"Research Implementation and Analysis of security issues in 5G Network Slicing using SDN and NFV"

Capstone Project MINT 709

by

Apoorva Dinesh Kumar

Master of Science in Internetworking

Department of Electrical and Computer Engineering

University of Alberta

Supervisor

Sandeep Kaur

Table of Contents

TABLE OF FIGURES

ABSTRACT:

The revolutionary part of 5G architecture makes it different from the previous generations of telecommunication, by providing the services such as enhanced mobile broadband(eMBB), ultra-reliable and low latency communications (URLCC), and massive machine type communications(mMTC). The challenging target of this novel technology is to provide massive communications between millions of devices with high data rate, more bandwidth and low latency using the protocol structure of Massive Internet of things (massive IoT). In terms of massive IoT application, the connected sensors and digital analytics that IoT technology offers can be used to track, sense and store as per the requirement. This includes the data flow from sensors to the IoT gateway and over the IP network, further to management and orchestration data centers in the h/w layer, virtualized layer and IoT applications. Here the technology has to meet extreme communications and traffic management. To meet these challenges, network should be highly programmable, automated, modular and use the concept of Network slicing. Network slices provides specific services over a single shared infrastructure. The idea is to focus on an isolated slice with a software defined framework assuaging various virtual network functions required for a massive IoT use case. Considering the architecture of a network slice (resource, network and service layers), we find difficulties in both implementation and security. The current use case has to decouple user plane and data plane making the application dynamic and to provide multi-tenancy in the packet flow between them. However, the focus is over the security threats that includes, poorly designed slice template and some threats include creating a fake slice. During the run-time of slice, it is exposed to variety of threats such as DoS, DDoS, performance attacks, data exposure and privacy breaks.

Chapter 1. Evolution of 5G Architecture.

1-1 History of Wireless Communication:

What differentiates us from animals is the way we communicate with each other. This whole journey of communication started with a cave man. Cave men used to gather around fire to discuss about their day-to-day activities, which we can compare to our modern-day social networking media such as Facebook. One certain day they decided to record their activities or knowledge by inscribing on caves, which we can now compare to modern day blogging. However, the problem with this method of communication was that it was localized. When people started moving out of caves, long distance communication became very important.

Furthermore, smoke signals were initiated, which was the first long distance communication. Smoke signals were used in the early 19th century by Northern Americans, where each tribe had its own signaling system. A smoke from top of the hill articulated incursion by tribes. It was also used in ancient China, where a soldier signified danger messages by sending smoke signals tower to tower. In this way, they could communicate as far as 750km within few hours. The smoke signal is still used by Catholic activists to indicate the selection of their new pope.

However, communication then evolved using pigeons. Pigeons, due to their natural roaming ability were extensively used for long distance communication. In 19th century pigeons were used to transmit star quotations from one city to another. Later Pony Express was the first courier service, where a human messenger relayed messages, mails, newspapers and small packages by horseback using small relay stations. This was used in the mid 19th century to convey messages between east coast and west coast of America. Later, semaphore flags which were a telegraphy system were used to convey information in the form of vision with handheld flags, rods and disks. It is still used during under wave replenishment at sea and as acceptable mode of communication in emergency.

So, in our analysis the last two hundred years are of primary importance. When we go back into ancient Greek, Roman and Chinese cultures, there were certain random experiments conducted which brought about the relationship between electricity and magnetism. These were certain foremost efforts posed to the development of wireless communications. Any of the experiments were not intentionally orchestrated to discover the wireless communication. Even in the 19th

century, when the connection between electricity and magnetism was first developed, the intuition of what it could achieve in future was naturally missing amongst all the researchers. Evidently, all the random experiments that were eventually conducted in 19th century led to the kind of communication system that we are experiencing today. Thus, in [1] explains ground-breaking experiments that paved way to create such intelligent technology we are experiencing today.

It was during the early 18th century, say the year 1820, a Danish physicist Hans Christian Orsted, during one of his ongoing lectures discovered that when current was on and off of a battery, the compass needle deflected. As per his observation, magnetic needle aligned itself perpendicularly to the current-carrying wire. This particular experiment invoiced the fact that an electric current created a magnetic field in a circular manner as it flows through a wire. This was clear evidence depicting the relationship between electricity and magnetism, which inspired the development of Electro-magnetic theory [1].

The connection between electricity and magnetism was of immense importance that rapidly led to further developments. From the years 1823 to 1826, Dominigue Francois Jean Arago discovered rotary magnetism (Arago's rotation), which also proved the relationship between electricity and magnetism. Further, Andre-Marie-Ampere, another French physicist and mathematician, contributed the Electromagnetism phenomenon. The phenomenon stated, when two wires parallel to each other carried current in same direction would attract each other and would repel if the parallel wires carried current in opposite direction.

Michael Faraday's contributions are of significant importance in this journey. He successfully built two devices to prove electromagnetic induction. When he passed current through a particular coil there was a small current induced in a nearby coil that was placed. From this experiment we could infer transmission of electrons from one charged device to another with air as a medium. This proved to be a strong base to send signals wireless via air. He also found that the plane of vibration of a beam of linearly polarized light incident on a piece of glass rotated when a magnetic field was applied in the direction of propagation of the beam [1]. The theory speculated that light could be vibration of electric and magnetic lines of force and is electro-magnetic disturbance of certain wavelength. Faraday predicted the existence of Electromagnetic waves that occupied the empty space around the conductor.

As stated in [1], Later Samuel Finley Breese Morse, an American painter utilized the concept of electromagnetism to build single wired electric telegraph system. Telegraph was the first attempt to use electromagnetism in an effort to communicate, which offered a new speed of information transmission. The working of telegraph was made more efficient by using Morse's code in which each letter was given with a code of dots or lines. The important parts of electric telegraph included electro-magnet, battery, Morse key and cable. When the Morse key was pressed, the circuit connected with a battery was completed through ground. The electro-magnet on the receiver side attracted the armature producing a click sound and the imprinting the code, which denoted the message that was supposed to be delivered from the sender's side. However, when the signal was down the problem was solved placing a repeater in the circuit. This is how the first electric long-distance communication took place.

Further James Maxwell also contributed to the progress of wireless communication, by proving the existence of electromagnetic waves and formulating the electromagnetic theory of light. His theory explained, a moving electric charge can create an electro-magnetic field which spreads out through space at a constant speed, basically the speed of light. Which stroke him to think about electromagnetic waves and light must actually be different forms of the same thing. He also predicted the existence of radio waves, which was very significant finding for the development of wireless communication [1].

Furthermore, in 1866, the first transatlantic telegraph cable was installed and operated by Morse code, with a speed of 5 words per minute. Then in 1895 was the discovery of transmission of Morse coded wireless wave transmission, using a spark gap transmitter. Further, Marconi demonstrated wireless communications by transmitting radio signals over long distance in 1920. He also concluded if the height of antenna could be raised, then the range of radio signal transmission could be extended, which was derived based on wireless telegraphy.

In 1920, we had our first commercial radio broadcast. In 1921, the police car dispatch radios came on the scene. In 1930, the television broadcast experiments were started by the BBC. In 1935, the first telephone call was made around the world. World War II led to rapid advancements in radio technology. In 1947, W. Tyrell proposed hybrid circuits for microwaves, and H.E. Kallaman constructed the VSWR indictor meter. In 1955, John R. Pierce proposed using satellites for communications. Sony marketed the first transistor radio. In 1957, the Soviet Union launched

Sputnik I, which transmitted telemetry signals for about five months. The carterfone was a device invented in 1968 by Thomas Carter, which connected a two-way radio to the telephone system, letting one person on the radio talk to another person on the phone [1].

1-2 FIRST-GENERATION SYSTEMS:

1-2-1 Introduction:

After the development of frequency modulation (FM) in the year 1930 [2], the technology helped many communications in Second World War. An incoming audio signal was modulated with a carrier frequency signal to improve its strength and transmit the signal to longer distances, which helped mobile telephony serve large cities. However, such systems were of limited capacities which introduced to the implementation of Advanced Mobile Phone Service (AMPS). AMPS was introduced by AT&T in 1983, where the service was a standard system for analog signal cellular telephone service that was implemented in United States and other parts of the countries. AMPS allocated frequency ranging within 800- and 900-Megahertz (MHz) spectrum to telephone. The bands within the spectrum were divided into 30 kHz sub-bands for reverse channels and for forward channels. The division of the spectrum was achieved using frequency division multiple access (FDMA). Meanwhile Europeans also contributed to the evolving communication technology by introducing Nordic Mobile Telephony (NMT) and NMT900 which operated in 450 MHz and 900 MHz bands respectively. NMT is an analog automated mobile telephone network assisting the long-distance calls, where anyone could call anyone else from anywhere. NMT voice channel was transmitted using FM and the signal transfer speed varying from 600-1200 bits per second, using Fast Frequency Shift Keying (FFSK) modulation. NMT also allowed multiplexing in the channels between base station and receivers. Also, the British introduced another technology called Total Access Communications System (TACS) which operated in 900 MHz band and was basically a modified version of AMPS. All these AMPS, NMT and TACS constituted the technology of first-generation communication system. First generation was the most prolific wireless voice communication platforms that highlight certain key concepts such as frequency reuse, mobility of the subscriber and hand-offs.

1-2-2 General 1G system architecture:

Typically, whenever we refer to cellular communication, it is usually associated with either AMPS or TACS [2].

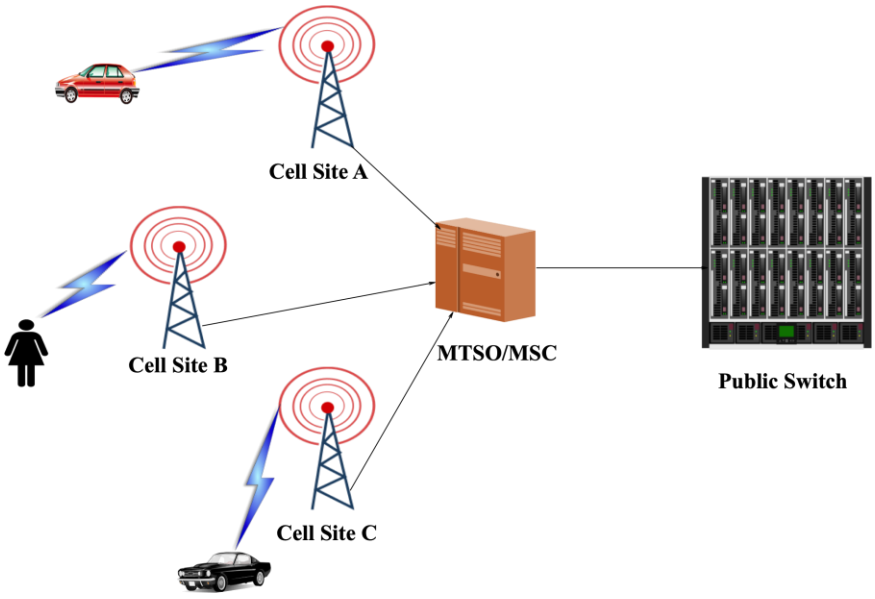

Figure 1. General 1G System Architecture [2]

As shown in figure 1, the architecture comprises of many high-level system blocks of cellular network. The architecture comprises of radio transmission between the mobile device and base station at every cell. Radio transmissions included full-duplex configuration, which means there was separate frequency band allocation to the transmitting and receiving signals. A cell site acts as a conduit between the mobile device and Mobile Telephone system office (MTSO), wherein the signal received from mobile device is conveyed to MTSO either through T1/E1 lines or microwave system. Also, from the end user side, mobile subscriber unit (MSU) consists of control unit and transceiver that transmits to cell site and receive from cell site [2].

Further in [1], MTSO effectively acts as a brain of the network, processing the calls and connecting cell site radio links to Public Service Telephone Network (PSTN). MTSO maintains call records and statuses of every single subscriber, also call routing and billing information. All the traffic between cellular network and PSTN or other networks passes through MTSO via landline cable connections, where MTSO converts the energy it receives from base stations to another medium. Architecture of MTSO comprises of Mobile Switching Centre (MSC), field monitoring and relay stations. The main function of MSC is to route mobile phone calls. MTSO also has an important feature that interconnects mobile telephone with land telephone network and provides mobile customers with services such as direct dialed mobile-to-mobile, mobile-to-land, and land-to-mobile callings.

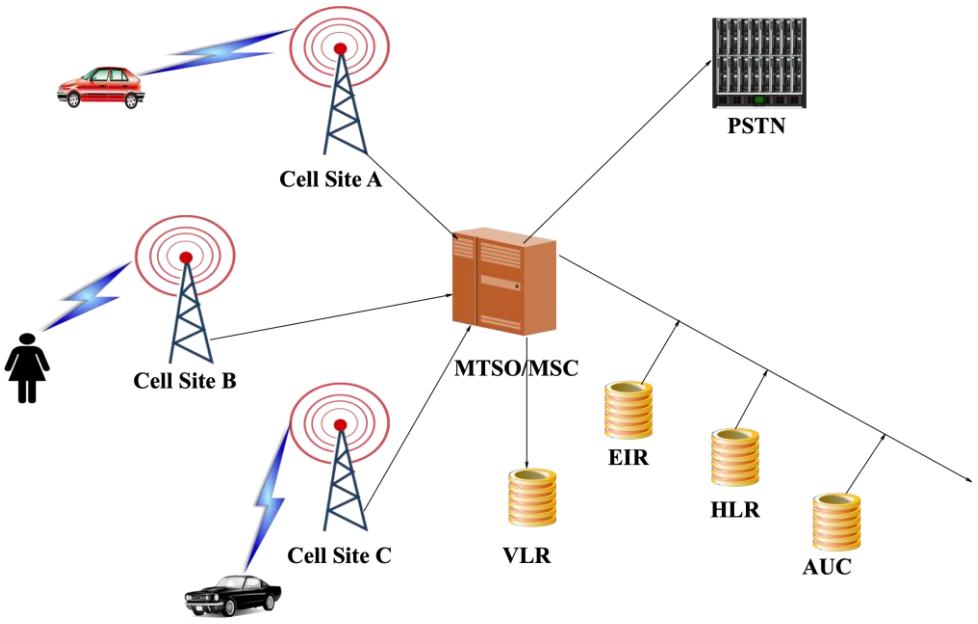

Figure 2. Components in MTSO/MSC [2]

As shown in the figure 2, MTSO consist of databases such as Home Location Register (HLR) and Visitor Location Register (VLR). HLR stores the documentations of the mobile subscribers of the network and VLR performs temporary caching of the details of subscribed users. MTSO consists of another database called Equipment Identity Register (EIR) to keep the record of blacklisted mobile phones. Also, there is Authentication center (AUC) for the purpose of authenticating every mobile user and encrypting the mobile communications between phone and network.

Further, PSTN has been evolving since Alexander Graham Bell made his first voice transmission over wire in 1876. PSTN mainly consists of transmission, switching, signaling and intelligent networks. Transmission network consists of multiplexers/demultiplexers as their nodes and optical fiber/coaxial cables to be their links. Multiplexer coagulates several signals into one signal, transmitting as sequenced frames over the shared channel and demultiplexer reverses the process multiplexer. The switching network uses circuit switching mechanism which provides connection-oriented services for voice subscribers using the channel of fixed bandwidth. Signaling network allows the analog signals to be passed within every switch present in the network, so that switches reserve resources during exchanging traffic. Basically, PSTN acts an edge gateway routing calls within same area code or if its outside there has to be an additional are code [2].

Figure 3. General Cell site Configuration [2]

As given in [2] every cell is of hexagonal shape covering large areas and placed at higher places preventing any gaps or overlapping. Every cell has one cell site, wherein each cell site consists of an antenna and few more electronic communications components. Transceiver locates receivers providing simultaneous two-way voice communication between cell site and subscriber phone. And regarding the spectrum allocation, cellular systems have been allocated with 25MHz for both band A and band B operators. The 25MHz is divided into 12.5 MHz each for transmitter and receiver.

One of the most important features of 1G technology includes the implementation of Handoff. There are several algorithms invoked to generate and process a handoff request and its order. Handoffs operate at low power levels and provide high capacity. As and when the RF signals from the mobile device keeps decreasing and reaches certain level, current base station contacts the adjacent one to transfer the mobile unit that has a call-in progress on a particular voice channel to another voice channel without interrupting the call [2].

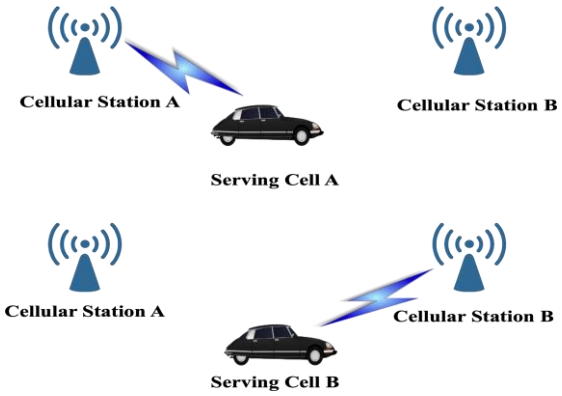

Figure 4. Analog Hand-off [2]

Another important feature includes frequency reuse to have higher capacity per geographic area than an MTS or IMTS. The concept mainly revolves around how we manage C/I signal levels for an analog system.

1-2-3 Frequency Division Multiple Access (FDMA):

FDMA in [4] is one of the most common multiple access procedures used in 1G suitable for analog communication, involves division of frequency bandwidth or a channel into multiple individual channels of certain bandwidth, such that each conversation is carried on a different frequency. Guard bands are maintained between each segment of the frequency band in order to avoid the cross talks. In FDMA, there is a central controller that allocates frequency bands to users, solely based on their needs. Once band has been allocated to the user, it remains until the entire flow of information has been done. We can divide the frequencies using frequency dividers such as flip-flop clock frequency dividers.

1-2-4 SECURITY IMPLEMENTATIONS IN 1G:

We live in a world where there is constant change in technology. Security in wireless cellular networks determines the system to manage, protect and distribute sensitive information in a secure way. Cellular Communication has become an important part of our daily life. The wireless medium has certain limitations over the wired medium such as open access, limited bandwidth and systems complexity. These limitations make it difficult although possible to provide security features such as authentication, integrity and confidentiality. Mobile networks started witnessing serious threats and challenges immediately after the introduction of the first generation (also called 1G) of mobile technology and has kept on growing as a complex and challenging threat landscape. 1G was primarily introduced to offer mobility for voice users. Consumers started witnessing the freedom to attend and make calls while mobile. Criminals discovered an opportunity and methods to commit mobile frauds and impersonate the legal subscribers to hack their phone to make free calls. Cell phone cloning became an industry by making and selling illegal cloned phones. Some hackers identified new ways to hijack and eavesdrop on the calls while being made and listen in to the private conversations for various nefarious reasons. In 1G, voice is transmitted as an analog signal avoiding encryption of scrambling. So, they can be eavesdropped upon using handheld scanners which are sold at places like Radio Shack [1].

.

1-3 SECOND - GENERATION SYSTEMS:

1-3-1 Introduction:

After the first generation, digital radio technology was implanted various modulation formats were utilized in order to increase the quality and capacity of the existing cellular systems. During the 2G systems, main service was mobile fax, where in it utilized 9.6b kbps to transport the information content. 2G systems deployed the utilization of digital radio technology in a Cellular, Personal Communication Services (PCS) and Specialized Mobile radio (SMR) in order to improvise the voice traffic throughput. The main difference between 1G and 2G is depicted in the following figure 5 [2].

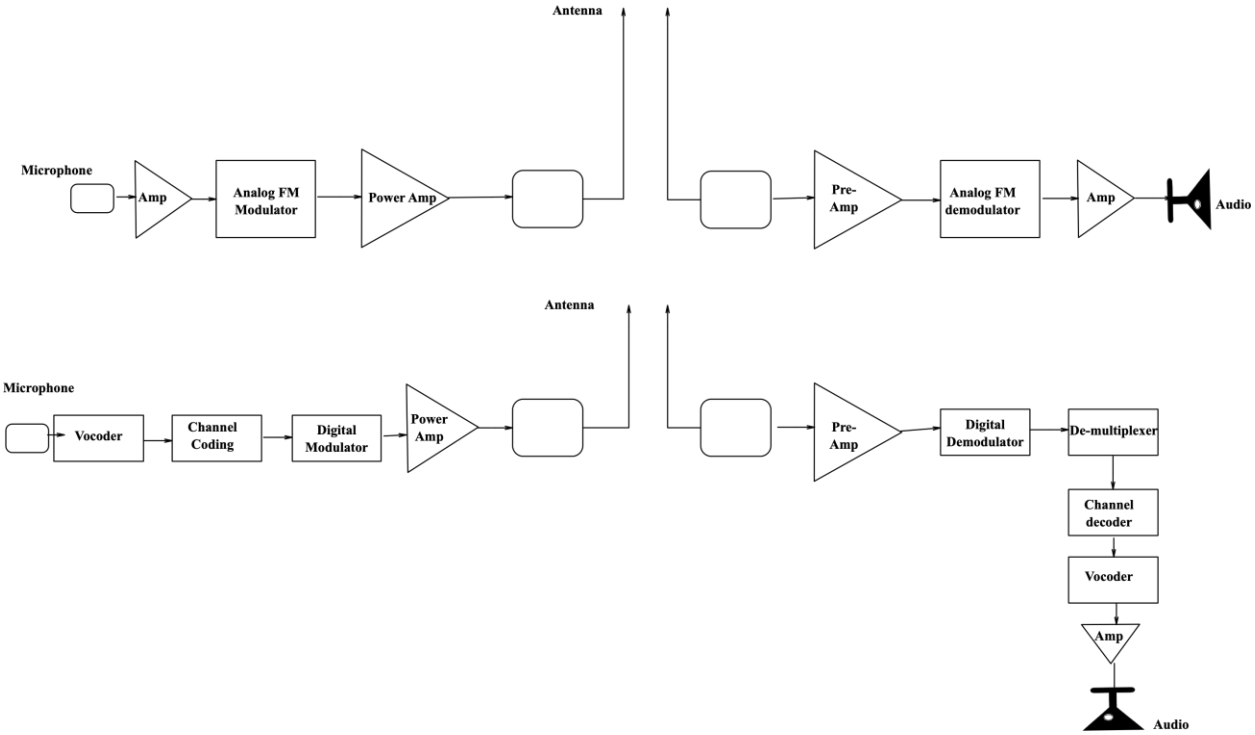

Figure 5. Analog and Digital Radio [2]

Reviewing the digital portion, voice being the initial information content to the microphone is being transmitted to vocoder. Vocoder converts the audio signal into data stream using a coding technique such as Pulse code modulation (sampling and quantizing), also performing audio data compression, and multiplexing also voice encryption. This helps in reducing the amount of data bits required to represent the audio. Further the digitized data is designated towards channel

encoder which encodes the data using any encoding mechanism to keep the data stolen and to remove redundancies from data, also it will be possible for the receiver to re-construct the data. Further, data signal is digitally modulated using ASK, FSK or PSK onto an RF carrier wave. The modulated carrier wave is then filtered and transmitted out from antenna [2].

On the other side, modulated RF carrier signal passes through filter and a pre-amp. And then to the digital de-modulator to down-convert the signal. Later the demodulated signal is sent to channel decoder to apply the inverse of encoding. Next to vocoder which converts the digital signal to analog signal and amplified for the user at the other end of communication.

So, why digital over analog? Digital signals are more secure with low bandwidths. Allows signals to be transmitted over lengthy distance with higher rate of transmission. By using digital waves, we can translate messages, audio, video into any device language [2].

The major benefits with 2G architecture include,

- Increased capacity over analog.
- Reduced capital infrastructure cost.
- Reduced per capital subscriber cost.
- Reduced cellular fraud.
- Improved features.
- Encryption.

However, people during implementation of 2G architecture had to face difficulties, as to how to include this architecture with the legacy ones.

The digital techniques for the cellular communication fall into two primary categories: AMPS and TACS spectrum. The Global system for Mobile Communications (GSM) is the chosen digital modulation technique for markets that use TACS spectrum allocation. The option is however, between Time division Multiple Access (TDMA) and Code division multiple access (CDMA) radio access platforms for the AMPS markets. The IDEN radio access platform is available in addition to AMPS/TACS spectrum decision, which operates in SMR band [3].

The radio channel is a communication medium that has to be shared by many subscribers in one cell. Mobile stations compete with another for the frequency resource to transmit their information

streams. However, there are high chances for the multiple access problem to occur, henceforth, special multiple access procedures are to be followed in order to divide the available frequency band. Following are the multiple access methods that were used in 2G technology.

1-3-2 Time Division Multiple Access (TDMA):

TDMA works on the principle that the complete bandwidth of a channel is allocated to all the users; however, users have limited time to perform their communication. The user time slots are combined into frames, as shown in the figure. Where each slot has been allocated to each user with a frame of six slots. The frames repeat after every Tf interval. Also, the guard bands must be allocated to prevent the interference between the users, which can be caused by variations of synchronization times [4].

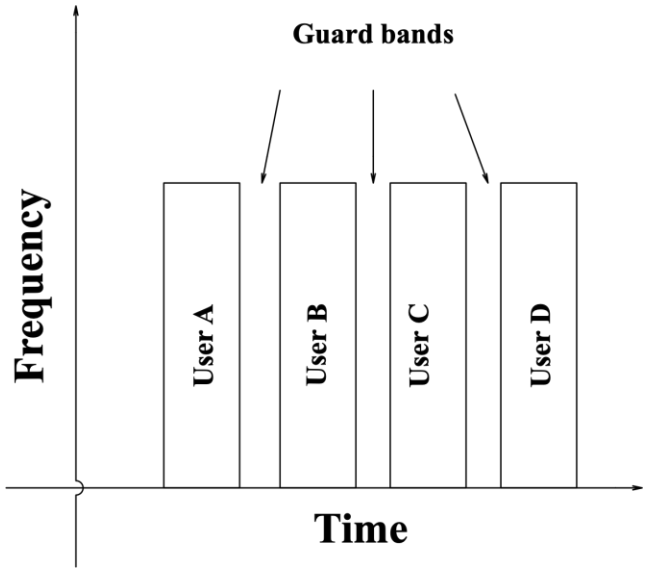

Figure 6 Time Division Multiple Access [4]

1-3-3 Code Division Multiple Access (CDMA):

CDMA allows one channel to carry all the transmission simultaneously, without any collisions. Consider there are four stations, station 1, 2, 3 and 4. Each and every station has a unique code, say C1, C2, C3 and C4. All these stations will be transmitting the data D1, D2, D3 and D4. Algebraic sum of the product of code and data say, C1*D1, C2*D2, C3*D3 and C4*D4 will be sent within the channel. Later in [4] at receiver side, whichever user wants to receive data from any other user, it has to multiply algebraic sum with the code of user it wants to listen to. Further,

sum it up and divide by number of users. This technique as a result gives the sender's data to receiver.

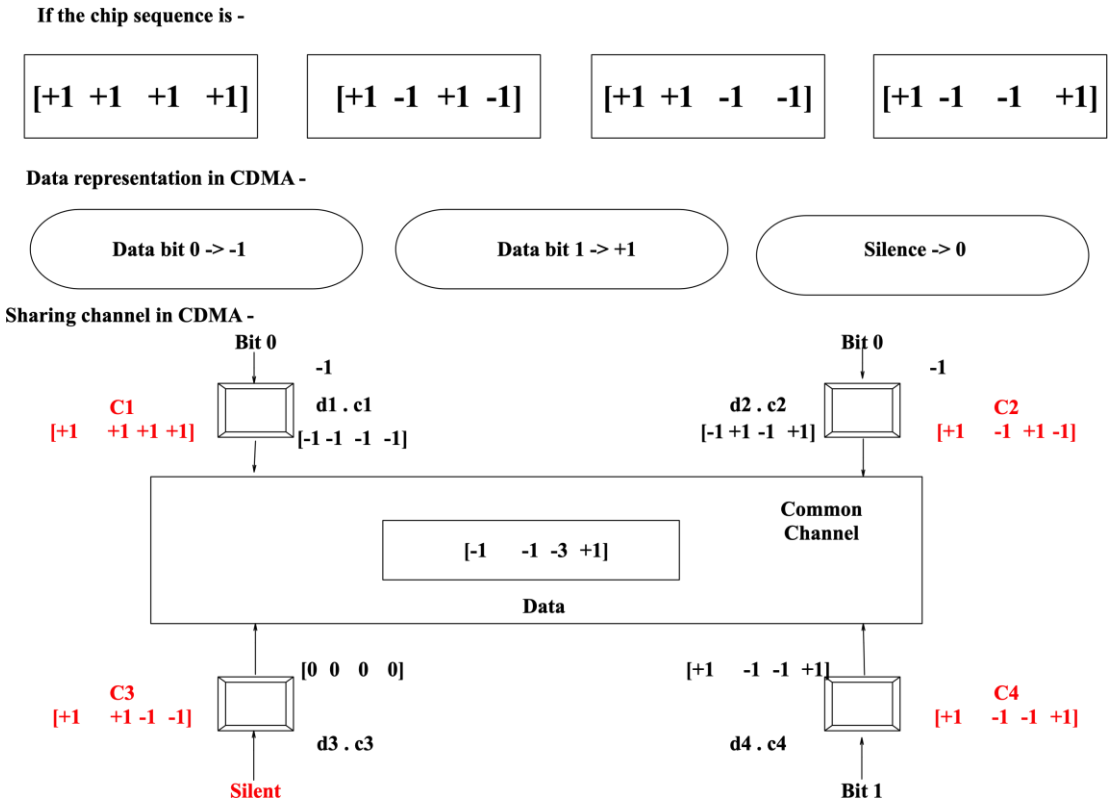

Figure 7 CDMA technology [4]

1-3-4 Global System for Mobile Communication (GSM):

The fundamental concepts of a GSM network are as shown in the following diagram. A user carries a mobile station which communicates over air interface as a physical medium with a base station called Base Transceiver Station (BTS). Every BTS consists of antennas and amplifiers which acts as transmitter or receiver equipment, also performs signal and protocol processing. Several base transceiver stations are controlled by a single Base station controller (BSC). BSC performs radio channel allocation, channel set-up also handovers. BTS and BSC are connected by fixed lines or point-to-point radio links, which together form radio access network [3].

The data sent by the users are routed through a switch called the Mobile Switching Center (MSC). It performs some of the switching functions such as path search, data forwarding, and service feature processing. MSC also performs additional functional such as location registration of users

and handovers of connection that is changing from one cell to another. Calls from another network such as PSTN, first arrive to Gateway MSC (GMSC), which basically queries HLR database to determine the location of a subscriber. The interworking of fixed and cellular is performed by Interworking Function (IWF). IWF performs mapping of protocols of cellular network onto those of respective fixed network. Connections to international networks are typically routed over the International Switching Center (ISC) of the respective country.

Further, GSM [3] consists of several databases such as Home Location Register (HLR) and the Visited Location Register (VLR) in order to store the current location of a mobile user. It helps to locate the customer, which cell exactly the person is located and to which base station he will be transmitting signals. Also, these databases are required to store the profiles of users, which are used for charging, billing and other administrative issues. Further two databases perform security related functions. Authentication Center (AUC) stores security related data such as keys for the purpose of authentication and encryption and Equipment Identity Register (EIR) stores equipment data. Further, the network management is performed from a central place called the Operation and Maintenance Center (OMC). As a whole, GSM network has been divided into radio access network, core network and the management network.

Initially in a GSM architecture, if a sender has to send a message to a receiver, mobile station wirelessly contacts immediate base stations. The radio access part of GSM that includes BTS and BSC routes the user information to MSC. MSC queries the HLR and VLR for the subscriber authentication. Later it locates the receiver's identity and switches the data to the respective base station controller upon encrypting the data using standard algorithms. Later base station locates forwards data to respective mobile station [3].

The following figure 8 shows the GSM system hierarchy, which includes each cell group assigned to a BSC, with at least one BSC present in a Location Area (LA).

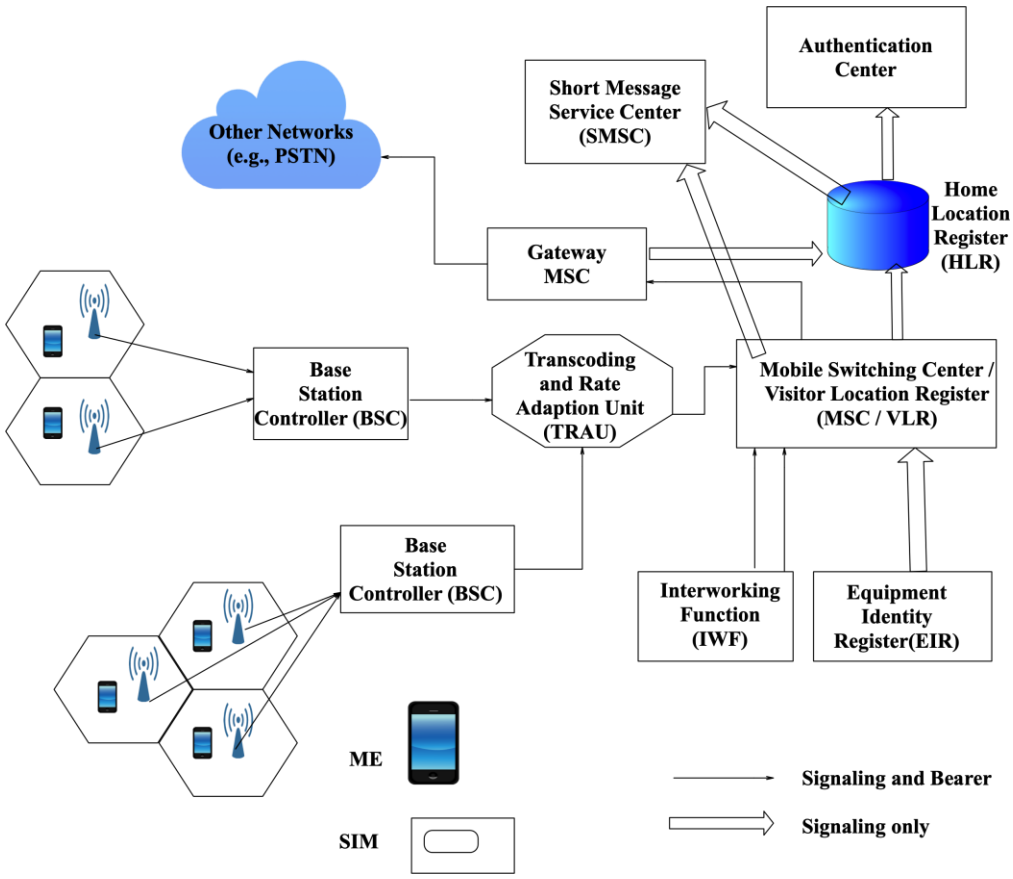

Figure 8 GSM System Architecture [2].

There has been a personal chip card plugged within every mobile station called Subscriber Identity Module (SIM). SIM turns out the piece of mobile equipment into complete mobile station with network usage privileges that allows making calls or receiving calls. Using SIM card, we can register to available local network onto different mobile stations. This enables international roaming independent of mobile equipment and network technology, provided that interface between SIM and end terminal is standardized. SIM contributes to personalization of mobile terminal storing short messages, telephone numbers also network specific data such as list of carrier frequencies used by network to broadcast system information periodically. SIM is protected by a Personal Identification Number (PIN), also performs implementation of standard cryptographic algorithms to authenticate and encrypt subscriber identity [3].

The International Mobile Station Equipment Identity (IMEI) uniquely identifies a mobile station internationally and gives clues about its manufacturer and date of manufacturing. IMEI are stored in three categories within EIR, whitelist (registered equipment's), blacklist (suspended

equipments) and gray list (equipment that has network access, but its use is reported to operating personnel). IMEI consists of Type Approval Code (TAC), Final Assembly Code (FAC), serial number and spare. The International Mobile Subscriber Identity (IMSI) [3] is also stored in SIM which helps in correctly bill the associated subscriber. Mobile station can only be operated if a SIM with a valid IMSI is inserted into equipment. IMSI consists of three parts, Mobile Country Code (MCC), Mobile Network Code (MNC), and Mobile Subscriber Identification Number (MSIN). Other identities within GSM architecture include Mobile subscriber ISDN number, Mobile station roaming number (MSRN), and Location Area identity.

1-3-5 SECURITY IMPLEMENTATIONS IN 2G:

Due the massive cellular network architecture of 2G, it is open to several attacks such as [1],

- Denial of Service attack (DOS): This is the most popular attack, which is caused by overloading the network resources, by sending massive amount of data to the network even beyond the capacity of the network. This result in the subscribers unable to access the network resources.
- Distributed Denial of Service attack (DDOS): This is similar to DOS, since it is not possible for a single host to launch attack on a large scale, multiple hosts are involved.
- Channel Jamming: This involves denial of service over a wireless channel for any legitimate user of that particular network.
- Unauthorized Access: This involves poor construction of authorization mechanism, which allows any attacker to enter inside the network and cause performance of un-authorized activities.
- Eavesdropping: This involves lack of encryption of communication flow within the channels, which causes attacker to eavesdrop sensitive calls, or text messages.
- Message Forgery: If the communication channel is not secure, then an attacker can intercept messages in both directions and change the content without the users even knowing.
- Message Replay: Even if the communication is secure, an attacker can intercept an encrypted message and replay it back at a later time and the user might not know that the packet received is the right one.

- Man in the middle attack: Attacker sits in between the two modes of communication intercept the messages between them and manipulate them.
- Session Hijacking: A malicious user can highjack an already established session and fake as the base station.

Following are the security features of 2G in order to prevent the above attacks, A3 algorithm as shown in figure 2-5 is used for the authentication of legitimate users of that network. A5 algorithm is used to encrypt the communication flow between the two devices. A8 is used to generate the cipher key. During authentication, VLR sends random value RAND to the SIM. Further, mobile station sends SRES generated cipher key to VLR and VLR will now compare the two values, accepts the subscriber if the values match or else will reject the authentication [1].

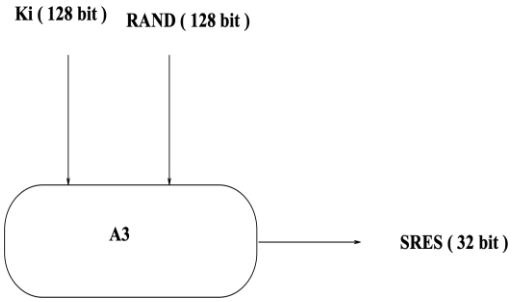

Figure 9 The A3 Algorithm [1]

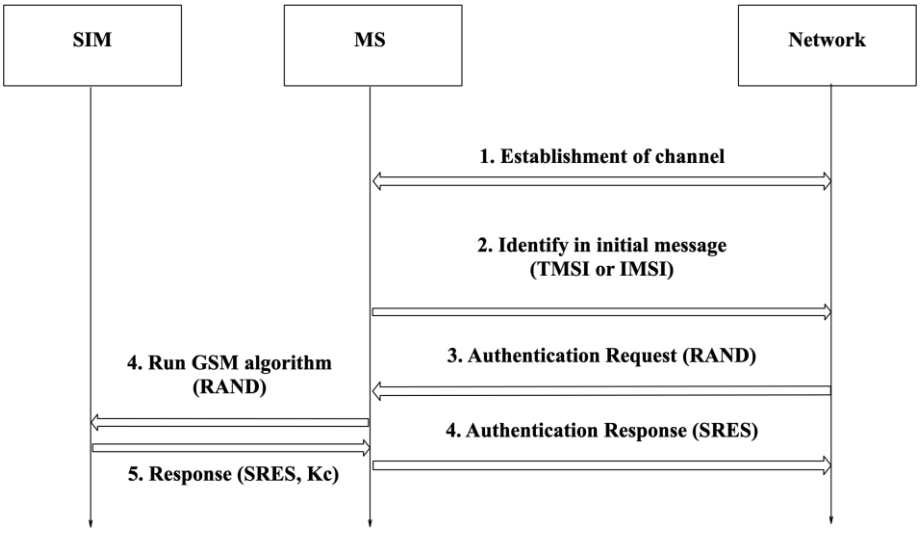

Figure 10 Working principle of A3 algorithm [1]

Further after authentication as shown in figure 10, base transceiver station and mobile station apply encryption to voice, data and signaling by applying the cipher key Kc [1]. Kc is generated by multiplexing random value (RAND) and key Ki using A8 algorithm. This confidentiality exists only between mobile station and base station transceiver, which is not within the whole GSM network. So, during the course, both SIM in the mobile station and VLR calculate the same Kc, which is used to encrypt or decrypt the messages using A5 algorithm.

1-4 THIRD - GENERATION SYSTEMS:

1-4-1 Introduction:

Based on the ITU specification, 3G can be defined as IMT-2000 which has received much attention as the enabler for high-speed data for the wireless mobility market. IMT-2000 is a radio and network access specification which has evolved through previous generations defining several methods or technologies to meet overall challenges of both mobile and high-speed data services [1].

IMT-2000 mandates data speeds of 144 Kbps at driving speeds, 384 Kbps for outside stationary use or walking speeds, and 2 Mbps for indoors. 3G arena had voice as the primary wireless application with the short message service (SMS) as the largest packet data service. However, it was a difficult decision that had to be made by any wireless operator, to choose transition method from 1G/2G to support various platforms that comprise IMT-2000. In order to understand this transition phase of the evolved generation, we need to understand the concept of 2.5G which allows to the network operator to decide, if he has to use cellular, PCS, or UMTS in order to deploy digital packet services.

Obviously [1], the decision on which platform to utilize involves guesswork and decisions based on a fundamental belief that particular platforms will enable services that are yet to be developed. And 2.5G acted as bridge between the already deployed 2G and those that were envisioned for 3G. Technologies implies in 2.5G includes General Packet Radio Service (GPRS)/High Speed Circuit Switched Data (HSCSD), Enhanced Data Rates for Global Evolution (EDGE), Code Division Multiple Access (CDMA 2000) (phase 1). 2.5 G is platform independent, data-play only technology. There were no changes in the wireless and fixed network access platforms. However, the main enhancement over 2G involved high data services (144.4 k).

The main objective of 2.5 G in [1] involved bridging of the existing 1G or 2G radio access platforms to that of 3G. It was the responsibility of the operator to choose the migration path based on company's resources, company's capital, spectrum, and manpower. However, the commonality that remained between the two generations involved deployment of packet-based data network regardless of which platform (GPRS/EDGE/CDMA) to be deployed.

Following figure 11 shows a table which illustrates relative advantages that each of the 2.5G platforms has over its fundamental underlying technology platform.

2G Technology	2.5G Technology	Enhancements	Migration-to-3G Platform
GSM	GPRS	• High speed packet data services (144.4K)	WCDMA
		• Uses existing radio spectrum	
IS-136	EDGE	• High speed packet data services (144.4K)	WCDMA
		• Uses existing radio spectrum	
CDMA	CDMA2000 (phase1)	• High speed packet data services (144.4K)	CDMA2000 –MC multi carrier
		• Uses existing radio spectrum	
		• 1XRTT used	

Figure 11 2G and 2.5G [2]

1-4-2 General Packet Radio Service (GPRS) Architecture:

In the Second Generation, GSM provided voice and data services over a circuit-switched network, which involved the entire dedicated modem between user device and the destination data network [5]. This was an inefficient mechanism to support the data traffic, also GSM provided the data speed rates up to 9.6 kbps only, which is considered to be very slow. GPRS was constructed as a solution to the inefficiency of 2G technology. GPRS provided more efficient packet-based data services at a higher data rate. GPRS is designed to provide packet data services at higher speeds than those of the standard GSM's circuit-based data services. Practically, it provides data speed up to 100kbps, with speeds of about 40kbps or 53 kbps more elastic, which is far better than the 9.6-kbps provided by standard GSM. Both GSM and GPRS have the same basic air interface with 200-kHz channel, divided into eight timeslots. Even though both have the same air interface, GPRS has different channel coding schemes. Most commonly GPRS uses Coding Scheme 2 (CS-2), which has less overheads and checks for error correction. GPRS even though provides comparatively more higher data transmission speed, it does not provide higher bandwidth which

was later implemented in 3G technology. However, the biggest advantage of GPRS include packet-based switching data, that allows a given user to consume RF resources only while sending or receiving the data. Suppose user is not sending or receiving the data, meanwhile another user can utilize the same RF resource improving the sharing of resource. For instance, if a user wants to request or send a web page, then MS has to request access to those resources and the network must allocate resources before the transfer can take place. Also, nothing is being transferred while the subscriber contemplates the content of page. But for the user it always appears to connect. During the web page is just being read, another user can utilize the same resource to do his job [5].

GPRS users can be grouped into three classes,

- Class A, which supports subscribers to use both voice and data services simultaneously over GPRS network.
- Class B supports simultaneous GPRS and GSM attach, but does not provide both voice and data services together.
- Class C provides both voice and data services together but can attach either GSM or GPRS.

Following diagram denotes the architecture of GPRS,

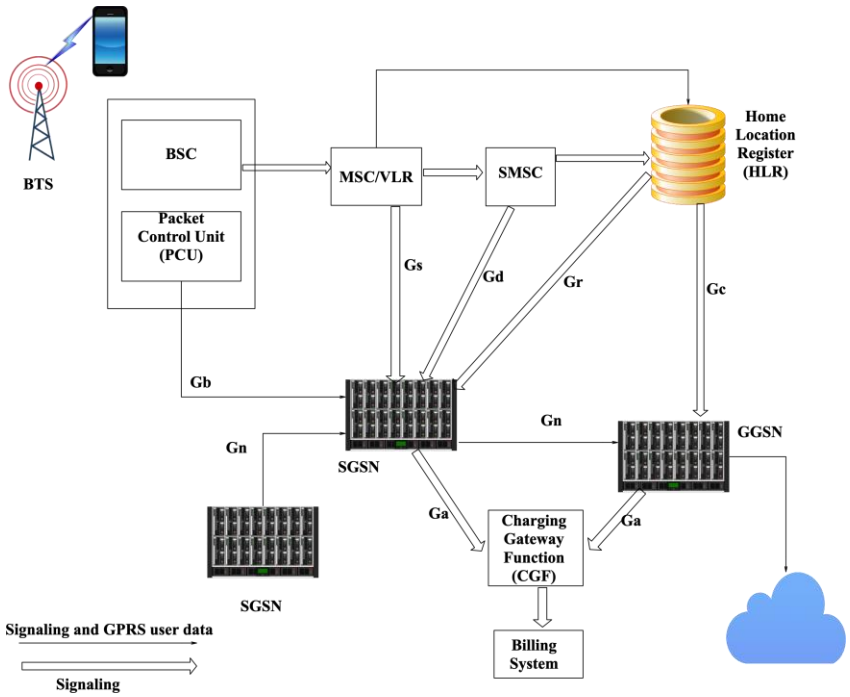

Figure 12 GPRS Network Architecture [2]

The PCU is the logical part of the network, performing functions such as access control of air interface, packet scheduling on the air interface, packet assembly and re-assembly [5]. PCU is physically integrated with the BSC. Packets from the mobile user is sent with overheads of every layer, as and when it reaches various components of the packetized data network, every overhead detail of the packet is read, recorded and updated. The GPRS support node (SGSN) is analogous to Mobile Switching Center (MSC)/Visitor Location Register (VLR) in the circuit-switched domain. SGSN performs mobility management, security, and access control functions. Unlike cellular areas in 1G/2G, GPRS has routing areas, where when a given mobile station changes between two regions, the mobile stations get registered even before the on-going session are hung, which in GPRS terms is known as Packet Data Protocol (PDP) context. The interface between SGSN and the BSC is Gb interface which is the frame relay-based interface using BSS GPRS protocol. This interface passes signals and control information along with data-traffic to and from SGSN.

Furthermore in [5], there is another Gs interface which is SS7-based interface which is located between SGSN and HLR/VLR. This performs location updates to HLR for GPRS subscribers and to retrieve GPRS-related subscription information for any GPRS subscriber that is located in the service area of the SGSN. Another optional interface, Gs interface lies between SGSN and MSC. The purpose of this interface is to provide co-ordination between the systems using both circuit-based and the packet-based systems.

SGSN interfaces with SMSC via Gd interface. This allows the subscribers to send and receive short messages over the GPRS network. Gateway GPRS support node (GGSN) acts as an external internet, thus, the user data enters and leaves the Public Land Mobile Network via GGSN. And the interface between SGSN and GGSN is known as Gn interface. GGSN will not query if the MS connected to SGSN belongs to the same GPRS routing area. However, if the MS belongs to external network, none of the SGSN connected to GGSN will have information about the respected MS, hence, GGSN queries HLR to know the information of the connected subscriber. Always a fixed address of IPv4 is assigned to the mobile station users [5].

Above physical layer (RF interface), there is Radio Link Control (RLC), and Medium Access Control (MAC) functions. And above these we have Logical Link Control (LLC), which provides logical link and framing structure for communication between MS and SGSN.

1-4-3 The EDGE Network Architecture:

As specified, and as used for GSM and GPRS, EDGE uses the same 200-kHz channels with eight timeslot structure. However, with EDGE, in addition to 0.3 Gaussian Minimum Shift Keying (GMSK) which was used in GSM, 8-PSK modulation is implemented.

0.3 GMSK means that there is a bandpass filter with a 3dB bandwidth of 81.25 kHz in the modulator. The objective with EDGE is to offer higher bandwidth efficiency, so that we can squeeze more user data from the same 200-kHz channel. This higher bandwidth efficiency is achieved through 8-PSK. For packet data services in an EDGE network, we refer to Enhanced GPRS (EGPRS) and the new coding schemes for EGPRS are termed Modulation and Coding Scheme-1 to Modulation and Coding Scheme-9 (MCS-1 to MCS-9) [2].

1-4-4 High-Speed Circuit Switched Data (HSCSD):

The need for higher data service speeds was well known before the introduction of GPRS or EDGE. At the time, only data services of up to 9.6 kbps were enabled by GSM, the that could be given was single time slot. The most obvious approach was a solution was by HSCSD to accommodate multiple number of slots with higher data rates. The original HSCSD versions permitted multiple time slots, each offering up to 9.6 kbps of user data. Four timeslots, for instance, could then give up to 38.4 kbps. A change in the channel coding scheme was subsequently proposed to allow 14.4 kbps of user data to be carried over a single time slot.

One of the main reasons for this change was to enable the mobile fax service to support a fax transmission at 14.4 Kbps over just a single timeslot [2]. Concatenation of four such timeslots could therefore offer speeds up to 57.6 Kbps. With the advent of the 8-PSK modulation that EDGE can provide, it is possible for HSCSD to achieve high throughput levels with fewer timeslots.

As implied previously, several deployment issues are associated with the introduction of CDMA2000-1x into a wireless system. Some of the obvious issues relate to the current spectrum usage that the operator has license control of. The spectrum usage considerations take on a different meaning depending on whether the system is new, that is, not a current infrastructure, or if it may or may not have the available spectrum from which to deploy the CDMA2000-1x channels.

The following diagram briefs the interaction between 1G, 2G, 2.5G and 3G platforms. So, the 3G implements two platforms WCDMA worked by 3GPP and CDMA 2000 worked by 3GPP2.

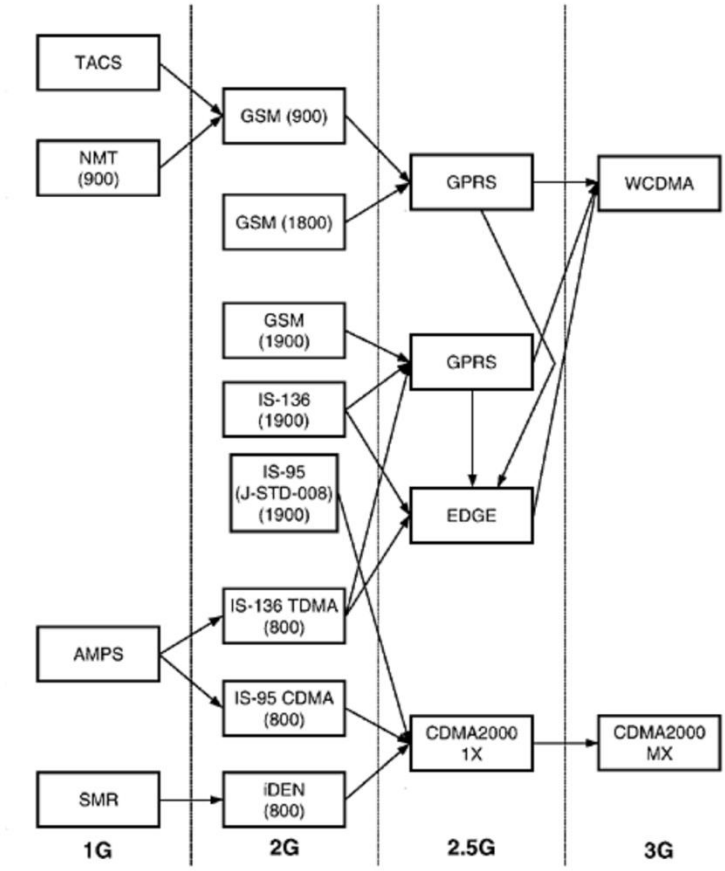

Figure 13 Migration Path [2]

1-4-5 Universal Mobile Telecommunications Service (UMTS):

The major transformations from 2.5G to 3G took place in access part of the network, and the radio access for UMTS is known as Universal Terrestrial Radio Access (UTRA). This implementation involves WCDMA, which includes both FDD and TDD modes to be the access method in the air interface. The core network of UMTS remained to be the evolution of the one in GSM. For several years, there has been enhancements in GSM architecture. Thus, for s given GSM specification, there were various versions such as Release 1996, Release 1997 and Release 1998 of 3GPP. 3GPP Release 1996 involved evolutions of access network to UTRA with also some of the existing GSM specifications (such as for the support of EDGE) [6]. The next release included 3GPP Release 2000, which included major changes to the core network. Since the changes were many there were two releases Release 4 and 5. Release 4 focuses more on changes needed to the core network and

Release 5 introduces a new call model, which means changes to user terminals, changes to core network, and some changes to access network.

UMTS [6] always aimed to provide the best service as compared to its previous generations with data rate up to 2Mbps. UMTS classifies its services based on the data availability for various types of users. For the conversational-based services, UMTS provided low delay tolerance, low jitter, low error tolerance and with less data, as the application only demands to be delay-sensitive. Furthermore, it provided services for interactive based which included request/response-type transactions. This involves low tolerance for errors but larger tolerance for delays than conversational services. Another service involves Streaming, which concerns one-way services using low-to-high bit rates. The UMTS air interface is a Direct-Sequence CDMA (DS-CDMA) system. DS-CDMA means that user data is spread over much wider bandwidth through multiplication by a sequence of pseudo-random bits called chips. The following figure explains you the conceptual depiction of spreading.

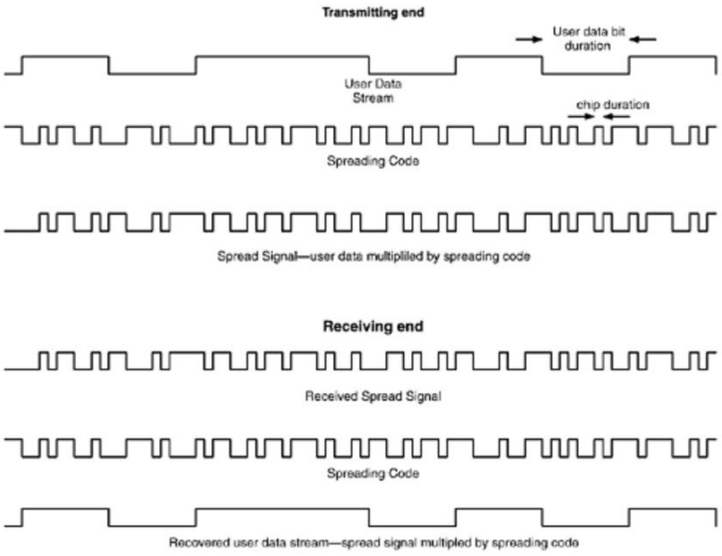

Figure 14 CDMA basic concept [2]

If multiple users are transmitting the data over the same frequency, then by applying pseudo-random sequence, the stream of data will actually be spread over higher bandwidth. It is very important to choose the particular type of sequence code, as it necessary to differentiate between two signals to cross-over or get interrupted by noise. Each user's spread data will be recovered by at the receiving end by dispreading the signals using appropriate spreading code. For instance, two

users (A and B) are transmitting data on the same frequency, but with two different spreading codes, then at the receiving end when the received signal is de-spread with the spreading code applicable to user A, then original data of user A is recovered. However, the data stream generated at the receiving end will have certain amount of noise which includes other signals that is the signal generated from user B. But the noise is relatively small. In other words, for a given bit of recovered user data, the signal-noise ratio must be sufficiently high. In CDMA, we refer to Eb/No, where Eb is the power density per bit of recovered user data and No is the noise power density. Provided that Eb/No is sufficiently large, then the user data can be recovered. Also, the ratio between the code rate and the user signal is constituted to be spreading factor. With the spreading being higher, higher will be the capability to recover data and higher will be the power of signal transmission. The chip rate in WCDMA is 3.84 * 10^6 chips/second. With the WCDMA FDD option, the paired 5-MHz carriers in the uplink and downlink are as follows: uplink—1920 MHz to 1980 MHz; downlink—2110 MHz to 2170 MHz. For the TDD option, a number of frequencies have been defined, including 1900 MHz to 1920 MHz, and 2010 MHz to 2025 MHz [6].

1-4-6 3GPP Release 1999 Network architecture:

Following shows the network architecture for 3GPP Release 1999, the first set of specifications for UMTS.

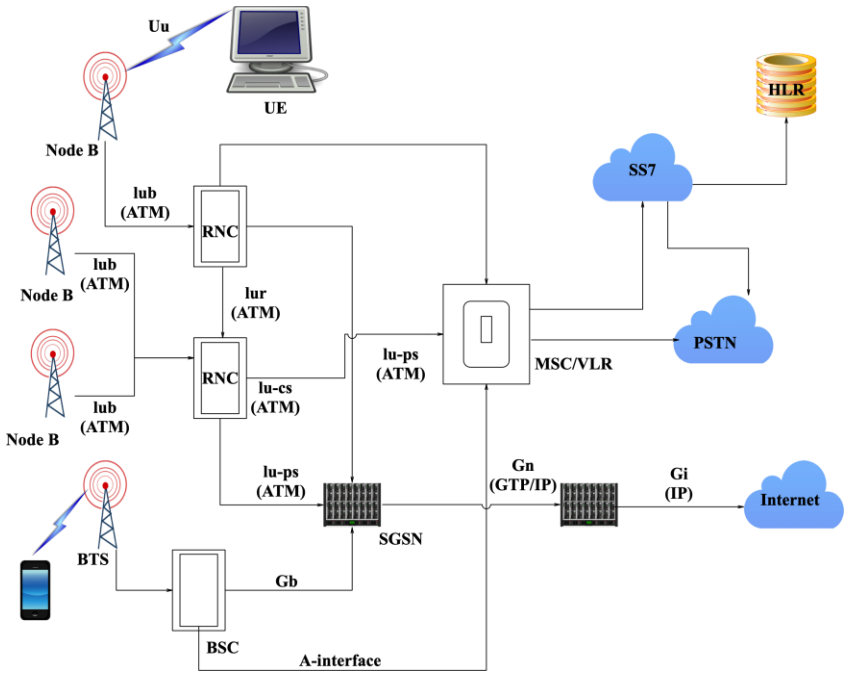

Figure 15 3GPP Release 1999 Network Architecture [2]

The interface between the User Equipment (UE) and the 3G network is called Uu interface, which is the WCDMA interface at the physical layer. And the base station in 3GPP is called Node B. There is a Radio Network Controller (RNC) which is analogous to BSC in GSM controls all the network resources of base stations (Node Bs) that are connected to it. Both RNC and Node B constitutes Radio Network Subsystem (RNS). There is lub interface between Node b and RNC, which is fully standardized, open and makes it possible for a Node B to connect any vendor's RNC.

Besides in [2] previous generations, in the Release 1999 both the UMTS RAN are connected with an interface called lur. The primary purpose of this interface is to support inter-RNC mobility and soft handover between Node Bs connected to different RNCs. Further there is a direct connection between UTRAN and core network via lu interface. Also, this interface is served for two purposes. One for the circuit-switched part of the core network via lu-CS interface, connecting RNC to single Mobile Switching Center (MSC)/Visitor Location Register (VLR). Another one for the packet-switched part of the core network via lu-PS connecting RNC to SGSN. Further, all the interfaces connecting each resources of the architecture are based on Asynchronous Transfer Mode (ATM), so that it can support both variable bit rate for packet-based services and constant bit rate for circuit-switched services. Also, the core network uses the same basic architecture as that of GSM/GPRS.

1-4-7 3GPP Release 4 Network Architecture:

The following figure 16 shows the basic architecture of 3GPP Release 4. In this architecture, there is evolution in the core network. The main difference between Release 1999 and Release 4 architecture is that the core network becomes a distributed network. Initially Release 1999 involved the implementation of traditional circuit switched MSCs, but in the Release 4 there is distributed switch architecture [2].

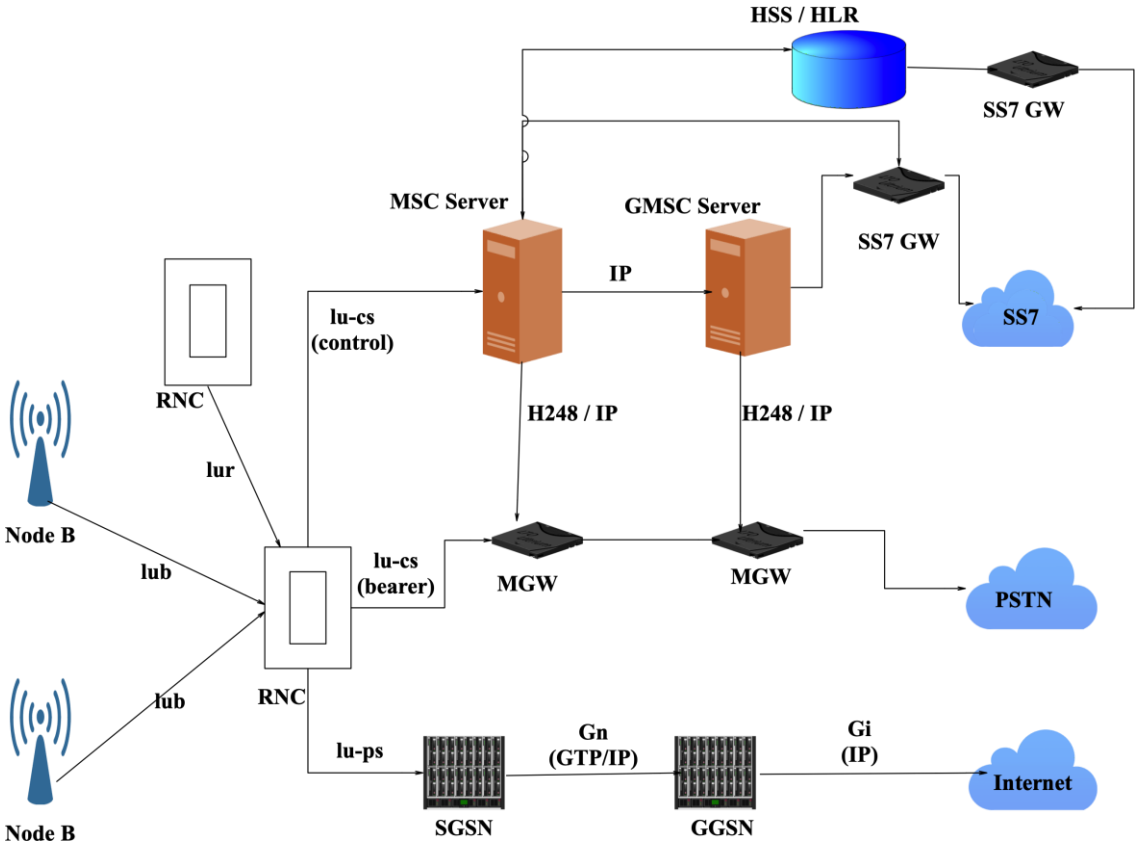

Figure 16 3GPP Release 4 Distributed Network Architecture [2]

Signals and data for the circuit-switched calls are transferred between the RNC and MSC server. MSC is basically divided into MSC server, which performs mobility management and call control logic and a media gateway (MGW), that contains switching matrix. Typically, MG takes up the calls from RNC via media path and routes those calls towards their destinations over a packet backbone. The packet backbone can either Real-Time Transport Protocol (RTP) or the Internet Protocol (IP), where in many cases circuit-switched calls picks up RTP. Further, the packet data traffic as shown in the figure will be transmitted from RNC into SGSN then to GGSN over IP backbone [2].

Furthermore, at the remote end, Gateway MSC Server (GMSC server) is utilized to handle the signals from media Gateway (MGW), in order to handoff the calls to another network such as PSTN. At this point there will be transcoding wherein MGW will convert the packetized voice to standard PCM for delivery to PSTN.

We can illustrate [2] the advantage of the distributed core network of Release 4 by considering the following scenario. Consider there has to be call set-up, with RNC being in city A and MSC located in city B. The call needs to travel from city A to city B, only to be connected back to local PSTN number in city A. With a distributed architecture, a call can be controlled by MSC in city A however, the media path can remain within city A, thereby reducing transmission requirements and reducing network operating costs.

Home Subscriber Server (HSS) is equivalent to HLR in GSM, with HSS using IP as packet-based transport protocol and HLR using Signaling System 7 (SS7) based interfaces. Also, we have SS7 gateway which on one side supports SS7 message over SS7 transport and on another side SS7 application messages over a packet network such as IP [2].

1-4-8 3GPP Release 5 All-IP Network Architecture:

Further evolutions in the UMTS architectures involves the implementation of an all-IP multimedia network architecture, involving the change in the overall call model.

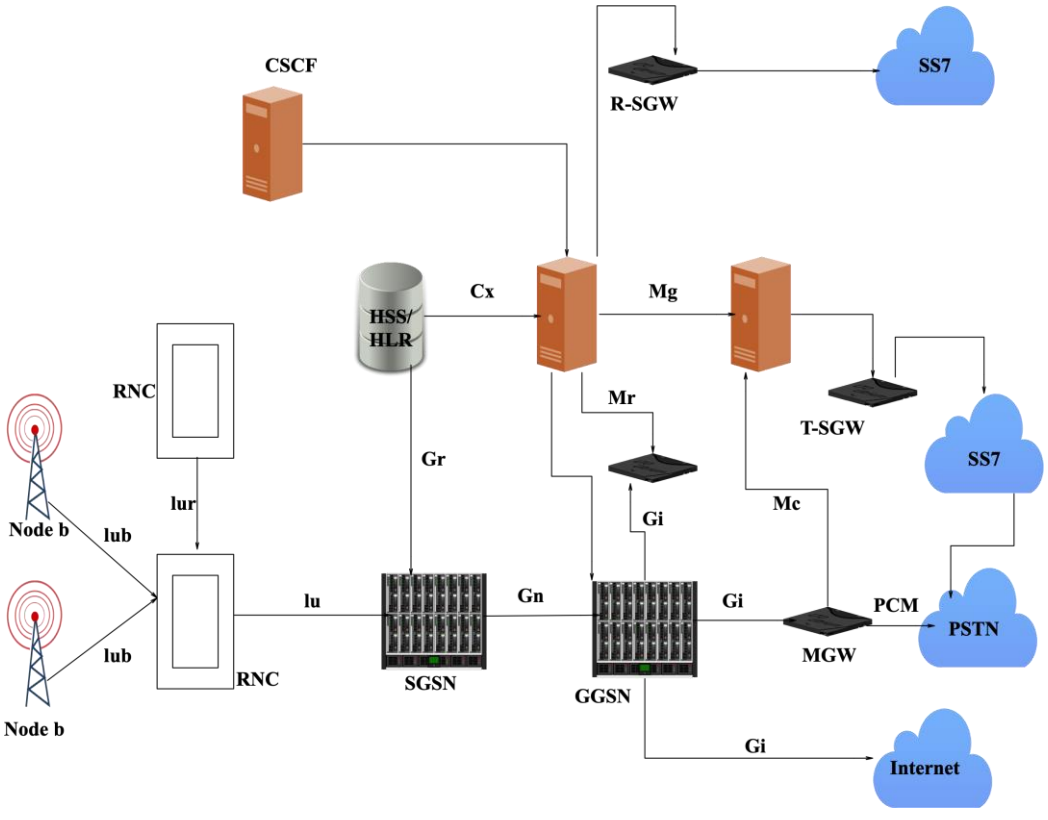

Figure 17 3GPP IP Multimedia Network Architecture [2]

There is an inclusion of various network elements such as Call State Control Function (CSCF), the Multimedia Resource Function (MRF), Media Gateway Control Function (MGCF), the Transport Signaling Gateway (T-SGW), and the Roaming Signaling Gateway (R-SGW).

Both the voice and data in this architecture will use the same interface, which is lu interface. User Equipment (UE) supports Session Initiation Protocol (SIP). The CSCF manages the establishment, maintenance, and release of multimedia sessions to and from user devices. This includes functions such as translation and routing.

The Multimedia Resource Function (MRF) is a conference bridging function used to support features such as multi-party calling and meet-me conference service [2].

The Transport Signaling Gateway (T-SGW) is an SS7 gateway that provides SS7 interworking with standard external networks such as the PSTN. The T-SGW will support Sigtran protocols. The Roaming Signaling Gateway (R-SGW) is a node that provides signaling interworking with legacy mobile networks that use standard SS7. In many cases, the T-SGW and R-SGW will exist within the same platform.

The media gateway (MGW) performs interworking with external networks at the media path level. The MGW in the 3GPP Release 5 network architecture is the same as the equivalent function within the 3GPP Release 4 architecture. The MGW is controlled by a Media Gateway Control Function (MGCF). The control protocol between these entities is ITU-T H.248. The MGCF also communicates with the CSCF. The protocol of choice for that interface is SIP [2].

1-4-9 CDMA 2000 Architecture:

A logical extension of an existing CDMA-one network [2] is the system architecture that will comprise a CDMA-2000 network, with the essential difference being the implementation of packet data services. For the purpose of managing packet data facilities, the implementation of the CDMA-2000 system is expected include improvements to BTS and BSC. The system architecture for a CDMA2000 network, due to packet data services, can be either centralized or distributed. The decision as to whether the system utilizes a distributed or centralized system is dependent upon the design requirements as well as operational issues. Hence, both of these WCDMA and

CDMA 2000 have certain commonalities in their construction, especially in the migration path from 2G to 3G technology.

1-5 CELLULAR SYSTEMS BEYOND 3G:

As the mobile devices were continuously integrating with various applications such as communications, information and medium of entertainments, there was drastic growth in the use of Internet. This growth led to the motivation for of mobile broadband construction. The Institute of Electrical and Electronic Engineers (IEEE) established 802.16, to develop a standard for the Wireless metropolitan area network (WMAN), which further was enhanced to support mobility. Further, Worldwide Interoperability for Microwave Access (WiMAX) was constructed to promote, develop, perform interoperability, conformance testing, and certify end-to-end wireless systems based on the IEEE 802.16 air-interface standards. [1] WiMAX network is designed using IP protocols, which does not offer circuit-switched voice telephony, however, it provided the voice services over Voice over Internet Protocol (VoIP). Later was the construction of LTE with OFDM, OFDMA technologies, based on the implementation of WiMAX. The beyond 3G system in 3GPP is called evolved universal terrestrial radio access (evolved UTRA) and is also widely referred to as LTE (Long-Term Evolution) while 3GPP2's version is called UMB (ultra-mobile broadband).

1-5-1 ORTHOGONAL FREQUENCY DIVISION MULTIPLEXING:

The traditional 3G systems utilized UMTS and CDMA 2000 technologies, which used CDM techniques [7]. However, Orthogonal Frequency Division Multiplexing (OFDM), was the key difference between the existing 3G systems and LTE providing high data rates with more advantages. OFDMA was the only solution to prevent the interventions of sessions caused due to multi-path, by using multi-carrier modulation. This methodology involves high bit rate data streams divided into several parallel lower bit rates. OFDMA also reduced the computational complexity, because of the implementation of Fast Fourier Transform (FFT).

1-5-2 MULTI-ANTENNA TECHNIQUE:

The multi-antenna technique increases the capacity of the link. Link robustness and spectral efficiency. It provides the solution for the multipath fading. Another important feature is multiuser MIMO, which allows multiple users in the uplink [7].

1-5-3 LONG TERM EVOLUTION (LTE):

There are few differences between the architecture of UMTS and LTE systems. Following figure denotes the differences between the network elements of both UMTS and LTE architectures.

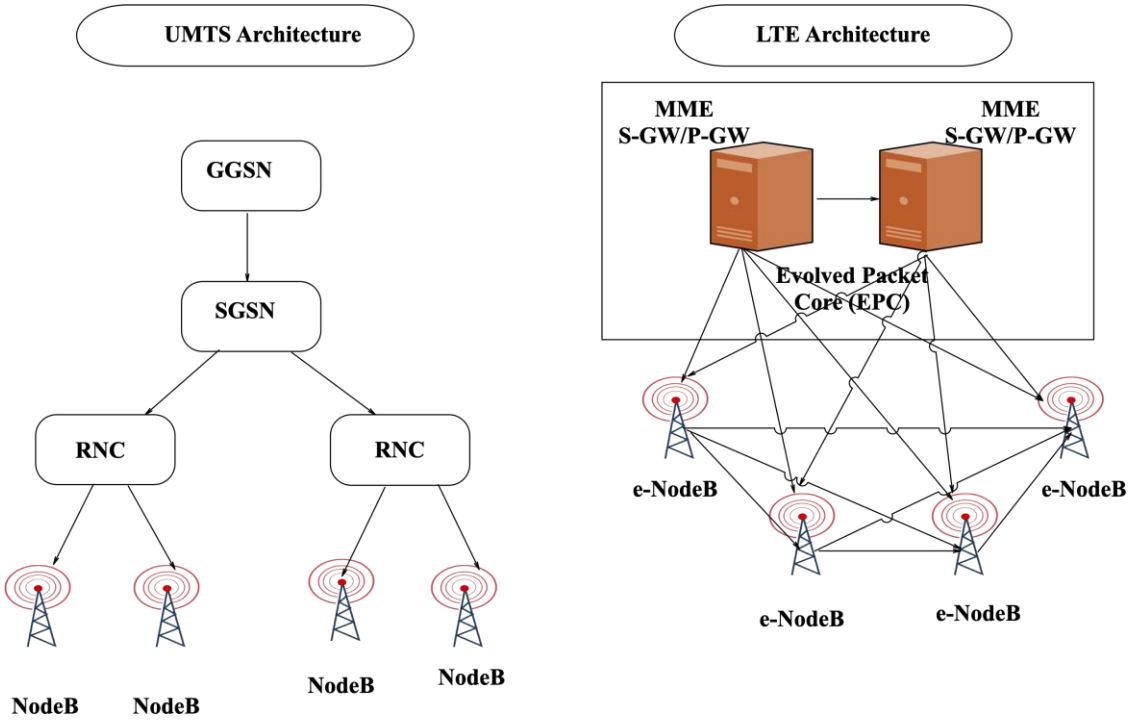

Figure 18 UMTS vs LTE Architecture [7]

LTE's objective is to provide high-data-rate, low-latency and packet-optimized radio-access infrastructure to enable scalable deployments of bandwidth. The system supports flexible bandwidths thanks to OFDMA and SC-FDMA access schemes [7]. In addition to FDD (frequency division duplexing) and TDD (time division duplexing), half- duplex FDD is allowed to support low cost UEs, with UE not required to transmit and receive at the same time.

In previous Generations, the base stations were controlled by a central device. In 2G, it was the base station controller, and in 3G it was the RNC. In LTE [7], this concept was abandoned, as it required significant resources, and the task was concentrated in few network nodes. Most applications on the device only transmit and receive information in bursts with long timeouts in between. During these times of inactivity, the air interface connection to the mobile device has to be changed to use the available bandwidth efficiently and to reduce the power consumption of mobile devices. So, the packet-switched connections generate a lot of signaling load because of

the frequent switching of the air interface state. So, these management tasks were distributed, to speed up the connection setup time and reduce the time required for handover, which is very crucial for real time services. Thus, making the LTE access network a simple flat network of interconnected Base Stations without a centralized controller like RNC, SGSN, and GGSN.

Radio Access Network [7] of the LTE is called Evolved UMTS Terrestrial Radio Access Network (e-UTRAN). The most complex node in the LTE network is the base station, which is also called as e-Node B (evolved-Node B). e-Node B consists of two parts, first is the Remote Radio Unit (RRU) which consists of the antennas responsible for modulation and de-modulations of the signals transferred over the air-interface. Second one consists of Baseband unit (BBU), which consists of the digital modules that processes the signals transmitted and received over the air-interface, also acts as interface to the core network. So, the most evolved part of LTE compared to previous generations is the e-Node B, which provides services of both autonomous unit of Node B of UMTS and the digital module of RNC, which includes Radio bearer controlling, Radio admission controlling, connection mobility control, and scheduling resources. It also provides services such as IP header compression and encryption of user data stream, routing of user plane data to serving gateway, scheduling and transmission of paging resources. e-Node B is connected to Serving Gateway (S-GW) to terminate interface towards the 3GPP radio access network and Packet Data Network Gateway (P-GW) to control IP data services, including routing, allocating of IP address, enforcing policy, and providing access to non-3GPP access network. Further MME performs authentication of users. We can see from the figure that S-GW and P-GW are connected and are called Evolved Packet Core (EPC).

1-6 FOURTH-GENERATION SYSTEMS:

1-6-1 4G System Architecture:

LTE network architecture [7] mainly supports packet-switched traffic providing services such as mobility, Quality of Service (QoS) with very less latency. Packet switching network allows its packets to choose any path to reach their destination unlike the circuit-switching network which requires dedicated path. In 4G architecture, packet-switching technology provided all the services including voice through packet connections. Major change in the 4G architecture includes elimination of Radio Network Controller (RNC), however its functions are embedded within eNB. If there was presence of RNC, then RNC had to control and manage multiple NBs, which would increase latency, it is of not much importance since LTE does not support macro-diversity or soft handoff. In the LTE architecture, all the interfaces are based on IP protocols. As shown in the figure, all the eNBs are interconnected using X2 interface and it is connected to MME/GW via S1 interface. MME/GW consists of two entities, serving gateway(S-GW) and packet data network gateway (P-GW). The S-GW serves UE, performing mobility functions which includes forwarding and receiving packets to and from eNB. The P-GW interfaces to the external data networks such as the Internet and the IMS, performing functions such as IP address allocation, policy enforcement, packet filtering and routing.

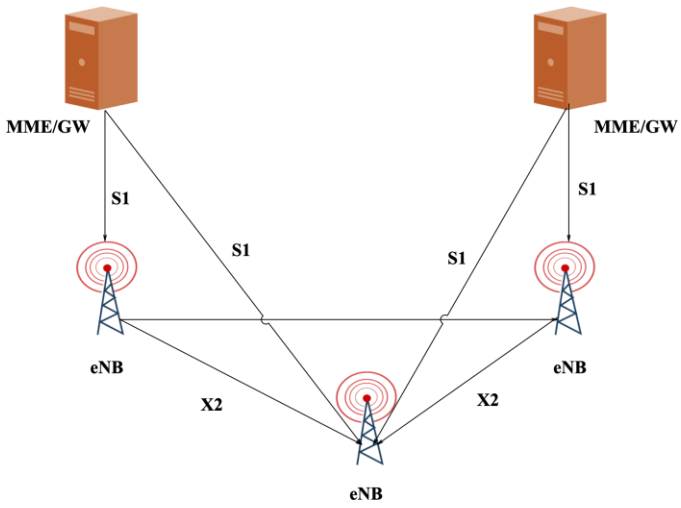

Figure 19 4G Network Architecture [7]

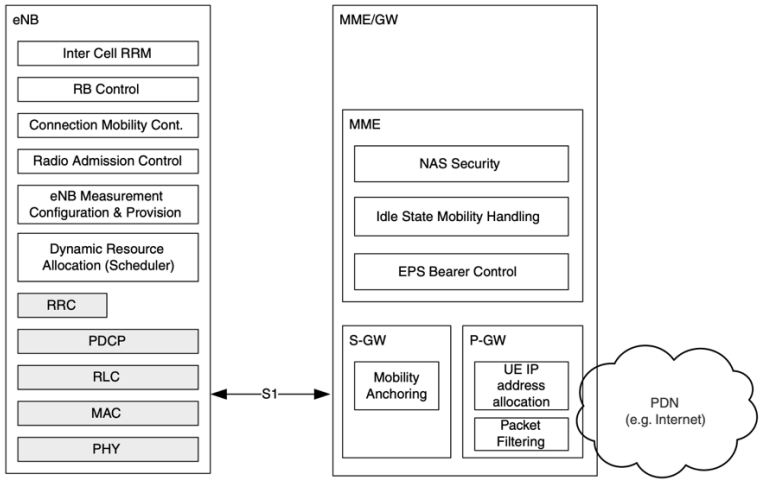

Figure 20 Functional split between eNB and MME/GW [7]

Evolved Node-B implements the services provided by node B in 3G technology and also incorporates the protocols traditionally implemented in RNC. Some of the functions provided by ENB include header compression, ciphering and reliable delivery of data. Also, on the control side it performs radio resource management and admission control [7].

MME is only signalling entity, providing services such as UE reachability, control and execution of paging transmission, tracking area list management, roaming, authentication, authorization, P-GW/S-GW selection, bearer management, security negotiations, and NAS signaling.

The following figures 21 and 22 shows the protocol stack of user plane and control pane.

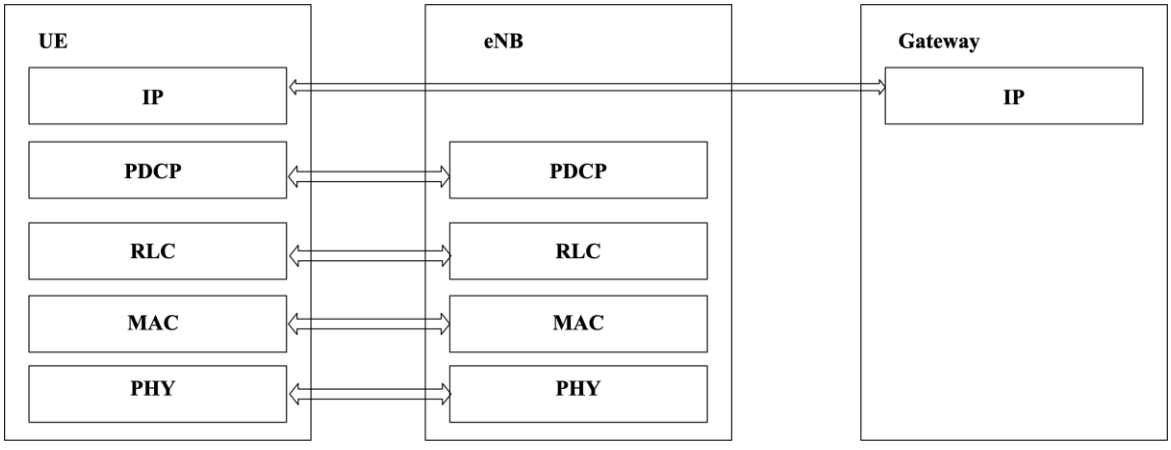

Figure 21 User Plane Control [7]

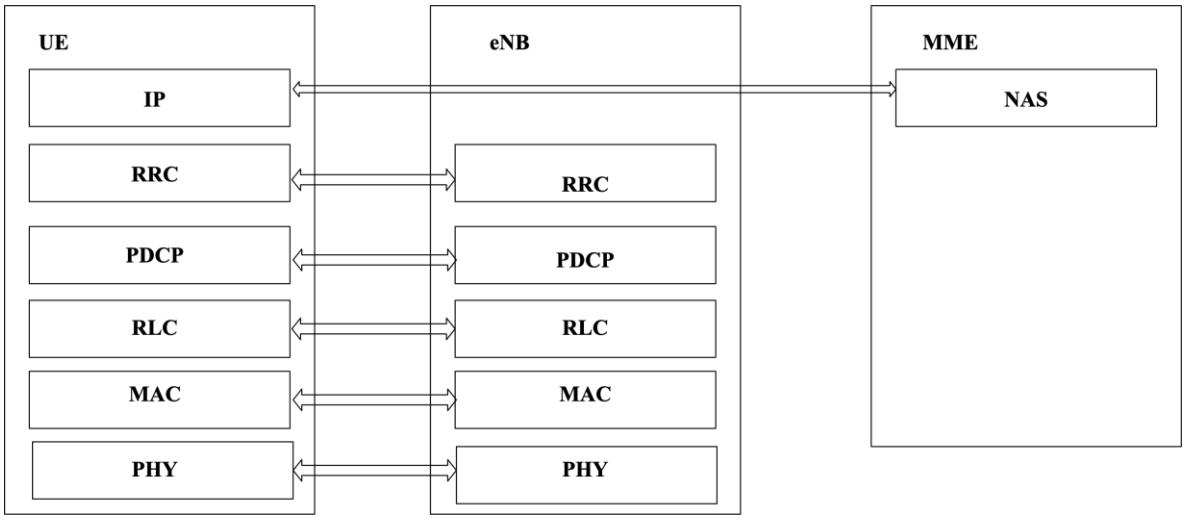

Figure 22 Control Plane stack [7]

From the figure 21 we see that, in user plane protocol stack, packet data convergence protocol (PDCP) and radio link control (RLC) in the previous generations terminates at RNC over the network side [7]. But in LTE the protocols terminate in eNB. In the control plane protocol, RRC functionality traditionally implemented in RNC is now incorporated into eNB. The RLC and MAC layers perform the same functions as they do for the user plane. The functions performed by the RRC include system information broadcast, paging, radio bearer control, RRC connection management, mobility functions and UE measurement reporting and control. The non-access stratum (NAS) protocol terminated in the MME on the network side and at the UE on the terminal side performs functions such as EPS (evolved packet system) bearer management, authentication and security control, etc.

The following diagrams 23 and 24 shows S1 and X2 interface protocol stacks [7]. The S1-U interface between the eNB and S-GW uses GPRS tunneling protocol (GTP-U) on UDP/IP transport layer. It allows non-guaranteed delivery of packet data between each set of terminals. Another S1-MME interface between eNB and MME uses Stream Control Transmission Protocol (SCTP) built over IP. At the transport layer SCTP acts as TCP with the congestion control service. The application layer signaling protocols are referred to as S1 application protocol (S1-AP) and X2 application protocol (X2-AP) for S1 and X2 interface control planes respectively.

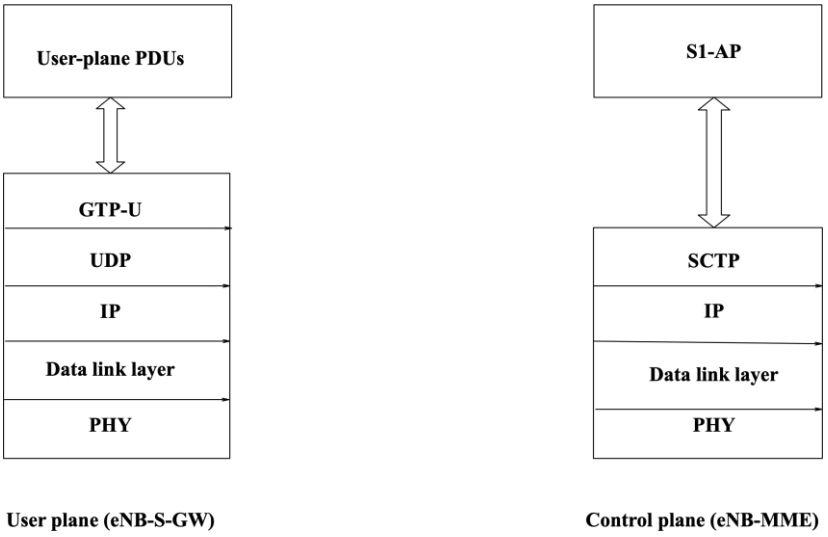

Figure 23 S1 interface user and control planes [7]

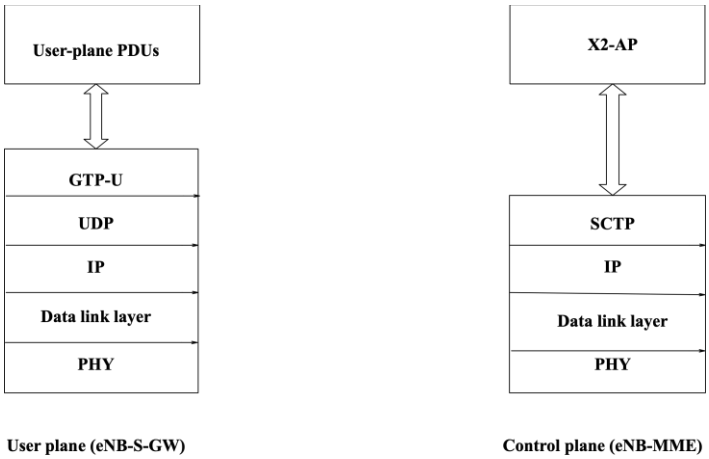

Figure 24 X2 interface user and control planes [7]

In the 4G architecture as given in [7] there was need for QoS for the services such as VoIP, web browsing, video telephony and video streaming. The QoS mechanism in the LTE, enables differentiation of packet flows based on QoS requirements. In the Evolved Packet System (EPS) bearer architecture, there is establishment of QoS flows which is also called EPS bearer between UE and P-GW as shown in the following figure. Further there is a radio bearer between UE and eNB to transport the packets. Each of the IP flow constitutes different EPS bearer and the network can prioritize traffic accordingly. Meanwhile, P-GW when it receives an IP packet from the internet, it classifies the received IP packets based on the predefined parameters, and it send it on

to the right EPS bearer. Based on the EPS bearer, eNB maps packets to the appropriate radio QoS bearer. There is one-to-one mapping between an EPS bearer and a radio bearer.

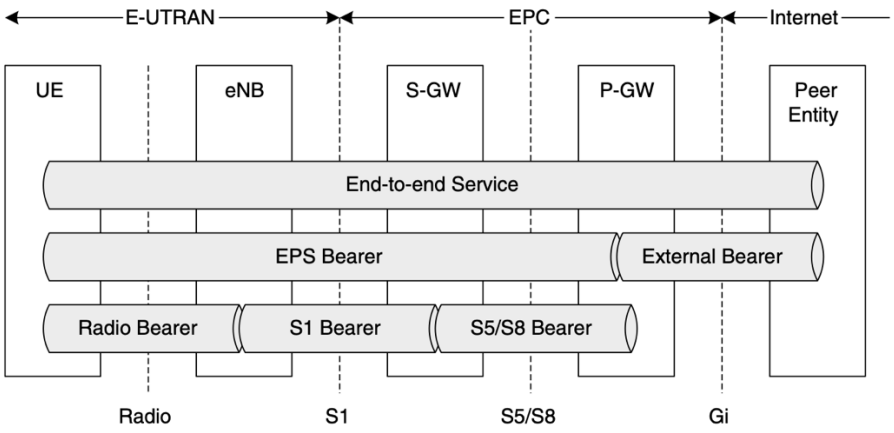

Figure 25 EPS Bearer service Architecture [7]

Following figures 26 and 27 depicts the downlink and uplink layer 2 structures. As seen from the figure, Layer 2 of the LTE consists of three sublayers namely medium access control (MAC), radio link control (RLC) and packet data convergence protocol (PDCP). Functions that are performed by the MAC layer includes mapping between logical (service access point (SAP) between MAC and RLC sublayers) and transport (SAP between PHY and MAC sublayers) channels, multiplexing of RLC packet data units (PDU), padding, transport format selection and hybrid ARQ (HARQ). However, in the downlink structure, MAC layer in addition to handling logical channels of single UE, it also handles priority among UEs [7].

The main services and functions of RLC sublayers include segmentation, ARQ in-sequence delivery and duplicate detection. The reliability of RLC can be configured to either acknowledge mode (AM) or un-acknowledge mode (UM) transfers. The UM mode can be used for radio bearers that can tolerate some loss. In AM mode, ARQ functionality of RLC retransmits transport blocks that fail recovery by HARQ. The recovery at HARQ may fail due to hybrid ARQ NACK to ACK error or because the maximum number of retransmission attempts is reached. In this case, the relevant transmitting ARQ entities are notified, and potential retransmissions and re-segmentation can be initiated. Also, PDCP layer performs header compression and decompression, ciphering and in-sequence delivery and duplicate detection at handover.

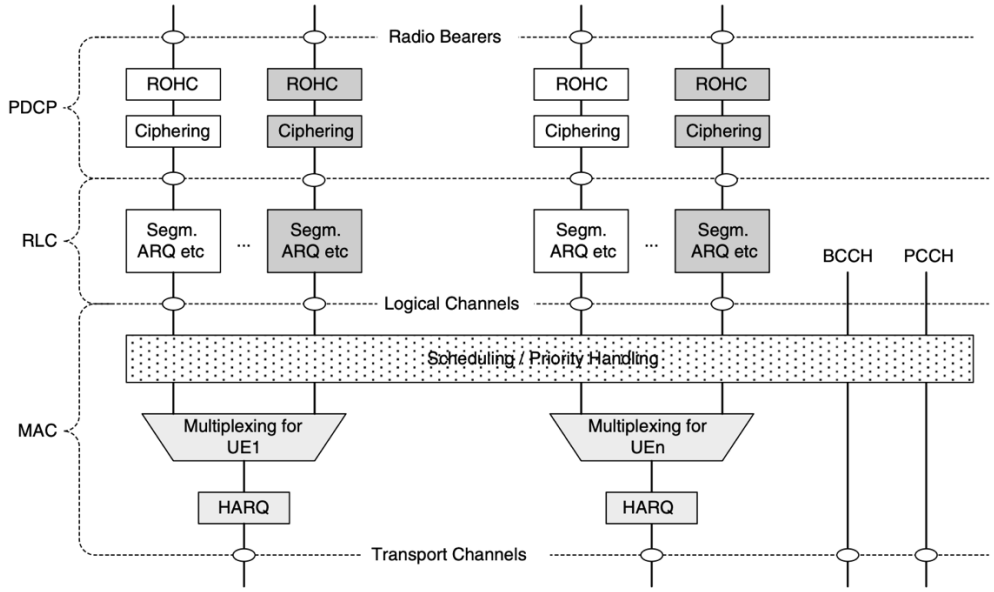

Figure 26 Downlink layer 2 structure [7]

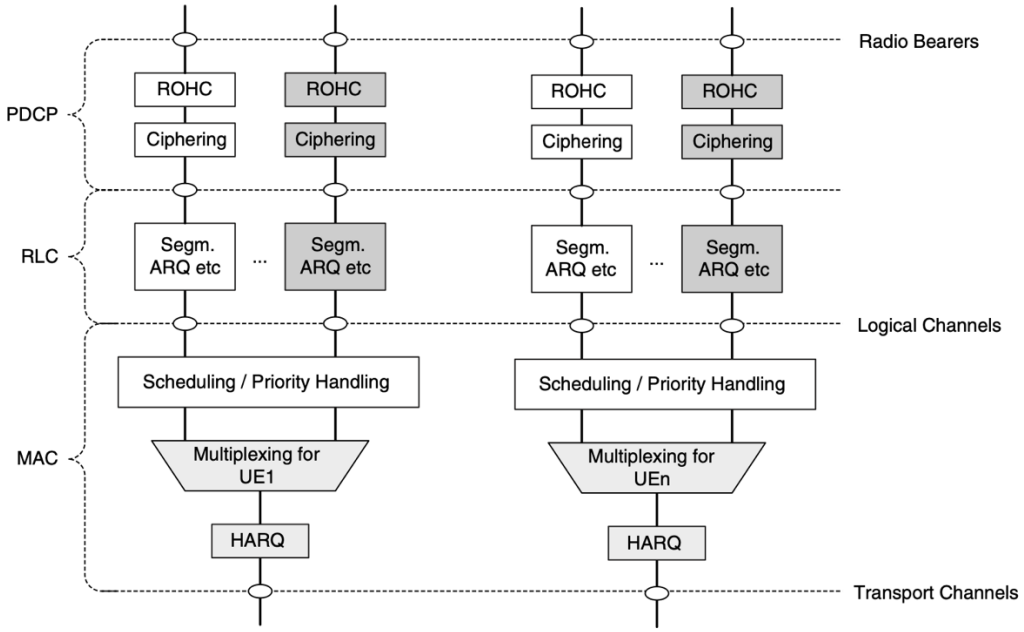

Figure 27 Uplink layer 2 structure [7]

The type of information it holds determines a logical channel [7]. The conceptual channels are further split into channels of control and channels of traffic. Control channels carry information about the control-plane, while traffic channels carry information about the user-plane.

The LTE system in [7] is based on highly simplified network architecture with only two types of nodes namely eNode-B and MME/GW. Fundamentally, it is a flattened architecture that enables simplified network design while still supporting seamless mobility and advanced QoS mechanisms. This is a major change relative to traditional wireless networks with many more network nodes using hierarchical network architecture.

Chapter 2. 5G Architecture implementation and it's technologies.

2-1 Introduction to 5G TECHNOLOGY:

Since, three decades of evolution we have seen constant modifications in the mobile cellular systems, having significantly changed from analog or circuit-based to packet-based communication systems. There are also humungous improvements in speed, bandwidth as well as the number of connected devices. As and when there is increase in the demand for number of users and the urge to meet the requirements of new services and communication trends from various industries such as agriculture, health, automotive, and transport, it required to admonish the huge challenge for 4G technology. Hence, there was development of fifth generation, in order to keep all the devices connected and provide the best possible communication [1].

Most promising and currently evolving fifth generation of communications as in [1] can be defined as "an end-to-end ecosystem to enable a full mobile and connected society. It empowers value creation towards customers and partners, through existing and emerging use cases delivered with consistent experiences and enabled by sustainable business models". Along with the existing Evolved Packet System services, 5G supports more additional services, which we can visualize based on the prospects such as system architecture perspective, spectrum perspective, user and customer perspective. Based on these discernments, we can substantiate why 5G was required over 4G.

From the system architecture perspective, 5G architecture was divided into non-standalone and standalone. Both of these architectures have multiple access technologies, with both the existing LTE's and New Radio (NR) for standalone 5G. In addition, Wireless Local Area Network (WLAN) technologies were augmented [1].

From the Spectrum perspective, since the future communication demands support for variety of use cases, there is demand for dense networks requiring high amount of spectrum. Traditionally, cellular frequency bands of 6GHz were provided, after certain investigations to suffice the on-demanding requirements World Radiocommunication Conference (WRC-15) approved for 24GHz

– 86GHz of spectrum. In addition, WRC-15 also identified a range of new spectrum bands below 6GHz that could be used for mobile services.

From a User and a Customer perspective, after 5G a subscriber could access any device from any location with high data rates, very low latency with very good Quality of Service (QoS). Also, 5G has been augmented with various new technologies such as, cloud computing, Software Defined Network (SDN), and Network Function Virtualization (NFV). All these technologies are assuaged into 5G architecture to create programmable software-centric system. Due to these facilities, one can experience ultra-high definition, 3-dimensional videos and autonomous driving specifications. However, there are large number of use-cases from different projects, organizations, and industrial vertical sectors, depending upon various needs and visions [1].

2-2 5G Use-cases:

There are three major categories of use cases, provided by most of the standardization groups, such as International Telecommunication Union (ITU) and 3GPP. Following are the use-cases of 5G technology [8],

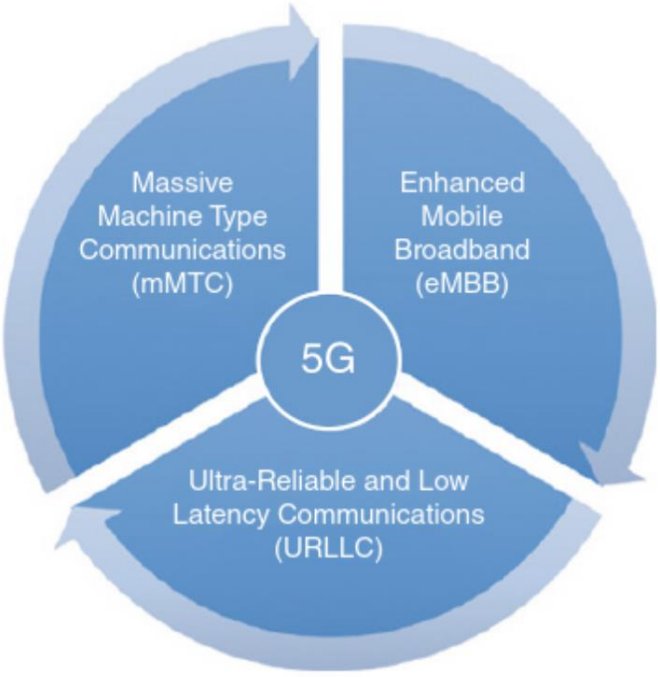

Figure 28 5G use-cases defined by ITU [8]

2-2-1 Enhanced Mobile Broadband (eMBB):

EMBB is a more or less direct development of mobile networks (as it was mentioned above MBB is an initial use case for LTE). Further network throughput could be pushed by 5G and user experience enhanced. The 5G target throughput in the downlink channel is up to 20 Gbps (i.e., from a base station to user equipment). In order to fulfil this requirement, 5G requires new frequency bands to form a channel with a bandwidth of up to 1 GHz. Multi-antenna transmission and beamforming are other enablers of such high throughputs. As most bands are already reserved, there are not so many open bands at low frequencies (<6 GHz). So, 5G networks will be launched using millimeter waves to provide ultra-wide channels (e.g., 28 GHz). There are also a number of free bands in this spectrum that can be assigned to provide wireless mobile services. Although low frequencies (ideally <1 GHz) are used to provide suitable coverage of the network [8].

2-2-2 Massive Machine Type Communications (mMTC):

MMTC is about a massive number of instruments, such as different sensors and metres, monitoring of remote equipment, etc. These devices are usually low-cost (<5$). The very low energy consumption and relatively limited volume of transmitted data are another distinctive feature of such devices. So high rates for these use cases are not needed. While the ability to support multiple devices that rarely transmit small packets (usually delay tolerant) and powerful UE (user equipment) are key building blocks in this area, energy saving features are key [8].

2-2-3 Ultra-Reliable and Low Latency Communications (URLLC):

These use cases are partly (or even mostly) for machine-to-machine communication as well as the above class of use cases. But this class is also about ultra-low latency (<1 ms in one direction) and extremely high reliability, as compared to the previous one. Mechanical automation, traffic safety, and Vehicle to Anything (V2X), robot control and remote medicine are examples that fall into this bucket. There is a particular collection of 5G features defined to address requirements posed by this class of use cases. For example, mini-slot support that allows data to be transmitted within a portion of the slot (aka TTI in LTE) [8]. That reduces the radio transmission time between gNB (base station) and UE on the radio connection (user equipment). In addition, the criteria for data processing time in

gNB and UE are far greater in 5G, i.e., allowed time to process data is much shorter (comparing to LTE).

Some of the requirements of 5G as identified by ITU-R include mobility, peak data rate, network, connection density and spectrum efficiency. The following diagram depicts the requirements of 5G based on user performance perspective and system management perspective [1].

Figure 29 5G key requirements [1]

In order to contribute for the development of next generation system i.e., 5G and to have the best experience in wireless communication, the construction has to suffice two major criteria such as data rate and latency. The data rate specifications are represented in terms of the peak data rate, which is the maximum data rate achievable for a user under ideal circumstances, and the data rate encountered by the user, which is the data rate achievable for a user in the actual network setting. Currently, 4G networks provide subscribers with a nominal peak data rate of 1 G bps, although the user's maximum data rate is about 10 Mbps. The implementation of new bandwidth-hungry technologies and services, such as virtual reality, ultra-high-definition video streaming (e.g., 4 K and 8 K), however, involves an extreme upgrade of 5 G networks relative to existing 4 G networks. The peak data rate is projected to be increased by up to 20 Gbps, while the data rate encountered by the user would increase 100 times over 4G networks and reach up to 1 Gbps [1].

Latency, which is usually expressed as end-to-end latency experienced by the end user, is another essential requirement. Minimizing latency is becoming more important with the emergence of newly defined services such as Tactile Internet, self-driving cars, and automated traffic control, which involve real-time responses and interactions. More precisely, in contrast to the 4G system, the 5G system is supposed to decrease latency 10 times in the user plane, down to 1 millisecond and half in the control plane, down to 50 milliseconds.

Massive networking refers to the need to accommodate a large number of wired devices in an area unit and thus a large or massive number of connections (e.g., connection per square kilometer). The growing number of network devices in the 5G era stems is not only from the introduction of new types of networks and new types of devices, such as sensors, metres, wearable devices and cars, but also from the exponential rise in the number of existing types of devices, such as smartphones and tablets [1].

The 5G system is supposed to accommodate a communication density of up to 1billion connected devices per square kilometer due to the abundance of these smart items, or to put it differently, 100 times more devices compared to 4G system. Network densification can also be expressed, in addition to the number of connected devices, by the traffic density determined by the total number of traffic transmitted by all the devices over the area considered. In the 5G age, the anticipated value for this metric is tens of Gbps per square kilometre. In addition to the need to accommodate a large number of connected devices, the 5G system is also intended to provide mobile consumers with a seamless service experience. However, not all devices and users in the 5G age are phones, so there is no need for seamless mobility. Depending on the types of devices and facilities, on-demand mobility applications should also be sponsored.

Two other important specifications as given in [1] that needs to be guaranteed in the 5G framework are reliability and high availability. In general, a system's reliability refers to its ability to guarantee the success rate of data transmission over a certain period of time under specified conditions (e.g., a latency budget). The reliability rate can vary based on different usage cases and facilities. As previously mentioned, in the third use scenario (i.e., ultra-reliable and low latency communications), there are a range of services and applications, such as public safety, eHealth, automated traffic control, and mission critical services, which require extremely high communication reliability. The 5G infrastructure is supposed to guarantee a reliability rate of up

to 99.999 percent in order to provide these kinds of services. In order to offer end-user services anywhere, the 5G infrastructure must guarantee its availability at all time, which refers to the ability to withstand future failure scenarios. Two network-driven criteria are flexibility and programmability. As the 5G infrastructure would incorporate various technologies to serve a wide number of devices and services, the network architecture should be versatile in order to accommodate a variety of different property-related specifications and attributes that the devices and services are exposed to. The flexibility of the network needs to support various RAN technologies, the capability of scaling the network, also between control plane and user plane, capability of installing new services in short duration, and capability of re-shaping network infrastructure to adapt to change in customer demands. Furthermore, the 5G network architecture should be programmable and identifiable. Indeed, the 5G network system would be built over the same physical infrastructure as a series of separate logical virtualized networks or slices [1].

In addition to improving the ability of the network and improving user experience, the design of 5G must take into account energy and cost efficiency. In particular, a 100-fold increase in energy efficiency compared to today's 4G system is anticipated for the 5G system. In the meantime, in order to guarantee the income of the mobile network, the cost efficiency that reflects the economic component of the 5G system must be improved. Furthermore, as previously mentioned, 5 G will be powered not only by human-centered devices such as smartphones, but also by a large number of "things" such as sensors, smart meters, etc. In order to work in the field, these things must have a much longer battery life, without any extra power source, such as at least 10 years of battery life.

Security is another significant factor that needs to be taken into account in the implementation of 5G, apart from the aforementioned requirements. Indeed, in the 5G era, the proliferation of diversified networks and devices would pose many challenges to ensuring protection. More precisely, 5G protection would need to be guaranteed at various levels, including the level of access, level of infrastructure, and level of service [1].

2-3 5G RADIO ACCESS NETWORK:

The main technologies supporting RAN of 5G include mmWave communication, massive Multiple Input Multiple Output (MIMO), ultra-dense cell, Machine-to-Machine (M2M) and Device-to-Device (D2D) communications, cloud-RAN, mobile edge and fog computing.

2-3-1 mmWave Communication:

One of the main advantages of the 5G system is to provide greater bandwidth, with the highest data rate of about tens of Gbps. More spectrum availability is needed in order to achieve those goals. Present wireless systems, however, usually run in a frequency band ranging from hundreds of MHz (e.g., 700 MHz) to below 3 GHz (e.g., 2.6 GHz). These uses of the spectrum are not necessary for 5G. One of the most effective solutions for expanding the bandwidth range is to exploit the very high spectrum bands, which have not been occupied yet (e.g., > 10 GHz). In particular, several proposed frequency bands above 10 GHz for 5G were accepted during the meeting at the WRC-15 conference hosted by ITU to be examined in advance of the next WRC conference in 2019, such as 24.25-27.5 GHz, 50.4-52.6 GHz, 81-86 GHz, etc. In this context, the best technology nominee is millimeter wave communication (mmWave) [1].

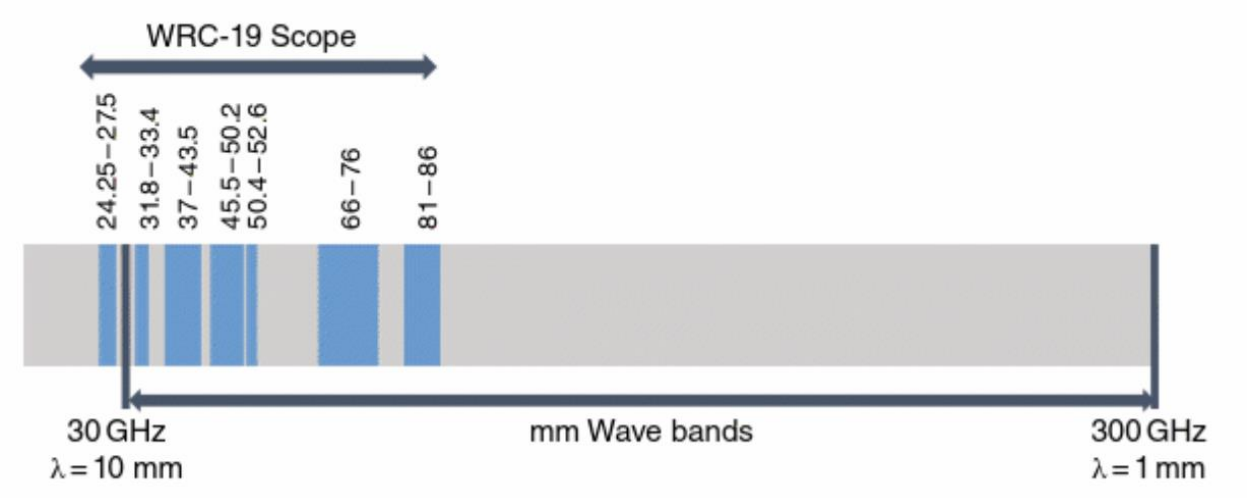

Figure 30 Millimeter-wave bands and potential 5G bands [1]

In 1897, the mmWave analysis was first performed by Jagadis Chandra Bose, referring to the use of frequencies in the 30 to 300 GHz range, with corresponding wavelengths between 10 mm and 1 mm. The mmWave communication has been widely used for indoor environments or backhaul

links due to certain factors, such as high propagation loss. However, several research projects have demonstrated the viability of mmWave technology for 5G mobile networks by implementing several recent developments to create a greater amount of bandwidth in propagation modelling or channel modelling. There are still a range of problems and open concerns that need to be tackled in the future, such as interference and fragmentation, in addition to the advantages of enabling wider bandwidth, higher data rate that makes mmWave a promising 5G technology.

2-3-2 Massive MIMO**:**

One of the most popular options is to densify the number of installed antennas, which refers to a technological solution called massive MIMO [9], in order to meet the 5G criteria in terms of network density and capability enhancement. Fundamentally, MIMO is a wireless communications antenna technology in which many antennas are used to transmit and receive data. In reality, in current 4G networks, the MIMO term has been widely used, referring to multi-user MIMO communication, where a multiple-antenna base station is supported simultaneously by many users; while massive MIMO is characterized as a multi-user MIMO system, where the number of antennas of the base station and the number of users is high. A function like having more antennas at the base station promises to improve the power and density of the network. More importantly, large MIMOs are said to dramatically boost the quality of spectrum and electricity.

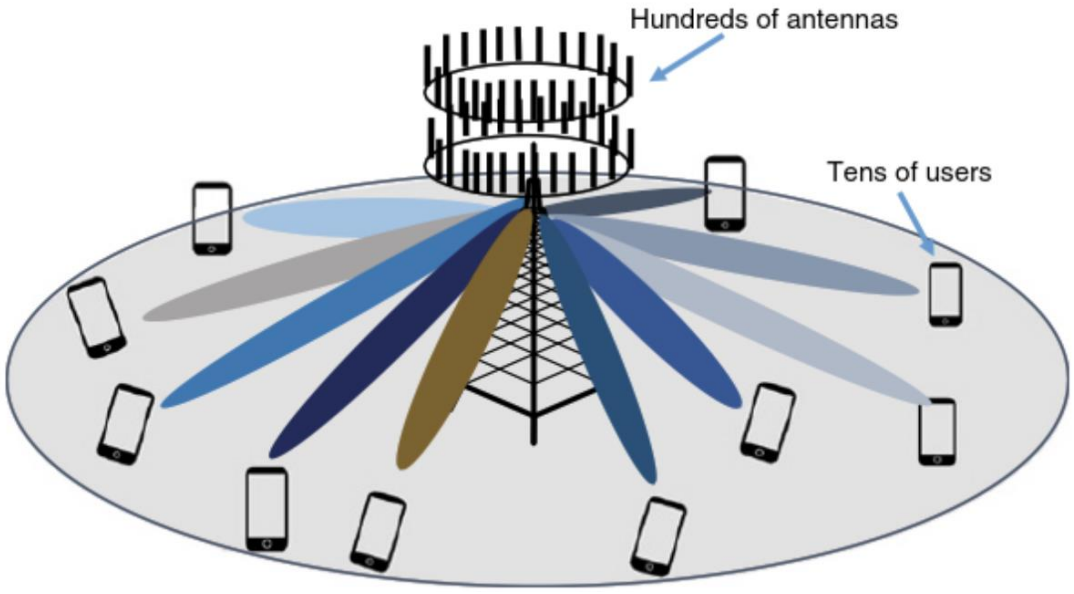

Figure 31 An illustration of massive MIMO concept [9]

These explanations make massive MIMO a relevant technology for 5G. The following figure depicts the huge MIMO design. There are still many research questions that need to be answered, apart from the advantages of massive MIMO, such as plot contamination mitigation, channel estimation, implementation-aware algorithmic architecture, etc. [9]

Very often radio-signals reaching the destinations degrades due to multi-path propagation which reduces link capacity and link reliability. So, in order to cope up with this degradation and enhance link performance we can utilize the concept of spatial diversity, antenna beamforming and massive MIMO. However, in spatial diversity, the principle is to provide the receiver the versions of the same signal which reduces the signal degradation effectively and improves the signal performance. This came be done by time diversity, where same messages are being transmitted at different time slots. Also, frequency diversity, where the same messages are being transmitted at different frequencies. And lastly, space diversity, where is uses separate antennas which are located at different positions to take advantage of different radio paths that exist in a typical terrestrial environment. And hence, this concept is used in MIMO.

Up until the1990's, space diversity was used in the systems that switched between two antennas or combined the signals to provide the best signal. Such systems demanded the high level of processing, but the systems processing was limited. However, with the improvement of the processing power, it was possible to implement MIMO. Between the transmitter and the receiver, the signal can take many paths due to obstacles and the objects in the medium. Traditionally, these multiple paths always introduced interference. By introduction of MIMO, these additional paths can be used to provide additional robustness to radio link by spatial diversity or to increase the link capacity by spatial multiplexing. So, in order to understand MIMO as given in [9], let us consider the following example. Consider a system where the data stream of 10111 is transmitted through a channel. Due to the fluctuation in the channel quality, the data stream might get lost, and the receiver might not recollect it. The solution to combat the rapid channel fluctuation is to add independent fading channels by increasing the number to transmitting antennas or the receiving antennas or both. So, in brief spatial diversity techniques where same information is sent or received across independent channels to combat fading. Here, diversity gain is defined as, Number of transmitting antennas (Tx) multiplied by number of receiving antennas (Rx).

(No. of Tx) * (No. of Rx) = 1 * 1

So, let us increase the number of receivers by one count, the chances of proper delivery of the data stream across the link is very high. Thus, additional fading channel increases reliability of overall transmission link. Here, diversity gain,

Diversity gain = 1 * 2 = 2

Figure 32 Diversity gain of 1*2 MIMO [9]

In this way, more diverse paths can be created by adding multiple antennas can be at the transmitter side as well. And now as per Shannon's channel capacity theorem, there is a limit on the capacity of the link based on the given bandwidth and the S/N ratio of the received signal.

Capacity (bits/sec) = BW * $\log_2(1+S/N)$ [9]

In order to have higher order of modulation scheme, to increase to capacity, we have to have greater bandwidth or higher S/N ratio, which is difficult, expensive and sometimes not compromising. So, another way to improve the data throughput of the individual channels is by using spatial multiplexing. Like previously, each transmitter or receiver will carry multiple antennas and each channel carry independent data, thereby increasing the data rate of the system. There is also antenna beamforming technique with MIMO, where smart antennas are used. In this technique we have phase array system, which has predefined patterns, wherein the required one is switched according to the direction required. Also, adaptive array system uses antennas performing adaptive beam forming, which has infinite number of patterns that can be adjusted to requirements in real-time. Traditional MIMO system have two, four or even eight. However, if the antennas increased up to ten, hundreds or even more, such cases are referred to as massive MIMO system. Massive MIMO system is very popular as it offers increased data rate, increased signal to noise

ratio and channel hardening [9]. So, we can see how robustness of the medium increases due to increase in number of antennas, allowing the spacing of the wavelengths of two beams projected from antennas.

2-3-3 Ultra-dense small cells:

The densification of the number of wireless nodes, which have a wider coverage range than the macro-cell base stations used in the legacy 3G and 4G networks, is another way of increasing network density and enhancing throughput. The scientific solution behind this concept is called the technology of small cells. The small cells' is an umbrella term for operator-controlled, low-powered radio access nodes with a coverage range between ten to several hundred meters, including those operating in licensed spectrum and unlicensed Wi-Fi carrier-grade, as specified by the Small Cell Forum. The following figure shows an example of small cell deployment. The size of the cell is decreased with small cells, which means they put the network far closer to the user, thus better serving high traffic areas such as indoor and hotspot areas. Furthermore, the higher number of low-powered transmission points on the small cell network makes the frequency resource available to be better used, thus improving the spectral quality. In addition, the 5G infrastructure would be designed in a heterogeneous way, where macro and small cells are co-located and maybe linked through wireless backhaul connections to each other, thereby providing increased levels of network capacity via traffic offloading [1].

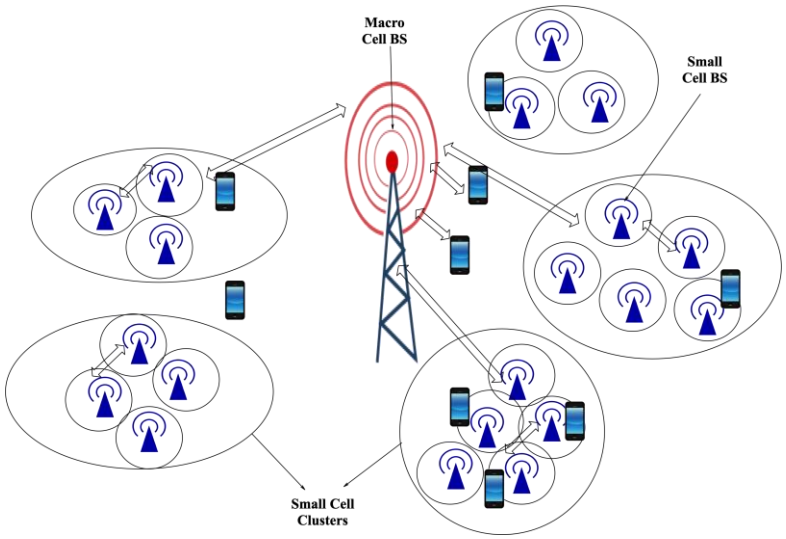

Figure 33 An illustration of small cells deployment [1]

Nevertheless, in terms of intervention and mobility management, the heterogeneity of small cells in the network can pose challenges, thus affecting device efficiency as a whole. In future studies, these problems will need to be discussed. Load balancing, wireless backhauling, mmWave and huge MIMO in small cells, etc., may include some other ongoing research on small cells for 5G [1].

2-3-4 M2M and D2D communications:

Two-thirds of the 5G use case categories would be linked to IoT and Machine Type Communication (MTC), including large and essential communications, as previously mentioned. Therefore, while the idea of M2M or MTC communication was introduced by 3GPP some time ago in 4G LTE systems, it is still regarded as one of the key enablers for 5G. M2M communication fundamentally refers to the automatic communication of data between devices and the underlying infrastructure for data transport. Data can be transmitted between an MTC device and a server, or between two MTC devices directly. As shown in the following figure, there are a range of services and applications allowed by M2M communication, such as monitoring and metering, home and industry automation, healthcare, and automotive. In future M2M related studies, many open issues and problems need to be addressed. For e.g., scalability, privacy and protection, energy conservation, etc. [1]

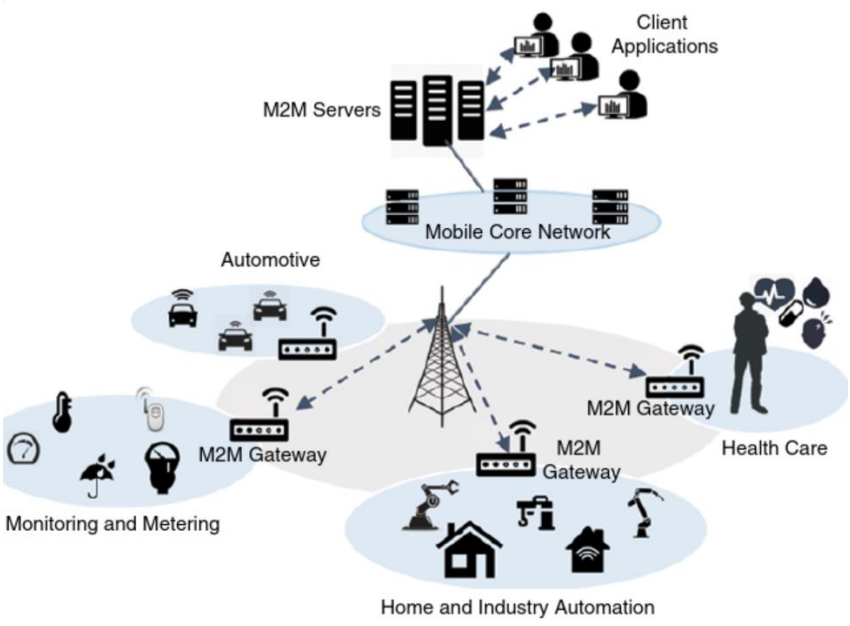

Figure 34 M2M communication [1]

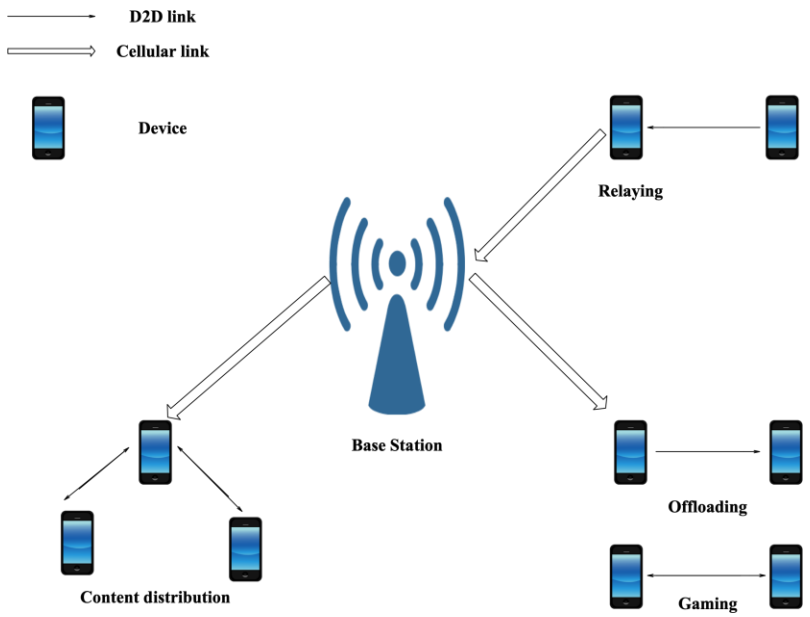

Figure 35 D2D communication [1]

D2D communication refers to direct communication without going through a network infrastructure between two mobile users / devices. It was defined in LTE Release 12 by 3GPP. D2D communication can help increase spectrum performance, user data rate gain, and reduce latency as well as energy consumption by leveraging direct communication between devices, thus being considered as one of the main components of the 5G framework. In general, on licensed cellular spectrum (e.g., LTE) and out-of-band D2D on unlicensed spectrum, the activity of D2D communication can be in-band D2D (e.g., Wi-Fi). As shown in the following figure, there are a range of use cases and application scenarios for D2D, such as proximity-based utilities, gaming, public safety, vehicular communications, and offloading. There are, however, a range of open problems, such as interference management, resource management services and system discovery, protection and privacy, that should be solved in the future. Integration of D2D communication with mmWave and large MIMO technologies will be some other avenues for future study [1].

2-3-5 Mobile edge and fog computing:

As we move to 5G, many of its services and applications will require very stringent latency in the order of milliseconds. One of the most prominent 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. Following figure illustrates the concept of MEC and

its architecture. As specified in [1] the ETSI white paper published 2015, the aim of MEC is to reduce latency, ensure highly efficient network operation and service delivery, and offer an improved 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. Some typical examples of services enabled by MEC are augmented reality, RAN-aware video optimization, connected cars, and IoT, etc.

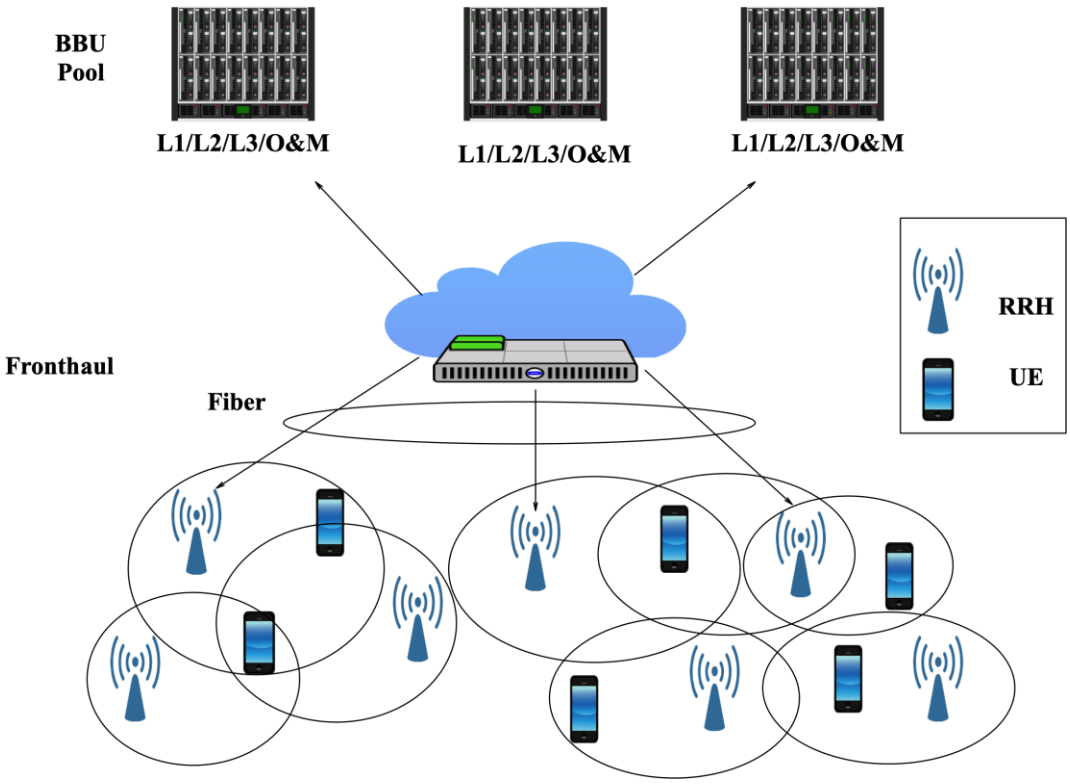

Figure 36 Cloud RAN concept [1]

A similar concept to MEC is Fog Computing (FC) defined by Cisco in 2012, a paradigm in which cloud computing resources are extended to the edge of the network, to create a highly virtualized platform that provides compute, storage, 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. Although both MEC and FC are extremely fit for the development of the 5G system, there are several research issues that need

to be further explored, such as the - interworking between edge clouds, between edge clouds and centralized clouds, mobility management to allow users to seamlessly access edge applications, and other open challenges such as security and performance [1].

Mobile edge computing (MEC) is an emerging technology which has revolutionized conventional solutions for cloud services. Mobile edge computing expands cloud computing by offering capabilities at the edge of the mobile network for processing, storage and networking. Delay-sensitive and context-aware apps can operate in close proximity to mobile users. Furthermore, the cloud services of today are not customised to individual specifications, but rather diversified towards a community of users. Service composition strategies should be implemented to ensure the delivery of user-specific applications on 5G networks. Cloud data is decomposed into a series of files and resources by the proposed solution, which are then distributed into MEC nodes. For quicker access, frequently requested files and resources are further cached on mobile user devices. Both MEC nodes and mobile users advertise their services in the collaborative edge/user space, where either composite or unrendered services are provided as requested by users. Service composition is accomplished through a workflow-net approach based on learning that relies on previous composition outcomes to create models of service composition to be used for new compositions. The presented solution provides guaranteed and fast delivery of the requested cloud composite services to end users while sustaining QoS requirements and load balancing among edge and mobile nodes [1].

2-4 5G CORE NETWORK:

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

2-4-1 Cloud computing:

For more than 20 years, the Internet has been increasing successfully [10]. The demand of growth has so far been met by adding even larger and larger routers. This has been useful for public networks and to scale them up. However, there is current need to implement more powerful packet networks within the metro and aggregation network domain, in order to meet today's increasingly rising demand for Internet connectivity and other packet-based services.

Cloud computing has become a widely used computing model to support cost-effective and efficient data processing using commodity servers. Cloud computing allows efficient use of distributed environments on massive data sets to solve large-scale computing issues. With cloud computing, there are numerous challenges, such as virtualization, isolation, efficiency, scalability, privacy, and protection. We will first include an overview of the cloud computing architecture in this section. Then, we will go further into different technology for virtualization and concentrate on virtualizing the network [10].

An on-demand, pay-as-you-go model of device and storage infrastructure as well as platform services is used by public cloud providers. Amazon Web Services (AWS) led the early cloud computing boom, beginning with their S3 service in 2006. Companies large and small, from backups and archival storage in S3, to EC2 computing, virtual private clouds, IAM authorization and authentication, and RDS managed databases, to name a few, have implemented their services. These services are simple to add, simple to consume, and can contribute to a sprawling, poorly documented system for customers. Cloud computing can be viewed as a layering architecture, as shown in figure 37 [10].

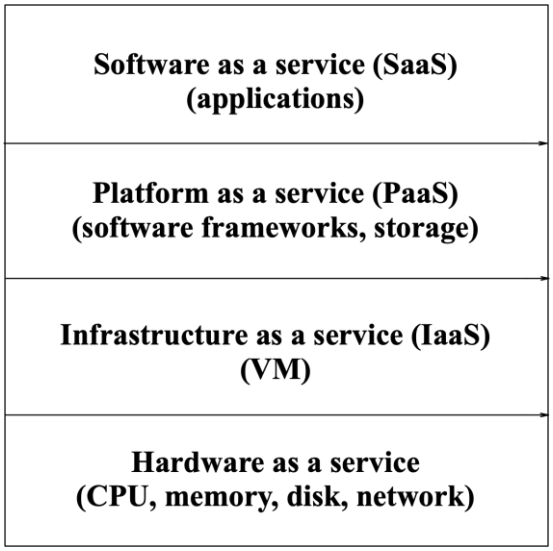

Figure 37 Cloud Computing Architecture [10]

- The hardware layer includes cloud physical infrastructure, that is, hosting facilities, servers, switches, routers, middleboxes for hardware, and support for power and cooling. As a pay-

as-you-go subscription service, the Hardware-as-a-Service (HaaS) model means purchasing IT hardware or parts of data centres. It shares the advantage of dynamically scaling up and down as the demand shifts, similar to other cloud computing layers. The hardware, which consists of thousands of servers in racks, is usually in the form of data centres. Different hardware management issues such as settings, fault tolerance, backup forces, and routine maintenance need to be addressed by the HaaS provider.

- The infrastructure layer is often referred to as the layer for virtualization. The Infrastructure-as-a-Service (IaaS) offers computing resources as a service. Virtualization is an elegant and transparent way to enable time sharing and resource sharing on the common hardware. This encourages consumers to pay as they expand. It also helps to make innovation quicker and to minimize go-to-market time by decoupling the hardware from the upper layer [10].
- The operating system and application frameworks (for example, the Java framework) and other system components are included in the platform layer (e.g., data base and file system). Many popular cloud services operate at this level. For example, Microsoft Azure, Google AppEngine, and Amazon offer APIs for implementing typical web services.
- The Software-as-a-Service (SaaS) model means that the provider offers the software on the common platform as well as the underlying database. This category is being moved by several conventional tech firms (e.g., IBM, Microsoft, and Oracle) and new players (e.g., Salesforce). As the demand shifts, cloud applications will automatically scale.

Compared to conventional computing models, the layering architecture of cloud computing offers a more flexible design. Whenever required to accomplish a particular mission, resources are drawn up-on request. Unneeded resources may be relinquished, and once the job is completed, the assigned resource is withdrawn. Cloud may be classified as private cloud, based on the business model, where the data and processes are handled within the organization, public cloud, where a third-party off-site provider provides/manages services and applications; and hybrid cloud, where both internal and external cloud providers operate [10].

Originally synonymous with public clouds, today cloud computing breaks down into three primary forms: public, private, and hybrid clouds. Public cloud is the most recognizable form of cloud computing to many consumers. In a public cloud, services, usually on a pay-as-you-use model, are

delivered as a service in a virtualized environment, created using a pool of shared physical resources, and available over the Internet. These clouds are more appropriate for businesses who need to rapidly test and improve application code and sell a service, need incremental ability, have less regulatory barriers to resolve, do joint projects, or want to outsource part of their IT requirements. In spite of their proliferation, a range of public cloud issues have emerged, including stability, privacy, and interoperability. Private clouds can be defined in contrast to public clouds. While a public cloud offers resources to multiple customers, a private cloud, as the name implies, ring-fences the resource pool using a shared infrastructure, providing a separate cloud network that can only be accessed by a single entity. Therefore, in a private cloud, on a private network, services and infrastructure are managed. Private clouds have the highest protection and control standard. On the other side, they enable the company to buy and manage its own software and facilities, which decreases cost effectiveness. In addition, to virtualize the business climate, they need a high degree of commitment from both management and IT departments. Such a cloud is tailored to companies that have highly sensitive software, must comply with strict legislation, or must adhere to strict security and data privacy concerns [10].

Both private and public cloud systems comprise a hybrid cloud. It is, therefore, necessary for businesses who want the freedom to switch between them to get the best of both worlds. A company, for instance, can run apps primarily on a private cloud, but rely on a public cloud to accommodate usage spikes. Similarly, by using public cloud services for non-sensitive activities, an enterprise can optimize productivity while relying on a private cloud only when necessary. Meanwhile, they need to ensure the smooth integration of all channels. Hybrid clouds are especially well suited to e-commerce, as their sites need to react on a regular and seasonal basis to fluctuating traffic. The company needs to keep track of many different security platforms on the downside to ensure that they can connect with each other. Regardless of its disadvantages, for many companies, the hybrid cloud seems to be the best choice. Similarly, by using public cloud services for non-sensitive activities, an enterprise can optimize productivity while relying on a private cloud only when necessary. Meanwhile, they need to ensure the smooth integration of all channels. Hybrid clouds are especially well suited to e-commerce, as their sites need to react on a regular and seasonal basis to fluctuating traffic. The company needs to keep track of many different security platforms on the downside to ensure that they can connect with each other. Regardless of its disadvantages, for many companies, the hybrid cloud seems to be the best choice [10].

Following diagram 38 depicts the demonstration of how cloud space and its resources are allocated for variable applications. The goal of the cloud computing is to provide services for various applications. And those applications either run on virtual machines or physical web application servers. At the very bottom of the cloud architecture of data center machines, we have different virtual machines running these services and those virtual machines run on the real servers.

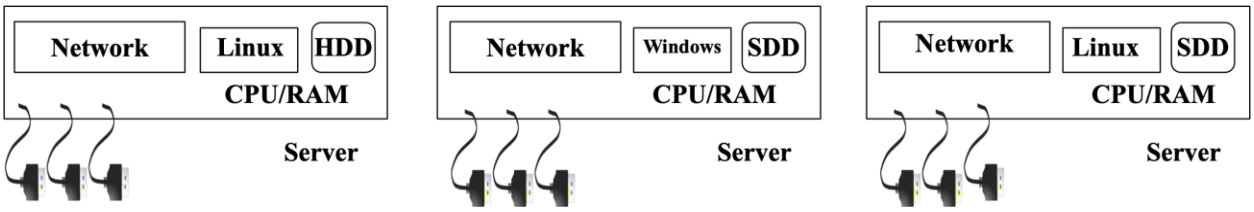

Figure 38 Physical servers on which cloud is deployed [10]

As per shown in the diagram, we have different physical servers connected to the internet or the network. We have the cables connecting to internet, inter-connection of servers and other network components. Apart from the network connection, these servers will be installed with hardware storage disks such as SSD and HDD. And these servers will run operating systems such as windows, Linux or MacOS. Each server will have its own CPU and RAM. And all these servers provided by the cloud will be hosted by those real physical servers but not directly, for the fact if the server wants to create web-application on the physical server, it will be an issue of security also manageability of the storage space and CPU. Hence, the cloud provider will create the virtualization layer as shown in the following figure.

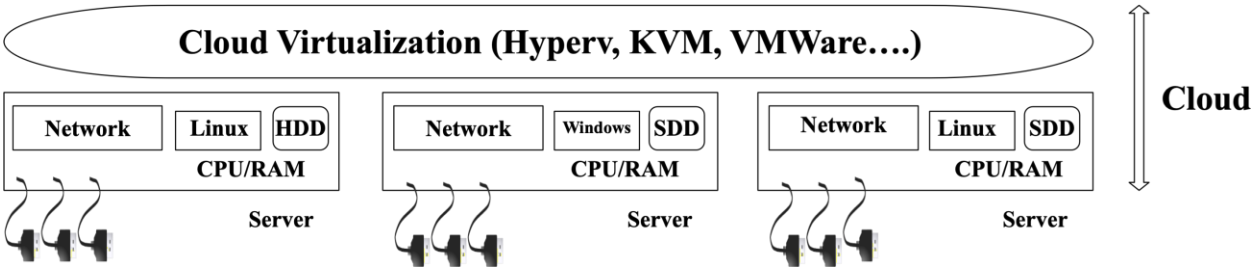

Figure 39 Virtualization layer [10]

So, the virtualization layer sits on top of the physical servers, which is called cloud virtualization. Virtualization manages to provide resources as requested by the virtual machines created to build use-case by the clients. This virtualization layer will manage anything that has to be deployed over the servers including disc, RAM, memory, OS, network and so on. Virtualization layers try to

deploy the virtual machines on the physical server using HyperView, KVM or virtual machines depending on the cloud providers. These two layers constitute the cloud layer.

On top of this architecture comes the virtual machines, which is the basic service of a cloud. Virtual machines will have their own operating system such as Windows, Linux or MAC. They utilize some the memory and CPU of the server that hosts this virtual machine, that has to go through the virtualization layer. The following figure depicts the virtual machine on top of the cloud layer. So, for instance consider an IoT use-case. A sensor has to depict a temperature and when it goes beyond a particular record it has to beep an alarm. In such cases, we can actually rent cloud server from cloud provider say AWS. The server has a virtual IoT machine, that can record temperature and module to compare the temperature with the limit. And another module to beep alarm if it exceeds. This is infrastructure as a service. So, we have provided the customer with a virtual machine with an operating system, CPU, storage and networking. It is the customer's responsibility to update the virtual machine. A virtual machine can be either a web application or even an SQL server. It is customer's responsibility to add resources over the virtual machine that has to be deployed over the service. However, it is difficult for a customer to manage all the resources, such as how much memory has to be applied for the web application based VMs and database VMs. And it is difficult to manage all the servers required for a particular service. So, for every service we have a module where for web applications we have google search or azure and for SQL we have AWS or MongoDB and so on. So, now these various modules can be depicted as platforms provided by the cloud providers itself. Wherein the customer need not have to manage any storage resources, or OS updating or certificate installments, everything will be handled by the cloud. This is called platform as a service. So, now is software as a service provider. This eases the customer more by managing the runtime of the applications. In this case, we only have the website to manage the applications. For instance, email, google drive etc. With all these various services provided we can choose the best to drive our network application. Following diagram depicts in detail view of various services.

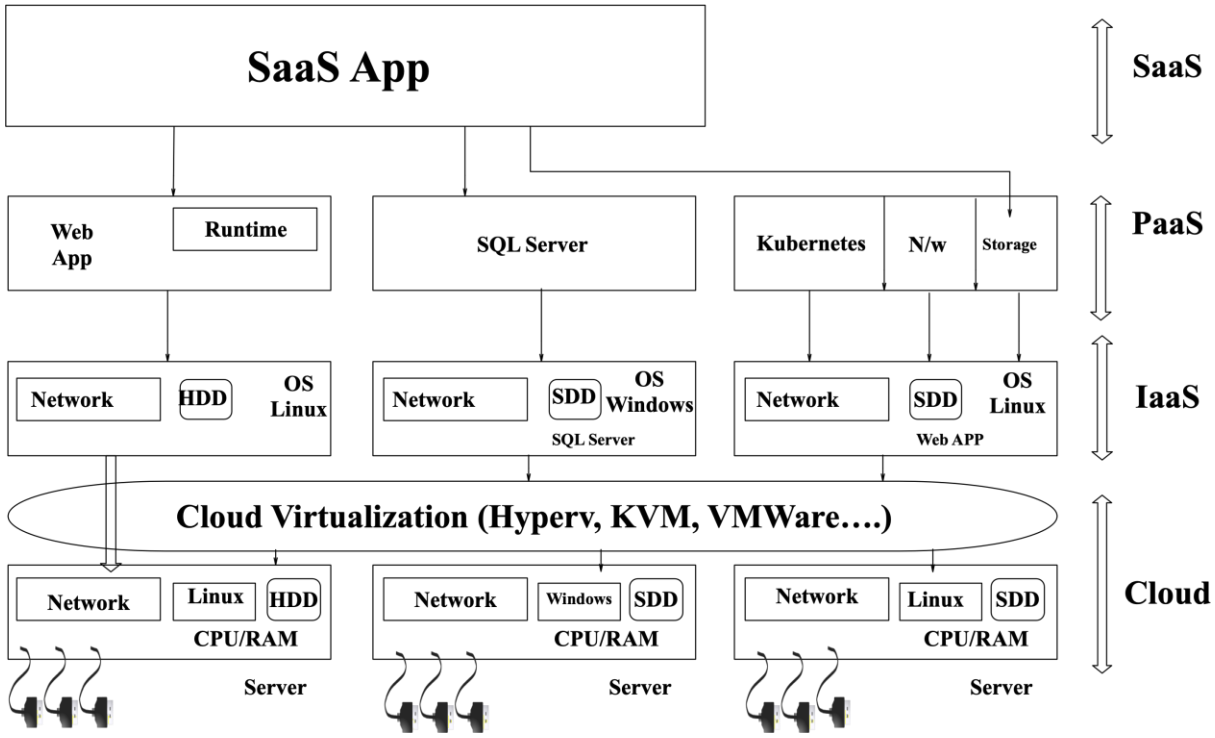

Figure 40 Cloud Services: IaaS, PaaS, SaaS [10]

Some of the key importance of cloud computing include virtualization. Virtualization is a technique that abstracts low-level physical hardware information and provides high-level applications with a clear, virtualized interface. Usually, a virtual machine (VM) refers to a virtualized server. The main enabler of cloud computing is virtualization. It offers the ability to share server clusters as a computing resource pool and the ability to map virtual resources to clients and applications dynamically. We review a few current host virtualization techniques in this section.

Further we have many more virtualization types. One such is a Container as a service (caas). In the virtualization space, containers are a game changer. It can't get simpler than this to launch a new virtualization. Containers are fast, easy to use, and ensure real portability of software. Each container has all of its essential dependencies and configurations for runtime. Containers running on the same engine share the same host server's Linux kernel and are run by Docker rather than hypervisor engines. In terms of capacity, containers are much smaller than virtual machines and have shorter start-up times. Docker containers are therefore viewed as an agile and lightweight solution. While the lightweight OS solutions have existed for a few years, Docket container is the one that leads to the mass adoption and the hype around containization. Docker container was

initially designed as a Go language building runtime. Today, the Open Container Initiative builds the ecosystem around Docker container. Recently, a container standardization is discussed, and a runtime implementation is produced, called runC. The following figure illustrates container architecture. The container does not have a per-VM guest OS, which is normally tens of gigabytes, relative to the conventional VM architecture. In addition to the binaries and libraries required for the applications, each VM instance has a complete guest OS image. It normally leads to high use of memory and disc and therefore a sluggish start-up time. In containers [10], by comparison, each application runs as an independent process in the host operating system's user space, sharing the kernel with other containers. Generally, each programme is in the order of megabytes, which is much lighter in weight.

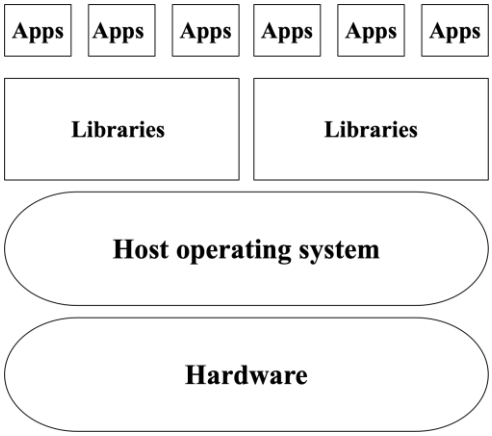

Figure 41 Docker container Architecture [10]

The main technologies behind container are namespace, control group, and union file system (UFS). Each container runs a set of namespaces, which provides the isolation. Each container can only access its own namespaces. The control group is used to set up access control to hardware resources for each container. The UFS provides lightweight and fast layering, which is the building block of container. Docker container builds on top of these techniques into the container format in the form of libcontainer. In addition, Docker also supports traditional Linux container LXC.

The entire Docker stack in a clustered dockerized environment contains five major layers: Cluster Manager layer, Node (or host) layer in the cluster, Docker Demon layer that runs on each node, Container’s layer created by the Docker Demon, and Applications layer that contains applications run on each container. Different Docker cluster managers are structured in different forms, for

example, on Swarm, the containers are part of its structure, but on Kubernetes, there is only a Pod that represents a set of containers, and the user can control the Docker containers just through the cluster Pods [10].

2-4-2 Network virtualization:

It refers to the technology that allows a set of network resources to be partitioned or aggregated and displayed to different users in a way that each user experiences an independent and specific view of the physical network. Core resources (i.e., links and nodes) or derived resources can be part of the abstraction of network resources (topologies). This technology can virtualize a connection (physical channel, data path, etc.) or a network to a network device (e.g., a router or NIC) [11].

Typically, the tunnelling device is agnostic at the endpoints. Usually, it uses a simple physical network infrastructure forwarding technique, such as shortest path routing, to allow packets in the tunnel to reach their intended VTEP. Usually, the network policy and access control are carried out on the virtual network layer. The mappings between hosts and VTEPs are maintained either centrally or spread across VTEPs in a distributed manner. Depending on the various forms of virtual networks, such information may be retained on edge switches or end-host virtual switches. Physical network routing and maintenance are isolated from virtual networks. The mappings between hosts and VTEPs are maintained either centrally or spread across VTEPs in a distributed manner. Depending on the various forms of virtual networks, such information may be retained on edge switches or end-host virtual switches. Physical network routing and maintenance are isolated from virtual networks. Typically, a separate administrative domain maintains it. The layout of the physical network is generally more stable. This separation allows the management of multiple layers in parallel. Without disputes, distinct strategies may be implemented. However, because of the separation, the events, state changes, and faults on the physical network cannot be conveyed to the virtual network in real time, and vice versa [11].

Digital network traffic is not supplied with enhanced facilities that are usually accessible from the physical network. Furthermore, classic network management services for FCAPS (Fault/Configuration/Accounting/Performance/Security) are not interrelated between virtual and physical domains. For example, overlay tunnel endpoints that trigger very large flows to the virtual network overlay have no knowledge of the underlying topology of the network and are unable to

optimally route the data flow or provide differential treatment quality of service (QoS) and can therefore create congestion that could otherwise be avoided. In the event of a connection breakdown on the physical network that impacts an overlay tunnel carrying virtual network traffic, another example is troubleshooting. Without a federation between control systems, it becomes very difficult for virtual network system administrators to troubleshoot or locate the root cause of the problem.

2-4-3 Network Function Virtualization:

NFs are implemented in non-virtualized networks as a mixture of vendor-specific software and hardware, also referred to as network nodes or elements of the network. For the various stakeholders in the telecommunication network system, Network Functions Virtualisation [11] represents a step forward. As such, in contrast to present practise, NFV incorporates a variety of differences in the way network service provisioning is realised. To sum up, it is possible to list these differences as:

- Software decoupling from hardware: Since the network component is no longer a set of interconnected hardware and software entities, the creation of both is independent of each other. This enables the software to progress separately from the hardware, and vice versa [11].
- Flexible implementation of the network function: the separation of software from hardware allows to reassign and share infrastructure resources, so that various tasks can be performed at different times together, hardware and software. The actual machine instantiation of the network feature will become more automatic, assuming that the pool of hardware or physical resources is already in place and mounted at certain NFVI-PoPs. The numerous cloud and network technologies currently available are leveraged by such automation. This also allows network operators to deploy fresh network services more efficiently over the same physical platform.
- Dynamic operation: The decoupling of the network function features into instantiable software components allows greater flexibility to scale the actual VNF output in a more dynamic manner and with finer granularity according to the actual traffic that the network operator requires to supply power, for example [11].

The implementation of NFs as software-only entities that operate over the NFV infrastructure is envisaged by Network Functions Virtualisation (NFVI). The figure below shows the high-level architecture for NFV. As such, in NFV, three primary working domains are identified:

- Virtualized Network Function, as the implementation of a network function by software that can operate over the NFVI.
- NFV Infrastructure (NFVI), including physical resource diversity and how to virtualize them. NFVI encourages the introduction of VNFs.
- NFV Management and Orchestration, which includes the orchestration and lifecycle management of infrastructure virtualization supporting physical and/or software resources, and the lifecycle management of VNFs. In the NFV context, NFV Management and Orchestration focuses on all virtualisation-specific management tasks needed.

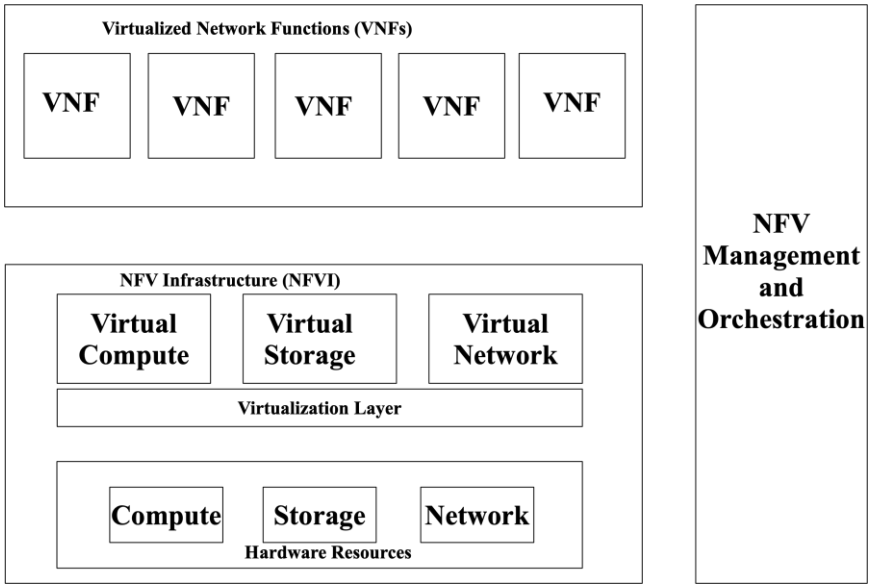

Figure 42 High-level NFV Framework [11]

The NFV framework enables dynamic construction and management of VNF instances and the relationships between them regarding data, control, management, dependencies and other attributes. To this end, there are at least three architectural views of VNFs that are centred around different perspectives and contexts of a VNF. These perspectives include:

- A virtualization deployment/on-boarding perspective where the context can be a VM.

- A vendor-developed software package perspective where the context can be several inter-connected VMs and a deployment template that describes their attributes.
- An operator perspective where the context can be the operation and management of a VNF received in the form of a vendor software package [11].

A network service can be perceived architecturally by endorsing network infrastructure as a forwarding graph of Network Functions (NFs) interconnected. These network functions may be implemented over a single network of operators or over an interface between different networks of operators. The behaviour of the underlying network function relates to the behaviour of the higher-level operation. Therefore, the behaviour of the network service is a mixture of the behaviour of its constituent functional blocks, which may include individual NFs, NF Sets, NF Forwarding Graphs, and/or network infrastructure. The network service's end points and network functions are represented as nodes and correspond to computers, software, and/or physical server applications. An NF Forwarding Graph as given in [11] may have network function nodes linked by unidirectional, bidirectional, multicast and/or broadcast logical links. A chain of network functions is a simple instance of a forwarding graph. A mobile, a wireless network, a firewall, a load balancer and a collection of CDN servers can be an example of such an end-to-end network operation. Inside the operator-owned resources is the NFV area of operation. Consequently, as an operator does not exert its authority over it, a customer-owned unit, e.g., a cell phone, is beyond the reach. Nonetheless, virtualization and network-hosting of client features is feasible and is within the scope of NFV.

The following figure 43 shows the representation of an end-to-end network service that includes a second nested NF Forwarding Graph in the middle of the figure interconnected by logical links as shown by the network function block nodes. The end points are connected via network infrastructure (wired or wireless) to network functions, resulting in a logical interface between the end point and the network function. In the figure with dotted lines, these logical interfaces are depicted. The outer end-to-end network service is composed of End Point A, the inner NF Forwarding Graph, and End Point B in the diagram below, while the inner NF Forwarding Graph is composed of NF1, NF2 and NF3 network functions. These are linked via the logical links provided by the Infrastructure Network 2 [11].

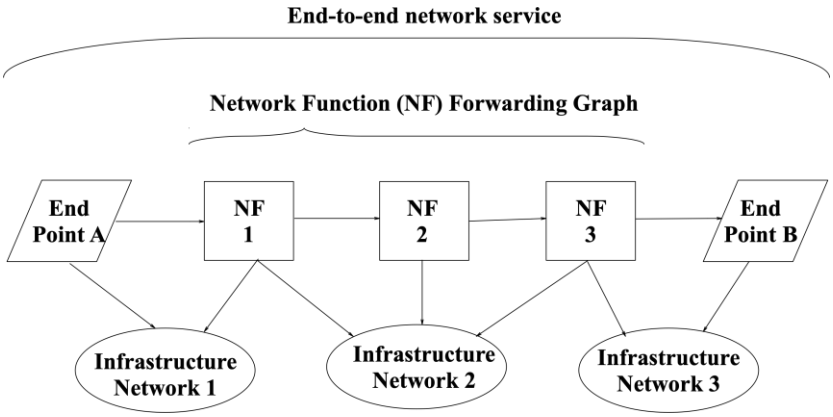

Figure 43 Graph representation of an end-to-end network service [11]

An example of an end-to-end network service and the various layers that are involved in its virtualization process is shown in the following figure. In this example, only VNFs and two end points can be composed of an end-to-end network service. A virtualisation layer realises the decoupling between hardware and software in the virtualisation of network functions. This layer abstracts the NFV Infrastructure's hardware resources. As shown in the figure below, the NFVI-PoPs involve computing, storage and networking services deployed by a network operator [11].

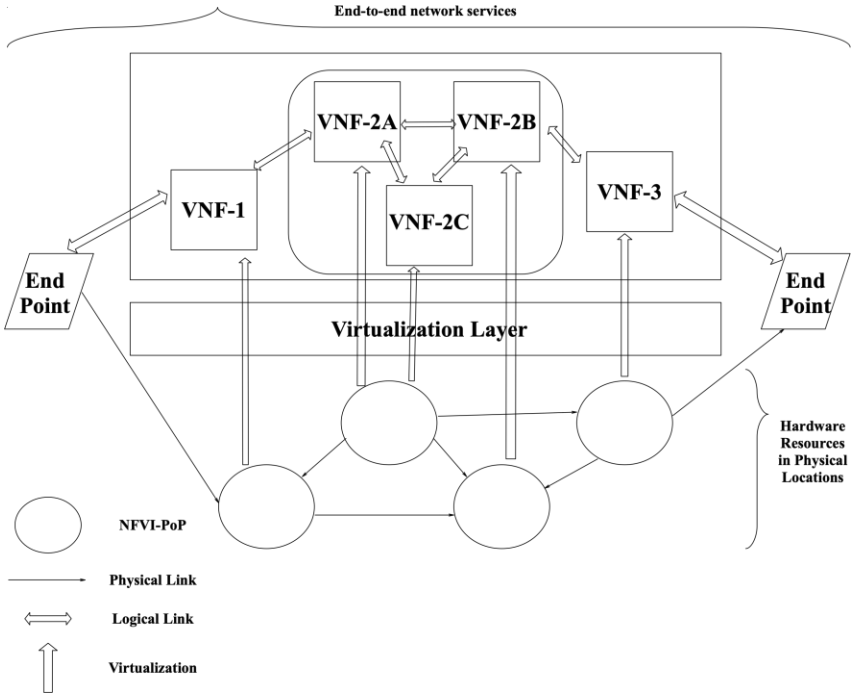

Figure 44 End-to-end network service with NFVs and nested forwarding graph [11]

Virtualized network functions, as shown by the arrow labelled "virtualisation" run on top of the virtualization layer that is part of the NFVI. As shown in the following figure, the VNF Forwarding Graph (VNF-FG) corresponding to the previous figure's network function forwarding graph is shown. The example of a nested VNF-FG (i.e., VNF-FG-2) constructed from other virtualized network functions is also shown in the figure (i.e., VNF-2A, VNF-2B and VNF-2C). The goal is to base the interfaces in a multi-vendor setting between NFs and/or VNFs and the infrastructure on agreed standards (e.g., standardised by an SDO, and/or open de-facto standard) [11].

NFV [11] stresses that the exact physical implementation of a VNF instance on the infrastructure is not observable from the E2E service point of view, with the exception of ensuring clear policy restrictions (e.g. position knowledge required for the deployment of a virtualized CDN cache node (see the use case CDN virtualization (vCDN) in ETSI GS NFV 001) or ensuring that redundant infrastructures are deployed on the infrastructure (see the use case CDN virtualization (vCDN) in ETSI GS NFV 001) This allows the deployment of a VNF instance on various physical resources, such as computing resources and hypervisors, and/or to be geographically distributed as long as its overall end-to-end service efficiency and other policy constraints are met. In any case, VNF instances and their supporting infrastructure need to be visible for configuration, diagnostic and troubleshooting purposes.

The NFV architectural framework [11] identifies functional blocks and the main reference points between such blocks. Some of these are already present in current deployments, whilst others might be necessary additions in order to support the virtualization process and consequent operation. The functional blocks are:

- Virtualised Network Function (VNF).
- Element Management System (EMS).
- NFV Infrastructure, including hardware and virtualized resources.
- Virtualisation Layer.
- Virtualised Infrastructure Manager(s).
- Orchestrator.
- VNF Manager(s).
- Service, VNF and Infrastructure Description.

- Operations and Business Support Systems (OSS/BSS).

The NFV architectural structure [11] that depicts the functional blocks and reference points in the NFV framework is shown in the following figure. Solid lines indicate the key (named) reference points and execution reference points and are within the scope of the NFV. These are future standardisation targets. In current implementations, the dotted reference points are available but may require extensions for virtualization of network feature handling. The dotted reference points, however, are not currently the principal subject of NFV. The architectural framework shown focuses on the functionalities required for the virtualization of the network of an operator and the consequent operation. It does not specify which network functions should be virtualised, as that is solely a decision of the owner of the network.

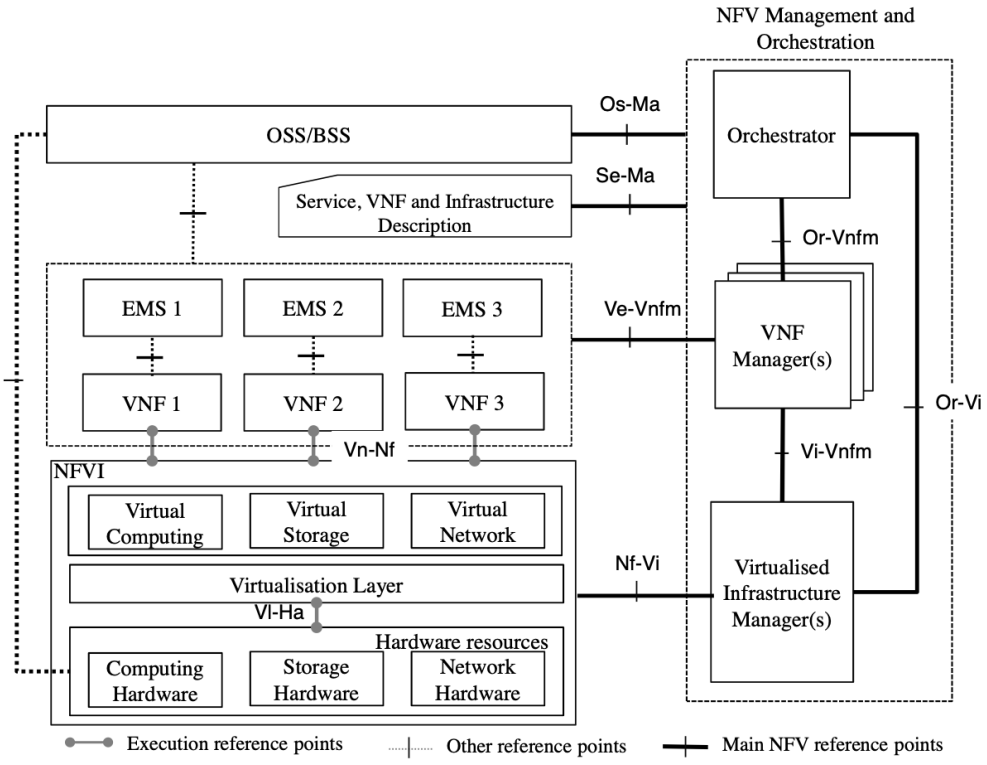

Figure 45 NFV reference architectural framework [11]

In a legacy non-virtualized network, a VNF is a virtualisation of a network element. 3GPPTM Evolved Packet Core (EPC) network components are examples of NFs, e.g., Entity for Mobility Management (MME), Serving Gateway (SGW), Packet Data Network Gateway (PGW); home

network components, e.g., Residential Gateway (RGW); and traditional features of the network, e.g., Servers, firewalls, etc. Dynamic Host Configuration Protocol (DHCP). A list of use cases and examples of target network functions (NFs) for virtualization is given in GS NFV 001. The functional behaviour and condition of an NF is largely independent of whether or not the NF is virtualized [11].

It is assumed that the functional behaviour and the external operating interfaces of a PNF and a VNF would be the same. Multiple internal components may consist of a VNF. One VNF, for example, can be distributed over multiple VMs, where a single portion of the VNF is hosted by each VM. In other instances, however, the entire VNF may also be deployed in a single VM. The precise methods of implementation are beyond the reach of the current text.

- The physical hardware resources in NFV include computing, storage and networking that provide VNFs with processing, storage and communication through the virtualization layer (e.g., hypervisor). As compared to purpose-built hardware, computing hardware is believed to be COTS. It is possible to distinguish storage capacity between shared network attached storage (NAS) and storage that resides on the server itself.
- The middle virtualization layer ensures that VNFs are decoupled from hardware resources and the program can therefore be deployed on various physical hardware resources. This kind of flexibility is usually given in the form of hypervisors and virtual machines for computing and storage resources (VMs). It is envisaged that a VNF will be deployed in one or more VMs [11].
- According to NFV, the management of virtualized infrastructure involves the functionalities used to control and manage a VNF's relationship with, and virtualization of, the computing, storage and network resources under its jurisdiction.
- The Orchestrator is in charge of the orchestration and management of NFV infrastructure and software resources and realizing network services on NFVI.
- A VNF Manager is responsible for VNF lifecycle management (e.g., instantiation, update, query, scaling, termination). Multiple VNF Managers may be deployed; a VNF Manager may be deployed for each VNF, or a VNF Manager may serve multiple VNFs.

- This dataset provides information regarding the VNF deployment template, VNF Forwarding Graph, service-related information, and NFV infrastructure information models.
- OSS/BSS in the figure refers to OSS/BSS of an operator.
- This reference point interfaces the virtualization layer with hardware resources in order to create a VNF execution environment and collect relevant information on the hardware resource state for VNF management without depending on any hardware platform.
- This reference point illustrates the execution environment supplied to the VNF by the NFVI. It does not presume any particular protocol for power. In order to ensure hardware-independent lifecycle, VNF performance and portability specifications, it is within the reach of NFV.
- Requests related to services, e.g., permission, validation, booking, allocation, by the VNF Manager (s). Sending configuration details to the VNF manager so that the VNF can be properly configured to operate within the network service's VNF Forwarding Graph. Set of VNF state information required for lifecycle management of network service.
- Resource allocation requests by the VNF Manager. Virtualized hardware resource configuration and state information (e.g., events) exchange [11].
- Resource reservation and/or allocation requests by the Orchestrator. Virtualized hardware resource configuration and state information (e.g., events) exchange.
- Specific assignment of virtualized resources in response to resource allocation requests. Forwarding of virtualized resources state information. Hardware resource configuration and state information (e.g., events) exchange.
- Requests for network service lifecycle management. Requests for VNF lifecycle management. Forwarding of NFV related state information.

Network functions as given in [11] are well-defined; thus, technical requirements record both their functional actions as well as their external interfaces. A VNF is a software package in NFV that implements certain functions of the network. The VNF software architecture itself needs more research, apart from the high-level NFV architecture perspective. Virtualization gives us the ability for traditional monolithic NFs to design modular and slimmer applications. A VNF for scalability, reusability, and/or faster response can be decomposed into smaller usable modules. Alternatively,

to decrease management and VNF Forwarding Graph complexity, several VNFs can be composed together.

2-4-4 Software Defined Networking:

In order to minimise the total cost of ownership (TCO) and boost average revenue per user (ARPU) and customer retention, the massive growth in subscribers, devices, applications, and traffic has imposed new challenges for service providers. To overcome some of these emerging problems, the SDN model emerged. SDN decouples the planes for control and forwarding. So that each plane can scale independently to decrease TCO. A collection of open APIs is supported by SDN so that network programmability has been implemented. It's possible to allow new services and creativity. SDN has been used to provide network virtualization for resource management with versatility in network access [12].

SDN can be built in three ways from a technological perspective,

- First, it can help to independently scale up the control and data plane. Independent scaling can help handle the increase in bandwidth. The increase in bandwidth comes from various kinds of applications. For example, there is a need for data plane scaling for the dominant traffic video applications. But scaling the control plane calls for machine-to-machine applications or VoIP traffic. In next-generation networks, these multiple forms of traffic will suggest independent control and data plane scaling to meet next-generation traffic growth and demands [12].
- Second, it can help increase the pace of operation and creativity. For service providers to use the network as a portal to extend their service models to include third-party applications, the network must be programmable. This allows the existing business models to grow. For the variety and quality of new network services, rapid development and deployment of not only in-house services but also third-party applications from other content suppliers is critical.
- Third, it can allow for network virtualization that is versatile and effective. Service providers typically deploy an overlay network for each type of network while running various network services, such as virtual private networks (VPNs) and video networks. They have one network for wireline applications, another for wireless applications, and

another for business applications, for example. As we move toward converged networks, by splitting this physical network into several virtual networks, one for each form of application, the aim is to build one network for several services. In view of the substantial interest due to the main drivers, SDN has gained a growing amount of publicity. Until now, however, there has been no consensus on SDN's principles and definitions. In the following, we take the initiative to outline the core aspects of SDN from the viewpoints of service providers and the four main principles defining the SDN in networks of service providers. Following figures illustrates the overview architecture and four concepts and the components in each layer in an SDN architecture [12].

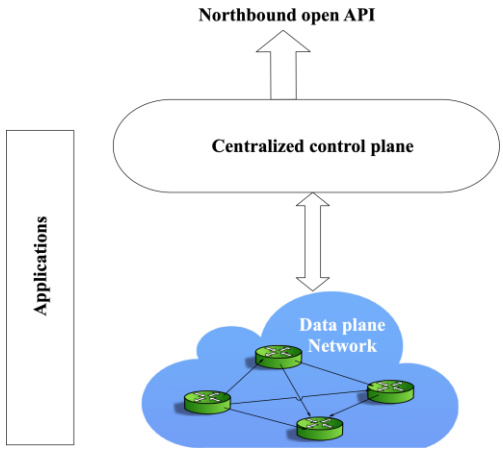

Figure 46 SDN Architecture [12]

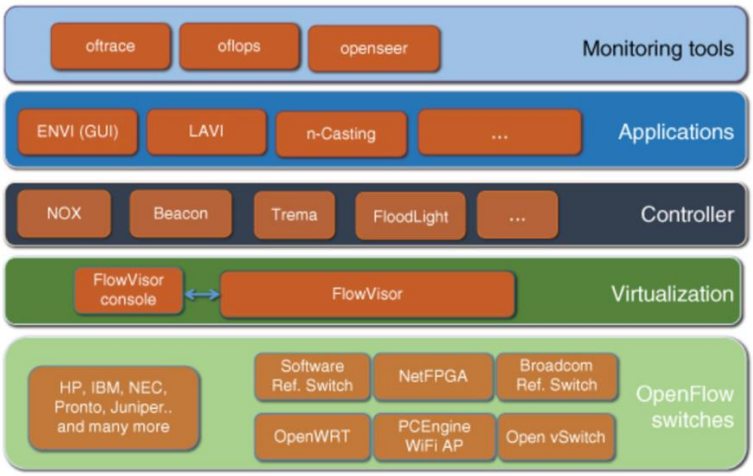

Figure 47 SDN components [12]

One of the main concepts of SDN is the isolation and centralization of the control plane program from the packet-forwarding data plane [12]. This varies from the conventional architecture of the distributed system in which the software of the control plane is distributed across all the network data plane devices. The transition from today's integrated box view to the view of separation is seen in the following figure. Centralized control in this architecture allows new services and applications to be implemented that are not feasible or quite complicated in conventional distributed networks. Compared to updating a whole networking system, the deployment of new services is much quicker as it is performed by current vendors. Independent and simultaneous optimizations of the two planes are made possible by the separation of control and data planes. We envisage that this route would lead to highly advanced and cost-effective high-performance data plane devices and control plane servers for packet-forwarding, thereby dramatically reducing service providers' CAPEX. Finally, because network information, applications, and resources are clustered at a centralized SDN location, operations such as network orchestration and monitoring are much simpler and cost-effective by eliminating individual software updates at multiple system locations that are prevalent in current distributed networks.

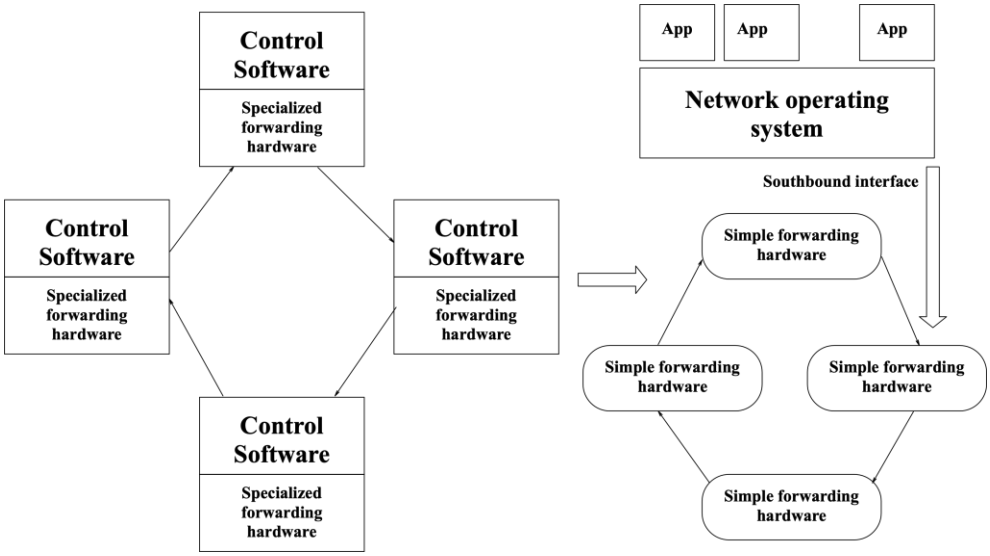

Figure 48 SDN comparison [12]

The isolation of the forwarding and control planes is the first fundamental feature of the SDN. The forwarding functionality is located in the forwarding plane, including logic and tables for choosing

how to deal with incoming packets based on features such as MAC address, IP address, and VLAN ID. It is possible to define the fundamental behaviour performed by the forwarding plane by the way it dispenses with arriving packets. It can forward an incoming packet, drop, consume, or duplicate it. For simple forwarding, by conducting a lookup in the address table in the hardware ASIC, the computer determines the correct output port. Due to buffer overload conditions or due to complex filtering resulting from a QoS rate-limiting feature, for example, a packet can be lost. Special-case packets that need processing are consumed and moved to the appropriate plane by the control or management planes. Finally, multicast is a special case of forwarding, where the incoming packet must be repeated before the multiple copies are forwarded from separate output ports [12].

In the control plane, the protocols, logic, and algorithms used to programme the forwarding plane reside. Many of these protocols and algorithms need the network's global awareness. How the forwarding tables and logic should be programmed or configured in the data plane is decided by the control plane. Since each system has its own control plane in a conventional network, the primary role of that control plane is to run routing or switching protocols so that all distributed forwarding tables on the devices remain synchronised across the network. Preventing loops is the simplest outcome of this synchronisation [12].

The next characteristic is the simplification of machines, which are then managed by a centralised system running management and control software, building on the principle of separation of forwarding and control planes. This software is removed from the computer and installed in a centralised controller instead of hundreds of thousands of lines of complex control plane software running on the device and enabling the device to act autonomously. The network is managed by this software-based controller using higher-level policies. In order to allow them to make quick decisions about how to deal with incoming packets, the controller then provides primitive instructions to the simplified devices when necessary.

2-4-4-1 Network Automation and Virtualization:

SDN can be derived precisely from distributed state, routing, and configuration abstractions. The current dynamic issue of network control faced by networks today is extracted from decomposing into simplifying abstractions. For a historical analogy, note that today's high-level programming languages reflect an evolution through the intermediate stage of languages such as C from their machine language origins, where today's languages facilitate great productivity gains by enabling the programmer to simply define complicated behaviour through abstractions of programming. The distributed state abstraction gives a global network view to the network programmer that protects the programmer from the reality of a network that actually consists of several machines, each with its own state, working together to solve network-wide issues. Without any knowledge of vendor-specific hardware, the forwarding abstraction enables the programmer to define the required forwarding behaviours. This implies that it is important to reflect a kind of lowest common denominator of network hardware forwarding capabilities regardless of language or languages emerging from abstraction. Finally, the abstraction of configuration, often referred to as the abstraction of specification, must be able to articulate the desired goals of the overall network without getting lost in the nuances of how those goals will be enforced by the actual network. To return to the example of programming, consider how unproductive developers of software would be if they were to be conscious of what is actually involved in writing a block of data to a hard disc while they are happily productive with file input and output abstraction instead. Via this configuration abstraction, operating with the network is simply network virtualization at the most basic level. At the heart of how we interpret Open SDN in this work lies this kind of virtualization [12].

In SDN, the unified software-based controller offers an open interface to the controller to allow automatic network control. The words northbound and southbound are also used in the sense of Open SDN to differentiate whether the interface is for applications or for computers. These concepts arise from the fact that the systems are represented above (i.e., to the north of) the controller in most diagrams, whereas the devices are represented below (i.e., to the south of) the controller. The OpenFlow interface [12] that the controller uses to programme the network devices is the southbound API. The controller provides a northbound API that enables software applications to be plugged into the controller, enabling the software to provide the algorithms and

protocols that will effectively manage the network. If the need arises, these applications can easily and dynamically make network changes. The controller's northbound API is supposed to provide an abstraction of the network devices and topology. That is, a generic interface is provided by the northbound API that enables the above programme to function without awareness of the individual characteristics and idiosyncrasies of the network devices themselves. In this way, applications can be built that operate over a wide range of equipment from manufacturers that can vary considerably in their details of implementation.

One of the outcomes of this level of abstraction is that it offers the opportunity to virtualize the network, decoupling the network service from the physical network underlying it. These services are also delivered to host devices in such a way that certain hosts are unaware that the network tools they use are virtual rather than the physical ones they were originally built for [12].

The behaviour and function of a Software-Defined Network is clear at a logical level. We provide a graphical representation of the operation of the basic SDN components in the figure below: the SDN modules, the controller, and the applications. Looking at it from the bottom up, beginning with the SDN unit, is the best way to understand the process. The SDN devices have forwarding functionality for determining what to do with each incoming packet, as shown in the following figure. The devices also contain the knowledge that drives certain decisions about forwarding. The data itself is actually represented, as shown in the upper-left portion of each unit, by the flows identified by the controller [12].

A flow defines a series of packets transmitted to another endpoint from one network endpoint (or set of endpoints) (or set of endpoints). IP address-TCP/UDP port pairs, VLAN endpoints, layer three tunnel endpoints, and input ports, among other things, can be described as endpoints. For all packets belonging to that flow, one set of rules specifies the forwarding behaviour that the system should take. A flow is unidirectional in that packets in the opposite direction travelling between the same two endpoints could each constitute a different flow. As a flow entry, flows are interpreted on a computer. A flow table is stored on the network interface and consists of a set of flow entries and actions to be performed when the device is reached by a packet matching the flow. When a packet is received by the SDN system, it consults its flow tables in search of a match. These flow tables were previously created when sufficient flow rules were downloaded to the

system by the controller. If a match is identified by the SDN system, the required configured action is performed, which normally includes forwarding the packet. If it does not find a match, depending on the version of OpenFlow and the configuration of the switch, the switch may either drop the packet or transfer it to the controller. A flow description is a relatively simple programming expression of what may be a very complex calculation of the control plane previously carried out by the controller. It is important to note that this complexity is such that it can actually not be done at line speeds and must instead be digested by the control plane and reduced to basic rules that can be processed at that pace for the reader who is less familiar with conventional switching hardware architecture. In Open SDN, the flow entry is this digested form [12].

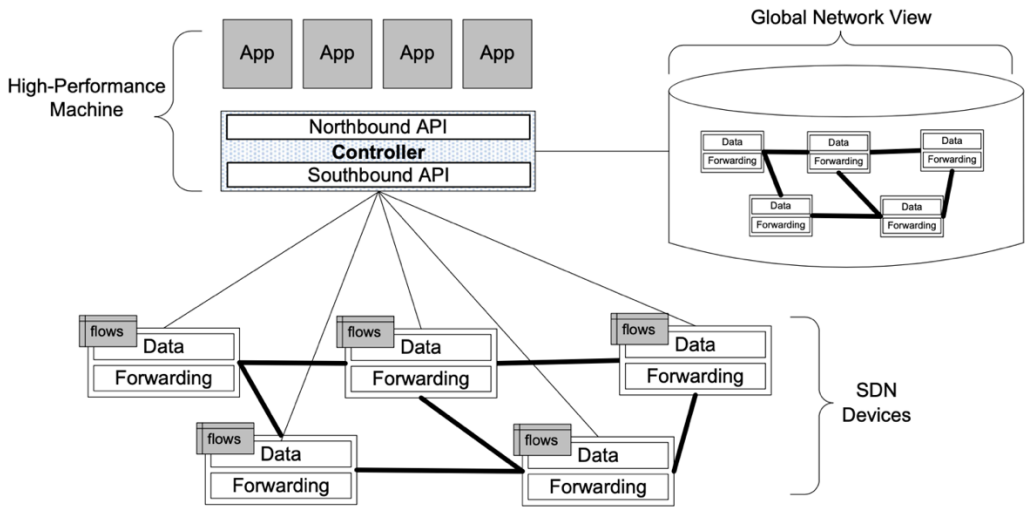

Figure 49 SDN Operation overview [12]

It is the duty of the SDN controller to abstract the network of SDN devices it manages and pre-send an abstraction of these network resources to the above-running SDN applications. The controller helps the SDN application to identify flows on devices and to help the application respond to packets that are forwarded to the controller by the SDN devices. On the right side of the controller, we see in Figure 49 that it retains a view of the whole network it controls. This allows optimal forward-looking solutions for the network to be determined in a deterministic, predictable way. Since a large number of network devices can be managed by one controller, these calculations are generally carried out on a high-performance computer with an order-of-magnitude

performance advantage over the CPU and memory capacity than is usually provided to the network devices themselves. For example, on an eight-core, 2-GHz Processor, a controller could be added versus the single-core, 1-GHz CPU that is more common on a switch. On top of the controller, SDN applications are constructed. These applications should not be confused with the layer of application described in the computer networking seven-layer OSI model. This definition is orthogonal to that of applications in the tight hierarchy of OSI protocol layers, because SDN applications are really part of network layers two and three. The SDN programme communicates with the controller and uses the controller to set up constructive system flows and to accept packets forwarded to the controller. The application sets constructive flows; usually, when the application begins, the application will set these flows, and the flows will remain until any configuration adjustment is made. A static flow is known as this sort of constructive flow. [12] Another form of constructive flow is where the controller intends to alter a flow based on the current traffic load pushed through a network system. In addition to the flows proactively specified by the application, in response to a packet forwarded to the controller, such flows are defined. Upon receipt of incoming packets forwarded to the controller, the SDN application will inform the controller how to respond to the packet and, if necessary, will generate new flows on the device so that the device can respond locally the next time the packet belonging to that flow is seen. Such flows are called reactive flows. In this way, you can now write software applications that, among others, implement forwarding, routing, overlay, multipath, and access control functions. The following figure depicts the OpenFlow protocol as the means of communication between the controller and the device.

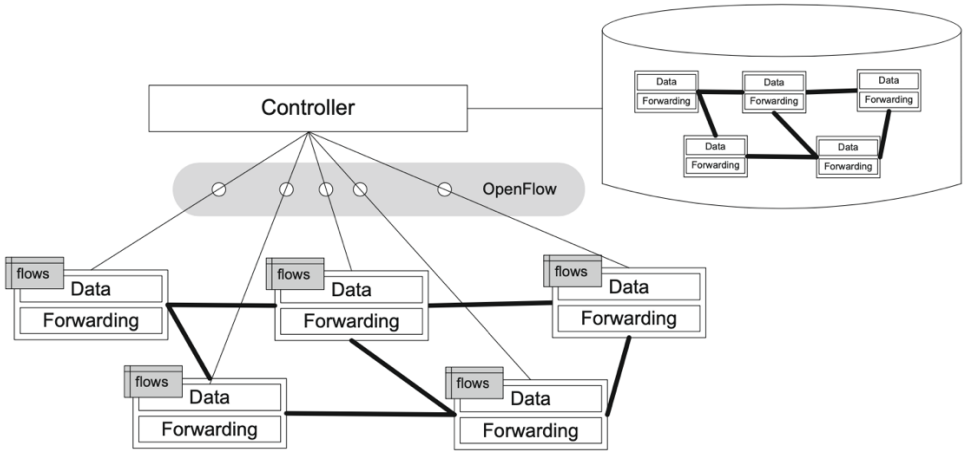

Figure 50 Controller-to-device communication [12]

2-4-4-2 SDN Devices:

An SDN system consists of a controller communication API [12], an abstraction layer, and a packet-processing feature. In the case of a virtual switch, this packet-processing feature is packet-processing software, as shown in the following figure in the case of a physical switch, as shown in the following figure, the packet-processing function is embodied in the packet-processing logic hardware. The abstraction layer embodies one or more flow tables. The logic of packet-processing consists of processes for taking action based on the results of analysing incoming packets and finding the highest priority match. The incoming packet is locally processed when a match is identified, unless it is directly forwarded to the controller. The packet can be copied to the controller for further processing when no match is found. This mechanism is often referred to as the packet consuming controller.

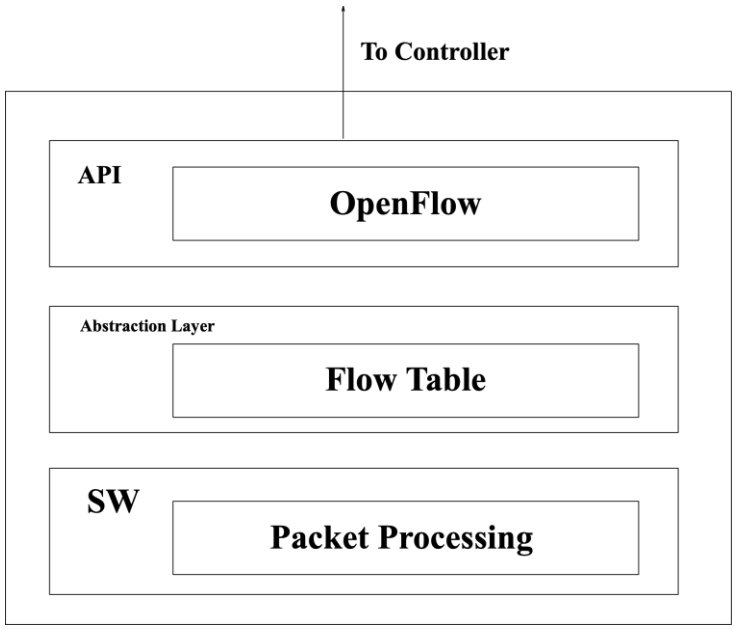

Figure 51 SDN Software switch anatomy [12]

2-4-4-4 Flow tables:

The fundamental data structures in an SDN device are flow tables. These flow tables allow the system to analyse incoming packets and take appropriate action on the basis of the packet content that has just been received [12]. Networking devices have historically received packets and measured them based on those fields. Based on the assessment, actions are taken. These behaviours

can include forwarding the packet to a particular port, dropping the packet, or flooding all ports with the packet, among other things. An SDN system is not radically different except that this simple operation has been made more generic and more programmable through the flow tables and their associated logic.

A number of prioritised flow entries are composed of flow tables, each of which usually consists of two components: match fields and behaviour. To compare against incoming packets, match fields are used. In priority order, an incoming packet is compared against the match fields, and the first full match is chosen. Actions are the instructions that should be executed by the network system if the incoming packet matches the match fields defined for the flow input [12].

For fields which are not applicable to a specific match, match fields may have wildcards. For instance, all other fields will be wildcarded when matching packets based solely on IP address or subnet. Similarly, if only the MAC address or UDP/TCP port is matched, the other fields are meaningless and are thus wildcarded. All fields may be relevant depending on the application needs, in which case there will be no wildcards. The flow table and flow entry constructs allow the developer of the SDN application to have a wide range of packet matching possibilities and take appropriate actions [12].

2-4-4-4 SDN controller:

The figure below illustrates the anatomy of an SDN controller. The figure shows the modules that provide the core functionality of the controller, both a northbound and a southbound API, and a couple of sample applications that may use the controller. The Southbound API [12] is used to communicate with the SDN computers, as we mentioned earlier. In the case of Open SDN or other proprietary alternative to other SDN solutions, this API is OpenFlow. It's worth noting that both OpenFlow and alternatives coexist on the same controller in some product offerings. In terms of its meaning and standardisation, early work on the Southbound API resulted in more sophistication of that interface. OpenFlow itself is the best example of this sophistication, but in the southbound-facing interface, de facto specifications such as the Cisco CLI and SNMP also reflect standardisation.

Unfortunately, the southbound OpenFlow standard or even the de facto legacy standards do not currently have a northbound equivalent. This absence of a controller-to-application interface specification is perceived to be a current SDN weakness, and some bodies are creating standardisation proposals. Northbound interfaces have been introduced in a variety of disparate ways in the absence of a standard that does not stand. A Java API and a Representational State Transfer (RESTful) API are included in the Floodlight controller, for example. For applications operating on different computers, the OpenDaylight controller provides the RESTful API. The Northbound API is an excellent opportunity for vendors and the open-source community to innovate and collaborate.

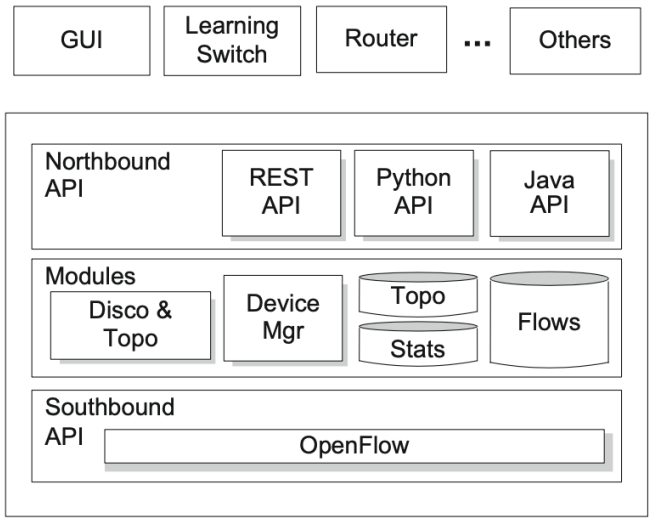

Figure 52 SDN controller anatomy [12]

2-4-4-5 The OpenFlow Switch:

The following figure shows the basic functions and the relationship between an OpenFlow V.1.0 switch and a controller. We see that in [12] core feature, as would be anticipated in a packet switch, is to take packets that arrive on one port (path X on port 2 in the figure) and forward them through another port (port N in the figure), making any required packet changes along the way. The packet-matching feature shown in Figure 5.2 embodies a special element of the OpenFlow switch. In the following, the big, grey, double arrow begins in the decision logic, shows a match with a specific

entry in that table, and guides the now-matched packet to the right-hand action box. For the disposition of this arriving packet, this action box has three simple options:

- Forward a local port to the packet, perhaps first changing such header fields.
- Break a bundle.
- Pass the controller to the packet.

The following figure 53 illustrates these three simple packet routes. The packet is transferred to the controller over the safe channel shown in the figure in the case of path C. The controller uses this same protected channel in the reverse direction when the controller has either a control message or a data packet to send to the switch. It uses the OpenFlow PACKET OUT message when the controller has a data packet to forward through the switch. Following figure shows that such a data packet coming from the controller can use the OpenFlow logic to take two separate paths, both denoted as Y. In the far-right case, the output port is explicitly specified by the controller, and the packet is transferred to that port N in the example. The controller indicates that it wants to postpone the forwarding decision to the packet-matching logic in the left-most direction Y event. An implementation of a given OpenFlow switch is either OpenFlow-only or OpenFlow-hybrid. An OpenFlow-only switch is one that, according to the OpenFlow logic mentioned above, forwards packets only. An OpenFlow hybrid is a switch that can also be used as an Ethernet switch or IP router in its legacy mode to switch packets. The hybrid case can be seen as an OpenFlow switch residing next to a conventional switch that is entirely separate. Such a hybrid switch requires a classification mechanism for preprocessing that guides packets to either OpenFlow processing or conventional processing of packets. During the migration to pure OpenFlow implementations, it is likely that hybrid switches will be the standard [12].

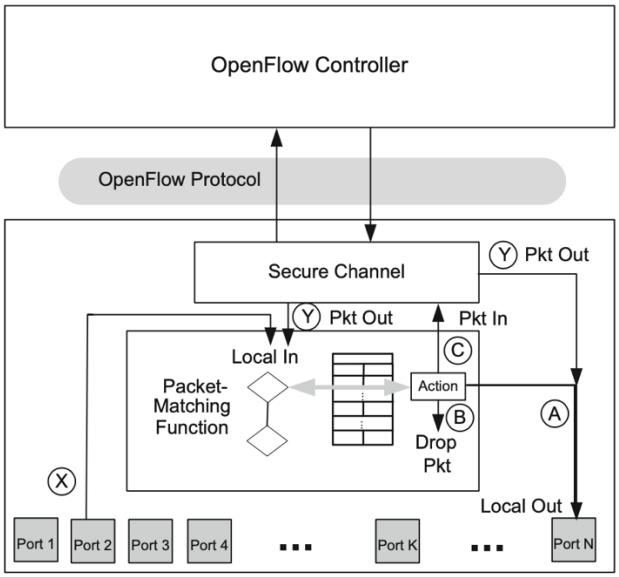

Figure 53 OpenFlow V.1.0 switch [12]

2-4-4-6 OpenFlow Controller:

Modern Internet switches make millions of decisions per second about whether to forward an Internet switch or not. The incoming packet, which output ports should be forwarded to, and the packet header fields may need to be changed, added, or deleted. This is a really difficult job. It is a technical marvel that this can be done at line speeds on multigigabit media. The switching industry has long recognised that not all functions can be done at line speeds on the switching data path, so the idea of separating the pure data plane from the control plane has long existed. The data plane matches headers, modifies and forwards packets based on a collection of forwarding tables and similar logic, and this is done very, very easily. The rate of decisions being taken via a 100 Gbps interface as packets stream into a switch is astoundingly high. To decide what the forwarding tables and logic in the data plane should be, the control plane runs routing and switching protocols and other logic. As the packets are being handled, this process is very complex and cannot be conducted at line speeds, and it is for this reason that we have seen the control plane isolated from the data plane, also in legacy network switches [12].

In three main respects, the OpenFlow control plane varies from the legacy control plane. First, with a common, standard language, OpenFlow, it can programme distinct data plane components.

Second, unlike conventional switches, it resides on a different hardware unit rather than the forwarding plane, where the control plane and the data plane are instantiated in the same physical package. This separation is made possible as the controller can remotely programme the elements of the data plane over the Internet. Third, the controller can programme multiple elements of the data plane from a single instance of the control plane [12].

It is the duty of the OpenFlow controller to programme all the packet matching and forwarding rules within the switch. Whereas a conventional router will run routing algorithms to decide how to programme its forwarding table, the controller is now performing that function or an analogous substitute for it. Any adjustments that result in recomputing routes will be configured by the controller on the switch.

2-4-4-7 OpenFlow Protocol:

The above figure shows that communication between an OpenFlow con-troller and an OpenFlow switch is described by the OpenFlow protocol. This protocol is what distinguishes OpenFlow technology most uniquely. The protocol consists, in essence, of a collection of messages sent from the controller to the switch and an appropriate set of messages sent in the opposite direction. The messages collectively allow the controller to programme the switch in order to enable fine-grained control over user traffic switching. Flows are specified, modified, and removed by the most fundamental programming. Remember that we described a flow in Chapter 4 as a collection of packets transmitted to another endpoint from one network endpoint (or set of endpoints) (or set of endpoints). The endpoints can be described as, among other things, IP address-TCP/UDP port pairs, VLAN endpoints, layer three tunnel endpoints, or input ports. For all packets belonging to that flow, one set of rules specifies the forwarding behaviour that the system should take. When a flow is described by the controller, it gives the switch the information it needs to know how to handle incoming packets that fit that flow. As the OpenFlow protocol has developed, the possibilities for treatment have become more complicated, but the simplest prescriptions for the treatment of the incoming packet are denoted in the figure above by paths A, B and C. These three choices include forwarding the packet to one or more output ports, dropping the packet, or transferring the packet to the exception handling controller [12].

2-5 NETWORK SLICING:

2-5-1 5G network slicing and beyond:

Three types of services are primarily intended for 5G networks: massive machine-type communications (mMTC), ultra-reliable low latency communications (URLLC) and enhanced mobile broadband communications (eMBB). The mMTC is distinguished by a large number of devices that communicate with each other and, along with long battery backup time, demands low costs. Other than that, URLLC requires low latency and ultra-reliability at the same time. On the other hand, along with a broad coverage area, eMBB requires higher data rates. It's imperative to upgrade the network infrastructure to enable users with eMBB, mMTC, and URLLC. Network slicing is a promising candidate for 5G networks to deliver a wide range of services and applications, such as e-health, virtual reality, smart transportation, smart banking, smart farming, and mobile gaming. The idea of network slicing is to utilize the resources of the network infrastructure to create several subnets for various types of services and applications. Then, to create an individual network for its applications, each subnetwork can perform slicing of the physical network resources. Slices of various types can be generated for mMTC, URRLC, and eMBB networks that use network resources in order to allow different 5G services. We may allocate complete end-to-end network resources to each slice type; however, due to the high cost, it does not seem realistic. On the other hand, using technologies such as network function virtualization and software defined networking, it would be feasible to allow sharing of network resources among multiple types of slices [13].

The idea of network slicing as in [13] is the virtual architecture of the network that enables, as seen in the following figure, a powerful and scalable ability to build several logical networks on top of the common physical infrastructure. Softwarization of the network is an evolving trend that uses software-based solutions to enable network slicing. Softwarization of the network can be accomplished by technologies such as virtualization of network functions and networking specified by software. Specifically, 5G network slicing would use software-defined networking, virtualization of network functions, cloud computing, and edge computing to allow scalable deployment on the same physical infrastructure for various types of services. As illustrated in the

figure below, network slicing enables the development of logical networks for various types of services. With the ability to grow on demand, each logical network would have independent power.

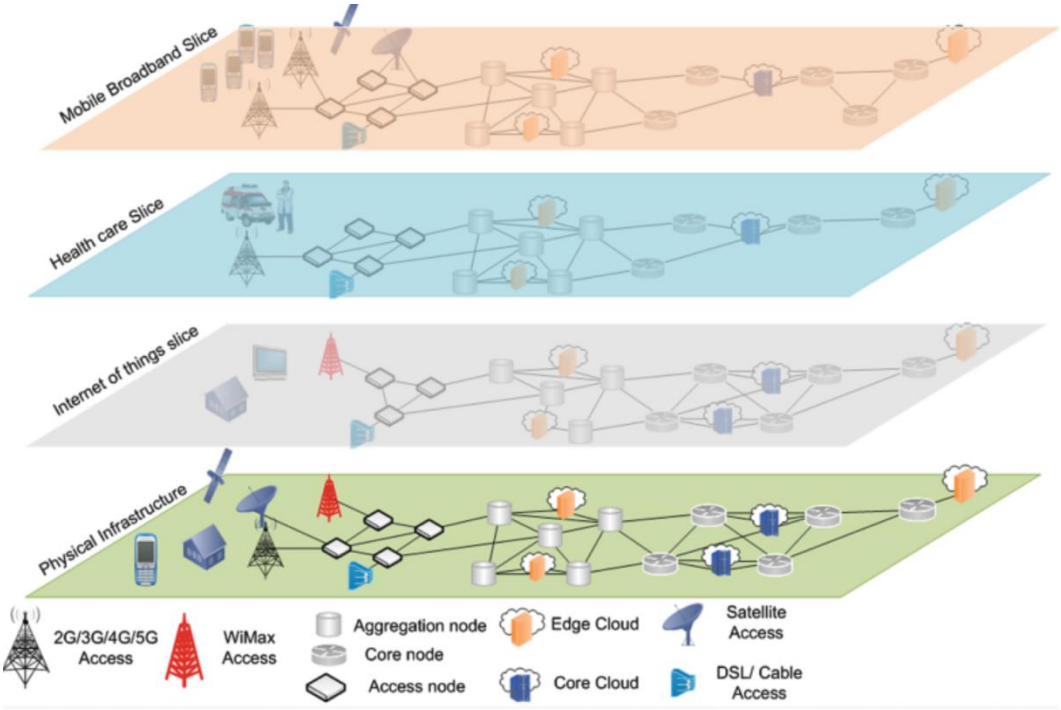

Figure 54 5G Network Slicing [13]

2-5-2 NETWORK SLICING ARCHITECTURE:

In general, slicing is to use virtualization technologies to architect, partition, and organise physical infrastructure computing and communication tools to enable diverse use case realisations to be flexibly supported. One physical network is split into several virtual networks with network slicing, each architected and configured for a particular requirement and/or specific application/service. In terms of service and traffic flow, a network slice is self-contained, and may have its own network architecture, engineering mechanisms, and network provision. 5G network slicing techniques are expected to simplify the development and operation of network slices and enable the physical network infrastructure to reuse functions and share resources [13].

We apply network slicing in two dimensions, as described in Section 1: vertical network slicing and horizontal network slicing. Objectives for vertical slicing to benefit vertical industries and markets. It allows the sharing of resources between services and applications and prevents or

simplifies a typical problem of QoS engineering. As a step forward, horizontal slicing aims to improve mobile devices' capabilities and enhance user experiences. Horizontal slicing goes around physical walls and beyond platforms. This allows resource sharing between network nodes and devices, i.e., sharing their resources with highly capable network nodes/devices (e.g., computing, communication, storage) to increase the capabilities of less capable network nodes/devices [13]. The end result of horizontal slicing is to spin off a new generation of moving network clusters, where terminals become moving network nodes. Over-the-air resource sharing across network nodes involves horizontal slicing. An integrated component and an enabler of horizontal slicing will be the 5G air interface. In contrast, horizontal slicing defines and enriches the 5G air interface.

Independent slices are created by vertical slicing and horizontal slicing. The end-to-end traffic flow normally transits between the core network and the terminal devices in a vertical slice. For example, between a wearable device and a portable device or between a wearable device and a small cell access point, the end-to-end traffic flow in a horizontal slice is usually local and transits between two ends of the horizontal slice. In vertical slicing, similar functions between slices are typically implemented by each of the network nodes. For the most part, the dynamic implementation of a network node is to dynamically assign resources to each slice. In horizontal slicing, however, when supporting a slice, new functions could be generated at a network node. For instance, to support various types of wearable devices, a portable device can require different functions. As well as the resource partition, the dynamic component could lie in the network functions [13].

The vertical and horizontal network slicing definition in [13]is demonstrated in Figure 55. In the vertical domain, through properly designed slice pairing functions, the physical computing/storage/radio processing resources in the network infrastructure (as denoted by servers and base stations) and the physical radio resources (in terms of time, frequency, and space) are sliced to form end-to-end vertical slices. The parameters may vary when the radio, the RAN and the CN are sliced. To pair the radio, RAN and CN slices to form end-to-end slices for different services and applications, slice pairing functions are specified. Examples of end-to-end vertical slices are shown in the following figure. Not necessarily, the mapping between RAN slices and CN slices 1:1. In the horizontal domain, the physical resources in the neighbouring layers of the network hierarchy are sliced to form horizontal slices (in terms of computing, storage, radio).

Multiple slices may work with a computer. For instance, on mobile broadband (MBB) service, a vertical slice on health care service, and a horizontal slice supporting wearable devices, a smart phone may function in a vertical slice

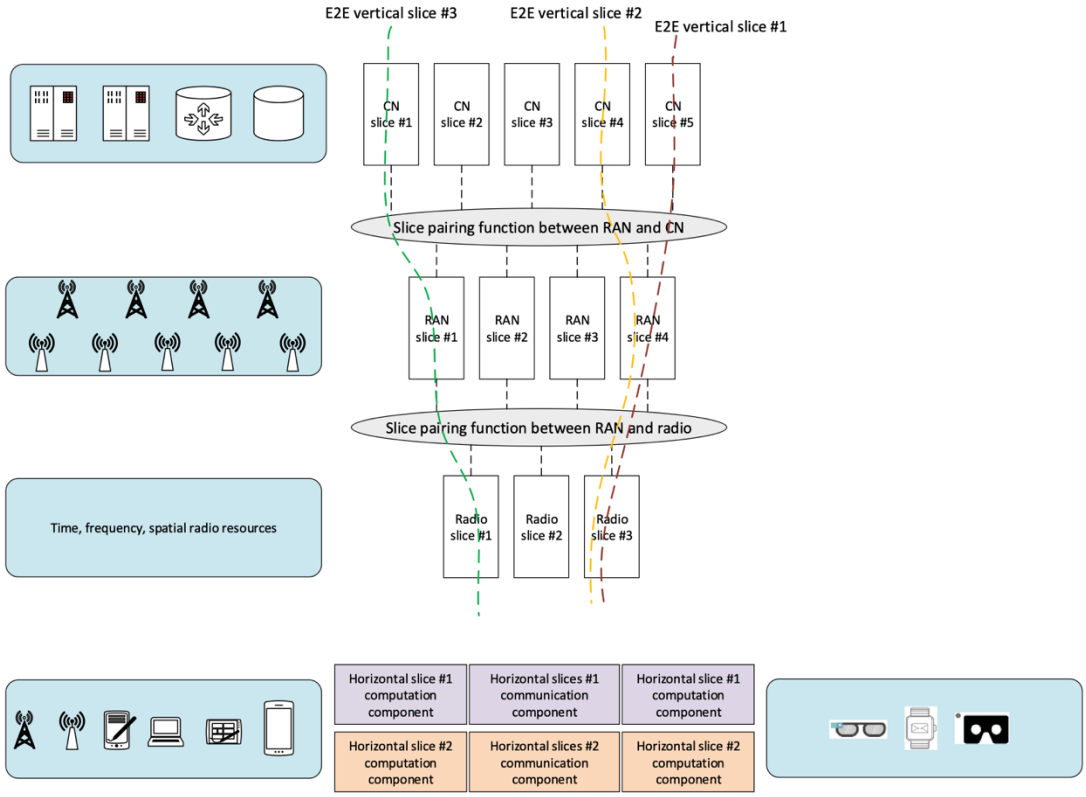

Figure 55 Illustration of vertical and horizontal network slicing [13]

2-5-3 ENABLING VERTICAL SLICING IN THE AIR INTERFACE:

In addition to meeting 5G specifications (e.g., data rate, latency, number of connections, etc.), the air interface's attractive features for network slicing and 5G in general include:

(a) Flexibility: Encouraging flexible distribution of radio capital between slices

B) Scalability: Scaling up quickly with the inclusion of new slices

(c) Efficiency: using radio and energy resources effectively [13].

The physical (PHY) and medium access (MAC) layer architecture, as shown in the following figure, offers one way to achieve the desired characteristics. A modular partition of the PHY resources, an abstraction of the physical PHY resources into PHY resource subsets for each one of the slices, and a compilation of MAC and higher layer operations based on the PHY resource subsets are the basic components. More precisely, the physical radio resource is divided into several parts, each specified by a numerology to satisfy certain communication requirements, such as low latency, wide coverage, etc., where numerology refers to the values of the basic physical transmission parameters that define the radio link, such as the waveform, the sampling rate, the duration of the symbol, the frame/subframe length A subset of physical resources taken from one or more segments of numerology occupies each slice. The MAC operation can then be divided into two layers on top of the physical resource subsets: Level-1 MAC performs intra-slice traffic multiplexing and scheduling; Level-2 MAC performs allocation of inter-slice services. The two-layer MAC partition eliminates the difficulty of several slices being scheduled jointly and enables greater scalability. A network slice ID (sNetID) is assigned to the network slice to define the network slice. The sNetID is known to all devices accessing the slice of the network and can be used to address all devices in the slice of the network. In order to indicate whether the slice is active in the BS, sNetID can be displayed in the device information. We may have slice-specific physical channels, as well as general physical channels, for main physical channels in the air interface, such as the physical downlink (DL) and uplink (UL) control channels, the physical random-access channel and the physical shared channel. Both slices may use the common physical channels, and the slice-specific physical channels are dedicated to the respective slices [13].

An example of the physical downlink control channel in one DL subframe, the physical uplink control channel in one UL subframe, and the physical random-access channel in one frame is shown in the following figure. Resource allocation information for the network slices is carried by the common physical DL control channel. To address the scheduled network slices, sNetID is used. The common physical control information addressed to the corresponding sNetID can be identified by all the devices accessing a planned network slice. In the radio services assigned to the network slice, the slice-specific physical downlink control channel for a network slice is located. In the network slice, the slice-specific physical downlink control channel carriers schedule data for the computers. Likewise, it is possible to design typical and slice-specific physical uplink control

channels in the uplink. The uplink control information can be aggregated by devices accessing multiple network slices and transmitted using the common physical control channel. Slice-specific random access channels can be used for random access to distinguish between dispute resolution and admission control of network slices, such that, for example, a crowded network slice with a high risk of random access collision does not impact devices that access another network slice, or a network slice that serves mission-critical services can be guaranteed low-latency access The slice-specific random access channel resource of a slice may be indicated in the broadcast downlink method, which requires the slice to be active in the BS or access point. The common random-access channel can be used for slices that have not been enabled at the BS or access point or for slices that do not need a slice-specific random-access channel. In this case, the typical random-access channel may also be used in the BS or access point as a way of triggering a slice [13].

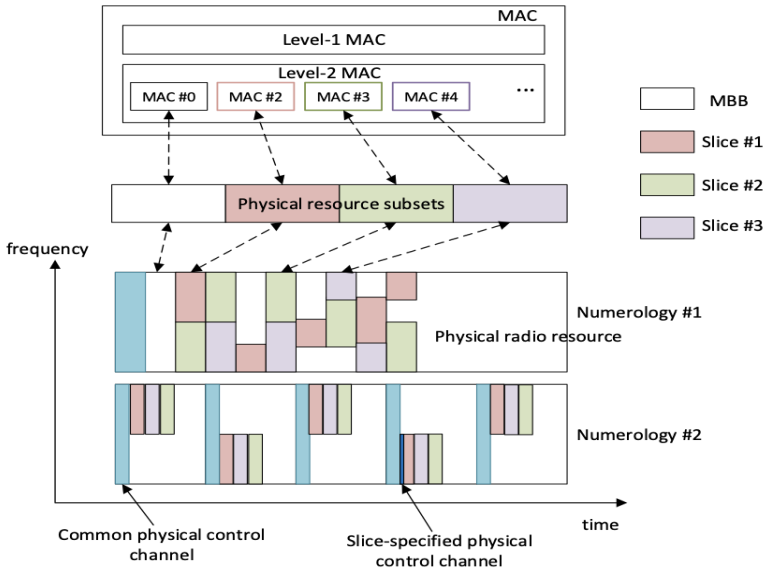

Figure 56 Air interface Slicing [13]

2-5-4 ENABLING NETWORK SLICING IN THE RAN:

As each slice would have its own RAN architecture, as is currently the case in mobile networks, RAN operations such as mobile association, access control and load balancing systems can be slice-specific instead of cell-specific. At any BS or access point, slice on/off operation would be allowed. Considering the slice-specific process, the control-plane (C-plane) and user-plane (U-plane) configurations could be customised. In a way, the slice-specific operation blurs the physical cell site definition and makes the operation of the network more service/traffic/user oriented rather than physical cell oriented [14].

The RAN architecture of each of the slices can be dynamically configured, based on factors such as traffic type, traffic load and QoS requirements. In one instance, for example, slice #1 can work only on a macro cell, slice #2 can only work on small cells, and slice #3 can work on both macro and small cells. Slice #1 may extend its operation to small cells in another case, while slice #3 may terminate operation on some of the small cells. The slice-specific RAN architecture will include slice-specific operation of the control-plane/user-plane, slice on/off operation, and access control and load balancing slice-based care.

Three options can be considered for the control/user-plane configuration of network slices: Option 1 refers to a case for each of the slices with a standard C-plane across network slices and a dedicated U-plane [14].

For each of the slices in [14], Option 2 relates to a case of dedicated control and user planes. Option 3 refers to a case with a standard C-plane for all slices and for each of the slices with a dedicated C/U-plane. Some control plane features, such as idle mode features (e.g. paging, cell reselection, monitoring area update) can be categorised into general C-plane slice features, while connected mode features (e.g. handover, dedicated bearer setup) can be categorised into slice-specific control plane features. Option 2 allows for a slice-specific C-plane, but it could be expensive for each slice to have an always-on C-plane. Option 1 can require less design effort, but for each slice, it lacks the flexibility to customise the C-plane. Option 3 makes C-plane-specific slice-specific always-on common C-plane and personalised C-plane.

Slice on/off, slice-based admission control and slice-specific load balancing are implicitly supported by the slice-specific RAN architecture. An access point will assign radio resources for the slice when turning on a slice and allow all slice-related radio and network functions, such as the and the corresponding physical channels, when turning on a slice. The triggers for an access point to turn on a slice could include: 1) that slice's traffic load goes beyond a certain threshold; 2) the number of active devices operating on that slice goes beyond a certain threshold; 3) the need to ensure continuity of service; 4) the need to meet certain QoS criteria, such as low latency, ultra-reliability, etc. Slice-on can be triggered by a computer or by the network at an access point. The network can decline the system slice-on request when triggered by a device if, for example, the network evaluates that the cost/overhead of serving the slice outweighs the service value.

Similarly, admission control and load balancing will be focused on the availability at the BS or access point of the requested slice, as well as the conditions of load and overall device performance. The BS or Access Point Device Information includes information about the active slices in the BS or Access Point. A computer may start the random-access procedure with a BS that has the intended slice actively running, based on the system details. If the intended slice is not supported by a BS, considering factors such as connection condition, QoS requirements, traffic load of neighbouring cells, etc., the UE can still start the random-access procedure. If the system makes the access request but the slice is not currently active in the BS or access point and the BS has agreed to approve the access request, using the slice on/off process, the BS/access point would need to turn on the slice [14].

2-5-5 ENABLING NETWORK SLICING IN THE CORE:

2-5-5-1 Core Network Slicing Enablers:

Wireless Network Virtualization uses five different means namely: radio spectrum sharing, infrastructure sharing, network slicing by operation, device or application, abstraction layer description that simplifies wireless access from heterogeneous networks, and wireless network programmability and management to achieve network sharing and RAN slicing. Virtualization, which happens to be a primary strategy, uses virtual networks to achieve wireless network virtualization behind network slicing [14].

Software Defined Networking is also an enabler, as the control plane is calculated from the data plane. A collection of network modules is thus committed to serving as the controller known as the SDN controller. The separation of the functionalities of the control plane from the functionalities of data forwarding brings the versatility required to achieve the most perfect implementation of RAN slicing. If SDN is carefully implemented to handle wireless network slices, it could turn out to be the required tool to ease the difficulty that could surround wireless network slice management and programmability.

Network Function Virtualization promotes the idea of re- moving network functions from dedicated physical network hardware equipment to run on any virtualization platform environment deployed in any location on the network. This will allow network functions operating on proprietary network devices to be decoupled to operate on decentralised and virtualized network servers that could be deployed throughout the network at any time with regard to network requirements and service requirements [14].

In order to support vertical networks, existing core networks typically use dedicated hardware. In 5G, different services and applications with varying specifications are required to support the communication system. The use cases of the next decade and beyond cannot be solved by the static, purpose-built vertical network as in present systems. It calls for the core network architecture and a more modular design to be further streamlined.

By flexibly identifying network functions in [14], processing specifications, security procedures and execution nodes based on the requirements of each industry and use cases, network slicing enables the support of multiple industries and different use cases in one network. Technical enablers for network slicing in the CN are NFV and SDN. The IT industry has developed SDN technology primarily for the efficient operation and maintenance of datacenters and large IP and Ethernet networks. SDN's aim is to isolate the control

Data plane aircraft and to make the control plane programmable through APIs in order to make networks deployed, controlled and managed versatile. NFV is mainly powered by providers of network services. NFV's aim is to virtualize network functionality into software applications that can be run on standard off-the-shelf servers in the industry or as virtual machines that run on those

servers. In order to easily allow configured/reused network elements and functions in each slice to fulfil their own requirements, NFV and SDN virtualize the network elements and functions. An example of precisely designed network slices based on common physical infrastructure comprising different network elements is shown in Figure 4. NFV virtualizes the build-up of physical network resources on which SDN performs the C-plane and U-plane dynamic configuration of each network slice. Based on the targeted networks and applications, each slice could have customised core network functions. A management plane may be needed to allocate and configure the network functions for the network slices. Slicing will meet different requirements in the core network, in the RAN and in the radio. Radio slicing, for example, may be based on communication parameters such as latency, efficiency, coverage, etc. A part of the radio resource is taken from each of the radio slices and one form of numerology is described. Service/application as well as contact specifications can be based on RAN slicing. The RAN architecture and C/U plane protocol could be customised for each of the RAN slices. To support various services/applications, the CN slices can be described. It could have customised core network functions for each of the CN slices. To pair the radio, RAN, and CN slices to form end-to-end slices, slice pairing functions are established. The pairing can be 1:1 or 1:M between radio/RAN/CN slices, e.g., a radio slice could have several RAN slices built on top; several CN slices could be built on top of a RAN slice [14].

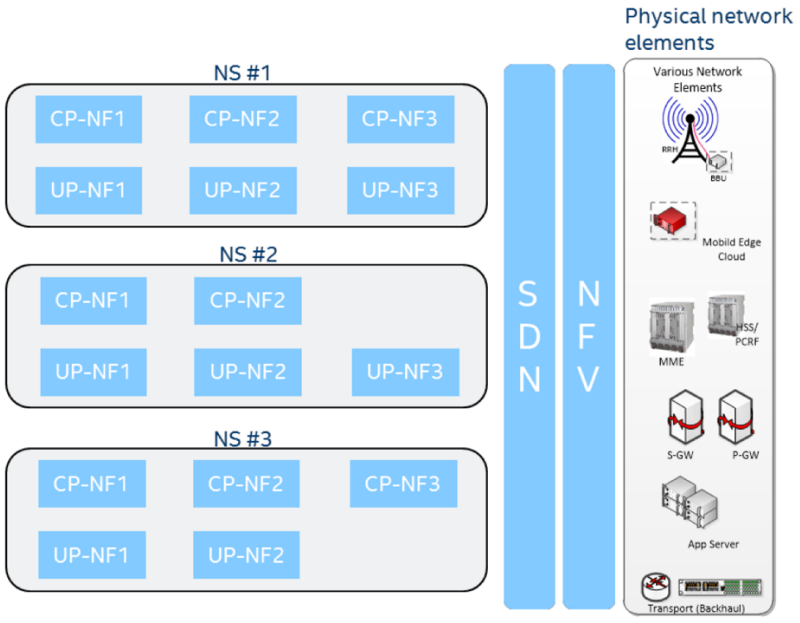

Figure 57 Flexible framework using 5G architectures [14]

2-6 Internet of Things:

The main objective of IoT is to "Connect the unconnected". In other words, objects can get connected to network which provide them an opportunity to interact and communicate with people and other objects. IoT is an emerging technology that helps devices to sense and control the physical world. This facilitates the path for designing advanced applications. To achieve this, there should be some sort of intelligence in devices plus joining them to the network.

So, IoT helps the smart objects to see, sense, make decisions and perform a sort of tasks and collaborate with other objects and share information with each other. In other words, IoT changes the old-fashioned traditional devices to smart ones by utilizing the underlying technologies like RFID, ubiquitous and pervasive computing, Wireless Sensor Networks, Embedded Systems, Internet protocols and so on [15].

The smart objects are generally designed based on their applications and use cases while the underlying technologies like ubiquitous computing and services can be utilized regardless of the type of application and use case. In order to have access to these benefits, the traditional Internet architecture needs to be revised and updated.

2-6-1 IoT Architecture:

Based on the different verticals and different fields of interest several groups developed different IoT architectures to fulfill various requirements. Below are some of the proposed architectures [16]:

- One M2M IoT Standardized architecture.
- IoT World Forum (IoTWF)Standardized architecture.
- Purdue Model for Control Hierarchy.Industrial
- Internet Reference Architecture (IIRA) by Industrial Internet Consortium (IIC).
- IoT-A (Internet of Things Architecture).

2-6-2 M2M STANDARDIZED ARCHITECTURE:

In 2008, ETSI created M2M Technical Committee that focused on creating a common architecture that would help to accelerate the adoption of M2M applications and devices. So, it was initially designed for Machine-to-Machine applications [15].

Later, in 2012, ETSI created a common service layer which can be embedded within field devices to allow communication with application servers. M2M architecture divides the IoT world into three different domains where each domain has certain functions and protocols. This architecture describes how things communicate.

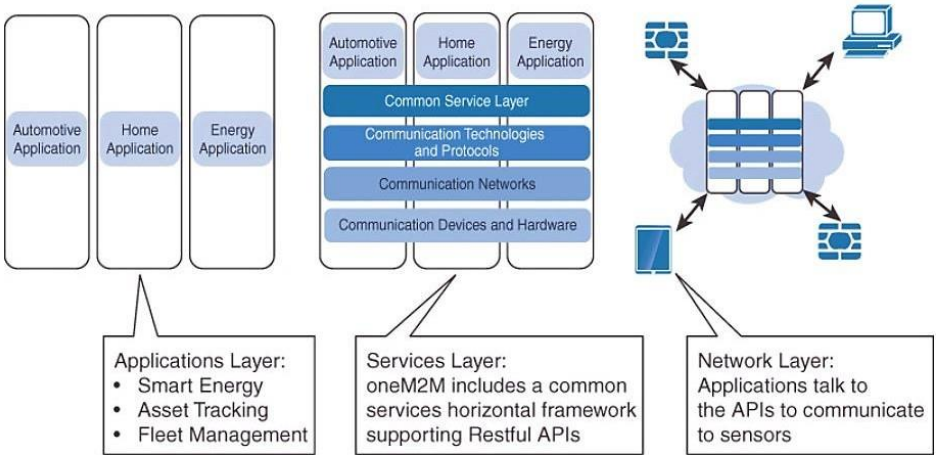

Figure 58 Overall M2M architecture layers, functions and protocols [15]

- M2M Device Domain (Access/M2M Area Networks): It is where IoT objects reside including sensors and IoT objects. Sensors communicate with each other and the network. Sensors need Gateways in order to communicate with network. In other words, this layer provide communication between IoT objects and endpoints using the communication network including wireless Mesh technologies IEEE 802.15.4, wired device connections like IEEE 1901 PLC and wireless point-to-multipoint like IEEE 802.11ah.
- Network Domain (Backhaul and core networks): This is where sensors can communicate over the Gateway and data can be forwarded to another network domain. This domain provides connectivity between Local Area and some applications somewhere in the cloud. The connectivity can take multiple forms. It can be radio-based

through the satellite, wired or wireless connection. Because of this variety, lots of different protocol exist including LTE, WiMAX and MPLS. In other words, multiple networking technologies can be leveraged to achieve connectivity between devices in the device domain and applications in the application domain. This domain includes modules for physical network that IoT applications can run on top of it using underlying management protocols and hardware [15].

- Application Domain: It is a place where applications are created that leverages data collected and taken from different devices and make use of data. At this level, management functions are bold. For instance: smart energy management, connectivity management, data analytics and fleet management. This layer provides connectivity between devices and their application and applications are industry specific. So, they will have their own data models.

2-6-3 IOT CONNECTIVITY PROTOCOL:

It is a good idea to share the differences and impact of each of short-range wireless network briefly. Short-range wireless networks divided into two parts: Wireless LAN (IEEE 802.11) and Wireless Personal Area Network (WPAN). Wireless LANs is just an improvement or option for LAN networks (IEEE 802.3) and they both can coexist together for better flexibility. WLAN intended to improve the data rate, mobility in LAN networks [20].

On the other hand, WPANs are designed for low-power, low-data rate wireless communications. And can be a replacement option for WLANs. Wireless PANs are divided into three categories:

- High rate (IEEE 802.15.3): This category has a data rate of 11-55 Mbps and is good for real-time wireless video applications.
- Medium rate (Bluetooth): It has data rate of 1-3 Mbps and is used for applications like wireless headsets for voice transmission [20].
- Low rate (IEEE 802.15.4): like ZigBee standard with data rate of 250 Kbps.

2-6-3-1 ZIGBEE PROTOCOL:

ZigBee is one of the protocols that implements IEEE 802.15.4 like 6LoWPAN and wireless HART protocols which in turn increases the features of IEEE 802.15.4 standard. In other words, it is a low cost, low power, low data rate wireless protocol for battery-powered devices. ZigBee lower data rate is good for wireless communications that sends and receives simple commands or information from sensors like temperature [20].

This protocol was introduced by ZigBee Alliance Group specializing in Adhoc control in 2003. ZigBee takes advantages of IEEE 802.15.4 MAC and PHY layers and on top of that it adds new functions as shown in Figure 2-3. It adds logical network, security and application software. It is a wireless mesh network with self-organizing specification [19].

Wireless devices using ZigBee standard, can operate in the following frequency bands: 868 MHz, 915 MHz and 2.4 GHz. As mentioned earlier, it has a built-in security. Network and security layers are responsible for setting up the network, configuring routing table, calculating routing path, neighbor discovery, topology formation and security communication.

In terms of security perspective, it applies IEEE 802.15.4 security approach in the MAC Layer and uses AES 128-bit key. It also applies security in the Application and Network layers [15]. ZigBee protocol defines the Network Layer specifications for star, tree and peer-to-peer network topologies and provides a framework for application programming in the Application layer.

ZigBee protocol works with the IEEE organization to ensure solution development. ZigBee goes to all seven layers. In ZigBee, many industry-specific applications are predefined, and vendors can create their own customized version. These industries include smart energy, security systems, medical data collection, utilities monitoring and control and home automation. One example of ZigBee application can be found in blood pressure in-home monitoring systems. So, the information related to pressure and heart rate will be collected by wearable device that is ZigBee-enabled [20]. Then, the information will be sent to central system like patient’s computer and important information will be sent to patient’s nurse wirelessly and by Internet. ZigBee structure only provides interoperability with vendors under the auspices of ZigBee alliance.

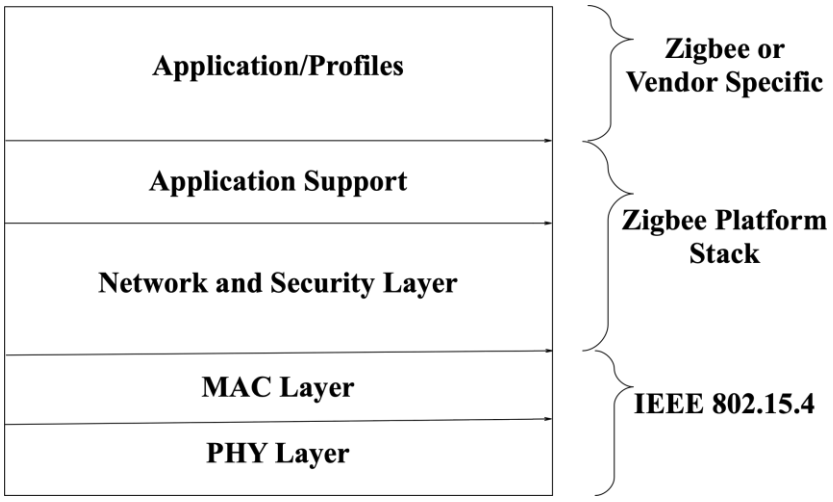

Figure 59 Zigbee Protocol Stack [15]

ZigBee focuses on two layers of OSI model: network and application layers. And IEEE 802.15.4 is the basis for ZigBee protocol. The following are the characteristics of ZigBee protocol:

- Low power consumption and requires at least two-year batter life for devices.
- Low cost because communication protocols are simplified, and data rate has been reduced.
- Low offered message throughput and has period, intermittent transmission.
- The minimum requirement for implementing ZigBee and IEEE 802.15.4 is simpler and less compared to IEEE 802.11 [20]
- Lower data rates and consists of simple devices
- Supports larger networks
- Transmission scheme is DSSS
- No guarantee for QoS
- Flexible protocol design
- Networks are secured by AES 128-bit encryption
- It uses reactive routing protocol and will establish a route if there is a demand (like

AODV protocols).

Also, ZigBee achieves to higher reliability by utilizing IEEE 802.15.4 standard with DSSS[1] and O-QPSK[2] approaches which provides a good performance in terms of [17, 18, 19]:

- Low signal-to-noise environments.
- CSMA/CA access method by listening to the media before transmitting the data.
- 16-bit CRC frame checksum.
- Acknowledgement per hop meaning that for each unsuccessful packet the retrial will be done up to maximum of three times and if the packet could not get delivered to destination, ZigBee informs the sending node about it.
- And mesh networks which brings the benefits like extended network range by multi-hop, forming Adhoc network, automatic route discovery and self-healing [17].

Downsides of ZigBee are relying on a single wireless link technology, it is tightly coupled with application profiles, limitations in integrating with Internet and it is not scalable, low data rates and it is not free [18, 19].

PHY Layer

As mentioned earlier, the PHY layer in ZigBee is IEEE 802.15.4 PHY Layer which is responsible for communicating and managing the radio transceiver. General tasks involve activation of radio for delivering packets, channel selection, chip modulation, transmission rate specifications.

MAC Layer

MAC Layer is responsible for association and disassociation of services and if the device is coordinator, it generates beacon frame and synchronizes the device with beacon. Using beacon provides timely schedule for all nodes. However, using beacon adds more complexity to ZigBee implementation [17].

MAC layer also controls the channel access for avoiding collisions using CSMA/CA. So, if a node wants to transmit, the MAC Layer senses the medium for any other radio transmissions and if the line is busy, it waits for some time [19]. This time is defined by the timer and back

off algorithm and then retries to transmit its data. Also, Mac Layer provides acknowledgement for all ZigBee packets in a hop-by-hop format. There are four types of MAC frames:

- Beacon frame: Used by PAN coordinator for synchronizing the clock of all the devices in the network. Also, it notifies the device if there is a pending data for him. If the answer is yes, the devices will pull the data from coordinator using indirect transmission.
- Data frame: It is used for data transmission. Data frame is the payload of network layer. Then, it becomes MAC layer payload.
- Acknowledge frame: It is used for the acknowledging the successful receipt of data frame. It does not carry any MAC payload [21].
- MAC frame: Also, called Command frame and includes commands like request to associate/disassociate with the network.

Network Layer

Like IP network, it performs routing and addressing. Here, ZigBee coordinators and ZigBee routers will perform route discovery each destination. Also, the type of the network, topology, network address assignment will be defined by coordinator [21].

There are two types of routing in ZigBee: Mesh routing and tree routing. Briefly speaking, mesh routing is done on point-to-point fashion between the ZigBee routers. And for mesh routing, AODV[1] routing protocol is adopted which is a reactive protocol and routes will be established upon they are needed. While in the tree routing, the coordinator becomes the root and end devices are the leaves. But this routing is good for small to medium network sizes and as the network grows using this approach by resource-constrained devices causes memory overflow [21].

Application Layer

Here, application objects reside. Application objects manage and control the protocol layers in a ZigBee device. Also, for each specific application, an application profile can be defined which specifies the message formats, processing actions [21].

Two types of application profiles exist: vendor-specific and public application profiles. Public

profiles are interoperable between different vendors while vendor-specific profiles are designed to be interoperable and used with products of one specific vendor.

Finally, ZigBee profiles are like port numbers in IP stack. So, when the application receives a packet, it demultiplexes based on the endpoint identifier. [20, 21]

2-6-4 IOT GATEWAY FUNCTIONALITY:

IoT GW functionality commonly includes,

(i) data pre- processing and transformations.

(ii) events/alerts and embedded actuation logic.

The realization of this functionality is typically considered at the edge of the network infrastructure, where local processing and decision making can support: (a) Low latency: low response times, enabling timely actuation, especially in mission critical cyber-physical systems (CPS), and (b) Scalability: opportunities for load reduction by enabling the aggregation and/or filtering of data prior to entering the transport domain and reaching centralized backend locations. Following diagram depicts the architecture of IoT gateway [14].

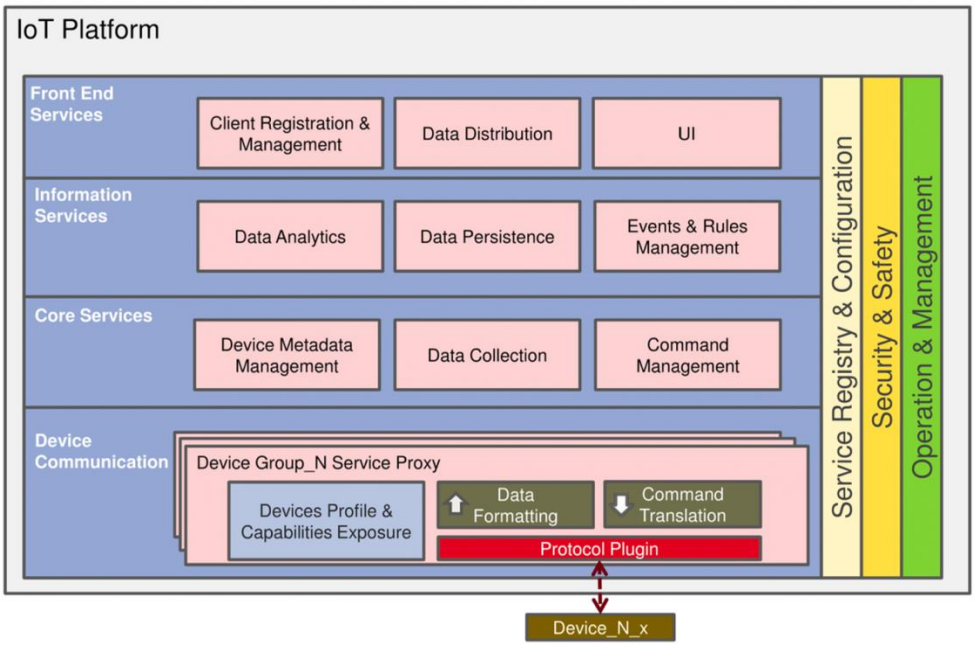

Figure 60 IoT Platform Stack [14]

It consists of several layers, with the layers closer to the IoT devices at the bottom and the layers closer to the applications at the top.

Device communication: The direct communication of the IoT GW with the IoT devices is the responsibility of this bottom layer. It translates messages to/from user-understandable protocols and communicates them to upper layers, formats data for upstream transmission to upper platform layers (device capabilities and measurements) and translates downstream commands received to device language [14].

Core Services: this layer serves as a reference point for upper layers, I collected data and command control from/to the communication layer of the device, and ii) management of metadata (device profiles, data types of measurements, etc.) as revealed by the communication layer of the device.

Information services: this layer deals with advanced data analysis (Data Analytics) of data collected, Data Persistence and Events and Rule Management, triggered by rules clearly specified by registered and approved users. Front End Services: provides a GUI, handles user registration for various services, and distributes data to northbound applications on the top layer of an IoT GW. Service Registry Configuration: vertical to all the above layers, this organization is responsible for component-level status and health checks and for application-wide configurations as a reference point.

There is also a vertical layer for Security and Safety, as well as a layer for Process Management, providing the platform operator with a management view.

2-6-5 Baseline mMTC Slicing Model:

We consider the capacity of the 5G network to devote wireless and wired networks, as well as computing and storage resources, as a simple mMTC network slicing strategy, to the connectivity needs of specific sets of IoT devices (UEs) owned, installed and controlled by prospective IoT platform operators. To support the corresponding connectivity needs, prospective IoT platform operators deploy their IoT devices based on the establishment of mIoT network slices. As a result, network slice instances are provided to IoT devices to be picked from, so that they are incorporated into the 5G facility control and data plane, e.g. system authentication, data forwarding [14].

Each slice allocates network resources for the transmission and distribution of IoT data, either from or to IoT devices (e.g., in case of actuators). Computing and storage resources, including IoT Gateway (GW) functionality, are also allocated to support the 5G core functions needed for this integration (if not realized as a dedicated PNF). The model mentioned is illustrated in the following figure. The exclusivity of per-slice tools simplifies management as well as the implementation of actuation logic and enables tenants to incorporate, control and customize their own IoT devices seamlessly.

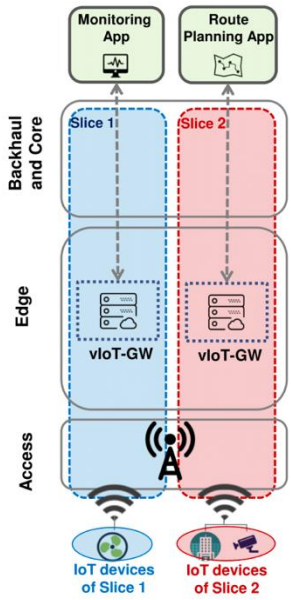

Figure 61 Baseline mMTC Slicing Model [14]

2-6-6 Applying SDN/NFV for 5G and IoT:

Firstly, it is necessary to apply SDN/NFV orchestration to the dynamic deployment of a virtual mobile network (VMN). SDN/NFV is expected to achieve VMN deployment scalability, cost-efficiency and versatility. These technologies are attractive in that they allow VMNs, based on the actual traffic load, to flexibly control their virtual core mobile network, RAN and Enhanced Packet Core (EPC).

Secondly, the ultra-low latency use of DCN (Data Center Network) is a critical solution that allows network functions and can be implemented on the core mobile network. As a consequence, this will include an opportunity inside the cloud to control network traffic. In addition, the reduction of DCN latency leads to a new way of processing and forwarding cloud flows that can accommodate IoT and mobile Internet applications.

Thirdly, the Full-SDMN architecture needs to adopt anything as a service (XaaS) to deal with on-demand services and multiple mobile network service providers in order to provide VMNs with virtual core and virtual radio access networks and to ensure separation between VMNs. Full-SDMN is also considered to be an optimal option for network operators to extend their footprint networks, which act as scalable services, including SDN controller applications and data plane functions, and the physical infrastructure is operated as a service because it offers many advantages, such as high scalability, low-cost services.

Finally, as a new aspect that encourages creativity, the comprehensive Full-SDMN architecture enables independent network protocols in the data plane. In other words, it provides controllers, programs, and data planes that are programmable.

2-6-7 Full SDMN Architecture Design [14]:

The proposed architecture, shown in the following figure 62, examines three key enablers, it has capabilities to provide the fully programmable on both the core and access mobile network and seamlessly connect across through access and core mobile network [14].

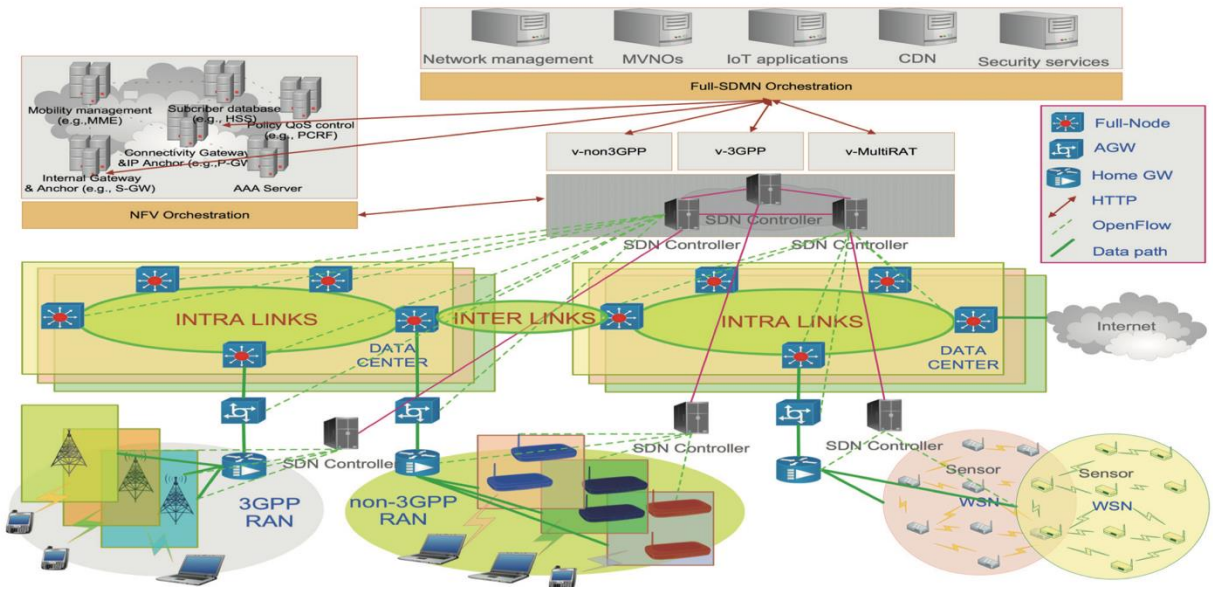

Figure 62 Full SDMN Architecture [14]

2-6-7-1 Softwarization of Core Mobile Network:

Firstly, network functions such as the policy and charging rules function (PCRF), home subscriber server (HSS), Authorization and Accounting (AAA), mobility management entity (MME), decoupled EPC control plane (S-GW, P-GW), and all signaling are running on commodity servers in DCN. This enables the NFV orchestration to quickly and easily manage certain components. Similarly, the decoupled data plane remains in the DCN infrastructure with the goal of using programmable full nodes to optimize network performance.

Secondly, DCN edge access gateways (AGWs) must be implemented to allow them to connect to different types of Radio Access Networks (RANs), such as 3GPP (e.g., LTE-E-UTRAN, UMTS-UTRAN, GPRS-GERAN), non-3GPP (e.g., WIMAX, CDMA2000, Wi-Fi), Wireless Sensor Network (WSN), and new RANs or Multi-RATs (e.g., WIMAX, CDMA2000, Wi-Fi) (Multi-Radio Access Technology) [14].

NFV orchestration, Full-SDMN orchestration and SDN controller work together to ensure that services and applications such as network management as a service, MVNOs (Mobile Virtual Network Operators), IoT applications, CDN (Content Delivery Network), protection, etc. are handled by full-SDMN orchestration. Full-SDMN orchestration creates streams to program the network components to create virtual core networks based on the service (VCNs). Full-SDMN orchestration can request information from virtual network functions like S-GW and P-GW to control different VCN parameters once a VCN is formed. Finally, NFV orchestration empowers SDN controllers via controller applications to program full-nodes and AGWs.

2-6-7-2 Full – Node Data Center:

The key feature of the data center is known to be the Full-Node, which is an integrated node. It includes the FPGA-based NIC (network interface card), all-optical rack top (ToR) switch, cluster top (ToC) switch with the ability to program the data plane with the use of new programmable optical technologies such as programmable transponders, selective spectrum switches, and the use of computing and programmable hardware (e.g., FPGA, CPU, NPU) and storage. ToC can combine various optical technologies, such as optical packet switching (OPS), optical burst switching (OBS), and optical circuit switching, which are fundamental hardware (OCS). On an embedded Linux server platform, a virtual OpenFlow Switch (vOFS) is also deployed. Each vOFS automatically controls the underlying hardware depending on the inserted flow entries to create cross-connections between their ports. It is anticipated that the adoption of VNODE and LIGHTNESS would provide deep programmability in the data plane,

slicing the network, and then running each slice with different protocols, as shown in the following figure 63 [14].

The logically centralized SDN controller manages each of the NICs, optical ToRs and ToCs, revealed by dedicated Full-Node agents, via a uniform control software interface. Agents are unique devices that use the proposed extended OpenFlow protocol to obtain useful hardware device information, so the SDN controller can update the current network status and then pass the control commands to the physical devices.

DCNs comprising a collection of Full-Nodes, optical fibre transmission and CORD, which can be chosen as an advanced solution, are connected by Inter-Links [14].

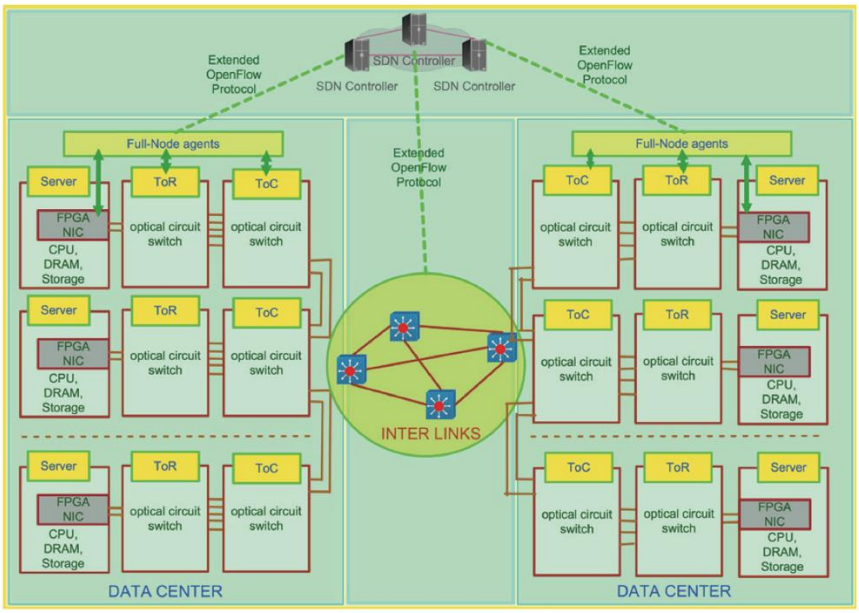

Figure 63 Full-Node and data center Architecture [22]

2-6-7-3 AGW:

An AGW is an access gateway that plays the role of both the S-GW and P-GW to communicate between VRANs and VCNs. The IEEE 802.1ad is known as Q-in-Q VLAN tunneling will be implemented in communication.

2-6-7-4 Full SDMN orchestration:

Full-SDMN Orchestration is a system for the assembly and composition of scalable services such as MVNOs, core networks, mobile network access, network management, IoT software,

etc. It unifies VNFs, controller applications (provided by SDN controllers), data planes, and cloud services (provided by NFV Orchestration). In this sense, SDN controllers receive full-SDMN orchestration specifications and then cooperate with each other to produce static flows that are pushed to create VCNs and VRANs in the physical network infrastructure [14].

2-6-7-5 NFV Orchestration:

VNFs are stored in the form of VM images or NFV orchestration containers and can also be deployed centrally on a single server or cluster under the operating mechanism of OpenStack. Moreover, they are also configured via SDN controllers to manage a virtual network. In this case, an individual instance for each virtual network is generated by NFV Orchestration.

In the architecture depicted in figure 63, all the features run on top of the SDN controller as controller applications instead of on the data plane. As a consequence, it is possible to cooperate with each other on similar applications. For example, when the UE is first attached, the AGW does not have any UE data, so the AGW will send a Packet-in message to the SDN controller. In this case, the AAA, which is used as an application for the SDN controller, is registered to listen to this form of message, and after checking that the UE is permitted to use the network services, the AAA feeds the UE profile to S-GW, P-GW for further processing. In addition, other applications such as Network Management, MVNOs, IoT applications, CDN, and security services can also be similarly deployed on other NFV orchestrations [14].

2-6-7-6 SDN Controllers:

Three important functions in the network are performed by SDN controllers, who are structured as a distributed control architecture: programming the data plane; controlling flows through network nodes; and providing the forum for hosting applications for controllers. In addition, SDN controllers work seamlessly with each other to implement v-3GPP, v-non3GPP, v-MultiRAT directly. The SDN controller receives a collection of specifications from the Full-SDMN Orchestration to monitor the VCNs and VRANs during service. Full-SDMN Orchestration, meanwhile, may also request VNF information to track the parameters of different VCNs and VRANs.

The SDN Controller has been expanded to support other devices on the network. Full-Node agents are specialised computers, such as the NIC OpenFlow agent, ToR and ToC OpenFlow

agents, for implementing the proposed extended OpenFlow protocol [14].

2-6-7-7 SOFTWARIZATION OF ACCESS MOBILE NETWORK:

The RAN also needs to be configured in the proposed mobile network architecture, similar to the core, to build virtual access networks based on particular requirements. In these conditions, it is also possible to abstract and share RAN resources consisting of physical connectivity resources for mobile network infrastructure (e.g., APs, eNodeBs, storage, computing, etc.) and physical radio resources (e.g., radio links between APs or eNodeBs and end-user devices) for better use. This also allows slices produced from the core network to be extended and mobile networks to be accessed, including end-network devices [14].

Chapter 3. Security threats and solutions of 5G Technologies

3-1 INTRODUCTION TO 5G SECURITY:

Cloud-native principles such as self-contained functions inside or through data centers (cloud) are used by the 5G architecture, communicating in a micro-service setting with all components working together to provide services and applications. The cloud-native 5G architecture offers an elastic, automated environment that can expand and contract network, device and storage resources as required. It is now possible to host several telecommunications and mobility functions as software services and dynamically instantiate them in various network segments. The complete 5G network needs to be scalable and is ultimately built to be configurable for software. The principle of disaggregation enables these two planes to exist on different devices or at separate locations within the network by the means of control plane and user plane separation. The separation of the control plane from the user plane allows the two planes to scale independently, without having to increase resources of one plane when the other needs additional resources. This separation makes it possible for the planes to fly at a distance from each other; they are no longer needed to be co-located, further enhancing scalability [22].

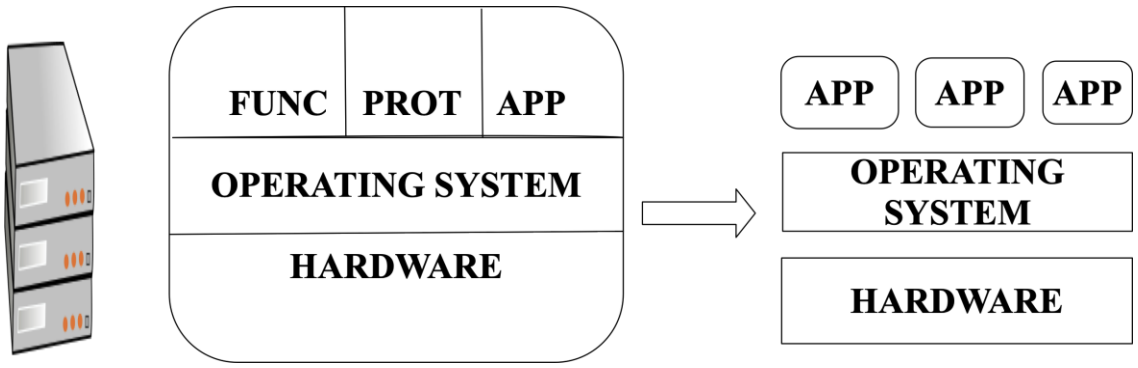

Figure 64 Monolithic vs Disaggregated Architecture [22]

In the Radio Access Network (RAN), cloud migration enables the remote radio unit (RRU) to be disaggregated from the baseband unit (BBU). It becomes possible to build a pool of BBU resources that support multiple distributed RRUs by separating these functions. Doing this enables RAN resources to be used more effectively. It also, however, creates problems, such as the need for fronthaul communication between the RRUs and the BBUs, where high bandwidth and low latency are needed for fronthaul. Via a new architecture that describes splits at different locations in the RAN, where RRUs connect to distributed units (DUs), then connect to centralized units, 5G tackles this frontal problem (CUs). This is referred to as the break of CU/DU and introduces the midhaul concept. The trade-offs between bandwidth requirements and the ability to centralize resources drive how and where an operator chooses to position the RRUs, DUs and CUs [22]

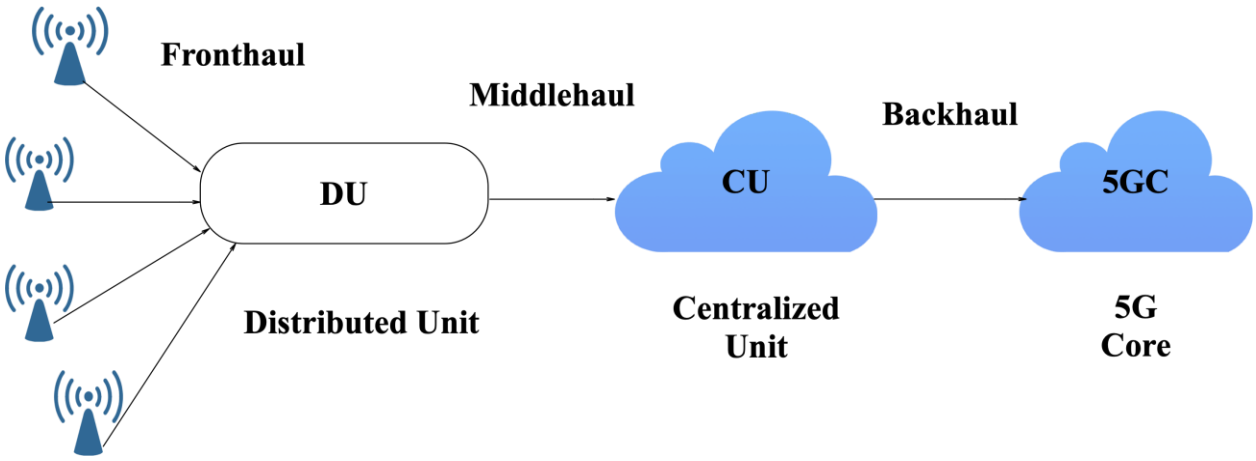

Figure 65 5G RAN splits resulting in fronthaul, midhaul and backhaul [22]

The electromagnetic spectrum being used is another significant difference between 5G and previous generations of cellular networks. 5G needs much greater quantities of bandwidth, and the amount of spectrum has now expanded, and its locations are therefore different. Some frequency ranges remain below 6 GHz, whereas access to far greater quantities of contiguous spectrum is given by moving into the millimeter wave (mmWave) range [22]. The shorter distances that their signals can travel and their failure to reach fixed structures are one downside of the higher frequency ranges (including rain fade). The quantity and location of base stations required for adequate coverage are affected by these limitations. With a greater number of smaller base stations, this can be solved, i.e., Small cells, Micro cells, Cells of Femto/Pico.

There are many services incorporated in the device for each generation of wireless networks. From validating whether a subscriber can connect to the network to handling the flow of user equipment traffic, these services perform different network functions. 5G introduces the idea of an SBA, one in which the individual functions or services are virtualized and "cloudified." As a result, an SBA allows all the functions to interact with each other through APIs across a packet-based network instead of a static route via specialized equipment and its physical arrangement depending on how it is actually linked.

Security attacks are divided mainly into two categories Passive and Active Attacks. Passive attacks are generally made to make unauthorized use of information of the authorized users and are not intended to interrupt the communication. Most famous passive attacks are Eavesdropping and Traffic Analysis. These attacks violets the privacy of users and Data Confidentiality. On the other hand, Active attacks are intended to interrupt the communication of legitimate users or to alter the communication. Active attack includes Distributed Denial of Service (DDoS) Attack, Denial of Service (DoS) Attack and Man in The Middle (MITM) Attack. To deal with these attacks many approaches are being used. These approaches can be categorized into two types, Cryptographic Approaches and Physical Layer Security (PLS) Approaches. Out of these two, Cryptographic approaches are conventionally being used at different layers of 5G Wireless network. Cryptographic approaches are considered into two types, Asymmetric Key Cryptography and Symmetric Key Cryptography. Symmetric Key Cryptography is referred to method of sharing a private key between each pair of communication parties for encryption and decryption [22].

In Asymmetric Key Cryptography, a private key is used to decrypt the data and a public is used to encrypt the data. As the name suggests public key is shared among all parties of network and each user will have his own private key. The performance of the Cryptographic approaches totally depends upon length of the key and complexity of algorithm. But due to the low latency and reduced power consumption requirements of 5G it will encounter problems in Cryptographic approaches. PLS can play interesting role here, it is proved that even if legitimate user uses worse channel than eavesdropper, it is possible to generate secrete key. There is lot of research done in 5G wireless network. Comparing these two strategies PLS will have lower computational complexity with larger scalability therefore PLS is ideal technique for 5G Wireless Networks [22].

3-2THREATS AND IT'S SOLUTION OF 5G NON-STANDALONE:

3-2-1 2G / 3G Downgrade Attack:

THREAT: Downgrade attacks allow opponents to push a UE-connected LTE to 2G or 3G, which has considerably less security controls. Ultimately, opponents may conduct active attacks and/or passive (e.g., eavesdropping) attacks by man-in-the-middle (MiTM) to gather sensitive data. This form of attack may be suggested by a client experiencing suspicious activity in their LTE connection. For example, if clients in a particular, well-known area (e.g., work, home, coffee shop, etc.) have historically encountered stable 4G/LTE connections, but their connections unexpectedly return to 2G/3G. Customers will usually tell from the connection indicator on their phones whether they are linked to 2G or 3G. The screen would usually display an "E", "G", or another symbol instead of signaling LTE and/or 4G [22].

SOLUTION: Third Generation Partnership Project (3GPP) Security against active IMSI catchers is included in release 15, the first 5G standard. This security is enforced such that the UE encrypts its identity with the HN's public key using public key encryption. The concealed identity is called subscription concealed identifier (SUCI). This defense only works when the SN is also a 5G entity since it would not be possible for an SN from LTE, 3G or GSM networks to process the SUCI. This means that an active IMSI catcher can mount a 5G UE downgrade attack so that it reflects an

LTE SN and exploits the vulnerability of LTE to steal the 5G UE IMSI. In order to escape this downgrade attack, we can use pseudonyms that have the same format as IMSI for LTE communication. The idea of using IMSI format pseudonyms is to confuse IMSI mobile network catchers. A pseudonym looks like a standard IMSI, but its last nine to ten digits are randomized and changes frequently. In the beginning, the UE is given with two pseudonyms and during AKA runs, it gets new pseudonyms. Instead of IMSI, it uses these pseudonyms to describe itself when connecting to an LTE SN. Instead of IMSI, the UE uses SUCI to connect to a 5G SN. This solution piggybacks on existing LTE and 5G authentication and key agreement (AKA) protocols for delivering new pseudonyms to the UE and does not require additional messages. Since IMSI-formatted pseudonym space is small, pseudonyms need to be reused, i.e., disassociated from one user and reassigned to another [22].

https://www.researchgate.net/publication/328781741_Defeating_the_Downgrade_Attack_on_Identity_Privacy_in_5G

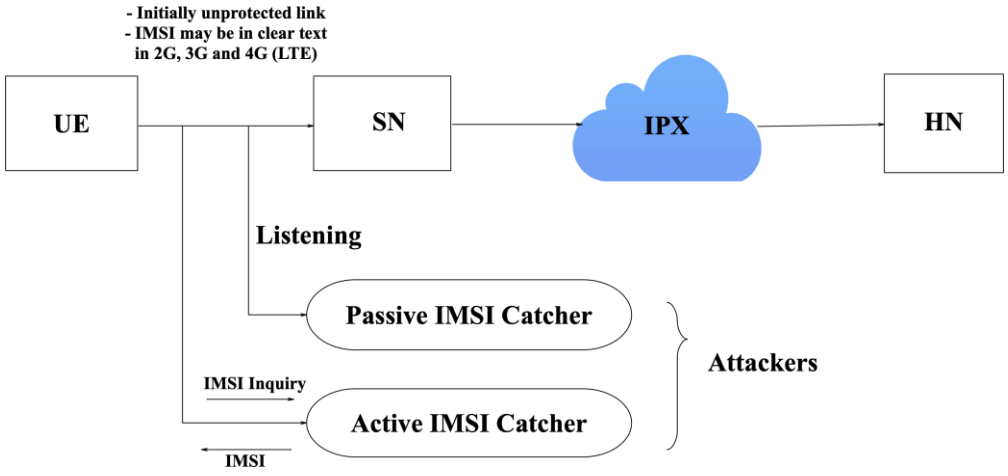

Figure 66 Schematic illustration of mobile network [23]

The figure 66 indicates a high-level mobile network architecture [23]. The UE of a user who has an HN (home network) subscription links to the SN (serving network) in order to obtain services. If the SN and the HN are separate networks, then the UE is said to be roaming. In that case, over the IP Exchange (IPX) network, the SN and the HN

communicate with each other. On LTE, 3G and GSM networks, the connection between the UE and the SN is initially unprotected. Any information sent over this initial connection may be shielded by confidentiality in 5G using the HN's public key pk. Information necessary for IPX routing cannot be shielded confidentially (otherwise, routing will not work). In the subscription database, the HN stores the (IMSI, K) pairs of all its subscribers. There is also a public/private key pair pk, sk on the 5G HN. The UE consists of ME and UICC. The UICC is tamper resistant: without sophisticated instruments, a malicious party cannot read its text. The Universal Subscriber Identity Module (USIM) is an application running on the UICC. The USIM stores the user's IMSI and subscriber-specific master key K in LTE. Note that if a UE uses K to mask the IMSI, the HN does not know which K to use to decrypt the message. In addition, UE is not equipped with any SN-specific keys. The IMSI is therefore initially sent unprotected over the connection in LTE between the UE and the SN. The IMSI is also initially sent unprotected over the connection between the UE and the SN in 3G and 2G. But the USIM can also store the HN's public key pk in 5G; and the UE can send the HN's encrypted (by the key pk) message before identifying itself. When a UE wants to connect to an SN, the SN wants to know the user's identity so that the user can be paid. The SN is forwarding an IMSI inquiry to the EU. Since the relation between the UE and the LTE (also 3G and 2G) SN is initially unprotected, the UE must reply in plain language with its IMSI. Listening to the radio channel, a passive IMSI catcher waits for IMSIs sent in plain text. The UE has no means of distinguishing an active IMSI catcher from a legitimate SN until authenticating the SN. The UE does not have a way of distinguishing an active IMSI catcher from a legitimate SN. Therefore, in plain text, the UE inevitably sends the IMSI to the attacker. Identification is accompanied by mutual authentication based on challenge and answer [22, 23]. The SN supplies the user's identity to the HN,

(i) HN then plan a challenge at random.

(ii) calculates the anticipated challenge response, an authentication token, and an anchor key.

(iii) any other necessary details may be computed. Using the master key K, the answer, authentication token and anchor key are computed.

An authentication token contains data that preserves the challenge's integrity and defeats the replay attack. Collectively, the challenge, answer, authentication token, anchor key and other related details are known as an authentication vector (AV). The HN sends the AV to the SN. The challenge and the authentication token are sent to the UE by the SN. Using the master key K, the UE verifies the honesty and freshness of the challenge. If the verification result is positive, using the master key K, the UE calculates the estimated answer to the challenge and sends the response to the SN. If the expected response received from the HN by the SN matches the response sent by the UE, the authentication will be successful [22].

3-2-2 IMSI Tracking (Privacy):

THREAT: The International Mobile Subscriber Identity (IMSI) is a special SIM (Subscriber Identity Module) card number that the network operator offers to the customer [23]. It includes 15-digit numbers used when it connects to any base station to identify the user unit (mobile tower). Three sections are included in the IMSI number:

A) 3-digit Code for Mobile Nation (MCC)

(b) 2/3-digit Code for Mobile Network (MNC)

C) 9/10-digit Identity of a Mobile Subscriber (MSI)

The IMSI catcher is a radio system designed to grab the IMSI number and intercept cell phone communications with special features. To exploit vulnerabilities in GSM networks and 4G/LTE networks, it can show itself as a fake cellular base station. A popular technique called man-in-the-middle (MITM) assault is used in the IMSI catcher [23].

When the IMSI catcher system is completely attached between the cell phone and the base station, its victims can do almost anything. For example, it can eavesdrop and record calls, intercept and redirect SMS messages, identify the location of the phone user, retrieve files from the target phone, including images, texts, turn on the microphone, camera, other target phone equipment, and so many others. There are numerous IMSI catcher names, such as Stingray, cell site simulator, and emulator of cell sites. There are also several styles that vary in size, price, and functionality. In the beginning, governments and law enforcement authorities

exclusively used such devices to track criminals and terrorists. These devices are now widely available, which means they can be used for various purposes by illegal persons, such as terrorists, criminal groups, drug traffickers and even individuals. Some websites offer IMSI catcher devices online and can be easily designed and configured with the use of available hardware and software with simple manuals. To differentiate real base stations from fake ones, most of the current solutions rely on cell tower features. The functionality of the mobile tower is minimal and can simply be simulated by IMSI catcher devices. This research focuses on a new approach for detecting and resolving IMSI catcher presence in a specific location area based on entire location area features [23].

When a mobile station (MS) attempts to connect to the base station (BS), the base station must be authenticated, but vice versa is not necessary. This is the vulnerability of GSM networks used to exploit network communications through IMSI catcher systems. The IMSI Catcher Software is a radio device that can be shown to exploit and hack vulnerabilities in GSM networks and intercept mobile phone calls and traffic in its vicinity as a cellular base station (BS). It is sometimes referred to as the emulator of cell sites, Stingray, or simulator of cell sites. GSM networks are vulnerable to various forms of attacks, such as breaking encryption, passive interception, and active interception attacks. As an active attack, the IMSI catcher attack is classified. It uses a popular technique called Man-In-The-Middle (MITM) attack, as seen in the figure 3-4, it acts simultaneously as a fake mobile phone to the actual base station and as a fake base station to the real mobile phone.

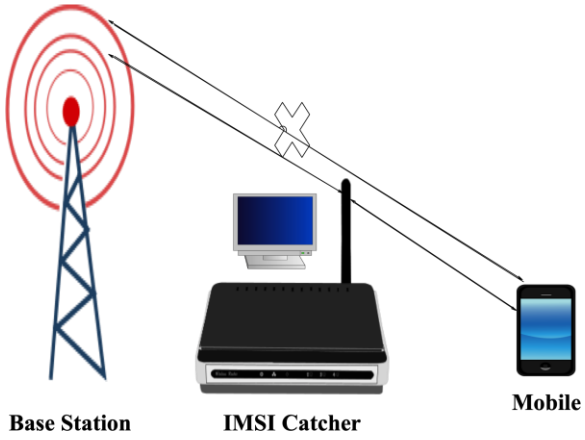

Figure 67 IMSI Catcher Attack (MITM) [23]

The IMSI catcher system can capture and intercept data traffic from IMSI numbers sent by cell phones in its area, which is why it is called the IMSI catcher. StingRay is one of the most common IMSI catchers, and the IMSI number is used to identify every cell phone on the network.

SOLUTION:

- The following diagram in the figure 3-5 depicts the cellular network divided into several location areas, with some base stations serving particular area.

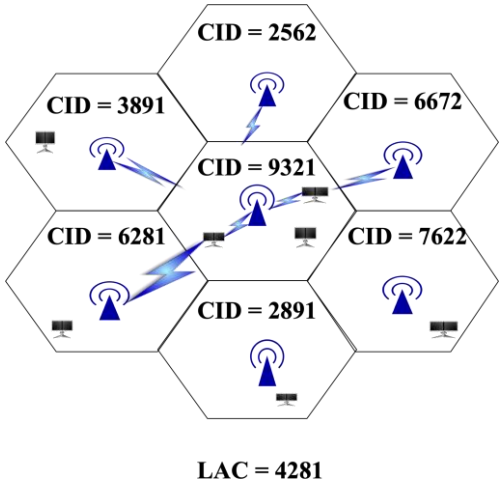

Figure 68 Cellular Network Location Area and its Cells [23]

- Each location area has its own location area code LAC different from other LACs in the network. Each base station inside this location area has also an unique ID number (cell id) CID.
- The LAC and CID are open to the public and can be interpreted in various ways, such as using certain network monitor apps, e.g., Application OpenSignal, App Cellidfinder. The proposed solution relies on finding a fingerprint (cellprint) in the cellular network for each location area and using it to identify the location area and detect any intruder within it before connecting to any mobile tower in that area. Base station features as a parameter to distinguish between real and fake mobile stations. These parameters may, however, simply be imitated by the IMSI catcher system, so it is not a realistic way to use only the cell tower

features for detection [23].

- The solution assumes that a specific cell print for each location area is designed using the characteristics of the entire location area, which includes multiple base stations. It will make it so hard for the IMSI catcher system to prevent detection of the task. The solution relies on finding a fingerprint (cellprint) in the cellular network for each location area and using it to identify the location area and detect any intruder inside it before connecting to any mobile tower in that area. Many of the previous articles suggested using base station features as parameters to distinguish between real and fake mobile stations. These parameters may, however, simply be imitated by the IMSI catcher system, so it is not a realistic way to use only the cell tower features for detection. The proposed solution assumes that a specific cell print for each location area is designed using the characteristics of the entire location area, which includes multiple base stations. It will make it so hard for the IMSI catcher system to prevent detection of the task [23].
- The cell print could be produced using the CID-based cell print generation algorithm by summing all CID numbers for each area of the site. It would be possible to hash and save the product of this summation. Thus, the cellprint for that region will differ from the original when the IMSI catcher system includes itself in a specific area. This enables the mobile user and the network operator, by cross checking the cellprint, to detect and address the presence of IMSI catcher devices in that location. Following figure debates the CID-based cell print generation algorithm.

Algorithm 2: Location-based cellprint generation algorithm

```
1  Start
2  Read the current LAC
3  For each BS belongs to current LAC (has same LAC)
4     Find the minimum spanning tree (MST)
      R = The length of MST
5  cellprint = R  # the value could be hashed
6  End
```

Figure 69 Location-based cellprint generation algorithm [23]

- This implies that there is an intruder within the area if the calculation result for

any position area is greater than the cell print, and the ID number for this intruder is the result of the new summation minus the cell print for that field.

3-2-3 Man in the Middle Attack:

THREATS: The over-the-air Consumer Plane traffic Access Stratum (AS) is not sufficiently protected by Integrity Defense security algorithms [24]. This theoretically leads to a situation where the message and/or contact flow of a customer could be intercepted between the UE and the server in the center. An adversary might manipulate the message and/or communication flow of the customer between the UE and the server. If end-to-end security encryption protocols (e.g., SSL, TLS, IPsec, VPN, etc.) secure the customer's communication, then this attack is impossible. Almost all communications in corporate, business, and social media (e.g., Corporate VPN, Banking, Facebook, Twitter, etc.). End-to-end protocols for security encryption are covered. The danger to customers and their privacy is limited, but worth noting, as not all Internet-destined communications are protected by encryption protocols for end-to-end protection. It is recommended that SSL certificates be checked by the customer. Such certificates should be valid for the traffic of their website and be careful of any websites where SSL certificates can expire, be self-signed or have questionable domain names. Customers can check the Extended Validation (EV), which offers the highest level of site protection, validates websites and/or businesses.

It is possible to sum up some of the strategies used for MITM attacks as follows:

1. ARP spoofing
2. IP address spoofing
3. DHCP spoofing
4. DNS spoofing
5. Gateway spoofing
6. ICMP redirection
7. Stealing of Ports [24]

1.ARP spoofing: The attacker also uses "ARP spoofing" to sniff LAN data frames and change the transmitted packets on the LAN. "ARP spoofing" is often referred to as "ARP Poisoning." The intruder will corrupt the victims' ARP caches that are directly connected and take over their IP address using this technique.

2. IP address spoofing: The attacker creates IP packets with a forged source IP address to mask

the sender's identity or impersonate another device. These types of attacks are usually aimed at manipulating the relationship of trust between victims' endpoints. One of the widely used devices is Hping.

3. DHCP spoofing: DHCP spoofing is an attack on a DHCP server where, using spoofed DHCP messages to gain access to server resources, the attacker tries to trick the server into gaining the IP address [24].

4. DNS spoofing: The intruder sniffs the DNS request ID and sends the victim in front of the real DNS Server with forged responses. DNS service records are often forged by the attacker to redirect the traffic to his servers.

5. Gateway spoofing: The idea is to cheat on the internal network PC gateway. The intruder constantly sends a series of incorrect router MAC addresses to the gateway inside the network so that the actual users are unable to update the address information stored in the router, resulting in the router sending data to the incorrect MAC address, allowing the regular PCs not to receive legitimate contact messages. The idea is to build a fake gateway using a forged gateway to trick the gateway PC to send false data rather than the usual way to access the router.

6. ICMP Redirection: The IP layer uses ICMP to send one-way notification messages to a host. Since ICMP does not support authentication, this can lead to attacks resulting in service denial or enabling packets to be intercepted by the attacker. Gateways commonly use the ICMP 'Redirect' message when a host assumes that the destination is not on the local network.

An attacker sends forged ICMP echo packets to networks' broadcast addresses during ICMP packet magnification. All devices on such networks send ICMP echo responses to the victim, absorb the bandwidth, and create real traffic denial of service.

7. Stealing of Ports: The intruder spoofs the switch forwarding database with this strategy and manipulates the switch port on L 2 switched victim networks to sniff the packets. Flooding the switch with forged ARP packets containing the same source MAC address as that of the victim host and the destination MAC address of the attacking host initiates such attacks.

As a result, the switch will change the MAC address that connects to one of these two ports repeatedly. Real packets from the victim host connected to one port and forged packets from the attacker on another port will be received by the switch. If the rate of arrival of the attacker's packets is faster, then the switch will give the attacker the packets intended for the victim host.

The attacker sniffs the received packet in the process, stops the flooding, and sends an ARP request for the IP address of the victim. The host of the victim sends the ARP response, which is used by the attacker to forward the forged packets to the victim [24].

SOLUTION: The majority of security mechanisms used against MITM attacks are strategies based on authentication, which may be based on the following:

- Passwords
- Secret questions
- Public key infrastructures
- Voice recognition
- Biometrics (Figure printing and Retina scan)
- Off-the-Record Messaging for instant messaging
- Off-the channel verification

A large number of newer techniques that have developed in recent times to provide users with greater login authentication are:

Using multi-factor authentication to secure logins: The simple provision of username and password is not enough to protect confidential data in current times with the advanced existence of security threats. In addition, several nations have implemented security legislation requiring strong authentication mechanisms to be provided by organizations.

Synchronized time, one-time password authenticators: A time-synchronized, one-time password authenticator system provides a unique Symmetric key that is combined with an algorithm to produce every 60 seconds a new one-time password (OTP). This handheld O.T.P. authentication system is synchronized to a security server that tests that 60-second window's password validity. For this reason, the hardware unit used may be as small as a keychain. Such devices may be used by the user to access the Internet from various places, to conduct high-value and high-risk transactions [24].

Software toolbars: There are software toolbars that function as a one-time password authenticator that is installed in a normal Internet browser such as Internet Explorer or Mozilla Firefox, similar to hardware devices. A new one-time password (OTP) is created by the software toolbar every 60 seconds.

Site-to-user authentication: A visible security reminder is provided to users at each login in this technique so that the user is confident that a valid website is being transacted. These security

reminders include a personal security picture and a caption pre-selected by the user at the previous login during the enrollment sessions. Once the website has proved its validity, for further processing, only the user needs to enter their password.

Multi-factor authentication to secure logins: The majority of one-time password (OTP) authentication solutions are based on something that is only accessible to the user as a PIN or something that is only available to the user (an authenticator). Every 60 seconds, the authenticator produces a new code, making it impossible for someone other than the actual user to enter the right token code at any given moment. Users simply combine the hidden Personal Identification Number (PIN) with the token code that appears on the authenticator devices' display at that time to gain entry. This results in a special, one-time combination of passwords that can guarantee the identity of a user confidently [24].

A combination of the following can be used along with the OTP scheme to simplify the procedure and at the same time to harden the system and vouch for the credibility of the user:

The following confidential information pertaining to a person that identifies him uniquely may be used to produce a particular code number:

- A 10-digit number of a bank account
- A Cell Phone Number of 10 digits
- Social Protection The user's number provided by the Government based / nominated agencies.

1. Procedure for the secret key to be generated: Assuming the persons individual details are as mentioned below, the users secrete key can be generated as explained,

   10-digit user bank account – 10 02 02 12 34

   10-digit mobile number - 98 25 01 23 45

   12-digit Aadhar social security no. – ABCD EFGH IJKL

   Picking the character at odd places in the 1 set of

   information we get - 0,2,2,2,4 [24]

   Picking the character at even places in the 2 set of information we get - 8,2,0,2,4

   Picking the character at every third places in the3 information we get C, F, I, L,

   Adding a character in the last place in the third set of information to represent the social id type. In this case it is character A representing the Aadhar unique

id is added and finally we have a matrix of 5 x 3

0,2,2,2,4 8,2,0,2,4 C, F, I, L A

Interchanging the rows and columns of the matrix for first time we get

0, 8, C 2, 2, F 2, 0, I 2, 2, L 4, 4, A

The secrete key can be written as

08C,22F,20I,22L,44A. Dropping the commas we get a 15-digit unique code like 08C22F20I22L44A

Assuming the persons individual details are as mentioned below, the users secrete key can be generated as explained the encrypted message can be decrypted at the receiving end using this hidden key. The secret key should be transmitted to the receiver on the user's mobile phone via offline communication such as SMS [24]. The mechanism referred to above will satisfy the following security dimensions:

- The message's confidentiality is kept since it is an encrypted contact.
- The sending party's credentials are frozen so that the sender is truly known as the true sender and not any unauthorized person or hacker.
- Since the identifiers used to create the secret key are checked by third parties who are legitimate user authenticators, the sender's conditions of authentication and honesty are also met.
- Because the transmission of the secret key from the sender to the receiver is using the cell phone in offline mode, which can be considered as the second type factor and is only visible to the receiver, the foolproof transfer of the secret key to the receiver can be done. Thus, correspondence can be secured, and the integrity of the message is maintained using the above mechanism.

2. Combination with other widely available parameters: In addition to the above, the monitoring mechanism to record the other parameters usually used by the legitimate user to access the internet and perform transactions such as:
    1. The IP address and the MAC address are merged.
    2. The most used operating system and browser mix.
    3. The Position of the Customer.

Any anomalies in the pattern of actions above may be

Processed accordingly to distinguish between lawful users and attacks by MITM. So, we can simplify the process of securing contact based on the above and quickly recognize any unauthorized individuals playing the MITM attacks [24].

3-2-4 LTE Roaming:

THREAT: LTE roaming as given in [25] is highly dependent upon the protocols of SS7 and Diameter. Initially developed in 1975 for the publicly switched telephone network (PSTN), SS7 was the foundation of 2G and 3G network voice communications. Diameter is a protocol of authentication and authorization defined to supersede the RADIUS protocol in 1988. The SS7 and Diameter protocols have both been used on a wide scale, and security flaws have been considered to be the target of attacks for years. Many operators have deployed voice over LTE (VoLTE), which uses Session Initiation Protocol / Real-time Transport Protocol (SIP-RTP) rather than SS7, as a switch from SS7. For authentication, authorization and Policy Charging and Control (PCC) functionality, Diameter is still used in LTE. Diameter and SS7, including voice calls, reading text messages and monitoring phones, are vulnerable to eavesdropping. VoLTE is not accepted by some LTE roaming mobile network operators and mobile virtual network operators, but even though an operator has deployed VoLTE and its customer roams to an MNO/MVNO network that does not support VoLTE, home networks must use SS7 for that roaming customer's voice services. Many operators have SS7 and/or Diameter firewalls, but a range of cross-protocol attacks are carried out against these firewalls.

SOLUTION: The sending node confirmation can be implemented as a partner node for any roaming signaling message exchange filter. In order to prevent illegal authentication, the location of the UE and the travel time for the change in location can also be compared before authenticating the UE. It is also possible for the attacker to send the attachment request claiming to be HPCRF, which can be prevented by testing the messages from the domain address itself. It is also possible to distinguish traffic depending on the subscribers visited and the roaming subscribers themselves. Instead of IMSI, another alternative is to use realm-based communication [25]

The author proposes a solution for scaling authentication requests using machine learning

techniques and algorithms. It is possible to scale a poor application as a negative rank and to scale a true request as a positive rank. Depending on the number of true or false attempts, the rank grows stronger or weaker. The IMSI is blocked or in the red warning zone if the false request seeks to surpass those thresholds. Likewise, if the actual attempts reach any limit, the IMSI is in the green or safe zone. In order to access the network, Green zone users would have more preferences. Targeting intelligent mobile roaming networks.

GSMA implements stronger roaming arrangements under Near Real Time Roaming Data Exchange (NRTRDE). Again, how strictly they obey and execute the requirements is up to the MNOs.

3-2 THREATS, VULNERABILITIES AND ATTACKS IN 5G STANDALONE:

The 5G ecosystem consists primarily of software that can run on general-purpose hardware that interacts with programming interfaces for applications (APIs). An area of vulnerability is the integrity of the software, especially from open-source locations and the general supply chain of software. 5G leverages Cloud-Native concepts where, in a dynamic way, resources can be built, dismantled, and constantly communicate with each other. To avoid unauthorized orders or unauthorized access to resources, all devices must be properly authenticated with secure communication (having the ability to instruct the system to exhaust its resources is one form of DoS attack). Users would have access to network-specific applications within the 5G architecture. In the 5G architecture, any existing hardware or software faults (including operating systems) would still occur. In the end, 5G hopes to use the SDN and NFV definitions, and both come with particular risks and vulnerabilities [22].

3-2-1 DOS/DDOS attack:

THREAT: Using security vulnerabilities, the intruder infiltrates and takes control of the IoT equipment. It creates, as shown in the following figure, its own Botnet network. This Botnet network is so wide in number, latitude and longitude that it is nearly impossible to retrieve it from attackers due to the variety of IoT devices and their geographical reach. Then, TCP SYN, DNS request, ICMP request, SIP request, and so on can attack the victim and interrupt the service, based on virtual requests that can include HTTP traffic. For various reasons, such as economic, political, etc., these bots may be on the dark internet with a very sophisticated device that the attacker takes control of it and executes attack commands on it to keep their core identity hidden. One of 5G's capabilities is to provide mobile equipment with connectivity and

to allow mobile equipment to retain its connection to the network on the move, but to change its geographical location. And carry out their assaults [22, 26].

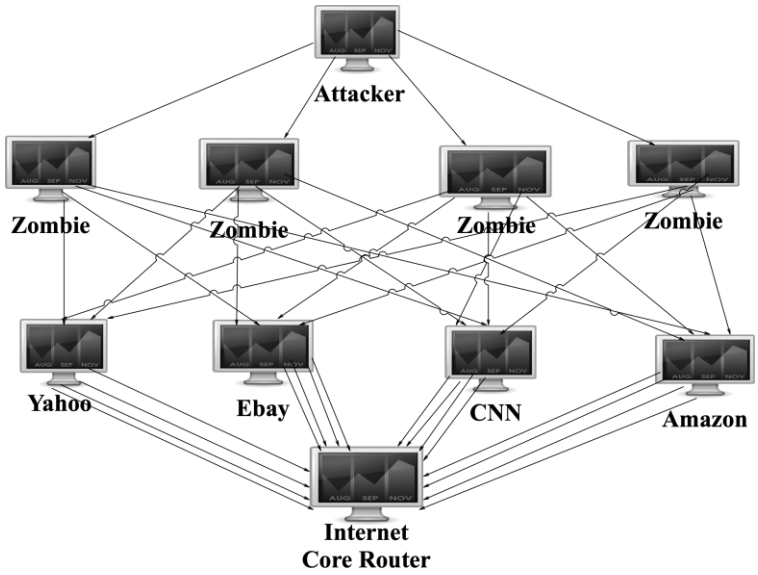

Figure 70 Schematic of a Botnet Network [26]

But here's a new scenario, given the 5G network expansion and interaction with older networks and the 5G function of the Gateway. In this case, the attacker takes control of these points and uses them as a C&C server provided the base stations that provide service on the IoT network, and the key point in this scenario is that because these points are service providers, they actually have their own default network or subscriber customers. So, there's no need to set up a botnet network for an attack to start. The versatility of these points can also lead to mobile C&C servers being developed. On the other hand, older networks can also join this cycle through the Internet and accessible botnet networks can be used [26].

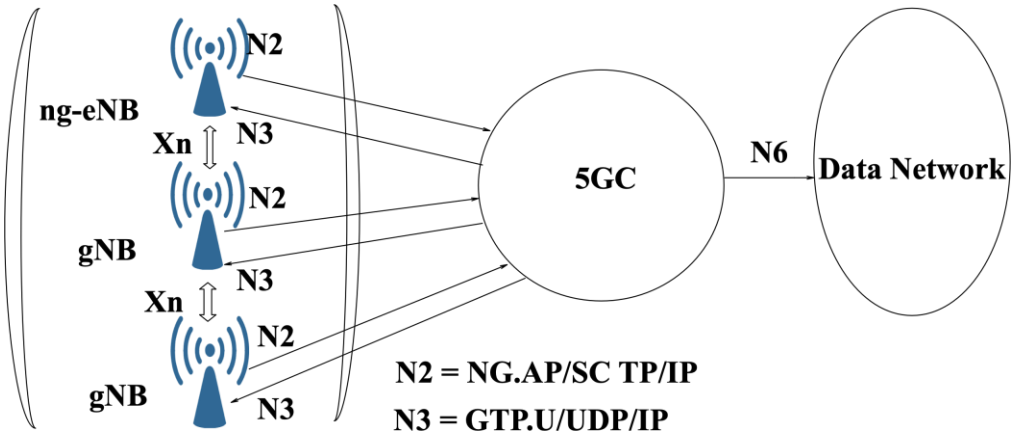

Figure 71 Points for attack in 5G [26]

A Web client typically uses a TCP SYN packet to send a request for a file to the Web server. The attacker sends a TCP SYN that pretends to create a link that makes it a server reserve buffer. The relation is not terminated by the attacker. Instead, it issues more TCP SYNs, resulting in the server naively wasting its memory with connections that have never been completed. Sending such SYN requests at a high rate drains the memory of the server and makes it unable to fulfil legitimate users' link requests. This is an instance of refusing the users' facilities. Other forms of flooding are TCP ACK or RST flooding, flooding of ICMP and UDP echo requests, and flooding of DNS requests. By no way is this list exhaustive. The attacker typically does not use the actual IP address of his or her own computer, but instead spoofs the source address of the attacking packets. This makes it more difficult to track the intruder down. If an attacker uses multiple hosts over the Internet to attack a victim, the DoS attack may be more serious. The attacker normally exploits several hosts in order to do this and deploys attacking agents on them. The intruder warns all agents to initiate an assault on a victim at the same time. This attack is known as a DDoS attack (Distributed DoS). In this case, it is more difficult to track the actual perpetrator, even though it is possible to trace the source(s). A reflector is like a light-reflecting mirror. On the Internet, several hosts can be used as reflectors, such as Web servers, DNS servers, and routers. In response to a question, the servers always address a SYN request. In response to particular IP packets, the routers send ICMP packets (time exceeded or host unreachable). These reflectors can be exploited by attackers to launch DDoS attacks. An attacking agent, for example, sends a SYN request to a reflector that specifies the IP address of the victim as the agent's source address. The reflector will give a SYN ACK to the victim without understanding this. There are millions of reflectors on the Internet, and any intruder can easily use these reflectors by sending large amounts of packets to flood the victim's network [26].

QoS attacks are another threat to computer networks, especially Quality of Service (QoS)-enabled networks such as Differentiated Services (DS) networks. The attacker is a normal network user in this environment, trying to get more money (better service class) than what he/she signed (paid) for. A QoS network offers various service groups at varying prices. Differences in the service classes' charging models will entice attackers to steal bandwidth and other network resources. These attacks allow use of known vulnerabilities to inject traffic or spoof the identity of legitimate customers with high QoS in firewall filter rules. Since the DS

architecture is based on aggregation of flows into service groups, due to the injected traffic, valid customer traffic may experience degraded QoS. The attacks, taken to an extreme, can result in a denial of service. This creates a need for an efficient protection mechanism to be developed that can automate the detection and response to QoS-provided network domain attacks. QoS attacks are divided into two kinds: attacking the process of network provisioning and attacking the process of data forwarding. Network provisioning requires the setup of QoS network routers. You may attack this mechanism by inserting bogus configuration messages, changing the content of actual configuration messages, and delaying or dropping messages. By using encryption of configuration messages of the signaling protocols, networks can be protected against such attacks. Attacks against the data transmission mechanism are of a more extreme sort. This attack involves injecting traffic into the network with the purpose of stealing bandwidth or causing QoS deterioration by causing longer delays, higher error rates and lower throughput for other customer flows. This mechanism calculates parameters such as delay, loss, and throughput in the Service Level Agreement (SLA) and contrasts these metrics with agreed values between the service provider and the customer. A service provider can detect any service breach within its network domain by implementing this monitoring technique. In addition, the monitor will verify whether or not excessive flows passing through its domain are intended for a specific domain of the network. DoS attacks aimed at its domain or any other downstream domains can result from this aggregation [26].

SOLUTION: The ways to detect and prevent DoS attacks are shown in the following figure. There are several ways to detect the source that causes a denial of service (DoS) attack. IP traceback is one of them. IP traceback can be done using either ICMP traceback messages or marking packets at the routers. The marking strategies at the routers can be of a deterministic and probabilistic type. Hash-based IP traceback provides source path isolation engine (SPIE) to track attackers even for low volume of packets. Monitoring a network can help to detect DoS attacks. One obvious way of monitoring is to log packets at various points of a network domain. For a QoS network, service level agreement (SLA) violation detection, can help to detect bandwidth theft and DoS attacks. It is to note that traceback is a detection approach rather than to prevent the attack. Filtering spoofed packets prevents networks from DoS attacks. Ingress/Egress filtering and route-based filtering are two approaches to prevent DoS attack on the Internet [26].

The ways to detect and prevent DoS attacks are shown in the following figure 3-9. There are several ways to detect the source that causes a denial of service (DoS) attack. IP traceback is one of them. IP traceback can be done using either ICMP traceback messages or marking packets at the routers. The marking strategies at the routers can be of a deterministic and probabilistic type. Hash-based IP traceback provides source path isolation engine (SPIE) to track attackers even for low volume of packets. Monitoring a network can help to detect DoS attacks. One obvious way of monitoring is to log packets at various points of a network domain. For a QoS network, service level agreement (SLA) violation detection, discussed in Section III, can help to detect bandwidth theft and DoS attacks. It is to note that traceback is a detection approach rather than to prevent the attack. Filtering spoofed packets prevents networks from DoS attacks. Ingress/Egress filtering and route-based filtering are two approaches to prevent DoS attack on the Internet [26, 27].

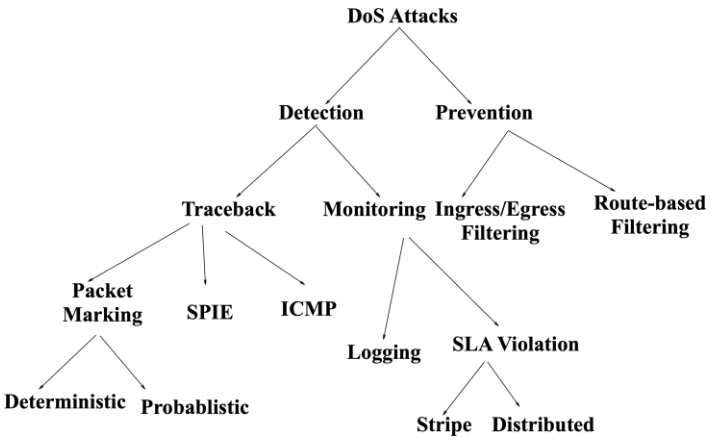

Figure 72 Classification of approaches to detect DoS attacks and service violation [27]

We use Figure 73 to demonstrate different ways to launch attacks and take actions against them.

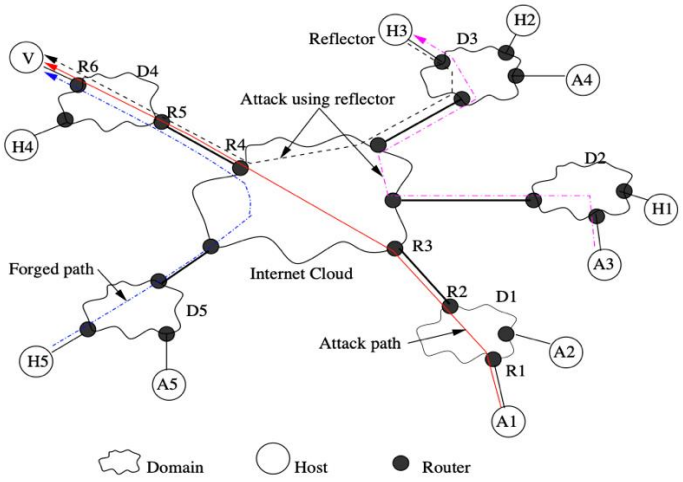

Figure 73 Different scenarios for DoS attacks [27]

In Figure 72, hosts Hs are connected to domain Ds, which are connected to the Internet cloud. When a host sends packets to other hosts on the Internet, it travels through several routers on its way. These routers are potential candidate to help in detecting and preventing DoS attacks. In Figure 73, A represents an attacker and V is the victim. Traceback is an effective scheme to determine the attacking source. It is difficult to trace the attack source because the attacker often spoofs the source IP address. In addition, the Internet is stateless, which means whenever a packet passes through a router, the router does not store any information (traces) about that packet. When a host sends a packet to another host over the Internet, it travels through several routers on its way, and we can trace the network path that the attack traffic follows. ICMP Traceback messages, involves every router to sample the forwarding packets with a very low probability (1 out of 20000) and sends an ICMP Traceback message to the destination. This message contains the previous and next hop addresses of the router, timestamp, part of the traced packet, and authentication information. In Figure 73 [27], while packets are traversing network paths from attacker A to the victim V the intermediate routers, Ri, sample some of these attack packets and send ICMP traceback messages to the destination, V. With enough ICMP traceback messages, the victim later can trace the network path, V-A1. This work shows a promising solution for constructing path from victim to the source involved in attacking. The disadvantage of this approach is that sometimes ICMP packets can be ignored at routers and these traceback packets can be dropped. The attacker/source can defeat the authentication mechanism by sending many false ICMP traceback messages to confuse the victim, since the routers send only few messages. In Figure 2, A3 sends a SYN request to specify as the source

address of this packet. H3 sends a SYN ACK to the victim V. According to the modification, routers on the path A3-H3 will send ICMP messages to the victim. This reverse trace enables the victim to identify the attacking agent(s) from these trace packets. The reverse trace mechanism is helpful for defending against DDoS attacks by reflectors and depends only on the number of attacking agents rather than the number of reflectors. This achieves scalability because number of available reflectors is much higher than number of attacking agents on the Internet.

3-3-2 SDN:

THREATS: Security risks from SDN would be much more concentrated relative to conventional network architectures, as opposed to the dispersion seen in the network elements of traditional networks. SDN's natural protection defects include [28]:

a) Vulnerable controller: The SDN controller focuses on most tasks, such as network information collection, network setup, and routing calculation. The nature of SDN provides a more oriented target for, and significantly decreases the complexity of such attacks. At the same time, cloud infrastructure architecture provides the attacker with very large-scale computing skills; attackers can easily execute attacks with the help of cloud computing platforms. If the attackers capture the controller of an SDN successfully, they will cause major paralysis of the network services and impact the entire network protected by the controller.

(b) Risks caused by open programmable interfaces: SDNs are more vulnerable to security threats because of their open existence. Second, it renders the SDN controller's software vulnerabilities completely exposed to attackers, since the latter would have enough data to devise an attack strategy. Second, a large number of programmable interfaces for the application layer are provided by the SDN controller, and this level of transparency can lead to interface misuse, such as embedding malicious code, such as a virus [28].

c) More points of attack: As the SDN is split into three layers, it is possible to spread the entities of each layer across various network locations; communication between these entities would be essential and regular. SDN therefore offers more potential attack points for attackers relative to conventional networks, as shown in figure 3-11 below. We point out six possible attack points using red stars in the SDN architecture in the figure, and we will identify them in the order label shown.

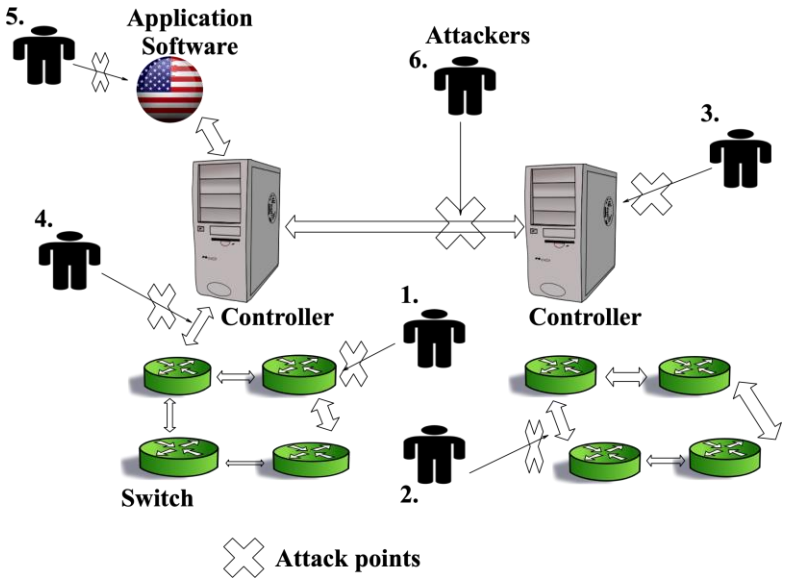

Figure 74 Possible attack points in SDN architecture [28]

- The SDN switch: Usually, an SDN switch is a separate system consisting of associated hardware and software that is vulnerable to attacks. The size limitation of Flow Tables is an illustrative weakness. Nearly all data packets sent between SDN switches are not encrypted and may contain confidential user information. These packets, particularly when the connections between switches are wireless media, can be easily intercepted by attackers.
- The SDN Controller: As mentioned earlier, the most desirable target for attackers is the controller. The software of the controller is unavoidably vulnerable due to the openness of programmability and the sophistication of its features, and this can be used for malicious attacks [28].
- The links between the controller and the switches: All forwarding rules are introduced by the controller into the switches. The data packets containing these rules can be exploited by the attacker by eavesdropping on the connection between the controller and the switch, resulting in the introduction of a bogus rule or the alteration of a malicious rule. Once the switch has installed deceptive rules, the data packets will not be forwarded correctly.
- The links between controllers: In a multi-controller system, it is important to communicate between different controllers in order to maintain the consistent state of the entire network. It is possible to intercept the data packets in the links

between the controllers, which may provide the attackers with possible clues to compromise the controllers.

- Application software: The application software is mounted directly on the controller and is normally installed on the same physical computer as the controller. Malicious code may be inserted in the controller when the application program invokes the controller's functions via the north-bound APIs. Therefore, the program for the application is considered the most convenient attack point for seizing the controllers [28].

3-3-2-1 Threats to the data forwarding layer:

At the bottom of the SDN architecture, the data forwarding layer is located and comprises thousands of interconnected switches. It is the duty of these switches to forward packets. The packets that pass through it will not be forwarded correctly if a switch is compromised. Furthermore, switches are the direct entry point for end-user network access, and attackers can assault a switch by simply adding a connection to a switch port. Recognizing security risks and identifying effective countermeasures for SDN switches is therefore very important. As seen in the following figure 75 [28], we consider the design and working concepts of SDN switches adhering to the OpenFlow specification. There are usually three feature modules in an OpenFlow switch, namely the OpenFlow client, the Flow Table and the Flow Buffer. If the switch receives a packet from an input port, the packet will be placed in the Flow Buffer and the Flow Table will look for a rule matching the message fields of the packet, such as a MAC/IP address and a TCP/UDP port. The packet will be removed from the Flow Buffer and forwarded to the output port if a suitable rule is found. Otherwise, the switch sends a Packet In message to the controller via the OpenFlow client to request instructions. The controller will perform a routing calculation after receiving the new message and insert a new rule into the Flow Table. Three key security threats can be defined according to the above process; they are a man-in-the-middle attack between the switch and the controller, aimed at interfering with rules, a DoS attack to overflow the Flow Table, and a DoS attack to overflow the Flow Buffer.

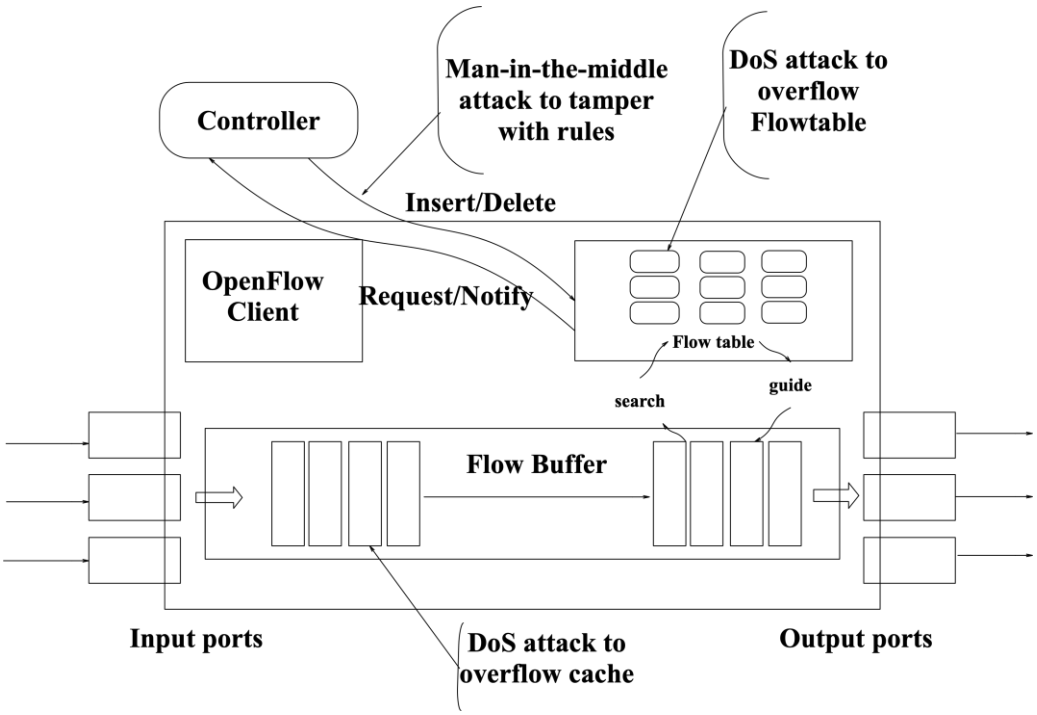

Figure 75 Working principles and security threats of OpenFlow Switches [28]

1. Man in the middle attack:

   THREAT: A man-in-the-middle attack is a traditional form of intrusion into the network, the key idea of which is to insert an agent node between the source and the destination node, which is used to intercept and exploit communication data without either communicating side noticing it. Unique attack methods include session hijacking, DNS spoofing, port mirroring, and so on for man-in-the-middle assaults. In order to gain control of network packet forwarding, a man-in-the-middle attack between the controller and switches is a perfect option for targeting an SDN, as it can be used to intercept and tamper with the forwarding rules given to the switch. Further attacks, such as black-hole attacks, may be introduced by attackers after this has been accomplished. Furthermore, we know that the controller and switches may not be physically directly connected, i.e., a packet may pass through several other switches from a switch to the controller. Therefore, in a man-in-the-middle attack, all switches and the hosts linked directly to them on the communication path are likely to be transformed to agent nodes [28].

   SOLUTION: Creating a safe channel between the controller and switches is the most

obvious solution. Transport Layer Security (TLS) was used in the OpenFlow v1.0 specification to safe controller-switch communication. Configuring TLS, however, is very difficult, and many vendors do not have TLS support for their OpenFlow switches. Later versions of the OpenFlow specification would then announce that TLS configuration is optional. In addition, no TCP-level protection can be offered by TLS, which means that the network is vulnerable to TCP-level attacks. In this situation, since TLS is not implemented, our primary safety challenge is to differentiate between regular and forged flow rules and to remove forged rules before they cause adverse effects [28].

Some alternative countermeasures to this challenge have been put forward. FlowChecker is a configuration validation tool capable of accurately detecting switches' internal configuration errors. In particular, it first generates models of all the interconnected switches and then, through a binary decision diagram and model testing technology, conducts end-to-end rapid analysis and inspection for all switch configurations through which misconfigurations can be identified. FortNOX offers a role-based authorization and authentication security enhancement strategy as a software extension module of the NOX controller. It can detect collisions with different forwarding rules through its novel analysis algorithm. The algorithm has strong robustness and, even in cases of malicious application attacks, executes its functions correctly. At the same time, FortNOX can check the validity of the changes through digital signatures or security restrictions before the applications change the forwarding rules. VeriFlow serves as the middle layer between the controller and the switches and is primarily responsible for dynamically checking network variables within the network as a whole, in particular when a new forwarding rule is inserted. Experiments have been conducted based on the Mininet, an OpenFlow simulation environment, and results show that VeriFlow can finish the detection of a new forwarding rule within a few hundred milliseconds by monitoring routing data, which is very effective.

Since controller communication is very important for the proper operation of switches, it is beneficial to mitigate the effects of man-in-the-middle attacks between the controller and the switches via redundant links or fast connection recovery mechanisms. The OpenFlow protocol itself has mechanisms for link stability checking, whereby each

switch periodically sends messages maintaining operation to the controller. It will automatically instruct the switch to connect to a backup controller if the master controller fails to respond. That is to say, if, within a certain period of time, the switch does not receive a response from the controller, the switch assumes that the controller has failed and will quickly connect to another controller, enabling the network to run continuously [28].

2. DOS attack to saturate the flow table and flow buffer:

   THREAT: OpenFlow's reactive rule architecture makes the switch susceptible to Denial of Service (DoS) attacks. Since packets with an unknown destination address can trigger a new rule to be introduced into the switch, in a short time, an attacker may produce large quantities of packets intended for unknown network hosts, quickly filling up the limited storage space of a switch's Flow Table. If the flow table is saturated with irregular traffic, the legal traffic will not be forwarded correctly, as the option to insert new rules will no longer be usable. Except for the Flow Table, the Flow Buffer is another object of DoS attacks. As mentioned above, they are buffered in the Flow Buffer until the packets are forwarded, waiting for the outcome of a rule quest or the insertion of a new rule. To free the storage space, packets in the Flow Buffer will be marked for deletion on a First in First Out (FIFO) basis. As in the case of the Flow Table, the Flow Buffer's storage capacity is also reduced. Large packets belonging to a different flow than that usually experienced by the switch may be flooded by attackers; the switch has to buffer these large packets, and this contributes to the Flow Buffer saturation. The Flow Buffer will not have enough space to store these packages when valid packets are received, and these new packets will have to be dropped [28].

   SOLUTION: FlowVisor can allow network operators to discern network packets according to packet header fields. FlowVisor serves as an agent between the switches and the controller; it recognizes and rewrites controller rules such that only the portion of the network that a controller is permitted to control is influenced by the resulting rules. For instance, the network segment comprised of all traffic to and from the web servers of an entity may be assigned to a controller. In response to a DoS attack, this controller would then create a rule to drop all UDP traffic. Once this rule is obtained by FlowVisor, it rewrites it to drop all UDP traffic to and from the web servers, leaving

the rest of the network unaffected.

The Virtual Source Address Validation Edge (VAVE) is a preemptive OpenFlow/NOX architecture security system designed to mitigate DoS attacks triggered by IP spoofing. In order to verify the source address, a new packet that does not fit any rule in the Flow Table will be sent to the controller during which IP spoofing can be identified, in which case the controller creates a rule in the FlowTable to avoid the particular flow from that source address. In the study of attack detection, this approach demonstrates good efficiency.

The use of detection systems for intrusion may help identify irregular traffic flows caused by DoS attacks. Such systems could be combined with similar mechanisms for dynamic access control of the behavior of the switches, such as rate limits for requests from the control layer. Resonance is such a mechanism that can reinforce the controller's complex access control policy. Based on real-time alerts and packet flow level information, the device issues dynamic safety policies directly to the forwarding data layer in the SDN architecture [28].

3-3-2-2 Threats to the control layer: SDN was initially developed as a single controller architecture, which lacks scalability and reliability, to mitigate the risk of having a single point failure in the controller. Therefore, a distributed control solution (controller clusters) was proposed, in which each individual instance of the controller acts as the master of certain switches and different controllers would interact with each other in order to manage the entire network collaboratively. However, the data forwarding layer should be visible to several physical controllers controlling the network instead of a single one, meaning the controllers need to act as a single controller for the entire network. In this case, many security concerns, such as authentication, authorization, and privacy issues during network information transmission, would need to be addressed by an application that spans several network control domains. Furthermore, the dynamic switch-over of the master controller and the coexistence of multiple controllers in a single network domain can cause configuration conflicts with the distributed collaboration of multiple controllers. Therefore, an inconsistent configuration is also a secret security hazard within the multi-controller architecture.

1. DOS/DDOS attack on controller [26, 28]:

   THREAT: Using their own host or controlling other distributed zombie hosts, an

attacker might generate enormous flooding traffic to an SDN-enabled network in a short time. This traffic will be mixed with regular traffic, and the distinction between the two forms will be difficult. If a switch does not know how to handle a new packet, it will first store this packet in its Flow Buffer according to the OpenFlow specification and then send a Packet In message to the controller to request instructions. The controller will therefore have to deal with an enormous number of Packet-In messages created by the flooding traffic in a short time in the case of a DoS attack, which can lead to an exhaustion of resources for processing normal traffic. At the same time, the bandwidth will be entirely filled by the attacking traffic between the controller and the switches, and this would significantly reduce the efficiency of the entire network. Following diagram explains the attack.

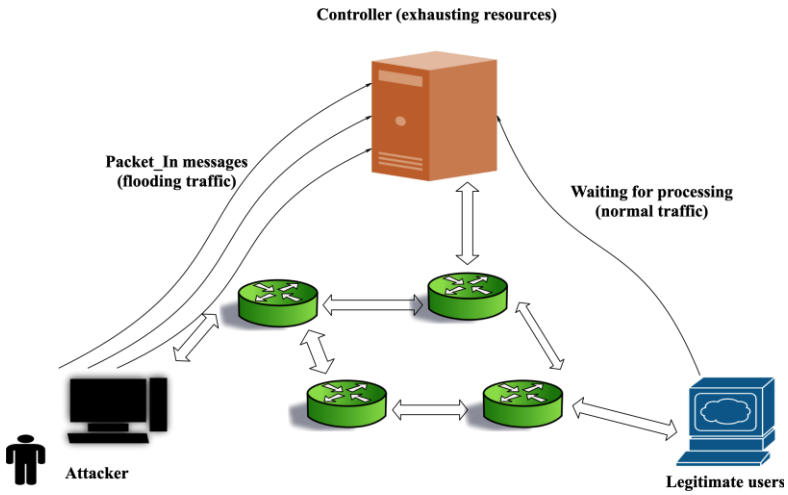

Figure 76 DoS / DDoS attack on controller [28]

SOLUTION: We may analyze the characteristics of traffic flows stored in the OpenFlow switches to detect this form of attack in order to mitigate the threat of DoS / DDoS on the controller. FloodGuard, which is protocol independent, is an SDN-oriented lightweight security system. There are two program modules in FloodGuard, the Active Flow Analyzer and Packet Migration, respectively. The Active Flow Analyzer performs a dynamic analysis based on the controller's real-time running logic to ensure the protection of the SDN, so that traffic flows triggered by DoS attacks can

be detected. Packet Migration is responsible for buffering the packets received and sending them via a rotation scheduling algorithm to the controller for processing at a restricted pace, which prevents the controller from consuming too much computational resources. The Packet Migration module will redirect a table-miss message to the data forwarding layer when the DoS attack is discovered. At the same time, the current network flow will be monitored by the Active Flow Analyzer to evaluate a set of sensitive parameters or variables from which the controller will create forward flow rules and proactively install them on the switch [26, 28].

A DDoS is a more effective DoS attack, and its concept is that a large number of compromised hosts can be hijacked and managed by the attacker to simultaneously generate large-scale distributed attack traffic to attack a target, such as the controller. Also, the DBA runs on the controller and, via a Locator/ID Separation Protocol (LISP), distinguishes between regular traffic and attack traffic. When the location of the network node changes, the DBA notifies the controller by locator of the corresponding update, which is the clue to the attack detection. Furthermore, if the transmission rate of the traffic reaches a certain amount, it will be considered by the controller as assaulting traffic and these packets will be dropped directly.

A Content-Oriented Networking Architecture (CONA) is a proxy node that can communicate with the controller and is situated between the client and the content server. In order to minimize the harm of DDoS attacks, content request messages from clients are intercepted, analyzed and filtered by CONA. A DDoS assault is considered to be in progress when the rate of request messages arriving at the content server reaches a certain value. In order to avoid spreading the attacking traffic, the controller will send a message to each relevant CONA agent and usual messages will be forwarded to a new server address. SDN is therefore ideal for defending against DoS attacks, and potential DoS security solutions can be built on the SDN architecture application layer [28].

3-3-2-3 Threats to the application layer:

1. Illegal access:

THREAT: According to the specification of OpenFlow, applications running on the controller are very versatile and extensible and have rights to access network resources and control network behavior. However, most of these applications are developed by

third-party entities, not controller vendors. Therefore, the lack of a consistent protection framework for SDN applications creates severe security threats.

SOLUTION: PermOF is a fine-grained permission scheme capable of providing OpenFlow controllers and applications running on top of it with privilege control. PermOF summarizes a collection of 18 permissions that the controller APIs had to implement and also proposes a tailored isolation system that maintains different application goals and isolates control traffic from data traffic in order to achieve robust resource isolation and control of access [28].

NICE is a new approach for implementing model checking with symbolic event handler execution that can rapidly explore the state space of unmodified controller programs written for the popular NOX platform. In order to check their accuracy, NICE can also be used as a method to automate testing of OpenFlow applications. Verificare is an instrument for using systematic verification techniques to model distributed systems. Authors demonstrated this tool by iteratively modelling an OpenFlow network to check its correctness and critical properties. VeriCon is a method of verification that can check the correctness of the programs of the controller. Via first-order logic and desired network-wide invariants, VeriCon implements classical FloydHoare-Dijkstra deductive verification. Experimental results show that VeriCon can easily verify accuracy and recognize bugs in large-scale SDN applications.

3-3-3 NFV:

THREATS: There are three attack profiles in NFV include,

- Intra-MVNO attacks.
- Inter-MVNO attacks.
- Attacks by end users [29].

Intra-MVNO attacks include attacks on an MVNO by its own employee to occupy and degrade network services. Inter-MVNO attack refers to any sort of attack from one MVNO against other MVNO(s) in order to extract competitor's information, corrupt or abuse their services. The last group includes the attacks that are triggered by mobile network end users within same MVNO or other MVNO.

In a cloud environment with NFV, network functions will be deployed as vNFs that carry security challenges such as,

1) Malicious loops that are triggered by routing loops, unavailability of management network due to network failure: To avoid these risks, network should be logically checked to be sure that management interfaces are available even if vNFs are down.

2) Inappropriate data removal due to VM crash, execution of malicious vNF and thus unauthorized modifications to BIOS/UEFI, hypervisor and OS: For mitigation safe boot i.e., trusted platform module (TPM) and crash security can be used.

3) Misuse of hypervisor resources by malicious VM (impacting other VMs) and quality of service (QoS) degradation: Performance isolation by segregating resources to each VM is recommended as prevention mechanism.

4) Insufficient vertical and horizontal VM AAA mechanism: to prevent this threat, AAA mechanisms among vNFs, between vNFs and application layer and between vNFs and management stations should be revised [29].

5) Software unreliability such as:

- Coding vulnerabilities that affect all MVNOs using same software: Correction and protection patches should be implemented on all VMs using same software.
- Configuration changes or correction patches that require reboot and trigger service interruption on MVNOs: Backup and load balancing are the mitigation mechanisms for such threat.
- Test and monitoring backdoors: Closing test and monitoring and debug interfaces is recommended.
- Stored password and private keys in VM images: Using unique private key for each image may avoid these threats [29].

Both the integrity of the code that comprises a virtualized function and the interaction between virtualized functions themselves is important. Open- source software is a concern in any environment where it can be used. The functions that comprised the 5G system can be composed of open-source software elements, but their security and integrity is not always known. The virtualized elements must communicate with each other in a standardized, API-style environment. The APIs themselves must adhere to standards but must also have safeguards in place to avoid being manipulated in unintended ways to cause disruption.

SOLUTIONS:

- Boot Integrity measurement using TPM: The measurement of device-sensitive components such as platform firmware, BIOS, bootloader, OS kernel and other system components can be safely stored and tested using the trusted platform module (TPM) as a hard-ware root of confidence. Only when the device is reset or rebooted can the platform measurement be taken; there is no way to write the new platform measurement in TPM during the run-time of the system. [29] The validation of the measurements of the platform can be carried out by the launch control policy (LCP) of TPM or by the remote attestation server.
- Protection of the hypervisor and virtual network: The hypervisor allows virtualization between the hardware and VMs underlying it. In the cloud, virtual networks use SDN to allow communication between VMs and outside networks as well. In order to secure the entire infrastructure, the protection of these elements is a must. One of the best practices for protection is to keep the hypervisor up to date by installing the released security patches on a regular basis. Failure to do so will result in potential exposure to safety risks. Disabling all programs that are not in operation is another best practice. SSH and remote access servers, for example, might not be required at all times, so it would be a good idea to allow these services only when necessary [29]. The gatekeepers of the entire infrastructure are cloud administrators, and the keys are their passwords. Safe admin accounts should be mandated by applying a firm password policy along with strictly following the security guidelines of an organization.
- Protection zoning: It is good practice to distinguish VM traffic and management traffic to prevent a VM from influencing other VMs or hosts. This will stop attacks by VMs that tear into infrastructure management. The division of VLAN traffic into groups is also a good idea and all other VLANs that are not in use are disabled. Similarly, it is possible to group VMs with similar features into separate zones and isolate their traffic. Based on its appropriate security level, each zone can be secured using access control policies and a dedicated firewall. A demilitarized zone is one example of these zones (DMZ).

- Linux Security: The host system kernel is a very important component of virtualized platforms, providing isolation between applications. The SELinux module, developed by the NSA, is implemented in the kernel and provides robust isolation between tenants when the host uses virtualization technology. Secure virtualization (sVirt) is a new type of SELinux that has been developed to combine compulsory access control protection with hyper-visors based on Linux. SVirt offers separation between data files and VM pro-ceases. Other kernel hardening tools may be useful outside these tools to protect the Linux kernel. Hidepd, which can be used to prevent unauthorized users from seeing the process information of other users, is a notable example. GRSecurity [29], which provides protection against attacks on corrupted memory, is another case.
- Introspection of hypervisors: Introspection of hypervisors may be used to scrutinize applications running within VMs to identify suspicious activities. It functions as a host-based IDS that has access to all VM states, so that it is not possible to easily hide the root kit and boot kit within VMs. Using introspection capabilities, the functionality of the hypervisor is improved, allowing it to track network traffic, access storage files, and read memory execution, among other things. Introspection APIs for hypervisors are important tools to perform deep VM analysis and potentially improve the security of VMs. They can also, however, be used as an exploit that allows the separation between VMs and the hypervisor to be broken and bypassed. LibVMI [29] is a hypervisor introspection library for different platforms implemented with Python bindings in the C language. It provides the hypervisor with the means to deeply inspect VMs (e.g., memory checking, vCPU register inspection, and recording trapping events).
- VNF volume encryption: Virtual volume discs associated with VNFs can contain confidential information. They need, therefore, to be covered. Through encrypting them and storing the cryptographic keys at secure places, the best practice for safeguarding the VNF volume is. To securely store these keys, the TPM module may also be used. In addition, in the event that a VNF is crashed or purposely damaged to avoid unauthorized access, the hypervisor should be

configured to safely wipe out the virtual volume disks. VM swapping is a method for memory management used to transfer memory segments from the main memory to the disc, which is used as a secondary memory to maximize the performance of the machine in the event that the system runs out of memory. These segments of transferred memory may contain sensitive data, such as passwords and certificates. Even after rebooting the system, they can be stored on the disc and remain permanent. This allows for an attack scenario that copies and examines a VM swap to recover some useful information. Encrypting VM swap areas is one way to prevent this sort of attack. For this reason, Linux based tools such as dm-crypt can be used [29].

- Security and Orchestration Management: One best practice is to develop an NFV orchestrator that implements the NFVI's security and trust specifications. Security feature orchestration and management includes integration by allowing interaction between the security orchestrator, the VNF manager and the element management systems (EMS). For example, setting scaling limits in the VNFD or network service descriptor (NSD) and making the NFVO enforce these constraints to protect against attacks such as a DNS amplification attack can achieve this form of security.

3-3-4 Cloud security:

THREATS:

1. Weak Identity and Access Management: To respond to the 5 W's (Who, what, where, where, why) of resource accessibility, identity and access management are very important. Poor management of identity and access can result in poor management of
    - Account Hijacking: If the Cloud Provider Console or API keys are lost, the cloud environment can be managed by a malicious entity outside an organization [30].
    - Data Breaches: Inadequate access control of object storage buckets and data stores leads to the publication of confidential information, which has been one

of the key causes of cloud data breaches.

- Malicious insiders: The loss of sensitive data and systems will result from malicious insiders attempting to take over admin/root privileges.
- Cloud services misuse and nefarious use: Cloud account hijacking will result in the malicious user using the compromised tools to launch DDOS, spam and phishing campaigns, making the company vulnerable to legal liability.
- Insufficient due diligence: Companies that manage information and comply with regulatory enforcement laws need a specific strategy to move to the cloud, otherwise this presents a danger to security and legal liability [30].

2. Risks to workload:
    - Persistent Advanced Threats (APT): Malware and Advanced persistent threats once enter an environment, adapt to the security measures and over time gain a foothold in the environment and propagate itself laterally and once it reaches the intended goal, it will exfiltrate sensitive data. These threats are difficult to identify and remediate.
    - Vulnerabilities: With multi-tenant cloud services, vulnerabilities that involve privilege escalation and VM boundary jumping can cause data breaches and leave vulnerable apps and workloads.
    - Insecure application services: Insecure APIs that support a separate application service will leave the application vulnerable to known attacks, resulting in downtime of the application or data breaches [30].
3. Risks from networks:
    - Attacks on DOS and DDOS: Poor network segmentation and firewall management lead to cloud resources being targeted by DOS and DDOS attacks which result in poor application performance and even application downtime.
    - Data exfiltration: Inadequate controls of outbound firewalls lead to attempts to exfiltrate data from compromised workloads.

SOLUTION: Traditional workloads tend to be long-lived, so an appropriate solution may be a full stack security solution with a large footprint, while cloud servers tend to be immutable and security solutions with a large memory/disk footprint are deployed, kernel dependence tends

to consume computing resources and is a ban on fast agile cloud deployment workflows.

The above-mentioned solution architecture indicates that it is very simple to replace vulnerable/compromised servers rather than patch them with a different approach to cloud protection, which is more detection-based rather than prevention.

The solution is constructed on cloud servers using both principles, i.e. Servers that are immutable and Zero-Trust [30].

1) Immutable Servers: As one of the remediation actions, the approach uses the idea of immutable servers. The basic principle is that it is easier to build a new one with pre-defined security controls and policies instead of modifying the current server when a possible threat is detected, as it minimizes the challenges of configuration management and increases the infrastructure's reliability.

2) Zero-Trust: The solution continuously analyses settings, user access logs, network logs, endpoint logs to give us the ability to implement the cloud security principle of Zero Trust, i.e., 'never trust, always verify.' We may take extra measures with persistent servers on the cloud to lock down settings, workloads, software, and users to the least privileged, established and agreed use patterns. In the above solution, any deviation from the preconfigured policies without further decision-making is considered to be out of compliance, which immediately activates programmed remediation behavior such as network isolation, termination or privilege revocation as appropriate to the resource.

3) Cloud Threat Defense Model [30]

1) Discover Assets: Cloud Vulnerability Protection identifies cloud workloads, access logs, network logs, antimalware endpoint security solutions logs and syslogs. It uses cloud-native services such as CloudWatch, CloudTrail, VPC, AWS Lambda, Kinesis, S3, Redshift, and Machine Learning to create a data lake that can be linked to analytics and machine learning to detect patterns and threats on Endpoint and cloud services.

2) Assess Security Posture: Cloud Threat Protection evaluates the cloud workloads created in terms of firewall controls on the network segment on which it is installed, known vulnerabilities based on the versions of the OS/kernel and the status of anti-malware solutions deployed by third parties.

3) Identify Security Threats: Applies data analytics to Cloud Vulnerability

Protection to define possible security vulnerabilities and provide a risk score for the identified threats, and machine learning on the information obtained from the above steps in the security model. It also recommends effective remediation steps to minimize and eliminate similar potential safety problems.

4) Remediate Security Issues: Cloud Threat Protection offers corrective steps to address the problems detected and provides the mechanism through mechanisms such as isolating vulnerable resources, upgrading endpoint security solutions, auto-correcting any unsafe changes to firewall settings to auto-remedy the problems.

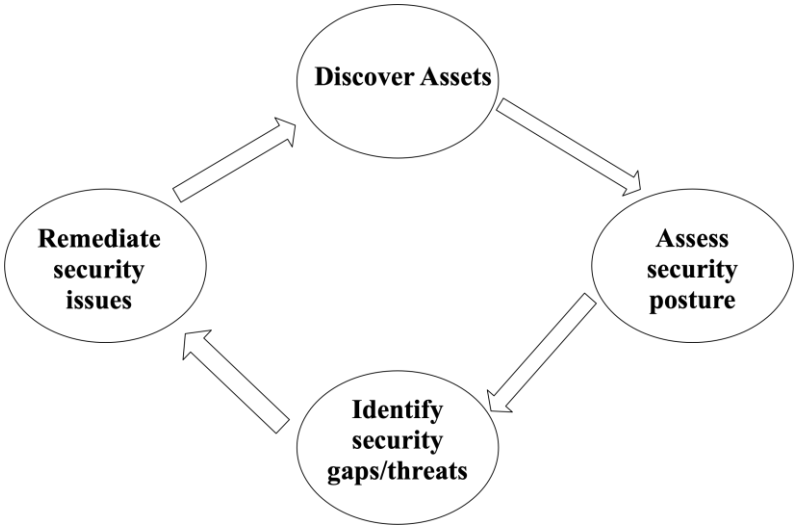

Figure 77Cloud Threat Defense – Security Model [30]

3-4 Layer wise IoT threats and solution:

3-4-1 THREATS AT SENSING LAYER:

Here, in Sensing/Perception Layer, sensors and RFID tags will sense and monitor the surroundings, collect the data and exchange them among devices. These devices have designed with constraints in resources and have limited bandwidth and network connectivity. Here, security considerations classified to two parts: security at end-node and security at sensing layer. Security concerns at end-node considers physical protection of the device, access control, authentication, non-repudiation, confidentiality, integrity, availability and privacy. For instance, possible threats can be unauthorized access, availability, spoofing attack where the attacker will masquerade as an IoT device with false data, routing attack, transmission threat and malicious code using Virus, Trojan and junk data [31]. One of the main challenges at this

layer as is detection of sensor node that behaves abnormally and the type of the encryption algorithm and key management mechanism and vulnerabilities in devices [34]. So, the possible threats at this layer can be:

- Node Capture: In this attack, the intruder takes the control of the key node like Gateway and causes data leakage while sender and receiver are communicating and causes the whole network in danger.
- Fake Node: Here, a fake malicious node will be added to the network and can perform or run malicious codes and program and infect the whole network. So, a fake node with fake data will be added while preventing the real information to circulate in the network and takes the energy of other real nodes.
- Replay attack: Here, the attacker will eavesdrop or listens to conversion between both ends and captures the authentication information of sender and next time, it sends the same authenticated information to the victim target while the receiver will not be notified it is a fake sender.
- Denial of Service: It causes the resource outage and service unavailability.
- Spoofing attack: Here, the malicious node masquerades itself as IoT node or gateway node with false data.
- Transmission threat: It covers the attacks like data manipulation, interrupt and blocking.
- Routing attacks: It includes altering routing information, selective forwarding attacks, Hello Flood attacks, etc.
- Sleep deprivation attack: The attacker will maximize the power consumption of nodes, so their lifetime will be minimized and finally shut them down.

However, the research in [34], separated the attacks based on IoT end-node attacks and sensing layer attacks which can be shown in the figures 78 and 79.

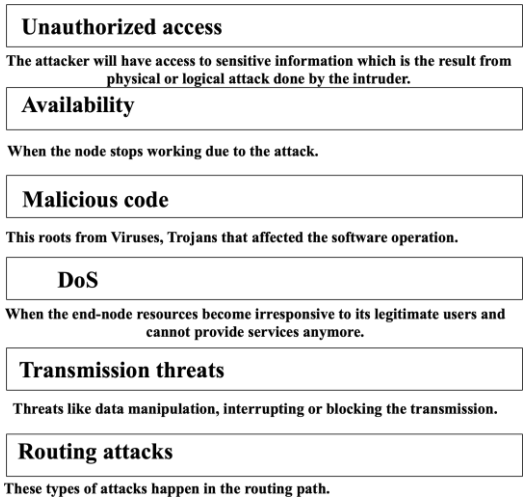

Figure 78 IoT end-node attacks [34]

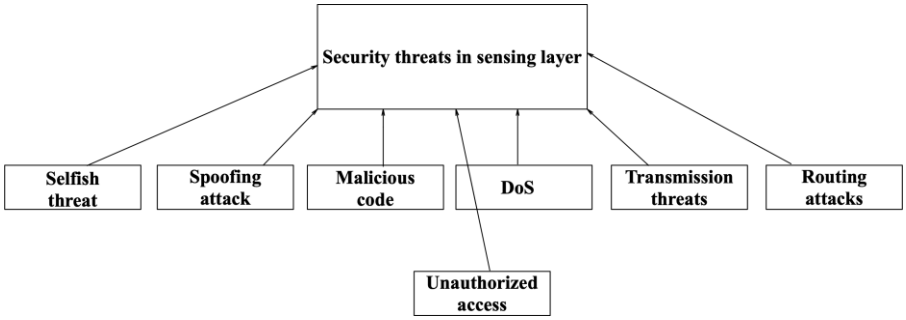

Figure 79 IoT Sensing layer attacks [34]

3-4-2 THREATS AT NETWORK LAYER:

In this layer, the information taken from sensors will be transmitted over wired or wireless medium. Here, the devices are interconnected, and data can be aggregated and delivered to other layers. Since the network deployments can be varied, it causes different security problems. So, challenges like network management technologies, energy efficiency in the network, quality of service, confidentiality of information, security and privacy are the top challenges at this layer. So, the type of attacks this layer may deal with can be eavesdropping, Man-in-the-middle, network intrusion and DoS/DDoS. These vulnerabilities root from the variety of technologies and protocols that shape IoT plus scalability since different devices and nodes join and leave the network, it increases the issues relevant to authentication. Besides, some of the security challenges as mentioned by [34] is enabling IPsec with IPv6 nodes. So, the common security attacks can be shown in the figure 80:

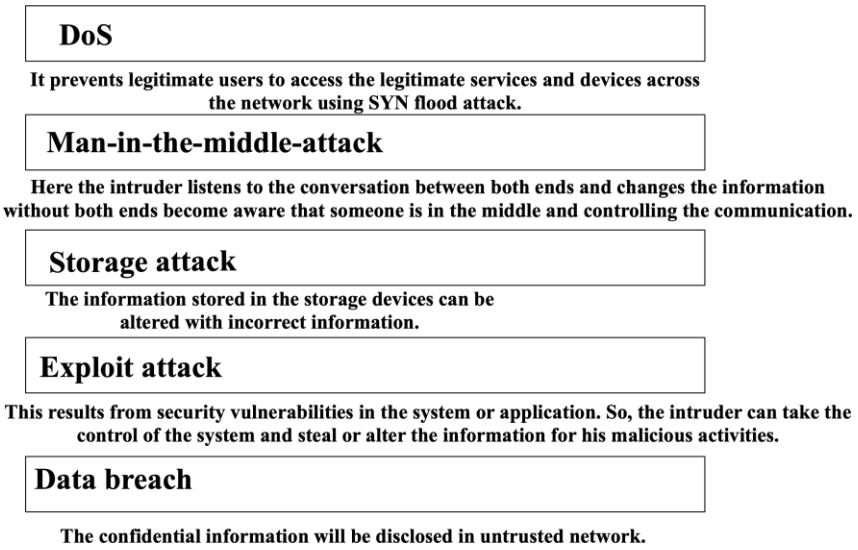

Figure 80 Security threats in Network layer [34]

However, the security consideration that can consider for this layer can be securing the traffic as the traffic flows between public and private networks. So, utilizing secure protocols like TLS/SSL encryption. But using heavy encryption technique is not appropriate for such devices with memory and processing constraints. For instance, Arduino takes around 3 minutes to encrypt a payload using **RSA** 1024- bit key length but encryption approaches that use elliptical curve digital signature is a good choice. So, one of the most popular method to provide security in the network layer is authentication. And some researchers encourage to use IPsec for Network Layer by introducing another Adaption Layer. Other approaches emphasized on utilization of lightweight authentication protocol based on the public key [34]. Also, security considerations should be considered for other standards rather than Wi-Fi like 6LoWPAN, ZigBee which will be discussed in the next sections.

3-4-3 THREATS AT SERVICE LAYER:

Service Layer is where the IoT management system will be represented. So, any services required by users and applications will happen at this layer. This layer utilizes middleware technology that allows reusability of software and hardware. Middleware is designed based on the requirements that are common among different applications plus APIs and service protocols.

In the service layer information exchange, data processing, service integrations, analytical services, User interface services, databases exist. So, the security requirements for this layer

include but not limited to service/group authentications, privacy, integrity, security of keys, non- repudiation, anti-replay, availability and authorization. So, the possible threats that can happen at this layer are as shown in figure 81. So, security consideration that can be done at this layer are providing secure transmission platform between service layer and other layers. Also, using service authentication, identification and access controls are highly recommended.

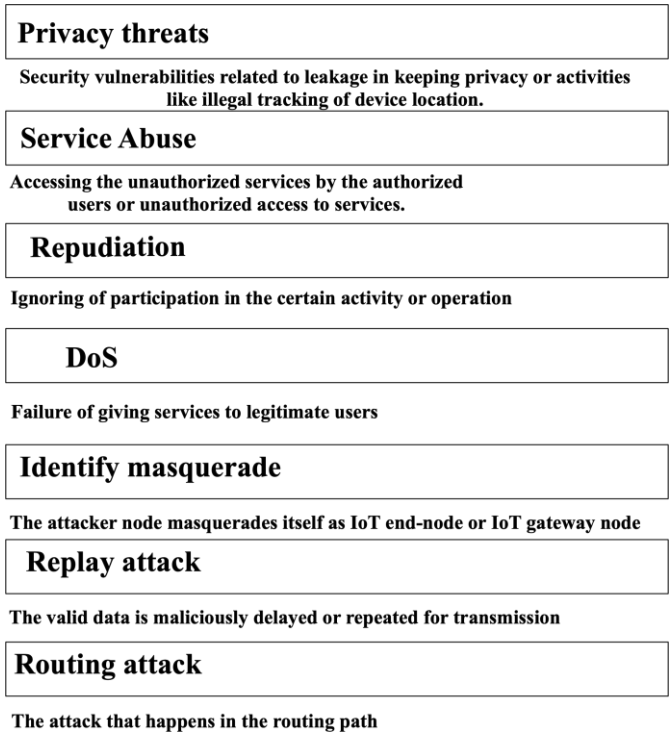

Figure 81 Possible threats at services layer [34]

3-4-4 THREATS AT APPLICATION LAYER:

The security consideration that can be considered at this layer are applying latest software updates, security patches, authentication, safe remote connection. So, the common attacks at this layer can be shown in figure 82:

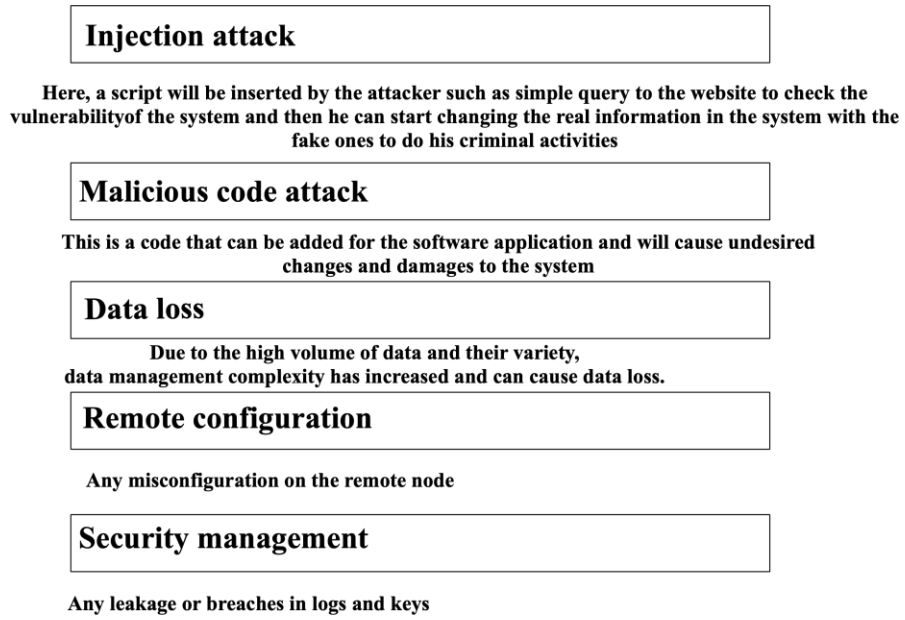

Figure 82 Security threats in Application layer [34]

3-4-5 THREATS IN COMMUNICATION TECHNOLOGIES OF IOT:

Encryption and cryptographic algorithms are divided in two types: symmetric encryption algorithm and public-key encryption algorithms. Using symmetric encryption algorithms makes several challenges [38]:

- Firstly, the key exchange protocol used in symmetric encryption is complex and requires more resources which is not suitable for WSN nature.
- Secondly, key should be kept confidential and secure while the WSN network is not secure and if one node is compromised, there is a high potential of affecting the entire network.
- Thirdly, using message authentication code which is used for authentication increases the communication overhead and needs more resources.

Therefore, using public key encryption introduced. In the public key encryption algorithm, each node holds its own private key and base stations hold the public keys for all nodes. This approach provides better scalability specially for unidirectional data transmission, but it adds more complexity and computational power [35]. So, symmetric approaches are appropriate for WSN environment but does not guarantee high security. On the other hand, asymmetric consumes more power and energy due to more computation and processing.

Key management: The aim of key management in WSN is to generate, distribute, store, update,

destruct and support code keys. Key management comes into two forms: static key management and dynamic key management.

In the static mode, the key's principals will be updated and adjusted before distributing keys and keys remain constant unless a network change happens. While in dynamic mode, key can be updating during the network's lifetime [35]. Some studies divided key management in to four groups [34]:

- Broadcast distribution of key where the key is responsible for protecting the station broadcasting the information to all nodes.
- Group key distribution: Here, a key will be used for protecting the nodes in a specific group.
- Node master key distribution: Here, the key is distributed between node and base station.
- Distribution of key shared between each pair of nodes.

Among these four classes, the distribution of key shared between each pair of nodes is common in WSN. Then, key management can be designed to use symmetric or asymmetric algorithms [35].

Secure routing: Unfortunately, most of the routing protocols in WSN designs concern about data transfer rather than secure data transfer.

Trust management: In addition to password and authentication-based approaches and using cryptographic algorithms, trust mechanisms should also exist to guarantee the security of WSN network along with energy consumption of nodes. Since WSN works based on the collaboration of all nodes for gathering data.

After talking about the security concerns in WSN, it is good idea to discuss and explain some of the most common security threats in WSN networks as shown in figure 82 [33,34,35]. In order to prevent these types of forwarding attacks, malicious nodes should be identified and removed from routing tables by using acknowledgment, neighbor node information or monitor nodes, multi- data flow, creating secure communication.

SOLUTION:

- Avoid using weak, hardcoded passwords which can be easily broken by brute force. So, relying on unchanged credentials is not recommended.
- Unnecessary services running on the device specially those are accessible through Internet should be controlled and monitored for any unauthorized access. Also, these services should be tracked if they compromise confidentiality, integrity, availability and privacyof data.
- Insecure web interfaces, cloud interfaces should be eliminated. These interfaces are open doors toward any malicious activities against the device, information and the related components. So, proper authentication, authorization, encryption and filtering should be considered and configured.
- Every device, hardware, software should be updated in a timely manner. Also, there should be a proper mechanism for secure device update, firmware validation, anti-rollback mechanism and a notification alarm if any sudden changes happen to the system.
- The existence any outdated, insecure software/libraries components which can make the device be vulnerable, should be prevented. Also, inclusion of third-party applications or software from insecure, compromised sources should be avoided.
- Confidential and sensitive information should not be stored in devices or systems with improper insecure privacy mechanisms.
- Data should be transmitted, processed and stored in a secure manner with proper encryption, access control mechanisms.
- There should be appropriate asset management, update management, secure system monitoring mechanisms.
- Systems and devices should not be left with their default setting configurations.
- Lack of proper physical hardening can cause security threats for devices being access and controlled either remotely or locally [35].

3-4-6 ARCHITECTURAL SECURITY DESIGN:

In order to provide high level of security in IoT, new lightweight security protocols and algorithms, efficient privacy mechanisms and physical safety mechanisms for keeping physical devices safe and secure, should be proposed and designed. But before taking further steps and actions, architectural security designs should be considered [35] since new security implementations should be mapped to IoT architecture and not to existing IT networks or WSN networks. Figure 4-1 shows the common protocols and standards used in IoT.

In this section, different architectural designs will be discussed, and their positive and negative aspects will be reviewed. In [35], three different security designs were suggested:

- End-to-End security in Things:

Having secure end-to-end communication is important for both the conventional IT network and IoT. Protocols like 6LoWPAN and IPv6 provides end-to-end communication in IoT. In order to provide security, it was recommended that each device be responsible and controls its security.

- For preparing end-to-end security one solution is increasing and facilitating the end devices or node with more memory and processing power and therefore applying the traditional security approaches like asymmetric cryptography.
- Another option is using hardware-based solution like PUF[1] to secure IoT devices and is used for authentication and secure communication. PUF only permits one-way functions

  [41] and has compatibility with IoT devices that have limited resources. PUF provides secure authentication with no means of cryptography on the device [35]. Well, each device chipset is unique and has its own fingerprint like humans. Then, PUF circuits can be added to chips. So, when the PUF circuits receives an input called challenges, the output is a set of bits called response. The interesting point is that chips will not create the same output or responses [35]. Authentication in PUF consists of two steps enrollment and authentication phases. In enrollment, the chip will connect to the server and chips that are PUF enabled are connected to the server directly. Then, they go through the process of challenge and response. Afterwards, the server will store the pair of challenge/response. So, next time the device wanted

to authenticate, the server will compare the response it with its table entries [35]. However, enabling chipset with PUF requires more memory capacity to save challenge/response pair and PUF circuit is added to memory chip. Besides, increasing the memory capacity in device will cause the increase in the cost of device and some devices may not support PUF.

- The next highly consistent approach for secure End-to-End connection between IoT objects is IP-based security solutions. If devices can support IP, they can support IP-enabled security solution as shown in figure 4-1. So, solutions like DTLS, IPsec, minimal IKEv2 and HIP DEX are highly desirable.

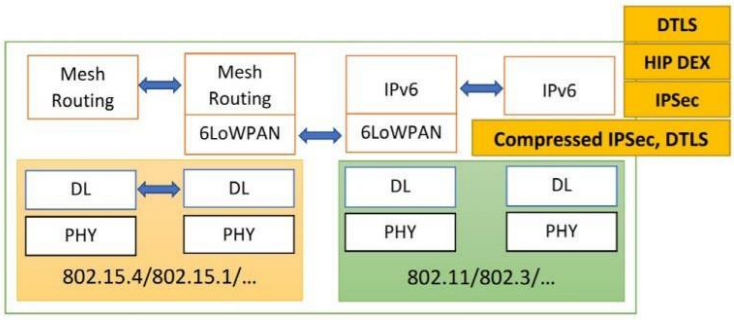

Figure 83 Security solutions for end-to-end [35]

As mentioned in [45], using public-key encryption is mandatory in HIP DEX and minimal IKEv2 while using public-key or pre-shared cryptography is optional in DTLS. Also, all the protocols use a variant of Diffie-Helman protocol. However, DTLS and minimal IKEv2 do not support multihoming and mobility.

- HIP DEX is an IETF protocol and belongs to HIP protocols family. HIP[1] is designed for establishing a secure channel between hosts but supports mobility and multihoming. It works based on the cryptographic namespace of stable host entities between network and transport layer [35].

- HIPv2 was designed to decrease the overhead caused by cryptography functions. This overhead will be reduced by removing hash functions and public key signatures. HIP DEX protocol is mainly designed as a key establishment protocol for devices with constrained resources and can be used as a key mechanism in IEEE 802.15.4 [44]. This protocol creates a secure communication channel between hosts by exchanging four packets between the session initiator and responder and their key agreement is based on Diffie-Hellman approach. This four-packet approach make

hosts secure toward DoS attack.

- DTLS is a secure point-to-point protocol that creates a secure communication channel between CoAP nodes by handshake as mentioned in previous chapters. CoAP does not provide the internal security by itself. Without DTLS, CoAP protocol is prone to Man-in-The-Middle and DDoS attacks. Although DTLS works on top of the UDP protocol, it provides data integrity, encryption and authentication. It protects devices against DDoS and Anti-Replay attacks.
- In other words, this protocol is designed for creating secure connection between constrained nodes. It works in four security modes: 1. NoSec when DTLS is not enabled in CoAP and communication is based on the UDP protocol. 2. Pre-Shared key in which the device is pre-installed with a list of
- symmetric keys. So, for communicating with a node, the device needs to use one of the keys in the list. 3. Raw Public key: a list of asymmetric key pairs is pre-installed. 4. Certificates: The device uses a pair of public keys and X.509 certificate for communication [35].
- However, this protocol has several challenges like lack of scalability and that's due to the resource limitations in end devices which restricts the number of DTLS sessions. Besides, it is not compatible with multicast traffic and caching [35].
- IPsec is also another protocol that provides end-to-end security for both IPv4 and IPv6 networks. It provides authentication and encryption for each packet in the communication. It is mainly targeted for powerful robust devices with no limitations in resources.
- IPsec composed of two primary parts: AH[1] and ESP [2]. AH authenticates the IP header and ESP performs the authentication and encryption of payload [59]. IPsec works in two modes: 1. Transport mode: the payload will be authenticated and encrypted. 2. Tunnel mode: the entire packet is encrypted and authenticated.
- However, it is not an appropriate choice for IoT networks that run over 6LowPAN since AH and ESP adds more bits and extra overhead which leads to more energy consumption. On the other hand, the key exchange process in IPsec is based on the

IKEv2 which is a heavy protocol. Hence, researchers tried to introduce another lightweight version of IPsec. One approach tried to compress the IPsec header using the same compression mechanisms utilized in compressing IPv6 like HC13 [35]. However, it worked based on Pre-shared key and needed to be further enhanced for IKEv2.

- From the above-mentioned solutions, devices should be able to support IPv6 and 6LowPAN protocols. The mentioned approaches add extra overhead to the network and packets transmissions. So, providing lightweight version of secure protocols are welcoming but needs more work and research to comply with IoT environment and constrained devices. Also, Pre- shared based or symmetric based cryptography is desirable for devices while it is not as much powerful as public key cryptography.

    - Deploying security service at edge:

Some objects are not enriched with enough resources to support end-to-end security like RFID tags, the security protection and responsibility will be moved to devices with richer resources like edge devices instead of end devices [35]. So, the end device should make a trust relationship with the edge device to oversee its security needs as shown in figure 84.

There are certain reasons that why security can be done at the edge devices or edge layer. Firstly, Edge layer devices have more resource capabilities. Secondly, it is close to IoT end devices and objects and mostly directly or within a few hops, they are away from each other. Besides, Edge devices have more information about the entire network. Also, Edge devices have high speed connection to cloud services and can ask them for further support.

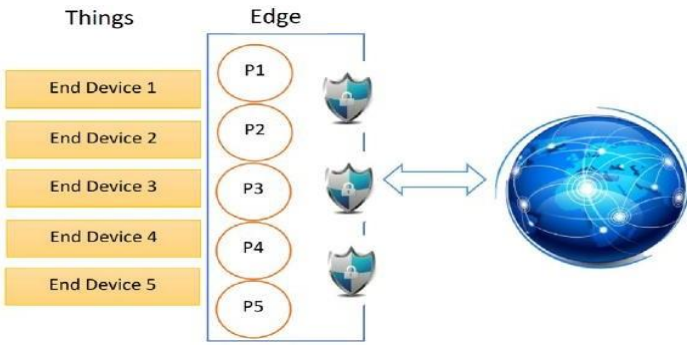

Figure 84 Providing security at edge device [35]

So, as shown in figure 4-2, the edge device will create a separate profile for each end device. So, any access or communication to end device, should pass through the edge device and authenticated mutually. So, through the process of authentication and authorization, the edge device can control and guide the end device to with which devices they can communicate and can access to their resources.

It is good to note that edge device can make use of Intrusion Detection algorithms for detecting attacks [34]. One of the approaches that applied in Edge device is called EdgeSec which was introduced in [35]. This approach was implemented at the Edge layer of IoT architecture. It is composed of seven modules including Security Profile Manager, Security analysis module, Protocol Mapping module, Interface manager, Security Simulation, Request Handler and User Interface. End devices will subscribe to edge device by Security Profile Manager Module. And by Security Analysis Manager, it checks what the security requirements are and how these security considerations should be implemented at the edge, devices or cloud layer. Then, based on the security requirements, it invokes Protocol mapping to activate and use relevant security protocols. The architecture of EdgeSec model is shown in figure 85 [35].

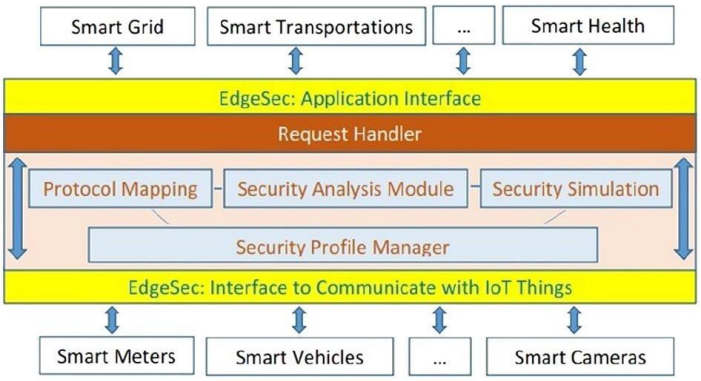

Figure 85 EdgeSec Security Architecture [35]

- Distributed security model:

Using edge device for security mechanisms is a good idea but there are several challenges toward it. Firstly, establishing trust between edge device and IoT end devices requires scalable authentication and authorization mechanisms using public key or private key

cryptography schemes which may add heavy load to end devices again.

So, the idea is before the end device and edge device start communicating with each other, the edge device sends the access request to cloud service and then in the cloud the authentication, authorization and trust process will take place. Finally, if the process passed successfully in cloud service, the edge device can trust the end device and starts communicating.

One example of this approach is Smart meter that was discussed in [35]. Here, the device will be authenticated by cloud-reader authentication protocol. Then, the reader will generate a one-time symmetric key for the device. This type of architecture will protect the device from eavesdropping, brute force, replay attack and device attack.

Authentication and Encryption in IoT: Authentication is the process that prevents any unauthorized access to the network through fake credentials. So, both the peer node and data extraction should be legitimate. Authentication process provides different IoT devices to integrate in different smart environments [35].

Authentication can be classified to integrity, message authentication and entity authentication, Key authentication, nonrepudiation and access control [35]. Besides, authentication often comes along with encryption and it requires key management techniques.

Key management mechanisms can be based on Symmetric-key cryptography and Asymmetric- key cryptography. In the first mode, the first option, a shared secret key will be distributed between two ends. While the Asymmetric-key cryptography will include the process of public and private keys.

Message's authentication provides integrity and data origin or in other words it provides data integrity.

Entity authentication also referred to as endpoint authentication which assures the identity and the presence of the entity claiming who it is.

Key authentication also takes care of both entity and the linked keys to that entity.

Authentication procedure can be one-way where only one party will authenticate himself

toward the other. Two-way authentication where both communication parties will authenticate each other. Finally, three-way authentication that a CA will authenticate both ends. In terms of IoT, since both ends will have a mutual communication, there should be mutual authentication in IoT. And authentication is the fundamental security mechanism that should be applied at different layers. Since, the end device needs to authenticate itself with edge device or gateway. Then, the gateways need to authenticate themselves toward the cloud service for sending the data and the application or that web interface needs to authenticate itself toward cloud to use the data for further processing and analysis [35].

However, authentication is one way but not the only way to provide security. Authentication prevents certain attacks like Man-in-the-Middle, Sybil, reply, impersonation attacks. In IoT, authentication can be provided using lightweight authentication and encryption protocols, multi- factor authentication using bio-hashing.

The protocols that can provide hardware-based authentication are PUF, TRNG and TPM which uses the physical characteristics of the hardware. One of the protocols that is widely used in authentication and encryption is TLS. The TLS version that can be applicable for IoT constrained devices are TLS-PSK, TLS-DE-RSA. TLS-PSK uses pre-shared keys while TLS-DE-SA uses RSA and Diffie-Hellman for key exchange and management.

There are three classes of authentication schemes in IoT including lightweight asymmetric-based, symmetric-based and hybrid protocols [35]. It is good t note that lightweight cryptographic algorithms can be achieved by lightweight block ciphers like smaller key sizes, smaller block sizes, simpler rounds, lightweight hash functions, smaller message sizes and so on. Some of the lightweight symmetric algorithms offered for IoT include AES-128, HIGHT[1], TEA, PRESENT, and RC5. AES-128 is utilized by CoAP protocol. HIGHT can be used in RFID systems. And lightweight asymmetric cryptography algorithms are ECC[2] and RSA, but RSA is so resource- intensive. ECC has smaller key size compared to RSA and needs less memory and processing power which can be applied for IoT nodes.

However, the mentioned authentication schemes are not strong enough toward attacks like Denial of Service, node capture, impersonation, replay attack to name but a few. Encryption is the way that provides end-to-end security. But the conventional cryptographic approaches are not applicable for heterogenous environment with constrained resources.

Glossary:

FDMA - Frequency Division Multiple Access
AMPS - Advanced Mobile Phone Service
MTSO - Mobile telecommunications switching office
GSM - Global Systems for Mobile Communications
SMS – Short message service
MMS – Multimedia Messaging Service
GPRS - General Packet Radio Service
EDGE - Enhanced Data rates for GSM Evolution
BSS - Base Station Subsystem
NSS - Network Switching Subsystem
OMSS - Operation and Maintenance Subsystem
MS - Mobile Station
BTS - Base Transceiver Station
BSC - Base Station Controller
MSC - Mobile Switching Center
ISDN - Integrated Services Digital Network
GMSC – Gateway Mobile Switching Center
ISC - International Switching Center
HLR - Home Location Register
VLR - Visited Location Register
AuC - Authentication Center
EIR - Equipment Identity Register
OMC - Operation and Maintenance Center
UMTS - Universal Mobile Telecommunications System
UE - User Equipment
AN - Access Network
CN - Core Network
WCDMA – Wideband Code Division Multiple Access
FDD - Frequency Division Duplex
TDD - Time Division Duplex
ME - Mobile Equipment
USIM - Universal Subscriber Identity Module
MT - Mobile Termination
TE - Terminal Equipment

IMSI – International Mobile Subscriber Identity
AN - Access Network
RNS - Radio Network Sub-system
RNC - Radio Network Controller
CN - Core Network
CS - Circuit Switched
PS - Packet Switched
PSTN - Public switched telephone network
PDN - Packet data network
SGSN - Serving GPRS Support Node
GGSN - Gateway GPRS Support Node
HSS - Home Subscriber Server
AS - Access Stratum
NAS - Non-Access Stratum
RRC - Radio Resource Control
LTE - Long-Term Evolution
E-UTRAN - Evolved UMTS terrestrial radio access network
EPC - Evolved packet core
APN - Access point name
S-GW - Serving gateway
MME - Mobility management entity
OFDM - Orthogonal frequency-division multiplexing
NR – New radio
mmWave - millimeter Wave
FR1 - Frequency Range 1
FR2 - Frequency Range 2
NSA - Non-Stand Alone
SA - Stand-Alone
EN-DC - E-UTRAN and NR Dual Connectivity
MN – Master node
SN - Secondary Node
NF - Network Function
NSSF - Network Slice Selection Function
AUSF - Authentication Server Function
PCF - Policy Control Function
AMF - Access and Mobility management Function

SMF - Session Management Function
UPF - User Plane Function
NRF - Network Repository Function
NEF - Network Exposure Function
UDM - Unified Data Management
UDR - Unified Data Repository
UDSF - Unstructured Data Storage Function
MVNO – Mobile virtual network operator
QoS – Quality of Service
UAV – Unmanned aerial vehicle
GPS – Global positioning system
IoT – Internet of things
PLMN – Public land mobile network
RSU - Road Site Unit
SST – Slice/Service Type
SDU - Service Data Unit
TB - Transport Block
HARQ - Hybrid Automatic Repeat request
RLC – Radio link control
PDCP - Packet Data Convergence Protocol
SRB - Signaling Radio Bearer
DRB - Data Radio Bearer
QFI - QoS flow ID
SIB2 - System Information Block Type 2
PDCCH DCI - Physical Downlink Control Channel Downlink Control Information
DL-SCH - Downlink Shared Channel
UL-SCH - Uplink shared channel
GTP - GPRS tunneling protocol
TEID - Tunnel Endpoint Identifier
PSS - Primary synchronization signal
SSS - Secondary synchronization signal
PCI - Physical Cell Identifier
Wi-Fi - Wireless Fidelity

Conclusion:

In order to secure the communication between the two devices, we must implement security mechanism at each layer of IoT architecture. So, considering the massive IoT architecture layers (application, cloud, network, gateway, communication and physical layers), we have to implement certain measures to eradicate the attacks. In application layer security, users gain access to some form of information via a user interface, GUI or apps. Authentication is the most necessary step in protecting IoT devices and systems which is usually missing or overlooked by developers. The cloud security layer must ensure proper protection of data and information, enforce privacy policies, and secure connections and the cloud network. The information transmission security layer is responsible for providing reliable secure data transfer within the entire system (i.e. computer network, wireless network and mobile network). The gateway information security layer deals with securing heterogeneous technologies at the edge. These technologies require control and protocol security. Internal communications security provides security to the system under the perimeter. Finally, the end device security layer secures the IoT devices.

Further, in order to isolate this entire IOT architecture or isolate the massive M2M communication network function, we implement slices. During preparation phase of a slice poorly designed, tampered with or improperly implemented network slice template (e.g., with design flaws, without up-to-date security patches, or injected malware) affects all the slices built from it. In addition to powerful active attacks that might damage the integrity of the template, content exposure might also disclose sensitive information. So cryptological protocols are used to provide confidentiality, integrity, and authenticity of network slice templates. The correctness of the network slice template must also be verified.

During the installation, configuration and activation phase of a slice, the point of attack is over the API, whose compromise would permit an adversary to interfere in the installation, configuration, or activation of a slice. So, we must implement mechanisms to secure APIs, such as access and operational rights. Good practices include the usage of TLS (for mutual authentication) or O-Auth (for authorization of service requests). Moreover, the API should permit auditing, monitoring, and reporting securely (e.g., traffic logs, APIs invocations). The general crypto-graphical techniques and real-time security analysis mentioned in the first phase remain useful in the second phase too.

During the runtime-phase of the slice, we have Distributed Denial-of-service attacks that are difficult to mitigate. During a flooding DDoS attack, the attacker will try to overwhelm the target network or service by sending a large amount of traffic. The challenge is to keep the target network or service available to the end users. Specific mitigation technique against DDOS is the implementation of Network slice. We need to create authenticity between every inter and intra slice. We need to make sure after when the data is used, the data has to be erased and network functions are to be de-allocated. Moreover, logging and auditing are of extreme importance. Different levels of logging must be implemented in distinct slices. Protecting the results of the logs and reports is of extreme importance, as their exposure would leak sensitive information. Usage of dedicated and isolated security zones during the whole life cycle is a good practice to mitigate security risks. In order to secure the cloud in edge service we must have compression and key management-based data security framework.

References:

[1] Liyanage, Madhusanka. *Comprehensive Guide to 5g Security.*.

[2] Smith, Clint, and Daniel Collins. *3g Wireless Networks.* New York: McGraw-Hill, 2002.

[3] Eberspächer, J. *Gsm : Architecture, Protocols and Services.* 3rd ed., English lang. ed. Chichester, U.K.: Wiley, 2009.

[4] Frenzel, Louis E. *Handbook of Serial Communications Interfaces : a Comprehensive Compendium of Serial Digital Input/output (i/o) Standards.* .

[5] Sanders, Geoffrey. *Gprs Networks.*

[6] Seidenberg, P., M. P Althoff, and Bernhard H Walke. *Umts : the Fundamentals.* Hoboken: Wiley [Imprint], n.20.

[7] Khan, Farooq. *Lte for 4g Mobile Broadband : Air Interface Technologies and Performance.* .

[8] X. Lin, J. G. Andrews, A. Ghosh and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," in *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40-48, April 2014, doi: 10.1109/MCOM.2014.6807945

[9] Mumtaz, Shahid, Jonathan Rodriguez, and Linglong Dai. *Mmwave Massive Mimo : a Paradigm for 5g.* First edition. .

<massive mimo>

[10] Ruparelia, Nayan. *Cloud Computing.* .

[11] NFV, Network Functions Virtualisation. "ETSI GS NFV 001 V1. 1.1 (2013-10)." (2013).

[12] Zhang, Ying. *Network Function Virtualization : Concepts and Applicability In 5g Networks.* .

[13] Kazmi, S. M. Ahsan, et al. *Network Slicing for 5g and Beyond Networks.* .

[14] Li, Qian, et al. "An end-to-end network slicing framework for 5G wireless communication systems." *arXiv preprint arXiv:1608.00572* (2016).

[15] . D. Hanes, G. Salgueiro , P. Grossetete, R. Barton and J. Henry , IoT Fundamentals: Networking Technologies, Protocols, and Use Cases for the Internet of Things, Cisco Press, 2017.

[16] R. Buyya and A. V. Dastjerdi, Internet of Things: Principles and Paradigms, San Francisco, CA: Morgan Kaufmann Publishers, 2016.

[17] S. Tennina, A. Koubâa, R. Daidone, M. Alves, P. Jurčík, R. Severino, M. Tiloca, J.-H. Hauer, N. Pereira, G. Dini, M. Bouroche and E. Tovar, IEEE 802.15.4 and ZigBee as Enabling Technologies for Low-Power Wireless Systems with Quality-of-Service Constraints, Berlin, Heidelberg: Springer, 2013.

[18] Z. Shelby and C. Bormann, 6LoWPAN: The Wireless Embedded Internet, Wiley, 2009.

[19] A. Minteer , Analytics for the Internet of Things (IoT): Intelligent analytics for your intelligent devices, 2017: Packt.

[20] S. Farahani, ZigBee Wireless Networks and Transceivers, Newnes, 2008.

[21] A. Dunkels and J.-P. Vasseur, Interconnecting Smart Objects with IP, Morgan Kaufmann, 2010.

[22] 5G Americas, "The Evolution of Security in 5G," October 2018. [Online]. Available: https:// www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_5G_Security_White_Paper_Final.pdf.

[23] Alrashede, Hamad, and Riaz Ahmed Shaikh. "IMSI Catcher Detection Method for Cellular Networks." *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2019.

[24] Pandey, Alok, and Jatinderkumar R. Saini. "A Simplified Defense Mechanism Against Man in the Middle Attack." *IJEIR*1.5 (2014): 2277-5668.

[25] Singh, Isha. "Signaling Security in LTE Roaming." (2019).

[26] Ghorbani, Hamidreza, M. Saeed Mohammadzadeh, and M. Hossein Ahmadzadegan. "DDoS Attacks on the IoT Network with the Emergence of 5G." *2020 International Conference on Technology and Entrepreneurship-Virtual (ICTE-V)*. IEEE, 2020.

[27] Habib, Ahsan, Mohamed Hefeeda, and Bharat K. Bhargava. "Detecting Service Violations and DoS Attacks." *NDSS*. 2003.

[28] Shu, Zhaogang, et al. "Security in software-defined networking: Threats and countermeasures." *Mobile Networks and Applications* 21.5 (2016): 764-776.

[29] Monshizadeh, Mehrnoosh, Vikramajeet Khatri, and Andrei Gurtov. "NFV security considerations for cloud-based mobile virtual network operators." *2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 2016.

[30] Lal, Shankar, Tarik Taleb, and Ashutosh Dutta. "NFV: Security threats and best practices." *IEEE Communications Magazine*55.8 (2017): 211-217.

[31] M. Burhan, . R. A. Rehman, B. Khan and . B.-S. Kim, "IoT Elements, Layered Architectures and Security Issues: A Comprehensive Survey," *Sensors (Basel)*, vol. 18, no. 9, p. Sensors (Basel), 2018.

[32] S. Li and L. D. Xu, Securing the Internet of Things, Elsevier Inc., 2017.

[33] A. Gerber, "Top 10 IoT security challenges," IBM, 2017.

[34] M. b. M. Noor and W. H. Hassan, "Current research on Internet of Things (IoT) security: A survey, "*Computer Networks,* vol. 148, pp. 283-294, 2019,.

[35] A. Čolaković and M. Hadžialić, "Internet of Things (IoT): A Review of Enabling Technologies, Challenges, and Open Research Issues," *Computer Networks,* vol. 144, pp. 17-39, 2018.