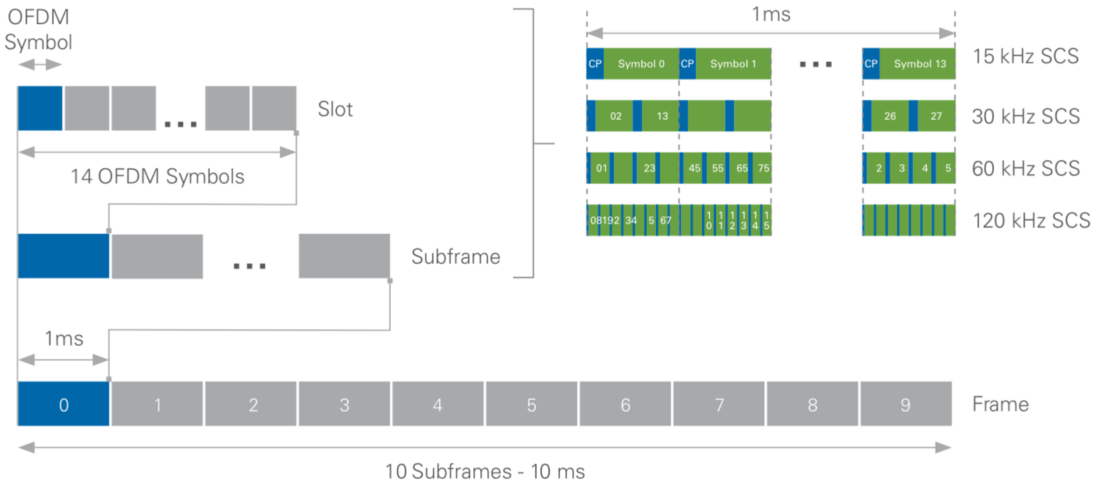

Figure 18: 5G NR frame structure

The time and frequency resource structure defines the 5G NR resource grid in Figure 19. Depending on SCS, the resource grid changes as the number of available subcarriers and OFDM symbols change. That is, for each numerology and carrier, 5G NR specifies a resource grid with a width given by the maximum number of resource blocks per SCS multiplied by the number of subcarriers per resource block, and a length given by the number of OFDM symbols per subframe [26].

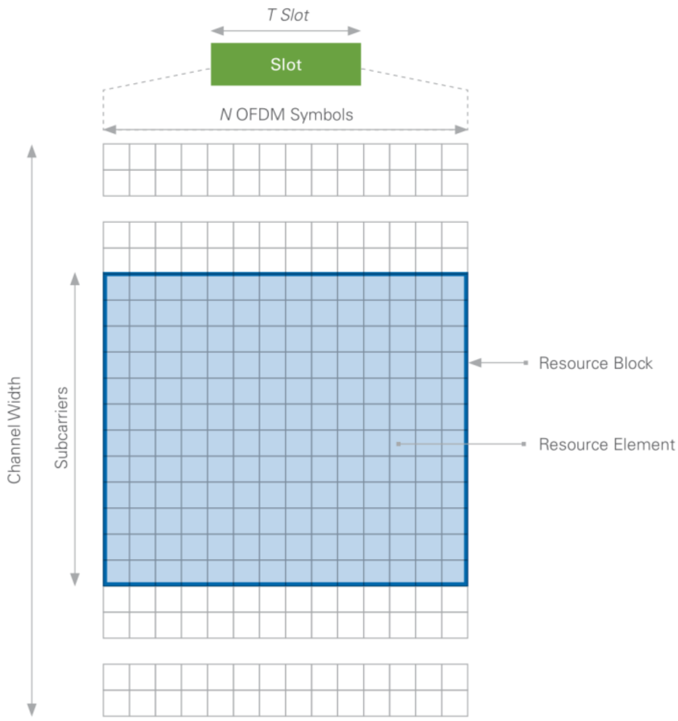

Figure 19: 5G NR resource grid

To enable agile and efficient use of TDD resources, NR also implements a flexible slot structure. Its subframe structure allows for dynamic assignments of the OFDM symbol link direction and control within the same subframe. By using this TDD mechanism, the network dynamically balances UL and DL traffic requirements and includes control and acknowledgment all in the same subframe. The slot format indicator (SFI) denotes whether a given OFDM symbol in a slot is used for uplink or downlink, or if it is flexible [25]. The system can allocate a slot as all DL, all UL, or a combination of DL and UL to service asymmetric traffic, as Figure 20 illustrates. DL control takes place at the beginning of the slot and UL control at the end. The system can either configure the mixed DL/UL slot statically, as in an 4G LTE DL/UL TDD configuration, or change the allocation of the DL/UL mix dynamically for better efficiency and scheduling based on traffic needs. To accomplish this, the 5G NR standard includes the SFI, a field that informs a UE whether an OFDM symbol contains DL, UL, or flexible (either DL or UL) slots. The SFI indicates the link direction of slots by indexing a row of the UE's preconfigured table of possible link direction assignments [26].

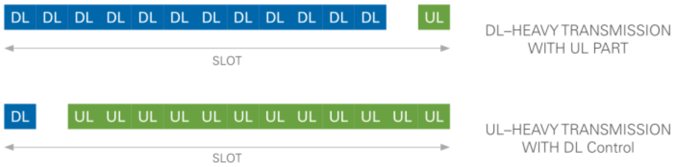

Figure 20: Flexible slot structure for dynamic management of TDD resources

Furthermore, when the system needs to work with large payloads that do not require the most urgent attention, the standard allows for slot aggregation. In the case of eMBB, for instance, having aggregated slots and longer transmission times meets application requirements while reducing TDD switching and signalling overhead [26].

The NR standard also uses mini-slots to support bursty, asynchronous transmissions with variable start positions and durations shorter than the typical, 14-symbol slot. A mini-slot represents the smallest possible scheduling unit, and it can last for 7, 4, or 2 OFDM symbols. Mini-slots are especially important for enabling low-latency transmissions [26]. URLLC, one of three primary 5G usage scenarios, is achieved partially through mini-slots. A mini-slot is shorter in duration than a standard slot and located anywhere within the slot. A mini-slot is 2, 4, or 7 OFDM symbols long. Mini-slots provide low-latency payloads with an immediate start time without the need to wait for the start of a slot boundary [25].

3.2.4 Bandwidth Parts

In 4G LTE, carriers are narrower in bandwidth — up to 20 MHz maximum. When aggregated, they create a wider channel bandwidth of up to 100 MHz. In 5G NR, the maximum carrier bandwidth is up to 100 MHz in FR1 and up to 400 MHz in FR2 [25]. Hence, in 5G applications, a wide range of devices and equipment will have to operate successfully across many different bands with varied spectrum availability. For instance, a UE with limited RF bandwidth might need to operate alongside a more powerful device that can fill a whole channel using carrier aggregation and a third device that can cover the whole channel with a single RF chain (Figure 21) [26],[29].

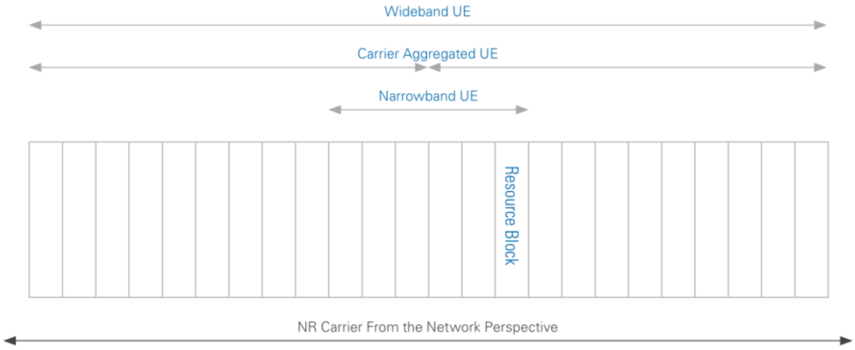

Figure 21: Parts for efficient spectrum management

Even though wide bandwidth operation has a direct effect on the data rates that users can experience, it comes at a cost. When a UE does not need high data rates, wide bandwidth leads to inefficient use of RF and baseband processing resources [26].

To tackle this issue, the 3GPP developed the concept of bandwidth parts (BWPs) for 5G NR [26]. A BWP refers to the subdivision of a carrier for different purposes. Each bandwidth part has its own numerology and is signalled independently [25]. The network can configure certain UE with one wideband carrier and other UE with a set of intra-band contiguous component carriers using carrier aggregation. This allows for a greater diversity of devices with varying capabilities to share the same wideband carrier. This kind of flexible network operation that adjusts to UE's differing RF capability does not exist in 4G LTE [26].

A BWP is composed of a group of contiguous PRBs. Each BWP has an associated SCS and CP (numerology). As a result, the system can use the BWP to reconfigure UE with a certain numerology. UE starts out with a default active BWP during the initial access until the system configures the UE's BWPs explicitly during or after connection establishment. Figure 22 shows that the network is allocating two BWPs (BWP 1 and 2) to one UE device while reserving a third, full-channel, overlapping BWP (BWP 3) for possible use by another higher bandwidth UE device or application [26].

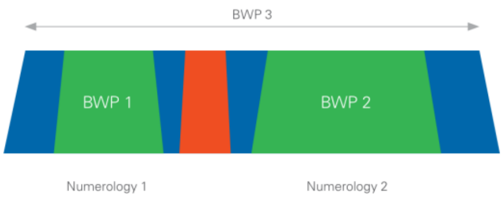

Figure 22: Bandwidth Parts

To summarise, 5G NR will have the flexibility to serve many different use cases effectively by using BWPs (Figure 23) [26].

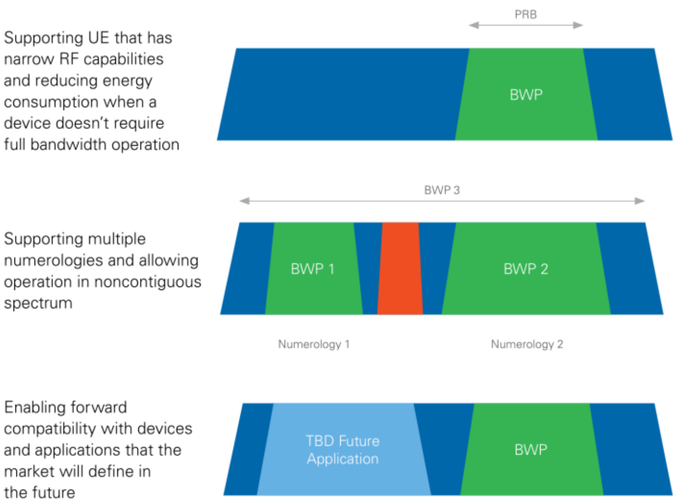

Figure 23: 5G NR can serve many use cases with BWPs

3.2.5 Reference Signals

To improve protocol efficiency, and to keep transmissions contained within a slot or beam without having to depend on other slots and beams, 5G NR introduces the following four reference signals. Unlike 4G LTE standard which constantly exchanges reference signals to manage a link, a 5G NR transmitter sends these reference signals only when necessary [26],[30].

Demodulation Reference Signal (DMRS): The DMRS is UE-specific, and a system uses this signal to estimate the radio channel. The system can beamform the DMRS, keep it within a scheduled resource, and transmit it only when necessary in either DL or UL. Moreover, multiple orthogonal DMRSs can be allocated to support MIMO transmission. The network presents users with DMRS information early on for the initial decoding requirement that low-latency applications need, but it only occasionally presents this information for low-speed scenarios in which the channel shows little change. Alternatively, tracking fast changes in high-mobility scenarios might increase the rate of DMRS transmission [26].

Phase Tracking Reference Signal (PTRS): As mentioned earlier, the phase noise of transmitters increases as the frequency of operation increases. The PTRS plays an important role, especially at mmWave frequencies, to minimise the effect of the oscillator phase noise on system performance. A main problem that phase noise introduces into an OFDM signal is common phase error (CPE), which appears as a common phase rotation of all subcarriers. The 5G NR system typically maps the PTRS information to a few subcarriers per symbol because the phase rotation affects all subcarriers within an OFDM symbol equally but shows low correlation from symbol to symbol. The system configures the PTRS depending on the quality of the oscillators, carrier frequency, SCS, and modulation and coding schemes that the transmission uses [26].

Sounding Reference Signal (SRS): The SRS is a UL-only signal that is transmitted by the UE to help the system obtain the CSI for each user. This information describes how the 5G NR signal propagates from the transmitter to the receiver and represents the combined effect of scattering, fading, and power decay with distance, for example. The system uses the SRS for resource scheduling, link adaptation, Massive MIMO, and beam management [26].

Channel State Information Reference Signal (CSI-RS): The CSI-RS is a DL-only signal that the UE receives and uses to estimate the channel so it can report channel quality information back to the base station (gNB). During MIMO operations, 5G NR uses different antenna approaches based on the carrier frequency. At lower frequencies, the system uses a modest number of active antennas for MU-MIMO and adds Frequency Division Duplex (FDD) operations. In this case, the UE needs the CSI-RS to calculate the CSI and report it back in the UL direction [26]. CSI is essential for 5G NR beamforming reliability. It uses channel estimation to intelligently change the precoding and adapt the beam to a specific user. The better and more precise the CSI is, the better the link adaptation [25].

3.2.6 Channel Coding

Channel coding, also referred to as error correction coding, plays a central role in determining the performance and implementation complexity of the physical layer. In 5G NR, Quasi Cyclic-Low Density Parity Check (QC-LDPC) codes are used for data channels as they provide latency and implementation advantages over other channel coding schemes at the same performance. Polar codes are used for encoding control information [18].

These specifications significantly increase the complexity of the 5G waveforms and introduce new challenges for device and component designers. Measurements become more critical, as does the need to validate protocols for the many different test cases and verify radio-frequency performance to deliver the expected quality of service. Massive MIMO and beam steering introduce challenges in beam management. The use of mmWave frequencies poses challenges in signal quality, and the requirement of over-the-air (OTA) testing makes validation even more difficult [25]. Researchers and engineers are, therefore, faced with the challenge of creating, distributing, and generating standard-compliant uplink and downlink signals that have larger bandwidths, more configurations, and more options, than ever before [31].

3.3 Comparison of 5G NR and 4G LTE

To conclude this chapter, the following fundamental features of 5G NR differentiate it from current 4G LTE implementations [26]:

Better Spectrum Utilisation: Wideband 5G carriers are designed to occupy up to 98 percent of the channel, avoiding large guard bands between carriers. This helps reduce channel overhead and allows for faster load balancing than 4G LTE aggregated carriers can implement. For example, Figure 24 compares five 20 MHz aggregated 4G LTE carriers versus a proposed single 98 MHz 5G NR carrier [26].

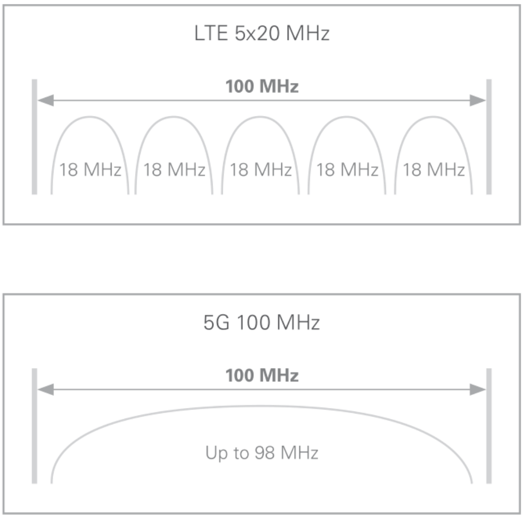

Figure 24: Improved channel utilisation with wideband 5G carriers

Flexible Numerology and Frame Structure: 4G LTE uses fixed 15 kHz SCS with a maximum of 1200 subcarriers in a 20 MHz channel. In contrast, 5G NR allows for greater spectrum utilisation with channels of various sizes, variable SCS and slot length, and a maximum of 3300 subcarriers per channel [26].

Enhanced Efficiency with Leaner Signalling: 5G has no cell-specific reference signals and synchronises and broadcasts every 20 ms, unlike 4G LTE, which transmits cell-specific reference signals four times per millisecond, synchronises every 5 ms, and broadcasts every 10 ms. This enables greater base station power savings (Figure 25) [26].

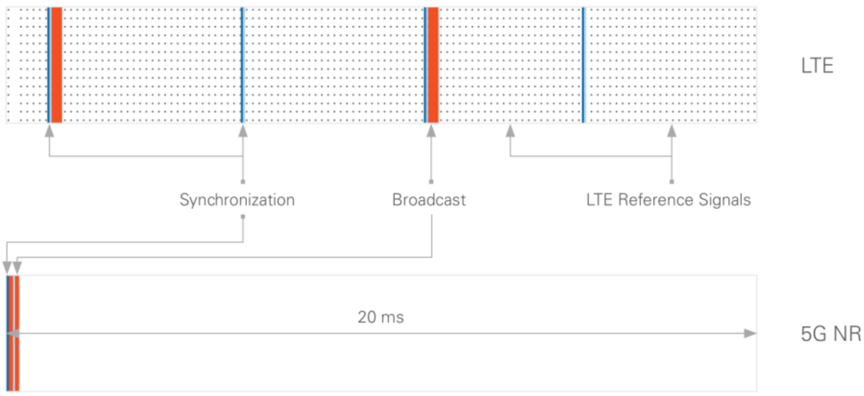

Figure 25: Signalling efficiency in 4G LTE vs. 5G NR

Manage TDD Resources Dynamically: 4G LTE has a fixed, static TDD structure that allocates slots to either DL, UL, or synchronisation and control signals. That is, within a radio frame, its TDD structure switches multiple times between DL and UL transmission and vice versa. On the other hand, within a slot, 5G NR can change dynamically between DL and UL to handle traffic demands in either direction (Figure 26) [26].

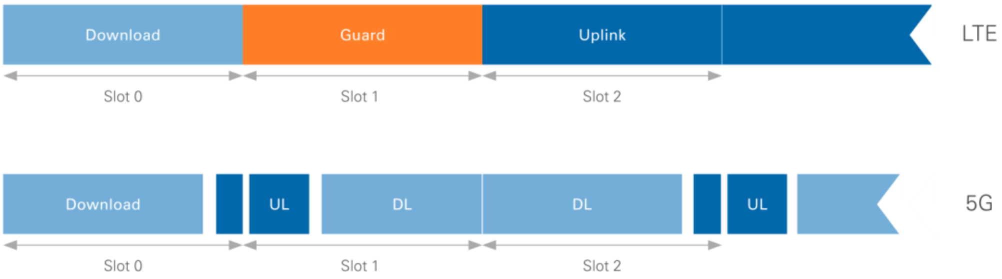

Figure 26: 5G NR Manages TDD Resources More Dynamically

Operation at mmWave Frequencies With Wider Channels: 4G LTE networks are limited to operating at a maximum frequency of around 3800 MHz whereas 5G NR networks will take advantage of both existing frequency bands and wide channels in newly licensed spectrum [26]. They would operate in sub-6 GHz bands and mmWave bands with appropriate handling of multipath delay spread, channel coherence time, and phase noise. Furthermore, 5G NR will support existing and new services with even higher data rates and address different latency and mobility requirements by changing the transmission turnaround time using variable SCS and by allocating wide channels. It will use the latest developments in Massive MIMO and beamforming technology to maximise spectral efficiency and guarantee better service for a higher number of users. In addition, the BWP concept in 5G NR will lead to more energy-efficient UE operation and superior spectrum management [26].

4. ENABLING TECHNOLOGIES

4.1 Introduction

To meet the needs of a networked society, next-generation 5G networks promise revolutionary improvements in network capacity, data rates and latency, with greatly increased network flexibility and efficiency. Likewise, MNOs will expect lower operational and infrastructure costs than today [32],[33]. Realising these challenging goals will require extensive and multi-faceted changes in all aspects of the cellular ecosystem, from chipsets and devices to base stations and small cells, from fronthaul and backhaul to network management and data centre performance [34].

5G support a multitude of new applications with a wide variety of requirements, including higher peak and user data rates, reduced latency, enhanced indoor coverage, increased number of devices, and so on. The expected traffic growth can be satisfied by the combined use of more spectrum, higher spectral efficiency, and densification of cells [35].

The 5G technology roadmap contains two parts: a new air interface (5G NR) and a 4G evolution air interface [36]. In the 5G NR technical framework, the key technology areas include massive MIMO, ultra-dense network (UDN), new multiple access and full spectrum access. As one of the most important enabling technologies for seamless wide area coverage, massive MIMO efficiently utilises spatial dimension resources to dramatically increase system spectral efficiency and enhance the user experience. UDN significantly reduces cell interference through inter-microcell collaboration and expands network capacity in local hot spots. Novel multiple access technologies increase equipment connectivity and reduce signalling overhead, as well as UE power consumption through grant-free mechanisms. Full spectrum access, which supports a hybrid network combining low and high frequency bands, can simultaneously meet the requirements of high data rates and large capacity [37].

The following sections will provide more details on the key technologies that will enable 5G NR to achieve new levels of performance and efficiency [1].

4.2 mmWave Spectrum

Despite industrial research efforts to deploy the most efficient technologies possible, the wireless industry always eventually faces overwhelming capacity demands for its currently deployed technologies, due to continued advancements and discoveries in computing and communications, and the emergence of new customer devices and use cases. In fact, the life cycle of every new generation of cellular technology is generally a decade or less due to the natural evolution of computer and communications technology. This trend has occurred for 4G LTE leading to the need to implement new technologies and architectures to properly serve the continuing demands of carriers and customers. [3]. Current 4G networks face a multitude of challenges and soaring demand for high-resolution multimedia applications brings these networks ever closer to their practical limits [38].

These challenges led to studies that proposed mmWave frequencies as a means of augmenting the saturated 700 MHz to 2.6 GHz radio spectrum bands for wireless communications [39]. The combination of cost-effective CMOS technology that operate well into the mmWave frequency bands, and high-gain, steerable antennas at the mobile and base station, strengthens the viability of mmWave wireless communications [40],[41]. Further, mmWave carrier frequencies allow for larger bandwidth allocations, which translate directly to higher data transfer rates. They would allow MNOs to significantly expand the channel bandwidths far beyond the 20 MHz channels used by

4G [42]. By increasing the RF channel bandwidth for mobile radio channels, the data capacity is greatly increased, while the latency for digital traffic is significantly decreased, thus supporting much better internet-based access and applications that require minimal latency. mmWave frequencies, due to their much smaller wavelengths, will exploit polarisation and new spatial processing techniques, such as massive MIMO and adaptive beamforming [43]. Given this substantial jump in bandwidth and new capabilities offered by mmWaves, the base station-to-device links, as well as backhaul links between base stations, will be able to handle much greater capacity than 4G networks in highly populated areas. Also, as MNOs continue to reduce cell coverage areas to exploit spatial reuse, and implement new cooperative architectures such as cooperative MIMO, relays, and interference mitigation between base stations, the cost per base station will drop as they become more plentiful and more densely distributed in urban areas, making wireless backhaul essential for flexibility, quick deployment, and reduced ongoing operating costs. Finally, as opposed to the fragmented spectrum employed by many MNOs today, where the coverage distances of cell sites vary widely over three octaves of frequency between 700 MHz and 2.6 GHz, the mmWave spectrum will have spectral allocations that are much closer together, making the propagation characteristics of different mmWave bands much more comparable and homogenous [3].

Moreover, fully realising the 5G vision will require much additional spectrum. Although additional spectrum below 6 GHz has been identified and, in some countries, already allocated for cellular communications, much larger contiguous spectrum is available in the centimetre and mmWave bands above 24 GHz. Figure 27 shows some of the candidate mmWave bands in various regions of the world for 5G NR [34],[44].

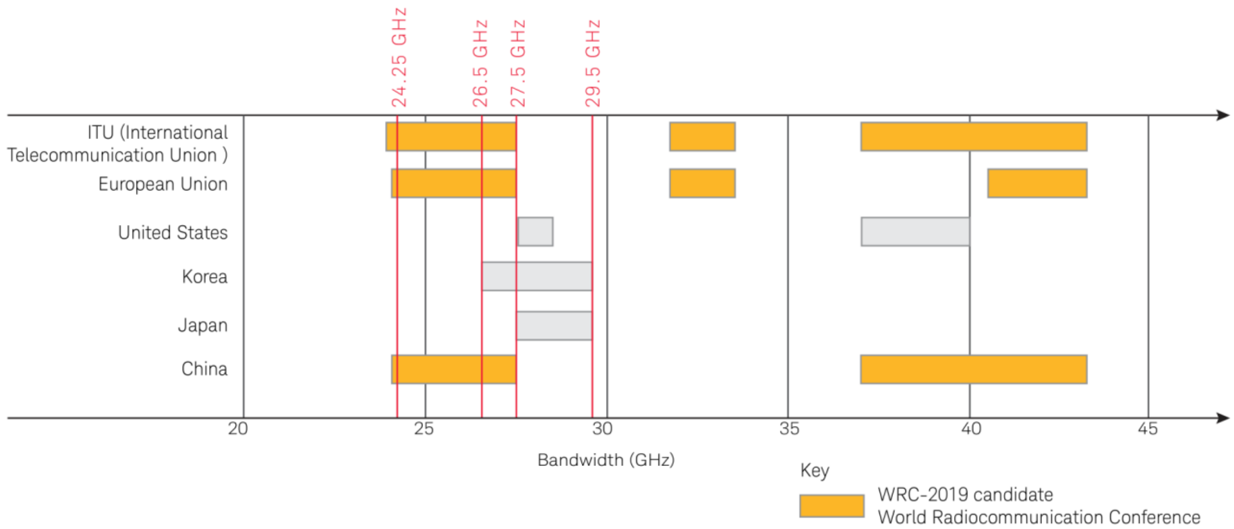

Figure 27: Candidate bands for 5G NR deployment above 6 GHz

5G networks are envisioned to ease the burden on the current infrastructure by offering significantly higher data rates through these increased channel bandwidths. mmWave bands will become a suitable alternative, given the shortage of available frequencies traditionally used for mobile communications. The large bandwidth available at these frequencies helps to offer data rates that satisfy 5G demands [38].

There is a huge amount of spectrum at mmWave frequencies ranging from 3 to 300 GHz. Many bands therein seem promising, including the local multipoint distribution service at 28–30 GHz, the license-free band at 60 GHz, and the E-band at 71–76 GHz, 81–86 GHz, and 92–95 GHz. [45].

By opening new mmWave opportunities for mobile broadband, 5G NR will deliver extreme data speeds and capacity that will reshape the mobile experience. However, mobilising mmWave bands comes with its own set of challenges. Transmissions in these higher bands suffer from significantly higher path loss as well as susceptibility to blockage, while meeting the power and form-factor requirements of mobile devices has also proven to be challenging. Thus, traditional mmWave implementations have been limited to mostly stationary applications such as shorter-range wireless docking, enabled by technologies like 802.11ad that operates in the 60 GHz band [1].

Because the mobile environment at mmWave bands is far more complex than at currently used frequencies, an updated network infrastructure and new hardware concepts are required to tackle the higher propagation losses [38]. Thankfully with the recent advancements in signal processing and antenna technologies, the idea of mobilising these bands is no longer out of reach [1].

Antenna arrays are a key feature in mmWave systems [45]. By employing a large number of antenna elements in both the base station and the device, along with intelligent beamforming and beam tracking algorithms, 5G mmWave can provide increased coverage, reduced interference, and a continuous connectivity experience even for non-line-of-sight (NLoS) communications and device mobility [1]. Using large arrays will keep the antenna aperture constant, eliminating the frequency dependence of path loss relative to omni-directional antennas (when used at one side of the link) and providing a net array gain to counter the larger thermal noise bandwidth (when used at both sides of the link). Adaptive arrays with narrow beams also lessen the impact of interference, meaning that mmWave systems could more often operate in noise-limited rather than interference-limited conditions [45].

4.2.1 Challenges at mmWave

The main reason that mmWave spectrum has not been exploited, until now, is that it had been deemed unsuitable for mobile communications because of rather hostile propagation qualities, including strong path loss, atmospheric and rain absorption, low diffraction around obstacles and penetration through objects, and, further, because of strong phase noise and exorbitant equipment costs. The prevailing view had therefore been that such frequencies, and especially the large unlicensed band around 60 GHz [46], were suitable mainly for very-short-range transmission [47]–[49]. Hence, the emphasis had been on WiFi (with the WiGig standard in the 60 GHz band) and on fixed-wireless applications in the 28, 38, 71–76 and 81–86 GHz [8].

Therefore, the main challenges of using mmWave bands are discussed in detail below.

Propagation Issues at mmWave: The main issues under investigation with respect to mmWave propagation for 5G cellular communication are path loss, blocking, and atmospheric and rain absorption. These propagation losses for mmWave frequencies are surmountable but require large antenna arrays to steer the beam energy and collect it coherently [8]. For a better understanding of these issues, we consider them individually and understand their impacts.

- **Path loss modelling** - One of the challenges of mobile communications in mmWave bands for outdoor access will be to overcome the expected difficulties in propagation conditions. Understanding the propagation conditions will be critical to designing an appropriate air interface and determining the type of hardware (particularly the array size) needed for reliable communications [50]. Ignoring atmospheric effects, the received power for a transmitter and a receiver communicating via free space is calculated using the Friis transmission equation [34],[51]:

$$\frac{P_r}{P_t} = \left(\frac{c}{4\pi Rf}\right)^2 G_t G_r$$

where P_r is the received power, P_t is the transmitted power, R is the distance between the transmitter and receiver, f is the frequency, c is the speed of light, G_t is the gain of the transmitter antenna, and G_r is the gain of the receiver antenna [34].

P_r is thus inversely proportional to the square of the frequency, f. For example, moving from 3 to 30 GHz using the same antenna gain adds 20 dB of path loss that, without mitigation, would severely impact network performance. In addition to the free space path loss, over-the-air communications at mmWave also need to consider the effects of atmospheric absorption, humidity, blocking, precipitation, and other factors which can increase the effective path loss between transmitter and receiver. For indoor channels, path loss is less significant but other effects like blocking need to be mitigated [34].

Hence, relative to traditional cellular bands, the most obvious problem of mmWave bands (above 6 GHz) is their higher path loss. Examining only free-space path loss, the expected loss (dB) will be [50],[52]:

$$L_p = 92.4 + 20 \log f + 20 \log d$$

where f is the frequency (GHz) and d is the distance (km) between transmitter and receiver. For instance, an additional 22.9 dB and 30.9 dB of losses are expected in the ranges from 2 GHz to 28 GHz and 70 GHz, which can be compensated by deploying larger antenna array sizes with higher antenna gains and MIMO technologies [50].

The mechanism typically employed to overcome additional path loss at mmWave frequencies is to increase the gain and therefore the directivity of the antennas. Base stations in a 5G network are expected to have antennas with directivities greater than 20 dBi while 5G UE will have directional antennas incorporated them, though with more moderate directivity [34].

The 3GPP co-existence research assumes gNB directivity of 30 dBi and UE of 25 dBi although practical UE directivity may be much lower except for fixed wireless access applications. In addition to countering the path loss of the mmWave channel, directivity also supports spatial reuse of the channel allowing the gNB to support multiple users in one cell with spatially separate beams. A further benefit of directivity is reduced interference to neighbour cells in the network which generally improves spectral efficiency. Consequently, 5G networks utilising mmWave frequencies will require both base stations and UE to steer relatively narrow beams toward each other (and via reflective surfaces) to optimise the link budget and provide effective communication. Developing protocols that enable base stations and UE to find and track each other and perform inter-cell handovers represents a key technical challenge of mmWave communications for 5G [34],[53].

- **Atmospheric and other losses** - Environmental effects such as gaseous (oxygen and water vapour) absorption, rain loss and foliage loss also pose major challenges. Path losses due to snow and fog are usually minor but may be important in rare cases [50].

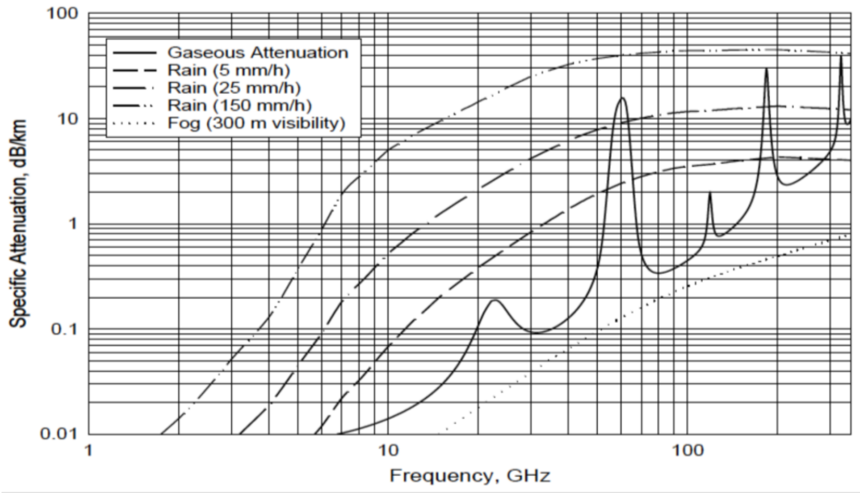

Figure 28: Atmospheric attenuation vs. frequency

Figure 28 shows the specific gas, rain, and fog attenuation (dB/km) as a function of frequency. As seen in the Figure, these effects exhibit a high degree of frequency dependent variation. Rain loss is more significant compared to other atmospheric losses, except close to the peaks in gaseous attenuation, but is relatively infrequent [50].

Atmospheric gas loss is roughly 0.10 dB/km and 0.15 dB/km at 28 GHz and 38 GHz, respectively, and about 0.40 dB/km at frequencies between 70 and 90 GHz. However, the gas attenuation could be of little concern since urban microcells will be designed for inter-site distances within 200 m for both backhaul and cellular. Thus, mmWave frequencies can provide access to spectrum with sufficiently low atmospheric attenuation in the case of small cells [50].

Usually, rain effects are analysed in terms of system outages which are typically required to be in the order of 0.01% or even 0.001%, a guarantee of 99.999% system availability, against expected mm/hr rainfall in a geographical region [54]. The effects are not insurmountable if cell radii are below 200 m. A 30 GHz carrier will see less than 1 dB loss over 200 m during heavy rainfall (25 mm/h), while a 60 GHz carrier would see less than 2 dB. However, tropical downpours, hurricanes, and the like where reliable mobile communications links are vital, must be considered [39],[50].

Even the frequencies, represented by the peaks in Figure 28, that are more susceptible to molecular oxygen and water vapour losses than other bands could still be used for small cell applications, where the atmospheric attenuation is not a limitation. In these frequency ranges, interfering signals are also limited by the inherent atmospheric loss, allowing tighter reuse of carrier frequencies [50]. Moreover, since the size of raindrops is similar to the wavelength around mmWave frequency bands, it causes the scattering effect of the signal propagation [50].

The total specific attenuation (dB/km) caused by atmospheric gasses and rain can be calculated from the following equation [50]:

$$\gamma = \gamma_G + \gamma_R = \gamma_0 + \gamma_w + k \cdot R^{\alpha}$$

where

γ_G: sum of the specific attenuations (dB/km) of oxygen and water vapour, respectively:

$$\gamma_G = \gamma_0 + \gamma_w$$

and

γ_R: specific attenuations (dB/km) due to the rainfall:

$$\gamma_R = k \cdot R^{\alpha}$$

R is the rain rate (mm/h); values for the coefficients k and α are determined as functions of frequency, f (GHz) and the polarisation type [50],[55].

For outdoor usage, other signal impairments like local environmental factors, such as trees and shrubs, which cause additional attenuation of the mmWave signal must be considered. The associated loss is produced by a combination of diffraction, ground reflection and through-vegetation scattering. For terrestrial links, the foliage loss is expressed as [50]:

$$L_{total} = -10\log_{10}\left\{10^{\left(\frac{-L_{sidea}}{10}\right)} + 10^{\left(\frac{-L_{sideb}}{10}\right)} + 10^{\left(\frac{-L_{top}}{10}\right)} + 10^{\left(\frac{-L_{ground}}{10}\right)} + 10^{\left(\frac{-L_{scat}}{10}\right)}\right\}$$

Where L_{sidea} and L_{sideb} represent loss due to diffraction from either side of the vegetation, L_{top} represents the diffraction loss above the vegetation, L_{ground} is the loss due to ground reflection, and L_{scat} is the loss due to scattering through the vegetation. Formulas for calculating these individual losses, which depend on frequency, depth of the vegetation, and kind of the vegetation, are included in [56].

Channel sounding and modelling at mmWave: Understanding the transmission properties of mmWaves in real-world environments is essential for the design of 5G NR UE and gNBs. As the wavelengths get smaller, physical processes such as diffraction, scattering, material penetration loss, and free space path loss, all make the channel properties of mmWave bands considerably different from sub-6 GHz bands [34].

Hence, the 3GPP study on 5G channel model for frequencies from 0.5 to 100 GHz [5] considers a number of scenarios including Urban Micro, Urban Macro, Indoor, Backhaul, Device-to-Device (D2D), Vehicle-to-Vehicle (V2V), and Stadium. The number of spatial clusters and multipath components per cluster in the mmWave channel, and the spatial dynamics, has profound implications on the design of network components. For instance, if the channel model defines a spatially rich channel, the antenna beam steering requirements are not so important, and many eigenmodes will be available for Single-User MIMO (SU-MIMO) but the resulting fast fading caused by the addition of numerous multipath signals will be complex. In contrast, a sparser channel will contain few eigenmodes, less fading but require much better beam steering. Therefore, realistic channel modelling is important for both device design and defining practical and useful test cases [34].

Since the publication of [57] mmWave channel modelling activities have been continuing at companies, universities, and at government institutions to develop a greater understanding of the mmWave channel and its behaviour [34].

In general, an appropriate channel model for mmWave bands should fulfil a set of requirements as follows [50]:

- Provide accurate space-time characteristics of the propagation channels in 3D space for LoS and NLoS conditions
- Support beamforming with steerable directional antennas on both transmitter and receiver sides with no limitation on the antenna type and technology

- Account for polarisation characteristics of antennas and signals
- Support non-stationary characteristics of the propagation channel arising from UE motion and non-stationary environment (e.g. moving people causing communication link attenuation or full blockage) [50].

Depending on the approach taken, channel models can be grouped into two categories: geometry-based stochastic approach and deterministic/quasi-deterministic. The geometry-based stochastic approach characterises the channel as several probabilistic rays with different delay and angular spread, representing a particular multipath propagation environment. The spatial and temporal statistics are collected in a measurement campaign for a given frequency in an environment like the desired deployment. On the other hand, the purely deterministic model would use electromagnetic equations and a model of the scatterer environment to predict the channel between the transmitted and the receiver, whereas the quasi-deterministic approach is based on the representation of the channel impulse response as superposition of a few quasi-deterministic strong rays and a number of relatively weak random rays, and the parameters for a given frequency in a particular environment are based on a measurement campaign. A great deal of work has already been done, by industry and academia, on channel measurements and modelling, that consider the multitude of different propagation environments and the large frequency range. The goal is to develop a set of channel models for mmWave propagation environments [50].

An advantage of mmWave systems is in the inherently small antennas required, which can be arranged in relatively small-footprint phased arrays for high directivity and beam steering [50].

Usually, root-mean-square (RMS) delay spread is increased for lower gain antennas which employ wider beams, as the wider profile collects signals from more directions with similar or equal gain to the boresight angle. This mainly applies to UE equipment whose size and power requirements do not support large arrays and have a more omni-directional pattern [50].

Conversely, RMS delay spread is decreased for higher gain antennas and the associated narrower beamwidth. The transmit beamwidth from the base station restricts the direction of the generated energy and thus the opportunities to scatter. Likewise, despite the higher gain, scattered energy of the multipath link may not be picked up by the spatial range of the receive antenna boresight [50].

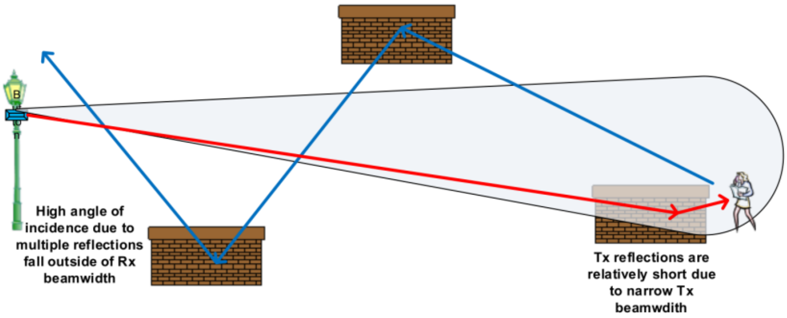

Figure 29: Scattering effects

Thus, for a given environment and use cases with different transmitter and receiver antenna radiation patterns, one may observe different scattering effects as illustrated in Figure 29. The key point is that delay spread is mitigated by the beamforming paradigm, which has been an important area of research, simulation, and prototyping [50].

Phased Array Technology, Beamforming and Beam Steering: Phased array antennas [51] are a practical and low-cost means of generating narrow beams (beamforming) and dynamically pointing them in the desired direction (beam steering). They enable beam steering without mechanical motion and are expected to be the principle mmWave antennas used for both base stations and UE. A phased array antenna is composed of an array of smaller antenna elements, such as individual patches or dipoles. By altering the relative phases and amplitudes of the signals applied to the individual elements, the antenna array can shape and steer a beam in a chosen direction. Figure 30 illustrates the basic operation of a phased array. A signal from a transmitter (Tx) is distributed to several antenna elements. Phase shifters are controlled to adjust the individual phase of the signal transmitted from each element, thereby enabling beamforming at a variable angle [34].

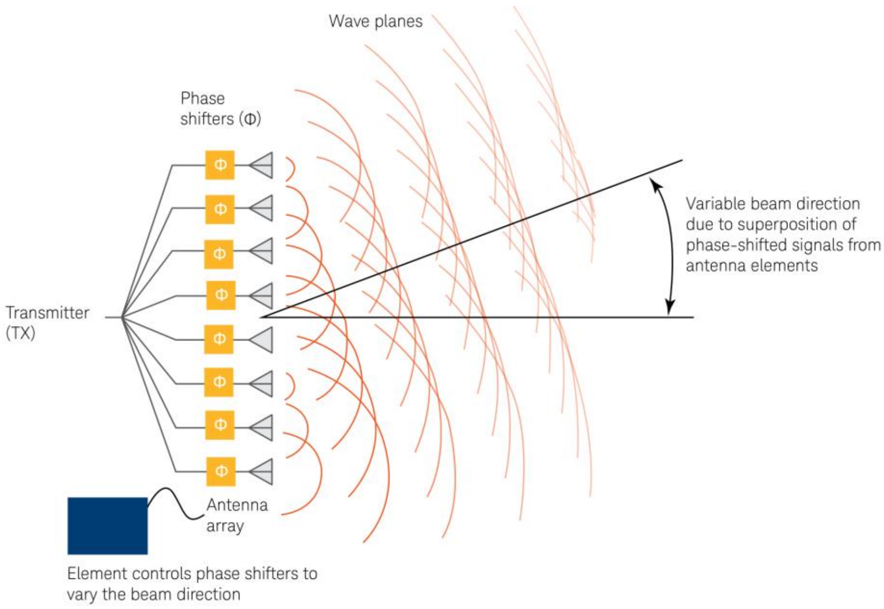

Figure 30: Basic operation of a phased array

However, building a cellular system out of narrow and focused beams is challenging and changes many traditional aspects of cellular system design. mmWave beams are highly directional, almost like flashlights, which completely changes the interference behaviour as well as the sensitivity to misaligned beams. The interference adopts an on/off behaviour where most beams do not interfere, but strong interference will occur occasionally. Overall, interference is de-emphasised and mmWave cellular links may often be noise-limited, which is a major difference from 4G. Therefore, the main issues will be link acquisition, leveraging the legacy 4G network, and the need for new transceiver architectures [8].

In this regard, three types of beamforming architectures are being considered for 5G networks: digital (implemented in baseband), analogue (implemented at IF or RF), and hybrid. Each has its relative merits. Due to cost and power reasons, beamforming in UE is expected to be analogue, whereas the beamforming of base stations may be either analogue, digital, or hybrid [34].

Radio and Antenna Integration, and the Elimination of RF Connectors: 5G mmWave devices will be highly integrated. This is driven by the combination of high frequencies, large numbers of antenna elements, the need to minimise signal path attenuation, and the need to reduce cost. A major consequence of the resulting integration is that traditional RF connectors at the boundary between the radio distribution network circuit and the antenna system are no longer possible to implement.

The distribution network employed to connect mmWave signals from the radio to the antenna must be extremely compact, especially in handheld devices and other user equipment. As a result, transceiver systems for 5G mmWave devices will be directly integrated with antenna arrays as shown in Figure 31. This will result in the devices that do not have connectors or probe points, that enable conducted tests, and can only be tested OTA [34].

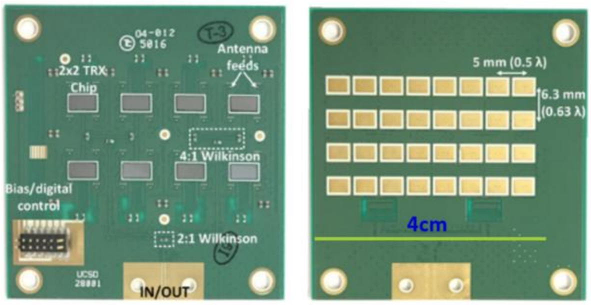

Figure 31: High-level integration of RFICs and antenna arrays

4.2.2 Adopting mmWaves – Important Factors and Use Cases

The use of mmWave bands for small cells is expected to provide the scalability, capacity and density required for a seamless integration of these cells into the cellular network infrastructure [39]. It offers the potential for increased network capacity as well as network densification. All these benefits, on the other hand, come at the expense of added system complexity particularly in terms of radio frequency (RF) front end, complex antenna design, and the need to combat higher atmospheric losses.

Therefore, the factors to consider while adopting mmWave spectrum are as follows:

Outdoor-to-outdoor coverage and link budget: The first consideration for link budget analysis is the signal power attenuation due to propagation loss over the air. The inverse Friis equation for isotropic radiators relates the free space path loss (FSPL) of an RF carrier as proportional to the square of its frequency. FSPL also increases proportionally to the square of the distance between the transmitter and receiver [52]. As such, a 30 GHz signal transmitted over 20 m loses 88 dB of power just covering this relatively short distance between transmitter and receiver. At 100 m, the loss is increased to 102 dB [50]. Coverage can also be analysed from the link budget perspective. Since the typical outdoor urban environments will include NLoS paths, the analysis should incorporate the NLoS cases [50].

Outdoor-to-indoor coverage: Due to cost reasons, it is better to provide indoor coverage using base stations located outdoors [50]:

Mobility: Since the Doppler shift is a linear function of the velocity and the operating frequency, mmWave bands will experience higher Doppler shift than sub-6 GHz bands [50].

The channel link between the base station and the mobile station typically includes multipath components which have different routes with different time delays, angles of departure and angles of arrival. From receiver's perspective, every path with a different route may result in a different Doppler shift because of potentially different angles of departure/arrival, relative to the direction of travel of the mobile station. Hence, multipath environments introduce a spread of Doppler shifts at the mobile station; the magnitude of Doppler shift is proportional to frequency and will therefore be larger at mmWave bands. As a result, the enlarged Doppler spread in these bands leads to faster channel fluctuations in the time domain than that in lower frequency bands [50].

These channel fluctuations in the time domain, in mobile communication systems, require a feedback mechanism for various information, like channel quality, as a means of properly adjusting system activity to the fluctuations so that the system's performance can be optimised. Since the latency of the feedback mechanism is a key factor in addressing the variation of the channel from mobility, it is important to minimise the latency. A limiting factor for the feedback latency is the symbol duration, which is directly determined by the bandwidth size. In mmWave bands, the available bandwidth could be much larger, and the symbol duration could be shortened, making it possible to minimise the feedback loop latency [50].

The number of multipaths is another factor that determines the amount of channel fluctuation. If the number of multipath components is decreased, the fluctuation of the channel becomes smoother in the time domain. Thus, if the multipath components impinging on the receiver are decreased by using narrow beamforming, the channel fluctuation will be less than in the case of wide beamforming. [50].

Impact of bandwidth: An advantage of adopting higher frequency for mobile communication is the ability to implement wide channel bandwidths with a bandpass filter. For FDD implementation, a duplex filter is needed. Tuneable and reconfigurable radio frequency (RF)/microwave duplex filters are critical building blocks in modern wireless systems used for channel selection and noise reduction but can also be a major bottleneck in performance. With an ever-increasing demand for the use of the mmWave spectrum for mobile services, there is an urgent need for tuneable and reconfigurable hardware to be used for different applications while enabling the effective usage of the available spectrum. This includes a broad range of novel, planar/hybrid tuneable circuit realisations for spectrum management and dynamic broadband filtering, as well as new, low-loss, micro-electromechanical (MEMS)-based reconfigurable duplex filters for wide tuning ranges and their implementation through state-of-the-art technologies. The current state-of-the-art approach enables a maximum duplex filter size of about 3-4 percent of the centre frequency of the band, which means that it is very difficult to implement a wider channel bandwidth. Carrier aggregation is required to achieve a specific data rate if centre frequency $F_c \times 0.03\sim0.04$ is less than the channel bandwidth necessary to meet the data rate, as shown in Figure 32 [50].

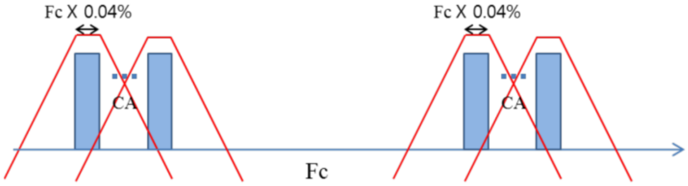

Figure 32: Impact of bandwidth at mmWave

Considering their peculiar characteristics, the following use cases of mmWave bands have been identified [50]:

Dense hotspot in an indoor shopping mall: Deployment of small cells is an effective solution to meet the ever-increasing demand for high data rate applications. Environments such as shopping malls may be situated outside the city centre, i.e. in rural and suburban areas, where the capacity of the macro-cell network may not be adequate. Bringing optical fibre to these locations may also be expensive and wireless backhaul/fronthaul solutions may be preferred to efficiently enable high data rate services in the mall. Inside such buildings, small cells are deployed to prevent outdoor to indoor propagation losses and to provide a favourable radio environment for enhanced wireless service. Essentially, technologies based on mmWave bands allow larger bandwidth at the cost of an increased number of cells [50].

Dense hotspot in an indoor enterprise environment: Providing universal coverage and high capacity in enterprise space is a major challenge for MNOs. Since enterprise buildings have different characteristics in terms of location, age, size, shape, number of rooms, etc., finding a unique solution to deliver high data rate mobile services may be difficult due to cost and scalability reasons. Technologies operating at mmWave bands can provide high data rate to users in their offices while other systems such as macro-cell networks using lower frequency bands can provide ubiquitous connectivity [50].

Dense hotspot in home and indoor environments: Wireless local area networks (WLANs) are a common means of accessing wireless applications/services at home. However, WLAN performance can suffer from interference. Moreover, current solutions do not provide seamless handover between cellular networks and WLANs. With use of indoor small cells at mmWave bands inside each apartment, small cells will benefit from the favourable radio environment characteristics that avoid interference between neighbouring apartments to offer an enhanced wireless service. Therefore, in places where dense deployment of small cells in existing frequency bands is not a viable solution due to the insufficient capacity and/or strong inter-cell interference, small cells at mmWave bands can be used as an alternative to provide high quality of experience (QoE) in indoor environment [50].

Dense urban hotspot in a square/street: This use case focuses on a square or street in the city centre where thousands of people will spend part of their everyday lives. The area is characterised by many potential indoor and outdoor hotspots, such as bus stops, restaurants, enterprises, and recreation parks. Due to the variety of uses in this setting and the high data rate requirements for multimedia broadband services, traditional solutions might not be adequate. For this dense area, MNOs can benefit greatly from the upgrade of their network through the deployment of small cells at mmWave bands, which will enhance the QoE of nearby users while providing sufficient capacity. Examples include cases where users sitting in a cafe or waiting for a bus can launch real-time video streaming apps, gaming, video calls, etc. [50].

Mobility in the city: A challenge for MNOs is the provision of high capacity inside public transport. In this case, a high number of users may require access to high data rate services in a relatively small indoor space that is characterised by high mobility (around 50 km/h or more). For instance, people using trams to move from home to work/the city centre may access the internet to read emails, update the software of their devices, download movies and files, or play videogames [50]. mmWave bands can be used to provide access and backhaul/fronthaul, dedicated to public transport, capable of delivering consistent high data rates to users when they enter and exit the public transport. Backhaul/fronthaul nodes distributed on the railways or on street level can use mmWave bands to transmit data to trams and other public transport along their routes. Also, devices installed in the public transport would provide connectivity to street level small-cell infrastructure [50]. The installation of access nodes on the main streets is another potential use case for the provision of connectivity to fast mobile users in cars and public transport, even if these vehicles are not equipped with small cells [50].

4.3 Beamforming Antenna Arrays and Massive MIMO

4.3.1 Introduction

Multiple-Input Multiple-Output (MIMO), beam steering, and beamforming are the most talked about technologies in 5G. They are essential for delivering the 100 times data rate and the 1000 times capacity goals specified in the IMT-2020 vision [58].

MIMO is an important approach to improve the capacity and efficiency of a network to meet these demands. It entails using multi-antenna technologies to support multiple frequency bands — from sub-6 GHz to mmWave frequencies — across many scenarios, including massive IoT connections and extreme data throughput [58].

For long, industry and academic researchers considered available mmWave bands as the next frontier to serve the data-hungry wireless applications of the future. New 5G systems operating at 28 GHz and above offer more available spectrum for larger channels, which are ideal for multi-Gbps links. Although these frequencies have less spectral crowding than those below 6 GHz, they experience different propagation effects such as higher free-space path loss and atmospheric attenuation, weak indoor penetration, and poor diffraction around objects [26].

5G NR Release 15 specifies frequency use up to 52.6 GHz with up to 400 MHz bandwidth per carrier and aggregation of multiple carriers for up to 800 MHz channel bandwidth. However, as stated already, operating at mmWave frequencies introduces path loss, blockage, and signal propagation challenges [25].

Understanding the challenges requires a basic knowledge of the techniques used to deliver high-quality, robust signals to and from a 5G device. Hence, there are various techniques for implementing MIMO, each offering distinct benefits and compromises [58].

The first is spatial diversity which helps improve reliability in many forms of RF communication. It entails sending multiple copies of the same signal via multiple antennas. This technique increases the chances of properly receiving the signal and thus improving reliability [58].

Another one is spatial multiplexing which is a multiple-antenna technique that feeds independent data into each antenna, with all antennas transmitting at the same frequency. It creates multiple channels with independent streams, which increases the overall data capacity [58].

Lastly, beam steering and beamforming use multiple antennas to create directional transmissions, increasing gain in exchange for a beam that must accurately point at the receiving antenna. Beamforming is more complex than beam steering, incorporating channel feedback to manipulate the beam shape and direction in real time. Spatial multiplexing with beamforming increases signal robustness with the added benefit of improved throughput. Also, multi-user MIMO can be implemented by directing multiple beams at different devices to achieve greater spectral efficiency [58].

Essentially, mmWave antenna arrays overcome undesired propagation effects by focusing their beams and taking advantage of the antenna array gain. Fortunately, the size of these antenna arrays decreases as the frequency of operation increases, allowing a mmWave antenna array with many elements to take up the same area as a single sub-6 GHz element (Figure 33) [26].

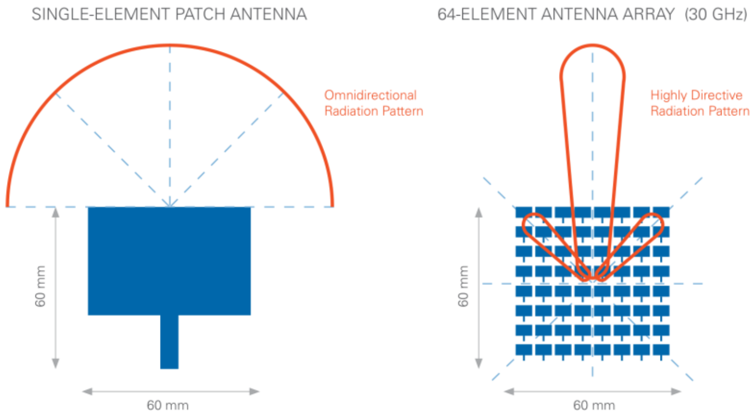

Figure 33: Comparison of mmWave and sub-6 GHz antenna arrays

5G NR also specifies new initial access procedures to ensure alignment of the directional transmissions used in beam steering. Figure 34 illustrates these initial access techniques where the base station uses beam sweeping to transmit multiple beams, identify the strongest beam, and establish a communication link. Initial access validation, beam management, and throughput achieved through the wireless link are crucial factors for successful implementation of beam steering in 5G [25].

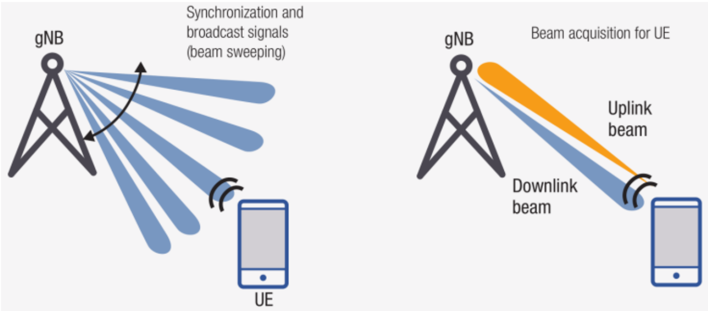

Figure 34: Beam sweeping and initial access

Evidently, a key area of innovation in wireless communications is in advanced antenna technologies. Network capacity and coverage can be improved by using more antennas intelligently. That is, more spatial data streams can significantly enhance spectral efficiency, allowing more bits to be transmitted per Hertz, and smart beamforming can extend the reach of base stations by focusing RF energy in specific directions. [1].

4.3.2 Beamforming Signal Generation, Propagation, and Management

Beamforming antenna arrays will play a significant role in 5G implementations. Apart from a higher directive gain, they offer complex beamforming capabilities. The direct targeting of user groups increases the capacity of cellular networks by improving on their signal to interference ratio (SIR). The narrow beams that are transmitted simultaneously reduce the amount of interference in the radio environment and make it possible to maintain sufficient signal power at the receiver terminal at longer distances in rural areas [38].

Several steps are required during beamforming and are explained in detail as follows:

Phase Coherent Signal Generation: A phase coherent signal is an essential prerequisite for any beamforming architecture. This term means that all RF carriers have a defined and stable phase relationship. As shown in Figure 35, a fixed delta phase between the carriers can be used to steer the main lobe to a desired direction [38].

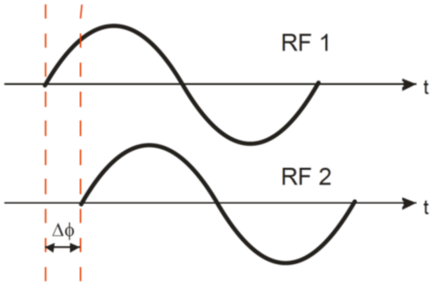

Figure 35: Phase coherent signals with phase offset

Signal Propagation: All signals radiated from any kind of antenna have the same basic characteristics. Multipath fading and delay spread diminish the capacity of a cellular network. Likewise, the practical network capacity is decreased further by congestion of the available channels and co-channel interference [38],[59].

- **Free Field Attenuation**—Electromagnetic waves are attenuated while travelling from the transmitter to the receiver. The free field attenuation describes the attenuation which the signal will experience due to the distance between the two stations [38].

  The Friis formula defines the free field attenuation [38]:

  $$P_{r,dB} = P_{t,dB} + G_{t,dB} + G_{r,dB} + 20log_{10}(\frac{\lambda}{4\pi R})$$

  Where $P_{r,dB}$ is the received power level in dB, $P_{t,dB}$ the transmitted power and $G_{r,dB}$ and $G_{t,dB}$ are the receive and transmit antenna gain in dBi [38].

  Figure 36 (left) illustrates the free field attenuation over a wide frequency band [38]. Even in case of a perfect LoS transmission, there are many distinct factors that affect the magnitude of the received signal. As shown in Figure 36 (right), the total attenuation varies greatly depending on the frequency and radiation environment [38].

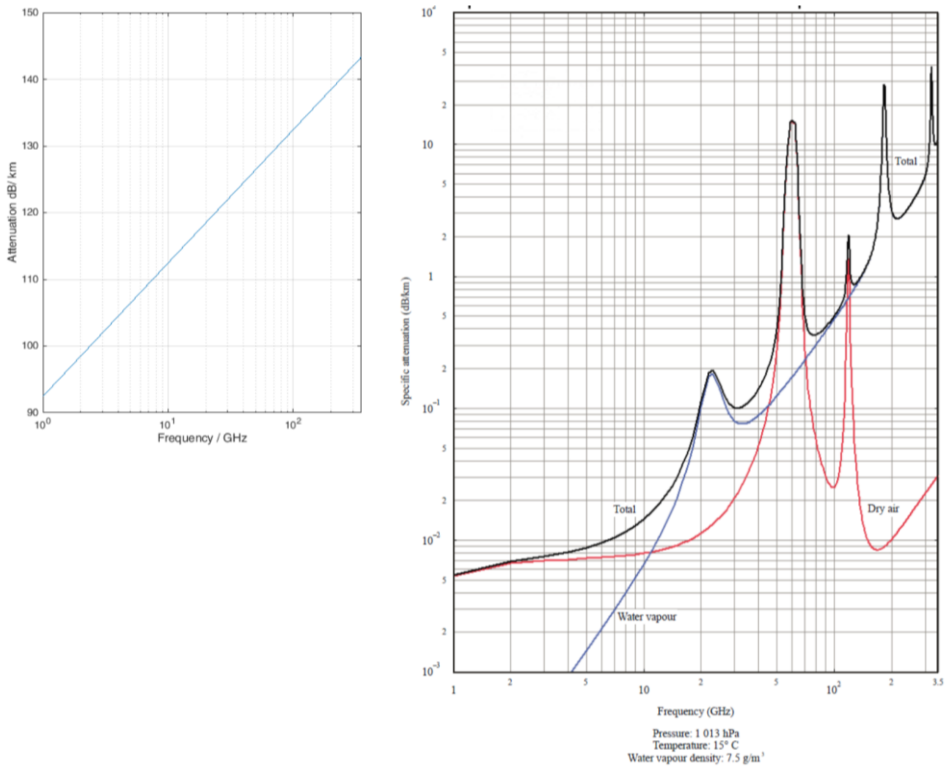

Figure 36: Attenuations - Free Field (left) and Atmospheric Gases (right)

- **Fading**—The phase shift in multipath signals is not fixed due to the time variant nature of the channel. The time-dependent received multipath signal, where the complex values $a_n(t)$ and $e^{-j\theta_n(t)}$ describe the change in amplitude and phase for the transmit path n, is given by the expression [38]:

  $$r(t) = s(t)\sum_{n=1}^{N}|a_n(t)|e^{-j\theta_n(t)}$$

  The signals add up constructively or destructively depending on the current phase shift. The received signal consists of a multitude of scattered components making it a random process. Based on a sufficient number of scattered components, this can be seen as a complex Gaussian process. This results in the formation of small fade zones in the coverage area which is called Rayleigh fading. A special case of fading is phase cancellation, which happens when multipath signals are 180° out of phase from each other. The cancellation, and therefore the attenuation of the signal, depends primarily on the amplitude and the phase balance. [38].

- **Delay Spread**—This effect is a result of the multipath nature of signal propagation. It describes the difference between the time of arrival of the earliest and latest significant multipath component. Usually, the earliest component is the LoS transmission. In the event of large delay spreads, the signal will be impaired by inter-symbol interferences which dramatically increase the bit error rate (BER) [38].

  Modern beamforming antenna architectures can help to mitigate these effects by adapting to the channel. In this way, delayed multipath components can be ignored or significantly reduced through beam steering. Antennas that are designed to adapt and alter their radiation pattern in order to adapt to the RF environment are called active phased array antennas [38],[59].

Gaining Access and Managing Beams: Managing large signal propagation loss at frequencies above 20 GHz is one of the major technical challenges of operating in mmWave bands. In concrete terms, this loss reduces the possible cell coverage area and range. To compensate for this, standard designers decided on beamforming technology with antenna arrays as a means of focusing RF energy on individual users and boosting signal gain. Hence, UE can no longer rely on the mmWave gNBs to broadcast omnidirectional signals to establish the first connection. The 5G NR standard, therefore, implements a new procedure for UE to gain initial access to the gNB. When a UE arrives at a new cell coverage area, it is blind to the location of the beam, ignoring the direction in which the gNB is currently transmitting to begin the network access procedure. The 5G NR initial access procedure offers an elegant solution for UE to establish communication with the gNB. It solves the problem of locating the gNB in the dark not only for mmWave operation but also for sub-6 GHz omnidirectional operation. This means that the initial access procedure must work in single-beam and multibeam scenarios. It also must support 5G NR and 4G LTE coexistence. The initial access procedure the steps depicted in Figure 37 and explained in detail as follows [26].

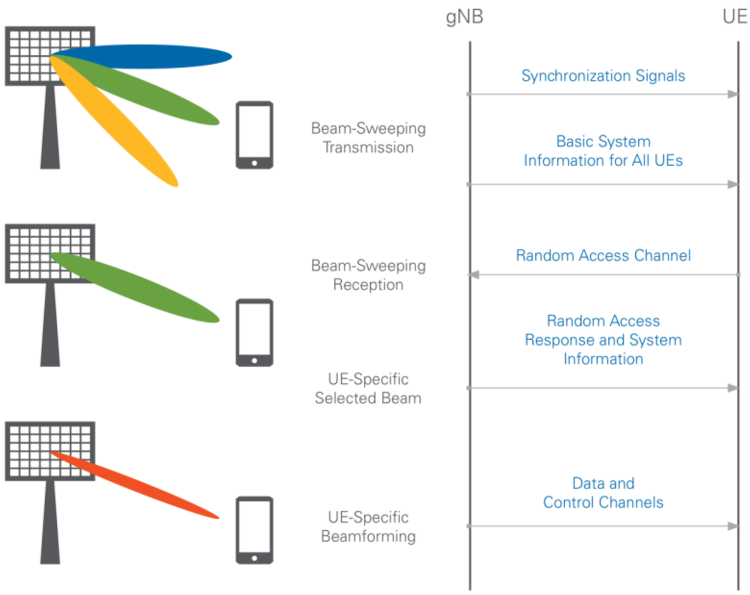

Figure 37: Initial access procedure

- **Beam-sweeping transmission**—The gNB transmits the physical broadcast channel (PBCH) in groups of four OFDM symbols called synchronisation signal (SS) blocks sequentially in multiple directions, as illustrated by the blue, green, and yellow beams, and maps each one to a different spatial direction. Using the concept of beam sweeping, the gNB transmits both the synchronisation signals and the system configuration information that the UE requires to access the network [26].
- **Beam-sweeping reception**—The UE detects the best SS block, which is the strongest detected beam, by listening until it matches the beam direction of the transmitter. This allows the UE to decode it and extract its time index. With the knowledge of when the gNB will use that beam direction again, the UE transmits to the gNB on the physical random-access channel (PRACH) at the right time. The gNB now knows in which direction and at what time the UE will transmit UL information [26].
- **UE-specific selected beam**—Once the UE and gNB establish communication via the best beam, the gNB sends the rest of the system information that the UE requires to set up a connection with the gNB [26].
- **UE-specific beamforming**—The system then proceeds to switch from a general, wider beam coverage to a UE-specific coverage with a narrower beam, using beam-refining procedures [26].

4.3.3 Beamforming Architectures

As already seen, mmWave bands potentially enable high bandwidths. To date, the limited use of these high frequencies is a result of adverse propagation effects especially due to obstacles in the LoS. Several transceiver architectures have been developed to compensate these issues by focusing the received or transmitted beams in a desired direction. All these solutions make use of small antenna element sizes due to higher carrier frequencies that enable the construction of large antenna arrays [38].

Amplitude and phase are the two most important variables required for beamforming. The combination of these two factors is used to improve side lobe suppression or steering nulls. The amplitude and phase for each antenna element n are combined in a complex weight w_n. The complex weight is then applied to the signal that is fed to the corresponding antenna [38].

Therefore, achieving a certain directivity or beamforming requires an antenna array where the RF signal at each antenna element is amplitude and phase weighted. There are three possible ways to apply amplitude and phase shifts [60].

Analogue beamforming: The approach makes use of an antenna array combined with phase shifters and power amplifiers to steer the beam in a desired direction, reducing sidelobes to a minimum. It is a cost-effective method due to the small amount of hardware that is required to build the beamforming array. Typically, an analogue beamforming array is connected to one RF chain generating only one beam at a time, and the range of the phase shifters used limits the applicable frequency range [60]. Figure 38 shows a basic implementation of an analogue beamforming transmitter architecture. It is composed of only one RF chain and multiple phase shifters that feed an antenna array [38].

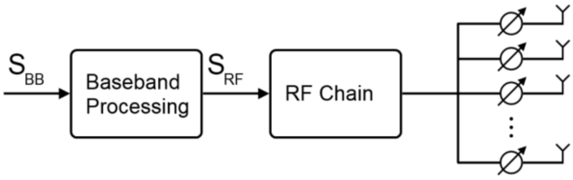

Figure 38: Analogue beamforming architecture

The first practical analogue beamforming antennas date back to 1961. The steering was done with a selective RF switch and fixed phase shifters [61]. The basics of this method are still used today, although with advanced hardware and improved precoding algorithms. These enhancements allow separate control of the phase of each element and, unlike early, passive architectures, the beam can be steered not only to discrete but virtually any angle using active beamforming antennas. Just as its name indicates, this type of beamforming is achieved in the analogue domain at RF frequencies or an intermediate frequency [38],[62].

In order to calculate the phase weightings of the analogue architecture, a uniformly spaced linear array with element spacing d is assumed. Considering the receive scenario shown in Figure 39, the antenna array has to be in the far-field of the incoming signal for the arriving wave front to be approximately planar. [38].

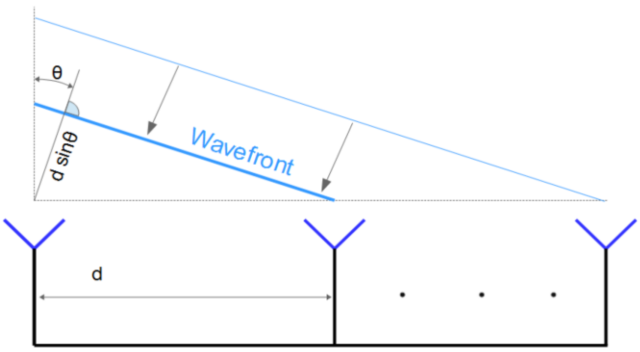

Figure 39: Additional travel distance when signal arrives off boresight

If the signal arrives at an angle θ off the antenna boresight, the wave must travel an additional distance $d * \sin\theta$ to arrive at each successive element as illustrated in Figure 39. This translates to an element specific delay which can be converted to a frequency dependent phase shift of the signal [38]:

$$\Delta\varphi = \frac{2\,\pi\, d\, \sin\theta}{\lambda}$$

The main lobe of an antenna array at a defined frequency can be steered to a certain angle using phase offsets calculated with the above expression. If the antenna elements are now fed with a signal of a different frequency, the main lobe will veer off by a certain angle. Since the phase relations were calculated with a specific carrier frequency in mind, the actual angle of the main lobe shifts according to the current frequency. This effect, which is frequency dependent, is called beam squint. By using time delays instead of frequency offsets, the previous expression can be converted to a frequency independent term [38]:

$$\Delta t = \frac{d * \sin\theta}{c}$$

This implies that the frequency dependency is eliminated if the setup is fitted with delay lines in place of phase shifters. The corresponding receiver setup is shown in Figure 40.

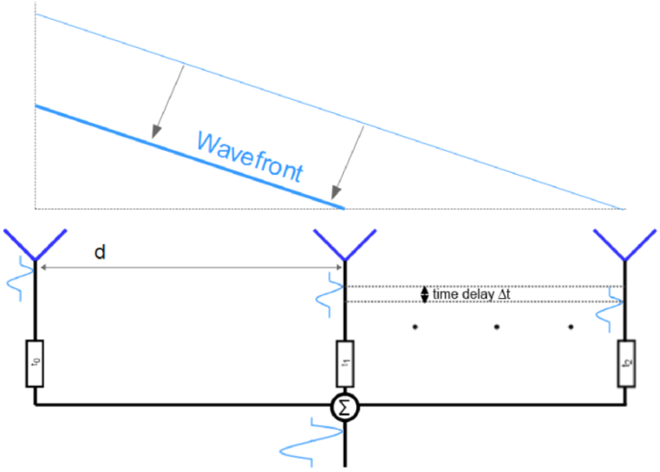

Figure 40: True time delay beam steering

The delay lines t_0 to t_2 compensate for the time delay Δt, which is an effect of the angle of the incident wave. Consequently, the received signals should be perfectly aligned and will thus add constructively when summed up [38].

The performance of the analogue architecture can be enhanced by changing the magnitude of the signals that are incident on the radiators [38].

Digital beamforming: This is a more advanced architecture in which phase and amplitude weighting is performed in the digital domain. With each antenna having its own transceiver and data converters, it is able to handle multiple data streams and generate multiple beams simultaneously from one array. In addition to beam steering, this architecture makes it feasible to implement null steering for interference reduction. However, it requires analogue-to-digital converters (ADCs), making it challenging at higher frequencies [60].

While analogue beamforming is typically restricted to one RF chain, even when using a large number of antenna arrays, digital beamforming in theory supports as many RF chains as there are antenna elements. If appropriate precoding is done in the digital baseband, this yields higher flexibility regarding the transmission and reception. The additional degree of freedom can be exploited to perform advanced techniques like multibeam MIMO. These benefits result in the highest theoretical performance possible compared to other beamforming architectures [63].

Figure 41 illustrates the general digital beamforming transmitter architecture with multiple RF chains [38].

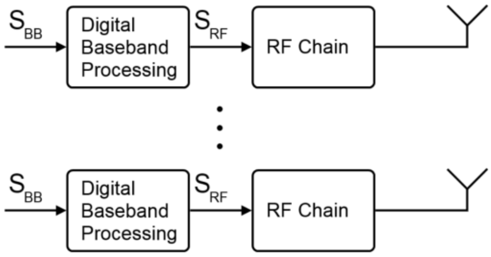

Figure 41: Digital beamforming architecture

Beam squint is a well-known challenge for analogue beamforming architectures using phase offsets, but with digital control of the RF chain, the phases can be optimised according to the frequency over a large band. Digital beamforming can also accommodate multi-stream transmission and serve multiple users simultaneously, which is a key driver of the technology. Nonetheless, it may not always be ideally suited for practical implementations regarding 5G applications. It has a high complexity in terms of its hardware requirements which would significantly increase cost, energy consumption and complicate integration of mobile devices. Thus, it is better suited for use in base stations since performance outweighs mobility in this case [38].

Hybrid beamforming: This beamforming architecture balances the advantages and disadvantages of analogue and digital beamforming. Targeting higher frequency ranges, such designs combine multiple antenna array elements into subarray modules that are connected to a digital pre-processing stage. With this approach, system designers are able to balance flexibility and cost trade-offs while still meeting the required performance parameters such as number of simultaneous beams and frequency range [60].

With hybrid beamforming, a substantial cost reduction can be achieved by reducing the number of complete RF chains, which also leads to lower overall power consumption. Compared to digital beamforming, there are less degrees of freedom for digital baseband processing since the number of converters is significantly lower than the number of antennas. Thus, the number of simultaneously supported streams is reduced as well. When they are operated at mmWave bands, the performance gap between both schemes should be relatively low due to the specific channel characteristics in those bands [38],[64]. The schematic architecture of a hybrid beamforming transmitter is shown in Figure 42.

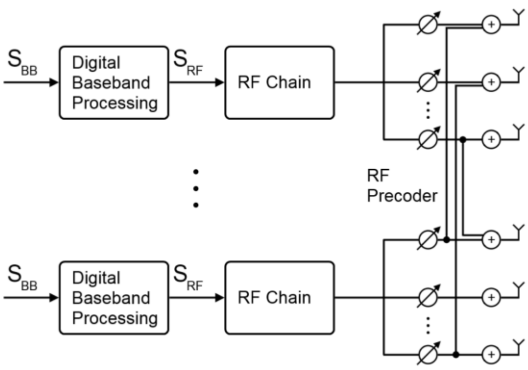

Figure 42: Hybrid beamforming architecture

The precoding is split between the analogue and digital domains. In theory, it is possible that every amplifier is interconnected to every radiating element [38].

4.3.4 Linear Array Antenna Theory

Antenna arrays can be arranged in several ways, the most common architectures being uniform linear arrays (ULAs), uniform rectangular arrays (URAs), uniform circular arrays (UCAs) and stacked uniform circular arrays [65]–[67]. Their system performance is usually measured in terms of beam gain and half power beamwidth in both azimuth and elevation [21].

Now, assume a linear antenna array with N equally spaced isotropic radiating elements. These elements can be imagined being placed along the x-axis of a spherical coordinate system, as shown in Figure 43 [38].

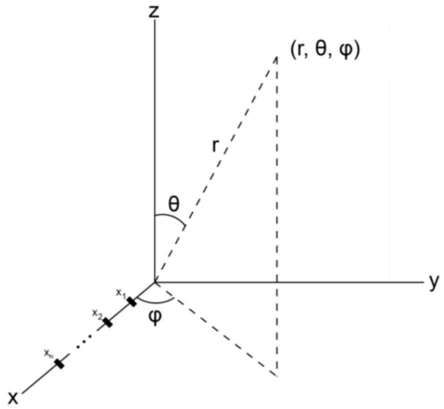

Figure 43: Linear antenna array

The radiation pattern F_{array} of a linear antenna array can be estimated by multiplying the array factor AF_{array} with the element radiation pattern $F_{element}$, that is considered equal for all elements assuming a large enough array, giving the equation [38],[68]:

$$F_{array}(\theta,\phi) = F_{element}(\theta,\phi) * AF_{array}(\theta,\phi)$$

The assumption of equal radiation patterns does not hold if the number of antenna elements is small. The outer elements may differ by a large degree from the pattern of the other antennas, which cannot be neglected in case of only a few elements. Thus, the previous equation is only applicable for coarse approximation in this case. Mutual coupling and losses in the elements are not considered as well. These effects contribute to a modified beam pattern with increased side lobe levels [38],[69].

Apart from the element radiation pattern $F_{element}$, the array factor AF_{array} is required to calculate F_{array}. For a linear array, this factor depends on the wavelength λ, the angle direction θ, the distance d between the elements and the number of elements N [38],[69]:

$$AF_{array}(\theta,\phi) = \sum_{n=1}^{N} a_n e^{jnkd\ sin\theta sin\phi} e^{j\Delta\varphi}\ ;\ k = 2 * \pi/\lambda$$

This can be simplified by introducing ψ, which describes the far-zone phase difference between adjacent elements [38],[70].

$$\psi = kd\ sin\theta sin\phi + \Delta\varphi$$

Resulting in: $AF_{array}(\theta,\phi) = \sum_{n=1}^{N} a_n e^{jn\psi}$

Further simplification and normalisation leads to the normalised array factor [70]:

$$\left|AF_{array}(\psi)\right| = \frac{1}{N}\left|\frac{\sin(N\psi/2)}{\sin(\psi/2)}\right|$$

The normalised array factor is periodic by 2π and allows to infer a lot of information about the characteristics of the linear antenna array [38].

The equations obtained show that the number of elements and their equidistant spacing have a great influence on the characteristics of a linear antenna array. The effects of modifying these two parameters will be explained by the example of Figure 44 [38].

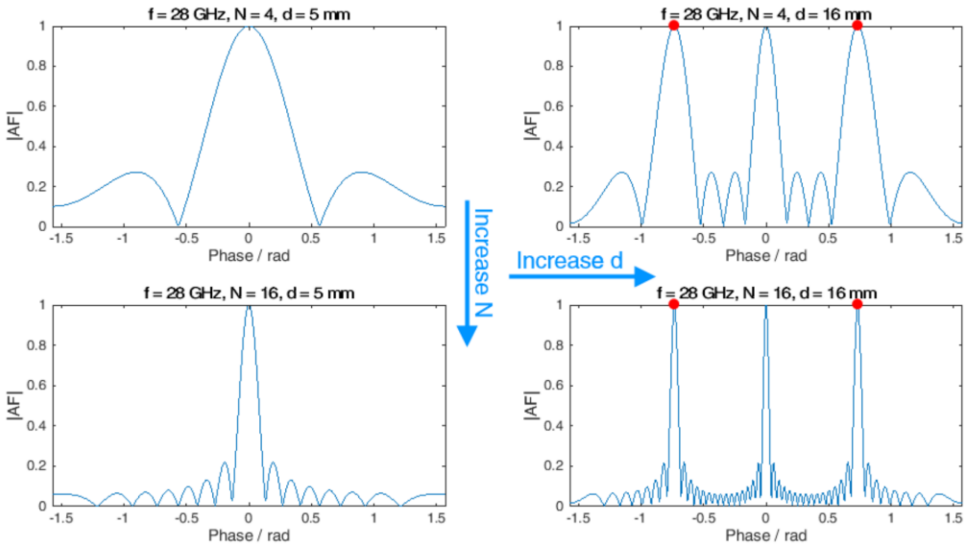

Figure 44: Normalised array factor for multiple configurations

The diagrams on the left show the normalised array factor $|AF_{array}(\psi)|$ for an antenna with an equidistant spacing of 5 mm between elements. The element distance is thus slightly less than 0.5λ at 28 GHz. The normalised array factor of an antenna with a spacing of 16 mm, which corresponds roughly to 1.5λ, is displayed on the right side. The diagrams on the upper half were calculated for an array of four elements, while the ones on the lower half belong to arrays consisting of 16 elements [38].

Comparing the upper and lower diagrams of Figure 44 illustrates the effect of increasing the number of elements while keeping the equidistant spacing constant. The main lobe width decreases for a larger element count. This means that the more elements a linear array has, the more directivity will be observed. Another effect of increasing the number of elements is a higher number of side lobes with an overall decrease in level [38].

The directivity of a linear array can also be improved by increasing the distance between elements, which generates a narrower main lobe. Similar to a higher number of elements, increasing the distance between elements increases the number of side lobes. On the contrary, a large inter-element gap generates side lobes that are of equal level compared to the main lobe. These side lobes, marked by red dots highlighting this effect for the antenna with a spacing of 1.5λ in Figure 44, are called grating lobes. Usually, these grating lobes are undesired as energy will be radiated to or received from unwanted directions. In applications that require large bandwidths, grating lobes can only affect part of the frequencies of operation [38].

Linear arrays with equidistant element spacing will generate grating lobes if the inter-element spacing exceeds half a wavelength. To prevent this phenomenon from occurring in the visible region, which is defined as the range [-90° 90°], the following condition must be kept [38]:

$$d < \frac{\lambda}{2}$$

If this condition is violated, grating lobes of increasing level begin to appear in the visible region. In a case where the distance between the elements exceeds one wavelength, the grating lobe levels start to equal the main lobe level [38].

Due to the periodicity of the array factor, grating lobes enter the visible region coming from the invisible region. If the grating lobes enter the visible region, the scan angle has to be restricted or the element spacing must be decreased. The maximum scan range $|\theta_0|$ for a given element distance d is specified in the grating lobe criteria [38],[71]:

$$sin|\theta_0| \leq \frac{\lambda}{d} - 1$$

Rearranging it helps to calculate the maximum value of d for a given scan range $|\theta_0|$:

$$\frac{d}{\lambda} < \frac{1}{1+sin|\theta_0|}$$

If a scan range of $|\theta_0| < 45°$ is selected, the element distance should not exceed 0.58λ in order to avoid grating lobes in this region [38].

Suppression and manipulation of grating lobes is a current research topic. It has been proven that the position and levels of grating lobes can be manipulated by modifying the element shape [72]. Using non-uniform element spacing allows to suppress certain grating lobes but adds a significant layer of complexity [38].

4.3.5 Antenna Array Technology

Antennas scale well with frequency and hence 5G UE and wireless access points will exploit the smaller required footprint by implementing additional elements. Considering a typical handset size of 120 mm × 60 mm, it would be feasible to fit upwards of 24 antenna elements across the shortest edge (60 mm) operating at 60 GHz. An 8 × 8 array would occupy a footprint of 20 mm × 20 mm [50].

The integration of higher bands with lower bands (such as 700 MHz, 1.9 GHz, 2.6 GHz) entails a much larger footprint, making it difficult to design antennas that can operate well at both 700 MHz and 60 GHz, and as such, it is possible that the two bands would require separate antennas [50].

The shorter wavelengths at mmWave frequency bands make it feasible to put more antenna elements in the limited size of the form factor. Therefore, antenna technology with an extended number of antenna elements can be used to provide high beamforming gain so that the increased path loss of mmWave frequency bands can be mitigated with beamforming techniques [50].

However, with the increased number of antennas with wider bandwidth, having an ADC/DAC per antenna element will be challenging because of the overall cost and power consumption. For this reason, communication systems at mmWave bands utilise a phased array architecture in RF or intermediate frequency (IF) instead of baseband, which reduces the number of ADCs/DACs while keeping a high beamforming gain [50].

Phased array beamforming is used to enhance the received signal power using beamforming gain while the baseband signal processing at the digital precoder is used to manage multiple streams for further improvement [50].

Even though greater antenna gains can be achieved with narrower beams, this might be detrimental to MIMO communications as some scattering environments require broader beam antennas to maximise channel capacity. In essence, narrow beams are not always ideal. If the beams are too narrow, then beamforming may not be able to properly synthesise a broad, omni-like aggregate pattern [50].

With this background in mind, here are some antenna array architectures that have been proposed.

Directional fixed-beam antenna array: Fixed-beam antenna array is a beamforming technology in which the direction of multiple beams are pre-adjusted and fixed in order to provide overall good service coverage within a cellular area. The fixed beamforming has the benefit of simplicity because it does not require phase shifters. Phase shifters usually require complex calibration process and precise adjustment that demand skilled experience [50].

Full adaptive antenna array: Traditional copper-based printed antenna designs remain viable at 60 GHz. The issues at 60 GHz are mainly the complexity and performance of the Tx/Rx modules. Hence, a fully adaptive antenna array is feasible strictly from the antenna element design perspective both in cost and performance. Some alternatives like reflect array, transmit array, and parasitic array antennas can be used to enhance the size and gain without additional transmission line loss [50].

Modular antenna array: As discussed earlier, there are many sources of propagation losses at mmWave frequencies, for which antennas with high directivity may be used for their compensation. Likewise, modern communication systems often require a base station that is capable of covering a relatively wide sector around it and communicating with other stations regardless of their locations. Traditional antenna architectures are generally not able to combine wide angle coverage with high directivity. On the other hand, reflective, parabolic dishes and lenses can create narrow beams that allow them to deliver the needed 30-40 dB antenna gain, but they lack the flexibility to cover wide angle and are relatively bulky. Phased patch antenna arrays allow steering a beam in a desired direction. However, to achieve the necessary directivity, the array must consist of a large number of elements (typically several hundreds to thousands) [50]. Meanwhile, antenna array architectures that are currently been mass produced for personal devices are comprised of a single module containing a radio-frequency integrated circuit (RFIC) chip that includes controlled analogue phase shifters capable of providing several discrete phase shifting levels. The antenna elements are connected to the RFIC chip via feeding lines. Because of loss on the feeding lines, this approach only allows the implementation of antenna arrays with relatively small dimensions of up to 8 × 8, thus achieving gains of about 15-20 dB [50].

Modular, composite antenna arrays are a novel antenna array architecture for mmWave band that provides simultaneous flexibility in form factor choice, beam steering, and high array gain. They are equally cost-efficient to construct. Each module is implemented in a traditional way with dedicated RFIC chip serving several antenna elements and an RF beamforming unit. The architecture is shown in Figure 45 [50].

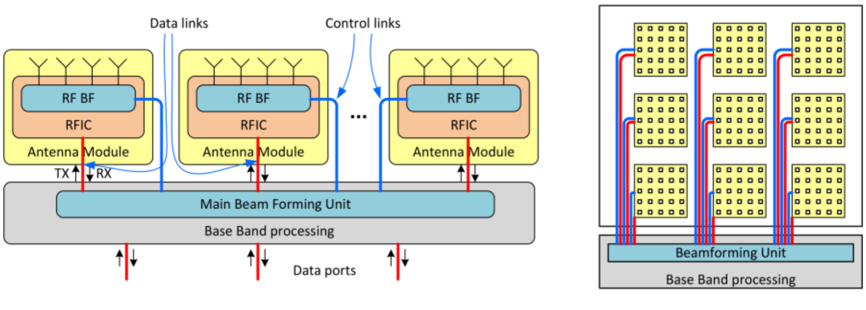

Figure 45: Modular antenna array

The aperture of the modular antenna array and total transmitted power exceeds that of an individual sub-array module in proportion to the number of the sub-array modules used. Hence, narrower beams are generated and, therefore, much greater antenna gains can be achieved with the modular array than with individual sub-arrays [50].

It is also possible that sectors of different sub-arrays can be configured in a way that the coverage angle of the composite array can be varied, thereby creating several coverage angles that would enable communication with several peer stations simultaneously [50].

Furthermore, multi-antenna transmission at mmWave frequencies can carried out using modular phased antenna structures [50]. Traditional multi-antenna transmission implementations assume each antenna can utilise an independently coded spatial stream with its own transceiver chain. However, this may be too challenging for these systems at mmWave frequencies due to the sheer number of elements involved for any degree of antenna gain [50].

Although it is true that coupling between antenna elements could be a challenge at mmWave frequencies, it is easier to mitigate such effects at these frequencies. Reducing spatial correlation between antenna elements is easier as frequencies increase. This could be done by increasing the separation between the antenna elements. In general, antenna designers can obtain superior MIMO performance at mmWave frequencies than lower frequencies because of the reduced space constraints on antenna footprint [50].

In the modular phased array architecture, each phased array module has a dedicated transceiver with an RF beamforming unit and can generate a dedicated RF beam as shown in Figure 46. The beamforming network is implemented at the RF or at the baseband of the transceiver unit [50].

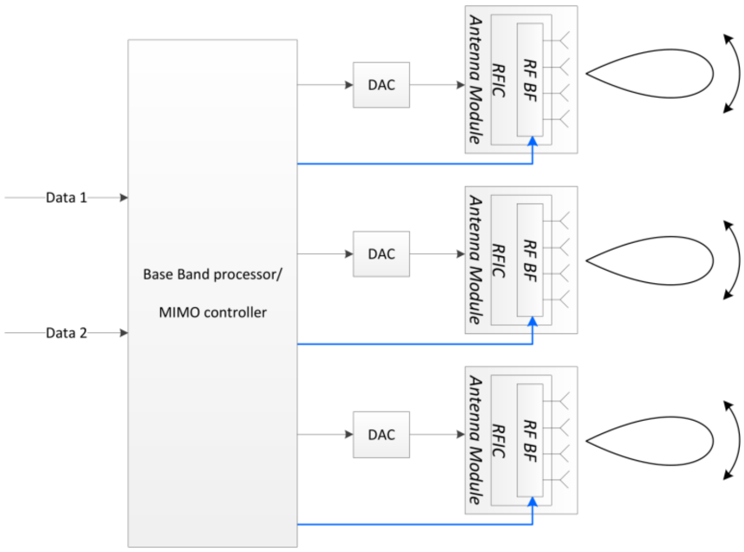

Figure 46: Multi-antenna transmission using modular phased arrays

Beams of individual sub-arrays can be steered in different directions to achieve a number of goals. For instance, one may want to steer all sub-array beams in various directions and configure each one to communicate with a different user. This will create the equivalent of an omnidirectional antenna pattern which may be useful to train antenna system of a peer station. Alternatively, it could be steered to communicate with other sub-arrays which are serving several users. This may substantially increase the throughput delivered to the users by the small cell [50].

Each antenna sub-array module could be imagined as a single antenna port in the context of a MIMO system. The beamforming gain is a result of proper array phasing of each element within a module and each module within the entire antenna system. Sub-arraying could lead to grating lobes, which can impact performance when trying to perform direction finding and/or null steering [50].

Apart from increasing throughput, beams could be combined as a means of extending coverage distance. The appropriate phasing of sub-arrays with respect to each other may result in a very narrow beam used for communication over extended distances and interference reduction in the entire deployment [50].

4.3.6 Massive MIMO

Multiple input multiple output (MIMO) is a technology that has been in use for many years. MIMO utilises many antennas at the transceiver to improve diversity gain. It also exploits uncorrelated propagation paths for higher efficiency and high throughput and/or to allow simultaneous access for different users [60].

Arising from research that blossomed in the late 1990s [73],[74], MIMO communication was introduced into WiFi systems around 2006 and into 3G cellular shortly afterwards. In essence, MIMO embodies the spatial dimension of the communication that arises once a large number of antennas are available at base stations and mobile units [8].

In single-user MIMO (SU-MIMO), the dimensions are constrained by the number of antennas that can be accommodated on a portable device. However, by having each base station communicate with several users simultaneously, the multiuser version of MIMO (MU-MIMO) can efficiently pull together the antennas at those users and overcome this bottleneck [8].

Well-established by the time 4G LTE was developed, MIMO became a key technology with two-to-four antennas per mobile unit and as many as eight per base station sector, and it seemed, because of form factors and other apparent limitations, that it was the extent to which it could be exploited. Marzetta was instrumental in formulating a theory in which the number of antennas increased by more than an order of magnitude, first in a 2007 presentation [75] with the details formalised in a landmark article [76]. The article made a bold proposal to equip base stations with a number of antennas much larger than the number of active users per time-frequency signalling resource. Given that under reasonable time-frequency selectivities accurate channel estimation can be conducted at most at tens of users per resource, this condition puts the number of antennas per base station into the hundreds. This brilliant idea, initially called "large-scale antenna systems" but now popularly known as "massive MIMO", offers enticing benefits [8]:

- Enormous enhancements in spectral efficiency without the need for increased base station densification [43],[77].
- Smoothed out channel responses because of the vast spatial diversity, which enables the favourable action of the law of large numbers. In principle, all small-scale randomness abates as the number of channel observations grows.
- Simple transmit/receive structures because of the quasi-orthogonal nature of the channels between each base station and the set of active users sharing the same signalling resource [8].

New 5G base stations (i.e. gNBs) will support Massive MIMO, which incorporates more than a hundred antennas, each transmitting a unique data stream. Massive MIMO designs allow the gNB to send or receive multiple signals to or from multiple users at once, resulting in higher spectral efficiency (bits per second per Hz of bandwidth) and higher signal to interference and noise ratio (SINR). This translates to higher user throughput and better cell coverage [15].

The high number of antenna elements, used at the transceivers of massive MIMO systems, allows two major concepts to be dynamically combined: beamforming and spatial multiplexing, both brought about by the ability of the many antenna elements to focus their energy towards smaller regions of space. If an antenna system can do this, it is referred to as massive MIMO. It is mainly applied at base stations whereas 5G UE may implement basic beamforming schemes [60].

The antenna arrays in massive MIMO systems serve numerous terminals simultaneously in the same time-frequency resource. The basic premise behind massive MIMO is to derive all the benefits of conventional MIMO, but on a much larger scale. Various configurations and deployment scenarios for the antenna arrays used by massive MIMO system have been proposed (Figure 47) [78].

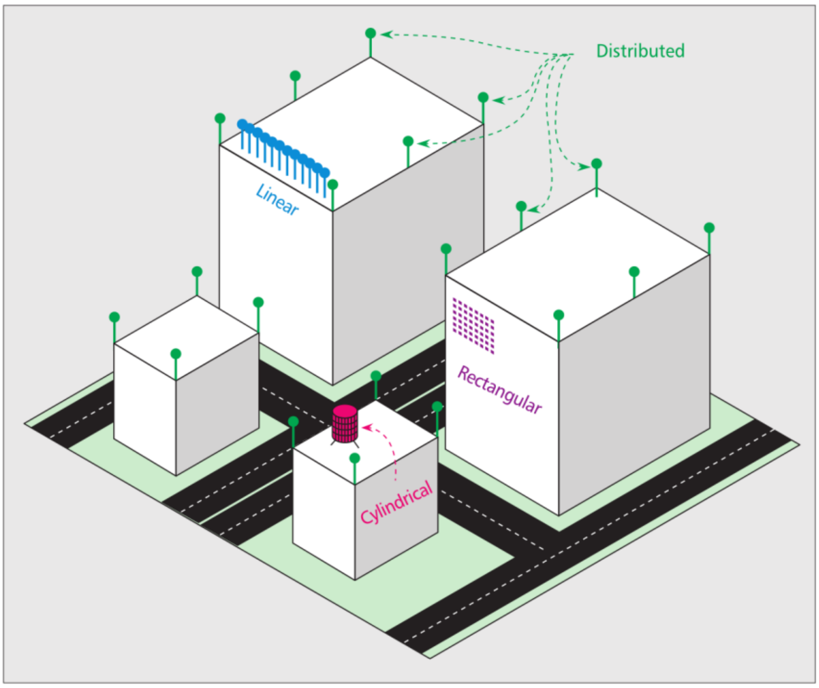

Figure 47: Possible massive MIMO antenna array configurations

The antenna units are small and active and could be fed with an optical or electric digital bus [78].

Massive MIMO relies on spatial multiplexing, which in turn relies on the base station with sufficient channel knowledge, on both the uplink and the downlink. On the uplink, this is easily done by having the terminals send pilots, based on which the base station reciprocates by estimating the channel responses for each of the terminals. A similar action is more difficult in the downlink. In conventional MIMO systems like in the 4G LTE standard, the base station sends out pilot waveforms, based on which the terminals estimate the channel responses, quantise the obtained estimates, and feed them back to the base station. This will not be viable in massive MIMO systems, especially when operating in a high-mobility environment, for two reasons. First, optimal downlink pilots should be mutually orthogonal between the antennas. This means that the amount of time-frequency resources required for downlink pilots scales with the number of antennas, so a massive MIMO system would need up to 100 times more resources than a conventional system. Second, the number of channel responses each terminal must estimate is also proportional to the number of base station antennas. Hence, the uplink resources required to notify the base station of the channel responses would be up to 100 times higher than in conventional systems. Generally, the solution is to operate in TDD mode, and depend on reciprocity between the uplink and downlink channels, although FDD operation may be possible in some cases [78],[79].

Active antenna systems (AAS) are essential for massive MIMO implementation. An AAS is made up of RF components, such as power amplifiers and transceivers, integrated with an array of antenna elements. This configuration not only reduces feeder cable losses, leading to improved performance and lower energy consumption, but also simplifies installation and minimises equipment space requirement [80].

By using AAS technology, it is possible to deploy massive MIMO antennas in 5G radio access networks. Due to its high beam gain, massive MIMO can be utilised to fulfil 5G requirements in terms of coverage and system capacity. Also, with reduced array size and more isolation of inter-cell interference, massive MIMO operating at mmWave bands will be more suitable for small cells. Furthermore, in a heterogeneous network comprised of macro and small cells, it will provide a flexible means of interference coordination/avoidance that considers practical factors like channel estimation and control signalling overhead to support large numbers of very narrow beams [80].

Benefits of a Massive MIMO System: Massive MIMO technology relies on phase coherent signals, which are simple to process, from all the antennas at the base station. Some specific benefits of a massive MIMO system are:

- It can increase network capacity 10 times or more and simultaneously improve radiated energy efficiency up to 100 times. The capacity increase results from the aggressive spatial multiplexing used in massive MIMO. The basic principle that makes the dramatic increase in energy efficiency possible is that with a large number of antennas, energy can be focused with extreme sharpness into small regions in space. The underlying physics is coherent superposition of wavefronts. By appropriately shaping the signals transmitted by the antennas, the base station can make sure that all wavefronts collectively emitted by all antennas add up constructively at the locations of the intended terminals, but destructively almost everywhere else [78].
- It can be built with cheap, low-power components. Massive MIMO is a game changing technology in terms of theory, systems, and implementation. With massive MIMO, the expensive 50 W amplifiers used in conventional systems are replaced with hundreds of cheap amplifiers with output power in the mW range. Expensive and bulky items like large coaxial cables can be eliminated as well [78].
- It enables a huge reduction of latency in the air interface. The performance of wireless communications systems is usually limited by fading. Fading can make the received signal strength very small at certain times. This happens when the signal sent from a base station travels through multiple paths before it reaches the terminal, and the waves derived from these multiple paths interfere destructively. Fading makes it hard to build low-latency wireless links. If the terminal is trapped in a fading dip, it has to wait until the propagation channel has sufficiently changed before any data can be received. Massive MIMO relies on the law of large numbers and beamforming to avoid fading dips, so that the impact of fading on latency is reduced [78].
- It simplifies the multiple access layer. With OFDM, each subcarrier in a massive MIMO system will have substantially the same channel gain. Most of the physical

layer control signalling becomes unnecessary since each terminal can be given the whole bandwidth [78].

- It increases network robustness against both unintended man-made interference and intentional jamming. Intentional jamming of civilian wireless systems is an increasing concern and a serious cybersecurity threat that is not well-known by the public. Simple jammers can be bought off the Internet at cheap prices or even put together, using off-the-shelf software radio-based platforms, at very low expense. Therefore, the only way of improving the robustness of wireless communications is to use multiple antennas because spreading information over frequency is not an option due to the scarcity of bandwidth. Massive MIMO offers many degrees of freedom that can be used to cancel signals from intentional jammers. If massive MIMO is implemented using uplink pilots for channel estimation, smart jammers could cause harmful interference with low transmission power. However, more ingenious implementations using joint channel estimation and decoding should be able to greatly reduce that problem [78].

Challenges: Even though it has many benefits, massive MIMO has many challenges as well. These include:

- **Pilot Contamination and Overhead Reduction**: Pilot transmissions are made orthogonal among users in the same cell to facilitate cleaner channel estimates [81],[82] but must be reused across cells, otherwise all available resources would end up being consumed by pilots. This unavoidably causes interference among pilots in different cells and hence reduces the quality of the channel estimates. This interference, referred to as pilot contamination, does not vanish as the number of base station antennas increases, and so is the one impairment that remains asymptotically [8].
- **Architectural Challenges**: Another more serious challenge to the realisation of massive MIMO systems has to do with their architecture. Dealing with this issue requires a complete redesign of base station structures where, instead of a few high-power amplifiers feeding a handful of sector antennas, there would be lots of tiny antennas fed by low-power amplifiers in their place. Therefore, some issues such as scalability, antenna correlations and mutual couplings, and cost must be sorted out [8]. For instance, mutual coupling between antenna elements results in energy loss and thus a reduction in their maximum range. Also, while moving from theory to practice, some antenna arrays will need to be designed in non-geometric shapes that may result in dissipating energy in undesired directions. Likewise, given the large number of antenna elements, antennas that are not calibrated properly will suffer from unwanted emissions in unwanted directions [60].
- **Channel Models**: Parallel to architectural issues, developing sound channel models would be challenging and would require extensive field measurements. Antenna correlations and couplings for massive arrays with relevant topologies must be determined, and a proper modelling of their impact must be established [8].
- **User Mobility**: User mobility can restrict how well massive MIMO solutions scale up in performance because channel coherence time decreases significantly at mmWave frequencies, which places tough restrictions on mobility applications. For correct channel estimation, the system needs to send UL pilots and payload in the UL direction. The faster UE moves, the shorter the channel coherence

time. For instance, in large coverage areas with fast UE, such as a car traveling at 120 km/h on a highway, the channel's coherence time at 2 GHz carrier frequency drops to around 1 ms. This makes the system to recalculate the channel 1000 times per second to track the UE as it moves and limit the multiplexing gain to a lesser number of terminals. Equally, in more controlled environments with little or no mobility such as fixed wireless access, the system can accommodate hundreds of terminals through spatial multiplexing using narrow beams [26].

Massive MIMO Implementation at Sub-6 GHz and mmWave bands: Massive MIMO was originally conceived for sub-6GHz frequencies, but it is also ideal for mmWave bands, i.e. frequency bands in the range 30-300GHz. Despite conceptual similarities, the way in which massive MIMO can be exploited in these bands is completely different, due to their specific propagation behaviours and hardware characteristics [83].

A key approach to increase the capacity of 5G and future wireless networks is the operation in mmWave bands. There is a huge amount of unused spectrum above 30GHz, which can be used as a complement to current sub-6GHz bands. Even though, path loss and blockage phenomena are more severe in mmWave bands, they can be overcome, at least partially, by keeping the same physical size of the antenna array as on lower frequencies, which is achieved using massive MIMO. However, there are fundamental differences between how massive MIMO technology can be designed, implemented, and exploited in sub-6GHz and mmWave bands [83].

- **First Difference – The Propagation Channel**: The propagation channels of both bands build on the same physics but are significantly different in terms of basic phenomena such as diffraction, attenuation, and Fresnel zones [83].

  Sub-6 GHz bands have more favourable propagation channels and spatial correlation, which have been widely studied for single-antenna and small-scale MIMO systems. The propagation relies on path loss, shadowing/large-scale fading, and multipath propagation/small-scale fading [83].

  Propagation in mmWave bands, however, experience higher attenuation and hence require more directivity. In these frequency bands, many objects behave as full blockers, including humans [84], and there is less diffraction. Specific frequencies suffer from absorption by gases with colliding resonance frequencies, such as 60GHz for oxygen. More than 40dB losses have been measured for propagation through windows, which is substantially higher than for sub-6 GHz frequencies. Therefore, outdoor-to-indoor coverage is limited in mmWave bands. Significant outdoor losses due to vegetation have also been observed [85]. Also, rain will cause higher attenuation with increasing frequencies, but its impact on the link budget is minor. Therefore, because of these unfavourable propagation effects, link budget is worse at mmWave bands than at sub-6 GHz, even if the physical sizes of base station antenna arrays are kept the same in both bands [83].

- **Second Difference – Hardware Implementation**: The hardware implementation architecture changes with increasing carrier frequency. Even though, more antennas can be integrated into a given area, however insertion losses, intrinsic power-overhead in RF generation, and amplification result in diminishing gains [83].

  Massive MIMO systems process a large number of antenna signals. Connecting these signals is the main hardware implementation challenge. This bottleneck can be circumvented in sub-6 GHz systems by bringing all individual signals to the same level using distributed processing [83],[86]. In mmWave systems, however, the connections to the antennas become extremely lossy because micro-strip lines behave as antennas, leading to losses of several dB/cm [87]. Hence, these systems will only benefit from more antennas if they can be integrated in a very compact way [83]. Interconnects are, therefore, the main bottleneck to exploiting the high bandwidth that can be achieved through the integration of many small antennas [83].

- **Third Difference – Signal Processing Algorithms**: The two previous differences, channel propagation and hardware implementation, have major influence on the algorithms needed for channel estimation, beamforming, and resource allocation in both sub-6 GHz and mmWave bands. At sub-6 GHz, channel estimation requires a lot of resources while beamforming is straightforward. On the other, mmWave channel estimation and beamforming are theoretically simpler since there are fewer propagation paths but become challenging if hybrid beamforming is used [83].

4.3.7 Massive MIMO for 5G

With the aim of using the spectrum more efficiently and serving more users, 5G NR plans to take full advantage of massive MIMO technology. Massive MIMO adds multiuser capabilities to MIMO by exploiting the distributed and uncorrelated spatial location of those multiple users. In this configuration, the gNB sends the CSI-RS to UE in the coverage area and based on the SRS response of each UE device, the gNB computes the spatial location of each receiver. The streams of data intended for each receiver go through a precoding matrix (W-Matrix), where the data symbols get combined into signals streaming to each of the elements of the gNB's antenna array (Figure 48) [26],[88]

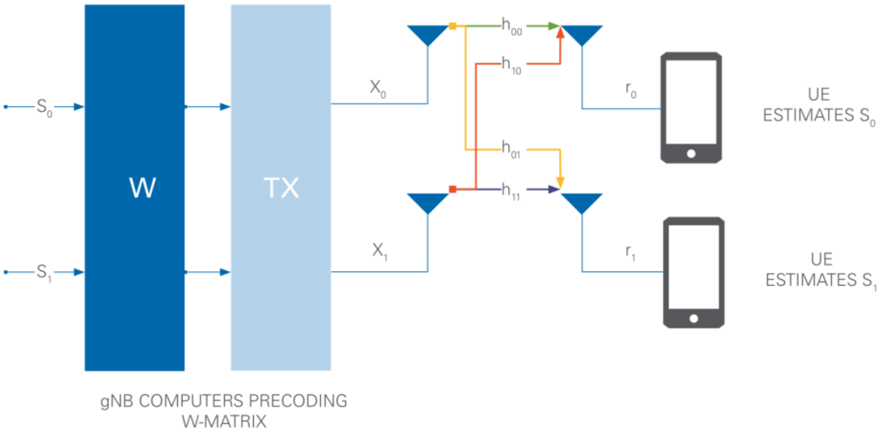

Figure 48: Representation of multi-user MIMO on the downlink

The multiple data streams have their own independent and appropriate weightings that apply different phase offsets to each stream so that the waveforms interfere constructively and arrive in phase at each receiver. This maximises the signal strength at each user's location while giving minimum signal strength in the directions of the other receivers, as Figure 49 indicates [26].

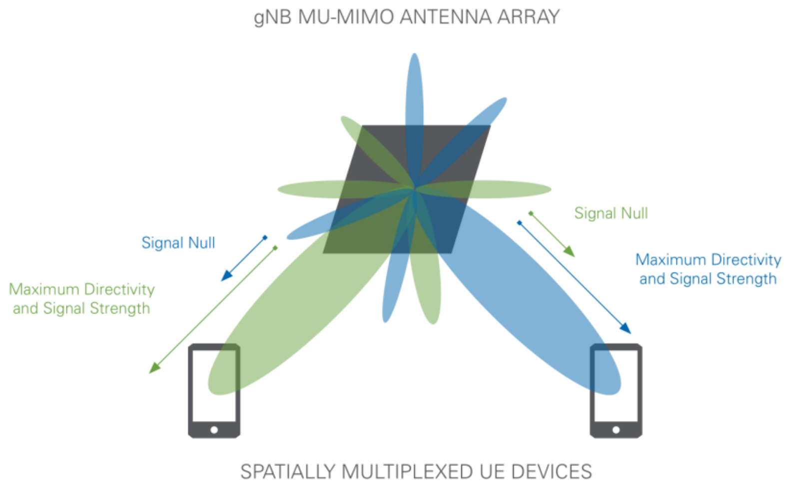

Figure 49: MIMO beamforming for spatial multiplexing

Consequently, the gNB communicates with multiple devices independently and simultaneously, effectively multiplexing them in space. As an extra benefit, in this massive MIMO implementation, the devices do not need any information about the channel or additional processing to get their data streams [26].

Massive MIMO on the DL enhances 5G NR system's capacity. The large number of gNB antennas relative to the number of devices can produce huge gains in spectral efficiency. In such conditions, the system can serve many more devices simultaneously within the same frequency band compared to 4G systems (Figure 50) [26].

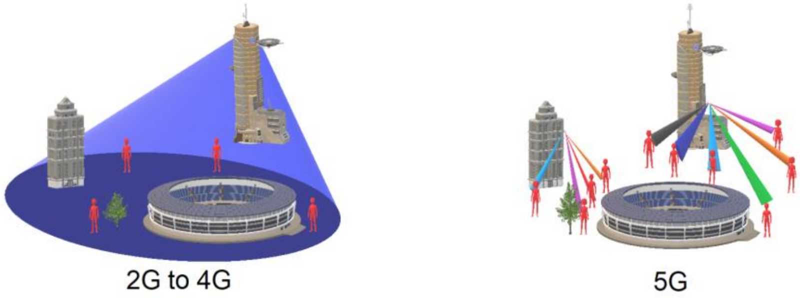

Figure 50: Multiantenna array for massive MIMO

Currently, the strongest case for massive MIMO operation is at sub-6 GHz frequencies. Spectrum is scarce and valuable in this region. In these bands, Massive MIMO systems can attain significant spectral efficiency by spatially multiplexing many terminals. The system can also achieve superior energy efficiency by exploiting large antenna array gains to reduce the amount of power that each front end must handle [26].

In massive MIMO systems, each antenna has its own RF and digital baseband chain. The gNB retains tight phase control and processes the signals from all antennas. Using digital processing, the system has a fuller picture of the channel response on the UL and responds quickly to changes in the channel. Massive MIMO operates mainly in TDD, which permits the assumption of channel reciprocity. That allows the system to estimate DL channels from UL pilots and removes the need for prior knowledge of the channel [26].

Another benefit of massive MIMO systems is that they will provide better and more reliable service to all UE in a coverage area. Because of an improved link budget and the ability to position a target UE precisely within the radiated beam while nulling a non-target UE, power control algorithms can achieve greater fairness among the UE [26].

In conclusion, massive MIMO systems have a huge potential as a key enabling technology for 5G networks. The technology offers huge advantages in terms of energy efficiency, spectral efficiency, robustness, and reliability. It allows for the use of low-cost hardware at both the base station and the UE. At the base station, expensive and powerful, but power-inefficient, hardware are replaced with parallel, low-cost, low-power units that operate coherently together. However, there are still challenges ahead to realise the full potential of massive MIMO technology, in terms of computational complexity, use of distributed processing algorithms, and synchronisation of antenna units [78].

4.4 Semiconductor Technology

Advancements in complementary metal oxide semiconductor (CMOS) semiconductor technology has facilitated the utilisation of mmWave spectrum bands and the integration of components like mixers, low noise amplifiers (LNA), power amplifiers (PA), and IF amplifiers in the same package [50]. Specifically, it has become possible to develop these two kinds of devices:

Device for Low Power Consumption: The use of RFICs is one of the key elements for communication systems operating at mmWave frequency bands. RFICs offer highly integrated solutions with benefits of low power consumption, small form factor, and low cost. They are manufactured using a cost-effective CMOS process [50].

Device for High Gain Beamforming: Next evolution in mmWave technology is modular antenna arrays [89],[90], consisting of large number of sub-array modules. Evidently, the traditional discrete frontend technology cannot be used for mmWave purposes because of weight, volume, and cost. The only viable option is the use of microwave monolithically integrated circuits (MMICs). From a simplified point of view, it is vital that the number of components on the printed circuit board (PCB) be kept to a minimum as losses in RF chip interconnection can become very high easily at mmWave frequencies. Hence, the level of integration should be as high as possible. In the modular antenna array, each module has built-in sub-array phase control and coarse beam steering capability. Its flexible and scalable architecture accomplishes a broad range of antenna gain [50].

4.5 Network Densification

Two approaches have been proposed for 5G network densification: deployment of small cells and device-to-device (D2D) communication. They are explained in more detail in the following sections.

4.5.1 Small Cells

A simple but extremely effective way to increase the network capacity is to make cells smaller. This approach has been demonstrated over several cellular generations [91],[92]. Cell shrinking has numerous advantages, the most important being the reuse of spectrum across a geographic area and the consequent reduction in the number of users competing for resources at each base station. In principle, cells can be shrunk almost indefinitely without a sacrifice in signal-to-interference ratio (SIR), until virtually every base station serves a single user (or is idle). This lets each base station to devote its resources, as well as its backhaul connection, to an ever-decreasing number of users [8].

Therefore, network densification through the deployment of large number of small cells is considered as one of the most effective ways for providing increased system spectral efficiency and satisfying the explosive traffic demand. The system capacity per square kilometre can be almost linearly increased as the number of deployed small cells increases if suitable interference management techniques are exploited [80].

From MNOs' perspective, small cells will offer improvement of capacity, especially in urban environments. Crucially, they would integrate seamlessly with existing network deployment without causing undue interference. Also, small cell networks equipped with self-organising, self-optimising, and self-healing capabilities will reduce the operational expense (OPEX) required to maintain and operate dense deployments [80].

The main distinctive attribute of a small cell is its range, which is typically around 10 to 200 m under NLoS conditions. This is much shorter than the range of a macro cell which might be several kilometres [50],[93].

They can be managed or unmanaged. Managed small cells are those that are deployed under the control of the MNO. Unmanaged small cells are those deployed by end users, such as home base stations. More so, it is beneficial to have a self-organising network (SON) capability for the vastly deployed small cells [50].

Furthermore, small cells can be deployed indoors or outdoors. When deployed outdoors, they are usually placed at a lower height than a macro cell (e.g. on street lampposts) and with lower transmit power to serve a targeted area. Therefore, a lot of small cells are needed especially in dense urban areas where more obstacles of signal propagation exist and where the mobile traffic keeps increasing [50].

They are deployed with one of two primary targets (Figure 51):

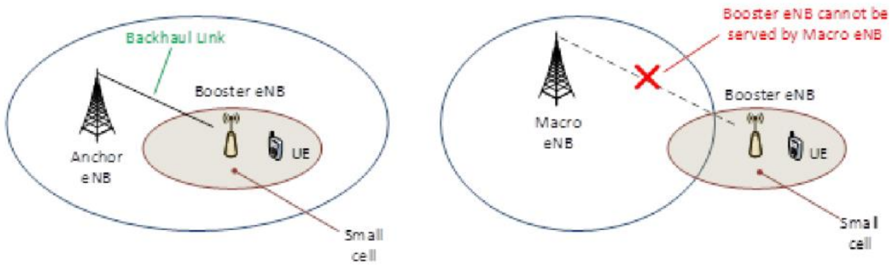

Figure 51: Capacity improvement (left) and coverage extension (right)

- **Coverage extension**: Small cells can be deployed at the edge of a macro cell to extend the coverage of a cellular communication system. The coverage of both cells may partially overlap. This type of small cell is designed to enhance user perceived experience in terms of service availability, and not primarily designed for targeted capacity. It can be deployed both indoors and outdoors and can be thought of as a range extension for macro cells, where peripheral coverage areas at cell edge require QoS and enhanced data throughput [50].
- **Capacity improvement**: Small cells can be deployed within the coverage of a macro cell to improve data throughput of a cellular communication system. Usually coverage of both cells overlap to a large degree [50].

In terms of deployment types, three categories of small cell deployment scenarios have been proposed [93]:

- **Hotspot**: This is a type of small cell deployed to ease congestion from the macro cell, within the macro cell coverage. It, therefore, provides targeted capacity in areas with high traffic density [50].
- **Indoor**: Indoor small cells are often deployed at indoor public spaces with steady daily, non-mobile traffic, and occasional peaks within the enclosed structures such as hotels and office spaces, that are often isolated from outdoor macro cell

coverage. As such, indoor small cells provide coverage enhancement. Indoor scenarios could be further divided into a) large indoor area such as shopping malls, airports, stadiums, etc. and b) multi-room scenario such office buildings [50].

- **Outdoor**: Outdoor small cells are deployed to provide coverage and/or capacity as a complement to macro cell coverage or in isolation of the macro cells such as in disaster recovery support and in rural areas. Although small cells are typically mounted on street facilities, at a lower height than macro cells, however in some scenarios their beams may be directed at higher locations, such as window panes (e.g. in order to cater for indoor coverage with or without additionally deployed repeaters). Outdoor scenarios could be further divided into a) contiguous coverage, b) non-contiguous coverage, and c) backhaul and fronthaul [50].

4.5.2 Device-to-Device (D2D) Communication

Besides cell shrinking, a second approach to densification exists in the form of direct D2D communication. This lets users in close proximity to establish direct communication, replacing two long hops via the base station with a single shorter hop. This can bring about lower power consumption and/or higher data rates, and a reduced latency [94]–[96].

The ever-increasing demand for higher data rates and capacity requires unconventional thinking for the next generation, 5G cellular systems. That is where cooperative communications proves to promising solution. It represents a new class of wireless communication techniques in which network nodes support each other in relaying information to realise spatial diversity advantages. This new transmission concept promises significant performance gains in terms of link reliability, spectral efficiency, system capacity, and transmission range [13].

In the first four generations of cellular networks, D2D communication functionality was been considered. This is mainly because it was seen as a tool to reduce the cost of local service provision, which used to be fractional in the past based on the MNOs' market statistics. The attitude of MNOs toward D2D functionality has been changing recently because of several trends in the wireless market [97]. For instance, the number of context-aware services and applications is growing rapidly. These applications require location discovery and communication with neighbouring devices, and the availability of such a functionality would lower the cost of communication among devices. D2D functionality can also play a crucial role in mobile cloud computing and facilitate effective sharing of resources (spectrum, computational power, applications, social contents, etc.) for users who are spatially close to each other. MNOs can take further advantage of D2D functionality to take some load off of the network in a local area such as a stadium or a big mall by permitting direct transmission among mobile phones and other devices. Furthermore, D2D communication can be of critical use in natural disasters. In an earthquake or hurricane, an urgent communication network can be quickly set up using D2D functionality, replacing the damaged communication network and Internet infrastructure [13].

Currently, technologies such as WiFi or Bluetooth provide some D2D communication functionality. However, they operate in unlicensed bands, and the interference is unmanageable. In addition, they cannot provide security and QoS guarantee like cellular networks. Not willing to lose the emerging D2D market, the MNOs and vendors are exploring the possibilities of introducing D2D communication capability into cellular networks [13].

Consequently, a two-tier 5G cellular network consisting of macro cell and device tiers has been proposed. The macro cell tier involves base station-to-device communications as in a conventional cellular system. The device tier involves D2D communications. If a device connects the cellular network through a base station, it is said to be operating in the macro cell tier. Whereas if it connects directly to another device or achieves its transmission through the assistance of other devices, these devices are said to be in the device tier. In such a system, the base stations will continue to serve the devices as usual. However, at cell edges or congested areas, devices will be permitted to communicate with each other, creating an ad hoc mesh network [13].

In the implementation of device-tier communications, the MNO might have different levels of control. Depending on the business model, it either exercises full or partial control over the resource allocation among source, destination, and relaying devices, or prefers not to have any control. Therefore, the following four types of device-tier communications are possible (Figures 52–55) [13]:

- **Device relaying with operator-controlled link establishment (DR-OC)**: A device at the edge of a cell or in a poor coverage area can communicate with the base station by relaying its information via other devices. This allows the device to achieve a higher QoS or more battery life. The MNO communicates with the relaying devices for partial or full control link establishment [13].

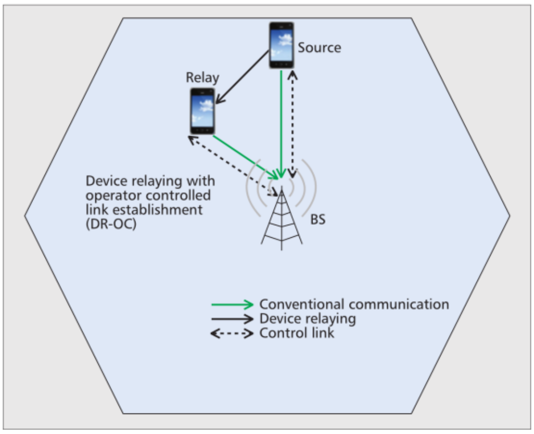

Figure 52: Device relaying with operator-controlled link

- **Direct D2D communication with operator-controlled link establishment (DC-OC)**: The source and destination devices communicate and exchange data with each other without the need for a base station, but they are assisted by the MNO for link establishment [13].

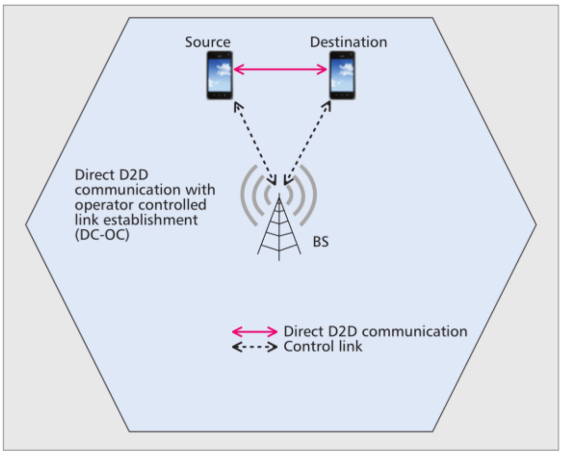

Figure 53: Direct D2D communication with operator-controlled link

- **Device relaying with device-controlled link establishment (DR-DC)**: The MNO is not involved in the process of link establishment. Therefore, source and destination devices coordinate their communication using relays between themselves [13].

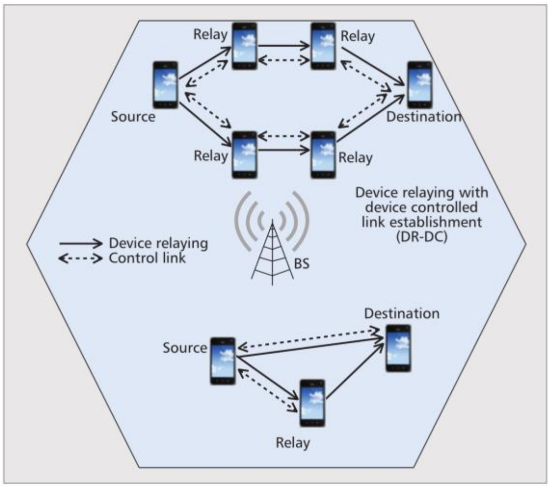

Figure 54: Device relaying with device-controlled link

- **Direct D2D communication with device-controlled link establishment (DC-DC)**: The source and destination devices have direct communication with each other without MNO control. Therefore, source and destination devices use the network resource in a manner that ensures limited interference with other devices in the same tier and the macro cell tier [13].

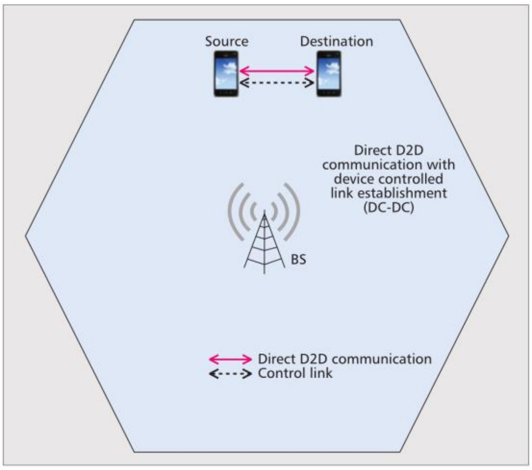

Figure 55: Direct D2D communication with device-controlled link

4.6 Network Function Virtualisation (NFV)

NFV enables network functions that were traditionally tied to hardware appliances to run on cloud computing infrastructure in a data centre. The main benefit will be the ability to elastically support network functional demands. Furthermore, this new architecture will provide agility through the creation of virtual networks and of new types of network services [98].

NFV replaces network functions on dedicated appliances like routers, load balancers, and firewalls, with virtualised instances running on commercial off-the-shelf hardware thereby lowering the cost of network changes and upgrades [7].

The NFV architecture [99] when applied on the 5G Core (5GC) network defines how virtualised software functions can share common physical resources of compute, storage, and networking through the creation of virtual machines (VMs). The VMs are instantiated either statically or dynamically through control functions defined in the Management and Orchestration (MANO) layer of the NFV framework [21].

Although it is theoretically possible to virtualise all Functional Entities (FEs) in the architecture and implement them on virtual machines, it may not always be the most ideal approach. Current research suggests that the next generation core network will comprise of both Virtualised Network Functions (VNFs) and Physical Network Functions (PNFs) [21].

4.7 Software Defined Networking (SDN)

SDN is a framework for creating intelligent programmable networks. Specifically, it is defined as an architecture where the control and data planes are decoupled, network intelligence and state are logically centralised, and the underlying network infrastructure is abstracted from the application [100]. The benefits of this architecture are the logical decoupling of the network intelligence to separate software-based controllers, exposing the network capabilities through an application programming interface (API), and allowing the application to request and manipulate services provided by the network [101].

The underlying motivation of SDN is programmability of networks. This is accomplished through the decoupling of the control and user plane. A highly scalable, distributed, stateless forwarding plane, is programmed with flow tables specifying how packets are to be treated. The forwarding tables are populated by a centralised control plane entity which supports functions like mobility management, policy, subscription control and is able to maintain end-to-end path information for every service that the network supports [21].

SDN allows the dynamic reconfiguration of network elements in real-time, enabling 5G networks to be controlled by software rather than hardware, improving network resilience, performance and QoS [7].

Furthermore, a clear separation between control plane and user plane functions leads to the following FEs defined for the 5G core: Mobility Management Control Function (MMCF), Session Management Control Function (SMCF), Policy Function (PF), Subscriber Database Function (SDBF), Authentication Function (AuF), Application Functions (AF) and User Plane Function (UPF) [102]. The resulting architecture is shown in Figure 56. The interface between the key FEs is called NGx, x being a number [21].

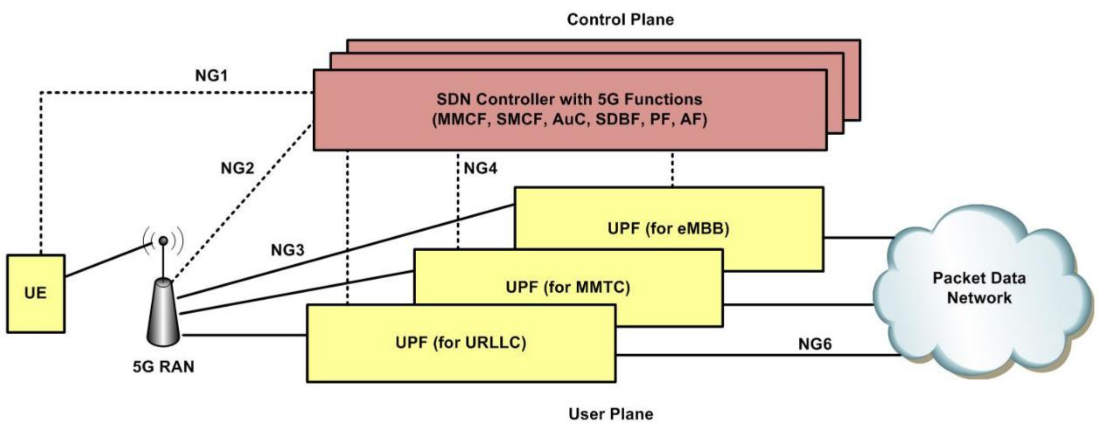

Figure 56: SDN-enabled 5G core network

In conclusion, the introduction of SDN in the transport network of 5G will make it possible to move from a collection of multi-technology networks to a single transport network, where Internet Protocol (IP), Optical and IP Microwave networks will be all combined and controlled by an SDN controller. This will permit the integration of different technologies from multiple vendors into a single end-to-end transport network. The benefits of SDN in the transport network include intelligent and automatic transport connectivity provision, monitoring, performance, troubleshooting and network optimisation. It also automates many functions which are currently done manually by operational staff. Lastly, SDN will facilitate the introduction of network slicing and new service categories required in 5G [16].

5. 5G ARCHITECTURE

5.1 5G Next Generation System (NGS) – An Introduction

The 5G NGS was introduced in 3GPP Releases 15 and 16, marking a radical departure from 4G and other previous networks. Besides specifying the 5G NR and RAN (with support for new frequency bands including mmWave frequencies up to 100 GHz), Release 15 defines a new, service-based network architecture that is designed with: network function virtualisation in mind; support for features including automation and orchestration, network discovery; support for multiple access mechanisms with a common control structure (including bringing in non-3GPP access networks); and network slicing to create multiple logical networks on the same physical network [18].

Release 16 is primarily focused on fixing, enhancing, and enabling new services on 5G NR. The primary features being developed are MIMO enhancements to increase efficiency in the mmWave bands, expanding vehicle-to-everything (V2X) beyond what is available using 4G LTE, 5G NR positioning enhancements to meet both commercial and regulatory requirements for voice services, enhancements to base station interference mitigation, improvements in UE power savings, additional Industrial IoT features, improvements to handover performance, studies on operation above 52.6 GHz, SON improvements, 5G NR carrier aggregation and dual connectivity improvements, defining UE testing in mmWave bands, 5G NR in unlicensed bands, and solutions for in-band backhaul [18].

The 5G NGS is composed of 5G Core (5GC), NextGen RAN (NG-RAN) and Transport (Fronthaul and Backhaul) networks. The 3GPP's architecture of the 5G network is shown in Figure 57 below [103].

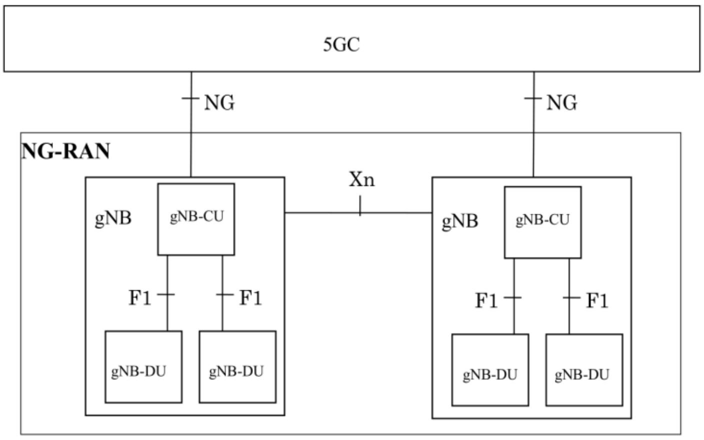

Figure 57: 3GPP NG-RAN architecture

The NG-RAN is comprised of a set of gNBs connected to the 5GC via the NG interface. The gNBs can be connected through the Xn interface. A gNB may consist of a gNB-Centralised Unit (gNB-CU) and gNB-Distributed Unit (gNB-DU). The CU processes non-real time protocols and services, and the DU processes physical (PHY) layer protocol and real time services. The gNB-CU and the gNB-DU units are connected via F1 logical interface. One gNB-DU is connected to only one gNB-CU. For flexibility, a gNB-DU may also be able to connect to another gNB-CU, if the primary gNB-CU fails, by appropriate implementation. NG, Xn and F1 are logical interfaces [104].

Furthermore, fronthaul is the network between Remote Radio Unit (RRU) and DU connected via the Common Public Radio Interface (CPRI) and/or enhanced CPRI (eCPRI). On the other hand, midhaul is the network between DU and CU connected via the F interface while backhaul is the network between CU and 5GC (NG interface) and between CUs (Xn interface). In some instances, CU and DU are co-located and form the gNB. In such scenarios, RRU to gNB is the fronthaul and gNB to 5GC is the backhaul. Moreover, the fronthaul would usually be based on lower-layer functional split while the midhaul be based on higher-layer functional split [104].

The following sections describe the individual components of the 5G NGS in detail.

5.2 5G Core (5GC) Network

There are two types of Core Networks:

- **Circuit-Switched Core** mainly used for voice in 2G and 3G.
- **Packet-Switched Core** mainly used in 2G, 3G (data only), 4G (data/voice) and 5G (data/voice).

In the Circuit-Switched Core, there are dedicated connections primarily to carry voice between users, and in the Packet-Switched Core, the communication is done through IP packets that are more efficient and support the move from Circuit Switched technology towards the move ubiquitous technology of IP [16].

In existing 4G mobile networks, the radio network works based on the Evolved LTE standard and the core network is referred to as the 4G Evolved Packet Core (EPC). In pure 5G, the equivalent standards will be 5G NR and a new 5G Core network (5GC). 5GC is a new architecture that has been defined from scratch by 3GPP. As well as being able to serve 5G NR, it introduces more flexibility, more openness and new protocols and will be able to function as a software defined network. It will offer a unified authentication framework, unified subscription control, unified QoS framework and charging, and will support the creation of network slices [16].

In 4G Evolved Packet System (EPS), the core network, EPC, is comprised of Mobility Management Entity (MME), Home Subscriber Server (HSS), Serving Gateway (SGW), PDN Gateway (PGW), and Policy and Charging Rules Function (PCRF) as illustrated in Figure 58 [2].

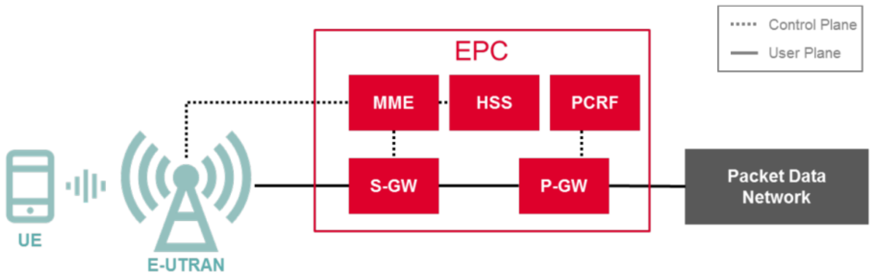

Figure 58: EPC architecture

While MME handles the control plane, SGW and PGW handle both control and user plane. PCRF is also an integral component of EPC that provides rules for policy (e.g., QoS) and charging. The EPC handles EPS Mobility Management (EMM) procedures and EPS Session Management (ESM) procedures [2].

On the other hand, 5GC, the core network of the 5G Next Generation System (NGS), proposes a Service-Based Architecture (SBA) that would provide unprecedented efficiency and flexibility for the network. SBA is an architecture for building system based on fine-grained, interaction of loosely coupled and autonomous components called services. This architectural model is designed to take full advantage of the latest virtualisation and software technologies [2].

Compared to EPC, the elements of 5GC's service-based architecture are designated as network functions (NFs) [2]. The NFs offer their services to any other applicable network functions through a common framework of interfaces accessible to all network functions. Network Repository Functions (NRFs) allow every network function to discover the services offered by other NFs present in the network whereas Network Exposure Functions (NEFs) permit NF discovery within the network, and also securely to external networks. The goal of this model to maximise the modularity, reusability, and self-containment of NFs, and at fostering the ability to grow flexibly while taking advantage of the latest virtualisation technology [18].

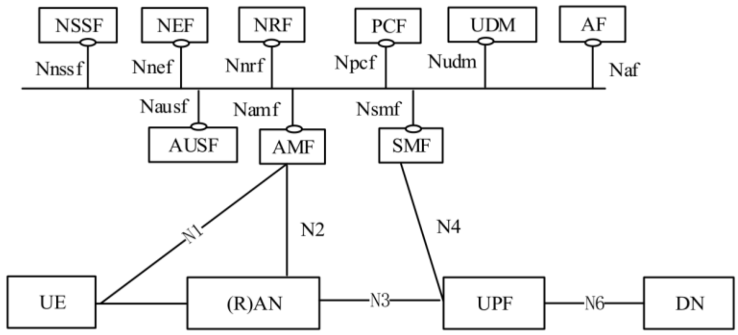

Figure 59: Service-based view of the 5G NGS architecture

Figure 59 depicts the service-based view of the 5G Next Generation Network, including the following NFs [2],[18]:

- AF: Application Function
- AMF: Access and Mobility Management Function
- AUSF: Authentication Server Function
- NEF: Network Exposure Function
- NRF: Network Repository Function
- NSSF: Network Slice Selection Function
- PCF: Policy Control Function
- SMF: Session Management Function
- SMSF: SMS (Short Message Service) Function
- UDM: User Data Management
- UPF: User Plane Function

A service (NF) is an atomised capability in a 5G network, characterised by high-cohesion, loose-coupling, and independent management from other services. This lets individual services to be updated independently with minimal impact to other services and deployed on demand. A service is managed based on the service framework including service registration, service authorisation, and service discovery. It offers a comprehensive and highly automated management mechanism implemented by NRF, which greatly reduces the complexity of network maintenance. A service will interact with other services in a light-weight manner, for instance via API invocation [2].

With NFV technologies, the NFs can be virtualised and hosted in a cloud environment. The physical boundary between traditional EPC network components such as MME, SGW, and PGW will be blurred with virtualisation and software. This has fostered the redesign of the 5GC to be open and flexible enough to meet the diversity of service and business requirements in the 5G era [2].

Hence, SBA is the natural step that allows 5G network functionality to become more granular and decoupled, bringing the following benefits to 5G:

- The network is highly efficient in rolling out new network features, allowing MNOs to quickly deploy new business and services and reduce the time to market for installing bug fixes and rolling out new network features [2].
- The network is extensible. Each service can interact directly with other services with a single light-weighted service-based interface. In that way, the service-based interface can be easily extended without introducing new reference points and corresponding message flows [2].
- The network will be modular and support reusability. It is composed of modularised services, which reflects the network capabilities. A service can be easily invoked by other services with appropriate authorisation, enabling each service to be reused as much as possible [2].
- The network is easily open. The information about a 5G network can be easily exposed to external users such as 3rd parties via a specific service without complicated protocol conversion [2].

The 5G SBA can be broadly mapped to a generic framework shown in Figure 60. It is composed of three main layers: the infrastructure layer, the network function layer, and the service (or business) layer. It also has a MANO entity that translates use cases and service models into network slices by connecting NFs, mapping them to infrastructure resources, configuring and monitoring each slice during its lifecycle [105].

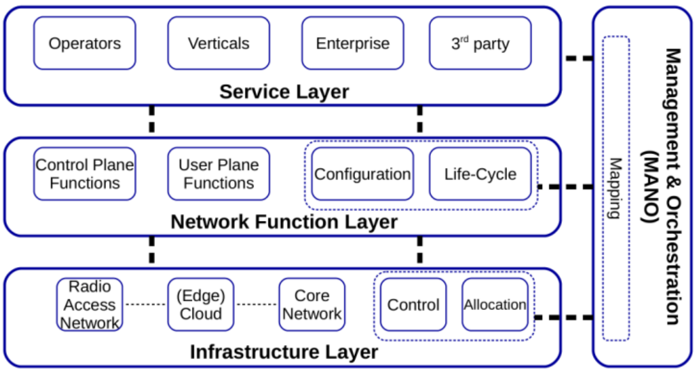

Figure 60: 5G service-based architecture

- **Infrastructure Layer**: The infrastructure layer broadly refers to the physical network infrastructure spanning both the RAN and the Core network. It also incorporates deployment, control, and management of the infrastructure; the allocation of resources (computing, storage, network, radio) to slices; and the manner in which these resources are revealed to and can be managed by the higher layers [105].
- **Network Function Layer**: The network function layer encapsulates all the operations that are related to the configuration and lifecycle management of the NFs that, after being optimally placed over the (virtual) infrastructure and chained together, provide an end-to-end service that satisfy certain requirements and constraints described in the service design of the network slice [105].

- **Service Layer and MANO**: Perhaps the most important element that distinguishes network slicing, in the context of 5G, from other forms of slicing that have been considered in the past, for instance cloud computing, is its end-to-end nature and the requirement to express a service through a high-level description and to flexibly map it to the appropriate infrastructural elements and NFs. This observation naturally leads to two new high-level concepts: (1) a service layer that is directly linked to the business model behind the creation of a network slice; and (2) network slice orchestration for the hypervision of a slice's lifecycle [105].

In conclusion, while EPC could be considered an evolution of previous generation packet core networks, the 5GC has been designed from its inception to be "cloud native", that is inheriting many of the technology solutions used in cloud computing and with virtualisation at its core. It also offers superior network slicing and QoS features. Another important characteristic is the separation of the control plane and user plane that adds flexibility in connecting the users, allows an easier way to support a multitude of access technologies and provides better support for network slicing and edge computing [2].

5.3 NextGen Radio Access Network (NG-RAN)

In a RAN, a radio access technology is used to connect a mobile phone to the radio mast. Each mast serves an area referred to as a cell. Cells are organised like the cells in a honeycomb hence the term "cellular networks". They work with each other to minimise radio interference through the careful allocation of radio channels. In rural areas, the cells can cover several kilometres, but in busy traffic areas, cells are subdivided into smaller cells to provide more capacity [16].

Existing 2G, 3G and 4G networks use different radio access technology. 5G NR is a new radio access technology offering data rates higher than 1Gbps. It can transmit more data in the same amount of spectrum while utilising more spectrum [16].

As illustrated in Figure 61, the main change in 5G NR, compared to 4G LTE, is that the original BBU function in 4G LTE is split into three parts: Central Unit (CU), Distributed Unit (DU), and Remote Radio Unit (RRU) [104]. This split function architecture enables a separation of User Plane (UP) and Control Plane (CP) in the NG-RAN thereby optimising the handling of 5G services, like ultra-low latency and ultra-high bandwidth [18].

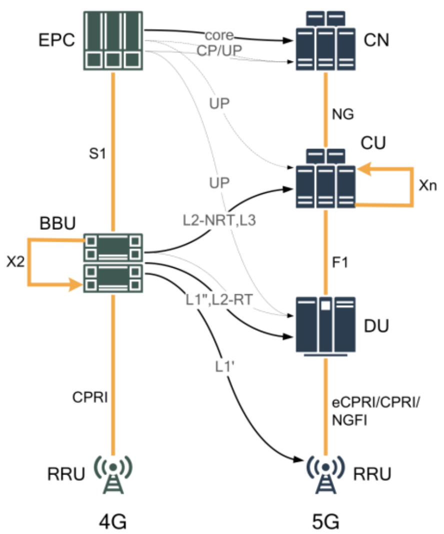

Figure 61: Evolution from single-node (4G) to split function architecture (5G)

Figure 62 depicts the evolved NG-RAN architecture [18]:

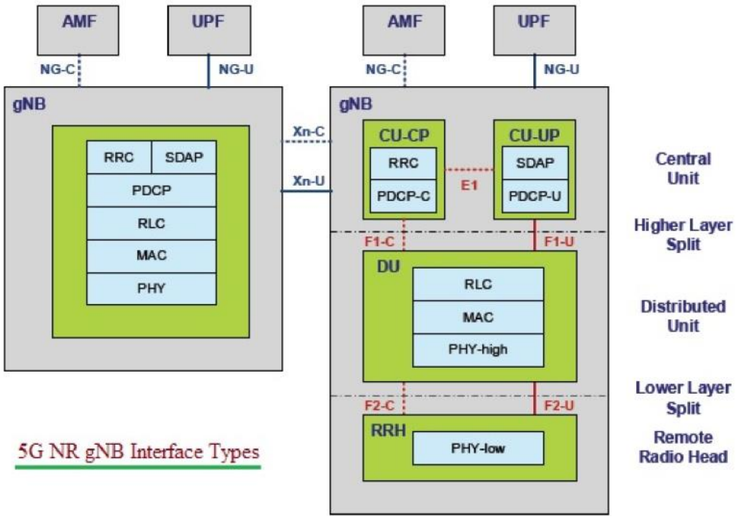

Figure 62: 5G NG-RAN architecture

It includes the following logical NG-RAN nodes, interfaces, and functions [18]:

- **Central Unit–User Plane (CU-UP)**: This is the Packet Processing Function (PPF) which encompasses user plane functions that are asynchronous to the Hybrid Automatic Repeat Request (HARQ) loop and includes the Packet Data Convergence Protocol (PDCP) layer such as encryption and the multipath handling function for the dual connectivity anchor point and data scheduling. CU-UP interfaces with the User Plane Function (UPF) of 5GC via the NG3 interface.
- **Central Unit–Control Plane (CU-CP)**: This is the Radio Control Function (RCF) which manages load sharing among system areas and different radio technologies, as well as the use of policies to control the schedulers in the Beamforming Processing Functions (BPFs) and Packet Processing Function (PPFs). At the user and bearer level, the CU-CP negotiates QoS and other policies with other domains and is responsible for the associated service level agreement (SLA) enforcement in the NG-RAN. It controls the overall NG-RAN performance in terms of service requirement, creates and manages analytics data, and is responsible for the NG-RAN SON functions. CU-CP interfaces with the Access and Mobility Management Function (AMF) of 5GC via the NG2 interface.
- **Distributed Unit (DU)**: This logical node includes a subset of gNB functions, depending on the functional split option. Its operation is controlled by the CU.
- **F1 Interface**: The F1 interface connects a CU and a DU. It is an open interface that supports the exchange of signalling information between endpoints. In addition, it supports data transmission to the respective endpoints. The interface supports Control Plane (F1-C) and User Plane (F1-U) separation. It also enables exchange of UE and non-UE associated information.
- **E1 Interface**: E1 is the point-to-point interface between a CU-CP and a CU-UP. This is an open interface that supports exchange of signalling information between the end points. It is a control interface and is not used for user data forwarding.
- **Xn interface**: Xn supports the exchange of signalling information between two NG-RAN nodes and the forwarding of Protocol Data Units (PDUs) to the

respective tunnel endpoints. From a logical perspective, the Xn is a point-to-point interface between two NG-RAN nodes. It supports procedures over the control plane (Xn-C) and user plane (Xn-U) and enables procedures for intra-NG-RAN mobility and dual connectivity between NG-RAN nodes.

5.4 Transport Network

Transport is the term that refers to the network infrastructure between the radio mast and the core network. These networks are also called access and backhaul networks and there could also be a midhaul network, in some cases. Connections are usually optical fibre, but microwave is used where a wireless connection is more suitable. In the previous generations of 2G to 4G, microwave connections have been key technologies for the final connection to the radio mast. With the introduction of 5G, innovative IP Microwave advances including mmWave technology provide the opportunity to support higher bandwidth and lower latency. The use of IP Microwave is expected to be balanced with the use of more optical fibre in the transport coupled with optical multiplexing technologies which have the potential to resolve the new bandwidth and delay requirements of 5G [16].

Fronthaul: Fronthaul is defined as “a network path between centralised radio controllers and remote radio units (RRU) of a base station function” [106]. This architecture permits the centralisation of all high layer processing functions at the expense of the most stringent fronthaul latency and bandwidth requirements. The rise in 5G data rates makes it impractical to continue with the conventional CPRI fronthaul implementation. Allocating more processing function to RRU would relax the latency and bandwidth requirements – but fewer processing functions can then be centralised. It is, therefore, important that the new functional split architecture considers technical and cost-effective trade-offs between throughput, latency, and functional centralisation [7],[104].

Backhaul: Backhaul networks connect the RAN to the core network. The ultra-high capacity, fast speeds, and low latency requirements of 5G require a backhaul network that can meet these high demands. Optical fibre is often considered the most suitable type of backhaul by MNOs due to its longevity, high capacity, high reliability, and ability to support very high capacity traffic [7].

However, fibre network coverage is not ubiquitous in all cities where 5G will be launched and even less so in suburban and rural areas. Installing new fibre networks in these areas might also be prohibitive in terms of cost for MNOs. Hence, a portfolio of wireless backhaul technologies such as point-to-multipoint (PMP) microwave and mmWave, could be used to complement optical fibre. PMP is capable of downstream throughput of 1 Gbps and latency of less than 1 ms per hop over a 2-4 km distance. mmWave has considerably lower latency and is capable of higher throughput speeds [7].

While most emphasis is placed on terrestrial technology, high altitude platform systems (HAPS) and satellite technology also have a role to play in 5G. HAPS and satellite systems (including non-geostationary constellations) can provide very high data rates (> 100 Mbps – 1 Gbps) to complement fixed or terrestrial wireless backhaul networks outside major urban/suburban areas and can deliver video transmission to fixed locations. They can be integrated with other networks to augment 5G service capability and address some major challenges such as the support of multimedia traffic growth, ubiquitous coverage, machine-to-machine communications (M2M), and critical telecom missions [7].

In general, the service of the backhaul transport is multipoint to multipoint whereas the service of both the fronthaul and the midhaul transport networks is point to point, with the assumption that a DU only belongs to one CU at a specific time and a RRU only belongs to one DU. Furthermore, both the fronthaul and the midhaul transport networks should provide a reasonably low latency to satisfy the requirements of latency sensitive services [104].

In terms of transport network deployment, four scenarios have been identified [104]:

- **Independent RRU, CU and DU locations** – In this scenario, there are fronthaul, midhaul and backhaul networks. The distance between an RRU and DU is at most 20 kilometres while the distance between the DU and CU is up to tens of kilometres [104].
- **Co-located CU and DU** – In this scenario, the CU and DU are located together, therefore there is no midhaul [104].
- **RRU and DU integration** – In this scenario, an RRU and DU are installed close to each other, maybe hundreds of metres, for instance in the same building. To reduce cost, an RRU is connected to a DU just through straight fibre and no transport equipment is needed. In this case, there are midhaul and backhaul networks [104].
- **RRU, DU and CU integration** – This network structure can be used for small cells and hotspots. There is only backhaul in this scenario [104].

Figure 63 depicts the transport network architecture for independent CU and DU deployment. Fronthaul transport is between RRUs and DUs, midhaul transport is between DUs and CUs, while backhaul transport is between CUs and Core Network [104].

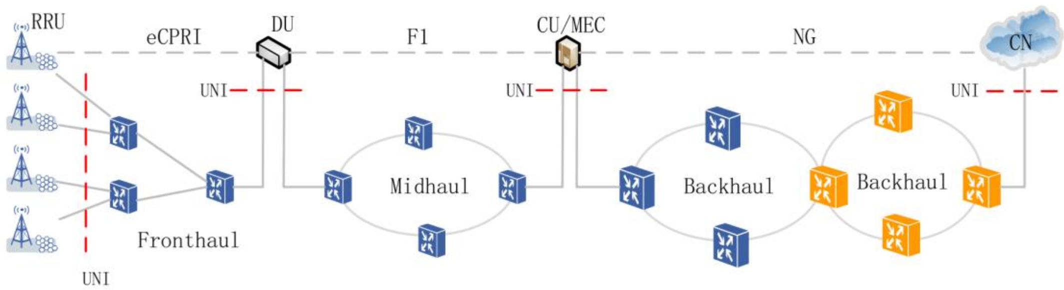

Figure 63: Transport network architecture for independent CU and DU deployment

For other deployments, the transport architecture could also take the following forms [104]:

- For co-located CU and DU deployment, there will be no midhaul transport network. Only the fronthaul and the backhaul will carry 5G traffic.
- For RRU and DU integration deployment, there will be no need of fronthaul transport network. Just the midhaul and the backhaul will transport traffic between RRU/DU to CU and CU to Core Network, respectively.
- For RRU, DU and CU integration deployment, only backhaul transport network will be left. This transport network architecture is almost the same as that of 4G [104].

In terms of network topology for the trio of transport networks [104]:

- For fronthaul transport networks, a star or ring network topology may be used.

- For midhaul transport network, a ring topology is normally used.
- For backhaul transport network, both ring and mesh topology are used [104].

In conclusion, one of the defining features of 5G networks will be end-to-end flexibility.[3] This flexibility is a result of the introduction of network softwarisation where the core network hardware and the software functions are separated. Network softwarisation, using NFV, SDN, network slicing and Cloud-RAN technologies, aims to increase both the pace of innovation and the speed at which mobile networks can be transformed. In addition, edge computing will be important for real-time and latency-sensitive applications. It brings data closer to end-user devices, providing computing power with extremely low latency for demanding applications, thereby speeding up the delivery of actionable data, cutting down on transport costs and optimising traffic routes [7].

[3] ITU: http://news.itu.int/5g-update-new-itu-standards-network-softwarization-fixed-mobile-convergence/

6. DEPLOYMENT OPTIONS AND NEW FEATURES

6.1 Standalone (SA) and Non-Standalone (NSA) Options

Just like previous generations, 3GPP defined both a new 5G core network, referred to as 5GC, as well as a new radio access technology (RAT) called 5G NR. However, unlike previous generations that required that both the access and core networks to be deployed are of the same generation (for instance, EPC and LTE together formed the 4G system), with 5G it is possible to combine elements of different generations in different configurations, namely [2]:

- Standalone (SA) using only one RAT
- Non-Standalone (NSA) combining multiple RATs [2].

The 5G ecosystem currently invests in and plans to roll out these two variants of 5G (Figure 64), both primarily based on 3GPP Release 15 [17].

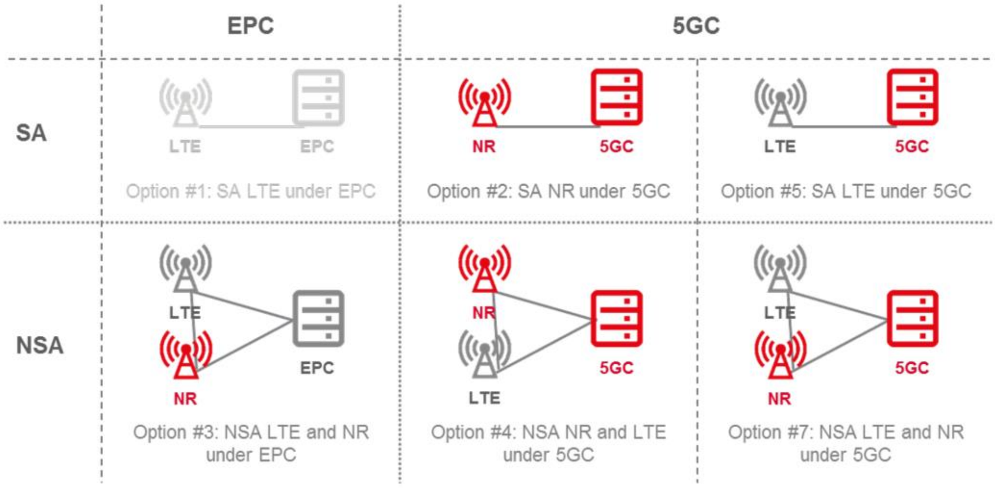

Figure 64: 5G deployment options

In a standalone (SA) scenario, the 5G NR or the 4G evolved LTE radio cells and the core network are operated alone. This means that the 5G NR or 4G evolved LTE radio cells are used for both control plane and user plane. The SA option is a simple solution for MNOs to manage and may be deployed as an independent network using normal inter-generation handover between 4G and 5G for service continuity [2].

3GPP has defined three variations of SA, which are [2]:

- Option 1 using EPC and 4G LTE evolved NodeB (eNB) access (i.e. similar to 4G LTE networks)
- Option 2 using 5GC and 5G NR next generation NodeB (gNB) access
- Option 5 using 5GC and 4G LTE ng-eNB[4] access [2]

In a non-standalone (NSA) scenario, the 5G NR radio cells are combined with 4G LTE radio cells using dual connectivity[5] to deliver radio access. Also, the core network could

---

[4] A version of 4G LTE eNB that is capable of being connected to 5GC

[5] Dual Connectivity is an operation where a given UE consumes radio resources provided by at least two different network points (e.g. 5G NR access from gNB and LTE access from eNB).

be either EPC or 5GC depending on the choice of the MNO. The NSA solution may be chosen by MNOs that want to leverage existing 4G deployments by combining 4G LTE and 5G NR radio resources with existing EPC and/or that want new 5GC to deliver 5G mobile services. This scenario will necessitate tight interworking with the 4G LTE RAN and end-user experience depend on the RAT(s) used [2].

3GPP has also defined three variations of NSA, which are [2]:

- Option 3 using EPC and an 4G LTE eNB acting as master and 5G NR en-gNB[6] acting as secondary
- Option 4 using 5GC and a 5G NR gNB acting as master and 4G LTE ng-eNB acting as secondary
- Option 7 using 5GC and an 4G LTE ng-eNB acting as master and an 5G NR gNB acting as secondary [2]

MNOs must therefore consider the feasibility of different options in meeting their intended initial use cases and interoperability of their choice with other options to ensure their networks deliver the use cases effectively while supporting global interoperability [2].

6.2 New Features

6.2.1 Cloud-RAN (C-RAN) and Functional Split

C-RAN vital to the realisation of 5G networks. It is a cloud-based radio network architecture that uses virtualisation techniques combined with centralised processing units. It offers a means of replacing the distributed signal processing units at mobile base stations and reducing the cost of deploying dense mobile networks based on small cells [7].

In a C-RAN architecture, the base station functions are split into the Remote Radio Unit (RRU) and Baseband Unit (BBU). The RRU is located at the base station site and the BBU is centralised in a data centre facility. Existing C-RAN solutions utilise an optical transmission link between the two components [21]. The main benefits of C-RAN are that it [21]:

- enhances the effectiveness of inter-site scheduling and cooperative techniques (because inter-site signalling is internalised to the BBU pool)
- enables efficiencies through statistical multiplexing gains of combined resources
- facilitates the benefits of NFV to be applied to some parts of the radio protocol stack [21].

The current C-RAN implementation approach adopts the transport of digitised I/Q samples between RRU and BBU. The encapsulation of the I/Q samples onto an optical transmission link is defined in the CPRI specification. The bandwidth needed for CPRI scales linearly with system bandwidth, antenna ports and sampling frequency. With 5G NR system parameters, CPRI line rates of around 12 Tbps will be required. Currently, the maximum line rate for CPRI is 24 Gbps [21],[107].

Furthermore, 3GPP is investigating new approaches of splitting signal processing functions between the RRU and the BBU. Various protocol splits between the RRC-

[6] A version of 5G NR gNB that is capable of being connected to EPC

PDCP-RLC-MAC-PHY layers are being considered. Each option results in different fronthaul bandwidth and delay requirements that vary considerably from tens of Mbps to tens of Gbps, and from 10 ms to 150 µs round trip time. The IEEE P1914 working group is creating specifications for transport of the RRU/BBU payload resulting from the various split options over packet transport networks. Meanwhile, the CPRI specification group has developed an enhanced transport solution called eCPRI which can use Ethernet transport as well as optical transport [21].

The diverse signal processing and protocol functions of a base station cannot be virtualised with a single hardware platform like in the Core Network. It will still rely on application-specific integrated circuits (ASICs) or field programmable gate arrays (FPGAs) for real-time computation of physical layer functions. Higher layer functions like Radio Link Control (RLC) could be implemented with general purpose compute resources. Consequently, the Network Function Virtualisation Infrastructure (NFVI) layer of a BBU datacentre is more complex than that proposed for the Core Network [21].

RAN virtualisation is based on the idea of base station softwarisation, which allows certain RAN functions to run at remote cloud platforms. This paradigm gained momentum with the emergence of the C-RAN concept [108],[109], where RAN functions are split between the BBU, hosted in the cloud, and RRUs that provide antenna equipment and radio access. As already stated, initial C-RAN deployments utilised a high capacity fronthaul network, typically based on optical technology to connect the BBU, that provides the corresponding RAN functionalities, with several RRUs. However, the wide availability of high-speed optical links, especially in urban small cell environments, cannot be guarantee. Hence, flexible solutions were explored aimed at moving some RAN functionalities to a cloud environment, giving rise to the concept of flexible functional split. Figure 65 provides an overview of the different functional split options [110].

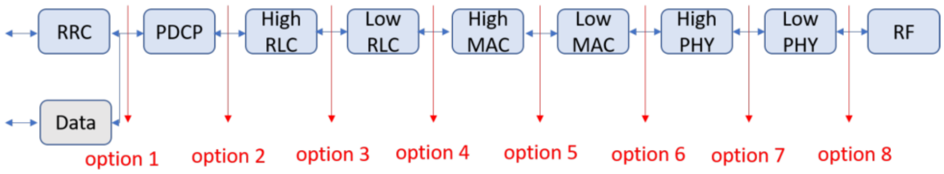

Figure 65: An overview of functional split options

Amongst these options, the most common ones are detailed below [110]:

- **PHY-layer option (option 6)**: this option provides the highest centralisation and can be realised only with an ideal fronthaul, i.e. a high data rate and low-latency optical fiber.
- **MAC-layer option (option 4)**: The MAC layer and the layers above it are virtualised and run on a BBU with real-time scheduling performed aggregately for multiple RRUs. This option leverages the gains of connecting distributed RRU physical layers to a common MAC, which allows coordinated scheduling and dynamic point selection. However, it requires a low-latency fronthaul as some of the MAC procedures are time-critical (e.g. UE scheduling) and need to generate a configuration at the Transmission Time Interval (TTI) level.
- **RLC-layer option (option 3)**: The RLC layer and other layers above it are virtualised at the BBU allowing various MAC entities to be associated with a common RLC entity. This option reduces the fronthaul latency limitations as real-time scheduling is performed locally in the RRU.

- **PDCP-layer option (option 2)**: This option is not time critical. It runs the PDCP functions at the BBU and may use any type of fronthaul network. Its main advantage is the possibility to have an aggregation of different RRU technologies (e.g. 5G NR, 4G LTE, and WiFi) [110].

6.2.2 Multi-access Edge Computing (MEC)

Multi-access edge computing (MEC) architecture is a promising way of supporting latency-sensitive applications. MEC places computing resources close to the edge of the network to facilitate multiple cell-site baseband processing. The physical proximity of computing resources leads to a decrease in transport latency for various latency-sensitive applications that a centralised network cannot support due to longer round-trip transit times [15].

In these applications, the maximum allowed end-to-end latency will be so limited that they have to be moved from the cloud to the edge, i.e. an infrastructure node which is located as close as possible to the devices utilising the application. An option would for instance be to have applications hosted directly on base stations [80].

There are four major contributors to end-to-end latency: the Service Application Processing, the Radio Interface, the Network (Transport and Core), and the Internet Segment. Of these four, Network latency can be reduced using MEC by deploying the Service Application and the Core Network closer to the customer. This also indirectly cancels the Internet Segment latency [16].

MEC is implemented based on a virtualised platform that relies on recent advancements in network functions virtualisation (NFV), information-centric networking (ICN) and software-defined networking (SDN). Specifically, NFV enables a single edge device to provide computing services to numerous mobile devices by creating many virtual machines (VMs)[7] for simultaneously performing different tasks or operating different network functions [111]. Whereas, ICN offers an alternative end-to-end service recognition model for MEC, shifting from a host-centric to an information-centric one for implementing context-aware computing. Lastly, SDN lets MEC network administrators to manage services through function abstraction, achieving scalable and dynamic computing [112]. In conclusion, MEC will enable numerous mobile applications such as Video Stream Analysis Service, Augmented Reality Service, IoT Applications, and Connected Vehicles [9].

6.2.3 Network Slicing

According to 3GPP, network slicing is a technology that "enables the operator to create networks, customised to provide optimised solutions for different market scenarios which demand diverse requirements, e.g. in terms of functionality, performance and isolation" [102],[110].

It permits a physical network to be separated into multiple virtual networks or logical segments that can support different RANs or several types of services for certain customer segments and greatly reducing network construction costs by using communication channels more efficiently [7].

[7] A VM is a virtual computer mapped to the physical machine's hardwares, providing virtual CPU, memory, hard drive, network interface, and other devices

It is expected to play a vital role in 5G networks because of the multitude of use cases and new services 5G will support. These new use cases and services will place different requirements on the network in terms of functionality and performance. By providing virtual or logical network slices over the same physical network, 5G makes it possible to offer specific services to different customers in terms of the performance characteristics they require such as guaranteed bandwidth or low latency and enabling new services and new devices that are dependent on these characteristics. This creation and management of end-to-end network slices, over the radio, transport, and core networks, is facilitated by networking technologies including SDN, NFV, and management via orchestration layers (MANO) [16]. Moreover, by virtualising one physical network into several different logical networks enables the service provider to ensure that one overloaded virtual network does not impact other virtual networks [15].

Furthermore, network slicing will enable value creation for vertical segments, application providers and third parties that lack physical network infrastructure. The VNFs, which constitute a network slice, may vary significantly depending on the service requirements of that specific slice. The type of service associated with a network slice would define the resources and service treatment the network slice would receive [110],[113].

For instance, an autonomous car will rely on V2X communication which requires low latency but not necessarily high throughput. A video streaming service, on the other hand, will require a high throughput and consistent latency in order to ensure an acceptable viewing experience. Both communications streams could be transmitted over the same common physical network but on different virtual or logical network slices. This optimises the use of the physical network and provides the specific characteristics required by the application [16].

According to [114], the network slicing process is broadly divided into three main layers, namely the service instance layer, the network slice instance layer, and the resource layer, as illustrated in Figure 66. Each service instance indicates a service provided by a vertical segment, application provider or MNO while the network slice instance represents a set of resources customised to accommodate the performance requirements of a specific service and may or may not contain one or a number of different sub-network instances that are either isolated or shared. A sub-network instance can be a network function, e.g. IP Multimedia Subsystem, or sub-set of network functions or network resources realising a part of a network slice instance [110].

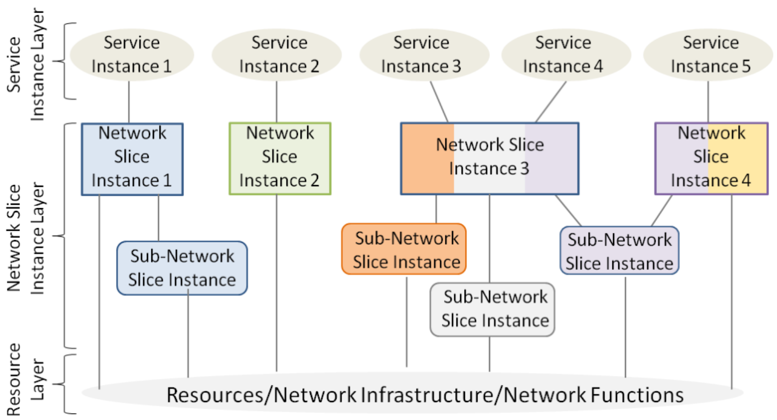

Figure 66: The NGMN network slicing concept

6.2.4 IoT Solutions

Narrow Band Internet-of-Things (NB-IoT) and CAT-M were first introduced in 3GPP Release 13, and both were selected as the IoT solutions for 5G, with improvements in data rates and low latency values in later releases [16].

NB-IoT was created based on the cellular IoT concept in a development led by Vodafone and is now the leading Low Power Wide Area (LPWA) technology globally. It is designed to support services requiring low throughput, extended coverage, and long battery life. Also, it enables very low-cost devices supporting use cases such as car park management, water meters or waste bin management, amongst other uses [16].

On the other hand, CAT-M is a simplified LTE design that complements NB-IoT and supports real-time voice, mobility, lower latency, and higher throughput in a smaller coverage area compared to NB-IoT. It considered as the best solution for applications such as wearables, electricity meters or elevator emergency services [16].

6.2.5 Integrated Access and Backhaul (IAB)

The IAB feature adds support for wireless backhauling of base stations in 5G NR. It is regarded as an enabler for further network densification without the need for fibre implementation in every base station. The solution is intended to allow flexible deployment of 5G NR base stations (called IAB nodes) that utilise the larger bandwidth on higher frequencies bands for wireless backhaul. Its architecture is based on the gNB split architecture where the gNB-DU functionality is terminated in the IAB node, while the gNB-CU functionality is terminated in a donor node. The IAB node re-uses existing procedures defined for the UE to connect to the donor node. IAB has minimal impact on the core network [17]. In summary, the main goals of IAB are to:

- improve capacity by supporting networks with a higher density of access points in areas with only sparse fibre availability.
- improve coverage by extending the range of the wireless network and by providing coverage for isolated coverage gaps. For instance, if a UE is behind a building (as shown in Figure 67), an access point can provide coverage to the UE by connecting wirelessly to a donor cell.

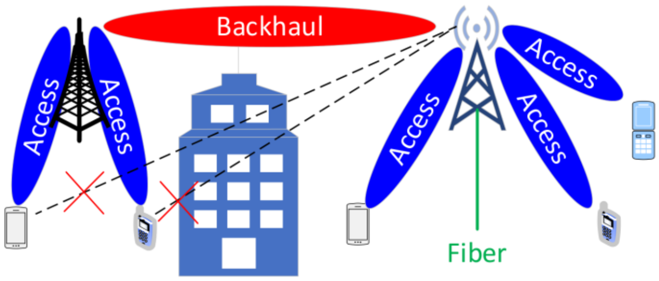

Figure 67: IAB coverage for isolated coverage gaps

- provide indoor coverage, for instance with an IAB access point on top of a building that serves users within the building [17].

7. TESTING AND VALIDATION

7.1 Introduction to 5G Testing

The goals for 5G are aggressive. The eMBB usage scenario targets peak data rates as high as 20 Gbps in the downlink (DL) and 10 Gbps in the uplink (UL) to support new applications such as high-speed streaming of 4K or 8K ultra-high definition (UHD) movies. While there are various ways to improve data rates, spectrum is at the core of enabling higher mobile broadband data rates. 5G NR specifies new frequency bands below 6 GHz and extends into mmWave frequencies where more contiguous bandwidth is available for sending lots of data. 5G NR Release 15 specifies mmWave operation up to 52.6 GHz with up to 800 MHz aggregated channel bandwidth. At mmWave frequencies, signals are more vulnerable to impairments. While consumers will appreciate the increased bandwidth, it introduces challenges pertaining to link quality requirements at mmWave frequencies. Impairments are not an issue at sub-6 GHz but become more problematic at mmWave frequencies. Hence, extra consideration is needed to determine test approaches that provide the precision required to correctly evaluate 5G components and devices [115].

Test technology development has accompanied the development of enabling technologies for every generation of mobile communications, leading to collaborative development. Various kinds of test instruments and systems support the numerous needs of the wireless communications industry, from research to verification and production. 5G test and measurement technologies are projected to appear before network and UE products, guiding product design and standard formulation [37].

In recent years, research institutions, MNOs, equipment, IC and instrument manufacturers around the world have performed major 5G technology validation and prototype testing in succession [116].

Compared to traditional testing in the 3G and 4G eras, three technical attributes of 5G impose enormous challenges on test instruments and methods [37]:

- the introduction of microwave and mmWave bands above 6 GHz
- the generation, reception, and storage of ultra-wideband (UWB) signals with bandwidths of hundreds of MHz or even GHz; and
- the design and application of large-scale antenna arrays (i.e. massive MIMO), with 64, 128 or more channels [37].

7.2 What is Tested: Test Categories

This section describes the influences these characteristics have on 5G test instruments and test technology with respect to various test scenarios for 5G wireless communications. Testing solutions are introduced as well [37].

7.2.1 5G Channel Sounding, Modelling and Emulation

The wireless channel is one of the fundamental components of the wireless communications system. Its physical properties of the wireless channel are characterised by a series of parameters, such as channel impulse response, path loss, Doppler delay, power delay profile and angle of arrival. Channel sounding can help extract these MIMO channel parameters and offer an important reference for subsequent 5G channel modelling and standardisation. The basic structure of the channel sounding system is illustrated in Figure 68, which incorporates the signal transmission/receiving instruments and measurement/analysis software [37].

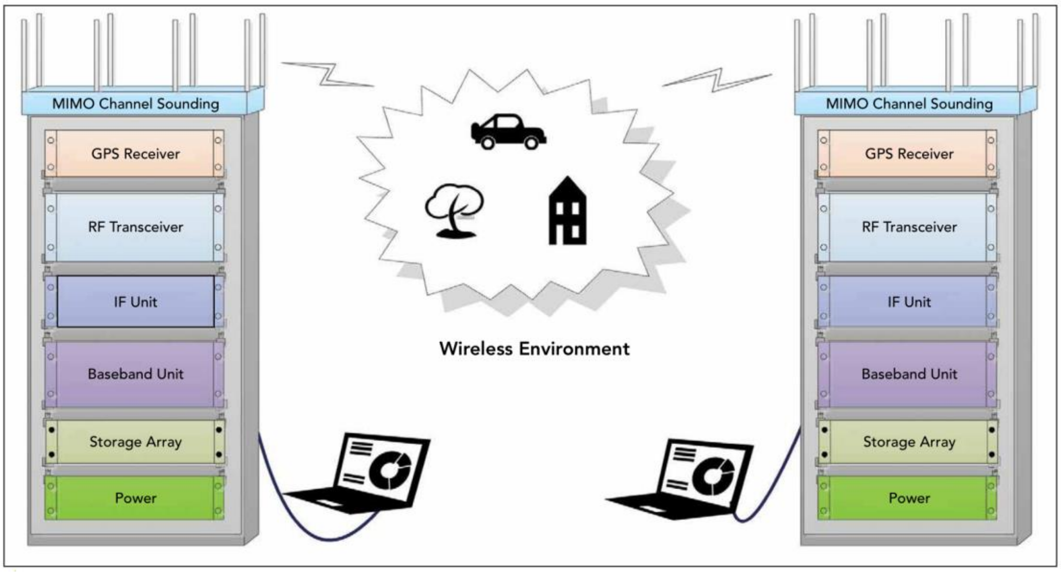

Figure 68: 5G channel sounding system

With the development of 5G technology, the traditional 3G/4G channel sounding system can neither handle new test challenges nor be updated to achieve better performance. Moreover, the characteristics of mmWave channels are not yet fully investigated or understood, and the update of traditional dedicated channel sounding equipment is likely to be expensive, due to inadequacies dealing with flexible testing from 6 to 100 GHz. Also, the ability to generate, receive and store ultra-high-speed baseband signals will need to be greatly enhanced, as well. Likewise, the introduction of large-scale antenna arrays (i.e. massive MIMO) substantially increases the required computational capacity of channel sounding instruments, making new multi-channel RF transceiver components the inevitable choice. Hardware and software platforms must, therefore, support massive data analysis and channel parameter extraction [37].

In recent years, a range of solutions have been proposed by enterprises and academic institutions around the world for 5G channel sounding based on a combination of existing products. For instance, Keysight Technologies proposed a mmWave MIMO channel sounding system whereas Rohde & Schwarz (R&S) proposed a scheme that enables fast measurement of both indoor and outdoor time domain channels with operating frequencies up to 100 GHz and bandwidths as wide as 2 GHz. Also, various research projects, such as METIS, NYU WIRELESS, mmMAGIC, MiWEBA and 3GPP, are investigating 5G mmWave channel measurements and modelling [3],[37].

Likewise, some popular channel models, like WINNER, COST 2100 and METIS 2020, have drawn more attention due to their scalability and reasonable complexity [117],[118]. These 5G channel models are customised for specific scenarios and frequency bands. Even though the mathematical methods used are not the same, these models are based on the analysis of a sizeable number of channel sounding results. At a 2016 3GPP RAN meeting held in South Korea, the first standard for the mobile broadband 5G high frequency (6 to 100 GHz) channel model was approved. However, there has been no unified 5G channel model combining both low and high frequency bands, which requires the corresponding aspects of 5G wireless technology to be considered. Extensive work on theoretical and practical channel modelling has continued in the past few years, but most research findings are constrained by spatial correlation and mutual coupling between adjacent antennas. Arrangement of antenna elements is fairly simple and assumptions about propagation conditions and antenna characteristics are too idealistic and limited to specific application scenarios. Hence, it is challenging to accurately extract a variety of actual channel characteristics [37].

When performing field tests for wireless communications systems in the real channel environment, there are numerous shortcomings, such as climate effects, poor mobility, high cost and unrepeatable test processes. The MIMO channel emulator (see Figure 69) allows researchers to emulate typical wireless channel environments in the lab, flexibly controlling and varying channel parameters, to identify performance problems as early as possible, lower test costs and significantly improve efficiency. Therefore, manufacturers have always included channel emulation as a vital part of the deployment of each generation of mobile communications technology, including 5G [37].

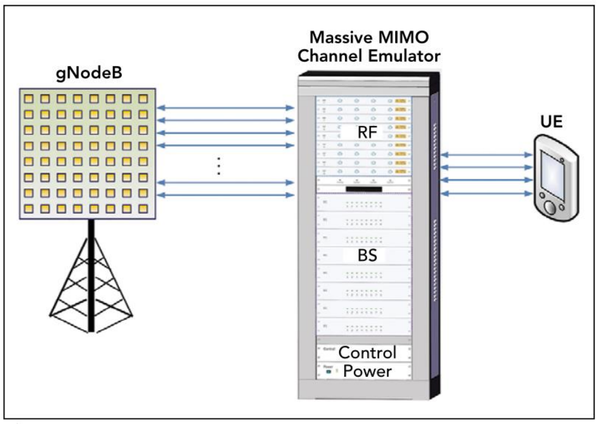

Figure 69: 5G channel emulator

As a key enabling technology for 5G, massive MIMO greatly reduces transmission power while improving channel capacity and spectral efficiency [76]. However, the required number of network equipment antennas goes from 10 to more than 100 times that of existing MIMO system antennas, which becomes a major bottleneck in the upgrade and optimisation of current 4G LTE channel emulators. The characteristic “pilot contamination” problem of massive MIMO technology directly affects the baseband channel estimation algorithm, feedback mechanism, interference control and synchronisation scheme [119-121]. Data throughput in the channel emulator rises sharply with expanding antenna array size, requiring extremely high-level computational resources, storage capacity and bus speed for the baseband processing unit. The RF system design must guarantee isolation and amplitude/phase consistency among multiple channels, which significantly increases channel calibration complexity [37].

Currently, Propsim F32, an advanced channel emulator from Keysight, is just able to support 32 RF channels, at most, and realise 64 × 8 MIMO channel emulations by combining multiple instruments. Unfortunately, this only covers operating sub-6 GHz frequencies and a maximum bandwidth 80 MHz; it is incapable of handling the high frequency and large bandwidth challenges of 5G test. An alternative channel emulator, Vertex, released by Spirent in 2016, is configured with 32 RF channels and 100 MHz bandwidth to meet the requirements of MIMO beamforming, MIMO over-the-air (OTA) and massive MIMO test, but the maximum operating frequency to 5.925 GHz only satisfies low frequency 5G test demands [37].

Over the years, a couple of Chinese instrument manufacturers have launched 8 × 8 MIMO channel emulators for 4G testing, laying a hardware platform and algorithm architecture foundation for the development of 5G channel emulators. Upcoming channel emulator technology have to achieve bandwidths of hundreds of MHz, cover frequency bands above 6 GHz and contain multiple channel models [37].

7.2.2 Radio Frequency (RF) Module and Antenna Array Test

Large-scale antenna arrays (i.e. massive MIMO) and RF frontends are indispensable 5G subsystems, comprising DACs, ADCs, frequency synthesisers and transmit/receive (T/R) multibeam antenna arrays. The T/R array includes RF components such as filters, mixers, power amplifiers and low noise amplifiers, each with its own set of performance specifications and corresponding test methods [37].

Figure 70 depicts the commonly used classes and functions of the excitation/source instruments and receiving/analysis instruments in RF test. Among them, the vector signal generator (VSG) and signal analyser provide the most complete measurement and analysis of a communication system's overall performance. The operating frequency ranges from 0 Hz to approximately 110 GHz, while supporting 200 MHz to several GHz of vector signal bandwidth. Regarding frequency coverage, the Keysight E8267D PSG and the R&S FSW85 signal analysers have reached 44 and 85 GHz, respectively. The maximum signal analysis bandwidths of two well-known signal analysers, i.e., FSW from R&S and UXA from Keysight, are 2 and 1 GHz, respectively. Additional bandwidth expansion requires the support of other components. To generate/analyse signals with ultra-large (GHz) bandwidths, the main technical hurdles include RF channel equalisation, high sampling rate ADCs, high speed digital signal processing, and high data rate transmission [37].

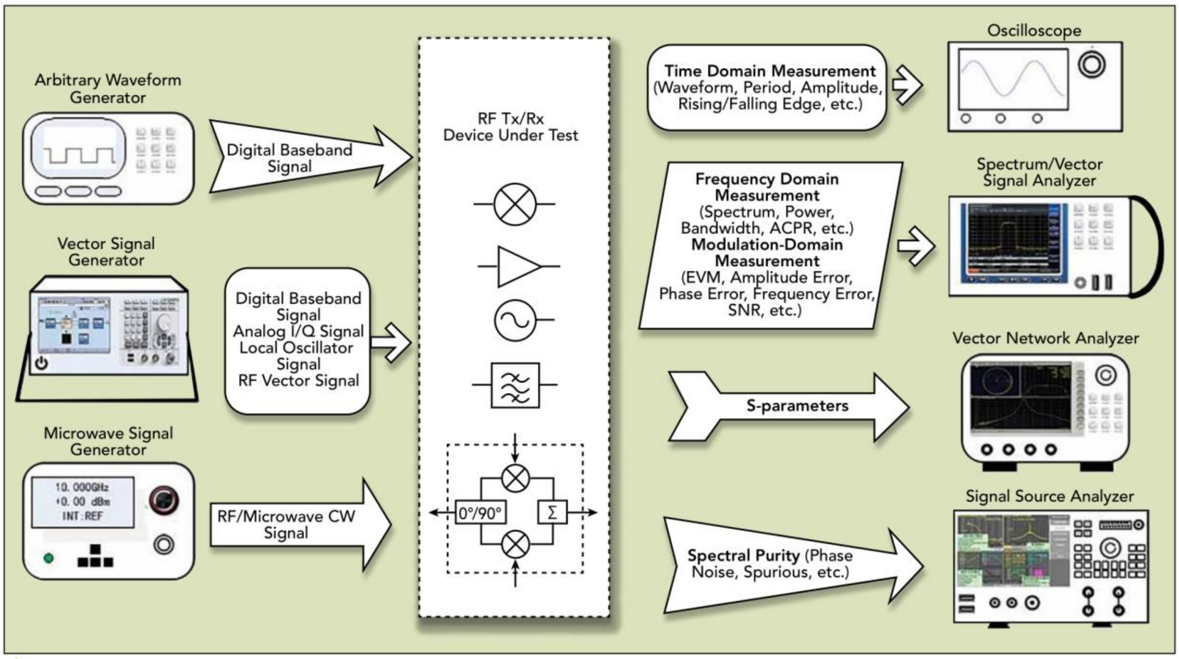

Figure 70: Instruments for testing RF transmit/receive components

For antenna array testing, the vector network analyser (VNA) is a key instrument. Due to the unavailability of a single, 64-port VNA, three methods are usually implemented. The first is a step-by-step test using a single, multi-port VNA, which is relatively cheap but sacrifices test speed and disregards the coupling characteristics between antenna elements. The second entails cascading several multi-port VNAs, for instance, a 64-element antenna array is tested with eight, cascaded 8-port VNAs (see Figure 71). This approach can precisely test the actual S-parameters of each antenna element after calibration and greatly increases test speed. However, there are some technical hurdles: the crosstalk between ports limits dynamic range, and calibration time impacts test efficiency. The third method uses the traditional dual-port VNA with a switch matrix, which is a compromise between the previous two options. Its cost is relatively low, but the speed is a bit slow, and the switch matrix introduces measurement errors. Some manufacturers are developing a single multi-port VNA to offer new solutions that address crosstalk between channels, fast calibration, cost, and other aspects [37].

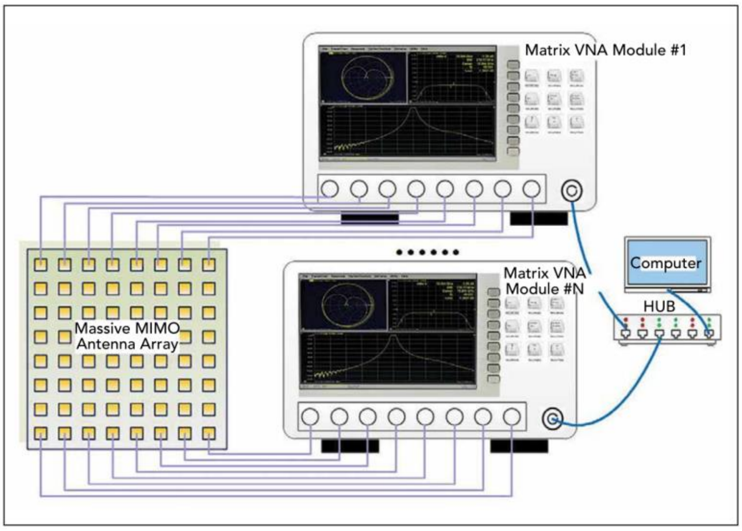

Figure 71:VNAs for testing massive MIMO antenna arrays

OTA test is another vital aspect of 5G antenna array testing for two reasons [122]. First, directional indicators of the antenna array, e.g., effective isotropic radiated power (EIRP) and effective isotropic sensitivity (EIS), have to be tested by OTA, which complies with 4G MIMO OTA test principles. Second, since 5G will utilise microwave and mmWave bands, the antenna array and T/R elements will be integrated to reduce loss and enhance matching. In this situation, most T/R component characteristics cannot be evaluated via wired conduction tests, and measures of performance like RF circuit transmit power and sensitivity may interact with the characteristics of the antenna, making individual assessment difficult [37].

7.2.3 Integrated Circuit (IC), Network and User Equipment (UE) Test

Although 5G-related technologies and standards are not yet finalised, IC, UE, and network equipment manufacturers, as well as MNOs, are rapidly conducting research and development (R&D) of 5G prototypes in a bid to launch competitive solutions. Amongst existing 5G prototype UE, some support high speed transmission of several Gbps and while others support as latencies as low as a millisecond. The battery life of some UE, especially IoT terminals, has been extended to nearly 10 years. Qualcomm, Spreadtrum, MediaTek and other IC manufacturers are developing 5G chips, of which Qualcomm has already announced prototypes. Also, there are plans to develop test technology for 128-channel integrated network equipment. However, UE diversity, scenario complexity and massive connections challenge the testing of ICs, UE, and network equipment [123], and existing 4G LTE test instruments can hardly fulfil these 5G testing tasks [37].

The integrated UE tester is used to emulate partial functioning of the network, then test the RF performance of the UE under network conditions or, with the signal generators, signal analysers (including spectrum analyser) and other conventional test instruments, perform conformance testing such as RF, protocol, and radio resource management. But emulating massive UE is likely to be a huge challenge for UE emulator design. 5G UE diversity implies that existing test instruments must have excellent scalability and compatibility. In addition, typical IoT application scenarios, like intelligent water meters and smart parking require low power consumption testing to assess battery self-discharge and sleep mechanisms. However, the industry presently lacks a mature test methodology to quantify UE power consumption characteristics. Other test systems, like the NV-IoT test system, the 5G terminal card interface test system and signalling monitors, are all essential components of 5G UE test. Hence, the complete 5G UE test system would need be an integrated and comprehensive test system that replaces multiple sets of discrete systems, as shown in Figure 72 [37].

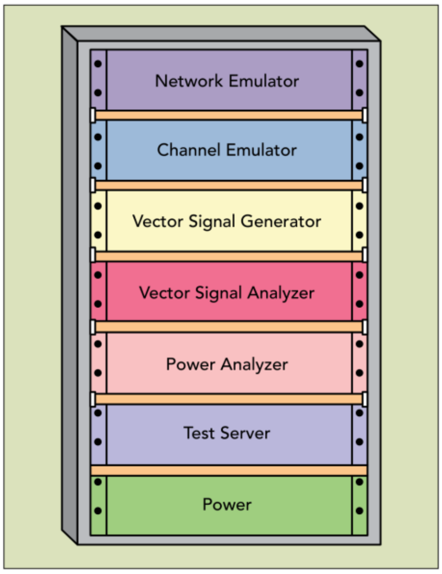

Figure 72: Integrated system for 5G terminal testing

Similarly, IC manufacturers have a strong aspiration to develop 5G test instruments at all phases, from chip development to product certification and mass production. Specifically, test instruments are expected to simulate network functions, verify and evaluate RF solutions, complete chip function/performance authentication, and perform final production test. To expedite the operation, further needs include installing software with a project configuration and results display, integrating other instruments to build test systems and supporting remote control. Also, the conventional IC test system is challenged by the need to be flexible, to reduce test costs, and to improve production efficiency which is driven by continuous module redesign, configuration changes, and reduced IC R&D cycle times and costs [37].

On the other hand, network equipment testing is used to confirm compliance with a communications system's quality specifications, interface requirements with other devices and electromagnetic compatibility, both intra-system and inter-system. A 5G network equipment test involves general instruments, such as a VSG, spectrum analyser, power meter, UE emulator, and channel emulator to build a massively connected test system with broad coverage. The objective is to test the load capacity limit and overload coordination capabilities of specific uplink and downlink service models and to assess system performance under different channel conditions. In the future, these test instruments must comply with corresponding 5G test specifications with continuous optimisation and upgrade capability and satisfy the performance needs of 5G while supporting a broader range of application scenarios [37].

7.3 How Testing is Done: Over-the-Air (OTA) Testing

5G has arrived and there is a lot of talk about OTA testing – that is, testing without the device under test (DUT) physically connected to the test equipment. This scenario differs noticeably from 4G, where most tests were carried out with cables. So, what is driving the emphasis on OTA testing in 5G? [24]

With devices and base stations operating in two frequency ranges, FR1 (450 MHz to 6 GHz) and FR2 (24.25 to 52.6 GHz), many of the low frequency tests will be similar to 4G LTE tests. However, base stations operating in FR2 will require radiated tests using OTA test methods. These methods will also be used to authenticate device and base station beam steering functionality in multi-element antenna arrays that are integrated into RFICs at FR1 and FR2 [23].

Multi-element antenna arrays will be used on 5G mobile devices to implement beamforming and/or beam steering. Phased array antennas are a practical and low-cost means of dynamically generating and pointing beams in a desired direction (i.e. beam steering). An array of smaller antenna elements forms a phased array antenna. By altering the relative phases and amplitudes of the signals applied to the individual elements, the antenna array can steer and shape a beam in a given direction. However, these arrays will be incorporated into RFICs and can only be tested using OTA methods, as there are no probe points to make conducted measurements [58].

The mmWave antenna arrays are quite small; for instance, a 24 GHz full-wave spacing is just 12.5 mm. In addition, they require physical bonding to the amplifier semiconductors for transmission and reception of radio signals. These integrated designs are impractical to probe, and cabled tests become impossible to test mmWave parameters [124].

Hence, OTA tests will be essential for the development, validation, and commercialisation of 5G devices. A standard OTA test will involve an anechoic chamber, different probing techniques, and test equipment to generate and analyse the radiated signals in a spatial setting. The anechoic chamber offers a non-reflective environment with shielding from external interferences so that radiated signals of known power and direction can be generated and measured in a controlled environment [24].

Therefore, to answer the question at the beginning of this section in an itemised form, the emphasis on OTA testing in 5G is driven by:

- **Increasing density of multi-element antenna arrays**: Phased-array antennas have become a preferred technique for mobile communications. They are used to create narrow beams and dynamically steer the beams in a desired direction when the relative phase and amplitudes of a signal applied to each antenna element are varied [24].

  They will be used in frequency bands below 6 GHz to boost capacity and in mmWave bands to overcome path loss issues. Due to the number of antenna elements required by the antennas to configure narrow beams, it will be difficult to fully characterise and validate beam performance. Designers will need to test and measure beam patterns using OTA test methods [24].

- **Shrinking size of mmWave components**: Higher frequencies imply shorter wavelengths. Multi-antenna arrays at mmWave frequencies will help to resolve signal propagation issues and provide directional antennas with higher gain. With shorter wavelengths, antenna elements can be tightly spaced, resulting in extremely compact arrays. Many vendors even choose to develop arrays that integrate into ICs. Of course, with highly integrated ICs, there are no probing points or place to put connectors. Consequently, it will become impractical to use traditional RF connectors between the radio circuit and the antenna, hence the need for OTA testing [24].

7.3.1 OTA Test Methods

OTA measurements are normally made in either the radiated near-field or the radiated far-field regions of an antenna system under test, as indicated in Figure 73 [34].

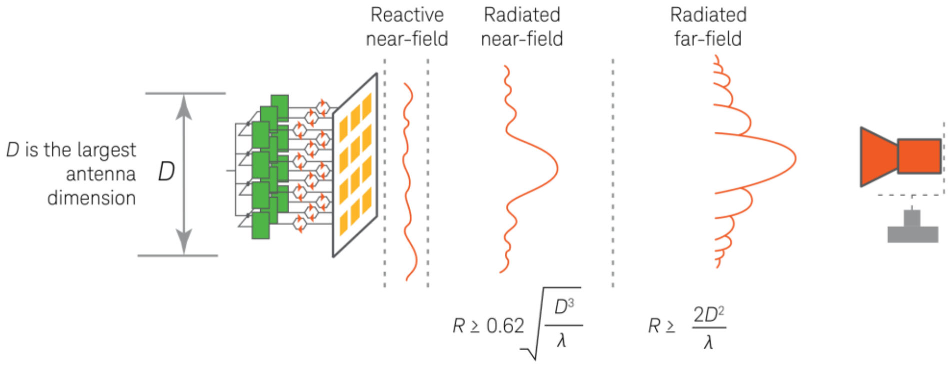

Figure 73: Beam evolution as a function of distance from an antenna

The reactive near-field is very close the antenna, usually in the order of a few wavelengths or less. In this region, absorption or coupling to objects (e.g., a probe antenna) has an effect on the E and H field components radiated from the antenna. As illustrated in Figure 73, this region extends to a distance approximately defined as [34]:

$$R = 0.62\sqrt{\frac{D^3}{\lambda}}$$

where D is the diameter of the smallest sphere encapsulating all the radiating elements of the antenna, including passive elements [34].

Further than this distance, a probe antenna will not affect the signal radiated by the antenna. In the radiated near-field, referred to as the Fresnel region, the radiated angular field distribution evolves with increasing distance from the antenna, even though the evolution slows at some point. The radial components of E and H fields may be significant in this region, diminishing in strength rapidly with distance R. For antennas where D >> λ, the far-field or Fraunhofer distance starts at $2D^2/\lambda$ from the antenna. Beyond this distance, the E and H field components are transverse and orthogonal, and the radial field components are negligible [34].

The boundary to the Fraunhofer region is not an abrupt transition in the evolution of the radiated patterns. The distribution of radiated power continues to evolve beyond this distance, so that an antenna pattern measured at $2D^2/\lambda$ will differ slightly from the pattern measured much further away, but it is close enough for various OTA test purposes. As the distance from the antenna increases, the amplitude of the peaks, sidelobes, and nulls in its radiation pattern stabilise [34].

With this in mind, 3GPP has defined three permitted test methods, which are: direct far-field method (DFF), indirect far-field method (IFF), and near-field to far-field transform (NFTF) [24]. They are described in detail as follows:

- **Direct Far-Field Test Method**: Far-field measurement is conceptually the simplest type of OTA measurement system and has been approved by 3GPP for base station RF measurements and as the baseline UE RF test method. A typical far-field anechoic chamber is shown in Figure 74 [34].

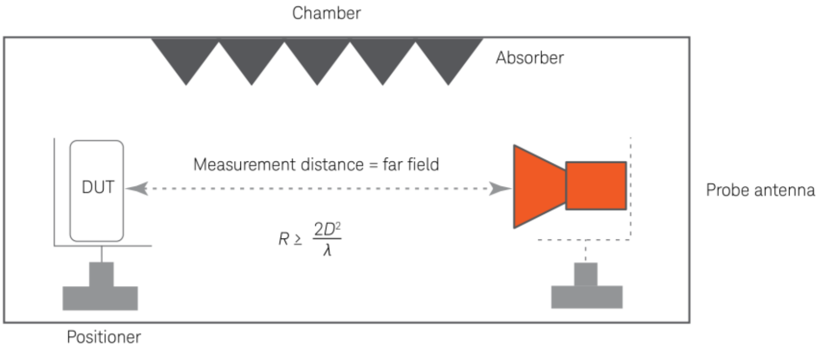

Figure 74: Far-field anechoic chamber

The DUT is mounted on a positioner that rotates in two planes, either azimuth and elevation or azimuth and roll, enabling measurement of the DUT at any 3D angle. Some chamber systems use an array of probe antennas instead of a single antenna, which can simplify the standard positioner movement at the cost of increased switching and calibration complexity [34].

Based on the far-field Fraunhofer distance equation, the far-field distance and therefore the size of the required chamber can vary considerably with antenna size and frequency. Table 3 gives some typical far-field distances at 5G associated frequencies. Since the direct far-field chamber has to be larger than this minimum distance, its size and cost can quickly escalate [34].

Table 3: Far-field distances for antennas of various sizes

	Far-Field Distance (m)		
D (mm)	28 GHz	39 GHz	60 GHz
50	0.47	0.65	1
100	1.9	2.6	4
150	4.2	5.9	9
200	7.5	10.4	16
300	16.8	23.4	36.0

That implies that typical active antenna array systems (AAS) for base stations, operating at mmWave frequencies, with sizes in the range between 100 mm and 300 mm would require large room-sized direct far-field anechoic chambers. However, using direct far-field anechoic chambers for mmWave 5G UE may be feasible since they would have antenna arrays that are significantly smaller than 100 mm in size [34].

Nevertheless, it is not always easy to define the D to be used in the far-field equation. For a typical antenna array, D is the diagonal extent of the array. When that array is integrated into a UE or a base station, the array may couple to the ground plane behind it and to nearby conducting materials causing the effective radiating area to increase. For mmWave 5G UE, another challenge in determining the value of D is that these devices are expected to have multiple antenna arrays to enable full spherical coverage and to help overcome situations where the signal is blocked by a human hand or head [125]. Figure 75 depicts examples of possible array placements [34].

Figure 75: Potential antenna array configurations for a mmWave 5G UE

Considering this, two approaches can be taken to measure devices with one or more antenna arrays. The first is a “white box” approach that requires prior knowledge of the antenna array position on the DUT. This position is determined either by design, declaration, or near-field scanning. The DUT is then positioned so that the centre of radiation is placed at the centre of the test zone. The far-field distance can then be computed with D being set to the largest dimension of the array (assuming no substantial ground plane effects). Typical array dimensions would then result in a far-field distance much less than 1 m for a UE. Testing the entire device would require repositioning the DUT for each array [34].

The second approach does not require prior knowledge of the antenna array position(s) and is referred to as the “black box” approach. In this case, the geometric centre of the DUT is placed at the centre of the test zone and the D used for computing the far-field distance is the maximum dimension of the DUT. The distance found using this approach is typically much larger than that found using the white box approach, but the device does not need to be repositioned even if its antenna arrays are active [34].

The white approach is desirable in a development environment since the far-field distance is much shorter and there is prior knowledge of the antenna structure. However, for conformance testing, 3GPP has decided that only the black box approach should be used. Moreover, the requirements for white box testing have not been universally accepted by UE vendors who prefer not to declare the antenna structure. In addition, there is no mechanism for the UE to signal when it changes its array and besides, the white box approach rules out use of more than one array at a time [34].

- **Indirect Far-Field Test Method:** This test method is based on a Compact Antenna Test Range (CATR) that uses a parabolic reflector to make the signals transmitted by the probe antenna parallel, thereby creating a far-field test environment. Even though, this method measures a single signal, it offers a much shorter distance and with less path loss than the direct far-field (DFF) method for measuring mmWave devices [124]. The basic configuration and operation are depicted in Figure 76 [34].

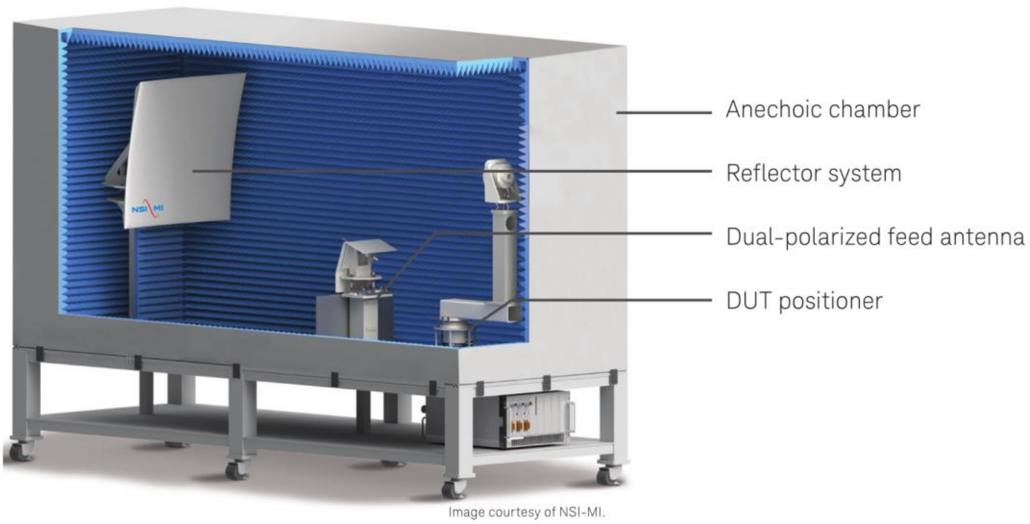

Figure 76: A compact antenna test range (CATR)

The key elements of the CATR method shown in Figure 76 are: a feed or probe antenna, a parabolic reflector system, and a rotating positioner to hold the DUT [34].

A diverging beam from the probe antenna located at the focus of the mirror illuminates the parabolic reflector, which then collimates the original beam and directs it to the DUT. The collimated beam has an approximately uniform amplitude and phase across its extent; it offers a nominally ideal plane-wave illumination to the DUT. The reflector enables the DUT to be tested under far-field plane wave conditions at a shorter distance than $2D^2/\lambda$, resulting in a method with potentially smaller footprint and lower path loss than the compared to the DFF method [34].

The volume in a CATR chamber is called a quiet zone. The quiet zone is usually cylindrical or elliptical in shape with the diameter or axis defined by the extent of the collimated beam within which the phase and amplitude are within certain limits. The zone starts at a short distance behind the probe antenna [126]. Typical quiet zone specifications are 10 degrees of phase variation, ± 0.5 dB of amplitude ripple, and 1 dB of amplitude taper, which is the roll-off toward the edges of the quiet zone [34].

Furthermore, correct device characterisation requires that the device’s entire radiating volume be within the quiet zone. The rotating positioner then allows characterisation of the device or antenna as a function of angle (azimuth and elevation). Since the configuration is reciprocal, the device can be measured either in transmit or receive mode without a need for repositioning [127].

- **Near-Field To Far-Field Transform Method:** This method samples the phase and amplitude of the electrical field in the radiated near-field region, also referred to as the Fresnel region, over a 2D surface using high-precision positioners. This surface could either be planar, cylindrical, or spherical. The far-field antenna pattern is then computed using Fourier transform algorithms [34].

Figure 77: Planar near-field scanner

This method can be achieved with comparatively small test chambers (see Figure 77) which are considerably cheaper than indoor far-field ranges. Measurement precision and the computed far-field antenna patterns can compare very well with far-field methods [34],[128].

The most suitable near-field system for a specific application will depend on the antenna characteristics of the DUT. Mathematically, the planar surface method is the simplest; it is appropriate for measurements of directional antennas with minimal backward lobe typical of many mmWave 5G gNB antenna arrays. Planar systems also allow the use of mechanical surface alignment to ensure accurate parallel movement of the scan probe with respect to the DUT array surface. For several 5G DUTs, spherical scan systems perform measurement of the DUT in a practically complete sphere around the DUT, allowing measurement of radiation from the edge or back of the array under test as well as in the main lobe. On the other hand, cylindrical scan systems may be appropriate for antennas arranged in a linear fashion or a sectorised layout [34].

After measuring the electric field at each point in the grid, the data undergoes a Fourier transform to obtain a linear combination of plane waves at various angles. These plane waves, which are represented in a spherical coordinate system, provide the far-field antenna patterns [51]. The transform is essentially a conversion of near-field measurements at multiple positions to far-field angular patterns [129]. In the case of the planar scan system, the extent of the scan is usually much larger than the DUT itself, and it is used to set the angular resolution in the far-field. Correct far-field results require compensation for the directive properties of the probe antenna [34].

For R&D purposes, near-field scanning is valuable as a diagnostic tool since transforms can be used to reveal the distribution of surface currents on an antenna array and identify faulty elements or other conditions. Moreover, it is a comparatively fast measurement method. Even though the near-field to far-field approach provides the most compact means for antenna characterisation, it poses challenges for mmWave device characterisation [34].

In conclusion, a brief comparison of the three 3GPP approved OTA test methods is shown in Table 4 below [22].

Table 4: Comparison of OTA test methods

Direct Far Field (DFF)	Indirect Far Field (IFF)	Near Field to Far Field Transformation (NFTF)
Simple and comprehensive	Near field to far field conversion enables compact antenna test range	Compact approach, can be lower cost
Can be very large with greater path for mmWave devices	Suitable for testing mmWave devices; not well suited for spatial RRM	Limited application: transceiver only; no receiver or RF parametric tests yet

7.3.2 OTA Test Challenges

Excessive path loss and distance at mmWave frequencies: As already explained, OTA tests are typically done in either the near-field or far-field regions of the antenna array. The characteristics of the transmitted electromagnetic wave vary depending on the distance from the transmitter. The signal becomes more developed as it propagates from the antenna array and the amplitude of the peaks, side lobes, and nulls of its radiation pattern evolve toward the far-field pattern [124].

While near-field measurements are suitable for some applications, 5G cellular communication link measurements will be based on far-field assumptions. Due to the nature of radiated waves, the far-field distance and associated path loss increase significantly with frequency. For instance, the far-field region of a 15 cm, 4G LTE device operating at 2 GHz starts at 0.3 metres and has a path loss of 28 dB. The far-field region of a 5G NR device operating at 28 GHz has a far-field distance of 4.2 metres and a path loss of 73 dB (see Table 5). This distance demands the use of an excessively large far-field test chamber, and the path loss is too great to make precise and repeatable measurements at mmWave frequencies. The distance also increases considerably as the source antenna grows bigger, worsening the size and path loss challenge. Hence, to successfully test RF performance measurements like transmitted power, transmit signal quality, and spurious emissions would entail overcoming the path loss problem [124].

Table 5: Estimated far-field distance and path loss for different radiating apertures

Size D (cm)	2 GHz Distance (m) Path loss (dB)	28 GHz Distance (m) Path loss (dB)	43 GHz Distance (m) Path loss (dB)
10	0.13 m 21 dB	1.87 m 66 dB	2.87 m 74 dB
15	0.30m 28 dB	4.2 m 73 dB	6.4 m 81 dB
20	0.53 m 33 dB	7.4 m 78 dB	11.4 m 86 dB

mmWave OTA test methods not fully defined: Low frequency tests in 5G are comparable to those in 4G. However, for mmWave, the following tests would require the use of OTA methods [124]:

- RF performance – minimum level of signal quality
- demodulation – data throughput performance
- radio resource management (RRM) – initial access, handover, and mobility
- signalling – upper layer signalling procedures

Amongst these, the only fully developed ones are the 5G RF performance test methods. 3GPP study groups are still defining test techniques for device demodulation and the more complex RRM [124].

Measuring device performance in real-world conditions: Performing OTA tests with the highest degree of performance and reliability requires a stable and controlled validation environment. This is where a channel emulator proves to be useful. It is a tool that simulates real-world conditions while offering control and repeatability for those conditions. It enables testing a variety of new technologies, including wider signal bandwidths, mmWave frequencies, and beam steering with signal propagation issues like path loss, multipath fading, and delay spread [124].

7.4 Why Testing is Done: Test Objectives

Conformance: Conformance testing entails determining the degree to which a system under test (SUT) meets the requirements defined in the implementation conformance statement (ICS). It proves that a product accurately implements the corresponding standard and is able to exchange instructions and information with other implementations via a known protocol or set of protocols. Conformance is a prerequisite for interoperability [130].

A major milestone in the development of devices and base stations is passing the conformance and compliance tests defined in the 3GPP RAN4 and RAN5 specifications. All UE and base stations have to pass the mandatory conformance tests before being released to the market. Conformance tests, however, only provide a minimum pass/fail result, giving no indication of how the device will perform when incorporated into a wireless communications system. Device and base station manufacturers will test a wider set of parameters using verification and regression testing to guarantee quality and sufficient margins. Pre-conformance testing is also carried out to check the confidence of a “pass” before conformance testing. This lowers the time and cost of rework in case the device fails official conformance tests (see Figure 78) [22].

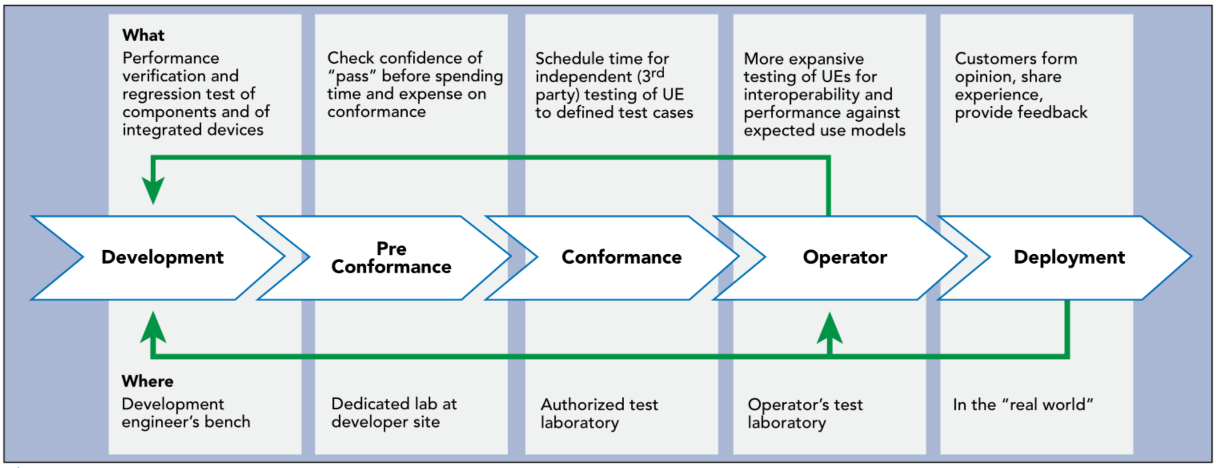

Figure 78: Typical test flow from development to deployment

Interoperability: The purpose of an interoperability test is to determine that products developed by various manufacturers communicate correctly with one another in accordance with the same communication standard or reference specification [130].

Certification: Certification suggests to end-users that a product delivers an acceptable level of quality in terms of performance and interoperability, allowing them to differentiate products with proven quality from those with unknown quality [130].

In communication context, product certification is necessary when there are safety, security, or connectivity implications. Its purpose is to ensure that products do not have issues that could adversely impact the health or safety of persons or the QoS for other communication network users [130].

Performance: This can be determined either in a laboratory and on the field.

- **Laboratory testing** – The DUT or SUT is connected to an emulated network simulates real-world propagation effects. During the testing, a trade-off between the number of significant test cases and the desired confidence level to comply with the given requirements needs to be considered, due to the high number of potential test cases. This testing is relevant to all ecosystem stakeholders as it gives indications of the real-world behaviour of the wireless solution [130].
- **Field trials and field testing** – The purpose is to verify the desired performance of the test object's system implementation in an environment which is largely similar to the target application scenario. Essentially, this entails the pre-commercial or commercial deployment of products in the desired use case scenario. During the trial, measured system KPIs are captured and then compared with predefined requirements. Field trials are performed on solutions that have been validated already. They are relevant to all ecosystem stakeholders because they provide information on the real-world behaviour of solutions [130].

Product validation: This is a qualitative assessment to establish whether a wireless communication module, a base station or a wireless device meets user requirements. It is done before commencing volume manufacture [130].

Product verification: This is a quantitative assessment to establish whether a wireless communication module, a base station or a wireless device meets design specifications. It is done after the product is manufactured [130].

Commissioning and acceptance: This involves the verification of a wireless system before it is put into operation. The deployment of the system, its configuration and characteristic parameters should be well-documented to serve as a reference for periodic audits or as a basis for extending or modifying the system [130].

Periodic audit: This is performed to achieve two objectives [130]

- To verify that products still comply with the requirements of conformance or certification testing. The aim is to ensure that modifications to the manufacturing process are not affecting the performance of the product and that it remains compliant. It is considered an aspect of quality control if it performed by the manufacturer. On the other, it is a means of monitoring production or an initial certification of the product design, if it is performed by the certification scheme owner or their representative [130].
- To verify that a wireless communication system still functions correctly after initial installation and has not degraded in performance while still fulfilling specified requirements. It is considered a part of maintenance if it is performed by the owner of the wireless communication system whereas it is a way of certifying the entire system if it is performed by the scheme owner [130].

Diagnosis and failure analysis: Communication systems must be constantly monitored to ensure reliable use. If there are deviations from typical behaviour, the reasons should be determined via appropriate tests. If communication fails, communication must be analysed so the cause is identified and resolved as quickly as possible [130].

7.5 OTE 5G Testbed

7.5.1 Introduction

Government policy makers and national regulatory authorities (NRAs) are encouraging technology pilots to stimulate early investment in 5G networks and infrastructure, and to bolster their understanding of 5G technologies. Moreover, the telecoms sector, comprising MNOs, vendors and research institutes, has been participating in 5G testbeds independently of NRA or government involvement [7].

Therefore, in line with other commercially led 5G testbeds all over the worldwide [7], the Hellenic Telecommunications Organisation (OTE) decided to set up a testbed, as a partner, within the 5G EVE[8] European project.

5G EVE[9] is the European 5G validation platform for extensive trials. It is one of three 5G Public-Private Partnership (5G PPP) infrastructure projects started on 1st July 2018. The 5G-EVE concept is based on further developing and interconnecting existing European sites to form a unique 5G end-to-end facility.

[8] https://www.5g-eve.eu/

[9] A large part of this section is adapted the 5G EVE document (https://www.5g-eve.eu/wp-content/uploads/2018/10/5geve_d2.1-initial-architectural-facilities-description.pdf)

The 5G end-to-end facility is composed of 4 interconnected sites facilities located in Greece, Italy, Spain, and France as shown in Figure 79.

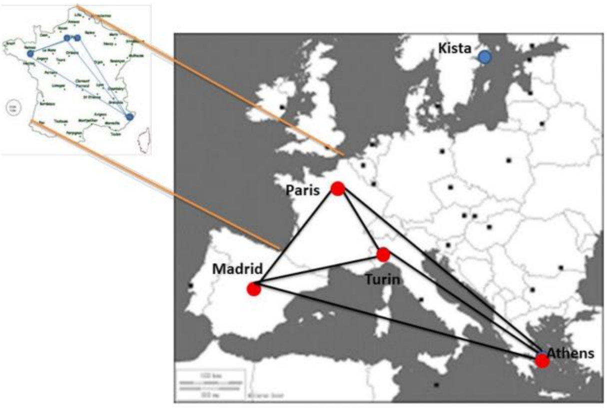

Figure 79: Location of 5G EVE site facilities

The objective is for the end-to-end facility to host a selection of use cases to be deployed by verticals. More specifically, the project targets as part of its workplan:

- Use case 1 – Smart Transport: Intelligent railway for smart mobility
- Use case 2 – Smart Tourism: Augmented Fair experience
- Use case 3 – Industry 4.0: Autonomous vehicles in manufacturing environments
- Use case 4 – Utilities (Smart Energy): Fault management for distributed electricity generation in smart grids
- Use case 5 – Smart cities: Safety and Environment - Smart Turin
- Use case 6 – Media & Entertainment: UHF Media, On-site Live Event Experience and Immersive and Integrated Media)

These use cases are applied to the different site facilities following the mapping presented in Table 6, which also details the guiding vertical partner in charge of use case planning and integration. The architecture of the site's facilities is very dependent on the requirements targeted by the service (and vertical) to be implemented (supported).

Table 6: Use case mapping

Use-case	France	Greece	Italy	Spain	Dominance
Use-case 1 - Smart Transport: Intelligent railway for smart mobility			Trenitalia		URLLC and mMTC
Use-case 2 - Smart Tourism: Augmented Fair experience				SEGITTUR	URLLC & eMBB
Use-case 3 - Industry 4.0: Autonomous vehicles in manufacturing environments		OTE/ Ericsson GR		ASTI	URLLC
Use-case 4 - Utilities (Smart Energy): Fault management for distributed electricity generation in smart grids	EDF	WINGS			URLLC and (critical) mMTC
Use-case 5 - Smart cities: Safety and Environment - Smart Turin		Nokia GR/ WINGS	Comune Torino		URLLC & mMTC
Use-case 6 - Media & Entertainment: UHF Media, On-site Live Event Experience and Immersive and Integrated Media)	ORANGE			Telefonica	eMBB, URLLC and mMTC

An overview of the administrative details of the four 5G EVE site facilities is shown in Table 7. It is worth noticing that the site owner is supported by a network operator and that each sites involve a number of partners covering the various phases of design, deployment, integration and operation with specifically assigned roles. The initial locations are also reported but could change since they are very dependent on the carrier frequencies made available by each national regulator.

Table 7: Overview of 5G EVE site facilities

Site Facility	Greece	Spain 5TONIC	France	Italy
Owner (operator)	OTE	Telefonica	Orange	TIM
Location(s)	Athens	Madrid	Nice, Paris, Châtillon & Rennes	Turin
Involved partners	Nokia, Ericsson, Wings	Ericsson, UC3M (IMDEA), Segittur, ASTI, Telcaria	Nokia, b-com, Eurecom, EDF, Orange	Ericsson IT, Nextworks, CNIT, Comune Torino, Trenitalia, Ares2t

For each site facility, a "Site Facility Manager" is defined, with two partners identified as primary and secondary to guarantee operation continuity (Table 8). Site facility managers are in charge of the site facility implementation, validation and operation.

Table 8: Site facilities management

Site Manager	France	Greece	Italy	Spain
Primary	Orange	OTE	TIM	IMDEA
Secondary	Nokia	WINGS	Ericsson	UC3M

7.5.2 Testbed Description

The Greek 5G EVE site facility covers a region of Northern Athens, around the R&D site of the Greek National Telecommunications Organisation (OTE). The OTE site serves as a testbed for services, equipment, and new features prior to their commercial release (including [pre-] 5G equipment), while it also maintains a connection to the commercial 4G+ network of OTE. The existing equipment and network functionality are a mix of Ericsson and Nokia technologies which have been progressively extended to support 5G during the lifetime of the project by both vendors. OTE, Ericsson GR and Nokia GR with the help of WINGS ICT SOLUTIONS (responsible for input integration, software development and pilot execution for one of the use cases) are responsible to prepare and upgrade the Greek site facility to be able to handle three 5G oriented vertical use cases, namely:

- UC3 – Industry 4.0 functionality with Automated Guided Vehicles (AGVs)
- UC4 – Utilities applications on Smart Energy grid monitoring & ultra-reliable / fast fault detection
- UC5 – Smart cities applications focused on Connected Ambulance

In the following sections, the high-level as well as the detailed architecture of the Greek 5G EVE testbed is described, including use case specific set-ups, hardware (HW) and software (SW) components presentation, RAN, Cloud and Core equipment and functionalities descriptions as well as potential configurations, settings and choices made for the operation of the network.

7.5.3 Testbed Architecture Overview

The Greek site consists of platforms and components of Ericsson and Nokia in terms of the access and core networks and have been upgraded with MANO ETSI compliant components. As it is explained in the following sections, during the initial stages of 5G EVE, Ericsson GR and Nokia GR will support their respective use cases with their own access and core networks (some level of interconnection is being considered at a later stage). The core networks (vEPCs) are part of a Distributed Mobile Broadband (DMBB) platform. The main logical architecture is shown in Figure 80.

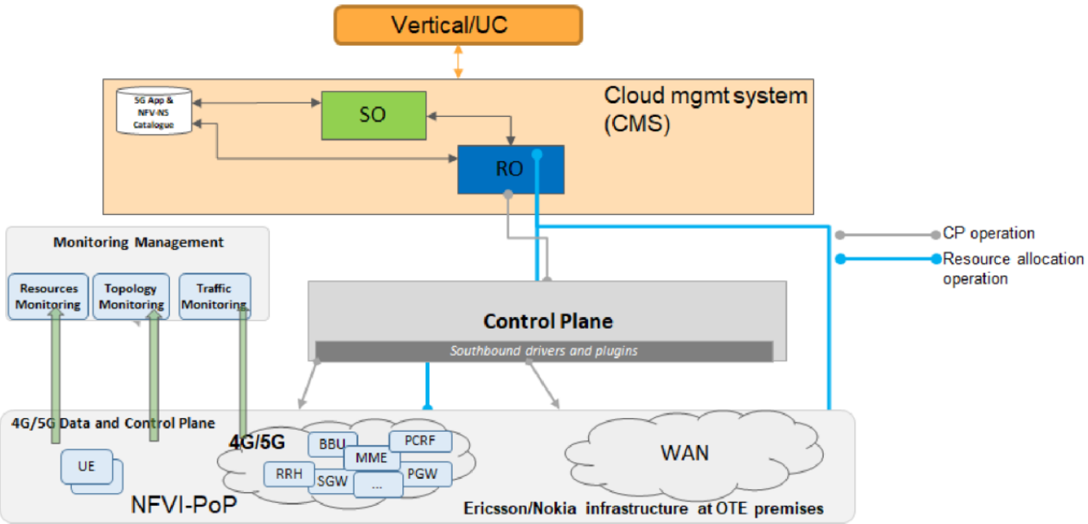

Figure 80: OTE testbed architecture

As it can be seen in Figure 80, the NFVI PoP consists of physical and virtual components from Ericsson GR and Nokia GR platforms that have been installed in OTE's premises as they explained in the following sections.

The RAN cloud consists of different nodes for Ericsson GR and Nokia GR. The core network consists of vEPCs from Ericsson GR and Nokia GR that in a second phase will be interconnected to each other. For the Ericsson GR case, a MEC cloud infrastructure has been used to advance the MEC service of AGV. The cloud management system consists of the service and resource orchestrators. The control system is responsible for controlling the VNFs' lifecycle and the traffic to the underlying data plane. The monitoring management system is responsible for collecting data on metrics, topology, and the resources in order to monitor the KPIs and adapt them to the needs of the project.

The three use cases that have been executed at a first stage over the Greek site facility utilise existing OTE facilities in combination with existing and new Ericsson and Nokia components and platforms. At the first stage of deployment (April 2019), some proprietary components were used by both vendors to enable the respective use cases that they are driving (i.e. Industry 4.0 for Ericsson GR and Smart City for Nokia GR), as well as to enable the execution of the utilities use case. Later on, the interconnection of equipment between the vendors will take place, enabled by OTE overlaid components. Hence, it operates as one unified facility capable of supporting multiple use cases. Figure 81 below depicts the use case oriented architectural overview of the Greek 5G EVE site facility, providing insights regarding the different vendor components co-existence and cross-utilisation and their respective usage by the envisioned verticals. More details regarding the detailed architecture from the point of view of each vertical and the kind of components and features to be used are provided in the following sub-sections.

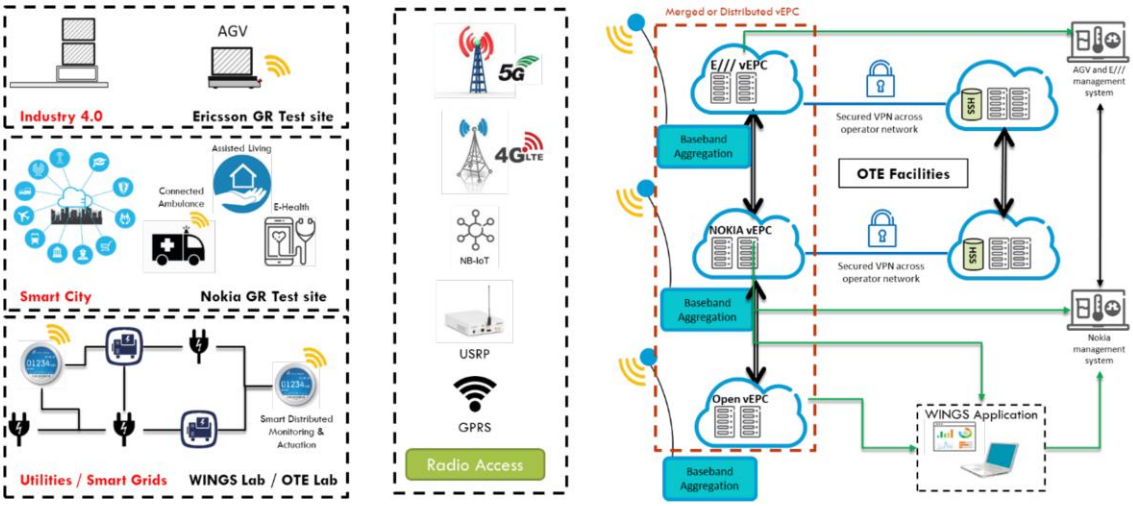

Figure 81: OTE testbed use cases

Industry 4.0 – Automated Guided Vehicle use case architecture: Mobile Cloud Robotics (MCR) in a Smart Wireless Logistic (SWL) facility is an exciting 5G platform that is being exploited by Ericsson GR and by the development partners. Mobile robots are used to transport goods between various stations in a process or to and from depots. Deploying mobile robots in logistics improves productivity and supports the implementation of effective lean manufacturing. As long as there are no constraints imposed in their movement capabilities caused by unexpected obstacles or dirt, robots can carry out any sequence of events to ensure that materials arrive at the right place just in time.

Realistic MCR scenarios are enabled through the replacement of traditional robots with new ones connected to the cloud. These new robots only include low level controls, sensors and actuators. So that, having their intelligence in the cloud means that they have access to almost unlimited computing power. Altogether, they are more flexible, more usable and more affordable to own and operate. The connection between MCR robots and the cloud is provided through the mobile network and will benefit from the expected 4G and 5G extremely low latency connections. The overall mobile network architecture for the Automated Guided Vehicle (AGV) is highlighted in Figure 82.

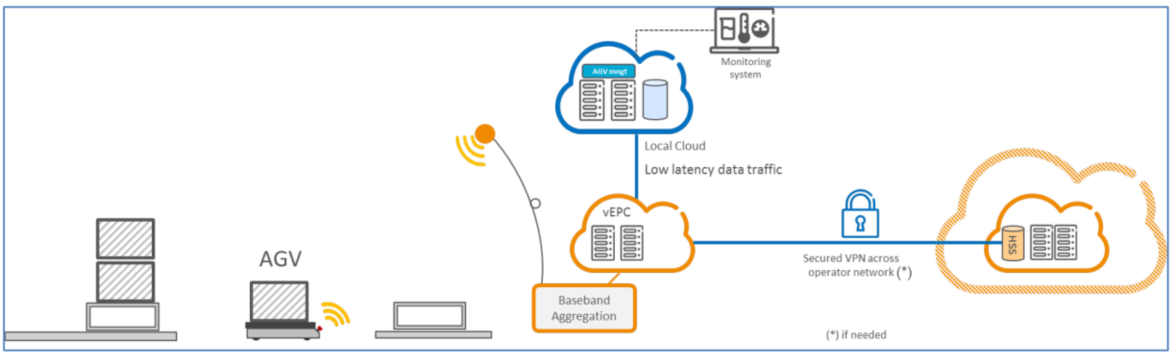

Figure 82: Automated Guided Vehicle (AGV) use case

The scope of the SWL is to build a stand-alone RAN and EPC solution able to fulfil the requirements of the planned use cases. The provided solution includes packaging and configuration of an Ericsson GR LTE+/5G system suitable for manufacturing/logistics process needs. The Smart Wireless Logistics solution end-to-end architecture is depicted in Figure 83.

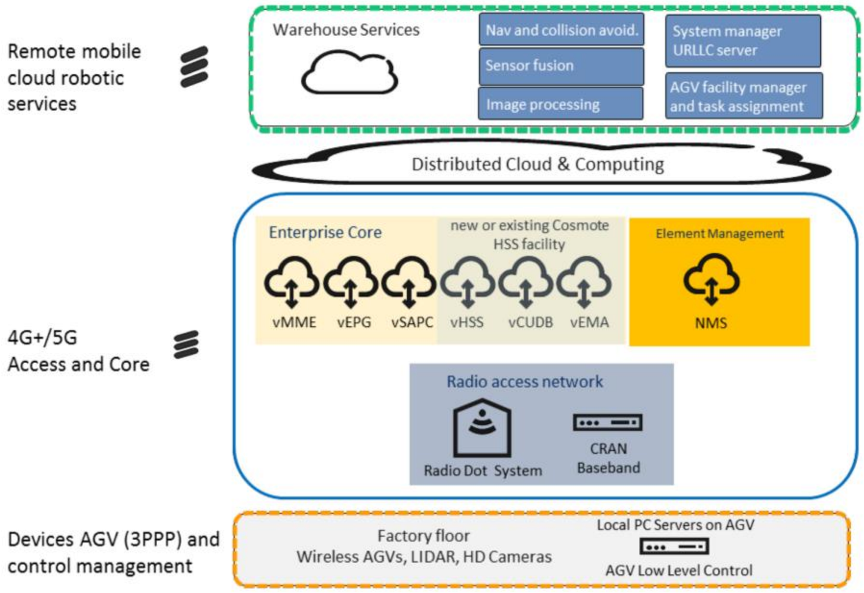

Figure 83: Smart Wireless Logistics (SWL) solution

The smart warehouse facility project can be easily extended to a modern factory covering an area up to of tens to hundreds of meters in manufacturing plants, several radio cells are required to operate simultaneously to serve the whole industrial area. At the same time, industrial applications need seamless, reliable and fast connectivity between the cloud and each individual robot in order to support high bandwidth and low latency. 4G+ and 5G can meet these requirements, since they can guarantee a smooth, seamless and lossless handover when robots move between radio cells and, at the same time, deliver good control over interfering signals from other machines and devices.

- **AGV Control Management** – On the lower layer of the architecture are the devices (AGVs) and their control management. The control functions are distributed partially on a remote cloud and partly on the AGV. The lower level functions, controlling sensors and actuators, are located on the AGV, while the rest resides in a powerful remote cloud.

  The benefit of this choice is mainly the enhancement of flexibility, exploiting the computation power of the cloud.

  The AGV computing and control architecture depicted in Figure 84 includes two computers. The vehicle controller supervises the communication with the AGV system management, controls the next step movement and handles sensors and actuators. It takes care of stopping the AGV in presence of a very close obstacle to avoid an immediate collision. It collects the LIDAR information to be sent to the AGV system management and then to the main control system for navigation and collision avoidance purposes. Collision avoidance should make use of cameras and LIDAR sensors fusion to determine the change of trajectory to avoid an obstacle.

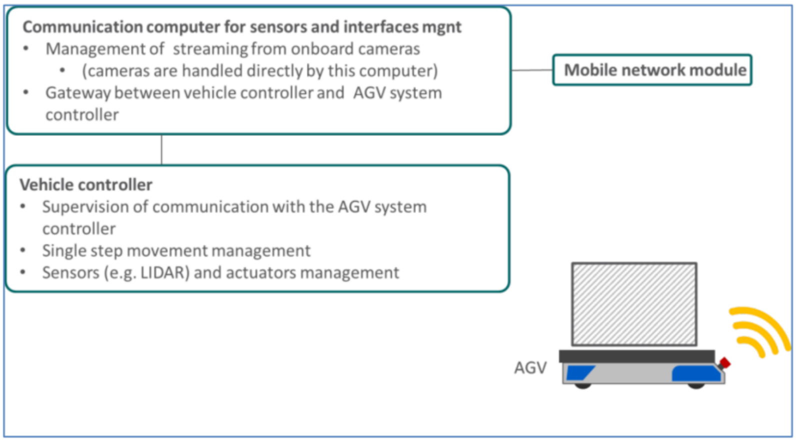

Figure 84: AGV computing and control architecture

The communication computer on the AGV will act as gateway between the vehicle controller and the AGV system management taking care of radio communication. It will also collect real time videos provided by the cameras installed on the AGV and will take care of streaming them, via the mobile network, towards the main control system for scenario recognition, navigation and collision avoidance purposes. The AGV would be equipped with at most four cameras, one per direction to have about a 360° view of the environment. The cameras are placed on the front, on the back and on both sides of the robot. The generated streams are sent to the main control system for processing and integration.

- **Mobile Cloud Remote (MCR) Control System** – The remote-control system will reside in a local cloud hosted on several VMs (see Figure 85). XEN[10] is selected as hypervisor because it has both an open source version and a commercial and professional one, so making more robust the choice for any future need.

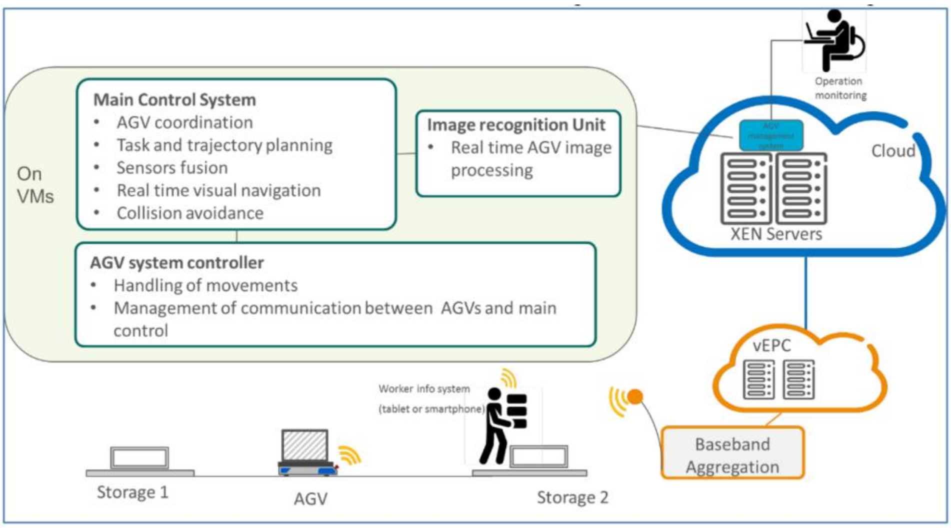

Figure 85: Mobile Cloud Remote (MCR) control system

[10] XEN Hypervisor: The open source standard for hardware virtualization https://www.xenproject.org/

The AGV system controller, the image recognition unit and the main control system resides in the cloud. The AGV system controller is a SW provided by the AGV manufacturer to handle the AGVs movements and robotic communication protocols. It behaves as an interface between the main control system and the AGV. The main control system manages the incoming service requests sent by the workers using an app running on a tablet or smartphone. A mission entails sending the AGV to a specific area to pick a box and shuttling it to a destination storage area. Once a request is received, the main control system decides the task be done by the AGV and plan its route. Then the AGV is asked to start its mission. The AGV navigation in the environment is controlled in real time by the navigation function. It relies on sensors information pre-processed by the image processing unit and the sensor fusion module. The video streams coming from the AGV are processed to detect and recognise objects and identify its position in the environment. Images will also be used to determine visual odometry information. The processed imaged data will then be integrated with the information coming from the LIDAR in the sensor fusion module. Positioning related to the environment and odometry data is used by the navigation and the obstacle avoidance modules that will compute the next movement step in real time.

The workers in the warehouse will have a tablet or a smartphone with a specific App for asking for services and monitoring operations. The tablet/smartphone is connected to the mobile network and interacts with the main control system in the cloud.

Smart City – Connected Ambulance use case architecture: One of the Smart City use cases that is demonstrated in the Greek site facility is the "Connected Ambulance" (also touching upon the e-Health vertical). The concept of the use case is the exploitation of the ambulance as a communication hub that will collect and transmit the patient's vital data and health stats, both from the accident scene as well as on route to the hospital. This will enable teleconsultation to paramedics, treatment during transport and timely informing of healthcare professionals in next point of definitive care (hospital or other health units). 5G network benefits such as capacity, coverage, reliability, mobility and virtualisation are demonstrated and verified.

As a first step, the initial architecture of the Greek site facility will aim to cover basic connectivity functionality and verification for the Nokia GR smart city use case. A high-level overview of the corresponding architecture for the initial phase is given in Figure 86.

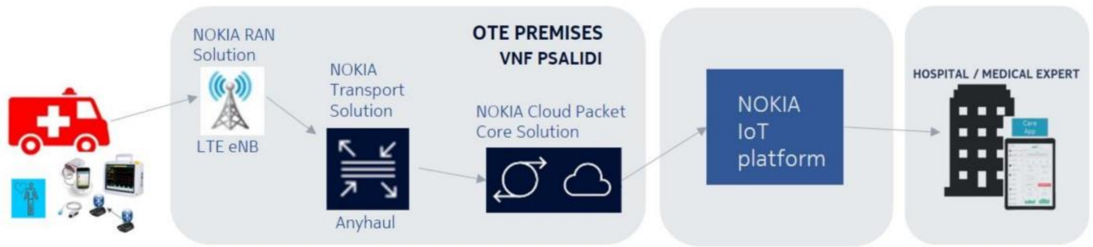

Figure 86: Connected Ambulance use case

The NOKIA IoT platform solution will provide NB-IoT compatible connectivity and traffic selection between the sensors transmitting the patient's vital data to the telecommunication network and the end user facility. Additionally, it will provide end user connectivity with IoT application(s), by which the data from the sensors and the IoT devices are collected and processed.

Utilities – Smart grid monitoring & fault detection use case architecture: The utilities use case is executed in OTE and WINGS facilities utilising the RAN, Cloud and Core components provided by the Greek facility partners and the WINGS Cloud IoT platform and specialised components developed during the project. Distributed energy production and consumptions points are monitored in real time using smart energy meters (recording multiple relevant KPIs) with programmable interfaces and a "live" analysis of the measurements using Big Data analytics techniques and AI will take place in order to achieve ultra-reliable and ultra-low latency fault detection and restoration.

At a first stage, the measurements from the various smart meters are transmitted through the live 4G+ access network provided by OTE, utilising NB-IoT access where available (specific OTE testing locations), while pre-5G and 5G components provided by Ericsson GR / Nokia GR will also be used for the transmission of the measurements with 5G connectivity at a later stage, once the Greek site facility is 5G ready. The evolution of the measurement transmission and actuation commands (GPRS / USRP, 4G+, NB-IoT, 5G) will provide significant insights and benchmarking data for such a use case, where the URLLC needs of the vertical play a crucial role to the successful completion of the use case and the satisfaction of the industry's stringent requirements. Figure 87 below depicts the high-level envisioned architecture for the utilities use case and the corresponding utilisation of the Greek 5G EVE site facility components.

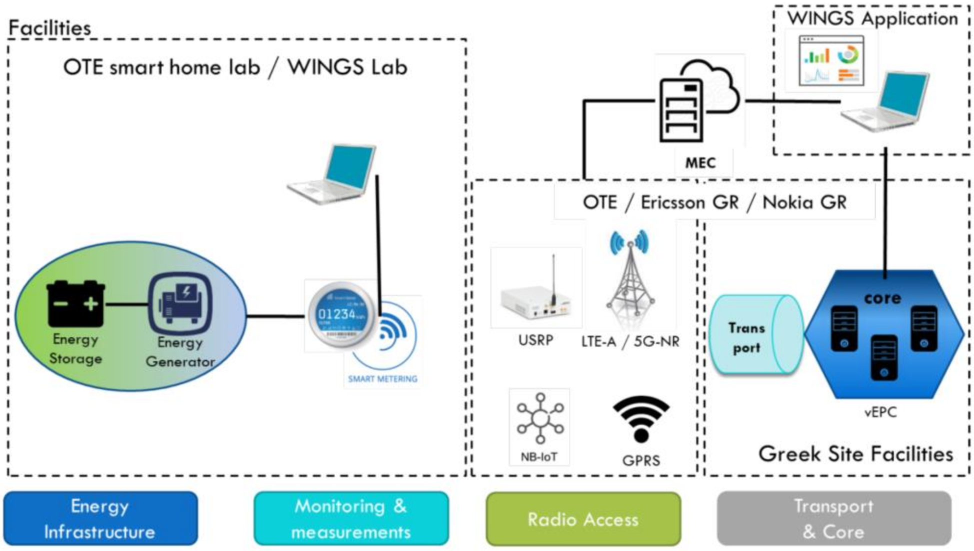

Figure 87: Smart grid monitoring use case

7.5.4 RAN Architecture Description

The RAN base is constituted of two parts: a hardware (HW) and a software (SW) base. The HW comprises the baseband node and one or more radio units depending on the manufacturing or warehouse coverage area. The software includes the components needed to operate the 3GPP wireless system including LTE (up to 3GPP Rel.14) and 5G (Rel. 15 and upwards). The baseband unit is common across different radio configurations. It provides the baseband processing resources for the encoding and decoding of the uplink and downlink radio signals, the radio control processing, the radio network synchronisation, the IP and the O&M interface for the Ericsson Radio System.

It is noted that for the first project milestone in April 2019, LTE connectivity is assumed. In the rest of the sections any reference to 5G technology is given to describe the expandability of the system in future site facility activities. The Ericsson RAN architecture builds on the Cloud RAN concept, which incorporates important aspects of both the network architecture and RAN functionality – today and on the road to 5G.

One of the key aspects in the Cloud RAN concept is Coordination and is provided by using advanced network coordination functionality such as for example, Carrier Aggregation (CA) and Coordinated Multipoint (CoMP) to maximise network performance and minimise interference. Low latency transport network as well as phase and time synchronisation are necessary components to support the time critical L1 and L2 RAN functionality. The Greek site facility architecture is based on the Centralised RAN approach.

Warehouse cellular coverage to enable the AGVs service is provided by deploying Ericsson's indoor Radio Dot System (RDS) solution. This is a high performance distributed active radio antenna system based on a centralised RAN architecture. A simplified diagram is shown in Figure 88.

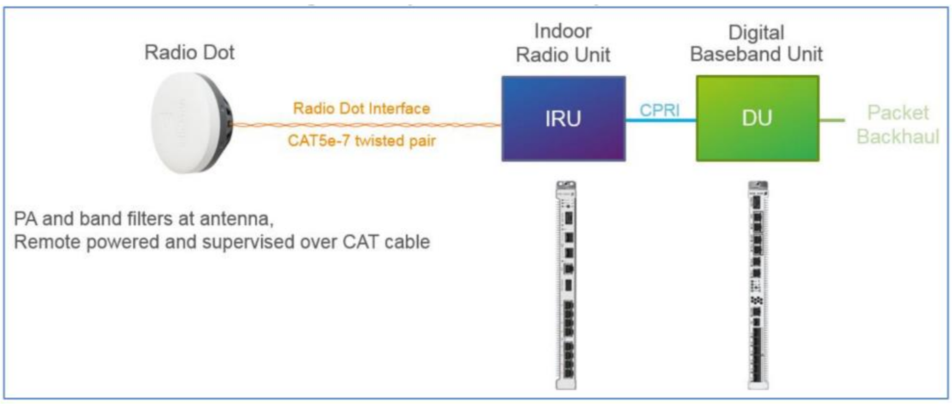

Figure 88: Ericsson Radio Dot System (RDS)

The Ericsson RDS consists of 3 key components:

- **Radio Dot (RD)** – It contains the power amplifier and filters for the frequency band(s). RDs are power fed from the Indoor Radio Unit (IRU) over up to 200 m LAN cabling. It is designed for deployment in indoor environment, in single, dual band and 5G variants.
- **Indoor Radio Unit (IRU)** – The IRU provides the power and control for the RDS. It instantiates the RD interface on 8x RJ45 ports and connects to the Dots over standard enterprise LAN cables.
- **Digital Unit (DU) or Baseband** – The Baseband connects to the IRUs over the CPRI interface. The Baseband runs the 4G+ SW features. It supports key coordination features for running small cells in large multi-antenna indoor environments. Features include Combined Cell, Carrier Aggregation, Lean Carrier, Uplink Comp. The Baseband provides synchronisation and transport security functionality and aggregates the radio traffic onto a common backhaul connection.

The provided RAN system is fully compliant with 3GPP R15 and later. The first phase of RAN SW/HW deployment will comprise functionality compliant to LTE Advanced Pro technology included up to 3GPP R14 specifications. The initial RDS architecture is also in line with the radio network architecture of 5G. It can coexist with pure 5G NR by including additional RDs optimised for 5G NR. The system has been upgraded to include 5G NR SW/HW compatible to 3GPP R15 specifications, Option 3x.

A summary of the system's HW/SW capabilities can be seen in Figure 89.

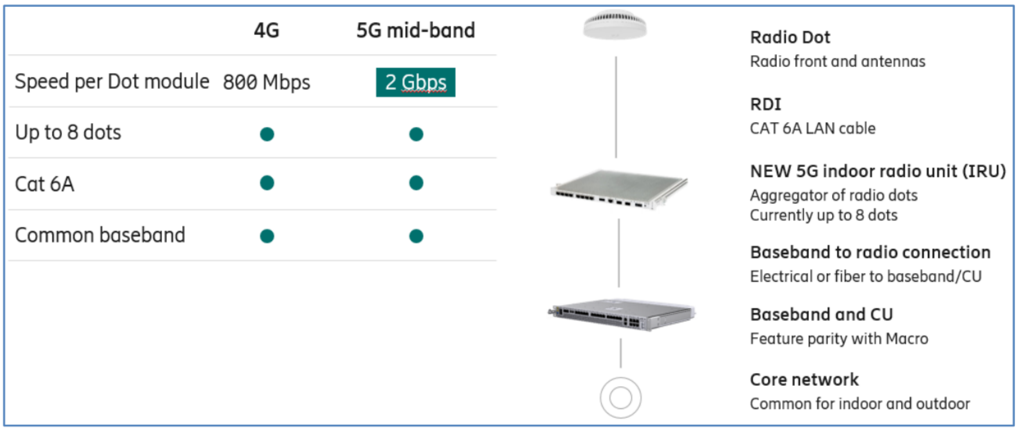

Figure 89: RDS Specifications

RRU: The Radio Remote Head (RRU) solution that will initially be used offering 4G+ coverage is the (RD) type RD2243 which is a low power radio transmitter designed to provide superior indoor mobile broadband coverage along with the Indoor Radio Unit (IRU). The RD2243 is a single band RD antenna with two Rx/Tx antenna branches. The RD uses the signal and power interface from the IRU over the Radio Dot Interface (RDI). Up to eight RDs can be connected to a single IRU. Using one Radio Dot and LTE 20 MHz spectrum capacity the following peak speed up can be delivered:

- 200 Mbps using MIMO 2x2 system (CAT4 device)
- 400 Mbps using MIMO 4x4 system (CAT9 device)
- The RD provides the following main features:
- Output power 2x17 dBm (2x50 mW) in a MIMO 2x2 configuration
- Higher Order Modulation (HOM) 256-QAM Downlink and 64-QAM Uplink
- Total Instantaneous Bandwidth IBW 40 MHz, 2x LTE FDD carriers of 5, 10, 15, 20 MHz
- Frequency conversion and amplifying functionality
- Single RDI cable for power and signal transmission - the RD has the power and the signal on the same LAN cable, so only one cable is required for the Dot.

Complementing the 4G Dot, RD4479 is a single band 5G RD with four Rx/Tx antenna branches which is used in the second phase to provide NR coverage offering peak downlink speed up to 2 Gbps as shown in Figure 90. The 5G NR RD delivers speeds up to 2Gbps and supports the new 5G mid-band 3-6GHz (n78 compatible). It fully reuses the currently installed architecture for 4G+ RD with the ease of installation and flexibility and functional parity with the macro network.

Figure 90: 5G NR Radio Dot

The IRU is an indispensable component of the overall RRU architecture and it provides baseband processing from a DU or Baseband unit, transceiver processing to the RD over the RDI and supplies power to the RD over the RDI. An analytical block diagram of the RDS frontend system is depicted in Figure 91.

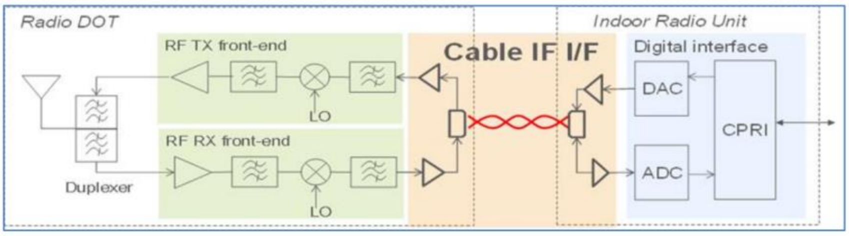

Figure 91: RDS front-end system

As shown in Figure 91, an IF-based design effectively extends the RD RF frontend over a LAN cable with an IF cable frontend interface, transporting the radio signals at low frequencies. The elegance and simplicity of such a design enables a neat and ultra-compact design of the RD. The design allows much lower power consumption compared to current Ethernet-based methods, and in addition to the radio and sync signals, control signalling, and power are provided by the same pairs. The RDS architecture design enables advanced features for cable equalisation, Automatic Gain Control (AGC), cable testing, troubleshooting, and different SON features adding value to the innovation. Further, it can support Single User (SU-MIMO 4x4) by utilising all four pairs of the cable which makes the RDS architecture fully compatible with 5G NR 3GPP R15.

The dimensioning of the RDS depends on the achievable indoor dominance coverage level and required cell capacity. A general recommendation for achieving a balanced UL/DL link is that a single RD antenna can provide coverage to a range between 500 to 800 m^2 area. Since one IRU can be connected to 8 distributed RDs, one IRU can provide coverage to a range of 4000 to 6000 m^2 space approximately.

BBU – Baseband Unit 6630: The baseband processing for the uplink and downlink of LTE and NR is provided by the baseband unit 6630 as shown in Figure 92. The RDS centralised baseband architecture enables coordination across the covered area. The baseband HW along with the interfaces front panel is shown in Figure 92 below, while its specifications are presented in Table 9.

Figure 92: Baseband Unit 6630 front panel

Table 9: Baseband Unit 6630 specifications

Baseband specifications for existing LTE SW (L18.Q3)	Baseband 6630
Downlink maximum throughput (Mbps)	2000
Uplink maximum throughput (Mbps)	500
Number of VoIP users (FDD/TDD)	2000/1000
Number of connected users	8000
Aggregated antenna bandwidth (MHz)	960 – 1560
Number of Cell (FDD/TDD)	24
CPRI radio interface - Hardware Prepared for NR (5G) and e-CPRI	15x SFP/SFP+
Transport interface, optical 1/10Gbps SFP/SFP+ ports	2

Frequency bands: The frequency bands that are used for 4G+ Ericsson RAN solution as described in this section are:

- The B7 which is, FDD 2600 MHz with 20 MHz spectrum deployment, based on the selection criterion to minimise interference and overlapping coverage with the commercial OTE/COSMOTE MBB network.
- The B38 (TDD, 2600 MHz) and B42 (TDD, 3500 MHz) are considered for deployment of the 5G-NR in order to provide 5G access at the second stage of the Greek site facility upgrades.
- The frequency bands that are used by the Nokia RAN solution (Nokia LTE eNB) in this initial phase are:
- Band Class 1 LTE access at 1900MHz for UL and 2100 MHz for DL,
- For 5G access available access at 3.4-3.8 GHz is considered, depending on use case evolution.

Subscriber Separation: To ensure a good performance control of the AGV service, a separation among mobile broadband subscribers and smart facility subscribers needs to be in place. The smart facility service will only target the enterprise user providing a viable alternative to fixed access (e.g. typically xDSL, fiber) with comparable or better quality. Subscriber separation solution main goals are:

- Ensure that mobile broadband subscribers will stay only on macro LTE FDD without camping/using enterprise network resources.
- Ensure that enterprise users will stay only on the indoor LTE FDD network without camping/using the production network.
- Cause no degradation service level to enterprise usage.

7.5.5 Distributed Cloud / MEC

Ericsson solution: Distributed Cloud covers a number of aspects of local access in a small-scale system. It is typical that the operator distributed network is connected to the operator's central network via a poor (i.e. limited bandwidth and/or high latency) backhaul, e.g. satellite or E1 links. The operator distributed network may be connected to one or several different types of networks local to the distributed site. This distributed cloud provides a small, optimised solution for best possible end user experience with local network connectivity. It enables carriers to address remote areas with very high transport costs, but also to provide a better user experience increasing the MBB subscription growth possibilities.

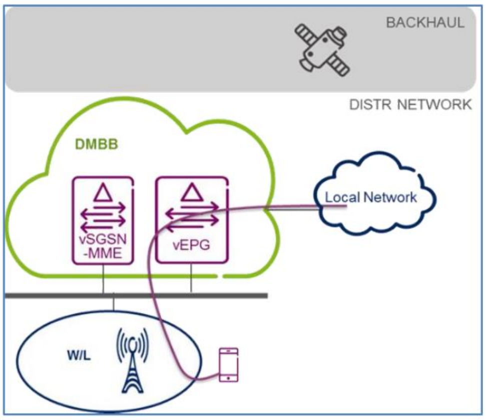

Figure 93: Distributed Cloud

The user experience is enhanced by using local network breakout (see Figure 93). This is possible thanks to distributing the vEPG. In this case the mobile operator can offload traffic from the backhaul by routing traffic directly to locally available internet access or peering partners. The distributed vEPG has full control of policy and charging for connections and services, which reduces payload over the poor backhaul.

Nokia solution: The NOKIA Cloud Packet Core at OTE premises will comprise of a NOKIA Cloud Mobility Manager (CMM) solution, as well as a Cloud Mobile Gateway (CMG) solution, with 3GPP Rel-14 compatibility in terms of connectivity with the IoT platform. The physical resources and infrastructure for the VNF deployed in the cloud is the NOKIA Airframe. Also, the required HSS packet core solution is provided through a Linux based emulator, which is NOKIA product compatible, providing required real-time aspects and offering real physical interfaces towards other network elements.

7.5.6 CORE Architecture Description

The solution for the core part of the Greek site facility is based on a 5G virtual EPC. Nevertheless, Radio Access Network connectivity for the first phase of the site facility is LTE based. A 5G EPC-in-a-box is proposed that fulfils the requirements for cost-effective test systems with few subscribers and minimal footprint. It is a further evolution of the Virtual Network Function (VNF) single server deployment enabling multiple VNFs on a single server. The deployment contains vEPG, vSGSN-MME and vSAPC.

EPC-in-a-box, as depicted in Figure 94, is designed to run on top of Ericsson OpenStack IaaS i.e., Cloud Execution Environment (CEE) and can use either HDS 8000 CRU or Dell 630 as HW. It is also tuned to be as efficient as possible when all VNFs are running at the same time. However, there is no requirement that all VNFs must be deployed. For example, it is possible to only deploy vEPG.

Deployment Overview: EPC-in-a-box deployment is built on Ericson Cloud Execution Environment (CEE) which includes the following functions necessary for EPC-in-a-box:

- Virtualised CIC (Cloud Infrastructure Controller)
- Support to run the vCIC in a non-redundant single-vCIC mode
- Hyper-Threading
- Pinning of the VM vCPUs to specific Hyper Threads (HTs), ePC, slicing

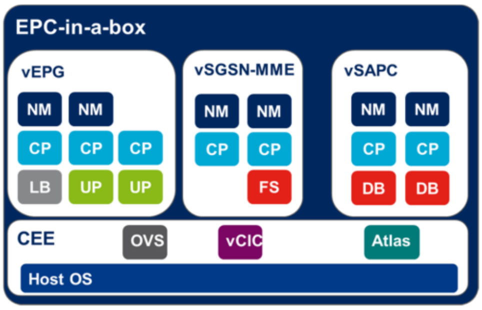

Figure 94: EPC-in-a-box

VM Distribution: EPC-in-a-box can be deployed on Ericsson HDS 8000 CRU hardware or Dell 630 COTS HW. Servers with minimum 12 Core processors are required. Each core has 2 Hyper-Threads. The VMs can be divided into two types:

- VMs requiring dedicated pCPU – these VMs have CPU resource intensive processes. Payload handling processes require dedicated pCPUs. These processes are always active (busy-looping), polling for incoming packets.

- VMs that can share pCPUs – these VMs are not CPU resource intensive. When there is nothing to process, they are not taking up any processor resources. Therefore, the vCPUs of these VMs can share a range of pCPUs.
- The technique used to allocate VMs to compute resources is called CPU pinning. EPC-in-a-box specifies in detail which pCPUs each VM should use. Without specifying anything CEE will have a 1:1 (v:p) ratio between vCPUs and pCPUs and assign pCPUs to vCPUs in a dynamic manner. CPU pinning is done in one of the following ways:
- Dedicated pCPU – A vCPU is pinned to one specific pCPU. The pCPU and memory allocation are optimised based on the VM type, vCPU role and the NUMA topology of the host.
- Shared pCPU – A vCPU is pinned to a specific range of pCPUs, where other vCPUs (within same VM and between different VMs) are pinned to same pCPU range. In this case the hypervisor scheduler decides dynamically which pCPU is used for each vCPU (i.e. vCPUs are floating). Sharing pCPUs between multiple vCPU is also known as CPU overcommit/oversubscription.

The EPC-in-a-box deployment provides scripts to generate the correct resource allocation by creating specific CEE flavours for each VM and a HOT file to deploy each VNF. The CPU pinning is controlled via OpenStack flavours. Figure 95 shows a high-level view of which VMs use dedicated respectively shared pCPUs. The gradient boxes show that some VMs use both dedicated and shared pCPUs. The VMs that are using both dedicated and shared pPCUs are still categorised into dedicated or shared group depending on if the majority of the VM vCPUs are using dedicated or shared pCPUs. E.g. the Load Balancer’s VMs are mainly using dedicated pCPU while Control Plane VMs mainly use shared pCPUs.

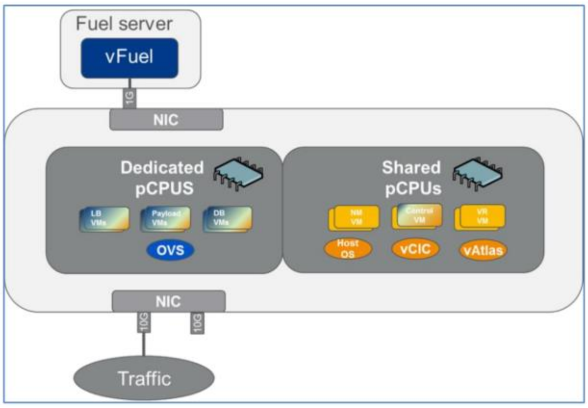

Figure 95: EPC-in-a-box CPU allocation

EPC Network Functions: The following key network functions are affected:

- MME (Software functionality of proposed SGSN-MME)
- S/PGW (Software functionality of proposed EPG)
- PCRF (evolved)

MME enabled for 5G: The SGSN-MME software has been enhanced to include support for 5G. The same platform and architecture are used with added 5G access and related functionality.

7.5.7 Orchestration

ETSI NFV MANO: The NFV management and orchestration is ETSI NFV MANO compatible and is provided by NOKIA CloudBand product family, namely CloudBand Network Director (CBND), CloudBand Application Manager (CBAM), and CloudBand Infrastructure Software (CBIS), as illustrated in Figure 96.

The CBND is an NFV resource and network service orchestrator (NFVO), built for OpenStack and VMware. It manages virtual resources across geo-distributed NFV infrastructure nodes, if needed. The CBAM is a VNF Manager (VNFM) built for OpenStack and VMware, which automates VNF lifecycle management actions by managing resources and applying associated workflows. The CBIS is a multi-purpose NFV infrastructure (NFVI) and virtualised infrastructure manager (VIM), built for OpenStack. It virtualises and manages compute, storage, and network resources. (Note: interfaces are MANO specific and can be described in one place for all testbeds).

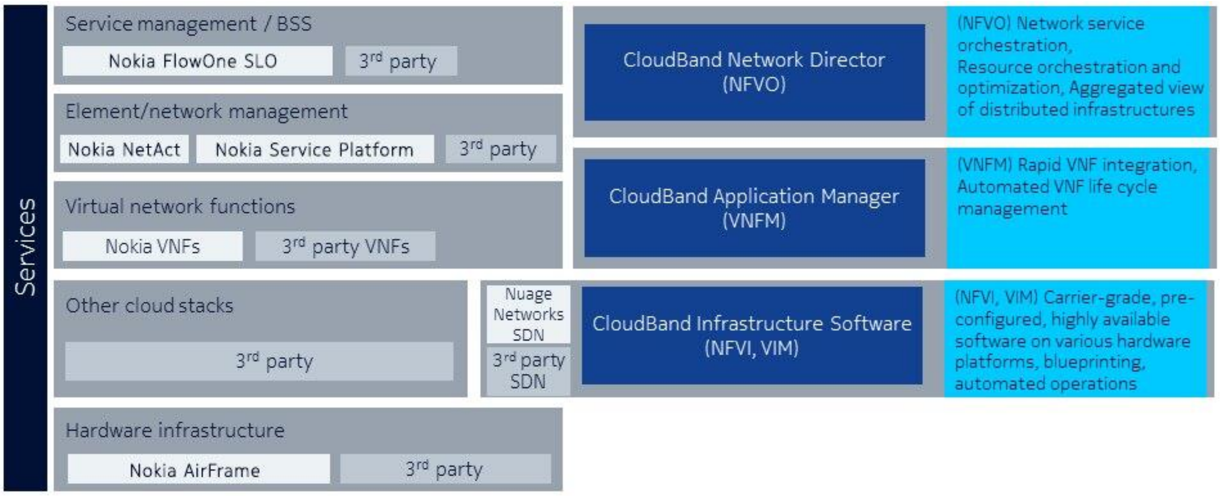

Figure 96: NOKIA CloudBand MANO

Network management transport: Both Nokia GR and Ericsson GR have deployed transport capabilities at OTE premises and the respective use case pilot sites. In the case of Nokia GR, the transport solution between RAN and Cloud Packet Core resides at OTE premises, combining different transport technologies, e.g. IP, optical, microwave. In the case of Ericsson GR, the data exchanged between the AGVs and the higher-level control function will rely on more than one L3 protocol. The traffic flow is mainly TCP and UDP based. So, they can travel directly on the mobile network without the need for further encapsulation. Traffic encapsulation could be required for some low-level communication between the AGVs and the upper layers. In such a case VXLAN or a VPN could be adopted.

7.5.8 Interfaces

The common interfaces that are utilised in the initial phase are illustrated in Figure 97.

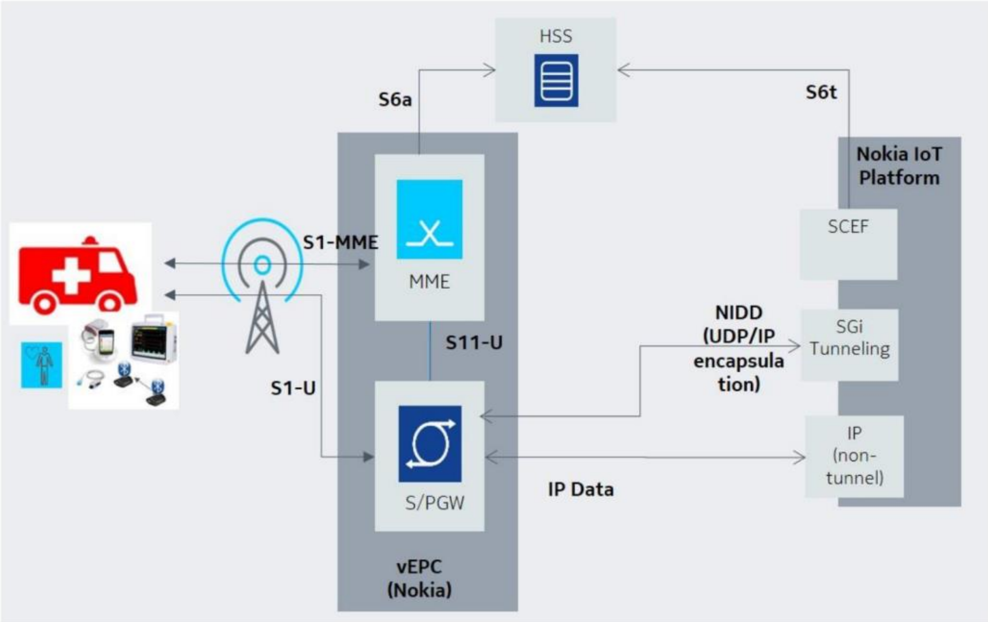

Figure 97: Initial phase interfaces

More specifically, those interfaces are:

- S1: eNB to MME and S-GW.
- S11: MME to S-GW.
- S6a: MME to HSS.
- S5: S-GW to P-GW.
- S6t: HSS to SCEF.

7.5.9 Development and Deployment Tools/Software

Integrated Base station Gateway (BGW) functionality in EPC-in-a-box for networking: EPC-in-a-box includes an integrated Base Station Gateway (BSW) functionality. This functionality is provided by the vEPG, which is based on the Ericsson Virtual Router (EVR) platform. By having an integrated BGW, the vEPC-in-a-box can be deployed without relying on an external BGW. This means that communication between VNFs within EPC-in-a-box do not traverse an external BGW. VNF communication to external networks traverses the integrated BGW (vEPG). The vEPG is configured with a single LB using a single vNIC, which is shared between the EPG functionality and the BGW functionality. The single vNIC is used for both vEPC internal as well as external traffic. This is made possible by using trunk VLANs.

The EPC-in-a-box internal networks between vSGSN-MME, vSAPC and the vEPG/EVR are preconfigured (see Figure 98). By default, a set of different Neutron networks are used to separate different logical networks. External traffic to and from the vEPG is routed by the vEPG through its LBs as usual. Between EVR and vSGSN-MME static routing is used. This static routing is preconfigured. BFD is used to supervise the static routes. All the service IP addresses of the vSGSN-MME need to be configured as static routes in the EVR.

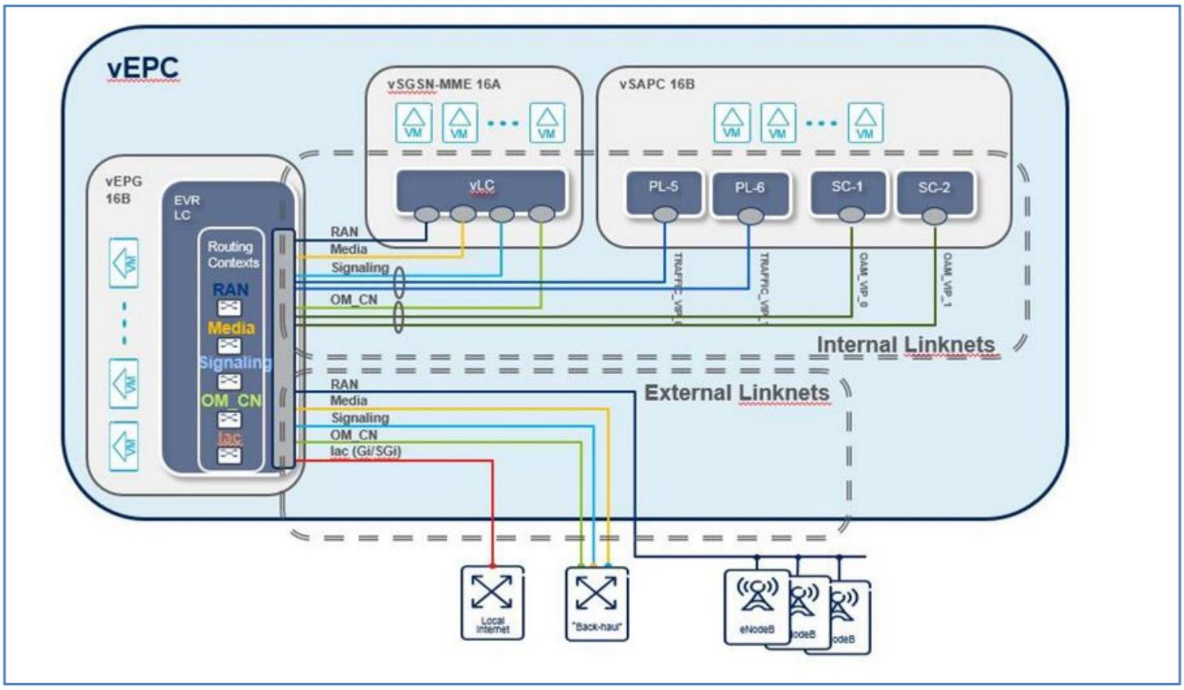

Figure 98: vEPC-in-a-box VLANs

OSPF routing is used between EVR and vSAPC. OSPF is used both for advertising the SAPC VIP addresses to EVR and to supervise the connectivity. All of the configurations for the vEPC-in-a-box internal routing, both static and OSPF, is preconfigured.

Externally VLAN separation is used for logical network separation. The box is preconfigured with a default set of networks which may be changed. In the future, it would be possible to use BGP/MPLS for network separation as well. The supported routing protocols for external connectivity are static routing, OSPFv2, OSPFv3 and BGP. All of these may be used together with BFD for fast failure detection.

Cloud Server Environment: The cloud part will make use of XEN server. XEN is a mature platform that has both an open source and a commercial proven enterprise use and its hypervisor was chosen for the cloud by Amazon and Google. A control management SW, XENcenter, is available for setting up and monitoring the cloud based on XEN server. However, the XEN environment comes with an API, provided in several programming languages that can be used to define a specific cloud management.

The App used by the workers on their tablet/smartphone to ask for services and monitoring is an Android application. It was developed using the Android Studio environment. The control and processing functions running in the cloud was developed mainly in python using the Anaconda environment. Specifically, dedicated libraries are used for processing acceleration. Numpy and Scipy are used for heavy mathematical operation and ski-image and OpenCV for the image processing. The FLANN library is used for implementing approximate nearest neighbour functions required by image recognition.

The application for monitoring and statistics could make use of relational Data Bases. In this case, mySQL is adopted. The COMAU's AGV comes with its own SW. It includes the vehicle system controller, the AGV simulator and an operator interface SW.

Radio Access System SW: The initial RAN SW builds on the Ericsson's 5G plug-ins. These are software-driven innovations that bring essential 5G technology concepts to today's 4G+ cellular networks enabling a flexible 5G evolution as well as improving operators' network mobile broadband performance allowing to introduce an array of new services and applications. The 5G plug-ins are built on 3GPP R13/15 specifications.

- **Reduced latency 5G plug-in for critical communications** – Smart warehouse facility implementation based on mobile cloud robotics requires high-resolution cameras and LIDAR sensors installed on the AGVs in order to enable real-time, autonomous operation controlled by cloud-based systems. The facility operator is able to dynamically assign different tasks to the robots and control their execution. The robots, being equipped with various sensors and actuators do not need guides on the floor and are able to avoid collisions with people and objects thanks to real-time remote processing of sensor data and images.

  Of paramount importance is a high-performance mobile network connecting the robotic vehicles to the cloud-based control system. For example, a UL speed of 60 Mbps or higher, maximum acceptable jitter of less than 5 ms, no data buffering and end-to-end latency of less than 10 msec is required to ensure seamless and safe operation. Ericsson's proposed 5G plug-in SW solution can meet these requirements by deploying advanced network functionality which greatly improves network performance. One of the proposed mechanisms is the reduced latency functionality which builds on two steps developed in 3GPP R14 and 15 specifications. Specifically, R14 concept of Instant Uplink Access (IUA) eliminates the need for explicit scheduling request and individual scheduling grants. Through pre-allocation of radio resources IUA can reduce the average radio Round Trip Time (RTT) latency (i.e., UL and DL) to 9 ms, which is a significant improvement compared to traditional LTE R13 RTT latency of 16 ms (Figure 99).

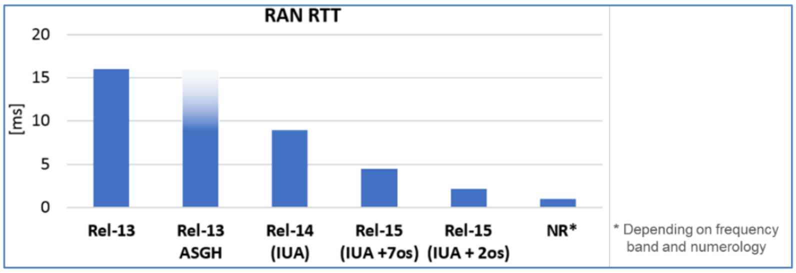

Figure 99: Reduced latency achieved with Ericsson 5G plug-ins

  The second method, which is specified in 3GPP R15, enables shorter transmission durations as illustrated in Figure 100. The concept is to compress the whole transmission chain while waiting for a transmit opportunity and transmitting the data. The associated control and feedback are performed faster.

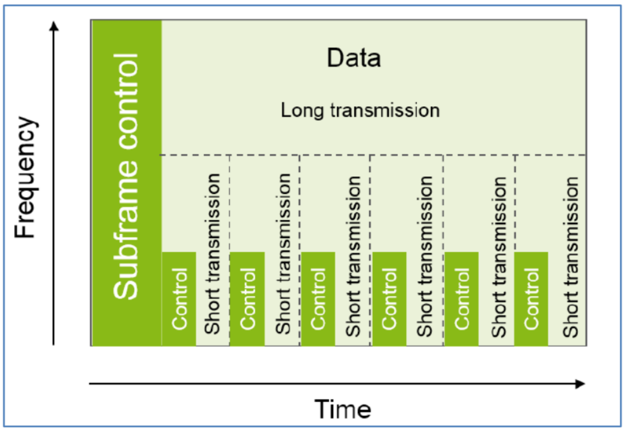

Figure 100: LTE downlink subframe with existing long and short transmissions

The compression is done by introducing transmissions with duration shorter than a subframe. In the downlink this is done by splitting the data part of the subframe into several parts. Each of these short transmission durations can be scheduled separately with a new in-band control channel. In addition to the downlink, the uplink subframe is split into multiple shorter transmission durations and is scheduled from the same in-band control channel. The subframes are either split into two parts, four parts or into roughly six parts for the lowest latency mode. At the highest splitting level, a one-way transmission can be done in a total of about 0.5 ms including processing of data.

QoS Mechanisms: The following QoS mechanisms are required on the site router and Mobile Backhaul network:

- **Classification and marking** – Traffic classification is based on the DiffServ model and IP packets are marked with a DSCP value. A DSCP value is configurable for most traffic by virtual EPC VNFs. The Site Router maps the DSCP value of an IP packet to its internal QoS parameters and treatments. For traffic lacking a DSCP value (for example Gx, Gy Rf and Gom), the Site Router is responsible for traffic classification with a suitable internal QoS priority and even marking the DSCP value for the IP packet. For cross-site traffic over an MPLS network, the QoS propagation to MPLS TC/EXP should be done by the Site Router.
- **Rate-limiting** – Rate-limiting should be set on the Site Router to reduce the risk of starving low-priority traffic and overloading the whole virtual EPC system (10 GE traffic interface versus 1Gbps throughput of the EPC capacity).
- **Queuing and Scheduling** – The Queuing and Scheduling are done on the Site Router based on different traffic classes. There is Strict Priority scheduling with bandwidth limit for inelastic traffic while WFQ/PWFQ scheduling with optional WRED for elastic traffic may also be applied.

7.5.10 Testing and KPIs

Table 10 shows the KPIs and tools that are used to assess the performance of the end-to-end system.

Table 10: Testing KPIs and tools

KPI	Definition	Tools
E2E DL latency	End to end mobile network downlink latency	Wireshark connected on both ends simultaneously and dedicated analysis SW to be developed
E2E UL latency	End to end mobile network uplink latency	Wireshark connected on both ends simultaneously and dedicated analysis SW to be developed
E2E DL latency variation	End to end mobile network downlink latency variation	Wireshark connected on both ends simultaneously and dedicated analysis SW to be developed
E2E UL latency variation	End to end mobile network uplink latency variation	Wireshark connected on both ends simultaneously and dedicated analysis SW to be developed
Uplink buffering delay	Buffering delay introduced by the network on camera streams sent by AGV to main control in uplink direction	Timing difference between packets coming through the mobile network and a direct cable connection to the camera. Wireshark and the dedicated SW developed for the latency analysis will be used
Number of collisions vs. number of obstacles	Ratio between the number of collisions of the AGV with respect to the number of encountered obstacles	Heuristic metric based on observations (use case specific)
Number optimal cruises vs. Total cruises	Ratio between the number of AGV travels following the optimal path with respect to the number of executed tasks	Heuristic metric based on observations (use case specific)

7.5.11 Testbed Interconnection

The Greek 5G EVE site facility is based on the existing OTE infrastructure and as such it may use the established 4G+ network of OTE to interconnect to the 'outside world' and to access / be accessed by external servers (i.e. using GEANT[11] based interconnection available).

In order to meet the strict deployment and operational schedule of the 5G Athens facility and to make sure that all three engaged use cases (see Section 2.1) are supported during the initial roll-out of the 5G EVE facilities, the Nokia GR and Ericsson GR respective equipment / feature installation and network upgrades will take place in parallel, each connecting to OTE's main facilities. At the first stage, there is no interconnection between the Ericsson GR and Nokia GR segments and the two network segments will operate separately (from RAN access to used vEPCs), each supporting the allocated use cases. At a later stage, and once the individual use cases are supported by the Greek facility, an interconnection of the two segments will take place.

[11] https://www.geant.org/

8. CONCLUSION

8.1 Measurement Results and Characterisation

Measurements from the testbed were obtained with the iPerf[12] software. A tabulated form of the results can be found in the Annex section of this thesis.

The peak data rates obtained were 1280 Mbps and 59 Mbps for the downlink (DL) and uplink (UL), respectively. On the other, the user data rates were 1150 Mbps and 51 Mbps for DL and UL, respectively.

The DL user date rate value is remarkable because it is more than 10x the minimum user data rate (100 Mbps) expected in 5G. However, the DL peak data rate is a far cry from the expected 5G peak data rate of 20 Gbps. This is understandable since the measurements were made indoors. It is also understandable that the UL values have much lower since the UL is traditionally the bottleneck in wireless communication. Moreover, the data rates could be much higher if the channel bandwidth is increased beyond its present value of 100 MHz. 5G is expected to have up to 800 MHz of contiguous spectrum.

The end-to-end (E2E) latency and round-trip time (RTT) were 5 ms (+-1 ms) and 10 ms (+-2 ms). The E2E latency (5 ms) is a 100% improvement over that of 4G (10 ms). However, further upgrades can be made to get closer or even attain the 1 ms over-the-air latency expected for a 5G system.

Finally, the system had 99% and 98% average availability and reliability, respectively.

Measurements were not conducted for capacity, device connection density, mobility, energy efficiency or spectrum efficiency.

This is because the test focus was on the Automated Guided Vehicle (AGV) use case, which is a URLLC scenario use case. In URLLC scenarios, low latency is of highest importance while high data rates could be less important, and the results obtained reflect this.

8.2 Recommendations for Future Work

In future measurements, I would recommend the inclusion of the measurements for the other key performance indicators that were excluded. Doing this will provide a complete picture of the 5G system. However, with values obtained, the testbed can effectively handle URLLC use cases.

My final recommendation is that outdoor testing should be performed as well to understand what the degree of improvements in the data rates will be.

[12] https://iperf.fr/

ABBREVIATIONS - ACRONYMS

3GPP	3rd Generation Partnership Project
5GC	5G Core
AAS	Active Antenna System
ADC	Analogue-to-Digital Converter
AF	Application Function
AMF	Access and Mobility Management Function
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuit
AuF	Authentication Function
AUSF	Authentication Server Function
BER	Bit Error Rate
BPF	Beamforming Processing Function
BWP	Bandwidth Parts
CA	Carrier Aggregation
CAGR	Compound Annual Growth Rate
CAPEX	Capital Expenditure
CATR	Compact Antenna Test Range
CMOS	Complementary Metal Oxide Semiconductor
CoMP	Coordinated Multipoint
CPE	Common Phase Error
CP-OFDM	Cyclic Prefix Orthogonal Frequency Division Multiplexing
CSI-RS	Channel State Information Reference Signal
CSP	Communications Service Provider
DAC	Digital-to-Analogue Converter
DFT-s-OFDM	Discrete Fourier Transform spread OFDM
DL	Downlink
DMRS	Demodulation Reference Signal
DUT	Device Under Test
E2E	End-to-End
eMBB	Enhanced Mobile Broadband
EMC	Electromagnetic Compatibility
EMM	EPS Mobility Management

EPC	Evolved Packet Core
EPS	Evolved Packet System
ESM	EPS Session Management
EVM	Error Vector Magnitude
FDD	Frequency Division Duplex
FPGA	Field Programmable Gate Array
FR	Frequency Range
FTTH	Fibre-to-the-Home
gNB	gNodeB
GSM	Global System for Mobile Communications
HAPS	High Altitude Platform System
HARQ	Hybrid Automatic Repeat Request
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High-Speed Uplink Packet Access
ICT	Information and Communication Technology
IMT	International Mobile Telecommunications
IoT	Internet of Things
ITU	International Telecommunication Union
ITU-R	ITU Radiocommunication Sector
KPI	Key Performance Indicator
LPWA	Low Power Wide Area
LTE	Long Term Evolution
MAC	Medium Access Control
MANO	Management and Orchestration
MIMO	Massive Input Massive Output
MMCF	Mobility Management Control Function
MME	Mobility Management Entity
MMIC	Monolithically Integrated Circuit
mMTC	Massive Machine Type Communication
mmWave	Millimetre Wave
MNO	Mobile Network Operator
NB-IoT	Narrowband IoT

NEF	Network Exposure Function
NF	Network Function
NFV	Network Function Virtualisation
NFVI	Network Function Virtualisation Infrastructure
NRF	Network Repository Function
NSSF	Network Slice Selection Function
OFDM	Orthogonal Frequency Division Multiplexing
OPEX	Operating Expenditure
OTE	Hellenic Telecommunications Organisation
PBCH	Physical Broadcast Channel
PCF	Policy Control Function
PCRF	Policy and Charging Rules Function
PDCP	Packet Data Convergence Protocol
PDN	Public Data Network
PDU	Protocol Data Unit
PF	Policy Function
PGW	PDN Gateway
PHY	Physical Layer
PMP	Point-to-Multipoint
PNF	Physical Network Function
PoC	Proof-of-Concept
PPF	Packet Processing Function
PRACH	Physical Random-Access Channel
PRB	Physical Resource Block
PTRS	Phase Tracking Reference Signal
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
R&D	Research and Development
RAT	Radio Access Technology
RCF	Radio Control Function
RF	Radio Frequency
RFIC	Radio Frequency Integrated Circuit
RLC	Radio Link Control

RRC	Radio Resource Control
RRM	Radio Resource Management
RRU	Remote Radio Unit
RTT	Round Trip Time
SBA	Service-Based Architecture
SCS	Subcarrier Spacing
SDBF	Subscriber Database Function
SDN	Software Defined Networking
SFI	Slot Format Indicator
SGW	Serving Gateway
SMCF	Session Management Control Function
SMF	Session Management Function
SMSF	Short Message Service Function
SRS	Sounding Reference Signal
SUT	System Under Test
T/R	Transmit/Receive
TDD	Time Division Duplex
TSG RAN	Technical Specification Group Radio Access Network
TTI	Transmission Time Interval
UCA	Uniform Circular Array
UDM	User Data Management
UE	User Equipment
UHD	Ultra-High Definition
UL	Uplink
ULA	Uniform Linear Array
UMTS	Universal Mobile Telecommunications Service
UPF	User Plane Function
URA	Uniform Rectangular Array
URLLC	Ultra-Reliable Low Latency Communication
V2X	Vehicle-to-Everything
VM	Virtual Machine
VNF	Virtual Network Function
VNI	Visual Network Index
VR	Virtual Reality

ANNEX

A. Testbed Measurement Data

		Greek site	
		OTE/Ericsson	Capability
Enabled 5G KPI Metrics: - Manual(M): Avail. Date/MS - Automated (A): Avail Date/MS	User Data Rate	M: MS8 A: MS9	DL: = 1150 Mbps B/W with iperf for 100 Mhz band UL: = 51 Mbps B/W with iperf for 100 Mhz band
	Peak Data Rate	M: MS8 A: MS9	DL: = 1280 Mbps B/W with iperf for 100 Mhz band UL: = 59 Mbps B/W with iperf for 100 Mhz band
	E2E Latency	M: MS8 A: MS9	RTT(avg): 10.0 ms +- 2.0 ms E2E(avg): 5.0 ms +- 1.0 ms RTT < 15 ms
	Reliability	M: MS8 A: MS9	98%(avg;for lat<25ms)
	Availability	M: MS8 A: MS9	99% (avg)
	Capacity (MS10)	M: MS10 A: MS10	
	Device Density (MS10)	M: MS10 A: MS10	
	Mobility (MS10)	M: MS10 A: MS10	

B. Photos of Testbed Components and Devices

Automated Guided Vehicle (AGV)

Ericsson Radio Dot

Sony 5G mmWave UE (front and back)

REFERENCES

[1] Qualcomm, “Making 5G NR a reality,” Qualcomm Technologies, Inc, Dec 2016. Available: https://www.qualcomm.com/media/documents/files/whitepaper-making-5g-nr-a-reality.pdf

[2] GSMA, “Road to 5G: Introduction and Migration,” GSMA, Apr. 2018. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2018/04/Road-to-5G-Introduction-and-Migration_FINAL.pdf

[3] T. S. Rappaport et al., "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," in IEEE Access, vol. 1, pp. 335-349, 2013

[4] A. F. Molisch, M. Steinbauer, M. Toeltsch, E. Bonek, and R. Thoma,‘‘Capacity of MIMO systems based on measured wireless channels,’’ IEEE J. Sel. Areas Commun., vol. 20, no. 3, pp. 561–569, Apr. 2002.

[5] J. Fuhl, A. F. Molisch, and E. Bonek, ‘‘A unified channel model for mobile radio systems with smart antennas,’’ Proc. Inst. Electr. Eng.-Radar, Sonar Navigat., Special Issue Antenna Array Process. Tech., vol. 145, no. 1, pp. 32–41, Feb. 1998.

[6] S. Rajagopal, S. Abu-Surra, Z. Pi, and F. Khan, ‘‘Antenna array design for multi-Gbps mmwave mobile broadband communication,’’ in Proc. IEEE Global Telecommun. Conf., Dec. 2011, pp. 1–6.

[7] ITU, “Setting the Scene for 5G: Opportunities and Challenges,” ITU, 2018. Available: https://www.itu.int/en/ITU-D/Documents/ITU_5G_REPORT-2018.pdf

[8] J. G. Andrews et al., "What Will 5G Be?," in IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1065-1082, June 2014

[9] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," in IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, Fourth Quarter 2017

[10] B. Clerckx, A. Lozano, S. Sesia, C. van Rensburg, and C. B. Papadias, “3GPP LTE and LTE-advanced,” EURASIP Journal on Wireless Communications and Networking, vol. 2009, 2009.

[11] Cisco, “Visual networking index,” white paper at Cisco.com, Feb. 2014.

[12] Ericsson Mobility Report. June 2019. https://www.ericsson.com/49d1d9/assets/local/mobility-report/documents/2019/ericsson-mobility-report-june-2019.pdf

[13] M. N. Tehrani, M. Uysal and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," in IEEE Communications Magazine, vol. 52, no. 5, pp. 86-92, May 2014

[14] 5G Americas White Paper, “5G Network Transformation,” 5G Americas, Dec. 2017. Available: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Network_Transformation_Final.pdf

[15] Keysight White Paper, “Five Challenges Facing Service Providers Rolling Out 5G,” Keysight Technologies, Available: https://stg2www.keysight.com/ie/en/assets/7018-07044/white-papers/5992-3269.pdf

[16] Vodafone, “Your 5G Questions: 5G Explained Pocket Guide,” Vodafone Group, Dec. 2018. Available: https://www.vodafone.com/business/media/document/1508881116678/vodafone-5g-explained-pocket-guide.pdf

[17] 5G Americas, “The 5G Evolution: 3GPP Releases 16-17,” 5G Americas, Jan. 2020. Available: https://www.5gamericas.org/wp-content/uploads/2020/01/5G-Evolution-3GPP-R16-R17-FINAL.pdf

[18] 5G Americas, “Wireless Technology Evolution: Transition from 4G to 5G,” 5G Americas, Oct. 2018. Available at: https://www.5gamericas.org/wp-content/uploads/2019/07/3GPP_Rel_14-16_10.22-final_for_upload.pdf

[19] IMT Vision White Paper, “Framework and overall objectives of the future development of IMT for 2020 and beyond,” ITU, Sept. 2015. Available: https://www.itu.int/rec/R-REC-M.2083-0-201509-I/en

[20] Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s). November 2017. https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf.

[21] M. Shafi et al., "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," in IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1201-1221, June 2017

[22] S. DeTomasi, “Navigating the 5G NR Standards,” Microwave Journal eBook, Apr. 2019, pp. 10-13.

[23] Keysight, “The ABCs of 5G New Radio Standards eBook,” Keysight Technologies, Nov. 2018. Available: https://connectlp.keysight.com/AMO_5GNR-ABCs-AppBR

[24] S. DeTomasi, “Why 5G is going to over-the-air testing,” Design World: Test & Measurement Handbook, pp. 8-11, June 2018. Available: https://www.designworldonline.com/june-2018-special-edition-test-measurement-handbook/

[25] Keysight White Paper, “Overcoming 5G NR Device Design Challenges Series – Part 1: 5G New Radio Standard,” Keysight Technologies, 2019. Available: https://www.keysight.com/zz/en/assets/7018-05995/white-papers/5992-2707.pdf

[26] NI White Paper, “5G New Radio: Introduction to the Physical Layer,” National Instruments, 2018. Available: https://download.ni.com/evaluation/rf/5G_New_Radio_WP.pdf

[27] 3GPP TS 38.101-1 V15.0.0 (2017-12) Technical Specification Radio Access Network; NR; User Equipment (UE) radio transmission and reception.
[28] 3GPP TS 38.211 V15.0.0 (2017-12) Technical Specification Radio Access Network; NR; Physical channels and modulation.
[29] 3GPP TS 38.213 V15.0.0 (2017-12) Technical Specification Radio Access Network; NR; Physical layer procedures for control.
[30] 3GPP TS 38.214 V15.0.0 (2017-12) Technical Specification Radio Access Network; NR; Physical layer procedures for data.
[31] NI White Paper, “Engineer’s Guide to 5G Semiconductor Test,” National Instruments, 2019. Available: http://download.ni.com/pub/white-paper/semiconductor/34424_5G_Semiconductor_Test_WP_en_Ltr_WR.pdf
[32] NGMN Alliance, “NGMN 5G White Paper,” Next Generation Mobile Networks, Ltd., 2015.
[33] GSMA, “Understanding 5G: Perspectives on future technological advances in mobile,” GSMA, 2014.
[34] Keysight White Paper, “OTA Test for Millimeter-Wave 5G NR Devices and Systems,” Keysight Technologies, Nov. 2017. Available: https://www.keysight.com/upload/cmc_upload/All/5992-2600EN_10-5-17_CS.pdf
[35] V. Jungnickel et al., "The role of small cells, coordinated multipoint, and massive MIMO in 5G," in IEEE Communications Magazine, vol. 52, no. 5, pp. 44-51, May 2014
[36] “IMT-2020 5G Wireless Technology Architecture,” IMT-2020 (5G) Promotion Group, May 2015, www.scribd.com/doc/294556768/WHITE-PAPER-ON-5G-WIRELESS-TECHNOLOGY-ARCHITECTURE-pdf
[37] Z. Jiang et al., “Progress and Challenges of Test Technologies for 5G,” Microwave Journal eBook, Apr. 2019, pp. 4-9.
[38] Rohde & Schwarz White Paper, “Millimeter-Wave Beamforming: Antenna Array Design Choices & Characterization,” Rohde & Schwarz, Oct. 2016. Available: https://www.rohde-schwarz.com/gr/applications/millimeter-wave-beamforming-antenna-array-design-choices-characterization-white-paper_230854-325249.html
[39] Z. Pi and F. Khan, ‘‘An introduction to millimeter-wave mobile broadband systems,’’ IEEE Commun. Mag., vol. 49, no. 6, pp. 101–107, Jun. 2011.
[40] F. Gutierrez, S. Agarwal, K. Parrish, and T. S. Rappaport, ‘‘On-chip integrated antenna structures in CMOS for 60 GHz WPAN systems,’’ IEEE J. Sel. Areas Commun., vol. 27, no. 8, pp. 1367–1378, Oct. 2009.
[41] T. S. Rappaport, E. Ben-Dor, J. N. Murdock, and Y. Qiao, ‘‘38 GHz and 60 GHz Angle-dependent Propagation for Cellular and peer-to-peer wireless communications,’’ in Proc. IEEE Int. Conf. Commun., Jun. 2012, pp. 4568–4573.
[42] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, ‘‘State of the art in 60 GHz integrated circuits & systems for wireless communications,’’ Proc. IEEE, vol. 99, no. 8, pp. 1390–1436, Aug. 2011.
[43] F. Rusek, D. Persson, B. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, ‘‘Scaling up MIMO: Opportunities and challenges with very large arrays,’’ IEEE Signal Process. Mag., vol. 30, no. 1, pp. 40–60, Jan. 2013.
[44] Samsung, SK Telecom, KT Corporation, LG Uplus, NTT Docomo, Inc., “R4-1704770: On band definition for 26.5-29.5 GHz,” 3GPP, 2017.
[45] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta and P. Popovski, "Five disruptive technology directions for 5G," in IEEE Communications Magazine, vol. 52, no. 2, pp. 74-80, February 2014
[46] R. C. Daniels and R. W. Heath, “60 GHz wireless communications: emerging requirements and design recommendations,” IEEE Vehicular Techn. Magazine, vol. 2, no. 3, pp. 41–50, Sep. 2007.
[47] S. J. Vaughan-Nicholos, “Gigabit Wi-Fi is on its way,” IEEE Computer, Nov. 2010.
[48] T. Baykas, C.-S. Sum, Z. Lan, and J. Wang, “IEEE 802.15.3c: The first IEEE wireless standard for data rates over 1 Gb/s,” IEEE Communications Magazine, vol. 49, no. 7, pp. 114–121, Jul. 2011.
[49] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, “State of the art in 60-GHz integrated circuits and systems for wireless communications,” Proceedings of the IEEE, vol. 99, no. 8, pp. 1390–1436, Aug. 2011.
[50] ITU Report, “Technical feasibility of IMT in bands above 6 GHz,” ITU, Jul. 2015. Available: https://www.itu.int/pub/R-REP-M.2376-2015
[51] C. Balanis, in Antenna Theory, John Wiley and Sons, 1997, pp. 86-88.
[52] ITU Recommendation, “P.525 : Calculation of free-space attenuation,” ITU, Aug. 2019. Available: https://www.itu.int/rec/R-REC-P.525/en
[53] 3GPP, “TSG RAN TR 38.803 v14.1.0 “Study on new radio access technology: Radio Frequency (RF) and co-existence aspects”,” 2017.
[54] ITU Recommendation, “P.837 : Characteristics of precipitation for propagation modelling,” ITU, June 2017. Available: https://www.itu.int/rec/R-REC-P.837/en

[55] ITU Recommendation, “P.838 : Specific attenuation model for rain for use in prediction methods,” ITU, Mar. 2005. Available: https://www.itu.int/rec/R-REC-P.838/en
[56] ITU Recommendation, “P.833 : Attenuation in vegetation,” ITU, Sept. 2016. Available: https://www.itu.int/rec/R-REC-P.833/en
[57] 3GPP, “TR 38.901 v14.1.1 TSG RAN “Study on channel model for frequencies from 0.5 to 100GHz”,” 2017.
[58] Keysight White Paper, “Overcoming 5G NR Device Design Challenges Series – Part 3: MIMO and Beamforming,” Keysight Technologies, 2019. Available: https://www.keysight.com/zz/en/assets/7018-06216/white-papers/5992-3070.pdf
[59] M. Naseef, G. Lloyd, and M. Reil, “Characterizing Active Phased Array Antennas,” Rohde & Schwarz GmbH & Co. KG, München, 2016.
[60] Rohde & Schwarz e-Guide, “ Massive MIMO: 8 things you need to know,” Rohde & Schwarz, Nov. 2018. Available: https://www.mobilewirelesstesting.com/wp-content/uploads/2018/11/OTA-Top-8-Things.pdf
[61] J. Butler and R. Lowe, “Beamforming matrix simplifies design of electronically scanned antennas,” 1961.
[62] C. Powell, “Technical Analysis: Beamforming vs. MIMO Antennas,” 2014.
[63] R. Wonil et al., “Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results,” IEEE 52, 2014.
[64] A. Alkhateeb, J. Mo, N. González-Prelcic, and Heath, Robert W. Jr., “MIMO Precoding and Combining Solutions for Millimeter-Wave Systems,” IEEE, 2014.
[65] C. Neil et al., “On the performance of spatially correlated large antenna arrays for millimeter-wave frequencies,” IEEE Trans. Antennas Propag., vol. 24, no. 2, pp. 106–112, Apr. 2017.
[66] J. Zhang, X. Ge, Q. Li, M. Guizani, and Y. Zhang, “5G millimeter-wave antenna array: Design and challenges,” IEEE Wireless Commun.,vol. 24, no. 2, pp. 106–112, Apr. 2017.
[67] R. S. Thoma, M. Landmann, G. Sommerkorn, and A. Richter, “Multidimensional high-resolution channel sounding in mobile radio,” in Proc. 21st IEEE Instrum. Meas. Technol. Conf., vol. 1. May 2004, pp. 257–262.
[68] M. I. Skolnik, Introduction To Radar Systems: Mcgraw Hill Book Co, 1961, pp. 280-286.
[69] V. Rabinovich and N. Alexandrov, Antenna Arrays and Automotive Applications: Springer, 2013, pp. 24-52.
[70] K. S. Das and A. Das, Antenna and Wave Propagation: Tata McGraw Hill Education Private Limited, 2013, pp. 153-163.
[71] I. V. Minin and Minin Oleg V., Basic Principles of Fresnel Antenna Arrays: Springer Science & Business Media, 2008, p. 12.
[72] S. I. Nikolov and H. Jensen, “Manipulation of Grating Lobes by Changing Element Shape,” 34, 2011.
[73] G. Foschini and M. J. Gans, “On limits of wireless communications in a fading environment when using multiple antennas,” Wireless Personal Communications, vol. 6, pp. 311–335, Mar. 1998.
[74] E. Telatar, “Capacity of multi-antenna Gaussian channels,” European Trans. Telecommun., vol. 6, pp. 585–95, Nov-Dec. 1999.
[75] T. L. Marzetta, “The case for MANY (greater than 16) antennas as the base station,” in Proc., Information Theory and its Applications (ITA), San Diego, CA, Jan. 2007.
[76] T. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” IEEE Trans. on Wireless Communications, vol. 9, no. 11, pp. 3590–3600, Sep. 2010.
[77] J. Hoydis, S. ten Brink, and M. Debbah, “Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?” IEEE Journal on Sel. Areas in Communications, vol. 31, no. 2, pp. 160–171, Feb. 2013.
[78] E. G. Larsson, O. Edfors, F. Tufvesson and T. L. Marzetta, "Massive MIMO for next generation wireless systems," in IEEE Communications Magazine, vol. 52, no. 2, pp. 186-195, February 2014
[79] J. Nam et al., “Joint Spatial Division and Multiplexing: Realizing Massive MIMO Gains with Limited Channel State Information,” 46th Annual Conf. Information Sciences and Systems, 2012.
[80] ITU Report, “Future technology trends of terrestrial IMT systems,” ITU, Nov. 2014. Available: https://www.itu.int/pub/R-REP-M.2320-2014
[81] B. Hassibi and B. Hochwald, “How much training is needed in multiple-antenna wireless links?” IEEE Trans. on Info. Theory, vol. 49, no. 4, pp. 951–963, Apr. 2003.
[82] N. Jindal and A. Lozano, “A unified treatment of optimum pilot overhead in multipath fading channels,” IEEE Trans. on Communications, vol. 58, no. 10, pp. 2939–2948, Oct. 2010.
[83] E. Bjornson, L. Van der Perre, S. Buzzi and E. G. Larsson, "Massive MIMO in Sub-6 GHz and mmWave: Physical, Practical, and Use-Case Differences," in IEEE Wireless Communications, vol. 26, no. 2, pp. 100-108, April 2019

[84] C. Gustafson and F. Tufvesson, “Characterization of 60 GHz shadowing by human bodies and simple phantoms," 6th European Conference on Antennas and Propagation (EUCAP), Prague, 2012, pp. 473-477.

[85] T. S. Rappaport, G. R. MacCartney, M. K. Samimi and S. Sun, “Wideband Millimeter-Wave Propagation Measurements and Channel Models for Future Wireless Communication System Design," IEEE Transactions on Communications, vol. 63, no. 9, pp. 3029-3056, Sept. 2015.

[86] H. Prabhu, J. N. Rodrigues, L. Liu and O. Edfors, “3.6 A 60pJ/b 300Mb/s 128ÃŮ8 Massive MIMO precoder-detector in 28nm FD-SOI," IEEE International Solid-State Circuits Conference (ISSCC), vol. 60, pp. 60-61, San Francisco, CA, Feb. 2017.

[87] S. Brebels, A. A. Enayati, C. Soens, W. De Raedt, L. Van der Perre and G. A. E. Vandenbosch, “Technologies for integrated mm-Wave antenna,” The 8th European Conference on Antennas and Propagation (EuCAP 2014) pp. 727-731, The Hague, Apr. 2014.

[88] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, “An introduction to the multi-user MIMO downlink, ” in IEEE Communications Magazine, vol. 42, no. 10, pp. 60–-67, Oct. 2004.

[89] A. Sadri, "Evolution of mmWave Technology from WiGig to Cellular and Backhaul Systems," IEEE The Brooklyn 5G Summit, Apr. 2014.

[90] A. Maltsev, A. Sadri, A. Pudeyev, R. Nicholls, R. Arefi, A. Davydov, I. Bolotin, G. Morozov, K. Sakaguchi and T. Haustein, "MmWave Smallcells is a Key Technology for Future 5G Wireless Communication Systems," European Conference on Networks and Communications (EuCNC’2014), Bologna, Italy, Jun. 2014

[91] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, “Femtocell networks: a survey,” IEEE Communications Magazine, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[92] M. Dohler, R. W. Heath, A. Lozano, C. B. Papadias, and R. A. Valenzuela, “Is the PHY layer dead?” IEEE Communications Magazine, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[93] “Backhaul Technologies for Small Cells Use cases, requirements and recommendations,” Small cell forum, draft white paper, Oct. 2012.

[94] B. Kaufman and B. Aazhang, “Cellular networks with an overlaid device to device network,” in Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Oct. 2008.

[95] C. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, “Resource sharing optimization for device-to-device communication underlaying cellular networks,” IEEE Wireless Communications, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[96] “Special issue on internet of things,” IEEE Wireless Communications Magazine, no. 6, Dec. 2010.

[97] D. Astely et al., “LTE Release 12 and Beyond,” IEEE Commun. Mag., vol. 51, no. 7, 2013, pp. 154–60.

[98] “Network function virtualisation - an introduction, benefit, enablers, challenges and call for action,” white paper, Oct. 2012. [Online]. Available: http://portal.etsi.org/NFV/NFV_White_Paper2.pdf

[99] Network Functions Virtualisation (NFV): Architecture Framework, document GS NFV002 v1.1.1, Group Specification ETSI, Oct. 2013.

[100] “Software-defined networking, the new norm for networks,” Apr. 2012. [Online]. Available: http://www.opennetworking.org

[101] S. Sezer, S. Scott-Hayward, P. Chouhan, B. Fraser, D. Lake, J. Foinnegan, N. Viljoen, M. Miller, and N. Rao, “Are we ready for SDN? implementation challenges for software-defined networks,” IEEE Communications Magazine, vol. 51, no. 7, pp. 36–43, Jul. 2013.

[102] “Study on architecture for next generation system, release 14,” 3GPP, Tech. Rep. TR23.799 v1.0.1, Sep. 2016. [Online]. Available: https://www.3GPP.org

[103] 3GPP TS 38.340, Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2

[104] ITU Technical Report, “GSTR-TN5G - Transport network support of IMT-2020/5G,” ITU, Oct. 2018. Available: https://www.itu.int/pub/T-TUT-HOME-2018-2

[105] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," in IEEE Communications Magazine, vol. 55, no. 5, pp. 94-100, May 2017

[106] ITU Recommendation, “Y.3100 : Terms and definitions for IMT-2020 network,” ITU, Sept. 2017. Available: https://www.itu.int/rec/T-REC-Y.3100-201709-I/en

[107] “Study on new radio access technology: Radio access architecture and interfaces,” 3GPP, Tech. Rep. TR 38.801, Mar. 2016. [Online]. Available: https://www.3GPP.org

[108] J. Wu, Z. Zhang, Y. Hong and Y. Wen, “Cloud radio access network (C-RAN): A Primer”, IEEE Network, Vol. 29, No. 1, pp. 35-41, Jan. 2015.

[109] China Mobile Research Institute, C-RAN The Road Towards Green RAN White Paper, Version 2.5, Oct. 2011.

[110] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," in IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2429-2453, Third Quarter 2018

[111] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, “Mobile edge computing—A key technology towards 5G,” ETSI White Paper, vol. 11, 2015

[112] C.-Y. Chang, K. Alexandris, N. Nikaein, K. Katsalis, and T. Spyropoulos, “MEC architectural implications for LTE/LTE-A networks,” in Proc. ACM Workshop on Mobility in the Evolving Internet Architecture (MobiArch), New York, NY, Oct. 2016, pp. 13–18.

[113] S. Shenker, “Fundamental Design Issues for the Future Internet”, IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, pp. 1176-1188, Sep. 1995.

[114] NGMN Alliance, Description of Network Slicing Concept, NGMN 5G P1 Requirements & Architecture, Work Stream End-to-End Architecture, Version 1.0, Jan. 2016.

[115] Keysight White Paper, “Overcoming 5G NR Device Design Challenges Series – Part 2: Millimeter-Wave Spectrum,” Keysight Technologies, 2019. Available: https://www.keysight.com/zz/en/assets/7018-06172/white-papers/5992-2997.pdf

[116] J. Gozalvez, “5G Tests and Demonstrations [Mobile Radio],” IEEE Vehicular Technology Magazine, Vol. 10, No. 2, June 2015, pp. 16-25.

[117] J. Medbo, K. Börner, K. Haneda, V. Hovinen, T. Imai, J. Järvelainen, T. Jämsä, A. Karttunen, K. Kusume, J. Kyröläinen, P. Kyösti, J. Meinilä, V. Nurmela, L. Raschkowski, A. Roivainen and J. Ylitalo, “Channel Modeling for the Fifth Generation Mobile Communications,” 8th European Conference on Antennas and Propagation, April 2014, pp. 219-223.

[118] Aalto University, AT&T, BUPT, CMCC, Ericsson, Huawei, Intel, KT Corporation, Nokia, NTT DOCOMO, New York University, Qualcomm, Samsung, University of Bristol, University of Southern California, “5G Channel Model for Bands Up to 100 GHz,” www.5gworkshops.com/5GCM.html.

[119] X. H. You, Z. W. Pan, X. Q. Gao, S. M. Cao and H. Q. Wu, “The 5G Mobile Communication: the Development Trends and its Emerging Key Techniques,” SCIENTIA SINICA Informationis, Vol. 44, No. 5, January 2014, pp. 551-563.

[120] P. Zhang, Y. Z. Tao and Z. Zhang, “Survey of Several Key Technologies for 5G,” Journal on Communications, Vol. 37, No. 7, July 2016, pp. 15-29.

[121] N. Liu and H. W. Yuan, “Research Status and Development Trends of Large Scale Antenna Systems in 5G Wireless Communications,” Electronic Science and Technology, Vol. 28, No. 4, April 2015, pp. 182-185.

[122] F. Y. Yang, T. Yang and W. L. Xie, “Study and Application on Test Methodology of Active Antenna System,” Telecommunication Science, Vol. 30, No. 2, February 2014, pp. 105-111.

[123] E. Hossain and M. Hasan, “5G Cellular: Key Enabling Technologies and Research Challenges,” IEEE Instrumentation & Measurement Magazine, Vol. 18, No. 3, June 2015, pp. 11-21.

[124] Keysight White Paper, “Overcoming 5G NR Device Design Challenges Series – Part 4: Over-the-Air Test,” Keysight Technologies, 2019. Available: https://www.keysight.com/zz/en/assets/7018-06219/white-papers/5992-3082.pdf

[125] W. Hong, et al., “Design and analysis of a low-profile 28 GHz beam steering antenna solution for future 5G cellular applications,” Microwave Symposium (IMS), 2014 IEEE MTT-S International, 2014.

[126] V. Rodrigues, “Basic Rules for Anechoic Chamber Design, Part Two: Compact Ranges and Near Field Measurements,” Microwave Journal, pp. 1-7, 13 July 2016.

[127] Keysight Technologies, “R5-1708557 “Applicability of CATR to Tx testing”,” 3GPP TSG-RAN WG4 Meeting #84, Berlin, 2017.

[128] J. Fordham and F. D’Agostino, “Sphiral near-field scanning for automotive antenna measurements,” in 10th European Conference on Antennas and Propagation (EuCAP), Davos, Switzerland, 2016.

[129] A. Yaghjian, “An Overview of Near-Field Antenna Measurements,” IEEE Trans. Antennas Propagation, vol. 34, pp 30-45, Jan. 1986.

[130] 5GACIA White Paper, “Selected Testing and Validation Considerations for Industrial Communication with 5G Technologies,” 5GACIA, Nov. 2019. Available: https://www.5g-acia.org/publications/selected-testing-and-validation-considerations-for-industrial-communication-with-5g-technologies/