Research on detection and prevention of rogue base stations in the 5G network.

Capstone Project

Presented by

POOJITHA PATTEM

University of Alberta
Master of Science in Internetworking
Edmonton, Canada

Supervisor
Sandeep Kaur

Acknowledgment

It gives me immense pleasure to express sincere gratitude to my mentor, **Ms. Sandeep Kaur**, for giving me constant support and invaluable guidance, comments, and suggestions throughout the project.

Also, I would like to sincerely thank **Prof. Shahnawaz Mir** and **Dr. Mike MacGregor** for providing me with this opportunity.

Finally, I would like to thank the Almighty for giving me strength and my family for their encouragement throughout my project which kept me motivated and helped me in the completion of this project.

ABSTRACT

Advance in a mobile communication network from 2G to 4G has brought unprecedented traffic growth, and moving from 4G to 5G will bring in more traffic with the expectation of using a 5G network in a large variety of industries. 5G addresses some of the primary objectives and demands, such as increased capacity, low latency, improved data rate, and better quality of services to better interconnect people and machines over the network. 5G-enabled Internet of Things (IoT) communication environment can be utilized in a wide range of applications, such as remote surgery, self-driving car, flying IoT drones, security, surveillance, and many more. This paper presents the result of a detailed survey on the fifth-generation (5G) cellular network architecture, massive multiple-input multiple-output technologies, D2D communication, network slicing, and cloud technologies on 5G radio access networks and software-defined networks. Also included in the paper are some of the vulnerabilities like impersonation, DDOS, malware, man-in-the-middle, and many more, causing security and privacy issues.

TABLE OF CONTENTS

LIST OF FIGURES

1 Chapter 1: INTRODUCTION

1.1 EVOLUTION OF WIRELESS TECHNOLOGIES

Cellular wireless networks have progressed by leaps and bounds since the development of the first 1G system in 1981, with a new mobile generation evolving roughly every decade. In a matter of three decades, the mobile industry has changed society through 4 or 5 generations of technological revolution and evolution, i.e., 1G, 2G, 3G, and 4G networking technologies (Figure 1). Through 1G, we obtained mass-market telephony. 2G established global interoperability and reliable mobile telephony and made SMS text messaging possible. 3G resulted in high-speed data transfer capability for downloads from the Internet. 4G provided a tangible improvement in data capability and speed and led to the possibility of freely available online platforms and high-speed mobile internet services. 5G technology arguably results in the most powerful cellular wireless networks with exceptional data capabilities, unlimited call volumes, and limitless data broadcast [1].

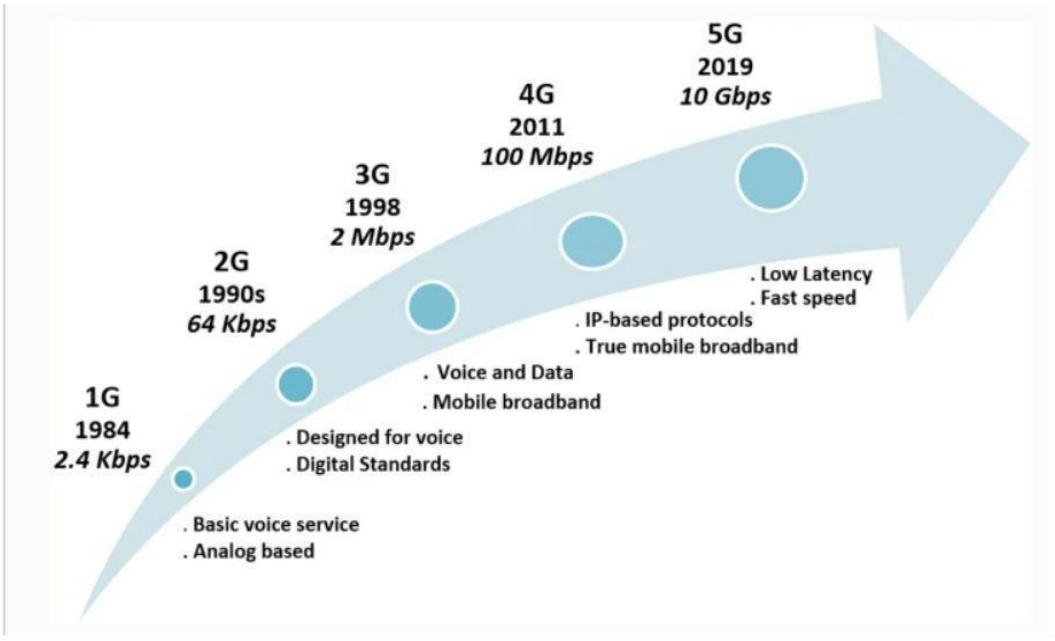

Figure 1: Evolution of mobile communications [1]

There has been exponential growth in the wireless industry in the past few years. Widespread wireless technologies, a rising variety of user-friendly and multimedia-enabled terminals, and

the broader accessibility of open-source tools for generating content have encouraged user-centric networks, leading to efficient network design requirements. As the mobile industry moved from fixed to mobile cellular telephony, Network Planning and Optimization related services have come into sharp focus. The progress of wireless access technology has already reached its fourth generation. Wireless access technology has resulted in divergent evolutionary paths, albeit with the common goal related to performance and efficiency. The First generation gave primary mobile voice, while the Second generation gave capacity and coverage. The focus of the Third generation was on higher data rate, multimedia support, and spread spectrum, followed by the Fourth generation providing access to a range of telecommunication services, including advanced mobile services, plus support for low to high mobility applications [2].

1.2 FIRST GENERATION (1G)

Starting in the late 1970s, when the cellular era began, mobile communication has evolved leaps and bounds every ten years or so in terms of technology and usage. Japan led the development in cellular technology, with a milestone being the initiation of the first cellular networks in Tokyo. Nordic Mobile Telephony (NMT) started cellular operations in Europe a few years following this. Additionally, systems like the AMPS (Advanced Mobile Phone Service) started in the USA, while TACS (Total Access Communication System) started in the UK. These together were a fraction of what was called ‘First Generation Mobile Systems’, which provided speech services based on analog transmission techniques.

The initial systems were incompatible with one another. Each network implemented its own protocols. Services such as roaming within the continent were impossible, and most countries had just one operator. They also could not establish much penetration; for example, penetration in Sweden was just 7%, whereas countries like Portugal had only 0.7 %. Handsets were priced over $1000, thus making them inaccessible. Even after putting aside high costs and incompatibility factors, first-generation technology also had inherent limitations in terms of channels, etc.

1G had significant setbacks compared to today's standards. Being on the listener's end over a 1G network was difficult due to low sound quality. The coverage was patchy, with heavy

amounts of static noise and background crackling. There was no roaming support either. Due to the lack of encryption, the concept of security over a 1G channel did not exist, and anyone with a radio scanner could eavesdrop on the call.

1.2.1 Nordic Mobile Telephone (NMT)

The development of the NMT mobile phone system in 1981 was a reaction to the rising congestion and mounting demand for more ARP (auto radio Puhelin, or car radiophone) mobile phones. Theoretical principles for NMT were ready by 1973, and specifications for base stations were available in 1977. There are two variants based on analog technology (first generation or 1G): NMT 450 and NMT 900. The numbers mark the frequency bands used. NMT 900 was initiated in 1986 as it carries more channels than the prior NMT 450 network. Nordic countries, Baltic countries, Russia, the Middle East, and Asia used the NMT network. Design of NMT with automatic switching built into the standard. In addition, the NMT standard specified billing and roaming. The NMT specifications were free and open; hence many companies produced NMT hardware and kept prices cheap.

One disadvantage of the first NMT specification is that there is no encrypted traffic. Hence, anyone wanting to eavesdrop would simply have to get a scanner and tune it to the right frequency. Some scanners had the NMT bands 'deleted' to prevent this accessibility; this is not very efficient as it is simple to get a scanner that does not have these limitations; it is also convenient to re-program a scanner to use 'deleted' bands again. Future versions of the NMT specifications defined optional analog encryption based on a two-band audio frequency inversion method. The base station (BS) and the mobile station (MS) both must support encryption and agree upon using it before starting a phone call. However, if two users had mobile stations providing encryption, they could use it during a conversation, even if the base stations did not aid it. Here, they encrypted audio between the two mobile stations. While the encryption method was not as robust as encryption in the latest digital phones, it did stop casual eavesdroppers with scanners.

Cell sizes in an NMT network span from 2 km to 30 km. With smaller ranges, the network can provide for a larger number of concurrent callers; for example, the range can be small for good service in a city. NMT used full-duplex transmission, providing for simultaneous transmission and reception of audio. Transmission power for car phone versions was 6 watts and for handsets

up to 1 watt. Control Signaling between BS and MS did not have a separate band but happened using the same RF channel for audio and the same 1200 bps (bits per second) FFSK modem. This led to frequent short noise bursts that were a distinctive sign of NMT sound.

1.2.2 Advanced Mobile Phone System (AMPS)

The initial cellular licenses in the US were issued in 1981, and services started in 1983 in Chicago and The Baltimore–Washington region with the AMPS. The AMPS was built on the FDMA technology, allowing many cell or cell sector users. Here, cell sizes were not rigid, and they employed an eight-mile radius in urban areas and a twenty-five-mile radius in rural areas. But as they added new users, new cells were initiated. As new cells were added, they re-evaluated the frequency plan to circumvent interference-related issues. This plan not only had capacity-related issues, but the security system was also bad. If one can get a person's serial code, they can make illegal calls from their ID. Despite efforts made to address these issues, especially those related to capacity, the results were poor, and the industry started to look in other directions, such as next-generation digital systems. The TACS was the same as the AMPS and operated in the 900 MHz frequency range [3].

1.2.3 Limitations

Even though the download speeds over 1G were also slow, with an upper limit that only reached 2.4kbps. The 1G mobile technology has many limitations such as poor sound quality, small coverage, inefficient use of the spectrum due to full analog mode of communication, low capacity, incompatibility between different 1G systems due to various frequencies of the systems, and having no roaming support among other operators [4].

1.3 SECOND GENERATION (2G)

The GSM system was implemented as a second-generation (2G) cellular phone technology. One of the goals was to provide a system that would build greater capacity than the prior first-generation analog systems. GSM did this with the help of digital TDMA (time division multiple access) models. With this approach, we can add more users within the available bandwidth. Plus, ciphering was used in digitally encoded speech to provide for privacy.

1.3.1 GSM Network Architecture

The GSM network architecture described in the GSM specifications is categorized into four main areas:

- Network and Switching Subsystem (NSS)
- Base-Station Subsystem (BSS)
- Mobile station (MS)
- Operation and Support Subsystem (OSS)

As the specifications and standards define the GSM network, it helps the system to operate flawlessly and in tandem, irrespective of the supplier of the different elements. A basic diagram of GSM Network Architecture (Figure 2) and the simplified architecture (Figure 3) are below:

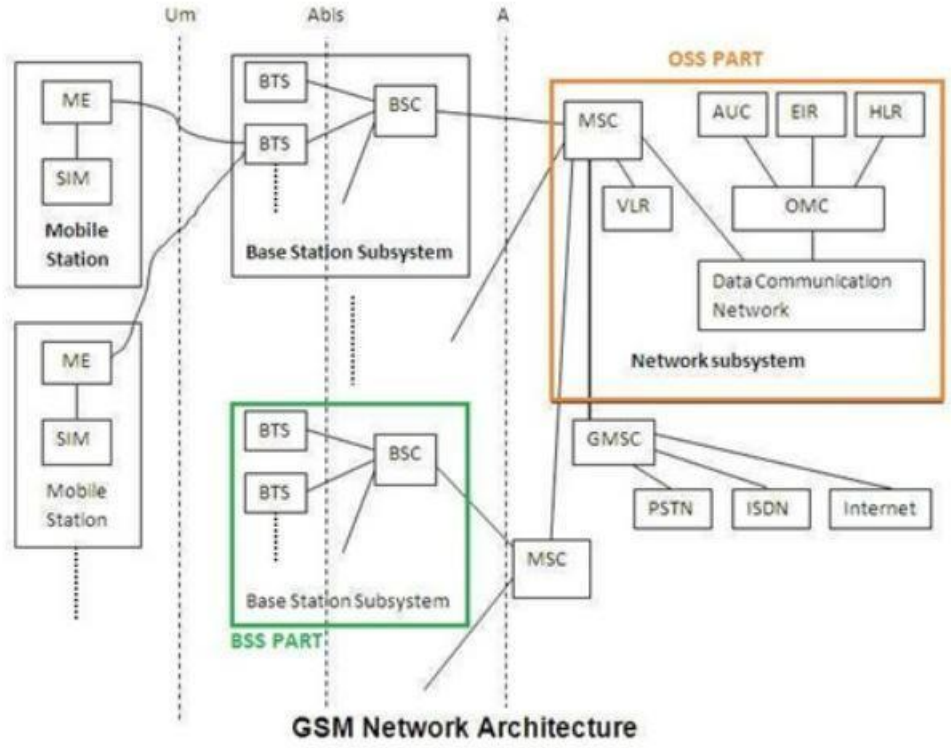

Figure 2: GSM Network Architecture [5]

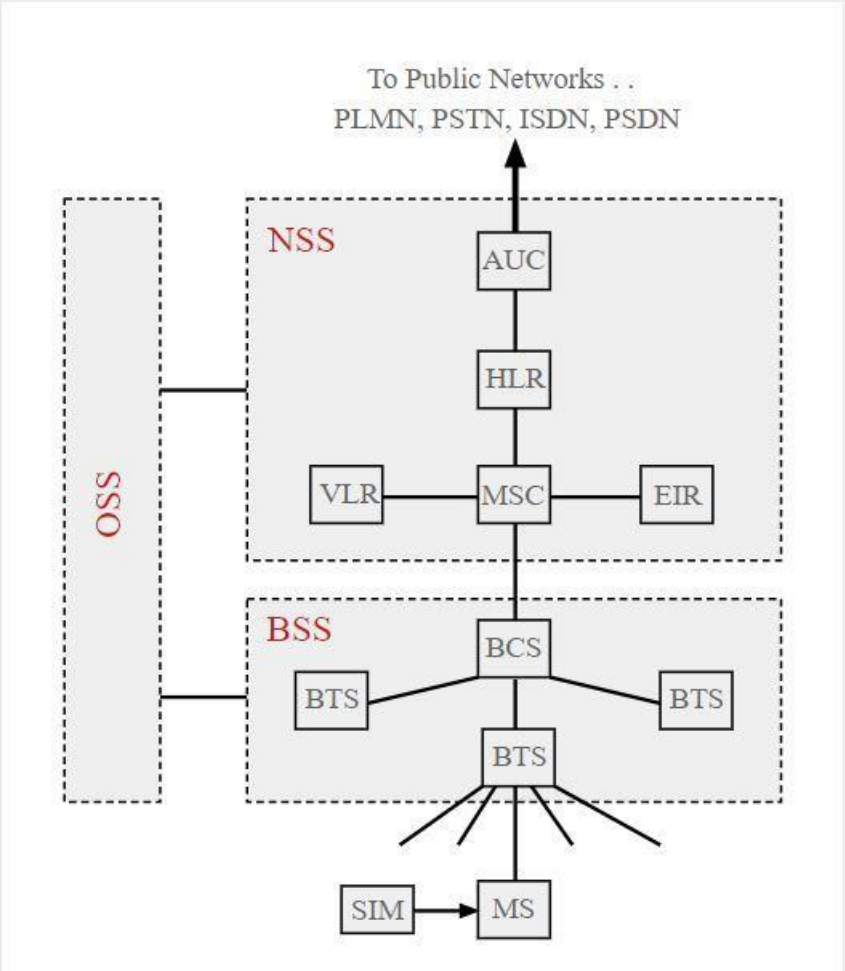

Figure 3: Simplified GSM Network Architecture [6]

1.3.1.1 Network Switching Subsystem (NSS)

The GSM system architecture holds an array of different elements and is often called the core network. It is a data network with various units that provide the main control and interfacing for the entire mobile network. The major components within the core network are:

- **Mobile Services Switching Centre (MSC):** The main element in the GSM core network architecture is the Mobile switching Services Centre (MSC) which does normal switching within a PSTN or ISDN, but it can provide additional functionality required to support mobile users. These additional functions include registration, call routing to a mobile subscriber, authentication, call location, and inter-MSC handovers. MSC also provides interfaces to route mobile calls to a landline over PSTN as well as route calls to mobiles on different networks.

- **Home Location Register (HLR):** The database stores the administrative information regarding every subscriber, including their previous location. So, the GSM network can route calls to the relevant base station for the MS. When a user turns on their phone, a suitable BTS is determined, and the network routes their incoming calls accordingly. As long as the phone is active (switched on), it's constant re-registers with the latest location to help the network location the mobile. A network has only one HLR for operational reasons, even though it spans different sub-centers.

- **Visitor Location Register (VLR):** The VLR has explicit data from the HLR that warrants the providing of opted services for the individual subscriber. It can be realized as a separate discrete, but the implemented VLR is an essential part of the MSC than a separate entity. In this way, access is made quicker and more convenient.

- **Equipment Identity Register (EIR):** The EIR is the unit that determines whether mobile equipment should be allowed onto the network. Each cellular equipment has an identifier number called the International Mobile Equipment Identity. This number is placed within the system and checked by the network during registration. Based on the information held in the EIR, the mobile may be allotted one of three states - allowed onto the network, barred access, or monitored in case it is suspicious.

- **Authentication Centre (AuC):** The AuC is a blanketed database that stores the secret key that is also stored on the user's SIM card. Auc authenticates the sim and uses the key to encrypt information on the radio channel.

- **Gateway Mobile Switching Centre (GMSC):** The term GMSC is misleading for the reason that the gateway operation is not related to or requires any linking to an MSC. The GMSC is responsible for getting from the HLR, the MSRN (Mobile Station Roaming Number), based on Mobile Station ISDN Number (MSISDN), the "directory number" of an MS. Initially, routing is done to any ME terminating call to GMSC without any knowledge of the MS's location. It is the GMSC that routes the call to the correct visited MSC.

- **SMS Gateway (SMS-G):** The SMS-G (Short Message Services Gateway) collectively describes the two SMS Gateways defined in the GSM standards. The two gateways are SMS-GMSC (Short Message Service Gateway Mobile Switching Centre) and SMS-IWMSC (Short Message Service Inter-Working Mobile Switching Centre). The role of SMS-GMSC is comparable to that of the GMSC, while the SMS-IWMSC provides a fixed access point to the SMS Centre. These gateways handle messages sent in different ways. SMS-GMSC is for sending short messages to an ME, while SMS-IWMSC is for short messages originating from a mobile on that network.

These were the principal entities used within the GSM network. These co-located, but mostly the overall core network is distributed around the network location region. This gave some resilience in case of failure.

1.3.1.2 Base Station Subsystem (BSS)

The BSS part of the 2G GSM network architecture is principally associated with communicating with the mobiles on the network. It consists of two elements:

- **Base Transceiver Station (BTS):** The BTS present in a GSM network includes the radio transmitter-receivers and their related antennas. The characteristic element for each cell is the BTS. It communicates with the mobiles. This interface is known as the Um interface with its relevant protocols.

- **Base Station Controller (BSC):** The BSC is the subsequent stage back to the GSM network. Located with a BTS in its respective group, it also controls that group of BTS's. It handles the radio resources and regulates items like the handover within the group of BTS's. It distributes channels and such. It relays information with the BTS's over the Abis interface.

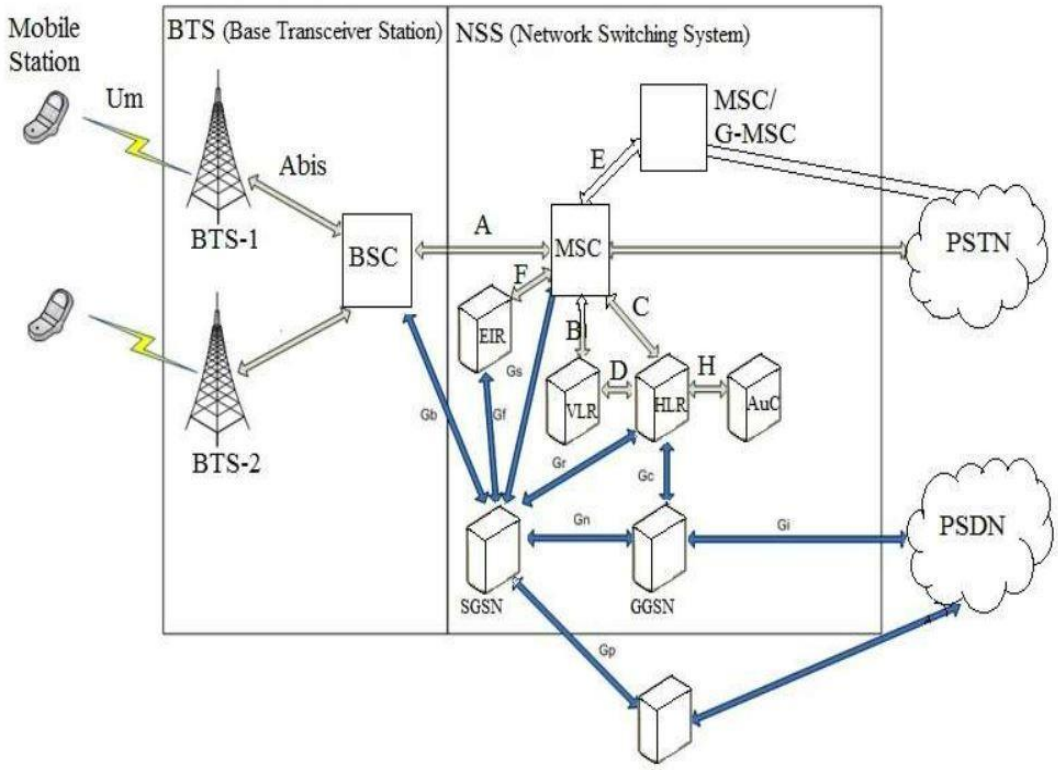

Figure 4: GSM Architecture on BSS Side [6]

The BSS element of the GSM network used radio access technology to empower multiple users to access the system simultaneously. Each channel could bear eight users maximum, and by enabling a base station to have a large number of channels, housing a large number of subscribers could be housed by each base station. Carefully considered locating base stations to enable complete coverage of an area. The area serviced by a base station is often called a cell.

There is always an overlap of adjacent cells to cover all areas properly. Channels used in a cell are not used in the next one to prevent this overlap from causing interference. This way, call quality is controlled while still maintaining sufficient frequency re-use.

It is essential to have the various BTSs linked with the BSS, and the BSSs linked back to the core network. A range of technologies was used to implement this. As the data rates employed within the GSM network were relatively low, E1 or T1 lines were frequently used, remarkably for linking the BSS back to the core network.

As more data was essential with the increasing perusal of the GSM network, and with other cellular technologies like 3G became more ubiquitous, multiple links used carrier-grade Ethernet.

Repeatedly remote BTSs linked with small microwave links decrease the need for installing specific lines in case of lack of availability. Base station placement often required good coverage rather than in locations where lines could be conveniently installed; the microwave link option opened up an attractive alternative mechanism of providing a link to the network.

1.3.1.3 Mobile Station (MS)

Mobile equipment (ME), mobile stations (MS), or more commonly known, mobile or cell phones are the component of a GSM mobile communications network that the user uses and operates. Recently, reducing their volume and size has been done substantially, whereas their functionality has greatly improved. Another advantage is that the time between recharges has significantly risen.

A cell phone has a number of components, although two main elements are the SIM and the main hardware. The hardware hosts the mobile phone's main elements, including the case, display, battery, and the electronics used to initiate the signal, process the data received, and be transmitted.

The ME or mobile station contains a number called the International Mobile Equipment Identity (IMEI) also. This is integrated into the phone during the manufacturing process and "cannot" be modified. The SIM or Subscriber Identity Module holds the data that provides the identity of the subscriber to the network. It holds a variety of information, including a number called the IMSI. As this is present in the SIM, it means that by changing the SIM card from one mobile to another, the user could easily switch mobiles. This ease of switching mobiles whilst keeping the same mobile number translates to this that people would regularly upgrade, hence generating a further revenue stream for network providers and also increasing the overall financial success of GSM.

1.3.1.4 Operation and Support Subsystem (OSS)

The operation or OSS support subsystem is an element in the general GSM mobile communications network architecture that is linked to components of the NSS and the BSC. This is to regulate and monitor the overall GSM network, and it is also employed to regulate the traffic load of the BSS. The point to be noted is that as the quantity of BS increases with the scaling of the user population, few of the maintenance tasks are transferred to the BTS, accumulating savings in the cost of ownership of the system.

The 2G GSM network architecture utilizes a logical method of operation. It is far easier than current mobile phone network architectures that use software-defined entities to warrant a very flexible operation. However, the 2G GSM architecture displays the voice and basic operational functions required and how they fall into place together. Given that the GSM system was all digital, the network was essentially a data network [5].

1.3.1.5 Advantages

- **Extensive Coverage:** The foremost advantage of GSM is its ubiquitous use throughout the world. To quote Gsmworld.com, GSM has a harmonized spectrum, implying that even though different countries may operate on different frequency bands, subscribers can transfer seamlessly among networks and still hold the same number resulting in GSM users essentially having coverage in over 218 countries.
- **Greater Phone Variety**: Another advantage of GSM is that as it is used all around the world, there is a greater range of phones that operate on GSM. Therefore, users have more flexibility while choosing a handset, one that accommodates their specific needs, and they are not limited to using phones only made in their respective countries.
- **No Roaming Charges on International Calls**: Because GSM is the same network worldwide, subscribers are not deducted a roaming fee for international calls. Still, most providers charge a service cut on international calls.

1.3.1.6 Disadvantages

- **Bandwidth Lag:** The greatest disadvantage of GSM is that many users share the same bandwidth. With a critical number of users, the transmission may encounter interference. Therefore, faster technologies, like 3G, have been developed on various

types of networks than GSM, such as CDMA, to circumvent such bandwidth limitations.

- **Causes Electronic Interference**: Another disadvantage of GSM is that it can interfere with certain other medical electronics, such as pacemakers and hearing aids. Such interference is resultant of the fact that GSM uses a pulse-transmission technology. Hence, many locations such as hospitals and airplanes require cell phones to be switched off [7].

1.3.2 2.5G – GPRS (General Packet Radio Service)

GPRS is an extension of the prior existing 2G network upgraded to have the capacity of launching packet-based services while simultaneously enhancing the data rates employed by these networks. The term 2.5G tells of 2G-Systems that have implemented a packet-switched domain along with a circuit-switched domain [8].

GPRS uses packet switching to transmit high-speed signaling and data more effectively than the GSM systems. It optimizes the network and reduces radio resource usage while maintaining a strict demarkation of the radio subsystem and network subsystem, enabling the network subsystem to be employed with other radio access technologies [9].

GPRS gives data rates from 56 Kbps up to 384 Kbps, with database HLR, VLR, EIR, and AuC. It provides facilities such as Multimedia Messaging Service (MMS), Wireless Application Protocol (WAP) access, and for internet communication services such as World-Wide Wireless Web (WWW) access and e-mail access. GPRS data transfer charges per megabyte of traffic transferred, while data communication via traditional circuit switching is charged per minute of connection time, irrespective of whether the user is utilizing the available bandwidth or idle. 2.5G networks might support services such as MMS, WAP, SMS mobile games, internet access, and search directory. Designed GPRS to support bursty and intermittent data transmission [8] [9].

1.3.3 2.75G - EDGE (Enhanced Data rates for GSM Evolution)

3GPP classified EDGE as part of the GSM family and is an upgrade that gives a potential three-fold rise in the function of GSM/GPRS networks. This specification allows for higher data rates up to 236.8 Kbits/s. EDGE technology is preferred over GSM due to its flexibility to carry circuit-switch data and packet-switch data [8].

1.4 THIRD GENERATION (3G)

1.4.1 UMTS WCDMA

The world's authoritative 3G system is the Universal Mobile Telecommunication System (UMTS). UMTS was created from GSM by completely modifying the technology used on the air interface while keeping the core network almost untouched. It was later enhanced for data applications by introducing the 3.5G technologies and high-speed downlink packet access (HSDPA)and high-speed uplink packet access (HSUPA), which are together known as high-speed packet access (HSPA). The UMTS air interface has two slightly variant implementations. Wideband code division multiple access (WCDMA) is initially specified version and the one used in most of the world currently [10]. WCDMA is strictly a Direct Spread Spectrum CDMA system, where user information is multiplied with pseudo-random codes to give way for synchronization, scrambling, and channelization. This system's design allows for it to operate in the 5 MHz bandwidth, which can hold 100 different voice calls concurrently. The peak data rate then ranges from 384 to 2048 kbps. Additionally, WCDMA enables using multi-code to improve a single user's data rate [11].

1.4.2 UMTS Network Architecture

Third-generation mobile networks design is for multimedia communication, thus improving the image and video quality and increasing data rates within public and private networks. WCDMA technology came out as the most widely adopted third-generation air interface in the standardization forums. The 3GPP produced this specification. The UMTS system architecture is in Figure 5. It includes a Radio Access Network (RAN) or UMTS Terrestrial RAN (UTRAN) that controls all radio-related operations; a Core Network (CN), which is responsible for routing data packets and switching calls to external network; and User Equipment (UE) that combines with the UMTS network. Entirely new protocols are designed based upon the new

WCDMA radio technology for the two, i.e., the UE and UTRAN, but bulk of the CN is from the GSM/GPRS networks.

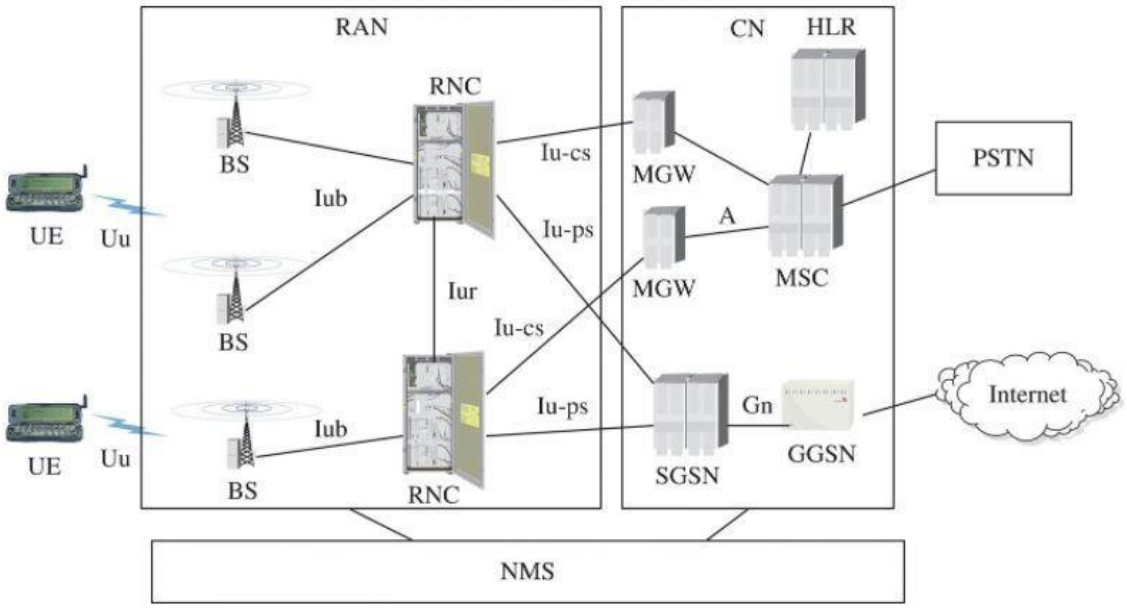

Figure 5: Third Generation Network Architecture [12]

1.4.2.1 User Equipment (UE)

- The UMTS Subscriber Identity Module (USIM) is a smartcard that holds the user identity, authentication and encryption keys, and a fraction of subscription information needed at the terminal. It also performs authentication algorithms.
- The Mobile Equipment (ME) is the radio terminal employed for radio communications over the UE interface.

1.4.2.2 UTRAN

UTRAN includes a set of Radio Network Subsystems (RNS) connected to the CN with the I_{ub} interface. An RNS is a Radio Network Controller (RNC) and one or multiple NodeBs. A NodeB is connects to the RNC via the I_{ub} interface.

- **Radio Access Network (RAN):** The main network elements in this fragment of the network are the radio network controller (RNC) and the base station (BS). The main

functions include telecommunication management and management of the radio resources.

- **Base Station (BS):** The BS in 3G is also called Node B. The BS is a major entity that acts as an interface between the network and the WCDMA air interface. Usually, as in second-generation networks, reception and transmission of the signals from the BS are done through omnidirectional or directional antennas. The main roles of the BS are channel coding, interleaving, spreading, rate adaptation, etc., and also the processing of the air interface.

- **Radio Network Controller (RNC)**: The role of an RNC is to act as an interface between the CN and the BS. RNC accounts for the control of the radio resources. And, counter to the case in GSM, the RNC, in conjunction with the BS, would be able to steer all the radio resource functions without involving the CN. The major roles of the RNC include load and congestion control of the cells, code allocation, admission control, routing of the data between the I_{ub} and I_{ur} interfaces, etc.

1.4.2.3 Core Network (CN)

The CN in 3G networks has two domains: a circuit-switched (CS) domain; and a second, packet-switched (PS) domain. The CS part deals with real-time traffic, and the PS part deals with non-real-time traffic. These two domains are attached to the other networks, e.g., CS to the PSTN and PS to the public IP network. Major elements of the CN include WMSC/VLR, MGW (Media Gateway) on the CS side, HLR, GGSN (Gateway GPRS Support Node), and SGSN (Serving GPRS Support Node) on the PS side. The protocol design of the UE and UTRAN is built on the new WCDMA technology, albeit the CN definitions are from the GSM specifications.

- **WCDMA Mobile Switching Centre (WMSC) and VLR (Visitor Location Register):** The switch and database are accountable for call control activities. WMSC is relied on for the CS transactions, and the VLR function has information on the subscriber visiting the region, including the mobile's location within the region.

- **Gateway Mobile Switching Centre (GMSC):** This is the interface between the external CS networks and the mobile network. This establishes call connections going in and out of the network. It also employs the correct WMSC/VLR for the call path connection.

- **Home Location Register (HLR):** This is the database that holds all the information corresponding to the mobile user and the kind of facilities subscribed to. A new database entry gets initiated when a new user adds to the system, which will be in the database until the user subscribes to the network. Also, it stores the UE location in the system.

- **Serving GPRS Support Node (SGSN):** The SGSN regulates an interface between the PS and the RAN domain of the network. It is mainly accountable for mobility management issues such as the registration and updating of the UE, security issues, and paging-related activities for the PS network.

- **Gateway GPRS Support Node (GGSN):** This functions as an interface between the 3G network and the external PS networks. Functions are the same as the GMSC in the CS domain of CN, but here it is for the PS domain.

1.4.2.4 Network Management System (NMS)

The network management system evolved along with the network technology from 2G to 3G. The NMS in 3G systems will be able to handle managing packet-switched data also, as opposed to voice and circuit-switched data in 2G systems. In 3G, the management systems would be more efficient, i.e., more work would be gleaned from the NMS rather than by visiting the sites. Such systems would improve the system quality and optimize it more efficiently. The management systems are also relied upon to control both the multi-technology, i.e., 2G to 3G, and multi-vendor environments.

- PS networks provide connections for packet data service, e.g., the Internet.
- CS networks provide circuit-switched connections, e.g., ISDN and PSTN [12] [9].

1.4.2.5 Advantages of 3G

- New radio spectrum to relieve overcrowding in already existing systems.
- Asymmetric data rates.

- More bandwidth, security, and reliability.
- Interoperability between service providers.
- Variable and fixed data rates.
- Rich multimedia services.
- Backward compatibility of devices with prior existing networks.
- Perennially online devices. 3G uses IP connectivity; IP is packet-based (not circuit-based).

1.4.2.6 Disadvantages of 3G

- The cost of upgrading cellular infrastructure and base stations to 3G is very high.
- The requirement of different handsets is always the issue of handset availability. 3G handsets will be a complicated product. Roaming and having both data/voice work has not yet been shown. Plus, the higher power requirements (more bits with the same energy/bit) require a larger handset, larger batteries, and shorter talk time.
- Base stations need to be nearer to each other (more cost).
- Tremendous spectrum-license costs and network deployment costs.
- Wireless service providers in Britain and Germany who won spectrum licenses in auctions shelled out astronomical prices for them. As a consequence, they have less money left to build the infrastructure. Hence delaying the deployment of 3G in Germany and Britain [13].

1.4.2.7 Beyond 3G – Mobile WiMAX

While HSPA and HRPD systems deploy after development, IEEE 802 LMSC (LAN/MAN Standard Committee) introduced the IEEE 802.16e standard for mobile broadband wireless access. It employs an enhancement to a prior IEEE 802.16 standard for fixed broadband wireless access. This 802.16e standard employed a variant access technology named OFDMA (orthogonal frequency division multiple access) and advertised better data rates and spectral efficiency than that which HSPA and HRPD gave. The IEEE 802.16 family of standards may be officially called WirelessMAN in IEEE, but it has been called WiMAX (worldwide interoperability for microwave access) by an industry group called the WiMAX Forum. The goal of the WiMAX Forum is to push for an increment and certify the interoperability and compatibility of broadband wireless access products. This WiMAX system upholding mobility

as in IEEE 802.16e standard is mobile WiMAX. Additionally, Mobile WiMAX also employs a simpler network architecture based on IP protocols to take advantage of the radio technology.

The introduction of Mobile WiMAX has led both 3GPP and 3GPP2 to engineer their version of beyond 3G systems, which bases on the OFDMA technology and network architecture similar to that in Mobile WiMAX. The beyond 3G system in 3GPP is named evolved universal terrestrial radio access (evolved UTRA) and is known as LTE (Long-Term Evolution), while 3GPP2's version is called UMB (ultra-mobile Broadband) [14].

1.5 FOURTH GENERATION (4G)

4G (Fourth Generation) cellular wireless systems are the most recent version of mobile technologies. 4G is defined to fulfill the requirements set up by the ITU (International Telecommunication Union) as a part of IMT Advanced. The major drivers for the network architecture evolution in 4G systems are all-IP (Internet protocol) -based, reduced data latencies and signaling load, reduced network cost, interworking mobility among other access networks in 3GPP and non-3GPP, worldwide roaming capability, and always-on user experience with the flexible quality of service (QoS) support. 4G systems include various access technologies:

- LTE and LTE-Advanced (long-term evolution) are part of 3GPP. LTE, as it is now, does match all IMT Advanced features. But, LTE-Advanced is a piece of a later 3GPP release and designed specially to meet 4G requirements.
- WiMAX 802.16m – The IEEE and the WiMAX Forum have identified 802.16m as offering a 4G system.
- UMB (ultra-mobile Broadband) – It is part of 3GPP2. Most network operators and vendors have decided to promote LTE instead.

1.5.1 LTE Network Architecture

The 4G Architecture is a new architecture evolved to give a higher level of performance in line with the needs of LTE, called SAE-System Architecture Evolution. The 4G Architecture has many advantages compared to 2G and 3G architectures like new routing techniques, efficient solutions for sharing dedicated frequency bands, bandwidth capacity, and increased mobility.

These needs achieve through several network elements with different roles. 3GPP identified in Release 8 the LTE system architecture that works as a base architecture network of 4G Network. The LTE architecture accounts for the Evolved Packet Core (EPC) network and the Evolved Universal Terrestrial Radio Access Network (E-UTRAN) [15].

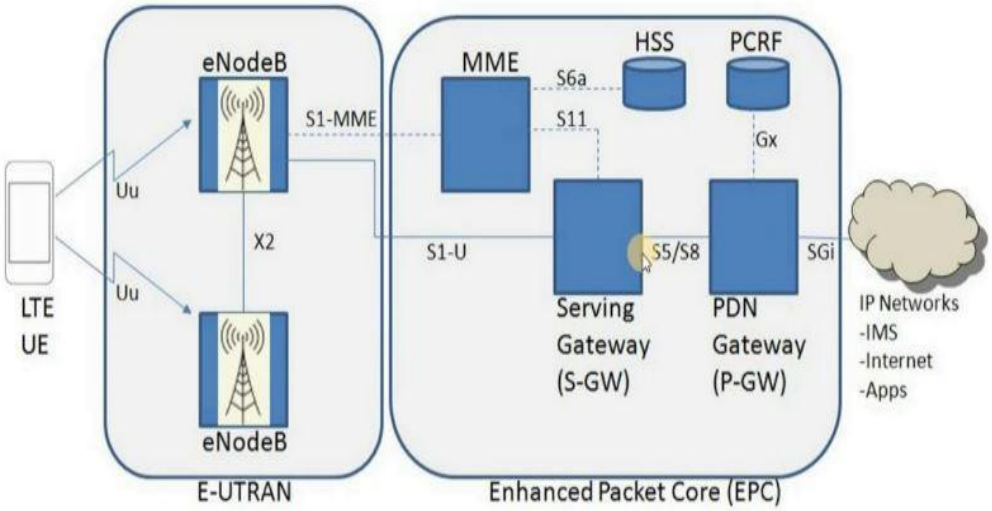

Figure 6: 4G LTE Architecture

The 3GPP has set performance targets for an LTE of peak data rates >100 Mbps in DL, and >50 Mbps in UL with latency less than 5ms on the air interface per link. The spectral efficiency of LTE can exceed the one of UMTS Release 6 by a factor of 3–4 in DL and a factor of 2–3 in UL. The access scheme in LTE is OFDMA in downlink and a SC-FDMA in uplink.

1.5.2 Architecture of EPS

The Evolved Packet System (EPS) is the Evolved UTRAN (E-UTRAN), Evolved Packet Core (EPC), and connectivity to legacy 3GPP access and non-3GPP access systems. The EPS architecture has a reduced number of network elements on the data path compared to GPRS/UMTS, RAN functionality supported in a single node, and the separation of the control and user-plane network elements (MME and Serving Gateway).

The new network elements are as follows:

- **Mobility Management Entity (MME):** It is the control plane (C-plane) functional element in EPC. MME stores and manages UE context, manages mobility and bearers, authenticates the user, generates temporary identities and allocates them to UEs, and acts as a termination point for Non-Access Stratum (NAS) signaling.
- **Serving Gateway (S-GW):** It is the user plane (U-plane) gateway to the E-UTRAN. S-GW provides an anchor point both for intra-3GPP mobility (i.e., inter-3GPP access mobility between LTE and 2G or 3G) and for inter-eNodeB (eNB) handover. It also regulates packet routing, forwarding, and buffering of downlink data for UEs which are in ECM-IDLE state.
- **Packet Data Network Gateway (P-GW):** Which is the U-plane gateway to the PDN (e.g., the operator or the Internet's IP Multimedia Subsystem (IMS)). P-GW is responsible for charging support, policy enforcement, and the user's IP address allocation.
- **E-UTRAN** is the radio access part of the LTE Network.

The legacy network elements interfacing LTE/SAE are as follows:

- **Gateway GPRS Support Node (GGSN):** It is responsible for terminating the Gi interface towards the PDN for legacy 2G/3G access networks. LTE/SAE interfaces this node only as a part of P-GW functionality and from the perspective of inter-system mobility management.
- **Serving GPRS Support Node (SGSN)**: is responsible for transferring packet data between the Core Network and the legacy 2G/3G RAN. LTE/SAE interfaces the SGSN only in the case of inter-system mobility management.
- **Home Subscriber Server (HSS)** is the IMS Core Network entity responsible for managing user profiles, performing user authentication and authorization. The user profiles regulated by HSS consist of subscription and security information and details on the physical location of the user. While IMS is not a mandatory network element, the HSS is a necessary node for the operation of the LTE system.
- **Policy Charging and Rules Function (PCRF)** is accountable for brokering QoS Policy and Charging Policy on a per-flow basis [16].

1.5.2.1 E-UTRAN: Evolution and Architecture

E-UTRAN is a radio access network of 3GPP's Long-Term Evolution (LTE). It is a new air interface system, which provides low latency, high-speed data rate, and optimizes for packet data. It uses SC-FDMA for the uplink and OFDMA radio access for the downlink. In LTE, two duplexing schemes are employed, time division duplexing (TDD) and frequency division duplexing (FDD). With LTE-TDD, a single frequency channel assigns to both the transmitter and the receiver. LTE-FDD requires paired spectrum with sufficient frequency separation to allow simultaneous transmission and reception. The E-UTRAN Network requires a high-speed data rate and reliable transmissions with bandwidth efficiency. For these requirements, we implement Multiple-input multiple-output (MIMO) systems where we can use multiple antennas in both transmitter and receiver, and a single LTE cell can use up to four antennas.

The radio access network E-UTRAN achieve many functionalities, including:

- Radio resource management (RRM)
- Provides initial access to the network, registration, and attach/detach to the network
- Mobility Management Functions
- Security Functions
- Terminal state transition
- Flexibility in spectrum usage
- Selection of an MME at UE attachment when the UE provides no MME information
- Handover Management – Intra-eNode [15]

1.5.2.2 Radio Protocol Architecture Access Stratum

The network protocols divide into the User plane and Control plane. The former is responsible for transporting user traffic, and the latter manages the transport bearer.

- **User Plane Protocols:** Figure 7 shows the protocol stacks for the user plane, where PDCP, RLC, MAC, and PHY sublayers (terminated at the eNB on the network side) perform functions like header compression, ciphering, scheduling.

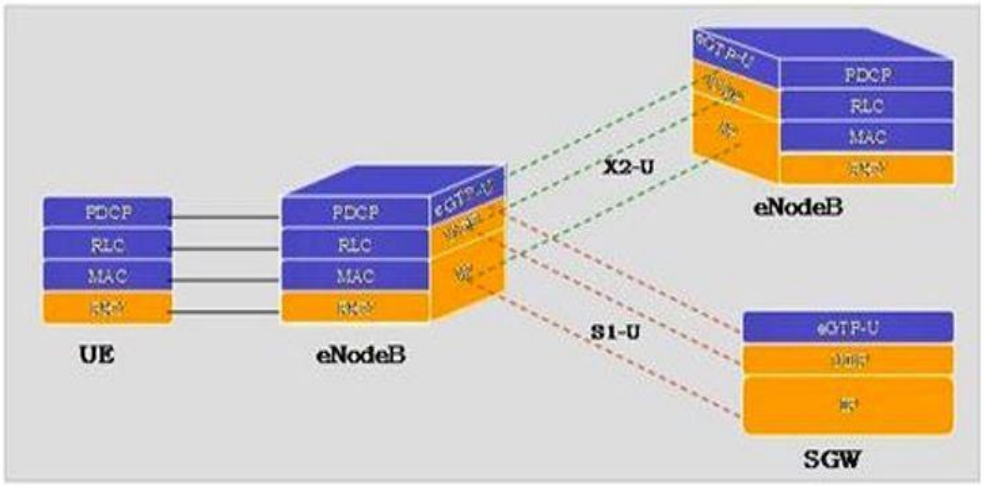

Figure 7: User Plane Protocol Stacks [17]

- Control Plane Protocols: Figure 8 shows the protocol stacks for the control plane, where: PDCP sublayer performs ciphering and integrity protection; RLC, MAC, and PHY sublayers perform the same functions as in the user plane. RRC performs functions like System Information Broadcast, Paging, RRC connection management, RB control, Mobility Control, and UE measurement reporting and control [17].

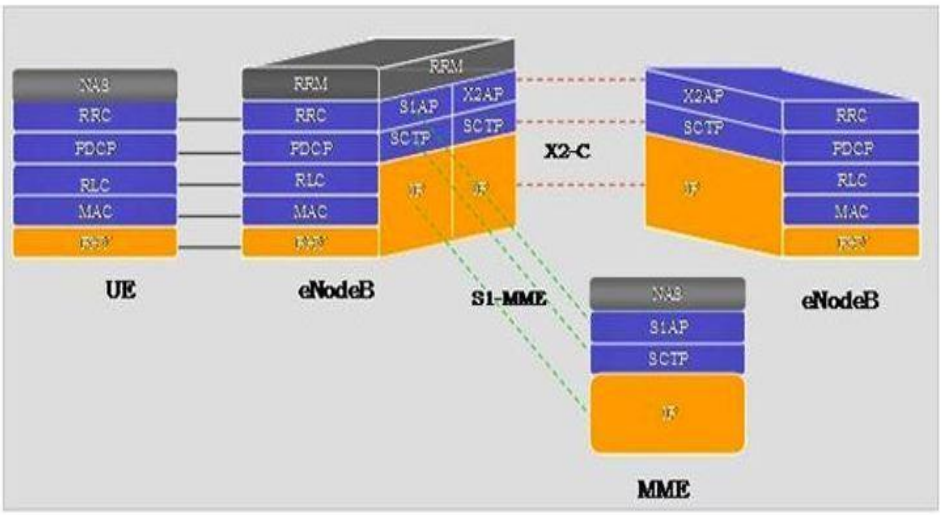

Figure 8: Control Plane Protocol Stacks [17]

1.5.2.3 E-UTRAN Architecture

The E-UTRAN architecture consists of eNodeBs that interface with the user equipment (UE) and provide user plane (PDCP/RLC/SMAC/PHY) and control plane (RRC) protocol terminations toward the user equipment (UE). eNodeB is a logical element that serves one or more E-UTRAN cells, and eNodeBs interconnects with one another with the help of an interface known as X2 and also connects using the S1 interface to the EPC, specifically to the Mobility Management Entity (MME) using the S1-MME interface and to the Serving Gateway (SGW) using the S1-U interface in Figure 9.

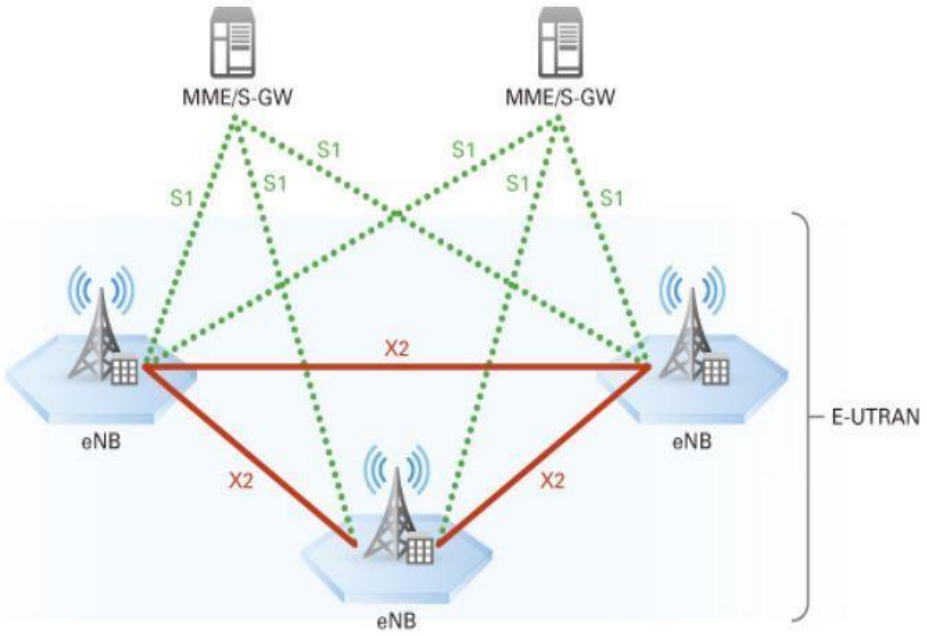

Figure 9: E-UTRAN Architecture

The E-UTRAN in LTE architecture includes a single node, i.e., the eNodeB, that interfaces with the user equipment (UE). The goal of this simplification is to reduce the latency of all radio interface operations. eNodeB connects to one another via the X2 interface, and theyconnect to the PS core network (EPC) via the S1 interface. The eNodeBs are generally interconnected with one another via an interface known as “X2” and to the EPC via the S1 interface. The eNBs are connected using the S1 interface to the Evolved Packet Core (EPC).

1.5.2.4 EPC Functions

From the latest evolution of the 3GPP core network architecture, the Evolved Packet Core (EPC) is a new, all-IP mobile core network developed by 3GPP to allow handover between different technologies. 3GPP Release 8 standards specify the EPC to improve network performance by separating control and data planes and through an IP architecture. The EPC Functions include:

- Network Access Control Functions
- Packet Routing and Transfer Functions
- Mobility Management Functions
- Security Functions.
- Radio Resource Management Functions
- Network Management Functions
- Charging Functions [15]

1.5.3 Advantages

- LTE network uses all IP network architecture. Due to this fact, it dedicates to packet-switched operations. It supports data as well as voice. The voice can be transported using voice-over LTE protocols (i.e., VOIP) and fall-back to legacy networks (i.e., 2G/3G).
- As LTE supports MIMO, a higher data rate can be realized.
- LTE uses SC-FDMA in the uplink so mobile terminals can have low power during transmissions, and thereby, battery life can be enhanced on the user side.
- As LTE downloads the files quicker, connection with the network gets released faster for every connection. This will decrease traffic load over the LTE network.
- LTE uses OFDMA in the downlink, which effectively uses channel resources. This increases the total user capacity of the LTE network as different users utilize different channels to access the system.
- It will not take much time for the user to open the browser and download a high bandwidth movie. This betters user experience to a great extent. This is due to the low latency in LTE.

- Due to improved architecture, the handoff is smooth from one region to the other. This results in smooth data streaming without any interruption of ongoing data transfers.
- Higher versions of LTE will further elevate the performance of existing LTE standard-based products.

1.5.4 Disadvantages

- The user needs to have a mobile phone which supports LTE functionality. This will incur a cost to the user to avail of the LTE service, as many phones in use at that time were not built to support LTE functionality.
- LTE network is a completely new network that requires the installation of antennas and equipment to make it operational.
- The LTE system is complicated. Hence, requires skilled engineers to manage and maintain the system. They need to be paid higher remunerations to retain them.
- As LTE service has started recently, it takes time to stabilize and have the LTE signal available freely. At its launch, the service was available only in some regions or cities. The problem could have been avoided by having multi-mode supported mobile phones so that users can avail other networks such as 2G, 3G, in case a 4G signal is not available [18].

2 Chapter 2: FIFTH GENERATION

2.1 INTRODUCTION

The tremendous rise in the number and variety of connected devices and the increase in user/network traffic volume and types, along with the performance limitations of 4G technologies, have motivated industry efforts and investments toward developing, defining, and deploying systems for the fifth generation (5G) of mobile networks [19].

Additionally, 5G will enable new services, including industrial Internet of Things (IoT) connectivity plus critical communication. 5G targets are set high with data rates up to 20Gbps and capacity increases of up to 1000 times with flexible platforms for device connectivity, ultra-low latency, and high reliability. A number of new use cases and applications can be run on top of 5G mobile networks. Expect that 5G can fundamentally impact all sections of society by improving efficiency, productivity, and safety. First, 2G, 3G, and 4G were predominantly about connecting people: they enabled people to call one another or access the Internet from virtually anywhere, anytime. 5G, with its capabilities for Ultra-Reliable Low Latency Communication (URLLC) connectivity, has been designed from the outset for high-performance IoT [20].

Due to this, there is a continuous need to push the performance envelope of the wireless systems to new limits to satisfy the demands for greater network capacity, higher user throughput, more efficient spectrum utilization, wider bandwidths, lower latency, more reliability, lowest power consumption, increased connection density, and higher mobility with virtualized and software-defined network (SDN) architectures. The 5G network architecture comprises modular network functions that can be deployed, configured, and scaled on-demand to accommodate a variety of use cases in a smart and cost-efficient manner [19].

2.2 5G Use Cases and Scenarios

The 3GPP study identified multiple potential markets and a high number of individual use cases for 5G and categorized those use cases into five families that share common requirements

and characteristics. The first three are the primary use case families for 5G, while the last two have been singled out from the others by 3GPP.

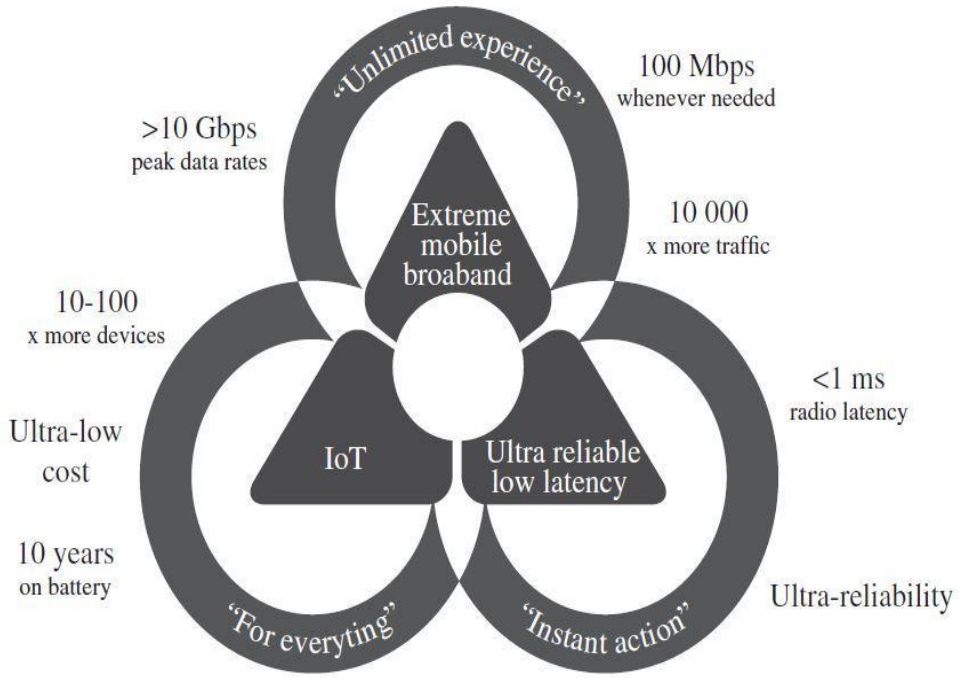

Figure 10: Main 5G Targets [20]

2.2.1 Enhanced Mobile Broadband (eMBB)

Enhanced mobile broadband (eMBB) addresses markets similar to LTE, but with improved Capabilities. Particularly, the eMBB use case calls for a higher data rate than LTE can provide, the expected data rate in more typical conditions, in terms of the peak data rate in ideal conditions, and the minimum in cases of poor network coverage. Applications are from the traditional business and consumer markets. Cases include data downloads and real-time video in an indoor office environment, traffic hotspots such as shopping centers and other high-density urban environments, public events such as sports events and concerts, and the provision of a consistent mobile broadband experience for consumers in rural areas and on public transport.

2.2.2 Massive Machine-type Communications (mMTC)

Machine-type Communications (MTC) is also known as the Internet of things (IoT). Both the terms allude to wireless communications between autonomous machine-type devices minus

any direct human interaction. However, the former is sometimes restricted to point-to-point communications, but the second always means that the devices communicate over the Internet. The terms massive Internet of things (MIoT) and massive machine-type communications (mMTC) both imply the aggregation and analysis of data from huge numbers of connected devices.

MTC applications are wide-ranging - from wireless tracking devices to geolocation of goods in a warehouse to cars in a delivery system or dogs in a local neighborhood. Some examples include wireless sensors, such as motion-triggered security cameras, environmental sensors, and strain gauges for bridges, buildings, and other civil engineering structures. A third application is in electronic health monitoring, including reporting medical information such as body temperature, blood sugar levels, and blood pressure levels to help permit independent living for the elderly and those with long-term medical conditions.

MTC has variant performance requirements from mobile broadband. The devices must be inexpensive and require low power consumption to ensure long battery life. The devices must also relay data successfully if the received signal is weak enough to support devices like smart meters that might be installed inaccessibly inside a building. The data rate of an individual machine-type device is usually low, but the collective data rate can still be high.

2.2.3 Ultra-reliable Low-latency Communication (URLLC)

Ultra-reliable low-latency communication (URLLC), also called critical communication (CriC), is a use case defined by the need for very low latency, usually in conjunction with very high reliability. LTE usually supports latencies of a few tens of milliseconds in the case of non-roaming mobiles, which are small enough for conversational VoIP and video over IP but are still large for other applications. For example, tactile networks are accountable over the timescales of human touch and require feedback in a few milliseconds to give an impression of immediate response. Virtual reality headsets find extensive consumer applications in immersive entertainment and online games that require similar latencies to curb feelings of nausea on the user's part.

URLLC is a use case in which 5G can afford to offer the best benefits over LTE because of its technical requirements and the opportunity it provides network operators to go beyond their traditional consumer markets.

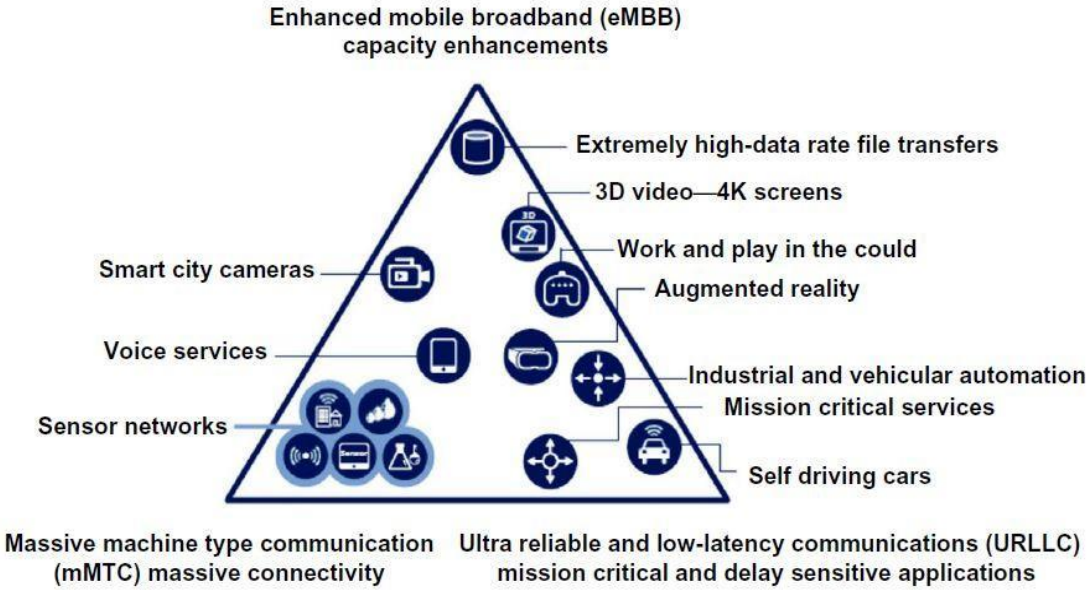

Figure 11: 5G Use Case Scenarios [19]

2.2.4 Vehicle-to-everything Communication (V2X)

Vehicle-to-everything (V2X) communication talks about exchanging information between the mobile telecommunication network and a road vehicle, plus with other vehicles and pedestrians nearby. It is often spoken of as an aspect of URLLC but is separated into its own use case by 3GPP [21].

2.2.5 Network Operation

Network Operation enhances network slicing, routing, migration and interworking and energy saving.

2.3 Requirements for 5G

ITU-R has identified some requirements for 5G depending on the end user's experience, system performance, services, and operation and management. These requirements aim to make sure

that technologies can fulfill the objectives of IMT-2020 and to set upon a specific level of performance that each proposed RIT/SRIT has to consider for IMT-2020.

- **Peak data rate:** It is the maximum data rate achievable under ideal conditions (in bit/s), this is the received data bits under the assumption of error-free conditions considered to a single mobile station, in the case all assignable radio resources for the corresponding link direction are in use (i.e., excluding radio resources used for physical layer synchronization, guard bands reference signals or pilots, and guard times). The minimum requirements for peak data rate are thus: Downlink peak data rate is 20 Gbit/s and Uplink peak data rate is 10 Gbit/s.

- **Peak spectral efficiency:** This is under ideal conditions, the maximum data rate normalized by channel bandwidth (in bit/s/Hz). The minimum requirements for peak spectral efficiencies are as follows: Downlink peak spectral efficiency is 30 bit/s/Hz and Uplink peak spectral efficiency is 15 bit/s/Hz.

- **User-experienced data rate:** The achievable data rate across the coverage area is available everywhere to a mobile user/device (in Mbit/s or Gbit/s). The minimum requirements for this are Downlink user experienced data rate of 100 Mbit/s, and Uplink user experienced data rate of 50 Mbit/s.

- **Peak spectral efficiency:** The maximum data rate under ideal conditions normalized by channel bandwidth (in bit/s/Hz). The minimum requirements are: Downlink peak spectral efficiency is 30 bit/s/Hz, and Uplink peak spectral efficiency is 15 bit/s/Hz.

- **Area traffic capacity:** The total traffic throughput served per geographic area (in Mbit/s/m2). The minimum target value for Area traffic capacity in downlink is 10 Mbit/s/m2 in the Indoor Hotspot – eMBB test environment.

- **User plane latency:** It is the time taken to successfully send a packet from the source till the time destination receives it (in ms). The minimum requirement for eMBB is 4ms, and for URLLC is 1ms.

- **Control plane latency:** It refers to the transition time from a most "battery efficient" state (e.g., Idle state) to the start of continuous data transfer (e.g., Active state). The minimum requirement for control plane latency is 20ms. Proponents are encouraged to consider lower control plane latency, i.e., 10ms.

- **Connection density** is the total number of devices fulfilling a specific quality of service (QoS) per unit area (per km2). The minimum requirement is 1,000,000 devices/ km2 in mMTC usage scenario.

- **Energy efficiency:** This is to minimize the energy consumed by RAN in relation to the traffic capacity provided by the capability of an RIT/SRIT measured in bit/Joule.

- **Reliability** relates to the capability to transmit a given amount of traffic in a predetermined time duration with a highly successful probability.

- **Mobility:** It is the maximum mobile station speed at which a defined QoS can be achieved (in km/h).

- **Spectrum and bandwidth flexibility:** It refers to the flexibility of the system design in handling various scenarios. In particular, the capability to operate at different frequency ranges, including higher frequencies and wider channel bandwidths than today. The bandwidth requirement is at least 100MHz [22].

2.4 Enabling Technologies for 5G

In order to meet the strict requirements discussed previously, a number of technology candidates have been considered and widely discussed. The development in technology will happen both in the radio access network (RAN), the core network, and the end-to-end system.

2.4.1 5G Radio Access Network

The enabling technologies for 5G RAN include mmWave communication, massive Multiple Input Multiple Output (MIMO), ultra-dense small cell, Machine-to-Machine (M2M), and

Device-to-Device (D2D) communications, cloud-RAN, and mobile edge and fog computing. These technologies are:

2.4.1.1 mmWave Communication

One of the key features of the 5G system is to have higher capacity in terms of data rate, for example, up to tens of Gbps at the peak data rate. To achieve those targets, we require more spectrum availability. However, current wireless systems typically operate in a spectrum band, ranging from hundreds of MHz (e.g., 700 MHz) to below 3 GHz (e.g., 2.6 GHz). These spectrum usages are not sufficient enough for 5G. One of the most effective solutions for expanding the bandwidth range is to exploit the very high spectrum bands (e.g., > 10 GHz). Due to some reasons, such as high propagation loss, mmWave communication is commonly used for indoor environments or backhaul links. However, many research initiatives have illustrated the feasibility of mmWave technology for 5G mobile networks by adopting many recent advances in propagation modeling or channel modeling to create a large amount of bandwidth.

Apart from the benefits of allowing larger bandwidth and higher data rate that makes the mmWave a promising technology for 5G, there are still some challenges and open issues that need to get solved in the future, such as interference and heterogeneity.

2.4.1.2 Massive MIMO

In order to meet the 5G requirements in terms of network density and capacity enhancement, one of the most prominent solutions is to densify the number of deployed antennas, which refers to a technical solution called massive MIMO. Fundamentally, MIMO is an antenna technology for wireless communications in which multiple antennas use to transmit and receive data. In fact, the MIMO concept has commonly used in current 4G networks, which refers to multi-user MIMO communication, where a multiple-antenna base station simultaneously serves several users; whereas, massive MIMO defines as a multi-user MIMO system, where the number of base station's antennas and the number of users are significant. Such a feature as having more antennas at the base station promises to increase the network capacity and density. More importantly, massive MIMO significantly enhances spectral and energy efficiency. These reasons make massive MIMO an essential technology for 5G. Figure 12 depicts the concept of massive MIMO. Apart from the benefits of massive MIMO, several research questions still

need to be addressed, such as mitigation of plot contamination, channel estimation, implementation-aware algorithmic design, etc.

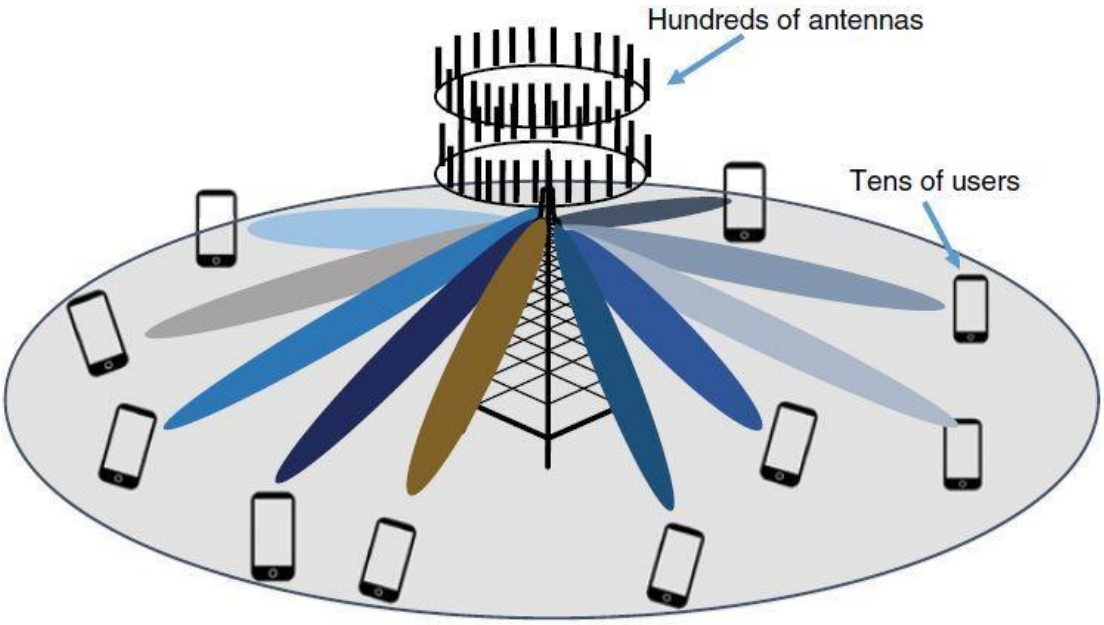

Figure 12: Massive MIMO concept illustration [11]

2.4.1.3 Beamforming

Beamforming is used with phased array antenna systems to focus the wireless signal in a chosen direction, normally towards a specific receiving device. This results in an improved signal at the user equipment (UE) and less interference between individual UE signals. Phased antenna arrays design is that the radiation patterns from each individual element combine constructively, with those from neighboring elements forming an effective radiation pattern - the main lobe - which transmits energy in the desired direction. At the same time, the antenna array design is in a way that signals sent in undesired directions destructively interfere with each other, forming nulls and side lobes.

The overall antenna array system maximizes the energy radiated in the main lobe while limiting the energy in the side lobes to an acceptable level. The direction of the main lobe, or beam, is controlled by manipulating the radio signals applied to each of the individual antenna elements in the array. Each antenna has the same transmitted signal, but the phase and amplitude of the signal fed to each element are adjusted, steering the beam in the desired direction.

Fast steering of the beam is achievable since the phase and amplitude of each signal can be in control electronically, allowing adjustments are to be done in nanoseconds [23].

2.4.1.4 Ultra-Dense Small Cells

Another way of increasing the network density and improving the throughput is to densify the number of wireless nodes, which have a smaller coverage range than the macro-cell base stations used in the 3G and 4G legacy systems. The technical solution behind this idea denotes small cell technology. The Small Cell Forum defines "small cells" as an umbrella term for operator-controlled, low-powered radio access nodes with a coverage range of between ten to several hundreds of meters, including those operating in licensed spectrum and unlicensed carrier-grade Wi-Fi. An example of small cell deployment is in Figure 13.

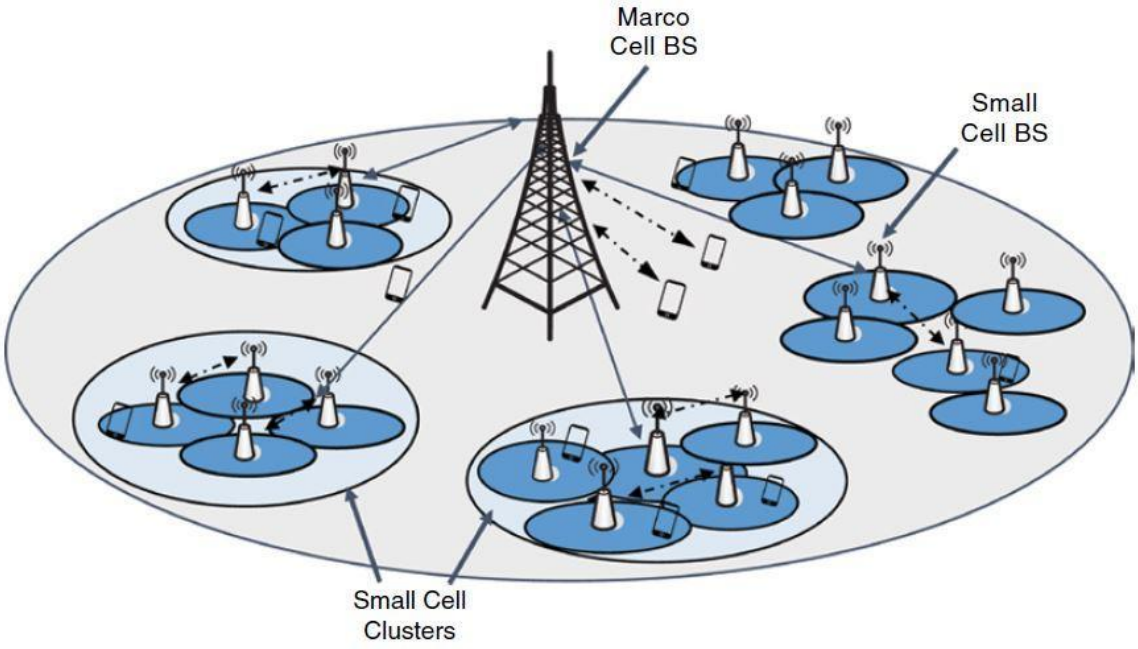

Figure 13: Small cells deployment illustration [11]

With small cells, the size of the cell reduces, meaning they bring the network much closer to the user, thus better serving high traffic areas such as indoor and hotspot areas. In addition, the higher number of low-powered transmission points on the small cell network enables better use of available frequency resources, thus improving spectral efficiency. Furthermore, the 5G system can construct in a heterogeneous fashion, where macro and small cells are co-located and may be connected via wireless backhaul links, thus providing increased levels of network

capacity through traffic offloading. However, the heterogeneity of small cells in the network will pose challenges in terms of interference and mobility management, thus affecting system performance as a whole.

2.4.1.5 M2M and D2D Communications

a. M2M Communication

As mentioned previously, two-thirds of the use case categories of 5G will be related to IoT and Machine Type Communication (MTC), including massive and critical communications. Therefore, although the concept of M2M or MTC communication was introduced in 4G LTE systems by 3GPP some time ago, it is still considered one of the key enablers for 5G. Fundamentally, M2M communication refers to the automated data communications among devices and the underlying data transport infrastructure. The data communications may occur between an MTC device and a server or directly between two MTC devices. There are several services and applications enabled by M2M communication, such as monitoring and metering, home and industry automation, health care, and automotive, as shown in Figure 14(a).

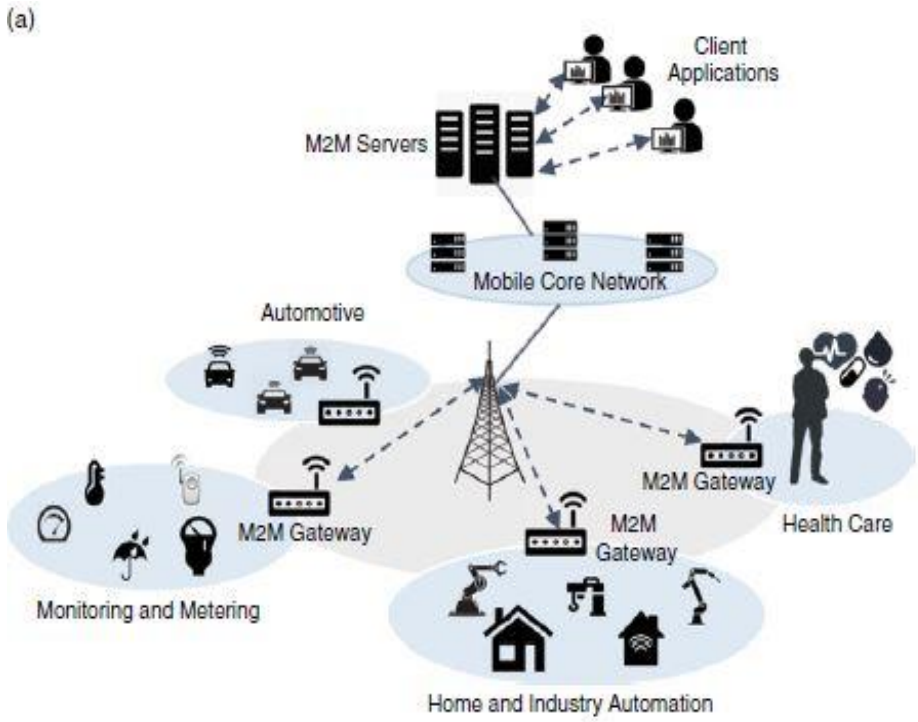

Figure 14(a): M2M Communications and use case scenario [11]

b. D2D Communication

D2D communication is direct communication between two mobile users/devices without traversing through a network infrastructure. 3GPP has specified it in LTE Release 12. By exploiting direct communication between devices, D2D communication can help improve spectrum efficiency, user data rate gain, and reduce latency and energy consumption, thus being considered one of the key components of the 5G system. In general, the operation of D2D communication can be in-band D2D on the licensed cellular spectrum (e.g., LTE) and out-of-band D2D on the unlicensed spectrum (e.g., Wi-Fi). There are some use cases and application scenarios for D2D, such as proximity-based services, gaming, public safety, vehicular communications, and offloading, as shown in Figure 14(b).

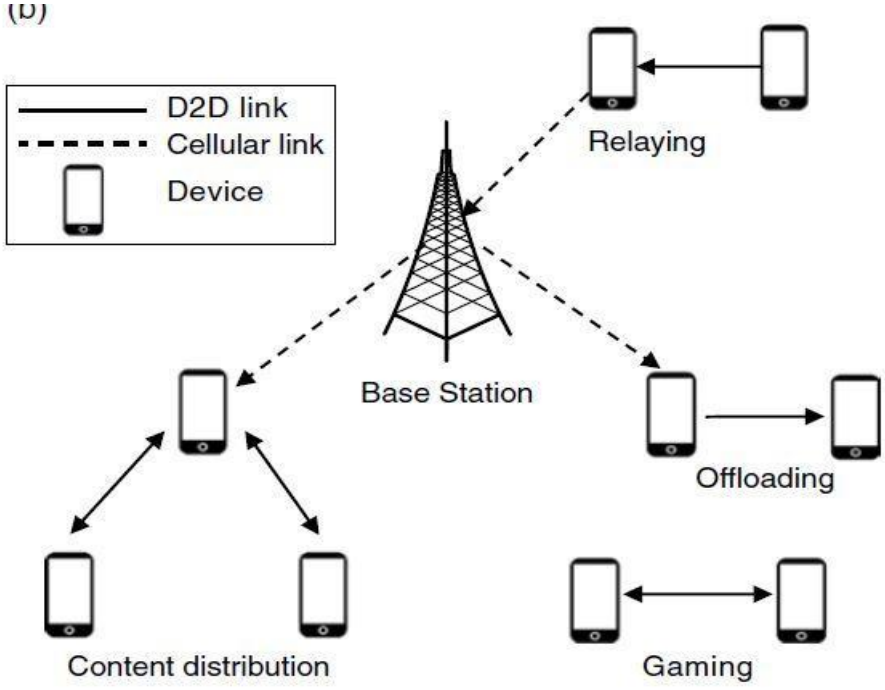

Figure 14(b): D2D Communications and use case scenario [11]

2.4.1.6 Cloud-based Radio Access Network

Cloud-based Radio Access Network (Cloud-RAN) is an ideal solution to design the radio access part of 5G networks since it enables energy efficiency, cost savings on baseband resources, improvements in network capacity, increased throughput, etc. The Cloud-RAN is essentially the decoupling of the Remote Radio Head (RRH) from the Base-Band Unit (BBU)

of a base station and the implementation of BU in a centralized cloud computing environment. RRHs connect to a BBU pool using high-speed fiber or microwave-link front-haul networks. Apart from the benefits that Cloud-RAN offers to the design of the 5G system, various challenges needed to have a look into before fully utilizing its benefits; such as front-haul constraints and performance optimization, placement optimization of RRHs, efficient scheduling, and elastic scaling of BBUs in the BBU pool. Some other research directions in the future could be the incorporation of C-RAN and distributed RAN (D-RAN) or research on heterogeneous CRAN (H-CRAN).

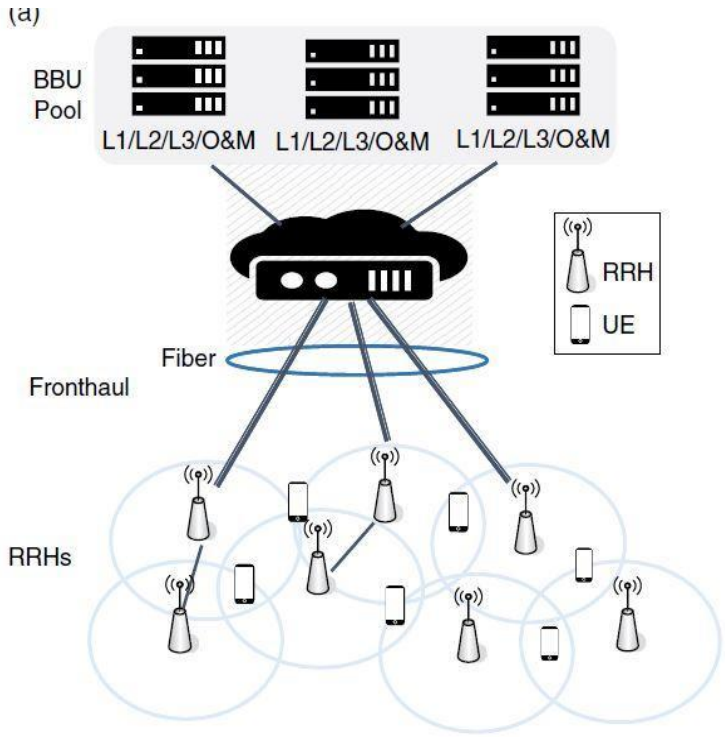

Figure 15: Cloud-RAN concept [11]

2.4.1.7 Mobile Edge and Fog Computing

As the world moves to 5G, many of its applications and services will require very stringent latency in the order of milliseconds. One of the most obvious solutions is to bring the IT services and processing capabilities down to the edge of the mobile network, within the RAN, and in close proximity to mobile users. This refers to the concept of Mobile Edge Computing (MEC) technology and its sibling Fog Computing. MEC aims to reduce latency, ensure highly efficient network operation and service delivery, and improve user experience. With this

capability, MEC will open new frontiers for network operators, application service providers, and content providers by enabling them to introduce innovative services and applications. Typical examples of MEC services are augmented reality, RAN-aware video optimization, connected cars, IoT, etc. A similar concept to MEC is Fog Computing (FC), a paradigm in which cloud computing resources are extended to the edge of the network to create a highly virtualized platform that accounts for storage, computing, and networking services between end-devices and traditional data centers. Some of the prominent features of FC, which are suitable for 5G communications, are low latency, location awareness, real-time interactions, mobility support, geographical distribution, and the predominance of wireless access.

2.4.2 5G Mobile Core Network

SDN, NFV, and cloud computing are considered key technologies to design the core part of 5G networks. These technologies are as follows:

2.4.2.1 Software Defined Networking (SDN)

In terms of network flexibility and programmability, SDN is the best technology candidate for developing 5G networks. SDN concept got first proposed in the campus and data center network areas. It features the separation of the data plane from the control plane and facilitates network management through the abstraction of network control functionalities, as shown in Figure 16. Being adopted by 5G, SDN will enable a more agile and flexible core network architecture. In addition, the programmability and openness characteristics of SDN will help mobile operators shorten the life cycle of introducing their new services and innovation into markets. By separating data planes and the control, the network infrastructure can be constructed on demand and based on service requirements (network-as-a-service), thus improving resource efficiency. We can also use the SDN concept in the RAN domain, where the SDN controller could control and schedule the radio resources for base stations, thus improving spectrum efficiency and mobility management.

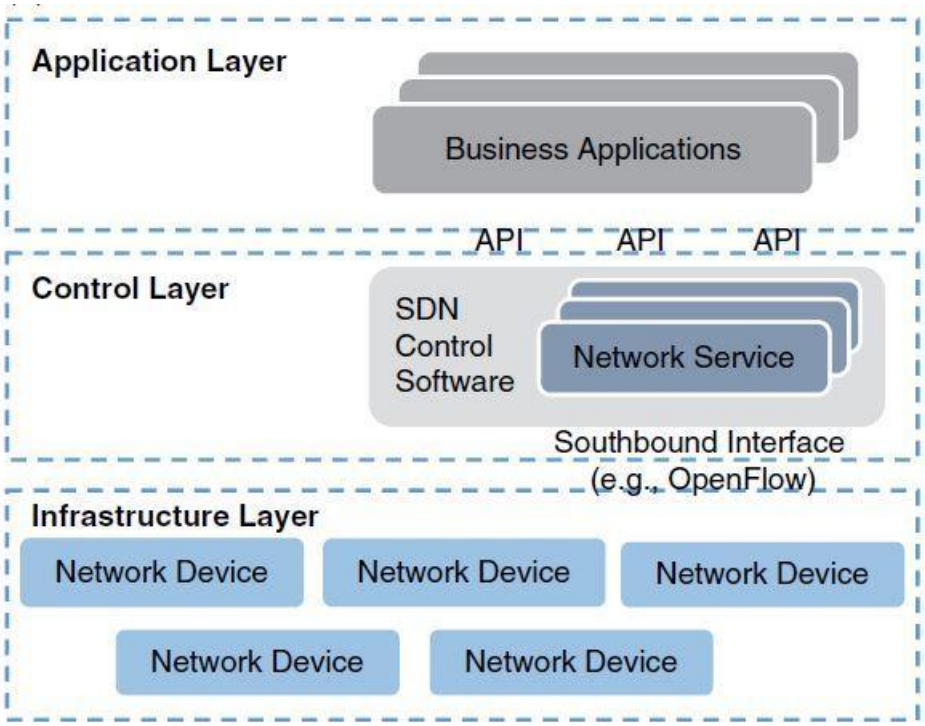

Figure 16: SDN Architecture [11]

However, there are still many challenges and issues with SDN, such as the scalability problem due to the centralization of network intelligence, extra latency between devices and the SDN controller, security problem of the communication channel between the control and data planes, the lack of standardization on designing the protocol communicating between the control and data planes, policy and charging enforcement. Other aspects related to the adoption of SDN into mobile networks, such as placement problems of SDN controllers, mobility management, load balancing, etc.

2.4.2.2 Network Function Virtualization (NFV)

As described previously, the 5G system is not just about high data rate, low latency, and flexibility. It is also about cost efficiency, which will impact the revenue of mobile operators. 5G mobile operators will expect the cost for the deployment, which refers to capital expense or CAPEX, and the cost for operation and management, which refers to operational expense or OPEX, to be as low as possible. NFV is the foundation for these capabilities and identifying as the cornerstone of 5G core network solutions. Figure 17 shows the reference architectural framework of NFV. Essentially, NFV refers to the relocation of network functions, which are traditionally implemented on dedicated costly Hardware platforms to software appliances

running in the cloud environment or on general-purpose commodity servers. By operating the network function as software, it is easier for mobile operators to dynamically scale the resources (computing, storage, and networking) according to changes in traffic demands and to faster time-to-market of new services. In addition, the combination of SDN and NFV has encouraged the development of new networking paradigms, such as network service chaining and network slicing. Although NFV proved to be the critical enabler for developing 5G, especially the core part, there are many challenges, such as optimizing network functions placement, resource allocation, management and orchestration, and network performance.

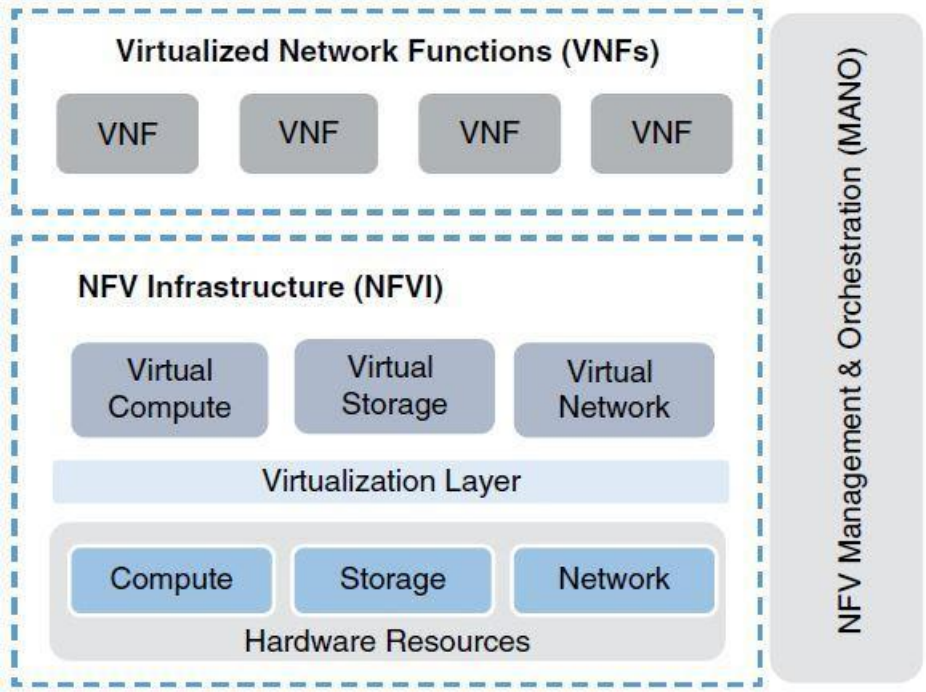

Figure 17: NFV Architecture [11]

2.4.2.3 Cloud Computing

As described in the previous section, cloud computing has been considered an ideal solution for re-designing the current RAN architecture. With its anticipated benefits, such as on-demand and elastic provisioning of services and resources over the Internet, cloud computing has made it one of the key enablers for designing 5G core networks. In this case, realizing 5G core network functions as virtual machines or containers controlled by the cloud manager. The capability of providing resources in a multi-tenant model of cloud computing allows mobile operators to implement the concept of mobile virtual network operators (MVNO) much more

easily than in the past. In addition, a pay-as-you-use business model and the ability to move and consolidate the resources offered by cloud computing can help mobile operators reduce their capital and optimize operational expenses.

Centralizing all resources will ease the management and provisioning process, but it will result in a long delay for end-to-end communication, which may not be suitable for some of the newly defined 5G services. Therefore, combining cloud computing and other computing paradigms, such as mobile cloud computing and edge computing, will be a promising direction to investigate in future research.

2.4.3 5G End-to-End System

The key technology enablers for constructing a 5G end-to-end system include network slicing, management, and orchestration.

2.4.3.1 Network Slicing

Today's 4G systems have been optimized mostly for serving human-to-human communication where mobile phones are the main players. However, we can expect that the 5G system can support diverse services and applications with various characteristics and requirements in the future, where IoT devices will become dominant. Such IoT-related services will require different features and network capabilities in terms of latency, data rate, mobility, reliability, security, etc. Therefore, to guarantee these requirements and improve the network performance and resource utilization, each service type should be provided as an end-to-end, isolated, and infrastructural environment to operate. In this sense, the network slicing concept will become the foundation of all capabilities.

Although network slicing is recognized widely as the key characteristic of 5G by many network operators and vendors, it still needs to be standardized. Thus, there are variants of defining what slicing is. Slicing is the basic concept of Network Softwarization; it lets logically isolated network partitions (LINPs) that have a slice to being considered as a unit of programmable resources such as computation, network, and storage. A network slice, namely “5G slice”, comprises a collection of 5G network functions and specific radio access technology settings that combine for the particular use case or business model. Figure 18 illustrates an example of

the network slicing concept, with three different slices corresponding to the three main 5G use case categories as discussed. To this end, implementing the network slicing concept is on an end-to-end basis. The 5G system will be composed of multiple end-to-end slices, where dedicated resources and quality of service (QoS) are guaranteed.

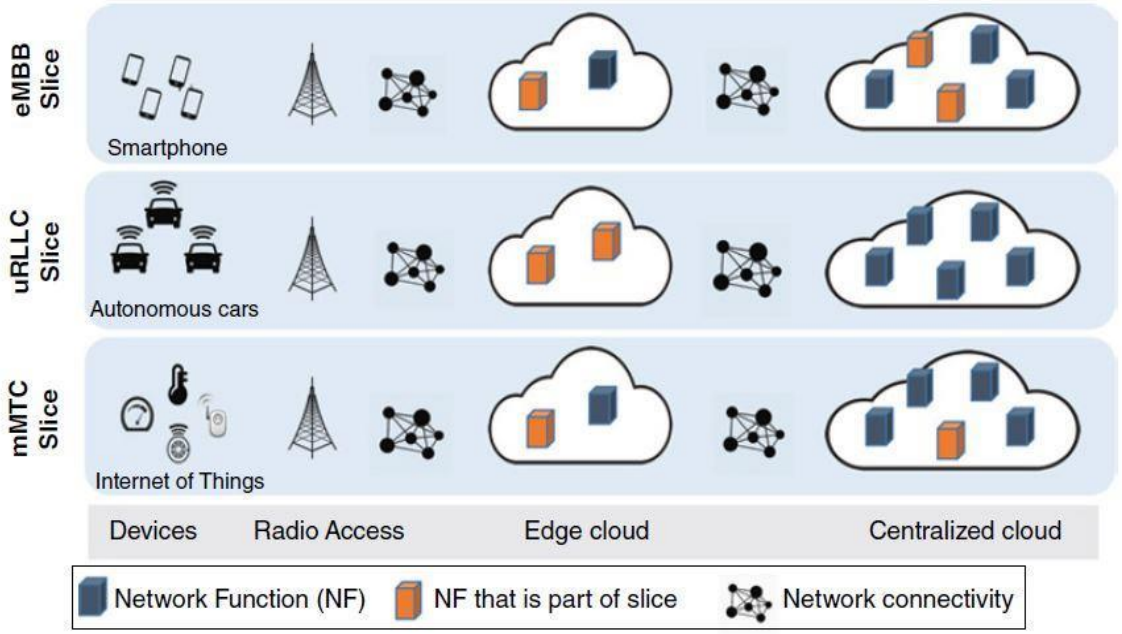

Figure 18: Network Slicing [11]

However, network slicing still presents many challenges and gaps that must be fulfilled in future studies, such as slice definition, lifecycle management of a slice, the resiliency of slice control, resource allocation and optimization within a slice and between slices, strongly guaranteeing security within a slice and between slices, end-to-end QoS management, integrating with other technologies (e.g., information-centric networking (ICN), D2D), etc.

2.4.3.2 Management and Orchestration

When the 5G mobile networking era arrives, due to the diversity of use cases, services, and network slices created with different resource requirements, the network's management and orchestration (MANO) becomes more and more crucial. The role of MANO will be managing the whole network infrastructure in terms of fault management, configuration, accounting, performance, and security. More importantly, MANO will be in charge of lifecycle

management and provisioning the network resources for the end-to-end connectivity of network slices in a dynamic, automated, and efficient manner. As illustrated in Figure 19, the end-to-end management and orchestration role will be multi-domain, multi-operators, and multi-technology spanning from the infrastructure layer to the application (service) layer and spanning from the RAN to the network core.

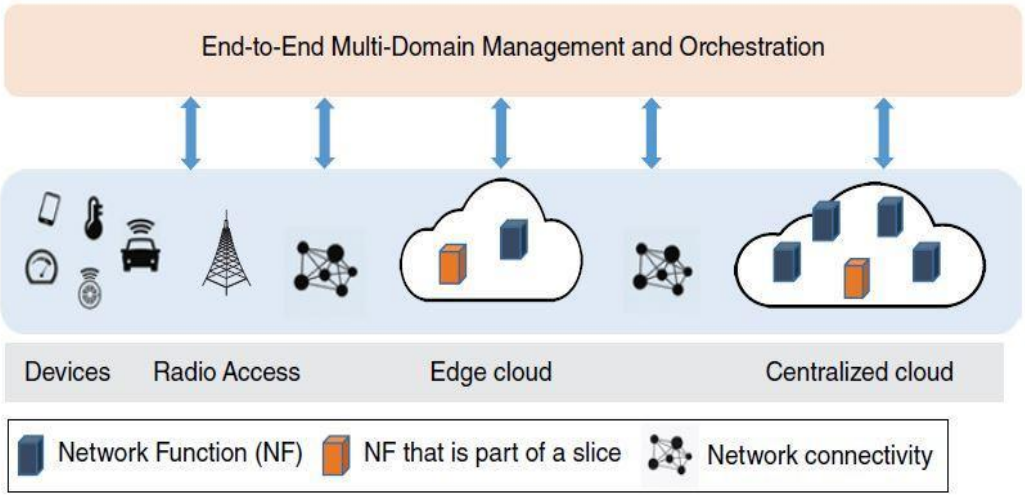

Figure 19: End-to-End multi-domain management and orchestration [11]

In 2014, the NFV MANO working group in the European Telecommunications Standards Institute (ETSI) had specified in its technical specification an architectural framework showing main components and their functionalities and the operations within the framework. In the meantime, there are several efforts that have implemented the NFV MANO concept as open-source platforms, such as OpenStack Tacker, OpenBaton, etc. These open-source platforms are being applied to today's 4G core network and will integrate into the 5G network architecture deployment. However, due to the diversity of network resources from RAN to the core, the current NFV MANO framework should be extended to manage virtualized network functions and resources and physical nodes. In addition, dynamically managing and orchestrating the network services and slices would also be challenging [11].

2.4.4 5G Network Challenges

As 5G Network uses Virtualization and SDN/NFV for infrastructure to provide services and use cases but service security cannot be provided unless the network infrastructure is secure. These are the main challenges we encounter in the development of 5G infrastructure:

- **Frequency:** Wireless carriers need to bid for higher spectrum bands in spectrum auctions to construct their 5G networks.
- **Deployment:** While higher frequencies transmit larger quantities of data, they are very susceptible to physical interference; therefore, a larger number of antennas and base stations need to get installed to establish sufficient coverage.
- **Cost:** Laying down the physical groundwork for 5G-enabled devices, robots, autonomous vehicles, appliances, and city infrastructure will need substantial, costly upgrades, estimated to be in the trillions of dollars.
- **Regulations:** Government regulators will burden with establishing extensive regulations for cybersecurity, spectrum availability, EMF radiation, and infrastructure sharing.
- **Security:** Carriers and network consortiums need to ensure that cloud-based and data virtualization services are uber-secure to cope with the increased connectivity that comes with the 5G rollout and sophisticated cybersecurity threats [24].

3 Chapter 3: 5G Network Architecture

3.1 INTRODUCTION

5G architecture is an evolution and amalgamation of current 4G architectures, but it establishes a Service-Based Architecture (SBA). The 3GPP says that a set delivers the SBA for a 5G core network of interconnected Network Functions (NFs) having the authorization to access one another's services. Few of the key differences/focus areas:

- Contrasting to a hard-wired, fixed-function, appliance-based architecture, which was the case for 4G LTE Core (or Evolved Packet Core (EPC)), full realization of the potential of 5G means moving to cloud-based open platforms and software.
- EPC (4G Core) elements were designed to be implemented on virtualized physical nodes but not constructed to be virtualized right from the outset.
- Network elements in 5G core are cloud-native, called "functions" vs. "nodes."
- Programmability and automation are an essential part of the target 5G architecture.
- With the flexibility, virtualization, and programmability, the new architecture would better support the possibility of diverging architectures for new services.

To summarize – 5G core is designed for three enhancements:

1. Control and User Plane Split: Mapping of 4G Core to 5G Core elements Access and Mobility Management Function (AMF), Session Management Function (SMF), User Plane Function (UPF).
2. Native support for Network Slicing for the 5G Use Cases including enhanced Mobile Broadband (eMBB), massive Machine Type Communications (mMTC) & critical MTC, and Ultra-Reliable and Low Latency Communications (URLLC).
3. Service-Based Architecture: A service-based architecture delivers services as a set of “Network Functions.”

3.1.1 4G Control and User Plane Separation (CUPS) EPC

The distinction of Control and User Plane for the 4G architecture was initiated with 3GPP Release 14. It distinguished the packet gateways into control and user planes, thus allowing for more flexible deployment and independent scaling, achieving benefits in OpEx and CapEx.

The next phase in the evolution to 5G was to give new names to core network entities and either split or merge them based on the functions within the consumer or control plane in the 5G architecture. For those with the 4G background, a few 4G CUPS Core elements can be freely mapped to renamed 5G Core elements. Some of the few key ones are:

- **Next Generation NB (gNB):** The new radio access technology is known as New Radio (NR) and replaces LTE. The new base radio station is called next-generation NB (gNB) (or gNodeB). It supersedes the eNB (or eNodeB or Evolved Node B) in 4G-LTE or NodeB in 3G-UMTS.

  The gNB regulates radio communications with the 5G capable User Equipment (UE) via the 5G New Radio (NR) air interface. Although, some types of gNB may connect to the 4G EPC instead of the 5G Core.

- **The Control Plane – AMF and SMF:** Mobility Management Entity (MME) in LTE is a signaling node for UE access and mobility, establishing the bearer path for UE's and mobility between LTE and 2G/3G access networks. Mobility Management Function in LTE replaces with:
  a. Access & Mobility Management Function (AMF): overseas connection, authentication, mobility management between device and network. It receives connection and session-related information through the UE.

  b. Session Management Function (SMF) – handles IP address allocation, session management, and policy enforcement control.

- **The Data Plane – User Plane Function (UPF):** The CUPS decouples Packet Gateway (PGW) control and user plane functions enable the data forwarding component (PGW-U) to be decentralized, then mapped to the UPF for the 5G Core. The user plane function includes a single entity User Plane Function (UPF); it puts together functionality from previous PDN-Gateway (P-GW) and EPC Serving-Gateway (S-GW). UPF handles packet routing and forwarding and Quality of Service (QoS) [25].

Different factors impacted the 5G network architecture. There were many factors, including the preparation for cloud-based implementations, readiness to deal with larger data rates and lower latencies compared to previous generations, enabling newer services, and the necessity to interwork with Long-Term Evolution (LTE), especially in the first phase. These factors put together have impacted 5G architecture. Apart from the radio technology development, defining a new 5G core network, which enables new service elements in terms of global and local services plus the novel concepts of network slicing or flow-based quality of service and multiple other features, thus enabling a more effective cloud-based implementation when compared to the LTE core (EPC, Evolved Packet Core).

3.2 5G Core Network Architecture

Recently, many communication service providers (CSPs) have been virtualizing their EPC, changing the deployment from a series of dedicated physical appliances (specific blade servers for each function) to virtualized network functions hosted on general-purpose servers in the cloud. These virtualized core networks allow CSPs to expand into new services, such as cellular Internet of Things (IoT), or to provide enterprise-specific core services or to expand legacy EPC networks. In practice, this virtualization of EPC does not work well in multi-vendor deployments.

Therefore, support for Network Function Virtualization and Software Defined Networking was one of the first architecture requirements for the 5G core network. In addition, it was equally important to require the capability for the 5G core network deploys in a cloud-native way. Applications designed for the cloud must also follow scaling design principles such as splitting the control plane and the user plane and separating compute and storage resources.

Separation of the control and user planes provides the ability to utilize distributed edge cloud architectures for bandwidth-intensive and latency-critical applications requiring independent distribution of the user plane functions from that of the control plane functions. For example, for certain low-latency use cases, it will be important to place user plane resources close to the access point where the user/device is attached, to reduce end-to-end latency and the transport network load. The 5G core network will further enable users to be connected simultaneously to the edge and central cloud and move between them.

The core network needs to support stateless network functions where the compute resource decouples from the storage resource. The separation of computing and storage resources enables unlimited linear scalability and extreme resilience. In the 5G architecture, the so-called Unified Data Management (UDM) will host subscriber data, session data, policies, operational data, charging, and accounting data. It will also enable an open ecosystem for data exposure and analytics.

Another major difference compared to EPC is that multi-access support is built-in right from the beginning, enabling efficient use of different fixed and mobile network assets to provide access-independent value-added services, maximized data rates, increased reliability, and improved user experience.

These new capabilities will further enable the building of end-to-end network slicing, helping public and private operators address various verticals' needs by dynamically adapting the network to their specific requirements. Network slices can be separated from each other in terms of connectivity and resource allocation, creating "virtual private service networks." These configure in specific ways to achieve, e.g., low latency, high reliability, or both. New network slices can be introduced easily via automation, thus enabling new business models that were not feasible with dedicated physical networks. The resulting 5G Core Network Architecture (simplified) is in Figure 20.

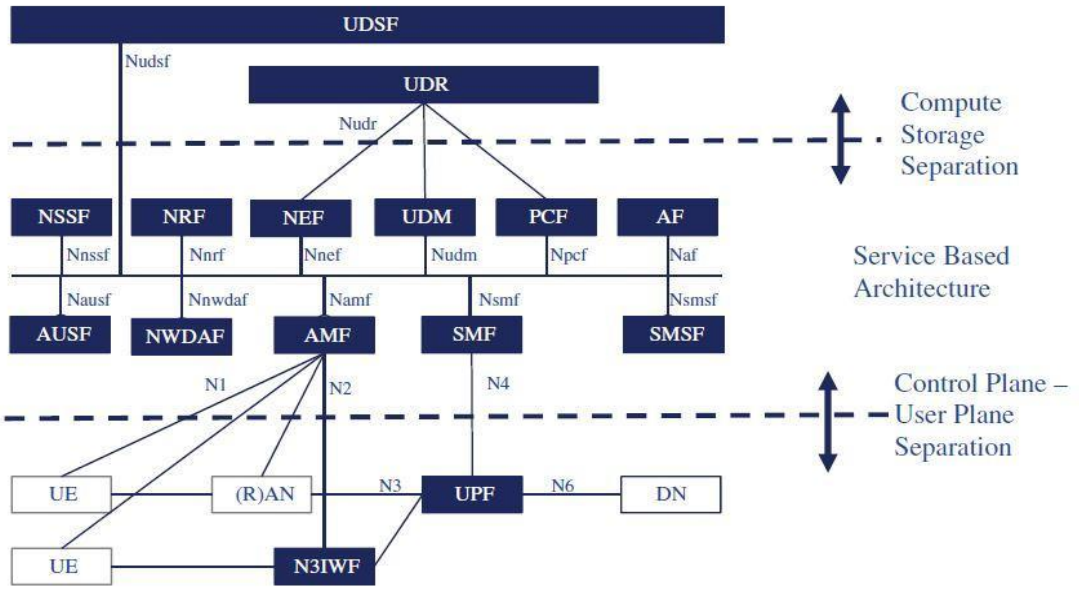

Figure 20: 5G Core Network Architecture [20]

The point-to-point architecture from 4G still applies to some interfaces in 5G. Different network functions connect over standardized interfaces that allow multi-vendor networks in the point-to-point model. This is well understood both conceptually and operationally; it has served mobile operators for decades. In the 5G system architecture, interfaces between the user equipment (UE) and the core (N1), the access network and the core (N2 and N3), and the control and user plane (N4) will still use a point-to-point architecture.

However, now, with the change to cloud infrastructure and the requirement for greater service agility, the point-to-point model is no longer the efficient option for all the core network interfaces. For CSPs that view 5G as an opportunity for transformative change both in terms of functionality and cost per bit, a service-based interface instead of point-to-point interfaces looks more attractive.

The challenge with the point-to-point architecture is that it contains many unique, or quasi-unique, interfaces between functional elements, each connected to multiple adjacent elements. This “tangle” of connections creates dependencies between functions and makes it difficult to change the deployed architecture. By introducing a new function or an existing function expanded or upgraded, the operators need to reconfigure multiple adjacent functions and test the new configuration before going live. This raises the business cases thresholds to experiment with and deploy new services.

In effect, the end-user service connects to the network, and consequently, the operator's addressable market is artificially limited. This is acceptable where the service set is well defined and relatively fixed (voice, broadband, etc.). But in the 5G era, where operators expect to offer a multitude of diverse services that must adapt to fast-changing demand or industry-specific requirements and need more dynamic and agile architecture.

The end-user service is decoupled by SBA from the underlying network and platform infrastructure to enable both functional and service agility. By the SBA operating on a cloud-native foundation, it is much simpler for an operator to add, remove, or modify virtual network functions (VNFs) from a network processing path (functional agility) and create new service paths on-demand (service agility). 3GPP has selected SBA for application in the 5G Core Control Plane (the middle part of Figure 20) Service-based interfaces are all based on HTTP and are denoted with the legend Nxxx in Figure 20.

3.2.1 Access and Mobility Management Function

The Access and Mobility Management Function is part of the control plane processing on the core side. The AMF receives the Non-Access Stratum (NAS) signaling from the UE via the access network and connects to the gNB for control plane signaling between the access network and the core network. The key functionalities are as follows:

- Terminates the (Radio) Access Network Control Plane signaling (N2).
- Terminates NAS Signaling (N1).
- Provides Access Authentication and Access Authorization.
- Provides NAS ciphering and integrity protection.
- Enables registration management, connection management, reachability management, and mobility management.
- Provides transport and proxying for session management (SM) messages between the UE and the Session Management Function (SMF).
- Transports SMS messages between the UE and the Short Message Service Function (SMSF).
- Supports location services (e.g., routes messages between the UE and the Location Management Function (LMF) as well as between the LMF and the RAN).

3.2.2 Session Management Function

The SMF takes care of the session management, such as session establishment, modification, and release, including maintenance of the tunnel between the User Plane Function (UPF) and the access network node, for example, the gNB. Further functionality includes the following:

- Provides UE IP address allocation and management.
- Selects and controls the UPF.
- Interfaces with policy functions.
- Controls and coordinates charging data collection at the UPF.
- Determines the Session and Service Continuity (SSC) mode of a protocol data unit (PDU) session.

3.2.3 User Plane Function

The UPF covers the user plane processing. There can be more than one UPF in the network. For example, there can be one UPF handling the local traffic and the other UPF handling all

other traffic. Services with very low latency requirements (those related to URLLC, etc.) are well suited for local UPF handling to avoid additional latency due to the long transmission distance. The key functionalities of the UPF are the following:

- Perform all UPFs on the packets: forwarding, routing, marking, quality of service (QoS), inspection, and so on.
- Enforce policies received from the Policy Control Function (PCF) (via the SMF).
- Send traffic usage reporting to the SMF for charging.
- Act as the anchor point for intra-/inter-radio access technology (RAT) mobility.
- Connect to external networks (N6) or other UPFs.

3.2.4 Data Storage Architecture

The Unified Data Repository (UDR) is a common database for all kinds of standardized data structures, including Subscription Data, Policy Data, Structured Data for exposure, and Application Data. Three different network functions can access the UDR to read, update, delete, and subscribe to notifications of data changes.

UDM hides application logic from the UDR. UDM takes care of, for example, the generation of authentication credentials, identity handling (both user and subscription), information storage about serving entities like the AMF and SMF, and SMS delivery support. In addition to standardized structure data, the 5G architecture enables any network function to store unstructured data in the Unstructured Data Storage Function (UDSF). This capability enables a network function to store all UE-related data in the UDSF, and thus the network function can become stateless. This capability defines for AMF in Release 15.

3.2.5 Policy Control Function

The PCF governs network behavior by applying a unified policy framework. Enabling this with the following primary capabilities:

- Creation of both SM and access management policy association.
- Provision and deletion of policy and charging management decisions.
- Provision and deletion of access and mobility management decisions.

- Delivery of UE access selection and PDU session selection policies to the UE via the AMF.
- Ability to utilize policy-control-related subscription information and application-specific information stored in the UDR.
- Ability to obtain spending limit reports from the Charging Function.
- Interactions with the NEF (Network Exposure Function), for example, to obtain Application-Function-influenced traffic steering authorization.

3.2.6 Network Exposure Function

- Exposes 5G Core capabilities and events to external applications. Uses the UDR as its data source.
- Allows third parties to securely provision information to the 5G Core (e.g., Expected UE behavior).
- Handles masking of sensitive network information toward external parties.

3.2.7 Network Repository Function

- Maintains the network function (NF) profile of available NF instances and their supported services and provides NF discovery and selection.
- Selection criteria can include a location (latency), load, Data Network Name (application), access network type, slice, etc.
- Provides much more granular policies and dynamic capabilities when compared to Domain Name System (DNS) selection.

3.2.8 Network Slice Selection

- Determines the candidate AMF(s) or AMF suitable to serve the UE.
- Select which slices the UE can connect.
- Takes care of mapping slice-specific identifiers.

3.2.9 Non-3GPP Interworking Function

The Non-3GPP Inter Working Function (N3IWF) provides support to connect UEs via non-3GPP access (only untrusted Wi-Fi access in Release 15). Enables this with the following primary capabilities:

- IPsec tunnel establishment supports UE to authenticate and authorize it to access the 5G core.
- Establishes IPsec Security Association to support PDU Session traffic.
- Performs tasks related to user plane traffic such as relaying uplink and downlink user plane packets between UE and UPF and packet marking.
- Handling of control plane traffic such as relaying NAS signaling between UE and AMF, and passing signaling from SMF related to PDU session and QoS.

3.2.10 Auxiliary 5G Core Functions

Figure 21 presents other network functions as introduced in 3GPP Release 15.

Name of network function	Primary functionality
Short Message Service Function (SMSF)	Provides necessary interworking functions to enable Short Message service between 5G UE and legacy SMS functionality in the network
Authentication Server Function (AUSF)	Authenticates the UE and provides related keying material for it
Security Edge Protection Proxy (SEPP)	Provides Message filtering and policing on inter-PLM control plane interfaces and hides network topology from other vendors as it can act as a single point of contact
Network Data Analytics Function (NWDAF)	Provides slice specific network data analytics to PCF and NSSF
Location Management Function (LMF)	Obtains location measurements equally from UE and Radio Access Network

Figure 21: Auxiliary 5G Core Network Functions [20]

3.3 5G RAN Architecture

The 5G radio access architecture consists of two basic elements as follows:

- Central Unit (CU) intends to handle relatively higher layers above the physical layer. The physical realization of CU could be either dedicated hardware or a radio-cloud type of implementation. The CU connects to the Distributed Unit (DU) and, on the other hand, to the 5G core network (Next-Generation Core (NGC), as shown in Figure 22.

- The location of DU is on the cell site with the antennas and the RF unit. It handles especially time-critical processing, which does not tolerate too much delay to enable processing away from the cell site.

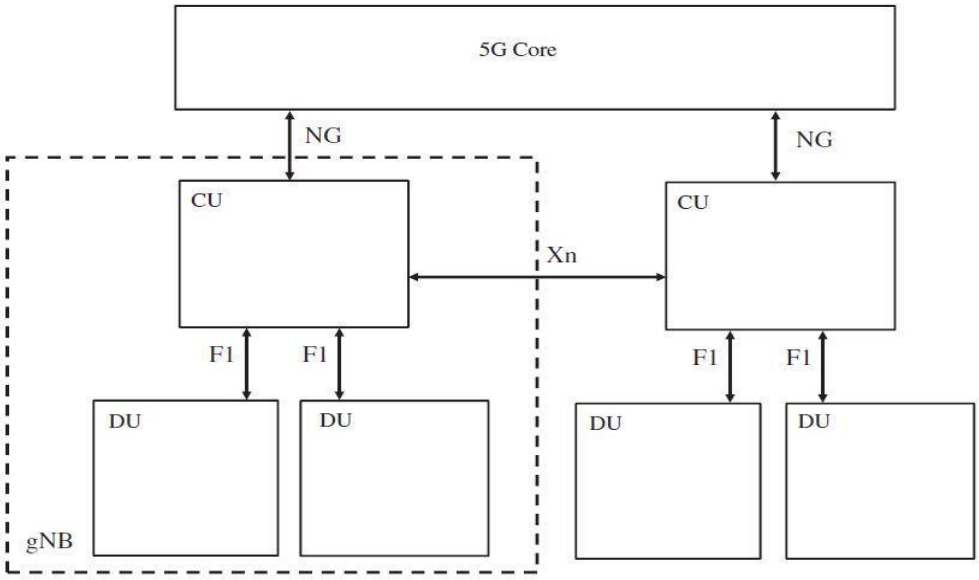

Figure 22: Overall 5G RAN Architecture [20]

The following interfaces are part of the 5G radio access architecture:

- The F1 interface connects the CU to one or more DUs. A single CU should be able to handle multiple DUs.
- The Xn-interface connects the different CU's. When considering the early deployment architecture with the LTE core (EPC), the interface between LTE eNB and 5G gNB is called the X2-interface.
- The E1 interface facilitates the separation of the CU's control and user plane processing for the CU-CP and CU-UP parts.

- The NG interface toward the 5G Core (5GC). It divides into the user plane (NG-U) and the control plane (NG-C). In the overall reference architecture in 3GPP, NG-U is called the N3 interface, and NG-C is called the N2 interface.

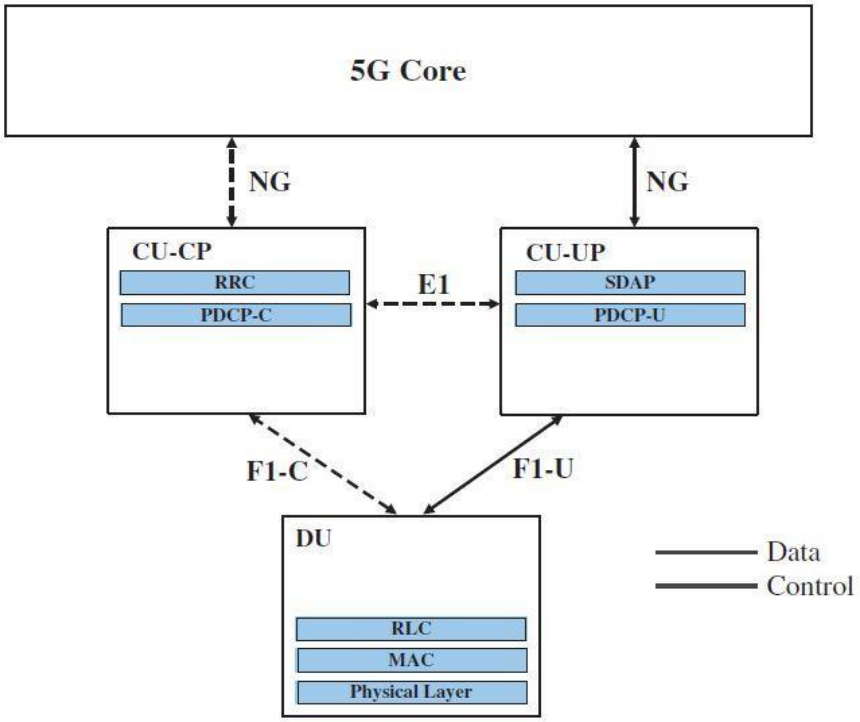

Figure 23: 5G RAN architecture with control and user plane separation for CU with the higher layer split [20]

The functional split shown in Figure 23 is also called the higher layer functional split. This is the only functional split defined in 3GPP in Release 15. This functional split is also suitable for the backhaul with some latency limitations since the retransmission is controlled locally in the DU. This also allows the CU to be relatively farther away than if the retransmissions controls are on the DU side.

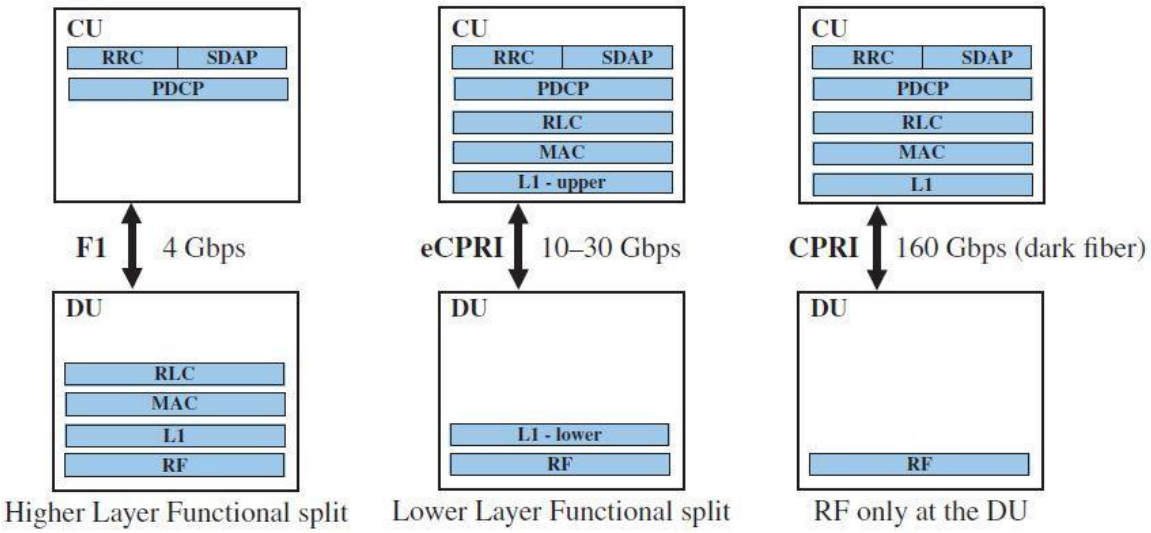

Figure 24: Impact on backhaul connectivity with different functional split options [20]

The lower layer functional split aspects have not been covered in 3GPP so far. The approach aimed for more functionality in the CU to facilitate more central handling of functionality and to have less functionality in the RF site and antennas and RF units. Some L1 functionality is necessary, in addition to the RF and antennas in the site, to avoid sending I/Q samples, as that would have resulted in massive requirements on the backhaul connectivity. The illustration in Figure 24 shows a factor of 40 difference in the data rates estimated when comparing sending bits after the Packet Data Convergence Protocol (PDCP) layer and when sending I/Q samples for the RF only in the site solution. Considering larger bandwidths than 100MHz and a larger number of antennas would result in an even bigger difference compared to RF only at the site solution. Moving MAC/RLC to the CU naturally results in higher data rates, as retransmission handling has moved to the CU. Thus, some packets end up being sent multiple times through the F1 interface in addition to the extra control signaling that becomes necessary. A lower layer split with enhanced CPRI (eCPRI) shown in Figure 24 is being considered in the O-RAN Alliance organization outside 3GPP, and the interface has also been defined in the Common Public Radio Interface (CPRI) under the name eCPRI, corresponding to the lower layer split options as studied in 3GPP. The key difference with the eCPRI interface is that, unlike the earlier CPRI interface versions, eCPRI is packet-based and thus facilitates options other than dark fiber, such as Ethernet. The term F2-interface referring to the eCPRI interface has been used outside 3GPP to refer to the interface between the DU (with RF and lower L1) and the

CU (upper L1 and other higher layer functions), but there are no studies on eCPRI/F2 interface by the 3GPP specifications.

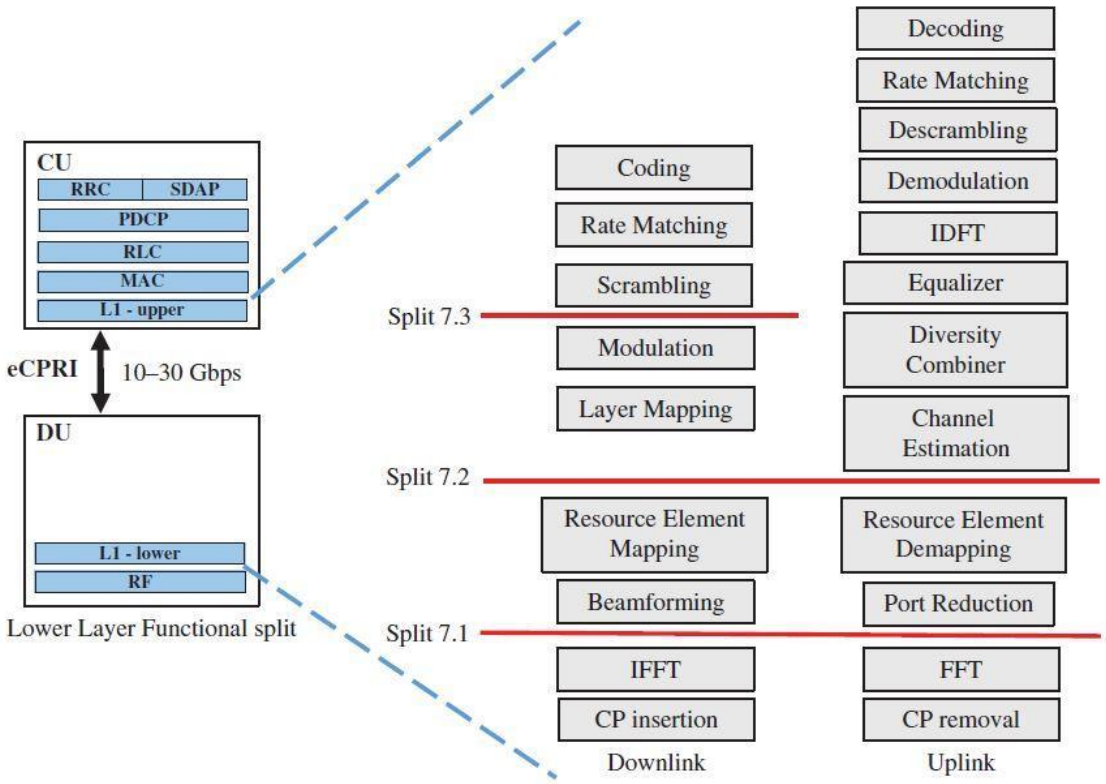

Figure 25: Lower layer functional slit alternatives [20]

In CPRI, multiple options are available for the lower layer functional split with eCPRI were defined and studied in 3GPP. The options include the following alternatives:

- Split 7.1, with the split in the downlink direction between the beamforming and Inverse Fast Fourier Transform (FFT) block, and in the uplink between FFT and beamforming.
- Split 7.2 with placement between the layer mapping and resource element mapping in the downlink and FFT and port reduction in the uplink direction.
- Split 7.3 is for the downlink only, between scrambling and modulation functionality.

There is an impact on the resulting data rates for the fronthaul interface as addressed, but clearly, the needed transmission capacity is always largely reduced compared to sending pure RF samples to the cell site. Between the different options, there is always the trade-off between how easy the network upgrades are to which operations are easier to perform closer to the RF

than further away. Work done in the O-RAN Alliance has taken split 7.2 as the basis for the work on the open fronthaul interface (Figure 25).

3.3.1 NG-Interface

The NG interface connects the gNB (5G-RAN) to the 5G Core on IP transport based on GTP-U and UDP. Similar to the S1-interface between LTE and the EPC core, the NG interface define to be an open multi-vendor interface. The NG interface has been designed in an access-agnostic way to facilitate convergence toward and use the 5G core for other access technologies. This would require the end devices to use 3GPP-defined NAS signaling for connection setup/registration.

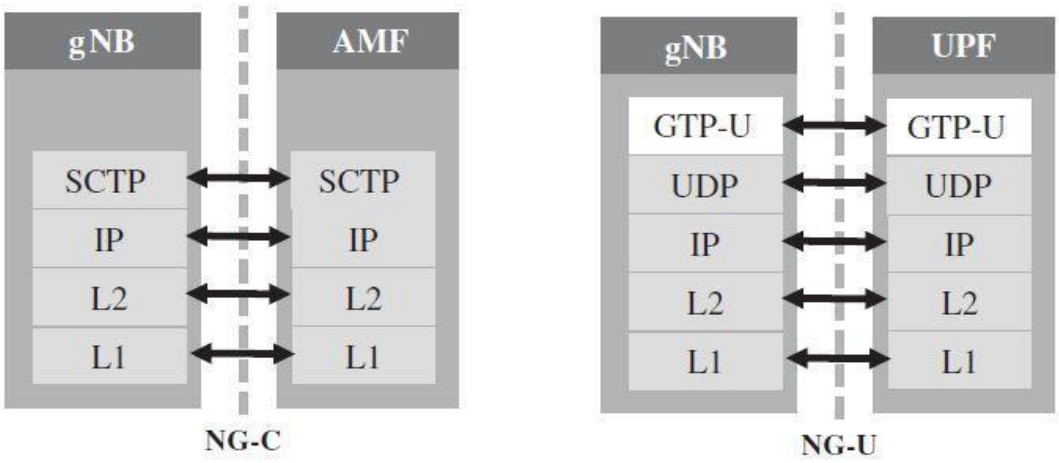

Figure 26: NG-interface control and user plane protocol stacks [20]

The NG-interface protocol stacks for the control and user planes are shown in Figure 26. The control plane uses the Stream Control Transmission Protocol (SCTP) on top of an IP connection. While the user plane part transports the user data, the key functions of the control plane part of the NG interface are as follows:

- NG interface management
- UE context and mobility management
- Transport of NAS messages
- Paging
- PDU session management

3.3.2 Xn-Interface

The Xn-interface connects gNBs to another gNB (or to eNB). We use the term Xn-interface when operating with the NG-core (5G-core). When architecture option three uses the 4G EPC core, the interface is called the X2-interface. The Xn-interface protocol stack is in Figure 27. As with the X2-interface in LTE networks, the Xn-interface enables operation as a multi-vendor interface. Figure 27 shows the Xn-protocol split. From the dimensioning point of view, the Xn-interface differs from the LTE-based X2 interface as the user data is routed via Xn/X2 more continuously than in LTE. The use of this interface in LTE is only for temporary packet forwarding.

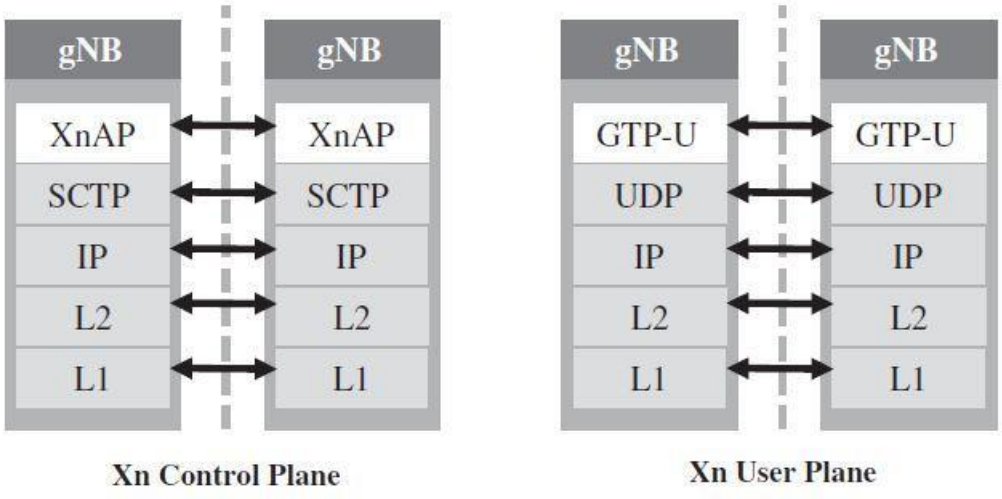

Figure 27: Xn control and user plane protocols [20]

For the user plane side, the key functionality is data forwarding and flow control, while on the control plane side of the Xn-interface key functionalities are as follows:

- Xn-interface management
- UE mobility management, including context transfer and RAN paging
- Dual connectivity

3.3.3 E1-Interface

The E1 interface connects the CU functionality's user plane and control plane parts. The E1 interface is in the June 2018 version of the specifications, together with the standalone 5G option and the 5G core. The E1 interface protocol stack is in Figure 28.

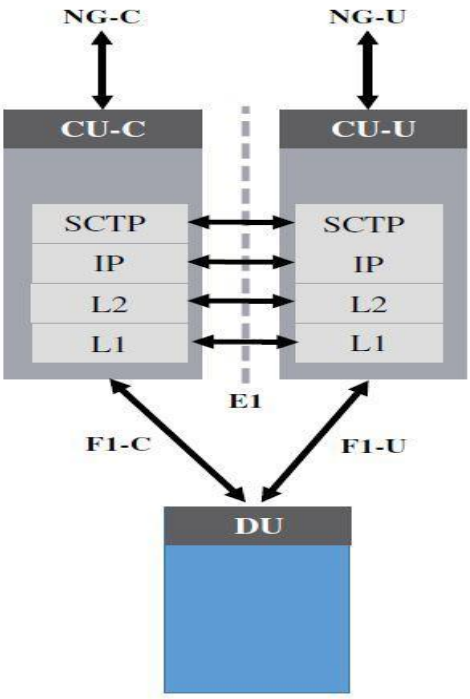

Figure 28: E1-interface protocol stack [20]

3.3.4 F1-Interface

The F1 interface connects the CU and DU. In Release 15, the supported functional split is the higher layer functional split, as presented earlier in this chapter. The F1-interface protocol stack is in Figure 29 [20].

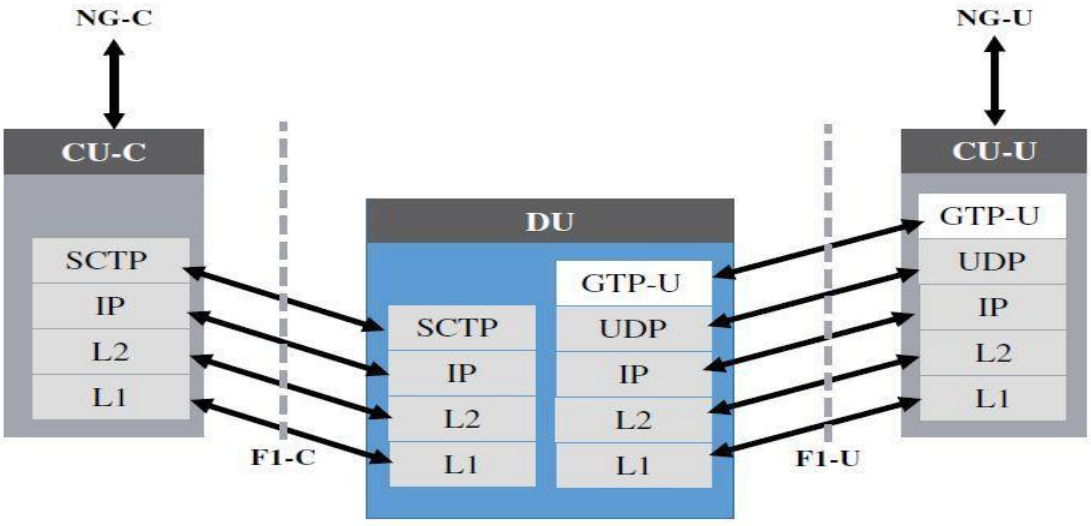

Figure 29: F1-interface protocol stack [20]

3.4 Virtualizing the RAN

But the task of virtualizing the RAN and the edge of the network is more excruciating but necessary to fulfill 5G deployments. As we distribute services with performance requirements end-to-end, the capacity to manage resources in the RAN in real-time becomes very important.

The virtualization of 5G networks has focused on the crucial concept of network slices. It is a virtual network explicitly dedicated to a specific customer or service. The network resources supporting the services are configured and employed exclusively to fulfill the performance profile needed. By logically demarcating the functions in the gNB and then implementing them as virtual software, the 5G RAN can uphold specific network slice requirements and the wide range of demands of each service simultaneously.

3.5 Optimizing the cost of 5G RAN

Virtualizing the RAN for 5G warrants new architecture and technologies, increasing costs. But the new architectures and technologies also provide opportunities to manage and optimize these costs.

3.6 The O-RAN Alliance

Formed because of the merger of the xRAN Forum and the C-RAN Alliance in 2018, the O-RAN Alliance is looking for an open, standards-based consensus for virtualizing the RAN. They want to build a more cost-effective and agile RAN through open interfaces while employing intelligence and deep learning techniques to automate operations.

The O-RAN Alliance lays down a design for a new architecture based on continuous data collection along with deep learning techniques. The RAN Intelligent Controller (RIC) is a primary component in the new architecture and is accountable for optimizing and managing the deployment and use of RAN resources.

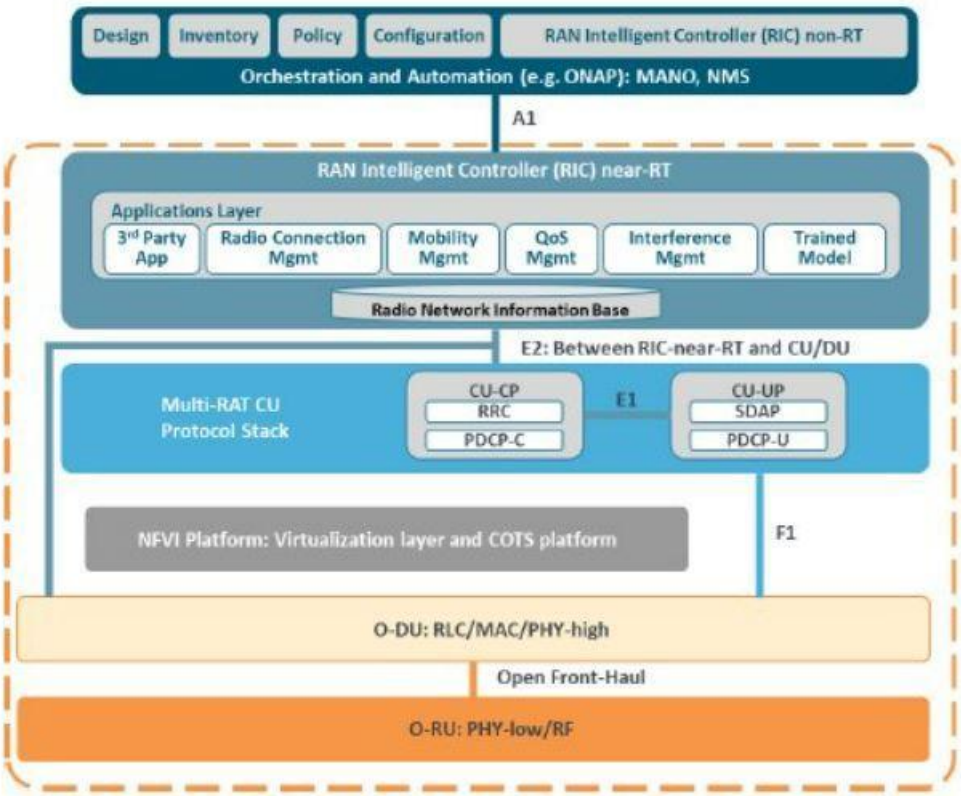

Figure 30: O-RAN Architecture [26]

The O-RAN Alliance focuses on opening the front-haul interface between the O-DU and O-RU, which is currently not standardized. This is critical to enabling multi-vendor O-DU and O-RU deployments and network sharing [26].

4 Chapter 4: Internet of Things (IoT)

The Internet of Things envisions a vast network of interlinked computing devices, including digital and mechanical devices. The connection between these devices power through the internet, where each device carries unique identifiers to communicate with other devices. Designing these “things” or objects is to engage in communications without manual intervention. The linked network, where automation systems and connected devices work together, provides a modern outlook for a futuristic world where data could be collected and analyzed to perform the right action [23].

When machines and objects can be sensed and controlled remotely across a network, seamless integration between the physical world and computers is possible. This allows for upgrades in automation, efficiency, accuracy, and the enablement of advanced applications [27].

4.1 IoT Architecture

One-size-fits-all architecture does not work for IoT projects due to the ever-growing nature of IoT devices and the wide variety of sensors. However, some basic building elements will almost be the same from project to project.

Firstly, you will have to build with scalability in mind. The quantity of data that you will collect over time will amount to enormous proportions, and you will end up needing a platform that can account for this in the long run.

One will also need to ensure high availability at any given time. System failures could lead to loss of business in the best case or, in the worst cases, have fatal consequences. In the end, you will need a flexible system to provide for quick and frequent changes. As your architecture changes as your business evolve, you will need to iterate quickly without breaking the existing architecture.

4.1.1 Three Layer IoT Architecture

While it is correct that no two IoT projects are exactly the same, the main layers have always stayed the same. Starting from when the first research on IoT was performed, the three-layer architecture has been the primary model for IoT applications. The three layers are Perception (or Devices), Network, and Application.

- **Perception Layer:** This layer has sensors and the data they provide. The data could be consolidated from any given quantity of sensors on the connected device.
- **Network Layer:** The network layer accounts for how high volumes of data flow throughout the application. This layer connects the various devices and channels the data to the appropriate backend services.
- **Application Layer:** This is the layer that the user interacts. This could be a dashboard showing the status of the devices which are part of a system or an application to control a device in a smart-home ecosystem.

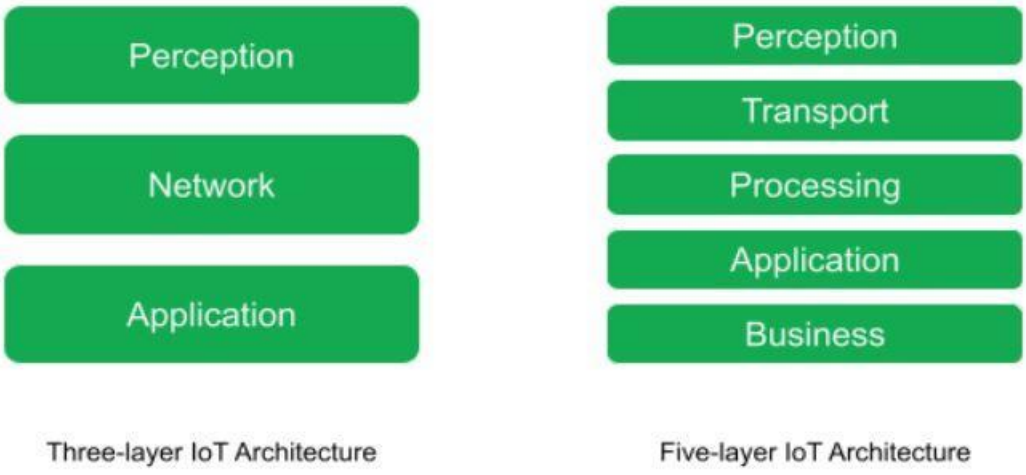

Figure 31: IoT Architecture [28]

4.1.2 Five Layer IoT Architecture

The traditional three-layer architecture is a great way to dissect an IoT project, but it is somewhat limited in scope. Hence, the development of alternate architecture models with new or modified layers. One popular architecture, known as the five-layer architecture, has three

new layers - Transport layer, Processing layer, and Business layer in addition to the original Perception and Application layers.

- **Transport:** The transport layer deals with the methods and modes of data transfer between the sensors and the Processing layer through different networks.
- **Processing:** The Processing layer saves, peruses, and pre-processes the data received from the Transport layer. For low latency purposes, this is often implemented on the edge of the cloud. This layer is referred to as the Middleware layer.
- **Business:** Business Layer presents data to the stakeholders based on what is found and consumed. This layer is placed higher than the application layer, and the stakeholders decide its functionality.

4.1.3 Four-Stage Approach to IoT Architecture

This is yet another way to describe the IoT architecture with emphasis placed on edge computing compared to others.

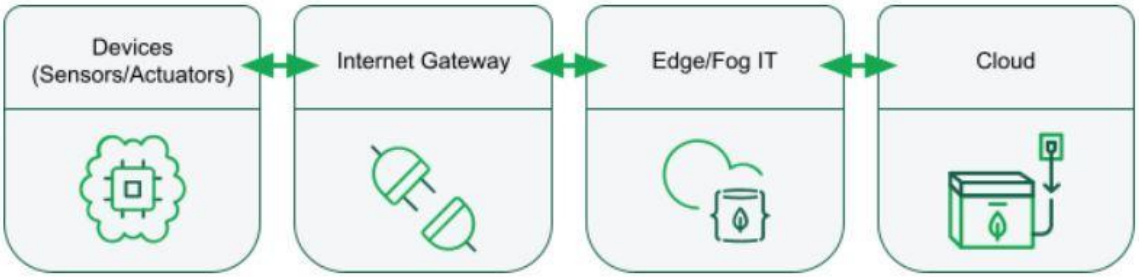

Figure 32: Building Blocks of IoT [28]

- **Devices:** This is the physical devices stage. The devices could be either sensors or actuators of the perception layer. The data gathered here is converted to digital form and transmitted to the next stage, i.e., the Internet Gateway stage.
- **Internet Gateways:** This stage pre-processes the raw data and transmits the data to the cloud. This stage can either be implemented onto the device or as a stand-alone device that communicates with the sensors and disseminates the data to the cloud.
- **Edge or Fog Computing:** The primary role of this layer is to process the data received by the cloud to act on time-critical operations as quickly as possible. Processing will

usually be on data received at the latest. In some cases, this layer may also pre-process the data to limit the size of the data reaching the cloud.

- **Cloud or Data Center:** This is the final stage where the stored data is processed, sent to applications, etc. We can perform deep analysis and other intensive operations like machine learning training in this layer.

4.1.4 IoT Architecture in Business

IoT use cases are diverse and can take various shapes. To understand the many layers, we can look at an example. Commercial airlines possess many planes, and each has a variety of sensors.

The Perception layer consists of all those sensors in the plane. These will indicate the present state of the aircraft, plus the data about the current flight. The sensors will survey things such as the altitude, the airspeed, the position, and the vertical speed. The rest of the sensors are collecting data to ensure that the plane's integrity is not compromised, monitoring feedback such as abnormalities in the engines.

This information, coming from multiple sensors from varied manufacturers, will be sent to a central unit on the aircraft as part of the Network layer. This data will be further converted into a standard format and pre-processed there. If any critical thing happens, such as an engine failure, actuators will trigger immediately instead of waiting for an entire round trip to the cloud. As soon as the plane has internet connectivity, the data will be sent to the cloud and moved to the Application layer.

As the data makes its way to the cloud, it can be processed and analyzed in the Application layer. Here, dashboards produce to verify anomalies, flag issues for maintenance, and provide business insights for the airline operator. The operator can make safer decisions and automate some tasks to improve flight security [28].

4.2 Types of IoT Network

Connectivity is one of the primary elements in IoT networking, given that no project can succeed with an unreliable connection between devices, sensors, and your IoT platform. There are four general types of IoT networks, commonly in deployment.

- **Cellular Networks**

Cellular networks use similar mobile networks as smartphones to communicate IoT devices. Although this type of network covers a vast region, cellular connectivity is often unavailable in places where IoT sensors are deployed periodically, such as the depths of a mine shaft or inside utility closets. Additionally, cellular-connected IoT devices need a lot more power and energy compared to certain other kinds of wireless networks.

The LTE-M and Narrowband IoT (NB-IoT) cellular wireless IoT protocols are rising in usage, with NB-IoT using lower amounts of battery power and costing less than LTE-M, but currently trailing in terms of coverage area.

- **Local and Personal Area Networks (LAN / PAN)**

LAN/PAN spans moderately short distances and provides a cheap solution but can be undependable in data transfer. Wi-Fi and Bluetooth are primary technologies used for IoT connectivity in networks of this kind. Wi-Fi can be employed for applications that run in a local set-up close to an access point or in a distributed setting with varied access points integrated into a bigger network.

Bluetooth Low Energy (BLE) is an energy-efficient wireless network protocol, but it has lower transmission rates and is more restricted in the amount of data that it's capable of transmitting.

- **Low Power Wide Area Networks (LPWAN)**

Created in response to cellular connectivity's initial challenges, LPWAN provides for a longer range than Wi-Fi and Bluetooth but utilizes less power than cellular. LoRaWAN (long-range wireless area network) is a regularly used IoT network protocol in this category, with low power needs and relatively cheap chipsets.

- **Mesh Networks**

All of the sensor nodes in a mesh network coordinate to distribute data amongst one another to reach the gateway. Such networks are very short range and may demand additional sensors throughout a building or repeaters to expand the area of coverage. IoT applications that need instant messaging (such as smart lighting) can lead to high power consumption. Mesh networks are robust and easy to install, making them a convenient choice for IoT deployment inside buildings [29].

4.3 IoT with 5G

5G is propelling change in the Internet of Things (IoT). It's a robust enabling technology meant for a new generation of use cases that will end up leveraging edge computing to make IoT more efficient and effective.

In multiple ways, the narrative of 5G is the interplay between two relentless forces: the increase in highly reliable, high-bandwidth communications and the quick spread of available computing power in the network. The computing power does not just end at the network, though. End-point devices that link to the network are also shaping to be smarter and more powerful.

The rising dynamic and powerful computational environment that is taking shape as telcos start to redesign their networks for 5G will exponentiate the uptake of IoT applications and services throughout the industry. We can expect that 5G will warrant new use cases in visual inspection and remote monitoring, autonomous operations in large-scale remote conditions such as mines, connected vehicles, and more.

The fast-expanding array of computing options needs a much more flexible approach to building and deploying applications and AI models that can make use of the most cost-efficient computing resources. Such AI models now even run-on edge devices that connect to the network edge, providing more secure and efficient data processing.

Before we dig deeper into the latest wave of IoT applications, let's talk about the advantages of the potent one-two punch of edge computing enabled by 5G, which delivers lower latency and higher bandwidth compared to today's 4G and LTE networks. The increased bandwidth

provides more data to relay during any given period, which is a prominently significant gain for demanding payloads like video. Due to this increment in capacity, we can expect the usage of video sensors and video data streaming to greatly rise, spurring broader usage of unmanned (robotic and drone) inspection and monitoring applications.

The decreased latency of 5G—the time required for data to reach its target destination—will let us respond to data faster. As the average latency of a 4G connection is 50–100 milliseconds, the similar figure for a 5G connection could be 10ms or less.

The lower latency and higher bandwidth of 5G will, in turn, speeds up the adoption of applications that employ edge computing, where the computational work is performed nearby to where created data and actions are being taken [30].

4.3.1 IoT Use Cases

It can convince everybody that 5G will be the next revolution. Some of the applications of the 5G-enabled IoT communication environment are,

- **Remote surgery:** One of the tremendous advantages of 5G is its low latency feature. It has a short time lag between a device pinging the network and getting the response, whereas it was the problem with 4G LTE. Because of the 5G network characteristics, a surgeon can now perform remote surgery with no physical presence in the same operation theatre. For example, the doctors at King's College London have demonstrated the surgery procedure where they used a dummy patient with the help of a "virtual reality headset" and "special glove". Through that, they have performed remote surgery with the help of a remote robotic arm.

- **Security and surveillance:** Surveillance and analytics is another application that will be very successful with 5G connectivity. Owing To the incremental threats to public safety and security in recent years, various governments and security agencies are installing public surveillance and security systems. The public video surveillance systems still depend on wired networks. However, adoption for wireless communications such as Wi-Fi or the 5G system is also gaining popularity due to the easy and fast setup with low cost. Closed-circuit television (CCTV) systems, cameras installed in vehicles (e.g., police cars), public transport, and surveillance drones are getting popular. Adopting a 5G communication system boosts the performance required

for sophisticated real-time video analysis and the deployment of massive cameras in the targeted regions.

- **Transforming Healthcare:** Most of the time, when somebody has some illness, and he/she needs some medical treatment and attention, they travel to a doctor's clinic or hospital. But it is very difficult for the people living in the rural area where that kind of medical facility is not available. Moreover, traveling in illness is challenging and time-consuming. However, with the advancement of information and communications technology (ICT), the latest tools and technologies, for example, telehealth and remote home monitoring systems, are available through which care can be received within the comfort of the homes. The remote home health monitoring system consists of health monitoring devices (i.e., implantable or wearable health devices) and other video & imaging facilities in which the communication technologies such as 5G and IoT can be utilized. These devices send and transmit the medical information of a patient (user) to a central authority (i.e., cloud server) through which a health expert (i.e., a doctor) can also access the medical data of the patient. Doctors can suggest medicine (treatment) prescriptions after a short video call remotely. Such kind of remote monitoring is equipped with sophisticated imaging equipment and produces an enormous amount of health sensing data which causes additional strain on the networks. This often increases congestion and slows down the data transfer speeds, especially in a large healthcare system that may interface with a huge number of patients in a single day. Therefore, a 5G mobile communication system that has very high data transfer speed along with low latency will be a very helpful for such a healthcare system.

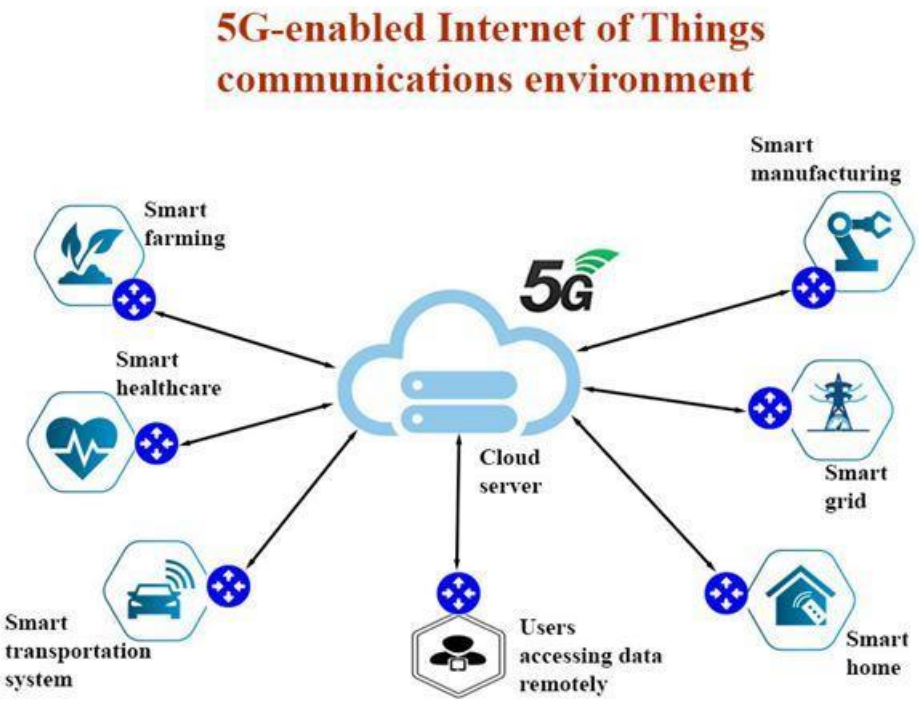

Figure 33: Network Model of 5G-enabled IoT Environment [31]

4.4 Security Challenges in IoT

- **Device Authorization and Authentication:** With multiple nodes of entries in an IoT system, authorization becomes a crucial factor. To evolve identities of devices to allow secure access to gateways and other IoT applications is crucial. Many devices have to suffer data security breaches due to weak authorization systems. A two-factor authentication model in IoT systems can sort out these issues with device authorization solutions that verify which devices can get access to the system.

- **Software Updates:** To update the firmware functioning on devices and gateways is a tough part too. Understanding the updates that are needed and applying them safely to the devices that are communicating with different networks is essential. For a few legacy devices, it is frequent that updates are unavailable and are unsupported by the manufacturers any longer. They should be tracked down and retired immediately.

- **Data Privacy:** Ensuring that the data is stored and processed safely through various networks is critical for privacy concerns. Applying data privacy into IoT systems includes editing and filtering out sensitive data to distinguish it from identifiable data. Systems with machine learning are in high demand, wherein the data is stored securely on the cloud. This ensures robust data privacy as an algorithm comes to the data versus traditional methods where data goes to the algorithm.

- **Limited Capacity of Devices:** Many IoT devices have limitations in the capabilities of storage, processing, and memory. These devices are incapable of providing security approaches such as encryption and many other complex processing functions to transfer data safely from one platform to another in real-time. It makes these devices more vulnerable to security attacks.

- **Vulnerability of IoT devices:** Connected IoT devices make it a complicated system where it becomes tough to analyze the consequences of the vulnerability that the devices are subjected to and manage the impact. The primary challenge is identifying the influenced devices within time, the services accessed, and the regions impacted so immediate measures can be taken to resolve the issue, dampening further impact.

Taking on a multi-layer security approach for IoT systems and devices is crucial to ensure security and easy transmission of critical data. It provides for managing services, devices, and apps easily. A compromising attitude towards security in IoT devices can cause large amounts of damage, such as data breaches, loss of data, and even system failure. Implementation of security features within the system that works as a default through the entire lifecycle of the system enhances secure function and maintains integrity [32].

5 Chapter 5: 5G Rogue Base Station (gNB)

5.1 5G Security Concerns

As new technology comes to the fore, it is justified to ask if it may pave the way to novel ways of collecting and processing personal data. Where 5G represents a significant shift in the use of mobile networks, existing data privacy regimes that are technology-neutral, not itself that address a whole wide range of benefits of data collected through apps, mobile device operating systems, social media, websites, and network operators and are likely to be sufficient to address the use of new 5G capabilities within the online ecosystem [33]. Therefore, it is crucial to shield them from potential attacks.

5G cyber security requires some significant improvements to circumvent the increasing risks of hacking. Some security worries arise from the network itself, while others involve the devices connecting to 5G. Nevertheless, both aspects put governments, consumers, and businesses at risk. When it comes to 5g and cybersecurity, some of the main concerns are as follows:

- **Decentralized security:** Pre 5G networks had fewer hardware traffic points of contact, making it simpler to do security checks and upkeep. 5G's dynamic software-based systems have far greater traffic routing points. To be fully secure, these need to be continually monitored. As this may prove difficult, any unsecured areas might put at risk other parts of the network.
- **More bandwidth will strain current security monitoring:** Whereas existing networks are limited in speed and capacity, this has aided providers in monitoring security in real-time. Hence, the benefits of an expanded 5G network might compromise cybersecurity. The elevated speed and volume will challenge security teams to come up with new methods for abetting threats.
- **Many IoT devices are manufactured with a lack of security:** All manufacturers are not prioritizing cybersecurity, as seen with multiple low-end smart devices. 5G implies more utility and potential for IoT. As more and more devices are encouraged to connect, billions of devices with varied security mean billions of potential breach points. Smart

TVs, speakers, door locks, refrigerators, and even minor devices like a thermostat can be a network weakness. A dearth of security standards for IoT devices implies network breaches and hacking might happen rampantly.

- **Lack of encryption early in the connection process:** Reveals device information that can be utilized for device-specific IoT targeted attacks. This information aids hackers in knowing which devices are connected to the network. Info such as operating system and device type (smartphone, vehicle modem, etc.) can aid hackers in planning their attacks precisely.

Cybersecurity vulnerabilities can result in a wide variety of attacks. Some of the known cyber threats include:

- **Distributed Denial-of-Service** (DDoS) overloads a network or website to make it go offline.
- **Botnet attacks** controls a network of connected devices to puppeteer a massive cyberattack.
- **Man-in-the-Middle (MiTM) attacks** quietly intercept and modify communications between two parties.
- **Location tracking and call interception** can be performed if someone knows even a small amount about broadcast paging protocols [34].

5.2 Rogue Base Station

A crucial attack vector is the radio interface between the base stations of a mobile network and the mobile phone. Radio devices utilizing this attack vector by impersonating a mobile network's base station towards a mobile phone are often called false base stations. These devices come in many flavors, a few of which also impersonate a mobile phone towards the mobile network. Attacks initiated using rogue base stations broadly fall under categories: privacy compromise of mobile phone usage, DoS on the mobile network, denial of service (DoS) on mobile phones, and frauds. However, the efficacy of these attacks largely varies between different generations of mobile networks as each latest generation becomes more robust than earlier. Despite this, due to multiple reasons like legacy networks that have overlived their time and interworking between a wide plethora of networks, the mobile network industry is only partially protected against all the types of attacks originating from rogue base

stations. Therefore, detection and subsequent protection against rogue base stations are essential topics for the mobile network industry and society.

Many approaches for detecting false base stations have been proposed and prototyped over the past few years. Most of them implement a data collection capability in the mobile phone, which either analyzes the data collected in the mobile phone or sends the aggregated data to a central server for analysis. The first part is User Equipment (UE)-based, and the latter is crowd-sourced detectors. What is usual for these types is that they decide whether a false base station is present by using the view of the network from the perspective of a mobile phone. This view is intrinsically limited compared to the view obtained from a network's perspective.

A mobile network knows not just the global state of the system, whereas a mobile phone only holds the knowledge of its local state, but the mobile network also has more information of a mobile phone's state than the mobile phone itself. The phone has sufficient knowledge of its state to operate when being asked by the network to take certain kinds of actions. One more drawback of these detection systems is that they require modified mobile phones. This means that end-users must download and then run a special application to assimilate and analyze measurements or detail the measurements to a central server on the internet for analysis. This dramatically reduces the number of measurements accessible for analysis. A few false base station detectors are network-based that rely on information collected by the mobile network. However, they either require a pre-existing network monitoring infrastructure or focus on a single 3GPP Radio Access Technologies (RAT).

5.2.1 BACKGROUND

Mobile networks have upgraded from 2G in 1991 to 3G in 2001 to 4G in 2008. The latest generation is the 5G in 2018, with the first commercial deployments of 5G occurring in 2020. A basic overview of the different generations of mobile networks is in Figure 34. On a high level, a mobile network consists of UEs, which refer to mobile phones that we use, Radio Access Network (RAN), which are network entities providing wireless radio access to the UEs, and Core Network (CN), which are the set of network functions that among other things do subscription handling, session and mobility management, and packet routing.

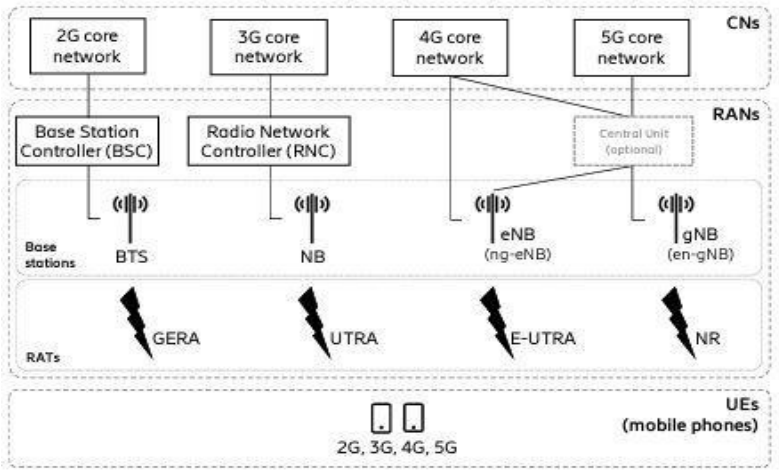

Figure 34: Simplified overview of mobile networks [35]

Historically, different generations of RAN offer radio access via a radio access technology (RAT) specific to that generation, i.e.,

- 2G RAT: GSM EDGE Radio Access (GERA)
- 3G RAT: Universal Terrestrial Radio Access (UTRA)
- 4G RAT: Evolved-UTRA (E-UTRA)

In 5G, however, two types of RAT can co-exist. One is 4G RAT (E-UTRA), and another is a new one, 5GRAT: New Radio (NR). The RAN uses base stations to offer these RATs. They are known as Base Transceiver Station (BTS) in 2G, NodeB (NB) in 3G, Evolved NB (eNB) in 4G, and next-Generation NB (gNB) in 5G. There also exists ng-eNB (Next-Generation eNB) that connects to a 5G core network and en-gNB (EUTRA New Radio gNB) that connects to a 4G core network. These base stations support one or more cells, a so-called cell being the smallest coverage area in which the base stations serve the UEs.

5.2.2 False Base Station Attacks

False base station is a broad name for a radio device that sets out to impersonate a legitimate base station. Although the name says “base station”, its attack capabilities have outgrown to also impersonate UEs towards the mobile network. It is widely known by other names such as IMSI catcher, Stingray, rogue base station, or cell-site simulator. A logical illustration of false base station attacks is shown in Figure 35.

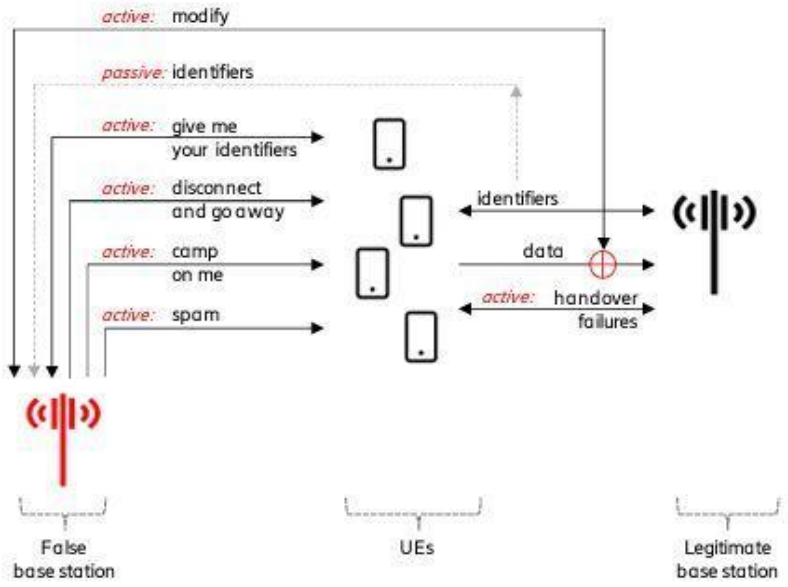

Figure 35: Logical illustration of False Base Station attacks [35]

One of the main attacks relates to the privacy of users, in which an attacker either passively eavesdrops on users' identifiers from the radio interface or actively obtains them by communicating with the UEs. The attacker then uses those identifiers to identify or track the users. The attacker might also try to fingerprint user traffic. Under stringent assumptions, a resourceful attacker may also exploit implementation flaws or vulnerabilities in application layer protocols like Domain Name System (DNS) and Internet Control Message Protocol (ICMP) by altering carefully chosen parts of the data in the radio interface. Another set of attacks relates to denial of service (DoS) on UEs and mobile networks. The attacker may use specific messages that the UEs and the network accept without authentication. The attacker may also create favorable radio conditions so that the UEs keep camping on the false base station, thereby cutting off all incoming communications from legitimate base stations. Radio conditions created by the attacker may also trigger certain events in the legitimate network, like handover failures. Because of these events, some implementations in the network may take disruptive steps such as barring even the legitimate base stations, thereby triggering service disruption.

At a basic level, an IMSI catcher has two main parts: a radio frontend for receiving and sending radio waves and a network backend for simulating a cellular core network. Today, anyone with a software-defined radio (SDR) and a computing device running an open-source base station

program (like OpenBTS) can effectively operate an IMSI catcher [36]. Therefore, the increased incentive for attackers due to ever-growing connectivity and increased feasibility of deploying a false base station are the main reasons that false base station attacks are more important now than ever. It has rightly got more attention on all fronts, media, hackers conferences, academia, standardization bodies, governments, law enforcement agencies, vendors, and operators.

5.2.3 Existing Countermeasures

While 2G remains most vulnerable among all generations, several attacks have become impractical, or the difficulty level for attackers has significantly increased with the newer generation of mobile networks. To Design the 5G with significant enhancements in terms of privacy and security against false base stations. E.g., encryption of permanent identifiers makes it more difficult for rogue base stations to track users by eavesdropping on this identifier over the radio interface; and integrity protection of user traffic enables detection of data alteration by false base stations in the radio interface. Nevertheless, eradicating attacks from a rogue base station is not completed yet. Some attacks are possible simply because newer networks require interworking with 2G networks. Also, attacks may be possible merely because of not updating security features in the devices. 3GPP is currently assessing if and type of further enhancements in 5G in terms of new protection and detection features [35].

5G standalone (SA) deployments promise a countermeasure against IMSI-Catchers: The Subscription Permanent Identifier (SUPI) (which is equivalent to the IMSI) should be encrypted) with the network operator's public key, yielding the so-called Subscription Concealed Identifier (SUCI). Only the operator can decrypt the identifier, and thus attackers cannot derive the permanent identity anymore. Furthermore, the user's device generates a fresh SUCI for every transmission. Thereby, users should be untraceable [37].

5.2.3.1 SUCI

In 5G networks, the UE holds the permanent identifier and key inside the USIM (Universal Subscriber Identity Module). These are the credentials utilized by the UE to set up mutual authentication with the 5G network. By using this process, three identifiers become critical -

the permanent identifier SUPI (4G: IMSI), the hidden/concealed identifier SUCI and the temporary identifier (5G: GUTI).

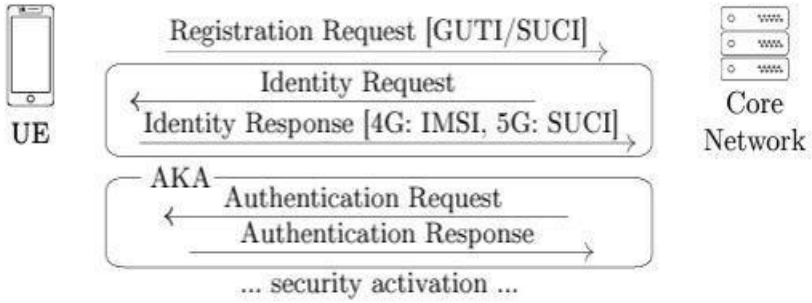

Figure 36: Authentication and Key Agreement (AKA) Process [37]

Figure 36 describes the basic message exchange for user registration and authentication. The initial stages need SUCI to be transmitted. But, if there is no established temporary identity, then the permanent identity is requested. This is similar to websites using cookies to keep a user logged in instead of getting them to authenticate frequently. This process should be a two-way street, with the network authenticating the user and the users authenticating the network; AKA messages are UNPROTECTED; encryption happens only AFTER agreeing on the session key.

5.2.3.2 SUCI vs. SUPI

In 5G networks, permanent identifiers prevent sending using the operator's public key; this is present in the USIM. Permanent SUPI is encrypted using this public key prior to transmission (aka SUCI). As the encrypted key is with the operator's public key, only the operator can read the SUPI to reveal information about the subscriber's identity. SUCI is regenerated before every usage to prevent linkage of SUCI (aka perfect forward secrecy), preventing the attacker from identifying if the SUCI refers to the same user (even if the user connects multiple times).

Since the SUCI varies, it gives the impression that various users connect to the network. SUPI concealment is an OPTIONAL feature, which has to be configured by the operator.

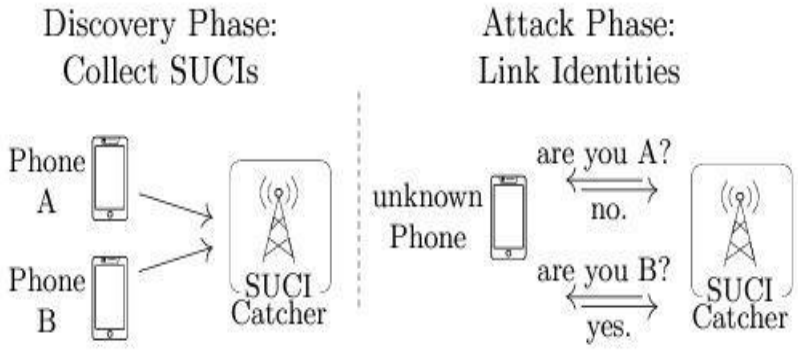

Figure 37: Two phases of SUCI-catcher attack [37]

5.2.3.3 SUCI Catcher Attack

- **Discovery Phase:**

  With the help of AKA linkability, the attack targets the UE revealing its own identity. This will only work when the attacker learns of any SUCI previously used by the target UE. This is done by sniffing the traffic for SUCI messages with the full knowledge of the location of the UE. Using the IMSI - the attacker can execute the encryption (either EC25519 or secp256r1) under the assumption that the operator's public key is available. The IMSI can be discovered using either a downgrade/SS7/mobile app-based attack.

- **Attack Phase**

  After completing the discovery phase, the attacker now has the SUCI of the target UE. When an unknown UE connects to the catcher, the attacker tries to find out if this unknown UE is identical to the subscriber. Using the obtained SUCI, the attacker makes a Registration Request (since the request requires no authentication to execute). This request will only respond to the Authentication Request, which is then responded by the UE associated.

However, there may be two outcomes with the Authentication Request. First, the unknown UE is actually the searched-for-UE authenticated successfully and responds with Authentication Response or Authentication Failure, with the reason Sync Failure (sequence number SQN needs to be synchronized). Secondly, if the searched-for-UE isn't the one, UE sends

Authentication Failure with the reason of MAC Failure to the SUCI catcher. As a way of handling the Sync failure, the attack precedes a reset stage which performs the successful AKA between the UE and then networks BEFORE the actual probe. This also handles the resynchronization of the sync number to handle Sync Failure errors. While the method highlights attacks for one UE, it scales well when searching for multiple UE [38].

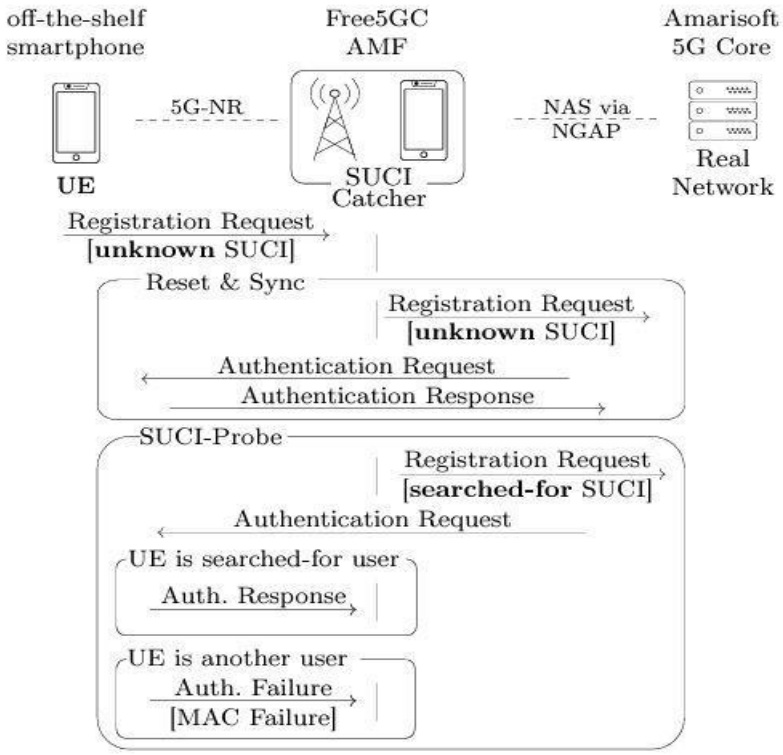

Figure 38: Attack Mechanism for SUCI Catchers [37]

5.2.4 Detection of IMSI Catchers

There are IMSI catcher detection apps procurable just for Android, but they need rooting of the device – which in itself is a compromise in security – to be able to access the cellular network messages availed from the diagnostic interface of the smartphone baseband. Sadly, detection is a mixed bag. As cellular standards vary greatly between countries and carriers and because very little is known for certain about how IMSI catchers work, there's no definitive list of heuristics that can be used here. Hence, each IMSI catcher detection app has its own list of indicators of IMSI catcher operation, like unexpected identity requests and removal of encryption from the cellular connection. Ending up with false positives is common, as

temporary equipment (for large events), testing equipment, and tower resuming tend to trigger user alerts.

More dependable hardware options are available for catching IMSI catchers, which is viable when protecting multiple smartphone users in a single site, for example, in a corporate headquarters or a military base. Usually, a setup like this involves a fixed, embedded system with sensor hardware and a cellular modem for perennially monitoring the broadcast signals of the nearby base stations, with a database that uploads data for analysis. As an IMSI catcher is detected, they send alerts to all of an organization's smartphone users.

Knowing that IMSI catchers make use of flaws inherent in cellular networks and are hard to detect, there's been a hard push by 3GPP, the organization in charge of specifying the 5G protocol, to drastically reduce the possibility of IMSI catchers from devices with this standard. Critically, designed 5G such that the IMSI (or another Subscription Permanent Identifier) is not disclosed in the clear when a mobile device is trying to establish a connection. Rather, 5G employs only a temporary paging identifier that needs to be refreshed after every use.

While this is a big leap towards privacy on cellular networks, some issues still mean that IMSI catchers will stick around a while.

- **Bugs:** As is usual with new protocols, security researchers are discovering scores of bugs in 5G, which includes a flaw in the Authentication and Key Agreement (AKA) protocol. As these are being quickly addressed, it is essential to remember that no standard is golden and that manufacturers of commercial IMSI catchers will certainly be leveraging these faults to develop 5G-specific models.

- **Poor carrier implementation:** Despite the 5G protocol being relatively secure, it is still in the hands of the carriers to implement it well. Some carriers have already mismanaged initial 5G rollouts in such a way that would let IMSI catchers change a device's stated category number in the connection process and hence operate as usual.

- **Downgrade attacks:** While carriers in the United States have the most deprecated 2G, it's still used throughout the world, implying that most phones are designed to work in a 2G network. Hence, downgrade attacks to 2G will be possible for the foreseeable future, even in non-2G environments.

5.2.4.1 Mitigation Steps

Despite the fight against IMSI catchers being largely unsuccessful, there are still a few strides that one (and any high-profile targets within any organization) can take to reduce personal and organizational risk:

- If your smartphone permits it, turn off 2G support. Doing so will significantly reduce the capabilities of IMSI catchers.
- When moving through chokepoints (like airports and border crossings) where there is generally a greater chance of IMSI catchers, switch off your smartphone or use an RF-shielding device, such as a Faraday bag. Neither of the options completely reduces RF emissions but can minimize them to a large degree.
- Use communication apps featuring end-to-end encryption that ensure that captured content cannot be deciphered easily by threat actors [36].

CONCLUSION

Despite the significant technology improvements from 2G networks to current LTE systems, cellular networks' overall architecture and functionality still contain strong ties to outdated legacy technologies. Certain simple features are now overdue for a systematic redesign that considers the current cyber-security landscape and the low-cost availability of leveraged tools to attack a mobile network. In the era of packet-switched traffic, redesign the systems accordingly to scale towards the massive connectivity goal of 5G systems.

This document summarizes 5G wireless network architecture with massive MIMO technology, network function virtualization (NFV) cloud, and device to device communications which helps to improve the performance requirements of 5G wireless cellular systems defined in terms of bandwidth, capacity, data rate, spectral efficiency, latency, MIMO and Cloud Technologies in general with radio access networks and SDN, energy efficiency, quality of service and short-range communication technologies, like Wi-Fi, Small cell, and mmWave communication technologies, which provides a promising future in terms of better quality and faster data rates for users and at the equivalent time reduces the pressure from the outside base stations and to reduce the vulnerabilities.

GLOSSARY

3GPP	3rd Generation Partnership Project
AMPS	Advanced Mobile Phone System
AuC	Authentication Centre
AMF	Access and Mobility Management Function
AKA	Authentication and Key Agreement
BSS	Base-Station Subsystem
BTS	Base Transceiver Station
BSC	Base Station Controller
BBU	Baseband Unit
BLE	Bluetooth Low Energy
CDMA	Code Division Multiple Access
CU	Centralized Unit
CN	Core Network
CS	Circuit-switched
CriC	Critical Communication
C-RAN	Centralized RAN or Cloud RAN
Cloud-RAN	Cloud-based Radio Access Network
CAPEX	Capital Expense
CPRI	Common Public Radio Interface
CUPS	Control-User Plane Separation
CSP	Communication Service Providers
CCTV	Closed-circuit Television
D2D	Device-to-Device
D-RAN	Distributed RAN
DNS	Domain Name System
DU	Distributed Unit
DOS	Denial of Service
DDOS	Distributed Denial-of-Service
EIR	Equipment Identity Register
EDGE	Enhanced Data rates for GSM Evolution

E-UTRA	Evolved Universal Terrestrial Radio Access
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
EPS	Evolved Packet System
EPC	Evolved Packet Core
eMBB	Enhanced Mobile Broadband
eNB	Evolved Node B
en-gNB	EUTRA New Radio gNB
FDMA	Frequency division multiple access
FDD	Frequency Division Duplexing
FC	Fog Computing
FFT	Fast Fourier Transform
GSM	Global System for Mobile Communications
GMSC	Gateway Mobile Switching Centre
GPRS	General Packet Radio Service
GGSN	Gateway GPRS Support Node
gNB	Next Generation Node B
GERA	GSM EDGW Radio Access
HLR	Home Location Register
HSDPA	High-speed downlink packet access
HSUPA	High-speed uplink packet access
HSPA	High-speed packet access
HSS	Home Subscriber Server
H-CRAN	Heterogeneous Cloud RAN
ISDN	Integrated Services Digital Networks
IMEI	International Mobile Equipment Identity
IP	Internet Protocol
IMS	IP Multimedia Subsystem
IoT	Internet of Things
IETF	Internet Engineering Task Force
IMT	International Mobile Telecommunications
ITU	International Telecommunication Union

ITU-R	International Telecommunication Union Radiocommunication Sector
ITU-T	Standardization of Telecommunications Sector
ICN	Information Cnetric Networking
ICMP	Internet Control Message Protocol
IMSI	International Mobile Subscriber Identity
LTE	Long Term Evolution
LINP	Logically Isolated Network Partitions
LMF	Location Management Function
LAN	Local Area Networks
LPWAN	Low Power Wide Area Networks
LoRaWAN	Long Range Wireless Area Networks
MS	Mobile station
MSC	Mobile Services Switching Centre
MSRN	Mobile Station Roaming Number
MSISDN	Mobile Station ISDN number
MMS	Multimedia Messaging Service
ME	Mobile Equipment
MGW	Media Gateway
MME	Mobility Management Entity
MAC	Medium Access Control
MIMO	Multiple Input Multiple Output
MTC	Machine Type Communications
mMTC	Massive Machine Type Communications
MIoT	Massive Internet of Things
M2M	Machine-to-Machine
MEC	Mobile Edge Computing
MANO	Management and Orchestration
MVNO	Mobile Virtual Network Operators
MiTM	Man-in-the-Middle
NTT	Nippon Telegraph and Telephone
NMT	Nordic Mobile Telephone

NSS	Network and Switching Subsystem
NMS	Network Management System
NAS	Non-Access Stratum
NF	Network Function
NFV	Network Function Virtualization
NR	New Radio
NEF	Network Exposure Function
NGC	Next Generation Core
NB-IoT	Narrowband IoT
OSS	Operation and Support Subsystem
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational Expense
O-RAN	Open Radio Access Network
PSTN	Public Switched Telecommunication Network
PLMN	Public Land Mobile Networks
PS	Packet-switched
PGW	Packet Data Network Gateway
PCRF	Policy and Charging Rules Function
PDCP	Packet Data Convergence Protocol
PHY	Physical Layer
PDU	Protocol Data Unit
PCF	Policy Control Function
PAN	Personal Area Networks
QoS	Quality of Service
RAN	Radio Access Network
RNS	Radio Network Subsystems
RNC	Radio Network Controller
RRM	Radio Resource Management
RRC	Radio Resource Control
RLC	Radio Link Control
RB	Radio Bearers
RIT	Radio Interface Technologies
RRH	Remote Radio Head

RAT	Radio Access Technology
RF	Radio Frequency
RU	Radio Unit
RRU	Remote Radio Unit
RBS	Rogue Base Station
SMS-G	SMS Gateway
SMS-GMSC	Short Message Service Gateway Mobile Switching Centre
SMS-IWMSC	Short Message Service Inter-Working Mobile Switching Centre
SIM	Subscriber Identity Module
SAE	System Architecture Evolution
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDN	Software Defined Network
SRIT	Set of Radio Interface Technologies
SDU	Service Data Unit
SDN	Software Defined Networking
SBA	Service-Based Architecture
SMF	Session Management Function
SSC	Session and Service Continuity
SUPI	Subscription Permanent Identifier
SUCI	Subscription Concealed Identifier
TACS	Total Access Communications System
TDMA	Time division multiple access
TDD	Time Division Duplexing
UMTS	Universal Mobile Telecommunications Systems
USIM	UMTS Subscriber Identity Module
UE	User Equipment
UTRA	Universal Terrestrial Radio Access
UMB	Ultra Mobile Broadband
UTRAN	UMTS Terrestrial RAN
URLLC	Ultra Reliable Low Latency Communication

UPF	User Plane Function
UDM	Unified Data Management
UDR	Unified Data Repository
UDSF	Unstructured Data Storage Function
VLR	Visitor Location Register
VoIP	Voice over Internet Protocol
V2X	Vehicle-to-everything
WAP	Wireless Application Protocol
WWW	World Wide Wireless Web
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WCDMA	Wideband Code Division Multiple Access
WMSC	WCDMA Mobile Switching Centre

REFERENCES

[1] M. Attaran, "The impact of 5G on the evolution of intelligent automation and industry digitization," *Journal of Ambient Intelligence and Humanized Computing*, 2021.

https://link.springer.com/article/10.1007/s12652-020-02521-x

[2] A. G. Roopali Sood, "Digital Society from 1G to 5G: A Comparative Study," vol. 3, no. 2, 2014.

[3] A. R. Mishra, Advanced Cellular Network Planning and Optimisation_ 2G_2.5G_3G...Evolution to 4G, Wiley, 2007 .

[4] R. Haverans, "brainbridge.be," Brainbridge Workforce Solutions, 27 May 2021. [Online]. Available: https://www.brainbridge.be/en/blog/1g-5g-brief-history-evolution-mobile-standards.

[5] "electronics-notes.com," About Electronics Notes, [Online]. Available: https://www.electronics-notes.com/articles/connectivity/2g-gsm/network-architecture.php.

[6] "rfwireless-world.com," RF Wireless World 2012, RF & Wireless Vendors and Resources, [Online]. Available: https://www.rfwireless-world.com/Tutorials/gsm-architecture.html.

[7] D. C. Wheeler, "techwalla," Leaf Group Ltd. Leaf Group Media, 2022. [Online]. Available: https://www.techwalla.com/articles/the-advantages-and-disadvantages-of-gsm.

[8] D. S. K. Mohammad Meraj ud in Mir, "Evolution of Mobile Wireless Technology from 0G to 5G," vol. 6 (3), 2015.

[9] Q. Z. Mooi Choo Chuah, Design and Performance of 3G Wireless Networks and Wireless LANs, Springer Science-l-Business Media, Inc., 2006.

[10] C. Cox, An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications, 2014: Wiley.

[11] I. A. A. B. A. A. G. M. Y. Madhusanka Liyanage, A Comprehensive Guide to 5G Security, Wiley, 2018.

[12] A. R. Mishra, Fundamentals of Network Planning and Optimisation 2G/3G/4G, Wiley, 2018.

[13] "silicon-press.com," Silicon Press, [Online]. Available: http://www.silicon-press.com/briefs/brief.3g/.

[14] F. KHAN, LTE for 4G Mobile Broadband, Cambridge University Press, 2009.

[15] N. A. M. O. Magri Hicham, "4G System: Network Architecture and Performance," 2015.

[16] A. Kukushkin, Introduction to Mobile Network Engineering, Wiley, 2018.

[17] V. K. A. K. Suyash Tripathi, "LTE E-UTRAN and its Access Side Protocols," Radisys Corporation, 2011.

[18] “rfwireless-world.com,” RF Wireless World 2012, RF & Wireless Vendors and Resources, [Online]. Available: https://www.rfwireless-world.com/Terminology/Advantages-and-Disadvantages-of-LTE.html.

[19] S. Ahmadi, 5G NR - Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards, Academic Press, 2019.

[20] A. T. T. N. Harri Holma, 5G technology _ 3GPP new radio, Wiley, 2020.

[21] C. Cox, An Introduction to 5G_ The New Radio, 5G Network and Beyond, Wiley, 2020.

[22] “blog.3g4g.co.uk,” 6 March 2017. [Online]. Available: https://blog.3g4g.co.uk/2017/03/imt-2020-5g-requirements.html.

[23] M. Ramgir, “Internet of Things,” Pearson Education India, 2019.

[24] “heavy.ai,” 2022. [Online]. Available: https://www.heavy.ai/technical-glossary/5g-infrastructure.

[25] M. Ivezic, “5G TECHNOLOGY PRIMER,” 2020.

https://5g.security/5g-technology/5g-core-sba-components-architecture/

[26] Comcores, “Accelerating 5G virtual RAN deployment,” Design And Reuse.

https://www.design-reuse.com/articles/48029/accelerating-5g-virtual-ran-deployment.html

[27] G. S. P. G. R. B. J. H. David Hanes, IoT Fundamentals Networking Technologies, Protocols, and Use Cases for the Internet of Things, Cisco Press, 2017.

[28] “mongodb.com,” MongoDB, Inc., 2021. [Online]. Available: https://www.mongodb.com/cloud-explained/iot-architecture.

[29] W. Goddard, “IoT Network & Architecture,” 2020.

https://itchronicles.com/iot/iot-network-architecture/

[30] S. M. Rishi Vaish, “Smarter Strategies”.

https://newsroom.ibm.com/5G-accelerate-IOT

[31] A. K. D. S. S. P. G. J. J. P. C. R. MOHAMMAD WAZID, “Security in 5G-Enabled Internet ofThings Communication: Issues,Challenges and Future ResearchRoadmap,” vol. 8, 2020.

[32] “igzy.com,” 2020. [Online]. Available: https://igzy.com/iot-application-security-issues/.

[33] “5G and Data Privacy: An overview for policymakers,” GSMA, 2020.

https://www.gsma.com/publicpolicy/wp-content/uploads/2020/07/GSMA_5G_and_Data_Privacy_July_20.pdf

[34] “kaspersky.com,” AO Kaspersky Lab, 2022. [Online]. Available: https://www.kaspersky.com/resource-center/threats/5g-pros-and-cons.

[35] M. A. E. E. U. S. K. N. Prajwol Kumar Nakarmi, “Murat: Multi-RAT False Base Station Detector,” 2021.

[36] M. Fong, “Protecting High-Level Personnel from IMSI Catchers,” 2020.

https://www.securitymagazine.com/articles/91767-protecting-high-level-personnel-from-imsi-catchers

[37] D. R. C. P. T. H. Merlin Chlosta, “5GSUCI-Catchers: Still catching them all,” 2021.

https://dl.acm.org/doi/pdf/10.1145/3448300.3467826

[38] D. S. Ramasamy, “linkedin.com,” 2021. [Online]. Available: https://www.linkedin.com/pulse/5-security-suci-catcher-ramasamy-cissp-cism-gcti-gnfa-gcda-cipm.

[39] “Why is 5G important?,” 2019.

https://www.verizon.com/about/our-company/5g/why-5g-important-discover-importance-5g-technology

[40] “avnet.com,” Avnet, Inc., 2021. [Online]. Available: https://www.avnet.com/wps/portal/abacus/solutions/markets/communications/5g-solutions/5g-beamforming/.

[41] “nokia.com,” NOKIA, [Online]. Available: https://www.nokia.com/networks/insights/privacy-challenges-security-solutions-5g-networks/.