5G for Telehealth: Utility of Local Mobile Network Operator Architecture and Network Slicing

Rakshitha De Silva

(169221F)

Degree of Master of Science

Department of Electronic and Telecommunication

University of Moratuwa
Moratuwa, Sri Lanka

Abstract

This thesis studies two novel technologies enabled by network softwarization, namely Local 5G Operator (L5GO) networks and network slicing for 5-th Generation (5G) and beyond 5G networks. This study proposes a L5GO network architecture for delay critical future telehealth services, considering two use cases on Augmented Reality Assisted Surgery (ARAS) and Robotic Aided Surgery (RAS). Study compares the latency performance of the proposed Local 5G Operator (L5GO) architecture with a traditional legacy network and a network equipped with Mobile Edge Computing (MEC). The results highlight the unique advantages of utilizing an L5GO to cater the communication needs of delay critical telehealth, compared to a traditional network. To the end of network slicing, which enables the creation of multiple logical independent networks on physical networking infrastructure can be classified as vertical and horizontal slicing. It is a key methodology to deliver guaranteed Quality of Service (QoS) to a multitude of use cases with varying resource requirements, and hence considered vital for 5G and beyond 5G mobile networks. This thesis focuses on the optimal resource allocation in a 5G network that utilizes network slicing, and caters to the communication requirements of a smart hospital. The study compares the performance of the slicing methods through solving two convex optimization problems, considering several smart hospital scenarios that differ from each other based on their medical speciality. The results are used to draw insights on the most appropriate slicing approach for each setup.

Keywords:
3GPP, 5G, local 5G operator, network slicing, smart hospital, telehealth, URLLC.

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature | Date

The above candidate has carried out research for the Masters thesis under my supervision.

Signature of the supervisor | Date

Acknowledgement

This work was carried out during the years 2017-2021 at the Department of Electronic and Telecommunication Engineering, University of Moratuwa.

I owe my deepest gratitude to my supervisors Dr. Madhushanka Liyanage and Dr. Tharaka Samarasinghe for their incredible guidance to successfully complete the research. A conference paper has been published based on the research outcomes of Local 5G Operator (L5GO) related work and another conference paper has been submitted presenting results of 5G mobile network slicing options. I would like to thank Mr. Yushan Siriwardhana for his valuable contribution as a co-author in these publications.

I would like to thank all the authors and owners of the materials that have been referred in this thesis and finally, I would like to thank all the people who supported to successfully complete the research.

Contents

List of Figures

List of Tables

Abbreviations

μO Micro Operator.

3D Three Dimension.

3GPP 3-rd Generation Partnership Project.

4G 4-th Generation.

5G 5-th Generation.

5GPP 5G Infrastructure Public Private Partnership.

AMF Access and Mobility Management Function.

API Application Programable Interface.

AR Augmented Reality.

ARAS Augmented Reality Assisted Surgery.

AUSF Authentication Server Function.

BSS Business Support Systems.

CNF Cloud-native Network Function.

CoMP Coordinated Multi-Point Transmission and Reception.

CP Control Plane.

CT Computer Tomography.

DS-TT Device-side TSN translator.

E2E End-to-End.

EIR Equipment Identity Register.

eMBB enhanced Mobile Broadband.

ETSI European Telecommunications Standards Institute.

FDD Frequency Division Duplexing.

GMC Grand Master Clock.

gNB next generation NodeB.

HDR High Dynamic Range.

IoMT Internet of Medical Things.

IoT Internet of Things.

L5GO Local 5G Operator.

M2M Machine to Machine.

MANO Management and Network Orchestration.

MEC Mobile Edge Computing.

MIMO Multiple input multiple output.

mMTC massive Machine Type Communication.

MNO Mobile Network Operator.

MR Magnetic Resonance.

NF Network Function.

NFV Network Function Virtualization.

NFVO Network Function Virtualization Orchestrator.

NG-RAN Next Generation-Radio Access Network.

NGMN Next Generation Mobile Network.

NSaaS Network Slice as a Service.

NW-TT Network-side TSN translator.

OSS Operating Support Systems.

PCF Policy Control Function.

PDU Protocol Data Unit.

PNF Physical Network Function.

PPMU Portable Patient Monitoring Unit.

QoS Quality of Service.

RAKT Robotic Aided Kidney Transplant.

RAN Radio Access Network.

RAS Robotic Aided Surgery.

RAT Radio Access Technology.

SDN Software Defined Networks.

SLA Service Level Agreement.

SMF Session Management Function.

TSN Time Sensitive Network.

TSN AF Time Sensitive Network Application Function.

UDM Unified Data Management.

UDN Ultra-Dense Network.

UP User Plane.

URLLC Ultra-reliable Low Latency Communication.

VAS Value Added Services.

VIM Virtualized Infrastructure Manager.

VNF Virtual Network Function.

VNFM Virtual Network Function Manager.

VR Virtual reality.

Chapter 1

Introduction

Telehealth, which allows access to health care services remotely, is considered to be a key application of next generation wireless networks [7]. The Coronavirus Disease 2019 (COVID-19) pandemic, has made telehealth a focal point, due to the critical requirement of minimizing personal interactions for decelerating the spread of the virus [8]. Furthermore, ensuring the health of the healthcare personnel has become paramount with healthcare facilities operating at their capacity. This thesis focuses on delay critical telehealth applications that may have inherent constraints in terms of direct personal interactions, and studies how the fifth generation wireless technology for digital communications, also known as 5G, can be used effectively in such a setting.

1.1 Background and Selected Scope

5-th Generation (5G) mobile networks are envisioned to bring about a paradigm shift in communications, providing ultra-low latency, higher bandwidth and massive connectivity. Innovative use cases that stem on the capabilities of Mobile Network Operator (MNO) are expected to emerge, with more emphasis on case and location specific communications. Facilities such as hospitals, shopping malls, sports arenas, universities, and factories, already demand for customized wireless connectivity solutions, with varying levels of reliability, data rates, latency and security [9]. Most of these differential Quality of Service (QoS) demands are beyond the capabilities of the conventional legacy MNO based networks. The focus of traditional MNOs is providing generic mobile communication services to mass geographical areas such as countries or provinces. The investment cycles of MNOs are generally long with a high infrastructure cost. Therefore, additional capital and operational expenditure, and increased effort, prevent the deployment of case and location specific network services by MNOs.

Mobile Edge Computing (MEC) brings the computing capabilities to the edge of the network, enabling latency aspects in MNO, while supporting high bandwidth. As MEC operates closer to the location of the applications, real-time altering of application aware performance optimization is possible. This is useful in medical applications that need low latency. However, most MEC configurations are likely to fail on providing low latency requirements that are mandatory for many telehealth use cases. These stringent QoS demands have made way to new players such as Local 5G Operator (L5GO), to enter the market with an aim of providing case and location specific communication services [10, 11].

The intention of L5GOs is to deploy localized communication services to satisfy the context-aware extreme service requirements for a selected group of users. Even though the L5GOs serve confined locations, they need to collaborate with the 5G MNOs, thus the L5GO deployments must follow the standards defined by standardization organizations such as the 3-rd Generation Partnership Project (3GPP) [9] [12]. In addition, the L5GO should lease spectrum from a regulator or an MNO, in the form of local spectrum licenses. The L5GO may maintain its own infrastructure or lease the infrastructure owned by an infrastructure vendor. The L5GO can provide roaming services to subscribers of other MNOs, upon their arrival to the L5GO coverage area, given that an agreement is in place between the two parties. Due to the communication demands of a L5GO network being highly variable, a tailored custom architecture that is ready to serve user needs is preferred over a universal network architecture.

Network slicing, empowered by Software Defined Networks (SDN) and Network Function Virtualization (NFV), allows the creation of multiple logical networks on top of shared physical infrastructure. It enables differentiated service delivery to the customers with guaranteed QoS in terms of functionality, performance, and isolation [6, 13]. Due to its flexibility and programmability, network slicing facilitates extreme resource requirements, creating and expanding numerous business opportunities for the future mobile network operators [6].

Network slicing methodologies in a 5G MNO environment have been discussed and classified considering different contexts [13–16]. One such classification is vertical and horizontal network slicing, which is the focus of this thesis. Vertical network slicing is the most widely discussed slicing approach, and the concept has been around since the 4-th Generation (4G) mobile networks. In a vertical slice, the End-to-End (E2E) traffic flow is usually between the MNO core network and the end user device, through the access network and the backhaul network. Hence, an E2E network slice created to fulfill the requirements of a given use case falls under vertical network slicing. Horizontal slicing decomposes 5G network into matching technical systems. These systems do not empower end user necessity individually, but when put together, they can deliver a user requirement E2E. 5G horizontal slicing focuses on enhancing the user experience through enabling resource sharing among network nodes and user devices. In horizontal network slicing, the 5G MNO resources including radio interfaces and end user devices resources are horizontally sliced, allowing 5G network nodes to access and process more end user traffic facilitating low latency communications [14, 16].

Smart hospital environments demand network services with Ultra-reliable Low Latency Communication (URLLC) for telehealth applications such as remote surgeries and patient health monitoring based on massive Machine Type Communication (mMTC) [17]. The resource consumption of these applications is high, which leads to frequent network congestion, limitations in terms of processing and storage, and Service Level Agreement (SLA) breaches. Therefore, resource allocation for an optimized 5G network is vital for both the MNO and the end user. To this end, network slicing facilitates SLAs to the end customer while significantly simplifying network QoS issues, in particular for a resource intense environment such as a smart hospital.

1.2 Contribution of the Thesis

This thesis explores telehealth deployment in 5G wireless networks considering two aspects. It presents a modified architecture for an emerging L5GO network, which provides ultra-low latency communication services suited for telehealth. Related sections in the thesis is based on two such use cases that require URLLC defined by 3GPP. They are, Augmented Reality Assisted Surgery (ARAS) and Robotic Aided Surgery (RAS). The proposed architecture comprises of both the core and the access 5G network functions needed to support communication.

In addition, the thesis explores the suitability of vertical and horizontal slicing for smart hospital environments that utilize telehealth applications such as ARAS, RAS and Internet of Medical Things (IoMT). The study formulates two convex optimization problems for optimal resource allocation for these applications, considering the two network slicing methodologies. Then, the author solves the optimization problems considering several smart hospital scenarios that differ from each other based on their medical speciality, *e.g.*, emergency services, infectious disease centers, pediatric hospitals, elderly care centers.

1.3 Organization of the Thesis

The thesis has six chapters, the remainder of the thesis is organized as follows. Chapter 2 presents the literature survey and related work. It articulates the 5G mobile network architecture and its key enablers, network slicing, L5GO networks and the telehealth use cases reinvent. Chapter 3 describes the L5GO network registration process and the Protocol Data Unit (PDU) session establishment, and presents the proposed L5GO architecture. Section 3.2 of Chapter 3 presents three 5G mobile network deployment models, which are compared in Section 3.3 through numerical results. Chapter 4 describes the system setup with the applications, services and resources used to build network slicing comparison model. Section 4.2 describes the related applications, services and resources studied in the thesis. Section 4.3 presents the numerical results. Chapter 5 resents applications of the derived models and future work. Chapter 6 concludes the thesis.

Publications

[1] R. De Silva, Y. Siriwardhana, T. Samarasinghe, M. Ylianttila, and M. Liyanage, "Local 5G Operator Architecture for Delay Critical Telehealth Applications," in *2020 IEEE 3rd 5G World Forum (5GWF)*, pp. 257–262, Sep. 2020.

[2] R. De Silva, Y. Siriwardhana, T. Samarasinghe, M. Liyanage, and M. Ylianttila, "Deployment Options of 5G Network Slicing for Smart Healthcare," in *2021 IEEE 7th Conference on Network Functions Virtualization and Software-Defined Networking (IEEE NFV-SDN 2021).* Under Review.

Chapter 2

Literature Survey

2.1 5G Mobile Networks

5G mobile networks represent a paradigm shift in wireless communication. It is the new global mobile wireless standard after 4G. 5G networks are digital cellular networks designed to connect everyone and everything. The three main features that characterize a 5G network are its ability to support enhanced Mobile Broadband (eMBB), massive Machine Type Communication (mMTC) and the provisioning of Ultra-reliable Low Latency Communication (URLLC) services [4]. Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS) has set 5G network deliverables to be [18] :

- 1000 times higher mobile data capacity for an area
- 10 - 100 times increment in the amount of connected devices
- 10 - 100 times increment in user data rate experience
- 10 times lower battery consumption in low power massive machine communications
- Millisecond level E2E latency.

compared to 4G networks. These are highly challenging specifications to achieve. Thus both Next Generation-Radio Access Network (NG-RAN) and the 5G core network have adopted numerous new technologies in order to deliver them.

2.1.1 5G Network Architecture

In order to provide guidelines towards 5G realization, standard bodies have defined their vision contemplating novel technologies and customer expectations. Table 2.1 points out

Table 2.1: Design principles of NGMN 5G network vision.

Radio	Operation and Management	Network	End-to-End network
1. Spectrum leverage (a) Utilize higher frequencies and unlicensed spectrum range (b) Control plane and data plane split (c) Up-link/Down-link split 2. Cost effective dense deployments (a) 3-rd party and user deployments (b) Network configuration, optimization automation (c) Enhance multi RAT coordination (d) Support multi operator deployments with shared network infrastructure 3. Interference management (a) Massive MIMO and CoMP (b) Make use of controlled non-orthogonal interfaces 4. Dynamic radio network topology support (a) Cells, relays and hubs movement (b) Device-to-device communication	1. Operation and management simplification (a) Self-healing and automation (b) Probeless monitoring (c) Integrated Operation and Management functionality (d) Cloud network orchestration of carrier networks	1. Common composable core (a) Entities and functionality count minimization (b) Control and user function split (c) No mandatory user plane functions (d) Legacy inter-working minimization (e) RAT agnostic core (f) Convergence of fixed and mobile core	1. Adaptation of flexible functions and its capabilities (a) Network slicing (b) Function variance (c) Flexible function and service allocation (d) NFV (e) SDN 2. Support new value propositions (a) Facilitates software/infrastructure as a service (b) Radio and network API exposure to 3-rd parties (c) Exploit big data and context awareness 3. Build in security for control plane and user identity protection

key design principles the Next Generation Mobile Network (NGMN) has considered for their E2E 5G network vision development [1].

Figure 2.1 [1] illustrates the 5G system architecture NGMN has proposed base on the above design principles. It is a combination of three layers, namely business application layer, business enablement layer and infrastructure resource layer. The business application layer includes applications and services of the operator, enterprise, verticals, and third parties that utilize the 5G network. The business enablement layer is a collection of functions a converged network requires to operate, which are in modular blocks. These functions can vary from software modules to configuration parameters, the same function can generate different levels of outputs based on the implementation and usage methods. The infrastructure resource layer contains the physical resources of a 5G network *e.g.,* end devices, access nodes, network nodes.

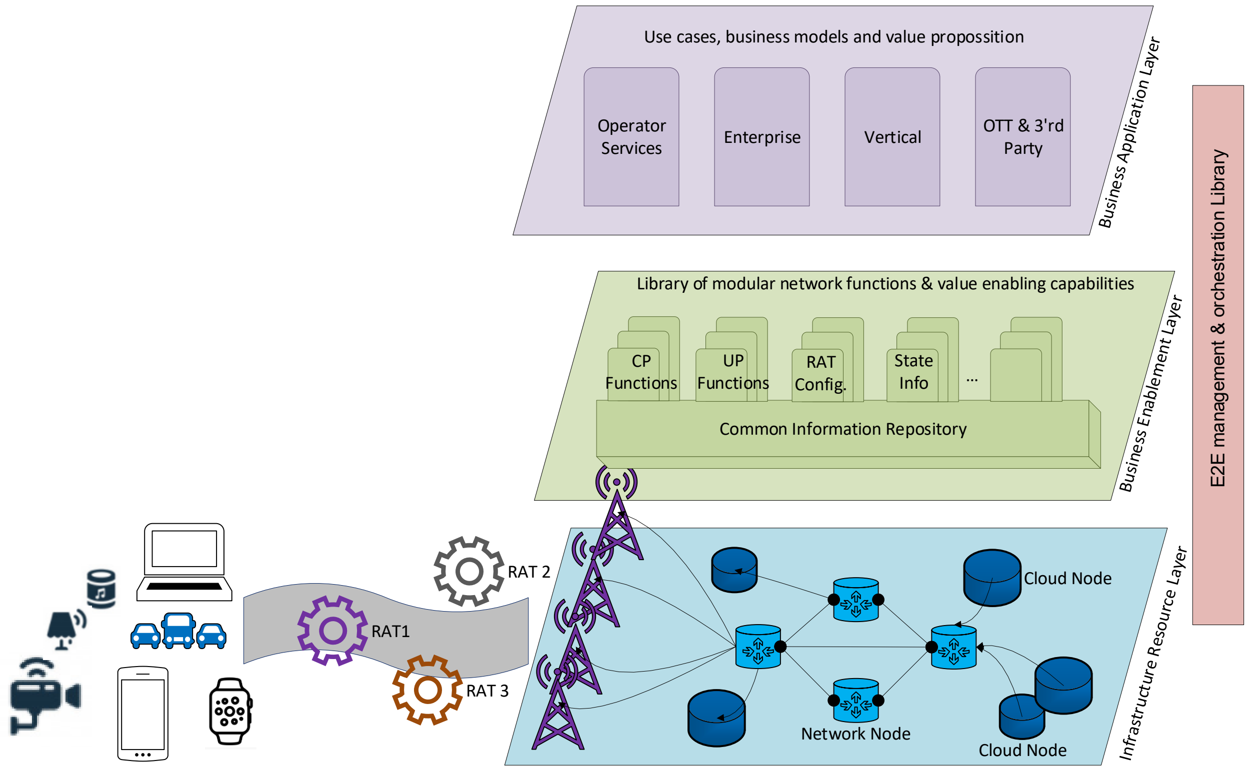


Figure 2.1: ETSI 5G Architecture [1].

3GPP, the main mobile network standardization body defined 5G mobile network architecture as given in Figure 2.2 [2]. It is a modularized design based on User Plane (UP) functions and Control Plane (CP) functions separation. The architecture includes a converged core network, which minimizes the dependencies between the access network and core network by providing a common interface for 3GPP and non-3GPP acess networks. Table 2.2 gives a description of Network Function (NF) included in the non-roaming system architecture.

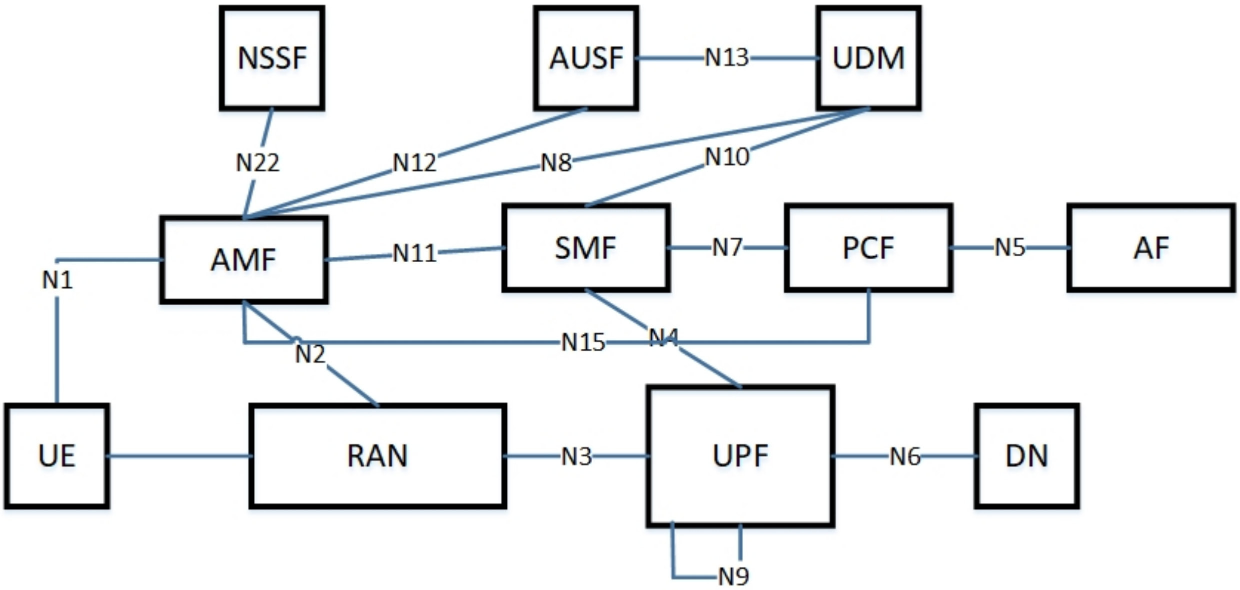


Figure 2.2: 3GPP non-Roaming 5G System Architecture in reference point representation [2].

2.1.2 Key enablers of 5G Networks

5G networks carry the advantages of millimeter wave frequencies, a key enabler of modern 5G networks which refer to 30-300 GHz frequency range. Mobile access networks can deliver gigabit range data rates with the use of mm wave to the end user devices increasing the user data rate [19]. Massive Multiple input multiple output (MIMO) with a large number of antennas at the next generation NodeB (gNB) enables spectral and energy efficiency compared to single antenna systems in 4G networks, thus empowering low power consumption [20]. Ultra-Dense Network (UDN) is another enabler of 5G networks, which empowers the realization of small cell network deployments leading to high data throughput per area [21].

Network Function Virtualization

NFV is considered as one of the key technology, in realizing modern 5G networks. NFV and its Management and Network Orchestration (MANO) offers less complexity in network maintenance while reducing the time to market of new services [4, 22]. Traditional mobile networks have dedicated hardware for each network element, which leads to under utilization of resources in most cases. In addition to that it leads to vendor locking, resulting capital and operational expenditure increases. NFV leads vendor neutrality by running mobile network functions on general purpose hardware, while also enabling flexible pay as you grow models to MNO [22, 23].

With the introduction of NFV to 5G, high availability features of cloud networks are

Table 2.2: Description of 5G Network Functions.

5G NF	Description
AMF	Access and Mobility Management function. Responsibilities include, but not limited to registration management, connection management, reachability management and mobility management. Transport Session Management (SM) messages and proxy for routing SM messages
SMF	Session Management function's main responsibilities are session establishment, user equipment Internet Protocol (IP) allocation and management, Dynamic Host Configuration Protocol (DHCP) functions, selection and control of UP function and charging data collection and support
UPF	User Plane Function acts as an anchor point for intra/inter RAT mobility, external PDU session point for data networks while handling the responsibilities of traffic usage reporting, packet inspection, UP Quality of Service (QoS) handling
PCF	Policy Control Function includes supporting policy framework to govern network behaviour, CP function policy generation and subscription information access
UDM	Unified Data Management provides support mainly for generation of authentication credentials, user identification handling, subscription management
AUSF	Authentication Server Function provides authentication for 3GPP and non-3GPP service access
NSSF	Network Slice Selection Function is responsible for network slicing in 5G network by determining the set of network slicing instances with allowed and configured Network Slice Selection Assistance Information (NSSAI)
AF	Application Functions communicate with 3GPP core network to provide application influence for routing traffic and also interact with the policy framework for policy control
DN	Data network, internet access or 3rd party services

adapted as an extra advantage. In NFV capable networks, network function deployment methods are mainly two folds, as Virtual Network Function (VNF) base on virtual machines or as Cloud-native Network Function (CNF) using container images. European Telecommunications Standards Institute (ETSI) has introduced a framework where such functions can be easily orchestrated and managed [3] which is illustrated in Figure 2.3. It consists of three main functional blocks namely the Virtualized Infrastructure Manager (VIM), the Virtual Network Function Manager (VNFM), and Network Function Virtualization Orchestrator (NFVO). The NFVO is responsible for network service management, thus it overlooks VNF on-boarding and overall E2E life cycle management. Since NFVO ensures service integrity, it offers reference points to external systems, Operating Support Systems (OSS) and Business Support Systems (BSS). Further NFVO is connected to databases and catalogs *e.g.*, network service catalog, VNF catalog and NFV infrastructure resource database, where it extracts information about relevant entities. The VNFM oversees the life-cycle of a VNF, therefore it is connected to VNFs and element managers for starting, configuring and scaling of associated entities. The VIM is connected to NFV infrastructure in order to manage VNFs and CNFs which are in virtual machines and containers, respectively.

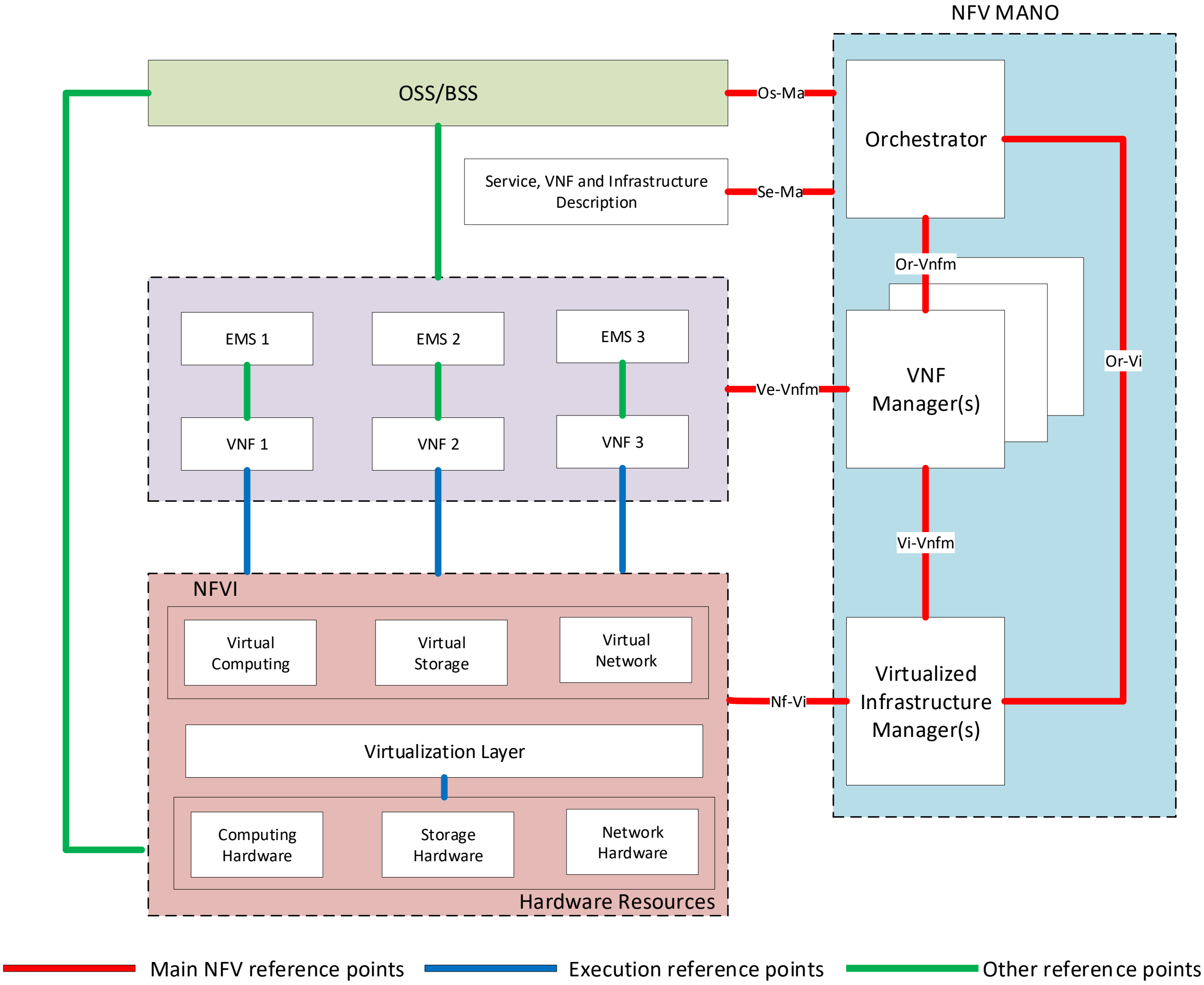


Figure 2.3: NFV reference architecture framework [3].

Software Defined Networks

SDN is another main enabler of modern 5G networks. The motivation behind SDN is to decouple the control plane and data plane from hardware based traditional MNO solutions [24]. Hardware and software of a MNO solution has different release cycles, therefore the introduction of SDN to 5G networks helps to get the full advantage of ever changing software. SDN brings network programmability to the picture with the decoupled control plane. This allows a dedicated SDN controller to program and optimize network resources to deliver a customize solution at much lesser time [4].

SDN deployment in a 5G MNO network can have multiple controllers established in the

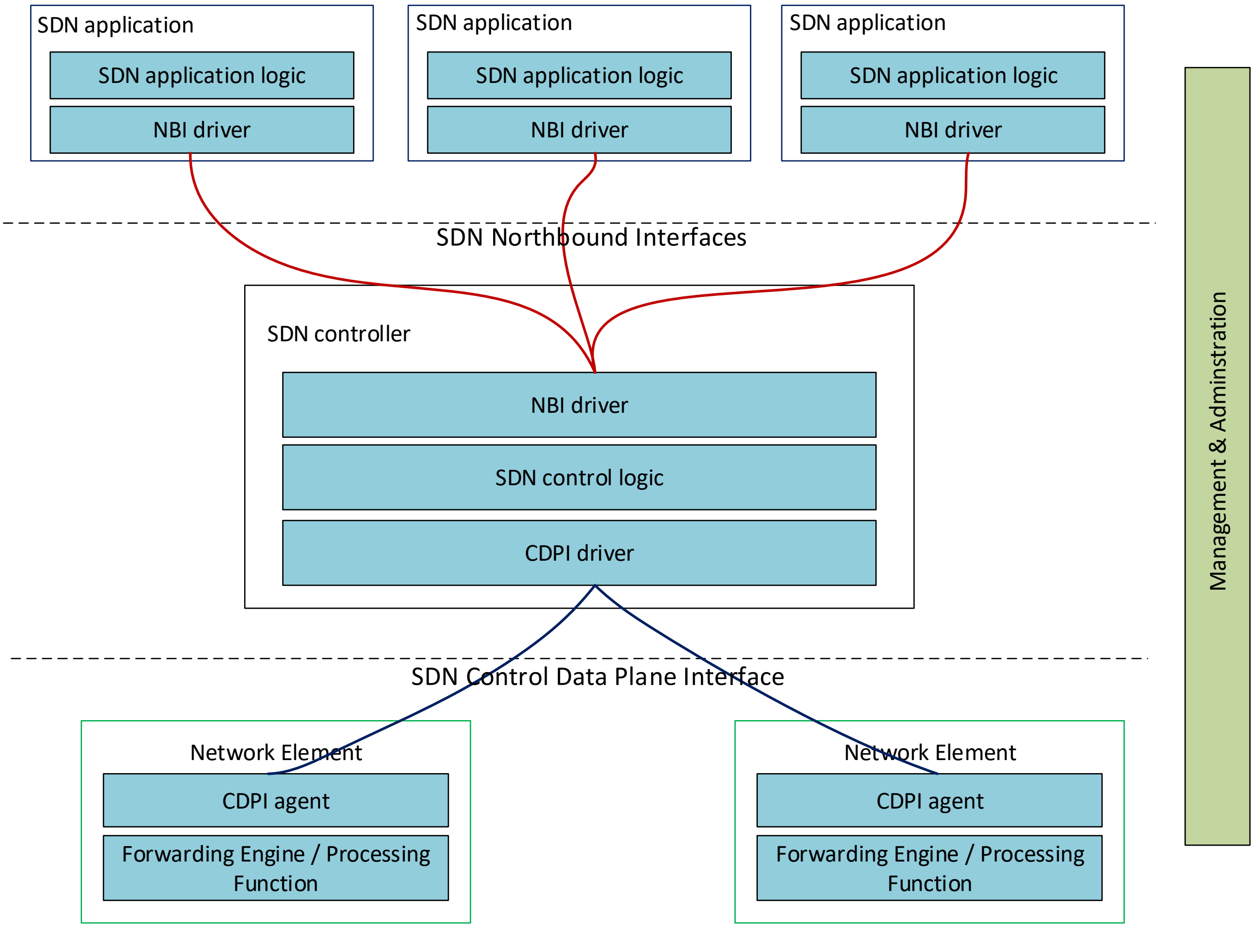


Figure 2.4: A high-level overview of SDN architecture [4].

network focusing on different sectors [4]. As illustrated in Figure 2.4, SDN architecture is developed around the controllers. Provisioning, management and control of services and other related resources are done by the controller. Service utilization happens in relevant network resources, controller can alter the traffic flow from application to the resources and vise versa. This simplify traffic handling inside a complex network [4, 25].

Mobile Edge Computing

MEC standardized by ETSI enables new services and yields promising business models to 5G MNOs, thus a 5G enabler. The most renowned advantage MEC brings to the table is reduction of network delay, which motivates URLLC service delivery through 5G networks [10]. MEC brings network processing and computing capabilities to the Radio Access Network (RAN) edge. This enables ultra-low latency and high bandwidth service delivery. In addition to that, MEC offers an open radio network edge platform, which encourages multi service and multi tenancy through allowing storage and processing access to 3-rd parties. [11] further points out the below MEC use cases and applications in relevance to 5G networks.

- Computation offloading
- Distributed content delivery and caching
- Web performance enhancements
- Internet of Things (IoT) and Big-data
- Smart city services
- Application awareness and content optimization

Figure 2.5 illustrates ETSI introduced generic reference MEC architecture [5]. It contains three reference point groups, namely reference points associated with the MEC platform functionality -Mp, management reference points -Mm, and the reference points for external entities -Mx. The multi access edge system contains MEC hosts and management, which are essential for MEC application performance. MEC host contains a MEC platform and the virtualization infrastructure. They are responsible to provide compute, network and storage resources for MEC applications. The MEC platform is a combination of vital functionalities MEC applications need to run on a virtualization infrastructure, and it enable MEC applications to provide and consume MEC services. MEC applications are installed on the virtualization infrastructure, and the MEC management is empowered with system level management and host level management. MEC system level management contains the Multi-access edge orchestrator, a core element which oversee the complete MEC system. MEC host level management comprises the platform manager and the virtualization infrastructure manager, they are responsible for specific functionality management of MEC hosts and applications [5, 10].

2.2 Network Slicing

Network slicing is a key concept of efficient agile network deployment, which supports diverse service delivery for emerging businesses. It enables slicing of a single network in to multiple instances, which are architected for specific application and/or service with defined network requirements. Network slicing introduces multiple self-contained, logical network deployment on a common physical infrastructure, leading to flexible network resource provisioning and dynamic network function assignments. Through the introduction of network hypervisor, network slicing allocates physical and virtual network resources to separate tenants. This enables resource sharing between virtual MNOs running on the same infrastructure. To facilitates service requests, network slicing generates E2E resource pool with bandwidth, processing, storage, network function as well as RAN [11]. Figure 2.6 gives an understanding

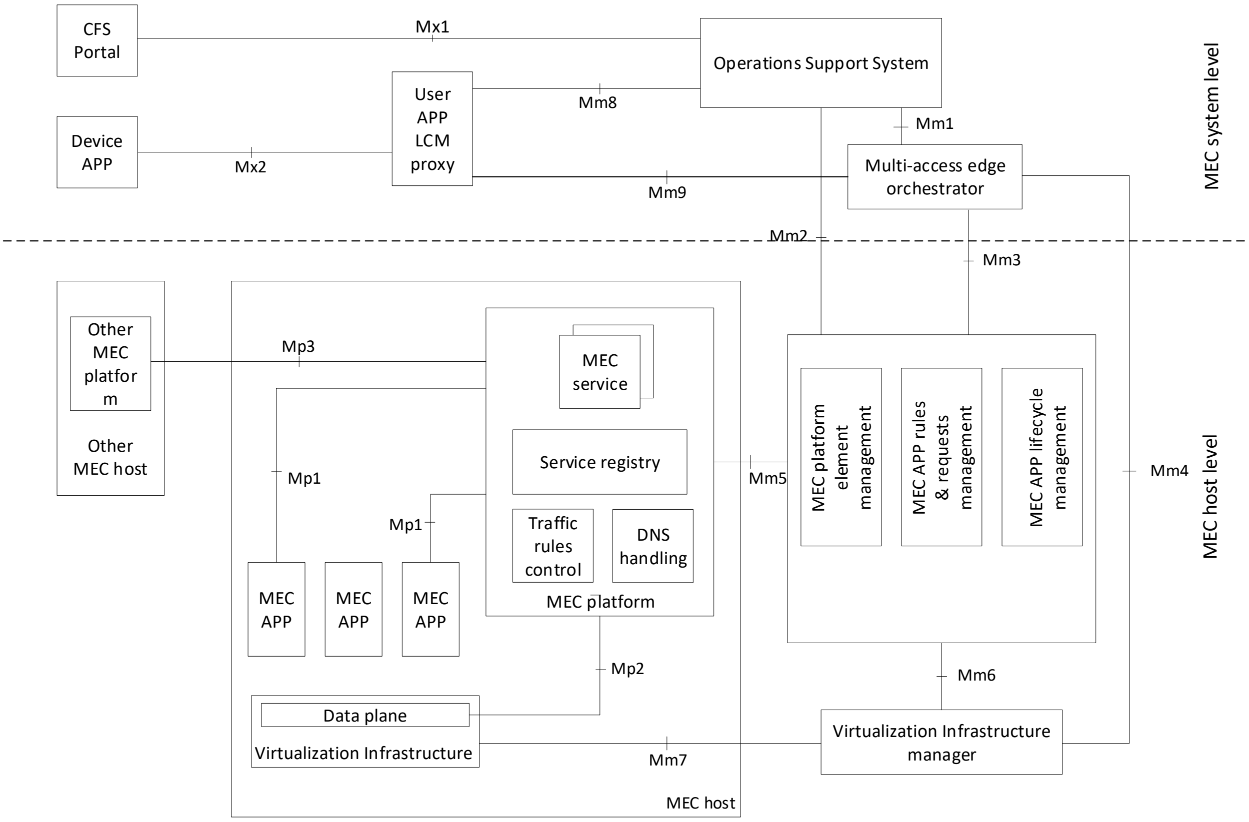


Figure 2.5: ETSI MEC architecture [5].

of the network slicing concept, and how customized service delivery occurs through different network slices.

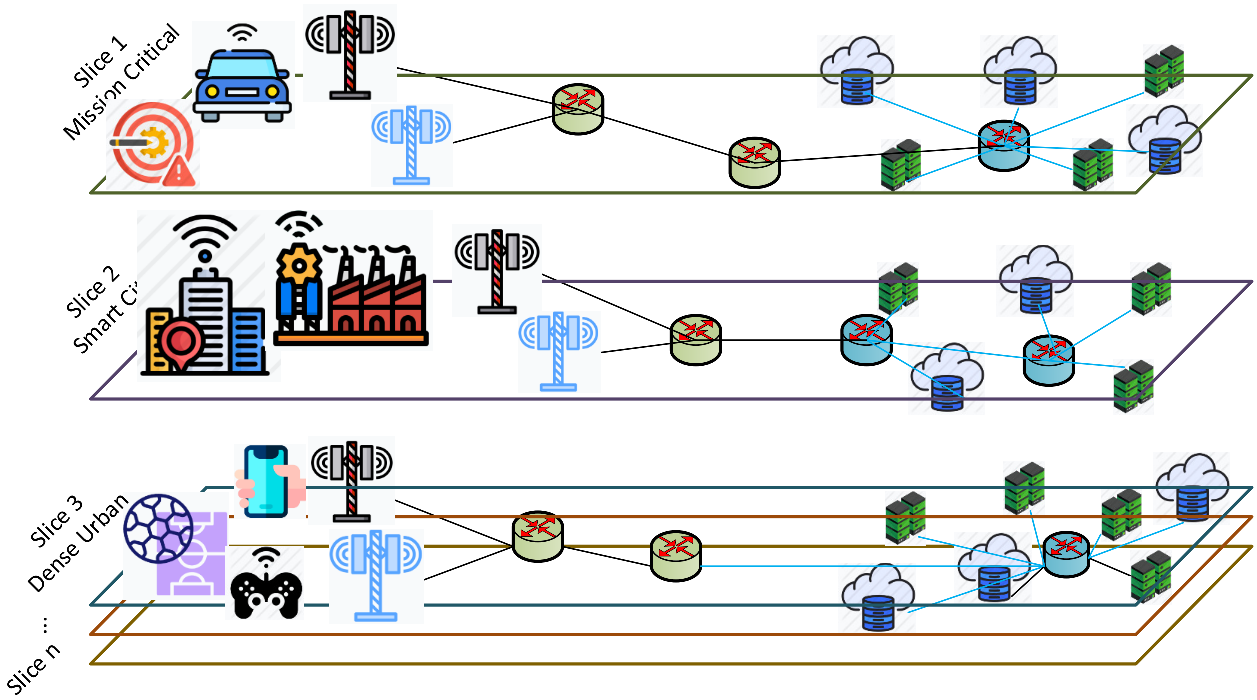


Figure 2.6: Concept of network slicing.

2.2.1 Network Slicing architecture

In [26], NGMN describe the network slicing concept, which consists of three main layers, namely service instance layer, network instance layer and resource layer. The service instance layer contains the end user and business services, which are facilitated by an MNO or a 3-rd party. An MNO creates the network slice instance using a network slice blueprint, which provides network characteristics a service instance needs to operate. The network slice instances can be created by an individual or several sub-network instance, also it can be composed with out any sub-network instance as illustrated in Figure 2.7. 3GPP defines network slicing as a technology that enables MNO to deliver optimized network solutions to a diverse customer base which demand various network requirements [6].

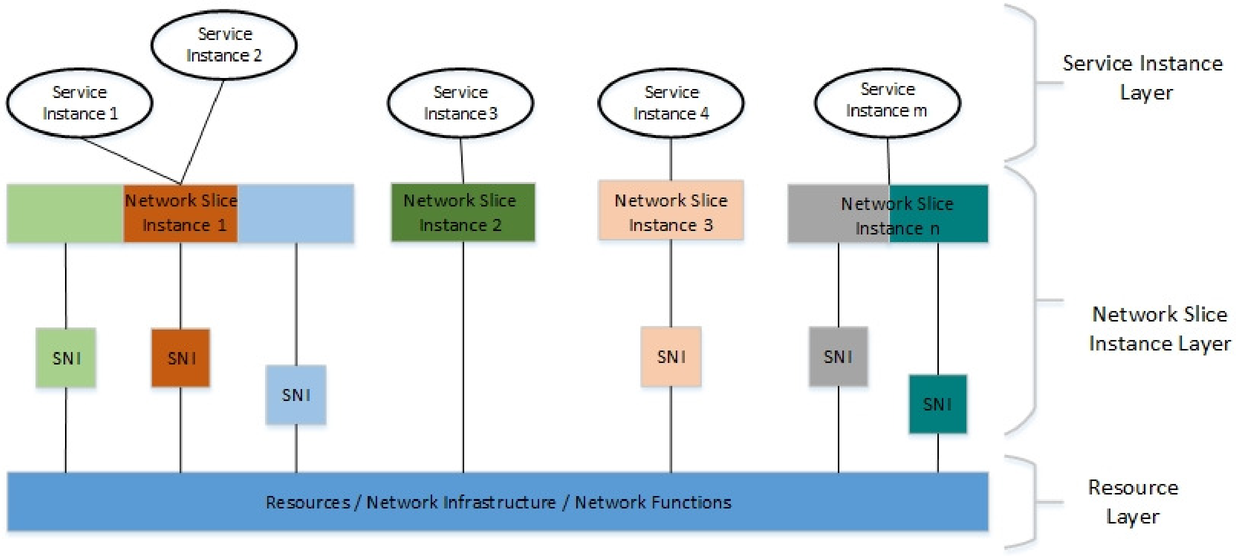


Figure 2.7: NGMN network slicing conceptual outline [6].

Network slicing principles

Network slicing is built base on the following principles [6] :

1. Automation : Empower an on-demand deployment/configuration of network slicing without manual engagement. This allows 3-rd parties to originate slice creations with the desired E2E network parameters such as capacity, latency and jitter.

2. Isolation : a basic property which enables performance guarantees and security of the network slices running on common network infrastructure. However, due to the defined resource separation, slice isolation can lead to resource under-utilization. Isolation deployment in shared resources can be achieved through resource sharing with respective policy guidance and with separation via virtualization.

3. Customization : ensure the allocated resources to a tenant are fully utilized. Slice customization can be achieved through

   (a) The separation of data and control plane in a network level abstract topology

   (b) On the control plane with the introduction of programmable policies

   (c) On the data plane with customized NF and data forwarding policies

   (d) Using Value Added Services (VAS) *e.g.*, big data, context awareness

4. Elasticity : is a property related to resource allocation to a particular network slice. It enables varying radio and network conditions, number of serving users and serving area to achive desired SLA. Elasticity of 5G network resources is realized through remodeling

allocated resources, network policy adjustments and, changing functionality of certain data and control plane elements.

5. Programmability : authorize 3-rd parties the control of network slice resources via a open Application Programable Interface (API), which discloses network properties for resource elasticity.

6. End-to-end : is a property related to service delivery from service providers to the end users. E2E property stretches along different domains and it merges divergent network layers and technologies *e.g.*, RAN, core network,transport network.

7. Hierarchical abstraction : is based on recursive virtualization, where resource abstraction approach is recurred to gain higher level of abstraction and broader scope with each repeat.

Network slice orchestration and management

5G network slice is a combination of VNF, CNF, Physical Network Function (PNF), VAS, transport network and RAN hardware. An illustration of network slice orchestration architecture is given in Figure 2.8.

The architecture mainly consists of [6]:

- End-to-End service management and orchestrator : performs multi-domain slicing, slice brokering, service mapping, policy provisioning and admission control. It is also responsible for receiving network slice requests from 3-rd parties and different verticals

- Virtual resource orchestration : oversees VNF embedding and instantiation of the virtual network service graph along with the life-cycle management of VNFs and VASs.

- Network resource programmable controller : empower VNF service chaining, QoS control along with the resource programmability. Network resource programmable controller can be a VNF allowing 3-rd parties to program the slice resources or a PNF providing network programmability in place of 3-rd parties.

- Life-cycle management : performs network element administration and management and policy provisioning.

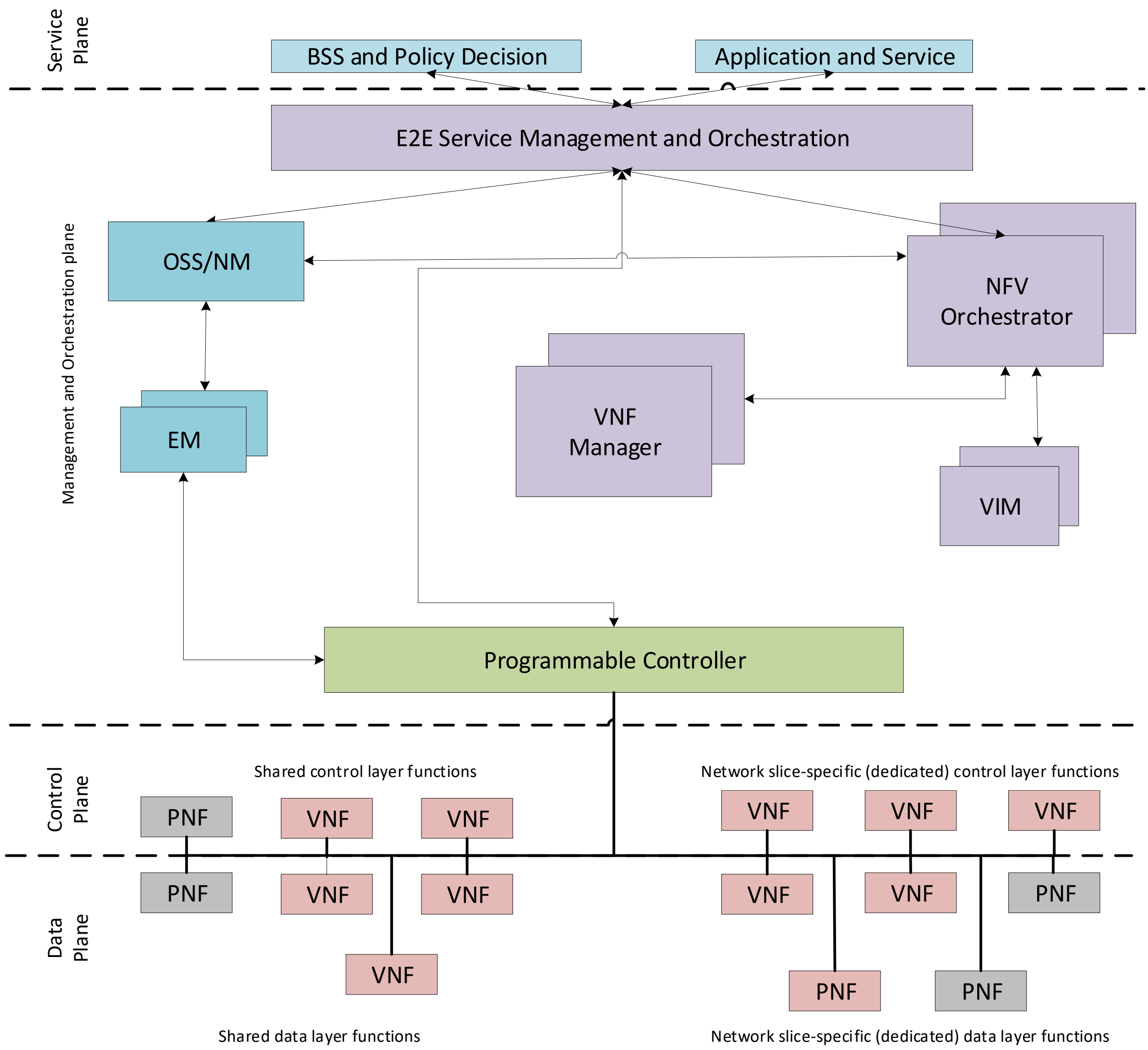


Figure 2.8: 5GPP network slice orchestration architecture [6].

2.2.2 Different network slicing methods

[13, 14, 16, 27] discuss different network slicing approaches which have unique advantages in service delivery. Based on characteristics such as service delivery, resource allocation and SLA network slicing methodologies are categorized and discussed in the literature.

- Slicing for QoS : Targets to create different types of network slices in order to provide diverse QoS range to the end customer, which can lead to numerous business opportunity generations. Network slices created specifically for live video streaming, medical surgeries and for IoT communications can be taken as examples for these type of network slices.

- Slicing for Infrastructure Sharing : The basic idea is to virtualize RAN domain of a mobile service provider network and share it between several operators. Infrastructure owner lease the slice to a tenet based on an agreement, which facilitates the control of both functions and infrastructure of that slice. This leads to a optimized scalable network, while creating numerous revenue generating opportunities.

- Static network slicing : In this scenario, network slices are pre-installed and pre-configured. The device needs to select the slice which it is going to establish a connection with.

- Dynamic network slicing : Can be identified as the most widely used network slicing methodology. This enables MNO to dynamically design, deploy, customise and optimize based on the customer service/business requirements and network conditions. The novel service concept Network Slice as a Service (NSaaS) is realized through dynamic slicing.

- Vertical slicing : This approach of network slicing has been in the picture since 4G network era. Its applications focus more towards core network slicing. However, it is applicable to whole 5G MNO, to that end each network domain is sliced separately and the slices are pared using slice paring function. To serve vertical industries (*e.g.*,automotive, manufacturing, eHealth, energy and smart cities) in a cost effective manner, vertical slicing methodology slices mobile networks vertically, which enables separation of sliced traffic from rest of the mobile traffic. This procedure reduces the complexity of traditional QoS problems.

- Horizontal slicing : Can be identified as the novel approach of network slicing which enabled through 5G mobile networks and related technologies such as edge computing and computation offloading. Edge computing and computation offloading make mobile traffic non-uniform at the edge. By doing so, it allows the edge network elements of 5G networks to process user generated traffic. It reduces the need of high computational and processing power requirements at the end user devices. Technologies like Augmented Reality (AR), Virtual reality (VR) require high computation capability and low latency, which ultimately increase the price tag of the devices and related applications. Horizontal slicing provide an advantage for such use cases as it reduces the traffic latency drastically, through providing traffic processing at the edge.

2.3 Local 5G Operator Networks

With the 5G mobile network deployments, world will move to a hyper connected environment. Promised low latency and high bandwidths in 5G will open-up innovative methods of connections in real time usage aspects as well as location specific. Facilities such as hospitals, shopping malls, sport arenas and industrial factories demand wireless connectivity with different levels of reliability, data rates, latency and security. Therefore it opens up gaps in the vertical sector, that needs to be filled by separate entities that have the required capabilities. Recently proposed L5GO networks come as a versatile concept to provide necessities to high demanding location specific wireless connections.

2.3.1 What are local 5G operator networks?

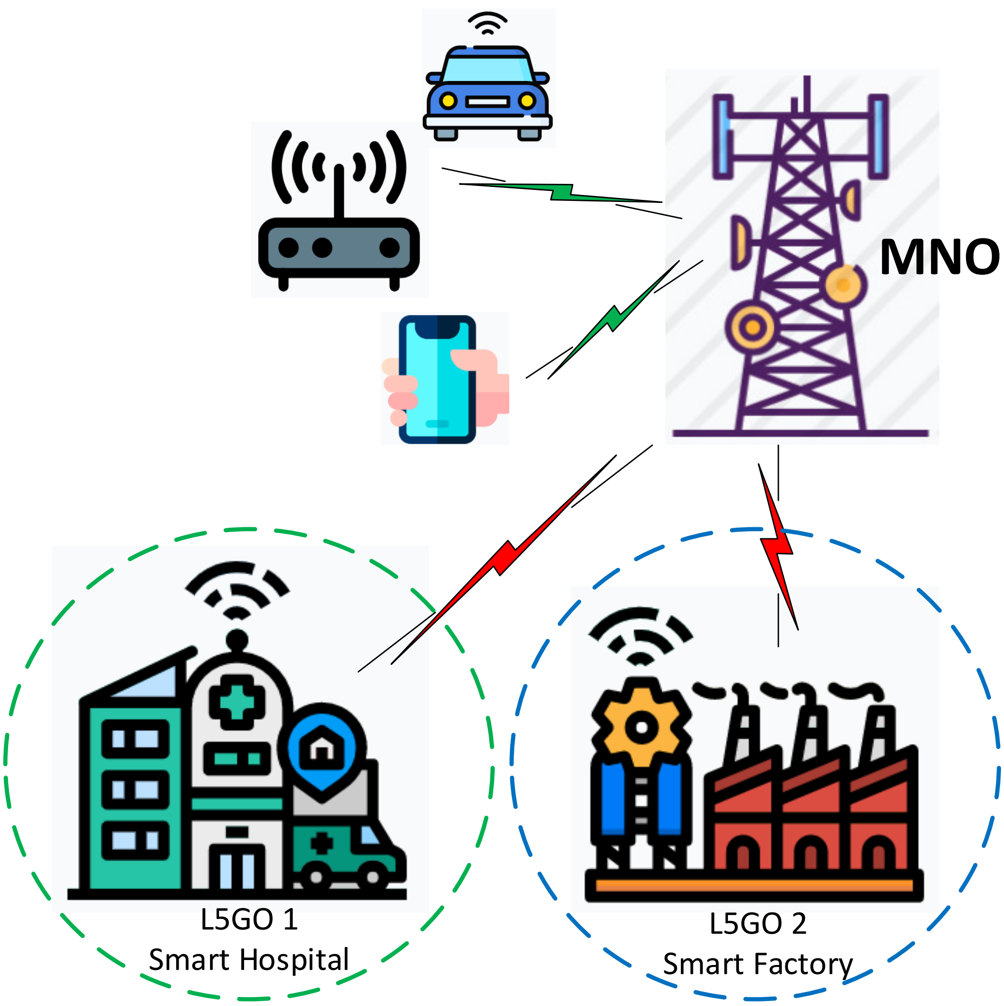


Figure 2.9: Collaborative service delivery by both L5GO and MNO.

Traditional MNOs are solely focused on the coverage area and the number of subscribers to the network. Even with the 5G networks there is a doubt about MNO's capability on providing low latency services to real time use cases and security they demand. Present deployments of 5G networks are based on mm wave frequency, local dense small cell networks with SDN and NFV for implementation and maintenance flexibilities . NFV enable separation of contrasting network functions to operate under different domains controlled by several stakeholders. SDN on the other hand gives the separation of control and data planes which ultimately leads to the full control of the network to the operator. Use of these technologies changes the ways of traditional MNO and will allow the partition of the network leading L5GOs to enter the domain. Complex environments like hospitals and industrial factories requires wireless solutions to connect their devices. Those connections also need tailored latency and reliability with specific security features. Figure 2.9 illustrates the collaborative service delivery between MNO and L5GO, where MNO provides a generic service while L5GOs focus to provide location and case specific services.

L5GO been a novel concept, there has to be an understanding on how it may prevail with the current operating models in the telecommunication industry. It has to provide local connectivity for MNO customers, therefore L5GOs have to have a strong collaboration with MNOs to share their infrastructure in the local environment. On the other hand, L5GOs have to cater specific requirements in order to satisfy this. Strong relationship with network infrastructure vendors, device and equipment vendors as well as content providing parties are essential. Most importantly, current regulatory organizations do not support L5GO concept. However to realize obtaining local spectrum licenses is a must. On the path to success, there are many facts to be considered, technology vise as well as financially. Successful deployments will lead to 5G network expansions.

The amount of traffic originated from diverse business environments grows exponentially. Thus timely implementation of communication services for such cases will generate more revenue opportunities for mobile service providers. However, focusing on location and case specific service delivery deviates MNO's business targets. Collaboration with L5GOs for such service deliveries will reduce capital and operational expenditure from MNO's standpoint, while providing full advantage of 5G networks to customers.

2.3.2 Research work carried out on local 5G operator networks

L5GO is a new and experimental concept for the telecommunication world. Table 2.3 contains a summary of important research work published on L5GO so far. It is important to note that the concept 5G Micro Operator (μO) is the same as L5GO. Initial research works carried around L5GO used the term 5G μO to name a local, case specific, location oriented local 5G network.

Table 2.3: Summary of important research works on L5GO.

Reference	Main contribution
[12]	Proposes μO concept for local service delivery in 5G for indoor small cell communication infrastructure and offers context related services and content. Also discussed regulation related aspects of local spectrum access rights, and technology related factors of flexible network implementation. Presents business opportunities for the μO concept such as operation of secure networks for vertical sector specific use, provisioning of hosted local connectivity to all MNOs in specific locations and offering of locally tailored content and services.
[28]	Expands the μO concept through recognizing and discussing key regulatory elements related to the mobile communication market. To this end, points out spectrum authorization for a quality guaranteed spectrum, the role of an MNO with rights and responsibilities, access rights to network infrastructure, and deployment of indoor networks.
[29]	Discusses cloud based deployments to empower connectivity and service provisioning in new mobile operator networks using μO concept. Points out the use of μO in supporting new application and services from verticals including related challenges, impacts on network security and the regulatory support from spectrum regulation bodies. Points out performance issues for fast and efficient deployment of 5G networks related to verticals and markets, providing tailored services and applications.
[9]	Proposes an architecture for 5G μOs focusing a smart factory environment, which supports industry 4.0 standards and discusses it in terms of network functions, and the operational units expands thorough the core and radio access networks. Produces simulation results for the latency measurements of the proposed μO architecture considering an AR use case in industrial internet. Points out benefits of μO driven local 5G networks for specialized user requirements compared to traditional MNO network deployments.
[30]	Discusses business model options for L5GOs in the context of contrasting network deployment options. Presents three generic 5G business models and respective value ecosystems, namely Vertical business model and ecosystem, Horizontal business model and ecosystem, and Oblique business model and ecosystem. In addition, the scalability, adaptability and sustainability of the business models and ecosystems are examined.

[31]	Proposes a deployment framework for indoor small cell campus networks favourable for μOs and MNOs.The framework is managed by the μO, and it uses network slicing to provide the campus with local customized service, while acting as a neutral host for MNOs. Discusses the advantages of μO deployment in terms of reduced costs, additional revenues, dedicated services, coverage, and spectrum utilization compared to a conventional Wi-Fi deployment.
[32]	Identifies the key elements of spectrums especially in the context of L5GO networks. Characterizes and identifies spectrum valuation elements for L5GO networks operating in shared spectrum bands. In order to do so, authors consider spectrum evaluation for 5G from the standpoint of the different stakeholder roles, including regulators, MNOs and entrant L5GOs.
[33]	In addition to the proposed L5GO architecture, authors point out hybrid architecture options, where both the L5GO and MNO collaboratively contribute to establish the core network to cater communications. Proposed architecture is discussed in terms of NFs considering smart factory with respect to an AR, massive wireless sensor networks, and mobile robots use cases.

2.3.3 Business opportunities and applications of L5GOs

[12] discuss four key business opportunities for L5GOs extensively, namely

1. Hosted local connectivity for MNOs in specific locations - Suitable in locations where it is neither feasible nor cost efficient for MNOs to build their own infrastructure.

2. Secure local networks for vertical specific needs - Solution for environments with highly specialized requirements *e.g.*, smart factories, smart hospitals.

3. Locally tailored services - To provide premium personalized content. *e.g.*, virtual reality, augmented reality services.

4. Acting as a "Mydata" operator - Control user specific data or application, and facilitates connectivity services on top of data that generates the central element of the business opportunity.

5G expands and supports diversified areas of usage scenarios and applications from verticals along with mobile broadband communications. Interactive real time communication including Machine to Machine (M2M), to deliver mission critical control and automation tasks can be taken as an example for this. Such services are required in geographically restricted places such as factories, harbours and mining sites. L5GOs can provide URLLC services, which traditional MNOs struggle to facilitate. Such services are required in places like smart factories and smart hospitals, where a 1 ms makes a big difference. [29] talks about such L5GO deployment opportunities and challenges associated with them.

2.4 Communication Services of Critical Medical applications in Smart Hospitals

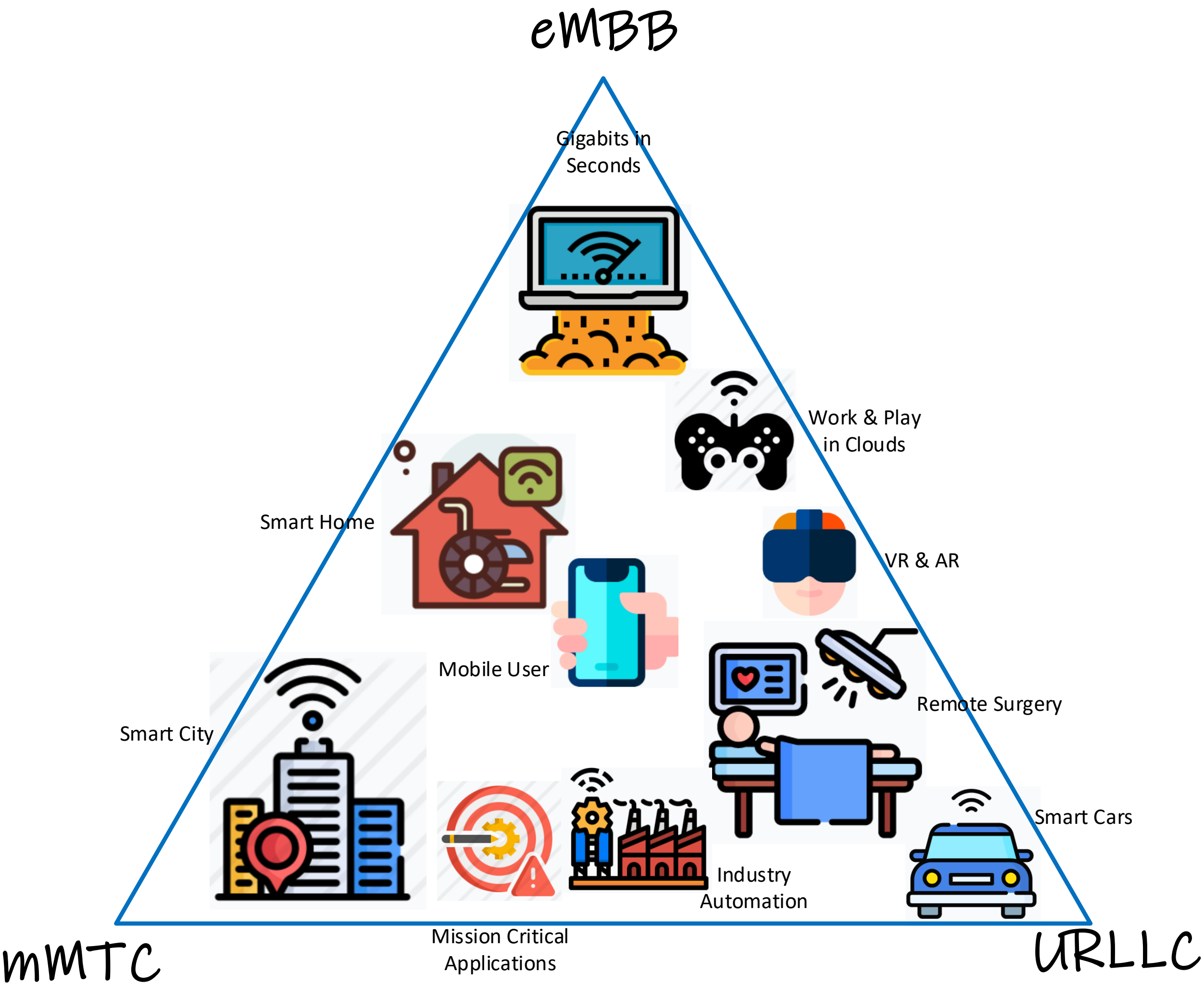


Figure 2.10: Use cases of 5G wireless networks.

5G mobile networks can provide reliable data in real time, which stamp its URLLC feature. Real time reliable data is vital for user applications such as power utility grids and healthcare facilities. Autonomous vehicles, a novel transport approach which require ms level data transmission latency with nearby vehicles and central control applications enabled by 5G's URLLC capabilities. It is expected to have 21 billion IoT devices installed around the world in year 2025 [34]. Such massive amount of sensor networks produce unprecedented amount of data. 5G networks carry the capability to support such large data transactions which resembles its mMTC abilities. enhanced Mobile Broadband (eMBB) refers to high data throughput expanding up to Gbps range. This enables cloud base business opportunities while expanding online gaming and AR, VR based use cases. Figure 2.10 refers to use cases of 5G wireless networks, which expand between the three main features discussed.

A country spends around 10% of its GDP on healthcare, which generate a $7,200 billion worldwide market for healthcare and related products [35]. Reducing the significant cost associated with medical procedures while maintaining the quality standards is a priority as it will broaden the availability of good healthcare facilities even among the people with a low income. To this end, guaranteeing safety and efficacy of the medical devices has been the main priority. However lately, the shift to value-based healthcare has led medical device vendors to deviate their business models from clinical value based to a cost efficiency approach. Premium wireless networks has led this business transformation from the front, enabling coordinated therapies, services, and health analytic. In addition traditional healthcare model has moved from paper records to electronic record keeping methods.

To this end 5G have generated an impact through personalization of healthcare by continuously collecting patient's monitoring data for processing and centralized storage. Also this remote monitoring has enabled a shift in care location from hospital to homes and other low cost facilities. Additionally, another approaches show that 5G can enable cost savings required by the medical industry can be found inside hospitals. For an example wireless transmission of low latency data improves operating room planning, enable efficient equipment usage and simplifies operating theater execution [35]. This study has taken in to account several use cases enabled by 5G networks, that can be applied in a future smart hospital environment. Through them the author tries to study the practicality of the proposed network models.

2.4.1 Augmented reality assisted surgery

Image guided surgery is used in medical industry commonly. Three Dimension (3D) imaging methods such as Magnetic Resonance (MR), Computer Tomography (CT) scans are typically used to give assistance to doctors in diagnosing patients. AR can be considered as an revolutionary methodology in real time imaging. There can be several ways of generating AR views and to identify and follow exact 3D position of a human body, but author is not addressing that in this study [35].

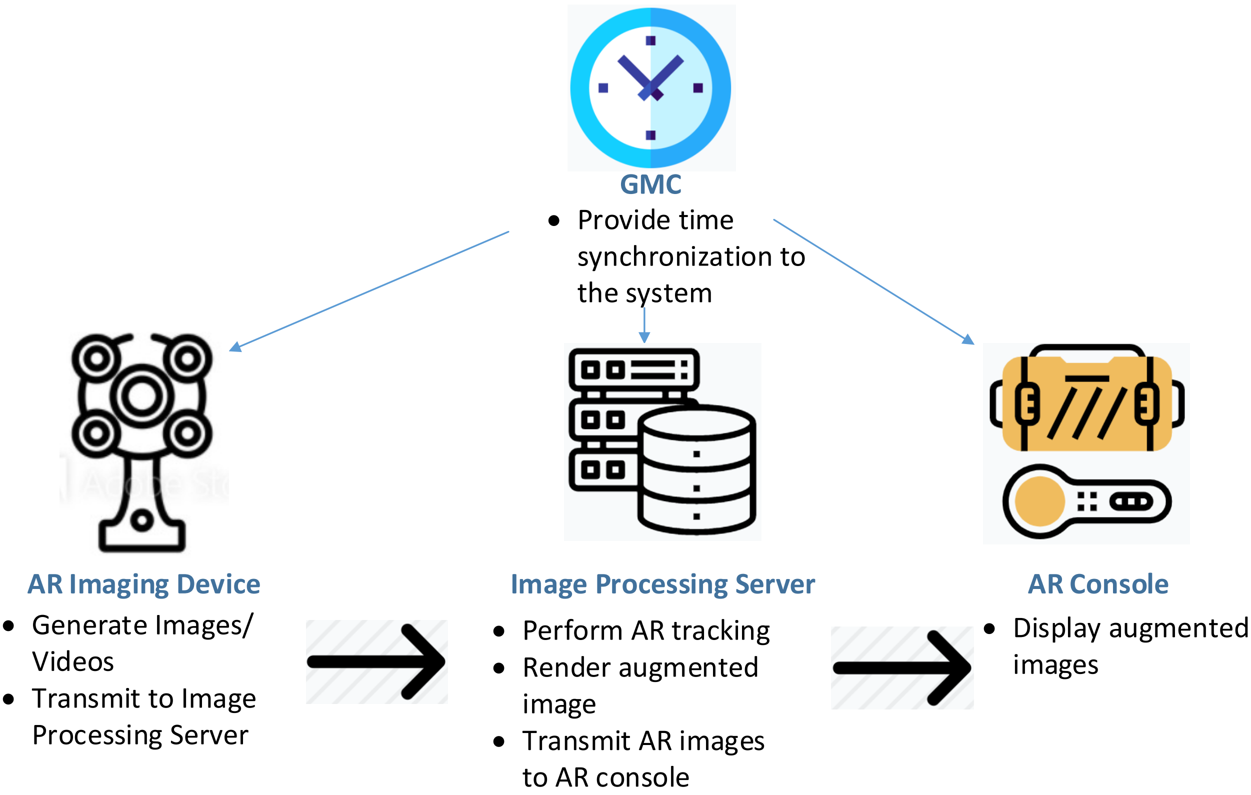


Figure 2.11: ARAS system model

This thesis concentrates on a scenario where AR imaging is generated by a device at the patient-end, and they are transferred to an image processing server to further augment the images, and send to a console device worn by a surgeon. As image processing can add latency to the total system, this thesis expect the scope at the patient to generate 4K uncompressed video compatible with High Dynamic Range (HDR) for up to 10 bits per channel color gamut management capabilities . A Grand Master Clock (GMC) is used to synchronize the setup to a tight accuracy. Figure 2.11 gives a further understanding of the use case system model.

2.4.2 Robotic aided surgery

Used in invasive surgical procedures such as delicate tissue replacements and situations in difficulty of exposure. Patient's end is equipped with a complex machinery that duplicate the surgeon's hand movements with a combination of tiny instruments which have the ability to bend and rotate in axes and areas impossible to maneuver a human hand. If surgeon can get sense of touch via this, it will improve the precision and shorten the learning curve. Thus to do so the control console is fed with haptic feedback to render force applied and the deflection by the tissue in the surgical procedure.

All equipment including the surgeon's console are synchronized with the GMC with an accuracy of 1μs. Robotic Aided Kidney Transplant (RAKT) is a typical surgery performed using this method. Such a surgery requires the transfer of 240 images per second over the network. Having two repetitive errors in any direction is fatal, as they may result in an incorrect command being sent to the actuators. By taking this into account, 3GPP specifies the message error rate to be lower than 0.0001 %, which sets the minimum requirement in terms of reliability, for a network utilized for a RAKT type surgical procedure. [35]

2.4.3 Internet of Medical Things

IoT devices are widely used in the healthcare sector to monitor patients. In certain cases, patients themselves use IoT for self monitoring of health as well. IoT applications made specifically for the medical industry or Internet of Medical Things (IoMT) can be categorized as wearable and ingestible. The most common wearable IoT devices monitor glucose levels, heart rates, blood pressure etc. Ingestible sensors are used to get readings such as blood PH levels and acid levels of the stomach [36], [37].

The conventional health monitoring systems generate discomfort as patients have to be attached to a dedicated machine at all the time of monitoring. IoMT allows patients to carry out regular tasks while enabling continuous remote health monitoring. A remote health monitoring system empowered with IoMT consists of a Portable Patient Monitoring Unit (PPMU) placed in a close proximity to the patient and a central monitoring system at the hospital.

Table 2.4 presents the network level requirements defined by 3GPP in [35] for above use cases.

Table 2.4: Network level requirements of use cases.

Use case	Characteristic parameter					Influence quantity				
	Communication service availability: target value in %	Communication service reliability: MTBF	E2E latency: maximum	Bit rate	Direction	Message Size [byte]	Survival time	UE speed	Num. of UEs connection	Service area
ARAS:4K 120 fps HDR 10bits real-time video stream with lossless compression	> 99.99999	> 1 day	< 750 μs	12 Gbits/s	UE to network	~ 1500 - 9000	~8 ms	stationary	1	100 m^2
RAS:4K 120 fps HDR 10bits real-time video stream with lossless compression	> 99.99999	> 1 day	< 50 ms	12 Gbits/s	UE to network	~ 1500 - 9000	~8 ms	stationary	<1	100
RAS:Motion control data stream	> 99.999999	> 1 day	< 2 ms	2 Mbits/s	UE to network	~ 250	~1 ms	stationary	<1	100 m^2
RAS:Haptic feedback data stream	> 99.999999	> 1 day	< 2 ms	16 Mbits/s	UE to network	~ 2000	~1 ms	stationary	<1	100 m^2
IoT:wireless wearable telemetry device	> 99.9999	> 1 year, ≫ 1 month	< 100 ms	0.5 Mbits/s	-	~ 1000	~1 s	≤ 500 km/h	<1000 per km^2	country wide

Chapter 3

Local 5G Operator Architecture for Delay Critical Telehealth Applications

3.1 The L5GO Network Architecture

This section proposes a possible Local 5G Operator (L5GO) network architectural model that facilitates delay critical 5G telehealth applications. The architectural model is set up to support the communication requirements of the two use cases presented in Section 2.4, namely ARAS and RAS. The network registration, which handles the registration of the terminal devices in the 5G network, and PDU session establishment procedures, which creates a data session between two hosts in the 5G network, are set based on the 3GPP specifications [38]

3.1.1 The L5GO Network Registration

The console sends the registration request via the L5GO access network to the gNB, and generates a registration request at the Access and Mobility Management Function (AMF). As an L5GO typically serves a location specific small user group, it is unlikely to have multiple AMFs in the network, thus eliminating the AMF selection procedure. The AMF exchanges identity request messages with the console in order to establish a connection. Then, the AMF validates the device through the Equipment Identity Register (EIR), and communicates with the Authentication Server Function (AUSF) for device authentication. The AUSF authenticates the device after communicating with the Unified Data Management (UDM) and retrieving the authentication data. The AUSF completes the authentication process with a response to the AMF. Upon successful authentication, the AMF contacts the Policy Control Function (PCF) for the respective policy association for the console. Then, the AMF sends an update to the Session Management Function (SMF) to notify the

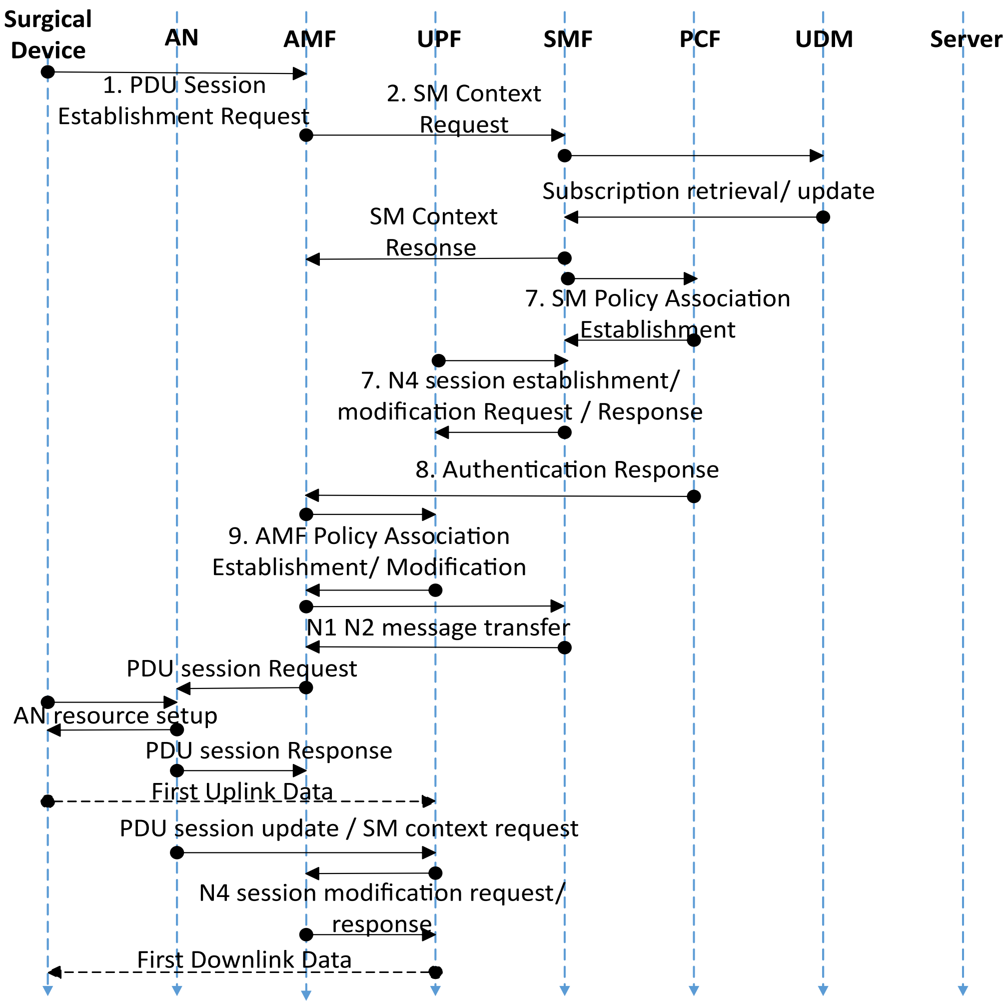


Figure 3.1: The message flow diagram for device registration.

session context. The AMF generating and sending a registration accept message to the console follows. The console replies with a registration complete message to the AMF, and completes the registration process [38].

3.1.2 The Protocol Data Unit Session Establishment

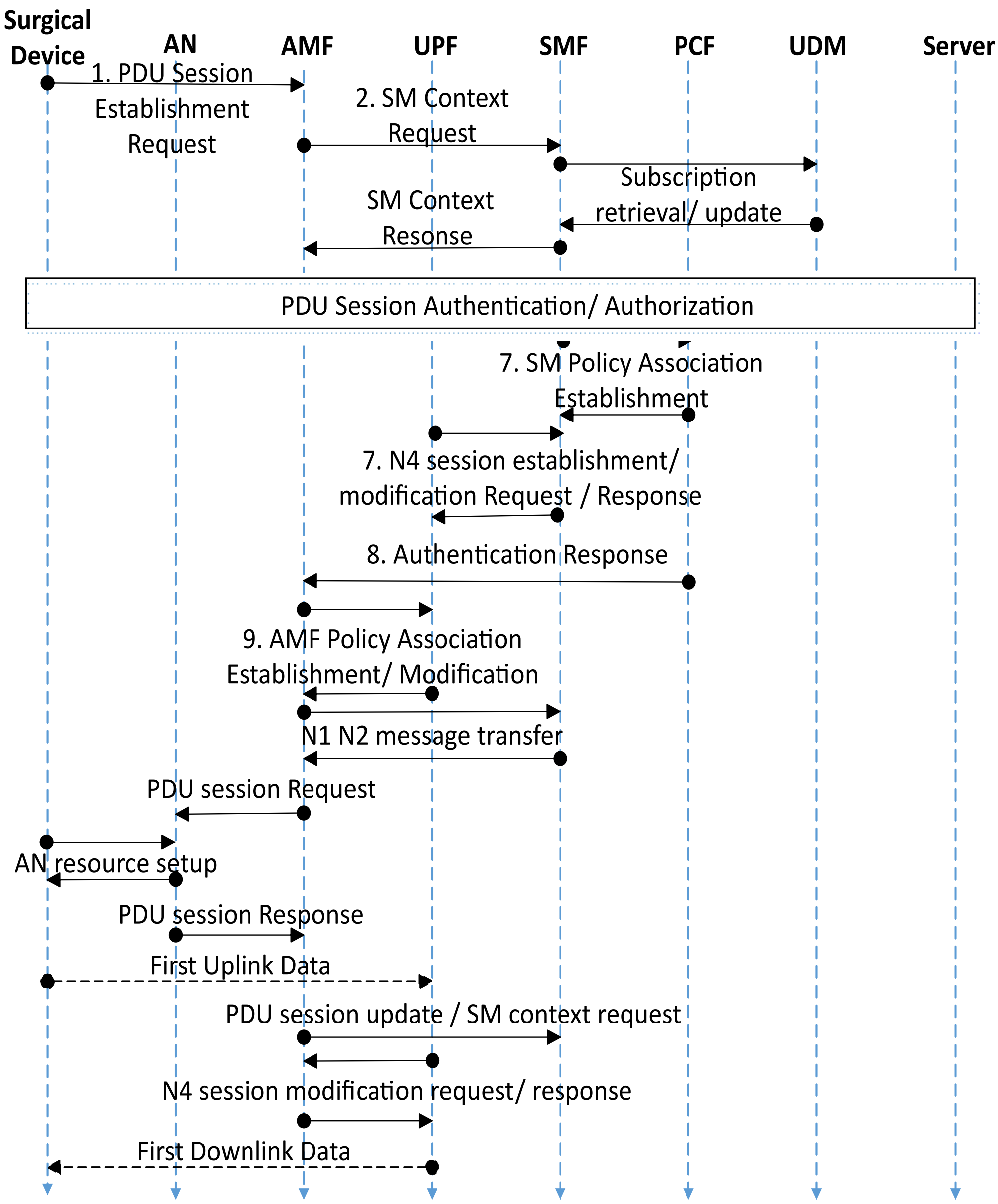


Figure 3.2: The message flow diagram for the PDU session establishment.

Upon successful registration, the console initiates the PDU session establishment procedure with the image processing server, according to the 3GPP specifications [38]. Firstly, the PDU session establishment request by the console is forwarded to the AMF via the L5GO gNB. The AMF generates and sends a PDU session create/session management context request to the SMF. The SMF interacts with the UDM for subscription retrieval, and updates related attributes for this particular session. Then, the SMF responds to the AMF with a session management context. Afterwards, the PDU session authentication/authorization takes place through message exchanges between the NFs in the surgical device and the imaging server, as shown in Figure 3.2. The SMF coordinates with the PCF for policy association for the session. The SMF and the UPF exchange session establishment/modification requests and responses. Then, an authentication response is sent to the AMF by the PCF. The AMF policy association establishment/modification messages interchange between the AMF and the UPF follows. Next, the SMF and the AMF initiate a message transfer session. The AMF transfers the PDU session ID details to the gNB so that the gNB can coordinate with the surgical device for the gNB defined resource setup. Based on that, the L5GO's AN sends a PDU session response to the AMF, and the AMF sends a PDU session update/session management context request to the SMF. The SMF then sends a request message to the UPF for session modification. Upon retrieval of this message by the UPF, the SMF transmits the response of the PDU session update to the AMF, which concludes the PDU session establishment process [38]

3.1.3 The Proposed Architecture

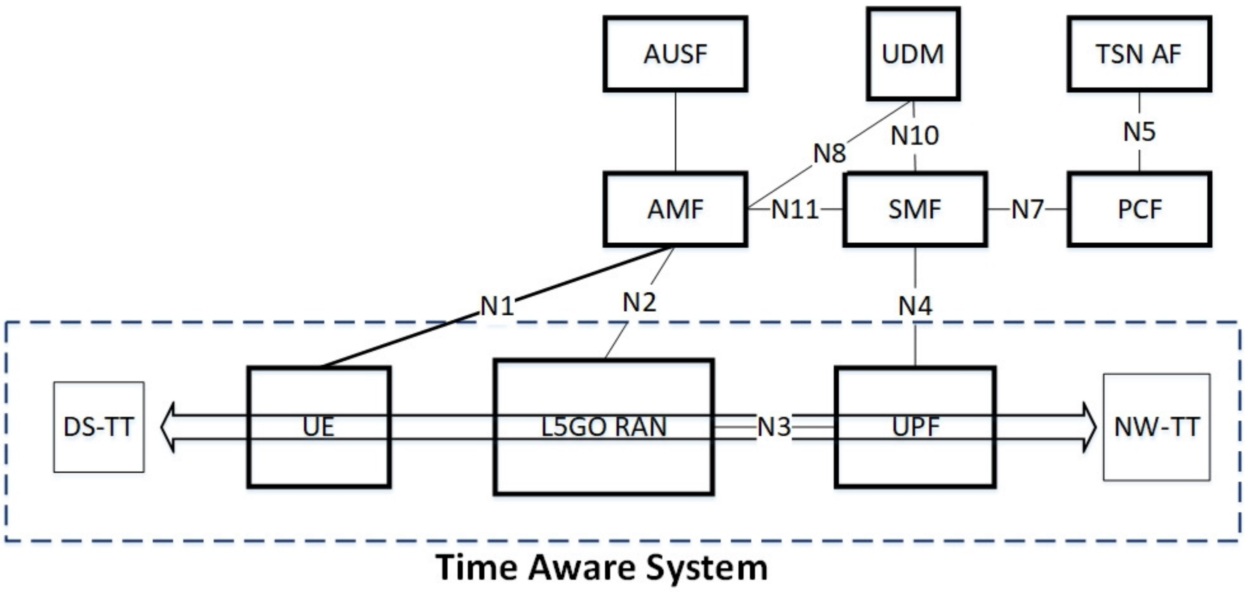


Figure 3.3: The proposed network architecture.

The proposed L5GO network architecture for delay critical telehealth applications is illustrated in Figure 3.3. It uses the Time Sensitive Network (TSN) logical bridge architecture defined by 3GPP [2] as the base, together with the IEEE 802.1AS time aware system. The IEEE 802.1 TSN enables transmission of various types of critical data over a bridged Ethernet network. Transmitted data can be of different QoS levels, regardless of the low latency end-to-end connectivity provided by the TSN [39].

The TSN bridge is between the Device-side TSN translator (DS-TT) and Network-side TSN translator (NW-TT) [39]. When operating the 5G core network elements, the RAN and the communication links are invisible to the TSN bridge because it works as an end-to-end tunnel. The entire system is synchronized by a GMC. The Time Sensitive Network Application Function (TSN AF) establishes the end-to-end tunnel connectivity and maintains it throughout the session, whereas the rest of the 5G network functions serve their original purpose.

A 5G network equipped with the TSN bridge features has the capability to deliver ultra-low latency services to the subscribers. As the use cases of interest are for surgeries of stationary organs, their E2E delay requirement is considered to be below 20 ms [35]. However, when it comes to cases with moving organs, the delay threshold decreases drastically.

3.2 The Network Deployment Models

The thesis considers three possible network deployment models;

- L5GO network
- Traditional MNO
- MEC enabled MNO network.

The image generating devices, the image processing server and the surgical devices are located inside the hospital premises in all three models. The study only focuses on the communication process associated with the data transfer, as it is the most critical stage with regards to ultra-low latency. This means, the delays associated with the device registration and the PDU session establishment are neglected. During the data transfer process, the generated data gets transferred to the image processing server via the access network and the core network function UPF. The image processing server augments the required information and sends the augmented information to the AR consoles, with the involvement of the UPF.

In the L5GO deployment model, a single L5GO provides connectivity to the hospital premises, through a multitude of gNBs with small coverage areas. The communication requirements of the surgical devices, the image processing server and the monitor consoles are catered through different gNBs, as illustrated in Figure 3.4. Most importantly, the L5GO core network is also located inside the hospital premises.

In the MNO deployment model, the MNO provides connectivity to all communicating devices in the hospital premises through its 5G network. Figure 3.5 depicts the MNO deployment model. The core network of the MNO resides outside the hospital facility. In a practical setting, there can be N such healthcare facilities located in a geographically distributed manner, and served by the same MNO.

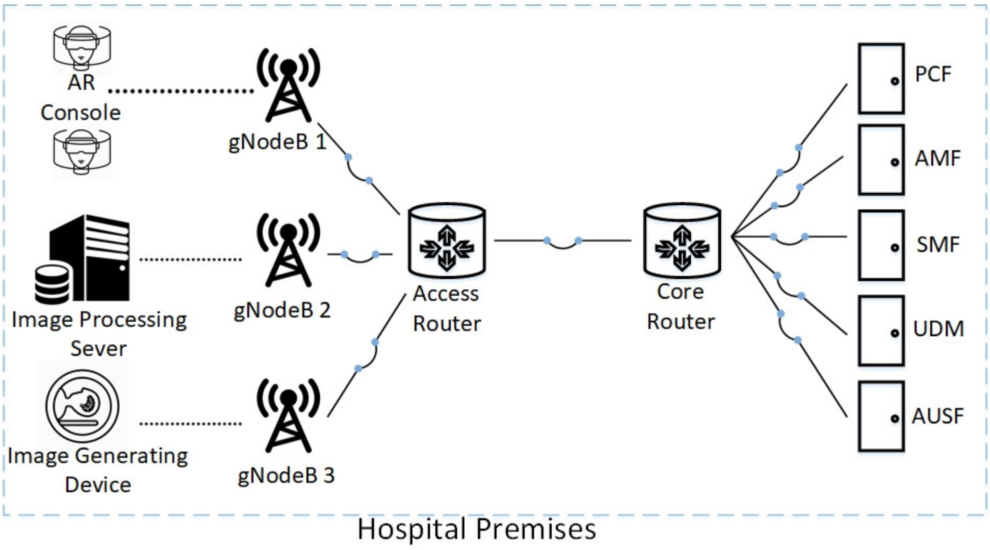


Figure 3.4: Deployment model for a L5GO serving a hospital premises.

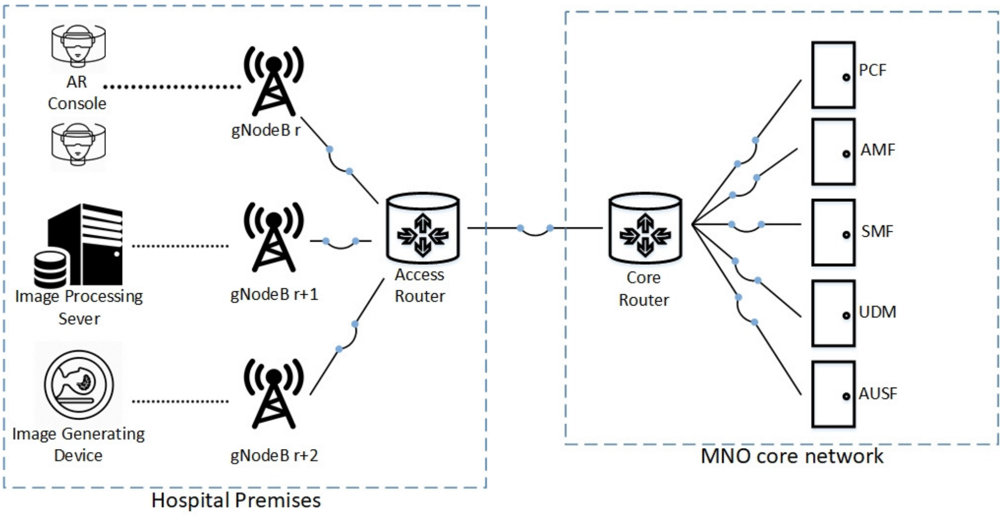


Figure 3.5: Deployment model for a MNO serving a hospital premises.

The third deployment model, which is the MEC network as illustrated in Figure 3.6, has a similar access network as the MNO model, but the core network functions are located at the edge, thus located closer to the hospital premises. The thesis assumes that M healthcare facilities with the similar setup are served by the MEC network, similar to the MNO model, and $M \leq N$.

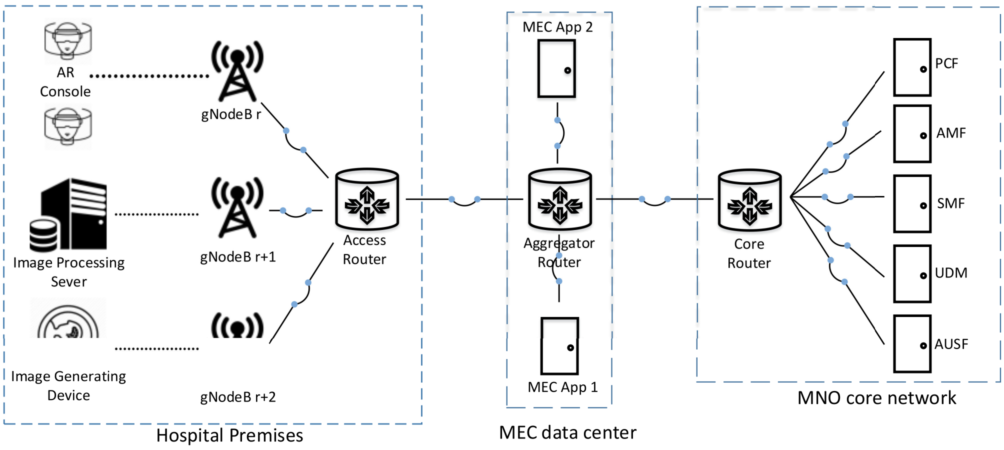


Figure 3.6: Deployment model for a MEC enabled network serving a hospital premises.

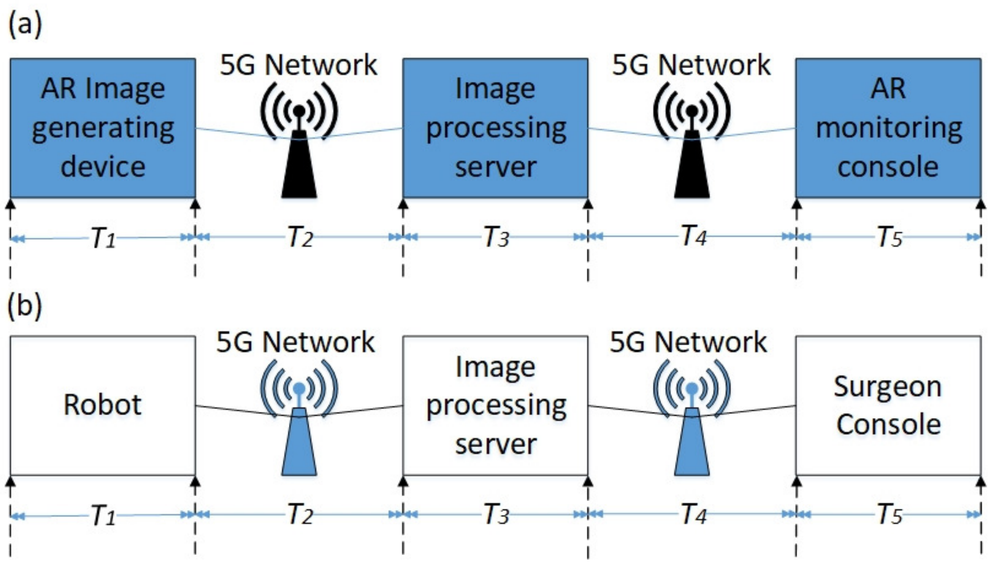


Figure 3.7: (a) Latency breakdown for ARAS use case (b) Latency breakdown for RAS use case.

The different latency components of the two use cases are illustrated in Figure 3.7. This thesis assumes Frequency Division Duplexing (FDD), and a 120 kHz, 7 s mini-slot bandwidth for the access network deployment. Moreover, the thesis assume a fiber backbone for the backhaul links to support high data rates and achieve ultra-low latency. Table 3.1 tabulates general calculation parameters considered for the two use cases.

Table 3.1: General calculation parameters.

Parameter	Value
Latency between the terminal devices and the AN T_{AN}	0.33 ms [40]
Latency between the AN and the core network T_{BH}	0.05 ms per km [41]
Mean service time of a gNB T_{NB}	0.1 ms
Mean service time of the access router T_{AR}	0.1 ms
Mean service time of the core router T_{CR}	0.05 ms
Mean service time of the 5G NF T_{NF}	0.01 ms
Maximum acceptable AR imaging system latency	14 ms [35]
Maximum acceptable Robotic Aided Surgery system latency	20 ms [35]
Distance to the L5GO core network	100 m [35]
Healthcare facilities served by the MNO (N)	10
Healthcare facilities served by the MEC (M)	3

3.3 Results analysis

3.3.1 The Impact of the Distance to the Core Network

The E2E latency of the system comprises of the image generation and processing delay, the operating time of the surgical devices and the latency of the communication infrastructure as illustrated in Fig. 3.7. The human reaction time is neglected. According to the latency breakdown in Fig. 3.7, the total E2E latency is given by

$$T_{lat} = \sum_{r=1}^{5} T_r, \tag{3.1}$$

where T_r denotes the latency of the r-th stage, and $r \in \{1, \ldots, 5\}$. For AR assisted surgery, the threshold value for T_{lat} is 14 ms [35], given that the surgery is performed on static organs, where the only moving object is the surgeon's hand. If the surgery is performed on a moving organ such as the heart, the threshold for T_{lat} should be lower than 14 ms to achieve sufficient precision. Further, it is known that $T_1 + T_3 + T_5 = 11$ ms for AR assisted surgery use cases [35]. The tolerable latency threshold for robotic aided surgery is slightly higher at 20 ms. This threshold also decreases when moving organs are considered. For robotic aided surgery, $T_1 + T_3 + T_5 = 16$ ms [35]. The latency in the 5G network can be calculated by

$$T_2 = T_4 = T_{AN} + T_{BH} + T_{NB} + T_{AR} + T_{CR} + T_{NF}, \tag{3.2}$$

where the symbols are defined in Table 3.1.

The E2E latency values for various network settings are presented in Fig. 3.8 and Fig.

3.9, for AR assisted surgery and robotic aided surgery, respectively. In particular, the figures show how the latency changes with the distance to the core network. For both use cases, three MEC core network distances (10 km, 25 km and 50 km) are considered. All these three settings will not satisfy the required average latency threshold for AR assisted surgery, whereas for robotic aided surgery, the setting with a distance of 10 km to the core network satisfies the latency constraint. We can observe that for the MNO network setting to work, the distance to the core network should be brought down to 100 m. Any distance greater than this violates the latency constraint. Finally, we can observe that the L5GO best suits the two use cases.

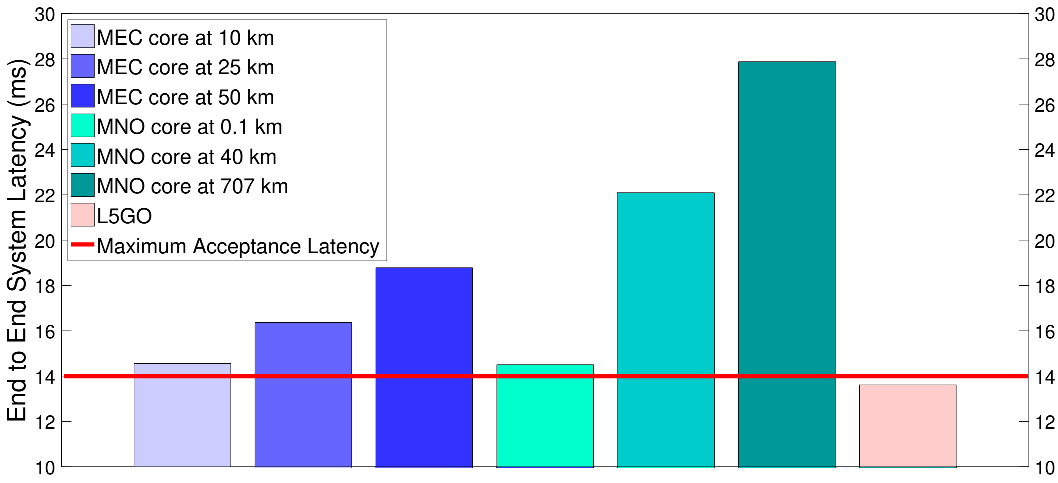


Figure 3.8: E2E latency of use case 2.4.1 with respect to distance to core network.

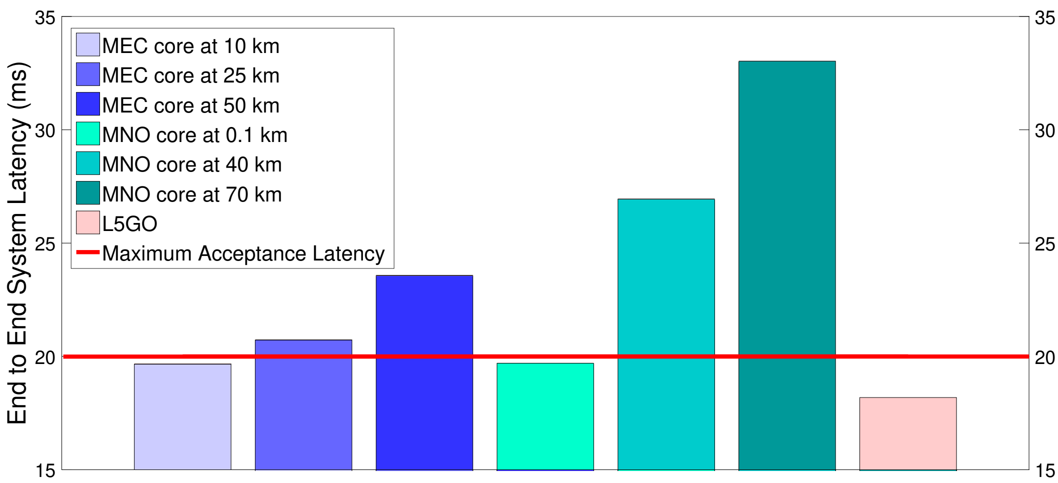


Figure 3.9: E2E latency of use case 2.4.2 with respect to distance to core network.

3.3.2 Impact of the Network Function Processing Delay

The Network Function (NF) processing delay T_{NF} is a crucial component in E2E latency. The NF processing delay mainly depends on the operator resources and the network traffic load. To this end, we have

$$T_{NF} \propto \frac{network\ load}{operator\ resources} \tag{3.3}$$

Therefore, the latency values related to the 5G network can be written as

$$T_2 \;=\; T_4 \;=\; T_{AN} + T_{BH} + T_{NB} + T_{AR} + T_{CR} + K.\frac{network\ load}{operator\ resources} \tag{3.4}$$

where $K = 1$ according to the proposed architecture in Section III.

An MNO typically possess higher computational and network resources compared to an L5GO, whereas an MEC possess more resources compared to an L5GO, but comparatively less resources compared to an MNO. Since the NF processing delay is inversely proportional to the operator resources, we consider four different resource levels for the MNO and the MEC networks as outlined in Table 3.2, where J is the resource multiplier. The total resources for an MNO or an MEC with a resource multiplier J, can be written as $J \times R_{L5GO}$, where R_{L5GO} denotes the total resources in the L5GO network. We check whether tolerable latency is achievable with the MNO and the MEC networks by increasing their resource levels compared to the L5GO.

Table 3.2: Increased resource levels at the MNO and the MEC networks compared to the L5GO.

Resource Level	Resource Multiplier for the MNO/MEC
L_1	$J = 1$
L_2	$J = 10$
L_3	$J = 100$
L_4	$J = 1000$

We consider values of 0.01 ms and 0.05 ms for the T_{NF} of the L5GO. Note that an MNO has $N = 10$ L5GOs in the network, and an MEC has $M = 3$ L5GOs in the network. Thus, it is not hard to see that the T_{NF} of an MNO can be calculated by multiplying the T_{NF} of the L5GO network by $\frac{N}{R}$, where R denotes the total resources in a given network. Similarly, the T_{NF} of an MEC can be calculated by multiplying the T_{NF} of the L5GO network by a factor $\frac{M}{R}$. Then, we calculate the maximum distance that the networks with the additional resources can have to their core networks, that is the length of the fiber backhaul. These values are tabulated in Table 3.3 for a T_{NF} of 0.01 ms at the L5GO, and in Table 3.4 for a

T_{NF} of 0.05 ms at the L5GO. The length of the backhaul is denoted by D_{BH}.

It is clear that the length of the backhaul, that is, where the core network can be deployed, varies based on the resource availability. It is possible for the MNO to support the required E2E latency, under all four resource levels. This is due to the underlying reason that the increased resource levels lead to a lower T_{NF} value at the MNO. However, it is interesting to consider the distance values to the core network although the delay constraints are satisfied. Obviously, the distance to the core network monotonically increases with the network resources at the MNO. However, it can be noted that the gradient of this increasing slope is rather small, thus the gain is not significant compared to the additional costs. As an example, considering the AR assisted surgery use case, for $J = 100$, D_{BH} is 2.2859 km, and for $J = 1000$ D_{BH} is 2.40985 km. The gain is approximately 120 m, which does not justify a 100-fold resource upgrade. It can also be seen that the backhaul delay T_{BH} is more prominent compared to the NF processing delay T_{NF} for the two use cases of interest.

Since the MEC only serves 3 health facilities simultaneously, it shows a higher gain in terms of D_{BH} compared to the MNO for similar resource levels. The MEC too satisfies the required latency thresholds when the resource levels are increased. For Robotic aided surgery, D_{BH} is 7.76625 km for $J = 10$, and D_{BH} is 7.8557 km for $J = 100$. Similar to the MNO, the change is insignificant when taking the cost of resource upgrade (increase of 90 m from 10-fold resource upgrade). At $T_{NF} = 0.01$ ms, the gain is about 170 m for AR assisted surgery, for the same 10-fold resource upgrade.

Table 3.3: Distance to the core network for ARAS and RAS use cases when L5GO $T_{NF} = 0.01$ ms.

Level	MNO			MEC		
	T_{NF} (ms)	ARAS D_{BH}(km)	RAS D_{BH}(km)	T_{NF} (ms)	ARAS D_{BH}(km)	RAS D_{BH}(km)
L_1	0.1	1.8024	7.1372	0.03	2.32045	7.76625
L_2	0.01	1.95065	7.1907	0.003	4.9829	7.8557
L_3	0.001	2.2859	7.3603	0.0003	2.7495	7.9232
L_4	0.0001	2.40985	7.57005	0.00003	2.91955	8.07075

Overall, our results show that increasing resources at the MNO and the MEC networks is not a financially feasible investment for delay critical telehealth applications. The results also further highlight the importance of the proposed L5GO architecture for delay critical telehealth applications.

Table 3.4: Distance to the core network for ARAS and RAS use cases when L5GO T_{NF} = 0.05 ms.

Level	**MNO**			**MEC**		
	T_{NF} (ms)	ARAS D_{BH}(km)	RAS D_{BH}(km)	T_{NF} (ms)	ARAS D_{BH}(km)	RAS D_{BH}(km)
L_1	0.5	0.67635	6.5351	0.15	1.47935	5.78545
L_2	0.05	1.7324	6.6696	0.015	1.99245	6.04125
L_3	0.005	1.806	7.29235	0.0015	2.126	7.6696
L_4	0.0005	2.0823	7.5262	0.00015	2.2269	8.26105

Chapter 4

Deployment Options of 5G Network Slicing for Smart Healthcare

This chapter explores the suitability of vertical and horizontal slicing for smart hospital environments that utilize telehealth applications such as ARAS, RAS and IoMT. Author tries to formulate two convex optimization problems for optimal resource allocation for these applications, considering the two network slicing methodologies.

4.1 System Model and Problem Formulation

The thesis consider a smart hospital environment that utilizes a 5G MNO network for connectivity. The 5G MNO facilitates numerous telehealth applications such as remote surgery, remote patient monitoring and management, and pharmacy management. An MNO can deliver guaranteed E2E SLA for such applications by utilizing network slicing. To this end, in modern 5G networks, the Next Generation - Radio Access Network (NG-RAN) to core network is softwarized. Thus the network resources such as bandwidth, storage, and processing can be partitioned into multiple slices based on user requirements. Network slicing is mainly twofold, vertical and horizontal. Figure 4.1 illustrates a vertical slicing approach for a smart hospital, where each individual telehealth application has been allocated a dedicated resource slice created in a 5G network. A network slice is created to deliver an application originated traffic across the network, with a guaranteed QoS for each telehealth application. Vertical slicing methods segregate general mobile traffic from a vertical application point of view and simplify the traditional QoS configuration in a complex network, enabling fast, and reliable service delivery [16].

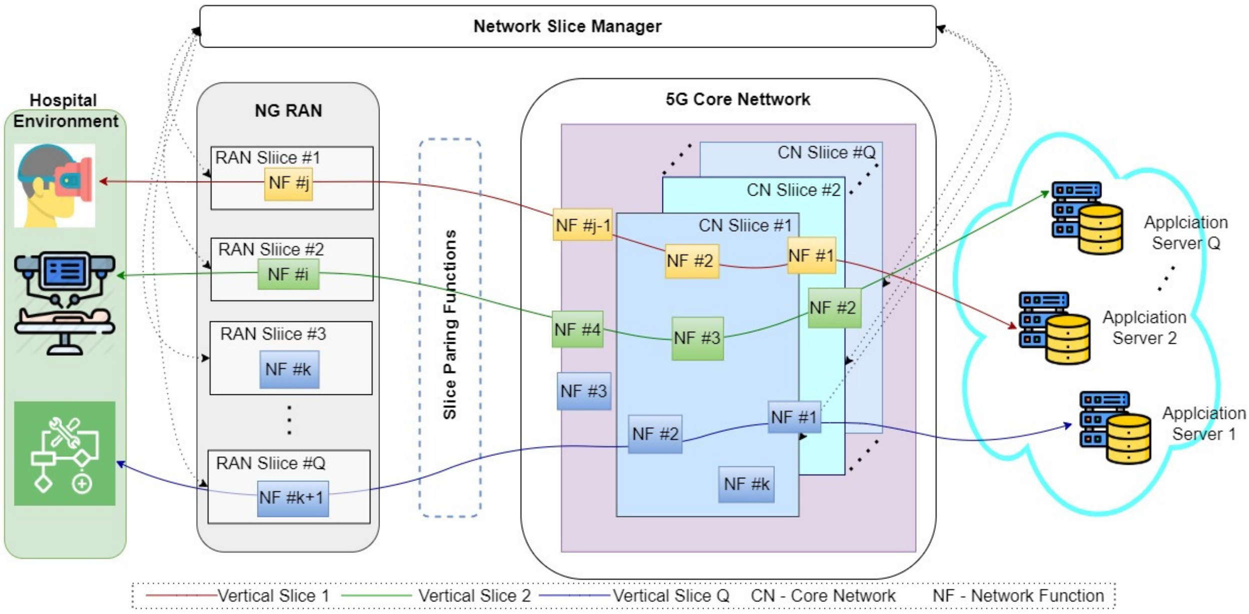


Figure 4.1: Vertical Slicing Architecture.

Horizontal slicing is a modern alternative approach in 5G network slicing that facilitates through computational offloading, capacity scaling and edge-cloud computing [16], [14]. Application generated traffic can be classified in terms of the services they utilize such as control, video and audio. Multiple applications can have common services associated with them. Fig. 4.2 [13] illustrates a horizontal slicing approach, where a dedicated resource slice is allocated for each service. Being different to vertical slicing, horizontal slices are connected to multiple end devices and application servers, and serve in a distributed manner in a 5G MNO. This thesis propose to map a service to each horizontal slice, and compare the network performance with vertical/application based slicing. The objective is to find out which slicing method serves better for a smart hospital environment under defined constraints. Formulation is subjected to network resource allocation prioritization, as different use cases require various QoS levels based on factors such as severity of the application, and the number of end users.

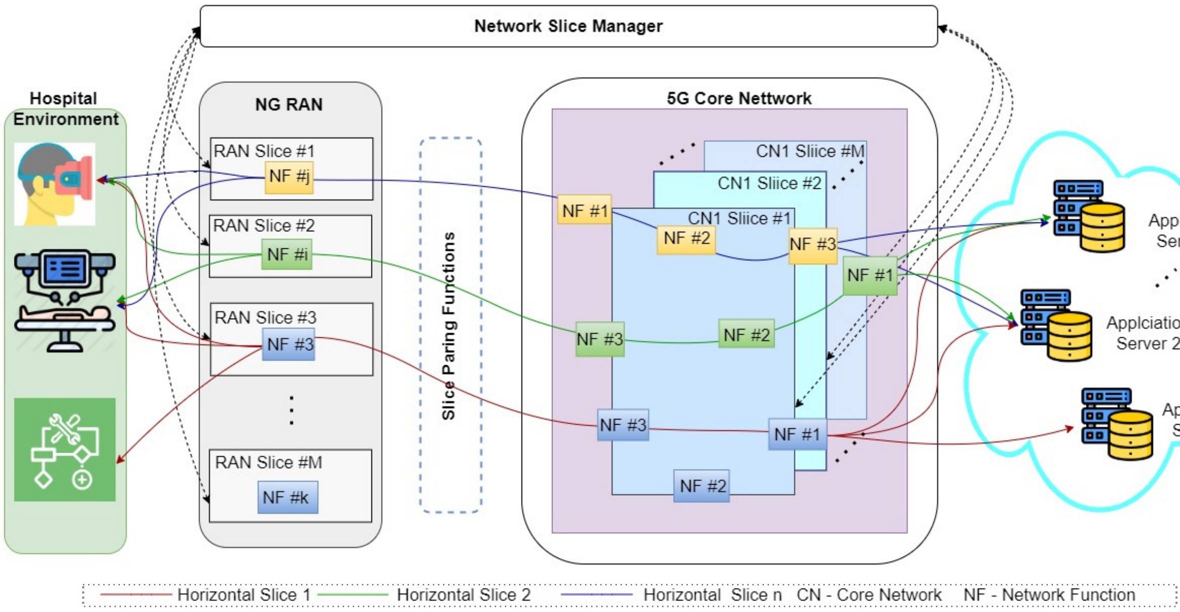


Figure 4.2: Horizontal Slicing Architecture.

The thesis consider a smart hospital with Q applications running simultaneously. The traffic from these Q applications can be segregated into M services. Each application utilizes a subset of these services. The event $E_{i,j}$ is true if the j-th service is used for the i-th application, where $j \in \mathcal{S} = \{1, \ldots, M\}$ and $i \in \mathcal{A} = \{1, \ldots, Q\}$. The network consists of T types of resources, $R_1, \ldots, R_T$, that need to sliced appropriately. For this purpose, the thesis introduce a cost based objective function that maximizes a weighted sum of the number of end users, given by $\sum_{i=1}^{Q} \beta_i U_i$, where U_i is the number of users utilizing the i-th telehealth application, and β_i is a cost associated with the i-th telehealth application, for $i \in \mathcal{A}$. Author assume that the number of users for the i-th application can vary between $U_i^{\min}$ and $U_i^{\max}$. Moreover, the costs can be set by the telehealth provider according to its revenue and service targets. Let $r_{i,j,k}$ denote the amount of the k-th resource allocated to a user utilizing the j-th service associated with the i-th application, where $i \in \mathcal{A}$,$j \in \mathcal{S}$, and $k \in \mathcal{E} = \{1, \ldots, T\}$.

Let $\alpha_{i,k}$ denote the fraction of the k-th resource allocated to the i-th application for $i \in \mathcal{A}$ and $k \in \mathcal{E}$. Then, the optimization problem in concern for vertical slicing can be written as follows.

$$
\begin{aligned}
\text{(P1):}\ \underset{\alpha_{i,k}}{\text{maximize}} \quad & \sum_{i=1}^{Q} \beta_i U_i && \text{(4.1a)}\\
\text{subject to} \quad & \\
& \sum_{j=1}^{M} r_{i,j,k} U_i . \mathbb{1}_{\{E_{i,j}\}} \leq \alpha_{i,k} R_k, \qquad \begin{matrix} i \in \mathcal{A}, \\ k \in \mathcal{E} \end{matrix}, && \text{(4.1b)}\\
& \sum_{i=1}^{Q} \alpha_{i,k} = 1, \quad k \in \mathcal{E}, && \text{(4.1c)}\\
& U_i^{\min} \leq U_i \leq U_i^{\max}, \quad i \in \mathcal{A}. && \text{(4.1d)}
\end{aligned}
$$

The indicator function is used to incorporate the event $E_{i,j}$ being true or false through (4.1b), where the indicator function outputs one if the event is true, and zero if the event is false. By solving this optimization problem, the MNO can obtain how the resources can be sliced vertically such that each vertical slice is allocated to a telehealth application.

Similarly, let $\gamma_{j,k}$ denote the fraction of the k-th resource allocated to the j-th service for $j \in \mathcal{S}$ and $k \in \mathcal{E}$. Then, the optimization problem in concern for horizontal slicing can be written as follows.

$$
\begin{aligned}
\text{(P2):}\ \underset{\gamma_{j,k}}{\text{maximize}} \quad & \sum_{i=1}^{Q} \beta_i U_i && \text{(4.2a)}\\
\text{subject to} \quad & \\
& \sum_{i=1}^{Q} r_{i,j,k} U_i . \mathbb{1}_{\{E_{i,j}\}} \leq \gamma_{j,k} R_k, \qquad \begin{matrix} j \in \mathcal{S}, \\ k \in \mathcal{E} \end{matrix}, && \text{(4.2b)}\\
& \sum_{j=1}^{M} \gamma_{j,k} = 1, \quad k \in \mathcal{E}, && \text{(4.2c)}\\
& U_i^{\min} \leq U_i \leq U_i^{\max}, \quad i \in \mathcal{A}. && \text{(4.2d)}
\end{aligned}
$$

By solving this optimization problem, the MNO can obtain how the resources can be sliced horizontally such that each horizontal slice is allocated to a service. The solutions of the two optimization problems are utilized to draw insights on the suitability of each slicing methodology considering various smart hospital environments.

4.2 Resources, Applications and Services in a Smart Hospital Environment

4.2.1 Resources

Resources required to transmit E2E traffic in a 5G network can be primarily categorized as:

1. Bandwidth
2. Processing time
3. Storage

according to our setup ($T = 3$). Telecommunication service providers allocate E2E network resources based on the SLA with the customer. Different applications and services utilize a subset of the aforementioned resources. Even though network slicing facilitates flexibility, there is always a maximum limit in terms of resource levels that can be allocated to a particular slice, which motivates the optimization problem of interest.

4.2.2 Applications

In this thesis, the author considers three potential applications ($Q = 3$) in a future smart hospital environment having 5G connectivity. The applications are as follows:

1. Augmented Reality Assisted Surgery (ARAS)
2. Robotic Aided Surgery (RAS)
3. Internet of Medical Things (IoMT)

Out of these applications, ARAS utilizes modern imaging technologies to examine internal body organs. The real-time images are augmented using image processing servers and transmitted to a console worn by a doctor. RAS is a similar but slightly advanced image guided surgical procedure which uses robotic arms to perform subtle surgeries [35], [17]. IoT applications made specifically for the medical industry or IoMT, which focus on remote health monitoring is also considered as the 3-rd telehealth application [36], [37]. These applications were further discussed in Section 2.4

The optimization problem (P1) in Section 4.1 studies mapping a dedicated vertical slice for each of these applications. Since both ARAS and RAS use ultra low latency 4K video streams to guide the remote end, the expected resource utilization for 5G MNO is high for these applications. On the other hand, resource requirement is considerably lower for IoMT when compared with the other two applications.

4.2.3 Services

The traffic generated from telehealth applications can be categorized according to the services they utilize. The applications discussed above produce 5G network traffic which can be primarily classified into three services ($M = 3$);

1. Audio
2. Video
3. Control

The optimization problem (P2) in Section 4.1 studies mapping a dedicated horizontal slice for each of these services. Video services are common for both ARAS and RAS. Audio service is only used for RAS, whereas control services are common for all three applications. It is evident that a horizontal slice for video may consume the majority of 5G network resources in a horizontal slicing scenario.

In summary, the association between applications and services for further calculations are as follows:

- ARAS $\Rightarrow$ Video + Control
- RAS $\Rightarrow$ Video + Audio + Control
- IoMT $\Rightarrow$ Control

4.3 Result Analysis

Requirements of the telehealth applications discussed in Section 4.2.2 vary based on numerous factors such as specialization of the hospital, emergency situations, public clinic dates, revenue targets and cost constraints. We consider five different smart hospital setups based on their specialization in terms of medical treatment as given below. We set the weights β_i, $i \in \{1, 2, 3\}$, to suit the requirements of each type of hospital.

- Elderly care hospital (W_1) - These hospitals require more patient monitoring systems than surgical facilities as elderly patients need to be continuously monitored.
- Emergency care hospital (W_2) - Likely to have more RAS and ARAS procedures for patients with accidental injuries.
- Infectious disease center (W_3) - These hospitals require to maintain distance between the healthcare professional and the patient, hence, it may require a higher number of RAS procedures.

- Teaching hospital (W_4) - Requires expert medical professionals to guide the students through surgical procedures, thus may lead to many ARAS and RAS procedures. Remote connectivity for consultant surgeons are also a possibility in such a setup. Also, as ARAS visualizes a detailed view of a patient's body, we can also expect a higher number of ARAS in a teaching hospital.
- Pediatric hospital (W_5) - Due to the accidental injuries may surgical procedures along with remote patient monitoring facilities.

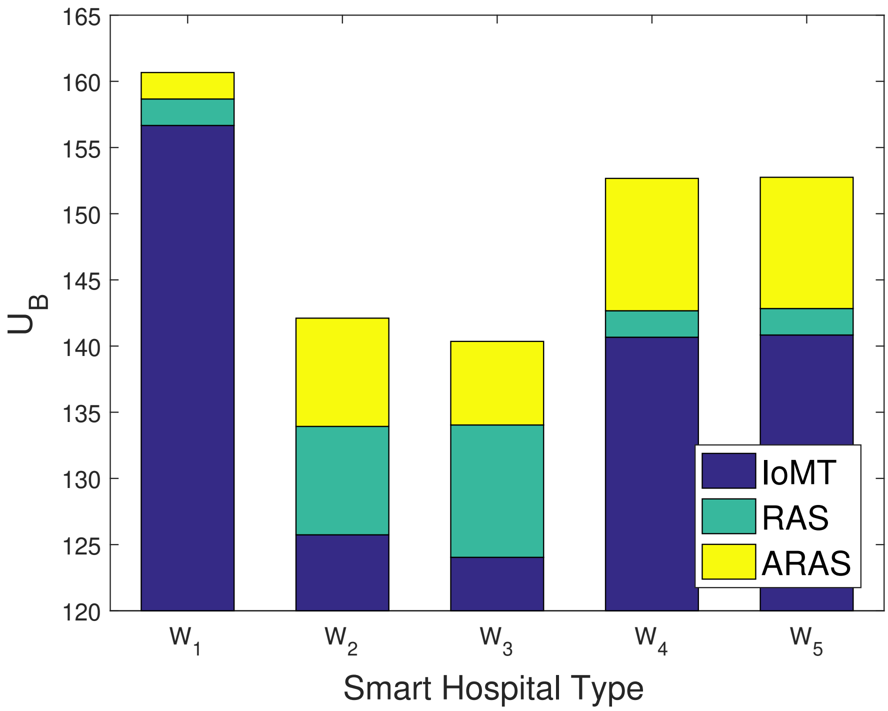
Table 4.1: Weight assignments and the respective values of the objective function.

Type	β_1	β_2	β_3	Vertical Slicing		Horizontal Slicing	
				U_A	U_B	U_A	U_B
W_1	1	1	1	161	161	482	482
W_2	2	3	1	166	142	488	482
W_3	2	7	1	206	140	496	459
W_4	4	2	1	184	152	489	466
W_5	2	2	1	164	152	486	482

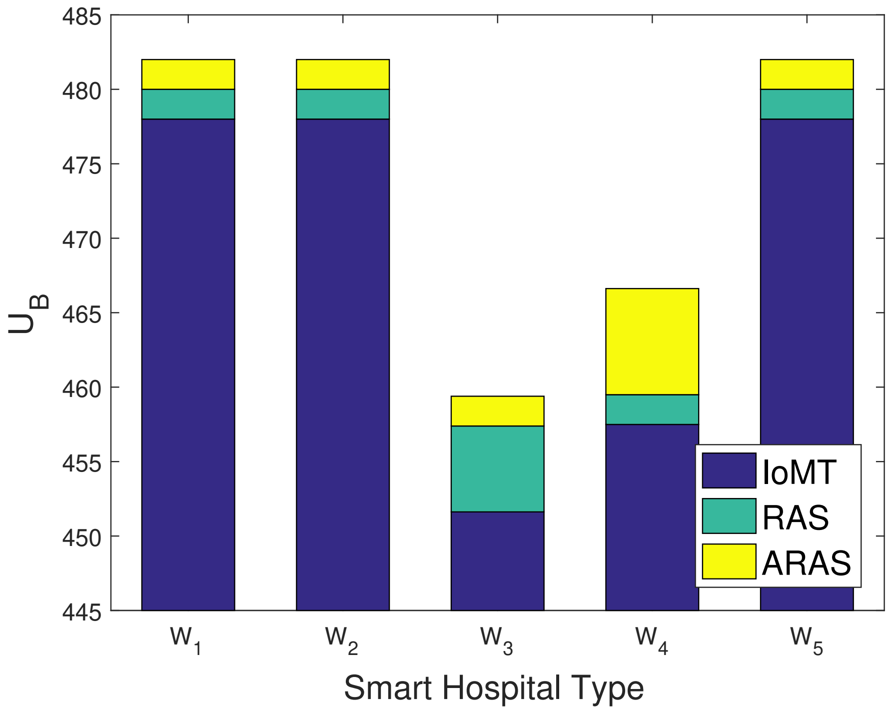
Table 4.2: Calculation parameters.

	Bandwidth (Mbps)	Processing (ms)	Storage (Mb)
Total	200000	1000	32000 [42]
For a 4K video transmission	12000 [35]	6 [35]	96
For a HD audio transmission	300 [35]	6	2.4
For control data stream transmission	2 [35]	6	0.2

It should be noted that hospitals have limited operating theatres, which leads to an upper limit for the number of concurrent surgeries. We set $U_1^{\min} = U_2^{\min} = 2$ and $U_1^{\max} = U_2^{\max} = 10$ in our results, and the rest of the calculation parameters are tabulated in Table 4.2. The IoMT device count in a smart hospital environment can reach up to large numbers compared to the other two telehealth applications, thus we omit an upper limit for that application. With these parameters, we solve the two optimization problems of interest. It is not hard to see that both optimization problems are convex, and they can be easily solved using techniques of linear programming.

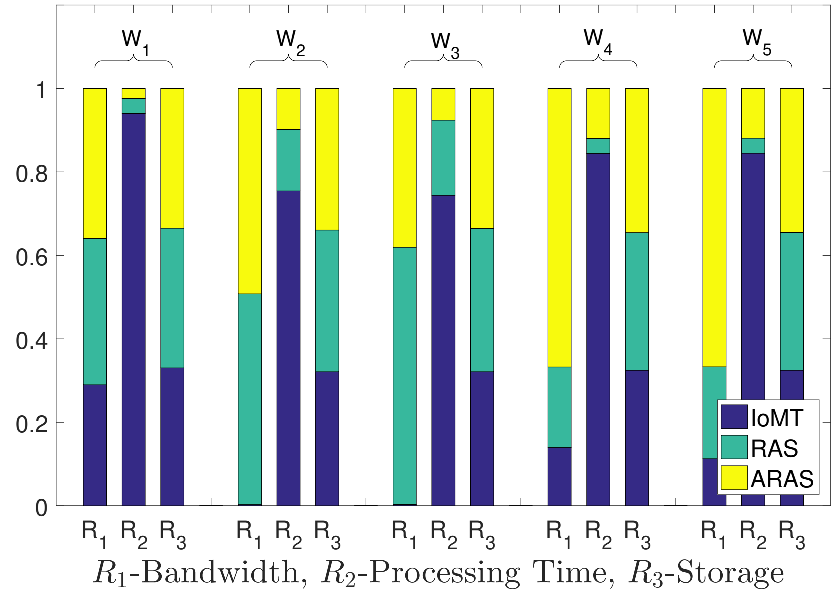


(a) Vertical slicing

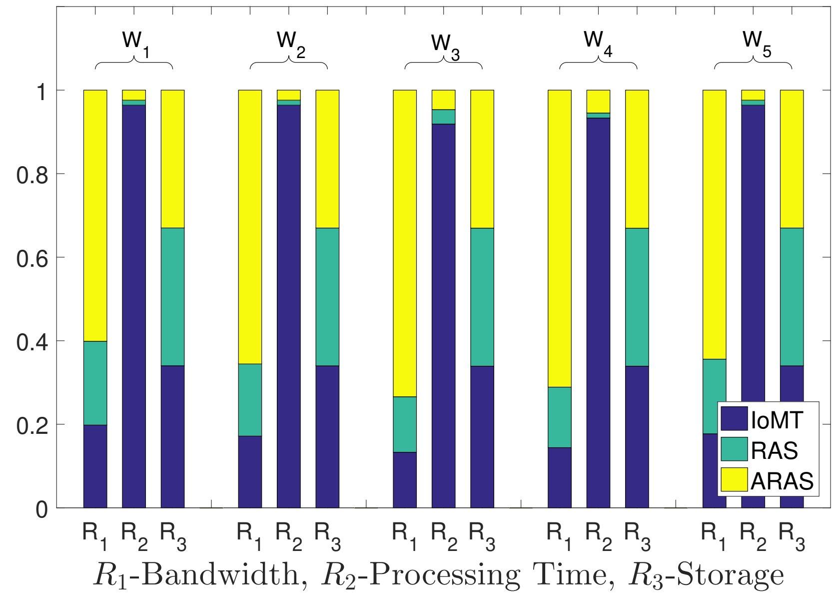


(b) Horizontal slicing

Figure 4.3: Number of users served for different weight types.



(a) Vertical slicing



(b) Horizontal slicing

Figure 4.4: 5G network resource allocation percentages for different weight types.

Figures 4.3a and 4.3b illustrate the number of users utilizing each application under the two slicing schemes, and Table 4.1, in which $U_A = \sum_{i=1}^{3} \beta_i U_i$ (weighted sum of users) and $U_B = \sum_{i=1}^{3} U_i$ (sum of users), tabulates the outputs of the objective function under optimal resource allocation. It is not hard to see that horizontal slicing facilitates more users for the three telehealth applications of interest. Since IoMT requires the lowest amount of resources when compared with the other two applications, the proportion of IoMT users within U_B is obviously high. Thus, comparing U_A values can be considered more meaningful practically.

For an elderly care (W_1) hospital, at optimal resource allocation, the number of users utilizing ARAS and RAS is $U_1^{\min}$ and $U_2^{\min}$ (minimum feasible values), respectively, whereas the number of users utilizing IoMT dominates the sum of users. Horizontal slicing seems preferable with an approximately threefold gain compared to vertical slicing. For emergency care (W_2) hospital, the priority is maximizing the RAS and ARAS user counts. It can be seen that the horizontal slicing fails to fulfil this requirement, with both U_1 and U_2 taking its minimum feasible values, similar to type W_1. However, we can observe that vertical slicing serves the purpose of type W_2 hospitals more appropriately.

For an infectious disease center (W_3), the goal is to prioritize the number of RAS users. It can be seen that the vertical slicing approach serves this purpose better with U_2 reaching the upper bound of 10. The vertical slicing approach serves ARAS favorably as well. For teaching hospitals, that is W_4 where the prioritization is for ARAS, U_1 reaches its maximum value with vertical slicing and performs better than horizontal slicing. However, U_2 is at its minimum with both slicing methods. It can be seen that the vertical slicing serves better for pediatric hospitals (W_5) as well.

Fig. 4.4 illustrates the bandwidth, processing time and storage resource dividend between ARAS, RAS and IoMT applications. It is important to scrutinize the numerical results to get an understanding of the 5G resource allocation in each slicing method. Horizontal slicing allocates a main portion of the bandwidth to ARAS for elderly care hospitals whereas vertical slicing distributes bandwidth resources approximately equally. Storage allocation is done almost in a uniform manner among the three applications, and the main portion of the processing resources are allocated for IoMT, under both slicing methods.

Vertical slicing allocates almost half of the bandwidth resources to RAS applications in emergency care (W_2) hospitals whereas ARAS consumes the main portion of the bandwidth in horizontal slicing. Since the priority is towards RAS in an infectious disease center, we can observe that more bandwidth is allocated to RAS with Vertical slicing, but this cannot be observed with horizontal slicing. In teaching hospital (W_4) and pediatric hospital (W_5) scenarios, ARAS consumes bulk of the bandwidth resources in both slicing methods. Even though IoMT uses more than 90% of processing time in W_3, W_4 and W_5 types hospitals with horizontal slicing, we can observe that vertical slicing uses lower processing resources. However, storage resource dividend among the telehealth applications happens almost equally for both slicing approaches.

Chapter 5

Discussion and Future Work

5.1 L5GO for Smart Hospitals

This study proposed a L5GO network architecture specifically designed for delay critical teleheath applications, which can be implemented on a smart hospital environment. The architecture is modeled base on two image guided surgeries, namely ARAS and RAS. Even though L5GO concept has been there for a while, not been much work carried out to analyze possible architecture options for L5GO networks. The thesis points out detailed message flow between L5GO core network and access network according to 3GPP standards. Identifying a L5GO as a 3GPP network reduces the implementation complexities of such network with a 5G MNO.

Fast growth of 5G networks empower new business opportunities with its URLLC, mMTC and eMBB features. Smart hospitals require ultra reliable and high bandwidth wireless network connections with low latency. L5GO is a novel concept with flexibility to provide content specific location oriented services. A smart hospital implemented with a L5GO network can bring numerous value additions to the patients, medical staff as well as hospital management. Specially in a pandemic situation like COVID-19, remote operated telehealth applications brings numerous advantages. It will enable the social distancing between the patient and doctor, which generates secure environment for the medical staff. Patients will have an advantage with the remote health monitoring capabilities as well as the with the accurate and precise remote surgical procedures. Remote monitoring will reduce the total operational cost of a hospital leading to a profitable business.

5.2 Vertical and Horizontal Network Slicing

Network slicing enabled through SDN and NFV technologies can optimize the E2E resource utilization of a 5G mobile network. Vertical network slicing, a well known network slicing methodology generates network slices specifically for different verticals. Horizontal slicing is novel slicing approach compared to vertical slicing, which focuses on traffic processing and resource slicing at the edge of the network.

The thesis proposes a network slice comparison model for horizontal and vertical slicing, using two convex optimization problems with an objective function of weighted sum of end user count. The thesis presents a general model as it can be applied to Q number of telehealth applications having M services, which use T number of resources. Therefore this model can be applied to any scenario of MNO's choosing and check the advantages and disadvantages of going with vertical slicing or horizontal slicing method. MNO can come to to an early conclusion of the network slicing method base on the results, it will help to reduce capital and operational cost in network deployments. Then the author solved the optimization problem using three use cases which are more likely to be implemented in a smart hospital, videlicet ARAS, RAS and IoMT. The study considers bandwidth, processing time and memory as the network resources, and audio, video and control as the services to solve the optimization problems. Generated results points out unique advantages of both slicing methods in terms of providing telehealth services over a 5G network.

5.3 Future Work

The two key outcomes of this research work are the proposed L5GO network architecture and the network slicing comparison model. In both cases several key assumptions were made, therefore generating a gap between the theoretical estimates that has been articulated throughout this research and practical applications of it. To this end, we need to further investigate applications of the proposed scheme and identify the shortfalls and possible solutions. Furthermore the research work has created a platform to explore new aspects of telehealth in 5G and beyond 5G networks. L5GO networks need to be further explored in terms of compatibility and harmony with MNO networks, and how networks beyond 5G can grasp the methodology to provide better customer centric service. Network slicing is a broad topic with numerous unexplored areas, there is a gap in a general methodology to compare all the network slicing methods which have been introduced. Teleheath is promised to become a multi billion dollar industry with the introduction of 5G. Security and reliability should be guaranteed in these critical medical applications. To that end there are new research work that need be carried out in the future.

Chapter 6

Conclusion

This thesis has explored telehealth deployments in 5G and beyond 5G network in two main aspects. It has proposed a Local 5G Operator (L5GO) architecture for delay critical telehealth applications, considering two delay critical healthcare use cases, *i.e.*, Augmented Reality Assisted Surgery (ARAS) and Robotic Aided Surgery (RAS). The proposed architecture has been designed based on the 5G Time Sensitive Network (TSN) architecture. The paper has highlighted the performance gains from the proposed L5GO network compared to the conventional Mobile Network Operator (MNO) based networks, and the Multi-access Edge Computing (MEC) based MNO networks, through an End-to-End (E2E) latency study. Numerical results have shown that the proposed L5GO architecture satisfies the stringent latency constraints of both use cases. On the other hand, the MNO or the MEC based network architectures fail to satisfy these constraints. It has also been shown that if it is required to use the MNO or the MEC networks for such delay critical use cases, the distance to the core network should be considerably decreased, and the total network resource levels should be considerably increased. The upgrade deems infeasible financially, which further highlights the utility of the proposed L5GO architecture for delay critical telehealth applications.

On the other hand, the thesis has introduced the formulation of two convex optimization problems for optimal resource allocation, and compares vertical and horizontal slicing methods, in a 5G MNO. An application has been provided for different types of smart hospital environments, considering three telehealth applications, which are, ARAS, RAS and IoMT The traffic generated from the telehealth applications have been considered to be comprised of three services, namely video, audio and control. In the comparison of the performance of network slicing, vertical slicing has been done by allocating a separate slice for each application, and horizontal slicing by allocating a network slice for each service. 5G network resources that include bandwidth, processing, and storage, have been allocated optimally for each slicing scenario. Numerical results have shown that horizontal slicing can support more telehealth application users than vertical slicing. However, the total end user count is dominated by IoMT users in both the slicing methods due to its low resource consumption. Both network slicing approaches have exhibited their unique advantages over one another depending on the type of the smart hospital environment. Bandwidth resource utilization has been dominated by ARAS and RAS, while IoMT has consumed the majority of processing. On the other hand, the storage utilization has been done approximately equally by the three telehealth applications.

References

[1] NGMN, "NGMN 5G White Paper," NGMN, Tech. Rep., Feb. 2015.

[2] 3GPP, "3GPP TS 23.501 ;System architecture for the 5G system (5GS);Stage 2 (Release 16)," Dec. 2019, technical specification. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3546

[3] ETSI, "Network Functions Virtualisation (NFV);Architectural Framework V1.1.1," ETSI, Tech. Rep., Oct. 2013.

[4] F. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key Technology Enablers for 5G Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2468–2478, 2017.

[5] ETSI, "Mobile Edge Computing (MEC); Technical Requirements, v1.1.1," Mar. 2016.

[6] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2429–2453, Mar. 2018.

[7] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[8] D. Soldani, "Fighting COVID-19 with 5G Enabled Technologies," White paper,Huawei Technologies, Tech. Rep., Apr. 2020.

[9] Y. Siriwardhana, P. Porambage, M. Liyanage, J. Walia, M. Matinmikko-Blue, and M. Ylianttila, "Micro-Operator Driven Local 5G Network Architecture for Industrial Internet," in *IEEE Wireless Communications and Networking Conference*, 2019, pp. 1–8.

[10] S. Kekki *et al.*, "MEC in 5G Networks," White paper No.28,ETSI, Tech. Rep., Jan. 2018, first edition.

[11] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[12] M. Matinmikko, M. Latva-aho, P. Ahokangas, S. Yrjola, and T. Koivumaki, "Micro Operators to Boost Local Service Delivery in 5G," in *Wireless Personal Communications*, vol. 95, 2017, pp. 69–82.

[13] S. Wijethilaka and M. Liyanage, "Survey on Network Slicing for Internet of Things Realization in 5G Networks," *IEEE Communications Surveys Tutorials*, pp. 1–1, Mar. 2021.

[14] M. A. Habibi, B. Han, and H. D. Schotten, "Network Slicing in 5G Mobile Communication Architecture, Profit Modeling, and Challenges," 2017, arXiv:1707.00852v1.

[15] S. Staff, "What is Dynamic Network Slicing?" Dec. 2017. [Online]. Available: https://www:sdxcentral:com/5g/definitions/dynamic-network-slicing/

[16] Q. Li, G. Wu, A. Papathanassiou, and U. Mukherjee, "An End-to-End Network Slicing Framework for 5G Wireless Communication Systems ," 2016, arXiv preprint arXiv:1608.00572.

[17] R. D. Silva, Y. Siriwardhana, T. Samarasinghe, M. Ylianttila, and M. Liyanage, " Local 5G Operator Architecture for Delay Critical Telehealth Applications," in *proc. IEEE 5G World Forum*, Sep. 2020.

[18] A. Osseiran *et al.*, "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014.

[19] Z. Pi and F. Khan, "An Introduction to Millimeter-wave Mobile Broadband Systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, 2011.

[20] L. Lu, G. Li, A. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, 2014.

[21] I. Hwang, B. Song, and S. Soliman, "A Holistic View on Hyper-dense Heterogeneous and Small Cell Networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, 2013.

[22] J. Matias, J. Garay, N. Toledo, J. Unzilla, and E. Jacob, "Toward an SDN-enabled NFV Architecture," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 187–193, 2015.

[23] G. C. Valastro, D. Panno, and S. Riolo, "A sdn/nfv based c-ran architecture for 5g mobile networks," in *2018 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, 2018, pp. 1–8.

[24] N. McKeown *et al.*, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, Mar. 2008.

[25] D. Kreutz, F. Ramos, P. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[26] P. Hedman *et al.*, "Description of Network Slicing Concept," NGMN Alliance, Tech. Rep., Jan. 2016.

[27] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource Slicing in Virtual Wireless Networks: A Survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, 2016.

[28] M. Matinmikko-Blue and M. Latva-aho, "Micro Operators Accelerating 5G Deployment," in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, Dec. 2017, pp. 1–5.

[29] A. Prasad, Z. Li, S. Holtmanns, and M. A. Uusitalo, "5G Micro-Operator Networks — A Key Enabler for New Verticals and Markets," in *2017 25th Telecommunication Forum (TELFOR)*, Nov. 2017, pp. 1–4.

[30] P. Ahokangas, M. Matinmikko-Blue, S. Yrjölä, V. Seppänen, H. Hämmäinen, R. Jurva, and M. Latva-aho, "Business Models for Local 5G Micro Operators," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 730–740, Mar. 2019.

[31] J. S. Walia, H. Hämmäinen, and M. Matinmikko, "5G Micro-Operators for the Future Campus: A Techno-Economic Study," in *2017 Internet of Things Business Models, Users, and Networks*, Nov. 2017, pp. 1–8.

[32] M. Matinmikko-Blue, S. Yrjölä, V. Seppänen, P. Ahokangas, H. Hämmäinen, and M. Latva-aho, "Analysis of Spectrum Valuation Approaches: The Viewpoint of Local 5G Networks in Shared Spectrum Bands," in *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Oct. 2018, pp. 1–9.

[33] Y. Siriwardhana, P. Porambage, M. Ylianttila, and M. Liyanage, "Performance Analysis of Local 5G Operator Architectures for Industrial Internet," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 559–11 575, Dec. 2020.

[34] B. Felter, "5 Exciting 5G Use Cases and Their Future Network Impact," Jul. 2020. [Online]. Available: https://www.vxchnge.com/blog/exciting-5g-use-cases

[35] 3GPP, "3GPP TR 22.826 : Study on Communication Services for Critical Medical Applications (Release 17)," Aug. 2019, technical Report. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3546

[36] S. Pradhan, "The Present and Future of Ingestible Sensors - The New Taste of Science," Jan. 2019. [Online]. Available: https://www.prescouter.com/2019/01/ingestible-sensors-innovations/

[37] S. Vishnu, S. R. J. Ramson, and R. Jegan, "Internet of Medical Things (IoMT) - An overview," in *proc. International Conference on Devices, Circuits and Systems*, Apr. 2020.

[38] 3GPP, "3GPP TS 23.502 ; System Aspects;System Architecture for the 5G System (5GS);Stage 2 (Release 16)," Dec. 2019, technical specification. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3546

[39] J. Farkas, L. Bello, and C. Gunther, "Time-Sensitive Networking Standards," *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 20–21, 2018.

[40] J. Sachs, G. Wikstrom, T. Dudd, R. Baldemair, and K. Kittichokecha, "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[41] M. Jaber, M. Imran, R. Tafazoll, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[42] Cisco, "Cisco ASR 1000 Series Aggregation Services Routers Data Sheet," Mar. 2019. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/routers/asr-1000-series-aggregation-services-routers/datasheet-c78-731632.html