UNIVERSITY
OF OULU

FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING
DEGREE PROGRAMME IN WIRELESS COMMUNICATIONS ENGINEERING

MASTER’S THESIS

Micro-Operator driven Local 5G Network Architecture for Industrial Internet Applications

Author	Yushan Siriwardhana
Supervisor	Mika Ylianttila
Second Examiner	Madhusanka Liyanage
Technical Supervisor	Pawani Porambage

March 2019

Siriwardhana Y. (2019) Micro-Operator driven Local 5G Network Architecture for Industrial Internet Applications. University of Oulu, Faculty of Information Technology and Electrical Engineering, Degree Programme in Wireless Communications Engineering. Master's Thesis, 70 p.

ABSTRACT

High degree of flexibility, customization and the rapid deployment methods are needed in future communication systems required by different vertical sectors. These requirements will be beyond the traditional mobile network operators' offerings. The novel concept called micro-operator enables a versatile set of stakeholders to operate local 5G networks within spatially confined environment with a guaranteed quality and reliability to complement mobile network operators' offerings. To enable the case specific requirements of different stakeholders, micro-operator architecture should be tailored to cater such requirements, so that the service is optimized. The novel micro-operator architecture proposed in this thesis using 5G access and core network functions, serves the communication needs of an Industry 4.0 environment having three use cases namely augmented reality, massive wireless sensor networks and mobile robots. Conceptual design of the proposed architecture is realized using simulation results for latency measurements, relating it with the results of a mobile network operator-based deployment. Latency analysis is carried out with respect to the core network distance and the processing delay of core network functions. Results demonstrate the advantages of the micro-operator deployment compared with mobile network operator deployment to cater specialized user requirements, thereby concluding that the micro-operator deployment is more beneficial.

Key words: Industry 4.0, Augmented Reality, network slicing, URLLC, mMTC

TABLE OF CONTENTS

FOREWORD

This master's thesis has been produced at the Centre for Wireless Communications, University of Oulu for the Degree Programme in Wireless Communications Engineering. The main objective of the thesis is to define the network architecture for a micro-operator providing case specific and localized communication services. The thesis discusses the architecture for a future industrial environment comprising of use cases augmented reality, massive wireless sensor networks and mobile robots. The architecture is defined in terms of network functions available in 3GPP 5G systems architecture.

This thesis work is supported by Business Finland in uO5G project, Academy of Finland in 6Genesis Flagship (grant no. 318927) and SecureConnect projects.

I express my sincere gratitude to my supervisor Prof. Mika Ylianttila for his support and guidance throughout the period of master's thesis. Also, I express my gratitude to Dr. Madhusanka Liyanage, for the immense support provided me as the second examiner of this thesis, and as a mentor throughout my master's studies. I extend my gratitude to Dr. Pawani Porambage, technical supervisor of this thesis, for providing valuable ideas and guiding me to realize the thesis objectives successfully. My sincere appreciation also goes to Dr. Marja Matinmikko, project manager of uO5G project, for providing the financial assistance and the research guidance. My appreciation also goes to Walia Jaspreet from Aalto University for his valuable ideas which contributed to the success of this thesis. I also thank my colleagues of Networks and Systems research team for creating a collaborative and research friendly environment. My thank also goes to Centre for Wireless Communications for providing me with the opportunity to conduct the research related to my thesis.

I also want to thank Archana, my loving wife, for all her contributions which helped me to realize my objectives. Finally, I would like to thank my parents for their encouragements which inspired me to achieve the goals of my life.

Oulu, 25 March 2019

Yushan Siriwardhana

LIST OF ABBREVIATIONS AND SYMBOLS

3GPP	3rd Generation Partnership Project
5GTN	5G Test Network
AF	Application Function
AGV	Automated Guided Vehicles
AMF	Access and Mobility Management Function
AN	Access Network
API	Application Programming Interface
AR	Augmented Reality
AUSF	Authentication Server Function
CN	Core Network
CP	Control Plane
CSMF	Communication Service Management Function
DHCP	Dynamic Host Configuration Protocol
DN	Data Network
E2E	End-to-End
eMBB	Enhanced Mobile Broadband
EPC	Evolved Packet Core
gNB	Next Generation NodeB
HD	High Definition
ICMP	Internet Control Message Protocol
IEEE	Institute of Electrical and Electronics Engineers
IPV4	Internet Protocol Version 4
IWN	Industrial Wireless Network
LTE	Long Term Evolution
M2M	Machine-to-Machine
MEC	Multi-Access Edge Computing
MIMO	Multiple Input Multiple Output
mMTC	Massive Machine Type Communication
MNO	Mobile Network Operator
NEF	Network Exposure Function
NF	Network Function
NFV	Network Function Virtualization
NG AN	Next Generation Access Network
NGMN	Next Generation Mobile Networks
NRF	Network Repository Function
NSI	Network Slice Instance
NSMF	Network Slice Management Function
NSSF	Network Slice Selection Function
NSSI	Network Slice Subnet Instance
NSSMF	Network Slice Subnet Management Function
OSPF	Open Shortest Path First
OTT	Over the TOP
PCF	Policy Control Function
PDU	Protocol Data Unit
QoS	Quality of Service
RAN	Radio Access Network

RAT	Radio Access Technology
SDN	Software Defined Networking
SIM	Subscriber Identity Module
SMF	Session Management Function
TCP	Transmission Control Protocol
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management
UDN	Ultra-Dense Network
UDP	User Datagram Protocol
UE	User Equipment
uO	Micro-Operator
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable and Low Latency Communication
WCNC	Wireless Communications and Networking Conference
VNF	Virtual Network Function

$D_{backhaul}$	Distance from factory to MNO core network
L_{dat}	Latency of AR data transfer process
L_{dat_r}	Latency of mobile robots data transfer process
L_{reg}	Latency of AR device registration process
L_{sensor}	Latency of sensor network communication process
N	Number of factories simultaneously served by MNO
T_{access}	Delay in access network
T_{alarm}	Processing time of alarm management server
$T_{backhaul}$	Delay in backhaul network
$T_{control}$	Processing time of controlling server
T_{NF}	Network function processing delay
T_{server}	Processing time of AR server

1 INTRODUCTION

The landscape of mobile communication services is rapidly changing with the proliferation of digitization technologies. Future Mobile communication systems will need to address the communication requirements such as extremely high data rates, extremely low latency communication, higher density of devices and the communications among massive number of devices. With these changes of the mobile communication ecosystem, more emphasis needs to be placed on location specific services in different vertical sectors such as industry, media, health, education and energy [1]. Factories, smart cities, shopping malls, hospitals and universities are identified as some of the most common locations, which would require location specific and case specific communication requirements [1]. These location specific services put greater demands on future mobile communication systems. Therefore, the advancements of the future mobile communication systems should be able to cater those requirements within a short period of time. These communication systems should also have higher reliability, enhanced privacy and security.

1.1 Background and Motivation

5G is the successor of the present fourth generation mobile communication systems and it is not merely an advancement of its predecessor, but it will be highly integrative with number of systems to provide user with a seamless experience [2]. Other than providing the traditional mobile broadband services to the generic customers, future 5G wireless systems focuses on addressing specific communication requirements of different verticals [3]. Each vertical sector and the locations belong to such vertical sectors are having different communication requirements. Moreover, the need for fast deployment of such services will become essential. Therefore, case specific and location specific requirements are expanding to a level such that the current capabilities of the traditional Mobile Network Operators (MNO) are not adequate for future service offerings. This occurs due to multiple factors. Generally, the main focus of MNOs is to provide services to masses. They usually have long investment cycles with high infrastructure costs and they have to utilize their available limited spectrum very effectively since the spectrum is scarce [4]. With the limited spectrum available to MNOs it is difficult to cater each and every location specific need and doing it will be highly inefficient in terms of operations. Managing small cell networks in number of locations will be an additional overhead for the big MNO operation. Therefore, a new player having the capability of establishing local 5G networks to cater those specific requirements will emerge in the future. This leads to the mobile communication market to be opened for local 5G network operators such as recently proposed in the micro-operator (uO) concept [5].

The focus of uO is to cater case specific and location specific communication requirements oppose to the wide area coverage provided by traditional MNOs [6]. The services provided by uO are tailored to the requirements of each use case to assure an optimal communication service. To support the uO objective of optimal delivery of case specific and location specific requirements, the system architecture of a uO should also be tailored based on the requirements. Considering the nature of the future service requirements of different verticals such as higher reliability, low latency and higher data rates, uO itself should be a 5G service provider, unless uO will not be able to satisfy those requirements. Therefore, the system architecture of uO should support this 5G communication scenarios.

1.2 Research Problem

uO is a 5G network operator and its system architecture should support 5G communications. However, because of the novelty of uO concept, system architecture for a uO is still not defined in a comprehensive manner. With this regard, in this thesis, we propose a descriptive architecture for emerging 5G uO, which is capable of providing user specific services in a spatially confined environment. Since uOs are specialized to provide tailored services, the system architecture of a uO and its deployment may also depend on the environment and the use case. Therefore, we define the architecture to cater the communication needs of a selected use case(s) of a selected environment.

Since uO is a 5G operator, the definition of uO architecture should be done with the aid of already defined 5G architectural components by 3rd Generation Partnership Project (3GPP) [7]. 3GPP definition of 5G access network (AN) and core network (CN) architecture consists of entities called network functions (NF). Creation of network functions is made possible by the techniques Network Function Virtualization (NFV) and Software Defined Networking (SDN) which will be discussed under literature review in detail.

Selecting the environment that the uO serves, identify the most common use cases exist within the selected environment, analyze the key communications steps to fulfill the requirements of those use cases, identify mandatory architectural components from 5G systems architecture defined by 3GPP and aggregate those components are the steps to be followed to realize the objective of defining uO architecture to cater case specific localized future communication services.

1.3 Selected Scope

Since uO is providing a specific service, tailoring the uO architecture to cater that specific service will enable efficient communication. Moreover, defining the uO architecture for a service which will be widely seen in future communications will yield more benefits in the case of real deployments. Future industrial environments will consist of ultra-dense sensor networks, intelligent machines, Automated Guided Vehicles (AGV) in their day-to-day operations. Therefore, future industrial environments will demand higher data rates, extremely low latencies and massive communications between devices. On the other hand, the operation of an industrial environment is most likely be confined to spatially. Therefore, a uO deployment to provide the future communication requirements of an industrial environment will be a useful implementation.

Considering these facts, for the thesis, we selected the smart factory environment which supports Industry 4.0 standards as the environment for which we define the uO architecture. Moreover, an industrial environment may consist of number of use cases. Therefore, deploying a uO to cater communication of most frequently seen use cases is important. Based on the 3GPP study on Communication for Automation in Vertical Domains [8], we select Augmented Reality (AR), massive wireless sensor networks and mobile robots as the use cases to be covered in the thesis. Hence, the uO architecture will be defined for an Industry 4.0 environment which consists of the above three use cases. It is assumed that the selected industry environment consists of all three use cases and the uO must cater them simultaneously.

We assume that the other generic communication requirements such as the factory workers' voice calls and internet access requirements are catered by MNO service offerings and uO is not involved to provide service for those communications. Hence, uO operation is spatially confined to the factory itself.

1.4 Methodology of the Work

To realize the objective of defining the uO architecture, we first selected one use case and identified the architectural components needed to cater the communication of the selected use case. In the process of identifying the architectural components, it is important to understand the typical communication steps of each use case and the message transfers between the terminal devices, access network and the core network. Message transfer procedures for general communication scenarios such as registering a device in the 5G network, Protocol Data Unit (PDU) session establishment between two devices, handover procedures are defined by 3GPP in Procedures for the 5G System [9]. Using those procedures we identified the message transfers needed for our use cases and then the network functions needed to support those message transfers. Once the network functions are identified the architectural components for that use case can be defined by combining them based on the 3GPP architecture definitions [7].

The task of identifying the architectural components is carried out for all the three use cases namely AR, massive wireless sensor networks and mobile robots. The final uO architecture should consist of all the network functions needed for the communications of all three use cases. Even though we assumed the factory consists of all three use cases, the requirements of each use case are different and therefore the resources for network functions needed in each use case also differs. Hence, to make the operations more efficient, communications of each use case must be done via different logical networks. One advantage of defining logical networks is to allocate optimal resources needed for each use case and make them independent of the resource utilization of the other use cases. Hence, we used network slicing concept to define logical networks and isolate the network function resources for each use case. Therefore, the final uO architecture should also comprise the network functions needed to implement network slicing.

To realize the conceptual design of the uO architecture we conduct simulations for the most important communication steps under each use case. The objective of the simulations is to identify the advantage of using locally deployed 5G uO to cater the future industry communications rather than obtaining the service from an MNO. For that, we compare two deployment models, one being the local uO serving the factory and the other being the traditional MNO serving the factory. In the local uO deployment, we assume that the core network of uO is located within the factory premises and for the MNO deployment we assume that it is located outside the factory premises at some distance. As the performance measurement, we take the latency parameters for each of the communication steps and we compare the latency advantage that the factory can get by deploying a local uO for the factory communications. We conduct experiments by moving the core network of MNO to different distances and observe the variation in latency. We also vary the resource level at MNO which results in different processing delays in network functions of MNO core network and observe latency variation.

In addition to that implement a local uO network using the commercially available Kuha base station and Cumucore Evolved Packet Code (EPC) and conduct experiments for direct communications between the devices in the uO network.

1.5 Contribution of the Thesis

The contribution of the thesis is a network architecture designed for a uO operating is a future industrial environment. The architecture is comprised of the network functions of the generic 5G architecture needed to cater selected use cases of the industrial environment. Performance of the proposed uO architecture is measured by comparing it with an MNO deployed model using simulations. Performance metric considered here is End-to-End (E2E) latency and it is measured against two parameters namely core network distance and network function processing delay. The thesis demonstrations that the uO based architecture is more effective in terms of latency than the MNO base deployment, for the multi use case industrial environment. Based on the work carried out during the period of the thesis work, a conference paper "Micro-Operator driven Local 5G Network Architecture for Industrial Internet" has been published in IEEE WCNC 2019 conference [10]. This paper illustrates the benefits of uO architecture for the augmented reality use case, rather than using MNO architecture deployment.

1.6 Organization of the Thesis

The thesis has seven chapters. The rest of the thesis is organized as follows.

Chapter 2 explains the existing literature and the state-of-the-art research which has already been carried out. The literature review has five sections and it first discusses the advancements of future 5G systems, future uses of 5G systems and key technologies. Then the 5G architecture defined by 3GPP, the enabling technologies such as NFV and SDN and Multi-access Edge Computing (MEC), NFs and their services. After that, it explains the uO concept and why to establish local 5G networks along with traditional MNOs. Then it describes the future smart factory environments and frequent Industry 4.0 use cases of AR, massive wireless sensor networks and mobile robots. Finally, it discusses the concept of network slicing which enables the creation logical networks on a common physical infrastructure.

Chapter 3 describes the conceptual architecture diagrams for each use case. Methodology used to derive the architecture from considering the use case, identify the communication steps of each use case, network functions needed to cater those communications, combining network functions to design the architecture and finally how network slicing is used to define the proposed uO architecture which caters all use cases simultaneously.

Chapter 4 elaborates the experimental setup which includes the descriptions of the two deployment models used for simulations, i.e. the uO model and the MNO model. In addition to that, it briefly explains about the simulation software used and the general simulation parameters and the values. A brief description of Kuha base station based uO network implementation is presented next.

Chapter 5 presents the results of the experiments for each industry 4.0 use cases. It illustrates how latency measurement varies with the independent variables namely core network distance and network function processing delay. It also discusses how uO deployment provides better performance than MNO deployment for each communication step. It also considers the critical communication step of each use case and which independent variable will be the dominant factor to decide the latency.

Chapter 6, discusses and critically analyses the work that has been done in this thesis. It includes a comparison of the thesis work with respect to the other research activities in the same domain. It also includes an analysis regarding "up to what degree the thesis objectives are met". Finally, it explains about the possible future research directions.

Chapter 7 includes a condensed summary of the research objectives, essential contents of the thesis, final results and what those results actually interpret.

2 LITERATURE REVIEW

This thesis focuses on defining an architecture for local 5G operator namely uO, who caters the future communication needs in an industry 4.0 environment for selected use cases. Therefore, the literature review first covers the future 5G networks in general, core and access network architecture of future 5G systems. Then, it discusses uO and the work that has been carried out regarding uO implementations. Since uO itself is a 5G network operator, extensive study has been carried out on the existing 5G network architecture released by 3GPP. Finally, this chapter covers the study on the industry 4.0 environment to identify the behaviour of communications future industries. It also discusses three common Industry 4.0 use cases called AR, massive wireless sensor networks and mobile robots, so that the defined uO architecture will support communication requirements of those industry 4.0 use cases.

2.1 Future 5G Networks

The last four generations of cellular technology have been an advancement of the previous technology and have each been a major paradigm shift. Indeed, 5G will be a paradigm shift from present wireless communication technologies, but it will also be highly integrative to provide universal high-rate coverage and a seamless user experience [2]. The fifth generation of mobile technology will not only be a mere evolution of the current network generation but can be considered as a revolution in the current information and communication technology field [11]. According to Next Generation Mobile Networks (NGMN) Alliance, the fifth generation of mobile technology will address the socio-economic demands and business contexts of 2020 and beyond [12]. By the design itself, 5G will include flexibility and scalability to enable wide range of use cases in the future. Key characteristics of 5G wireless systems are identified as extremely high data rates, ultra-reliability and low latency, and massive communication between devices [13]. Moreover, three specific areas of 5G services are diversified as Enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communication (URLLC) and Massive Machine Type Communication (mMTC) [14] [15]. The Mobile and wireless communication Enablers for the Twenty-twenty Information Society (METIS) project has derived 5G requirements into the following technical objectives [16].

- 1000 times higher mobile data volume per area
- 10 to 100 times higher number of connected devices
- 10 to 100 times higher user data rate
- 10 times longer battery life for low power massive machine communications
- 5 times reduced E2E latency

Key enabling technologies for future 5G wireless communications systems can be considered

- Software Defined Networking
- Network Function Virtualization
- Millimetre wave spectrum
- Massive Multiple Input Multiple Output (MIMO)
- Network ultra-densification
- Network Slicing
- Big data and mobile cloud computing and Multi-access Edge Computing [17] [18] [19] [20].

SDN decouples the control planes and the data planes of the network, hence the network intelligence and the state are logically centralized, and the underlying infrastructure is abstracted from the applications [21]. SDN focuses on four key features i.e. separation of control plane from the data plane, a centralized controller having the view on the network, open interfaces between the control plane and the data plane, and programmability of the network.

NFV, a complementary concept to SDN, enables the virtualization of entire network functions that were tied to hardware before to run on cloud infrastructure [17]. The main component of NFV is the Virtual Network Functions (VNF), which are software implemented network function rather than the dedicated network elements in the previous mobile communication architectures.

With the growth of the mobile data demand, the existing spectrum is getting crowded but much higher frequencies are still underutilized. 30-300 GHz frequencies are referred to as millimetre wave frequencies. The main advantage of millimetre wave communication systems is that they can achieve multigigabit data rates at a distance even up to a few kilometres in point-to-point communication [22].

Massive MIMO wireless communication refers to the idea equipping cellular base stations with a very large number of antennas. This will potentially allow for orders of magnitude improvement in spectral and energy efficiency using relatively simple processing [23]. Massive MIMO systems are highly energy efficient compared with their corresponding single antenna systems, thus contributing to the 5G technical objective of less power consumption.

Ultra-Dense Network (UDN) represents a new paradigm shift in future networks compared with the present mobile networks [24] [25]. The realization of this is done by the dense deployment of small cells in the places where enamours amount of traffic is generated [26]. Usually, the small cells need small amount of power for transmission, which limits the cell coverage.

The main feature of MEC is to push mobile computing, network control and storage to the network edges so that the computationally intensive latency critical applications to be processed at the edge of the network [27]. MEC has the advantages such as lower latency, energy efficiency for mobile devices, enhanced privacy and security for mobile applications [28].

Figure 1 shows how future 5G wireless communications will cater the needs of different use case scenarios.

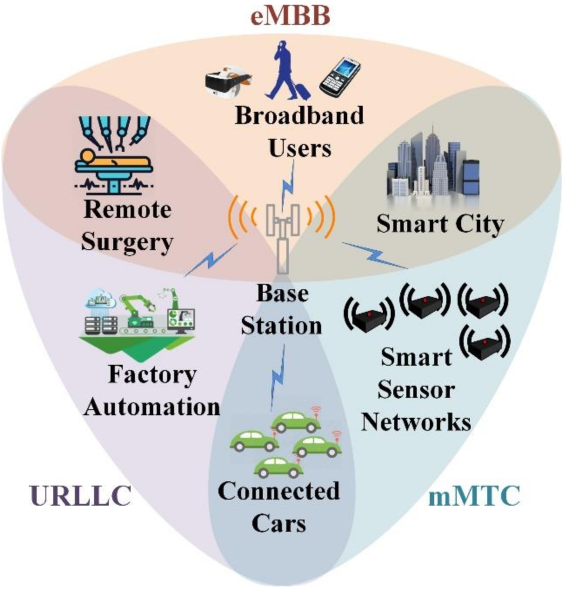


Figure 1. Use cases of future 5G wireless networks.

5G systems architecture is envisioned by NGMN as a combination of three layers [12]. Those layers are the infrastructure resource layer, business enablement layer and business application layer as depicted in Figure 2. The infrastructure resource layer consists of physical resources. It includes access nodes, 5G devices, cloud nodes, networking nodes and the connecting links. Virtualization principles enable the infrastructure layer to be exposed to upper layers.

The business enablement layer consists of all the functions required in a converged network. These functions are available in modular form and they together form the architecture. Network functions are realized by software modules using SDN technology, and they can be retrieved based on the need of the upper layer. More detailed analysis of each network function in this layer will be presented in Chapter 2.2. Business application layer consists of specific applications. These applications could be the services provided by the operator such as data services, content services, mobile services, application requirements from hospitals, campuses, industries, smart cities, vehicle-to-vehicle communications or any other third-party application requirements. One of the key characteristics of future 5G mobile communications systems is that it supports a multitude of applications and they are integrated to provide the user with a seamless experience. This way 5G has the ability of providing the specific communication needs of different verticals simultaneously to the provision of generic mobile communication services.

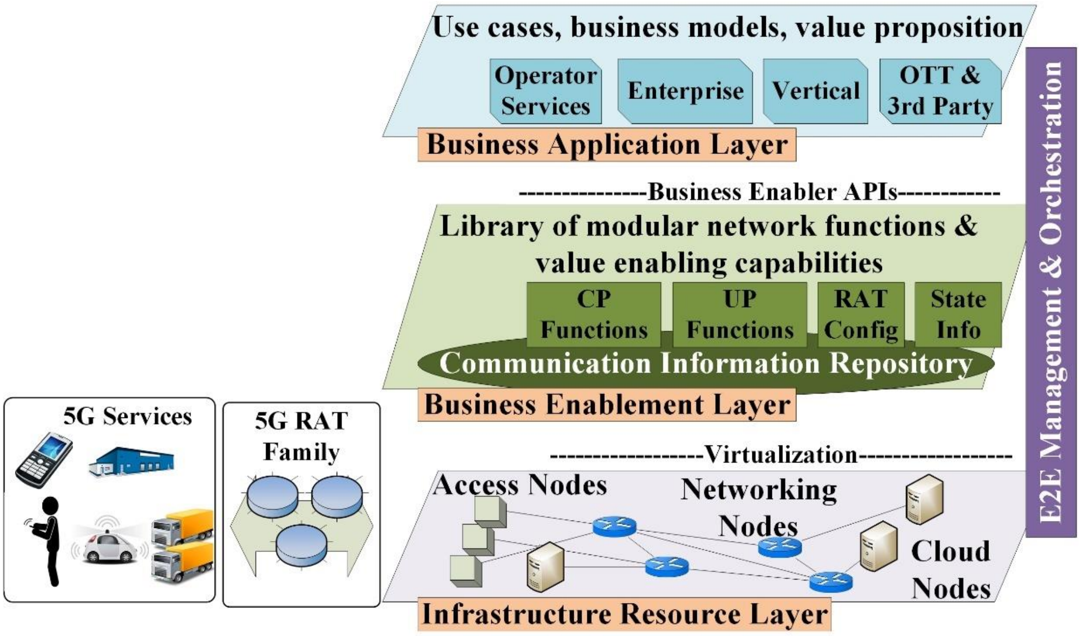

Use cases, business models, value proposition
Operator Services
Enterprise
Vertical
OTT & 3rd Party
Business Application Layer
---------------Business Enabler APIs------------
Library of modular network functions & value enabling capabilities
CP Functions
UP Functions
RAT Config
State Info
Communication Information Repository
Business Enablement Layer
------------------Virtualization------------------
Access Nodes
Networking Nodes
Cloud Nodes
Infrastructure Resource Layer
E2E Management & Orchestration
5G Services
5G RAT Family

Figure 2. NGMN view of 5G system architecture [22].

2.2 5G Network Architecture

uO itself is a 5G network operator. Therefore, the network architecture of 5G uO should also comprise the network functions of generic 5G architecture [5]. 3GPP has already released the specifications for 5G system architecture in its specification document 3GPP TS 23.501, System Architecture for the 5G System [7]. The main building block of 5G systems architecture is the NF. This contrasts with the network elements defined in EPC in 4G systems. SDN and NFV are involved in creating NF. There are multiple ways of implementing network functions in the 5G systems architecture. They can be implemented on a dedicated hardware or as a software instance on a dedicated hardware or as a virtualized function instantiated on an appropriate platform such as a cloud. The concept of network functions has led operators to add flexibility over the functionality of the underlying physical infrastructure of the 5G network.

5G network architecture separates the Control Plane (CP) functions from the User Plane (UP) functions. This enables network functions to be scaled and evolved independently and makes the deployment more flexible. 5G network architecture also modularizes the function design to enable flexible and efficient network slicing. The concept of network slicing will be discussed in detail in Chapter 2.5. Another important factor is ease of re-use of NFs in 5G, in contrast with 4G network elements [7].

3GPP specifications represent the 5G architecture in two ways. Service based representation shows how NFs within the control plane enable other authorized NFs to access their services. This representation shows the user plane functions connected as point to point links. Service based representation of 5G network architecture is depicted in Figure 3. Reference point representation shows the interaction between two NF as a point-to-point link between two NF. This representation is depicted in Figure 4.

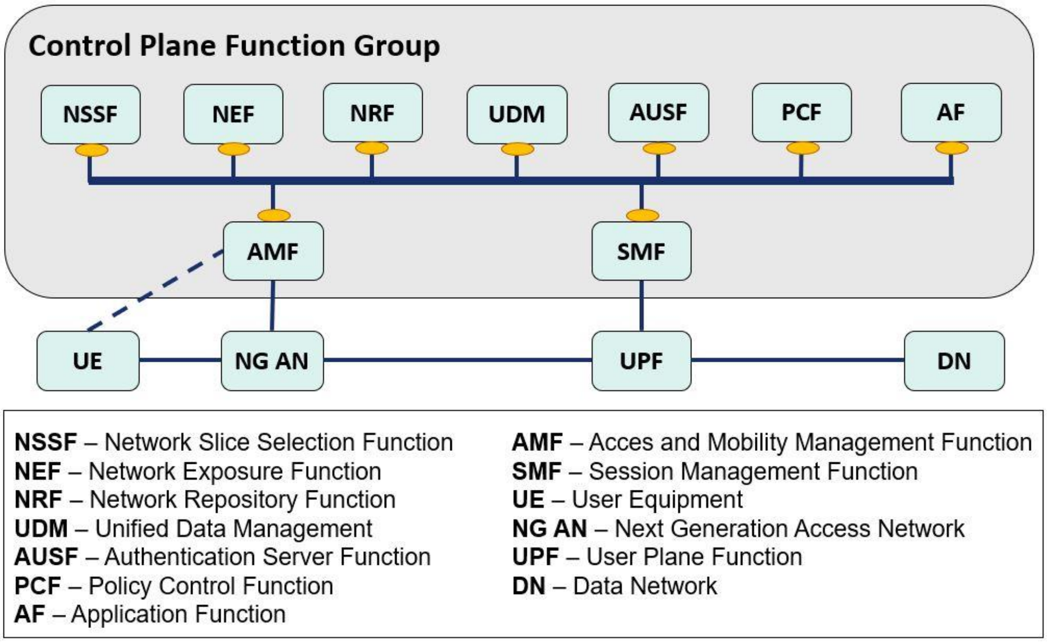

Control Plane Function Group
NSSF
NEF
NRF
UDM
AUSF
PCF
AF
AMF
SMF
UE
NG AN
UPF
DN
NSSF – Network Slice Selection Function
NEF – Network Exposure Function
NRF – Network Repository Function
UDM – Unified Data Management
AUSF – Authentication Server Function
PCF – Policy Control Function
AF – Application Function
AMF – Acces and Mobility Management Function
SMF – Session Management Function
UE – User Equipment
NG AN – Next Generation Access Network
UPF – User Plane Function
DN – Data Network

Figure 3. Service based representation of 5G system architecture [7].

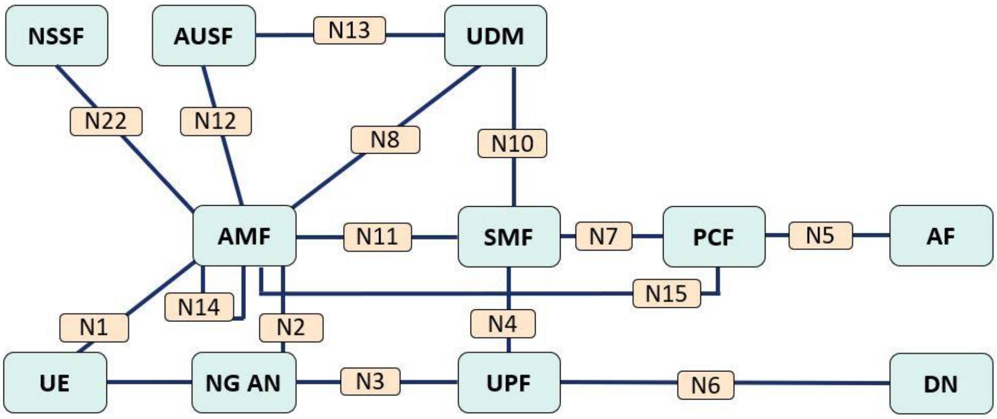


Figure 4. Reference point representation of 5G system architecture [7].

Each network function in 5G network architecture has its own functional description. Multiple instances of the same network function support the same functional description. A summary of the functional description of each NF is given in Table 1.

Table 1. A brief Description of 5G Network Functions

5G NF	Description
(R)AN	(Radio) Access Network
UPF	Supports: packet routing & forwarding, packet inspection, Quality of Service (QoS) handling, acts as external PDU session point of interconnect to Data Network (DN), and is an anchor point for intra- & inter-RAT mobility
DN	Operator services, internet access or 3rd party services
SMF	Supports: session management (session establishment, modification, release), UE IP address allocation & management, Dynamic Host Configuration Protocol (DHCP) functions. It also selects and controls the UPF for data transfer. If a UE has multiple sessions, different SMFs may be allocated to each session to manage them individually and possibly provide different functionalities per session
AMF	AMF provides UE-based authentication, authorization, mobility management. A UE even using multiple access technologies is basically connected to a single AMF because the AMF is independent of the access technologies
NSSF	Supports: selecting of the Network Slice Instance (NSI) to serve the UE, determining the AMF set to be used to serve the UE
NEF	Supports: exposure of capabilities and events, secure provision of information from external application to 3GPP network, translation of internal/external information. Ex: NEF handles masking of network and user sensitive information to external AF's according to the network policy
NRF	Maintains NF profile of available NF instances and their supported services. NF profile may include NF instance ID, NF Type, NF capacity information, Network Slice related identifiers.
UDM	UDM stores subscription data of UE
AUSF	Acts as an authentication server. Stores data for authentication of UE
PCF	Supports: unified policy framework, providing policy rules to CP functions, access subscription information for policy decisions in UDR
AF	Supports: application influence on traffic routing, accessing NEF, interaction with policy framework for policy control

Each control plane function in 5G network architecture has a service-based interface which facilitates the other control plane functions to access its services. Some of the most common service-based interfaces and its description is depicted in Table 2.

Table 2. Service-based Interfaces of 5G Systems Architecture

Interface	Description
Namf	Service-based interface exhibited by AMF
Nsmf	Service-based interface exhibited by SMF
Nnef	Service-based interface exhibited by NEF
Npcf	Service-based interface exhibited by PCF
Nudm	Service-based interface exhibited by UDM

Naf	Service-based interface exhibited by AF
Nnrf	Service-based interface exhibited by NRF
Nnssf	Service-based interface exhibited by NSSF
Nausf	Service-based interface exhibited by AUSF

Apart from the service-based interfaces, 5G system architecture also contains number of reference points, which represents the point-to-point interaction between two network functions. Few examples for 5G network architecture reference points are as follows:

- N1: Reference point between the UE and the AMF
- N2: Reference point between the (R)AN and the AMF
- N3: Reference point between the (R)AN and the UPF
- N4: Reference point between the SMF and the UPF
- N5: Reference point between the PCF and an AF
- N6: Reference point between the UPF and a DN

The knowledge of the 5G systems architecture is mandatory to develop the architecture for uO because uO itself functions as a 5G operator. Literature presented in this chapter was later used to derive the uO architecture for the selected use cases.

2.3 Micro Operators

Future wireless communications will require network operators to support more efficiently to the location specific needs than the current MNOs offerings. These location specific requirements will raise from environments such as factories, hospitals, campuses which will require higher connection densities in an indoor environment. These communications should be able to provide higher data rates, low latency and massive communication between devices.

Today, the mobile telecommunication market is run predominantly by MNOs and their key focus is to serve the masses. Usually, the investment cycle of an MNO is long, which makes it difficult for MNOs to respond rapidly for case specific, location specific needs [5] [29]. Also, it is difficult for new operators to enter the market because of the high investments and scarce spectrum resource. Apart from that, more traffic will be originated from indoors in the future and catering such indoor needs by already existing macro cells will be operationally difficult. On the other hand, each MNO catering the indoor need using small cells will not be a preferable approach. Therefore, the future communication requirements are unlikely to be met solely by MNOs. Changes in the future mobile connectivity market are illustrated in Table 3[5].

Table 3. Changes in Mobile Connectivity Market

Today	**Future**
MNOs dominate the market. They offer generic services to masses.	New local connectivity market is there with different verticals having different requirements which need to be deployed fast
Limited number of spectrum licenses	Changes in spectrum regulation. Mobile market will open for new entrants.
Entry barrier to mobile market is high due to high investment barrier to build own infrastructure.	Sharing of resources. Networks are scaled on demand and assets currently owned by others are shared.

Therefore, an entity other than the conventional MNOs should be there to cater future communication needs. This leads to the establishment of the local 5G networks. Local 5G networks are gaining increasing attention nowadays. This means that the entities other than MNOs can deploy 5G networks to cater case specific, location specific communication requirements. These local 5G networks can coexist with traditional MNOs in a way that the local 5G network provide the case specific requirement while the traditional MNO can cover the generic communication needs. A uO can establish small cell networks to support the communication needs while collaboratively working with network infrastructure vendors, regulators, facility owners and MNOs.

Figure 5 illustrates how uOs providing case specific and location specific services while MNO providing the generic services and how both uO and MNO can collaboratively work to achieve future commutation needs. In Figure 5, the specific communications in the factory and the hospital are served by the uO while the subscribers in the factory are served by the MNO for their day-to-day communication needs such as internet access.

uO relations with other stakeholders can be broadly categorized into three main domains called regulatory aspects [30], business aspects and technology aspects. The main problem in terms of regulator is the spectrum availability. The deployment of uO leads us to the question of what is the spectrum uO will use in their operation. Because the licensed spectrum is scarce and can only be shared by few operators and it is not possible to utilize by uOs. This leads to the discussion of micro licenses where uO can build and operate indoor small cell networks. For example, 3.5 GHz is a potential band for uO operations in Finland which is used for 5G test networks [5].

The growing amount of mobile traffic originated from the indoors and the timely implementation of communication solutions to cater those traffic requirements will become an increasingly important business opportunity. Because MNOs cannot respond to these rapid deployment needs, local operators who offer case specific and location specific services can make use of this opportunity much easier than conventional MNOs.

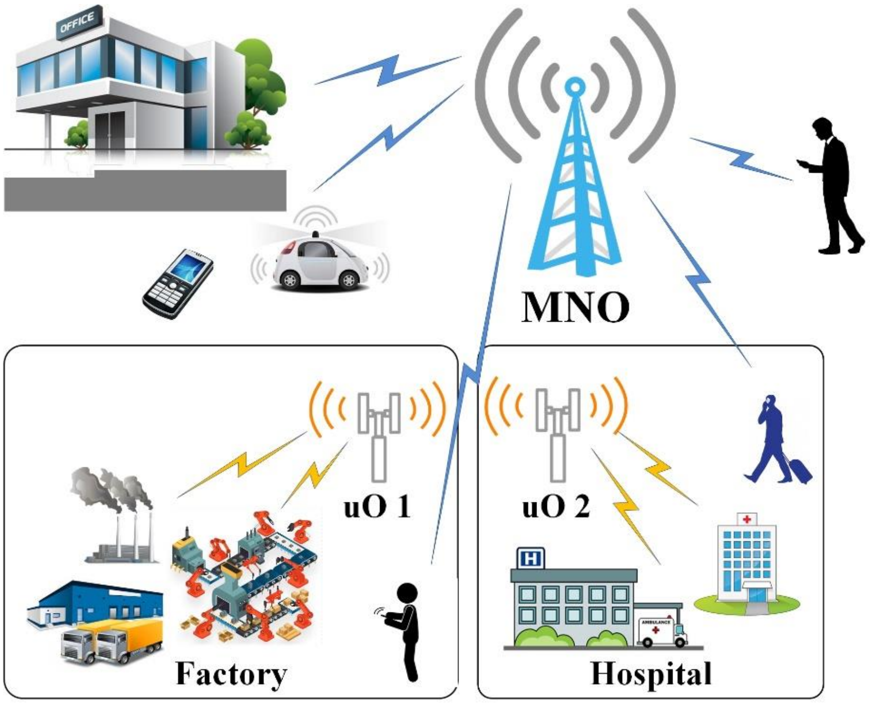


Figure 5. Collaborative service provision by both uO and MNO.

The major step of technical developments will come with 5G. 5G will support connecting vast number of devices connected to heterogeneous networks. Also, there will be new frequency ranges introduced which were not utilized in the air interface before. Apart from that NFV, SDN, new 5G architecture, massive MIMO, network slicing will be key technologies adopted by future uO. uO concept will benefit from the 5G developments to play its role as a local operator. For example, without having to invest in the whole infrastructure, uO can obtain a network slice from an existing infrastructure vendor to operate its own network. Figure 6 depicts the relations of a uO with other possible stakeholders.

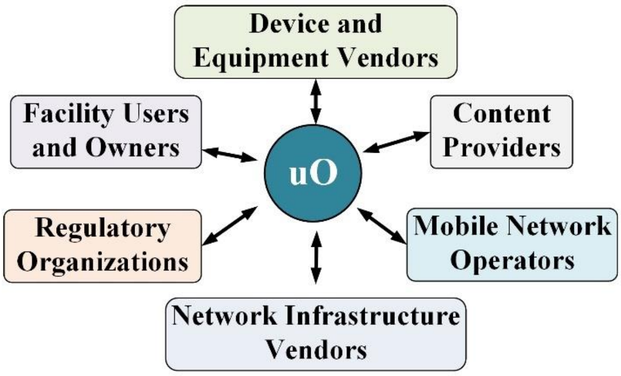

Device and Equipment Vendors
Facility Users and Owners
Content Providers
uO
Regulatory Organizations
Mobile Network Operators
Network Infrastructure Vendors

Figure 6. uO and its relations with other stakeholders [5].

2.4 Industry 4.0

Locations belong to different vertical sectors have different communication requirements, hence the network functions needed in catering those communication requirements will also be different. Therefore, the network architecture of a uO who is providing a case specific, location specific service, will also depend on the communication requirement.

Industry 4.0 refers to the advancement of industries into the next generation [31] [32] [33]. It aims to interconnect the factory devices, make them smart by adding more intelligence and ultimately resulting in improved adaptability, resource efficiency, and the supply and demand process between factories [34]. Typical industry 4.0 setting is depicted in Figure 7.

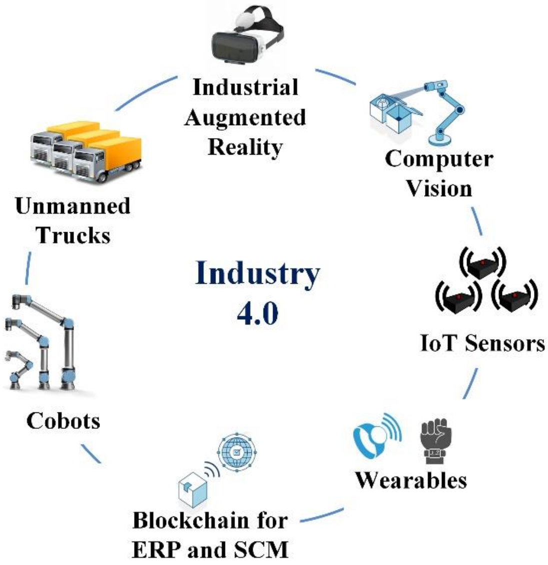


Figure 7. Industry 4.0 elements.

Machine-to-Machine (M2M) communication plays a critical role in Industry 4.0 which is also a key focus in 5G systems. Because of the massive machine type communication, Industry 4.0 environments put a stringent requirement on the communication infrastructure. Some of the key features of Industry 4.0 which makes it differ from the typical wireless sensor networks currently existing are latency, mobility, environment and capacity [35].

With the high density of machines in an Industry 4.0 environment, monitoring of those machines will become increasingly difficult. Apart from that there will be considerable increase in control traffic. However, in the future, all these must be carried out in real-time and therefore the communication puts a high demand on the latency of the communications. With the requirement of having low latency communications, the network architecture supporting the communications of an Industry 4.0 environment should also be very specific so that the communications requirements are met.

Mobility is another key characteristic where we will see significant changes when it comes to Industry 4.0 environments. As the number of moving nodes increases, the communication infrastructure catering the environment should also be designed in a way that it can handle these mobility requirements. Examples for moving nodes can be mobile products, mobile robots, AGVs, Unmanned Aerial Vehicles (UAV), workmen and other mobile devices. As an example, in a factory which has higher number of fast-moving robots, the network should be able to handle higher number of handovers in case of small cell deployment.

Environment will differ significantly in industrial wireless networks rather than wireless sensor networks. An industrial wireless network operates in a challenging environment because of various factors such as dust, vibration, heat, various obstacles, higher temperature and humidity. Additionally, signal interference generated from the motors will make the environment more challenging. Therefore, Industrial Wireless Networks (IWN) need additional strategies to ensure reliability and efficient communication.

Finally, the capacity requirement of an Industry 4.0 environment will be much higher in the future. There are multiple reasons which cause the increased capacity requirements. There will be a number of devices communicating with each other, in contrast to today’s environment. Consequently, IWN nodes require higher capacities. Therefore, the communication service provider who provides the connectivity should be able to cope with the fact that there will be very high capacity requirements in the future industrial networks.

As mentioned in the introduction, network architecture for a uO should be tailored to the environment and the use case to make the implementation and operation more efficient. For example, if the uO communications do not need to establish connectivity to the outside the factory, it needs comparatively less resources for UPF than a uO which needs internet access because of the less load in UPF in general. In this thesis, we select Industry 4.0 as our environment. We consider three most common Industry 4.0 cases called AR, massive wireless sensor networks and mobile robots [8] and examine their key characteristics.

2.4.1 Augmented Reality in Industrial Networks

Even though the workforce plays a substantial role in the future factories, to make the factory operations more efficient and operations more smoother factory workers will be supported by other entities. In this regard, AR plays a crucial role in an Industry 4.0 environment. It can be used to monitor the processes of the production flows, providing step-by-step guidance to perform pre-defined tasks by the workers and to obtain remote support for tasks such as maintenance or service checks. Therefore, AR can be considered as an application which will be heavily used in future industrial environments [8].

In this context, AR devices should be worn by the workers as a head-mounted device allowing them to use their hands to carry out their operational tasks. Therefore, these devices should be lightweight and highly energy efficient as frequent charging of these devices would be an operational overhead. To consume less energy, AR devices need to carry out minimal processing and more intensive tasks to be offloaded. For example, AR devices must take the responsibility of capturing the image, transmit it to external server and display the augmentations received from the server while the server carries out the more intensive tasks such as rendering the image and apply augmentations on top of the image. The closer the image processing server located to the AR devices, the lower the latency of the entire communication process can support. A diagram of a typical AR network is given in Figure 8. Possible communications within an Industry 4.0 AR network should be as depicted in Figure 9.

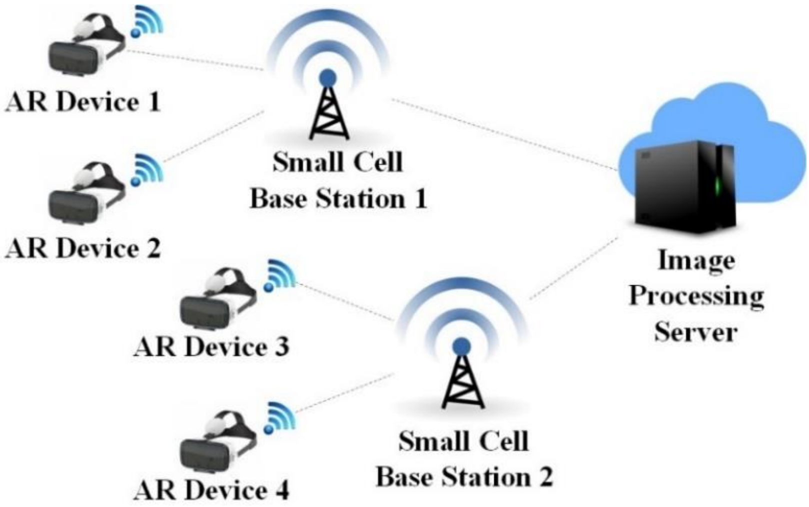


Figure 8. Components of a typical AR network.

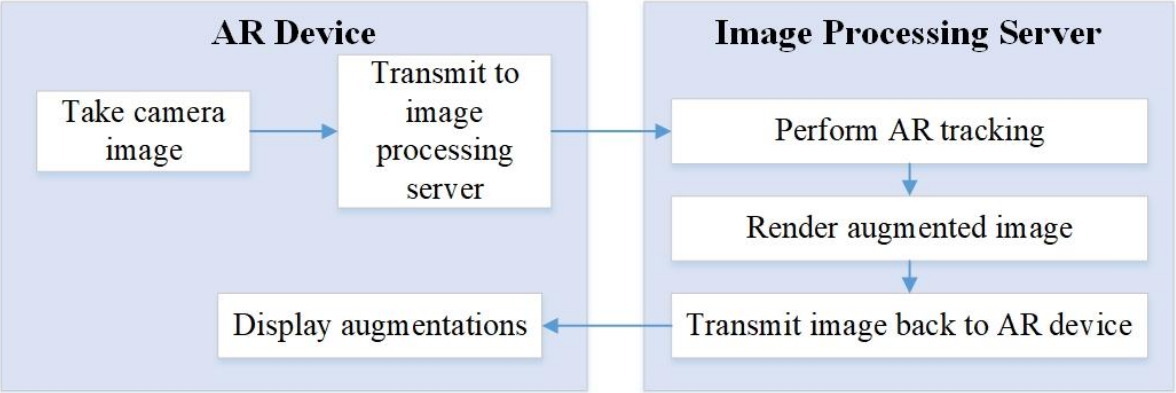


Figure 9. AR system model with offloaded processing [8].

The service flow of an AR system can be explained as follows.

1. A camera integrated with the AR device takes the images with a given frame rate and with an acceptable resolution for the application.
2. AR device continuously transmits all images to the image processing server via 5G system. Location of the image processing server can most probably be in a (local) edge cloud. This could vary based on the requirements for the implementation.

3. Image processing server does the processing of the images and determines the current field of view of the camera integrated into the AR device.
4. Image processing server determines the optimal placements of the augmentations and places augmentations in the current image.
5. The image processing server then renders the augmented image(s).
6. Image processing server transmits the rendered image(s) back to the AR device via the 5G system.
7. The AR device displays the augmented image.

The challenge associated with this AR operation is that the communication should support very stringent latency requirements while supporting a higher data rate. To make the data rate lower, AR devices can be equipped with advanced image compression techniques, but this affects the battery life of the device. This is against the concept of making AR devices lightweight. Other challenge is Mobility. Workers wearing the AR devices will always move along the factory floor and this will cause higher number of handovers in case of a small cell network.

Potential requirements of the AR use case were discussed in 3GPP technical report TR 22.804, Study on Communication for Automation in Vertical Domains. Table 4 provides a summary of these requirements.

Table 4. Potential requirements for Industrial AR Systems [8]

Description	Requirements
Frame rate and image resolution	Frame rate ≥ 60 Hz Resolution: HD (1280 x 720) or Full HD (1920 x 1080)
E2E latency between capturing a new image and displaying the augmented image	Round trip latency < 50 ms to avoid cyber sickness
One-way latency of the communication between AR device and image processing server	One-way latency ≤ 10 ms
Percentage of successfully delivered video frame within given latency constraints	Availability ≥ 99.9%
Number of AR devices	Should support at least 3 devices per base station
Handover from one base station to another	Seamless mobility without having observable impact on the application
Privacy and security	Video stream between the AR device and the image processing server shall be encrypted and authenticated by the 5G system

2.4.2 Massive Wireless Sensor Networks

In Industry 4.0, massive wireless sensor networks will be used for monitoring the parameters of the working environment. The sensors can monitor various types of parameters such as pressure, humidity, temperature, CO_2, sound and so on. The main purpose of having a sensor network to monitor the environment is to detect malfunctions in the surrounding. Once a malfunction is detected, appropriate actions must be taken by a decision-making entity so that the effect from the malfunction is mitigated. As an example, when an anomaly of the room temperature in detected, a machine can be triggered to its emergency stop. On the other hand, if the anomaly detected by the sensor network is not critical, the ongoing operation can be allowed to continue until certain solutions are provided by the experts. One key aspect of setting up of a sensor network is to decide where to implement the monitoring function. In case of the sensor nodes are less computationally complex, the functionality can be placed in a central cloud server. On the other hand, the functionality can be placed inside the factory environment in case of critical latency requirements. The idea behind an industrial wireless sensor network is depicted in Figure 10.

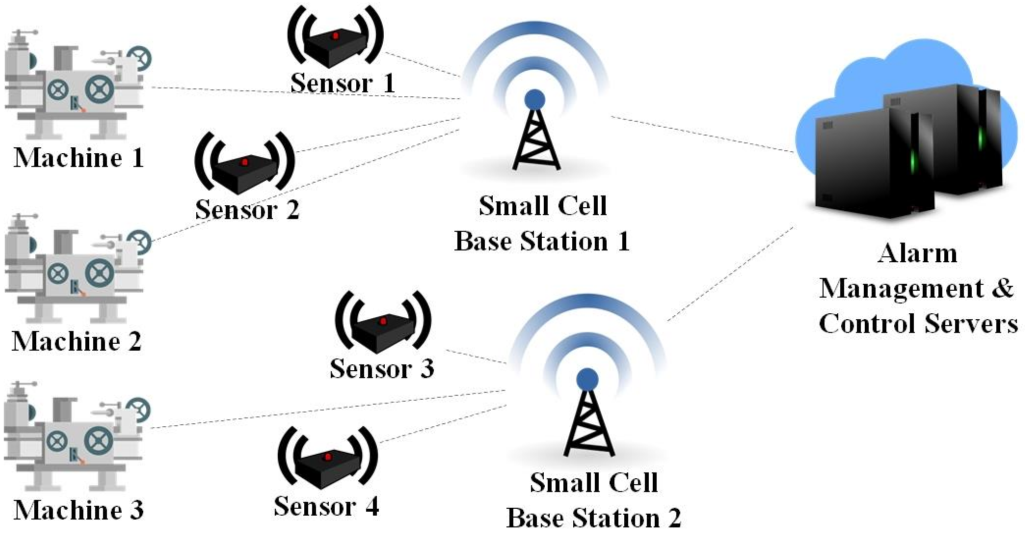


Figure 10. Wireless sensor networks to monitor industrial environment.

In an industrial wireless sensor network, measuring the environment can be realized in different ways. In the case of the simple sensor implementations, all the measurements taken by the sensors are transmitted to the management server and then the server decides on the controlling action based on the sensed values. In contrast, if more advanced sensors are deployed in the sensor network, the sensors can decide whether the measured data violates a given threshold and if violated, then they transmit the data towards the management server. This realization reduces the communication traffic.

Service requirements of the sensor networks can be categorized into three main areas based on the impact of the anomaly. The highest priority is given for condition monitoring for safety, and then for event-based condition monitoring and finally interval-based condition monitoring. Typical service monitoring requirements such as E2E latency, communication range per node, priority and service availability requirements are depicted in Table 5. E2E latency is considered for the entire communication even if there are other communication systems in the path (not necessarily 5G).

Table 5. Typical monitoring service requirements of future industrial environments[8]

Scenario	Priority	E2E Latency	Service Availability	Communication Range per Node
Condition monitoring for safety	Highest (1)	5 ms - 10 ms	> 99,9999% - 99,999999%	< 30 m

Event-based condition monitoring	High (2)	50 ms - 1 s	> 99,9%	< 30 m
Interval-based condition monitoring	Medium (3)	50 ms - 1 s	> 99,9%	< 30 m

The flow of actions for an anomaly can be explained in below steps.

1. Sensing devices continuously sends sensed data to the centralized server for learning the environment
2. An anomaly in the data is detected
3. Detected event is analysed at the management server and it is propagated to controlling instance
4. Controlling instance takes the action

Following are the challenges for the 5G system associated with implementing an industrial wireless sensor network.

- Large number of sensors per small cell base station
- Possible high data rates
- Low-latency requirements combined with high reliability

2.4.3 Mobile Robots

In future industrial environments, mobile robots such as AGV will be used in numerous applications and will play an extremely important role. Mobile robots can be programmed to execute multiple operations fulfilling number of tasks. Therefore, mobile robots can be used in transporting goods within the industrial environments instead of the mechanisms built with conveyor belts. These can also be used to assist human workers inside the factory environments. Such mobile robots are also called collaborative robots. They can sense and react with their environment therefore they operate more intelligently than the traditional machines programmed to travel through the pre-defined paths.

For the proper operation of mobile robots in an industrial environment, they should be monitored and controlled by a guidance control system. This will avoid collisions between robots, assign driving jobs and manage the traffic of mobile robots. Guidance for the mobile robots can be provided either using the tracks on the factory floor or the robot itself can guide itself using its own sensors such as cameras and lasers.

The communications of mobile robots can be categorized into three main cases based on the fact that who is communicating to whom. They are;

1. Communication between mobile robot and guidance control system
   Having communications between guidance control system and the mobile robots is extremely important. The guidance system can instruct the mobile robot to stop in case of an emergency. Communication from the robots to the control system is necessary to transfer data such as video or image data to the control system.
2. Communication between mobile robots
   The Main objective of the communication between mobile robots is to create a collision-free environment. Also, it helps to perform synchronized actions in case multiple robots are parts of a particular operation.
3. Communication between mobile robots and peripheral facilities
   This kind of communication helps to perform actions on the environment such as opening and closing doors.

Following aspects can be considered as the challenges to the 5G systems, which connects these mobile robots to the other elements in the networks

- Very stringent requirements on latency and communication service availability
- Very stringent requirements on synchronization between mobile robots
- Seamless mobility support
- Potentially high density of mobile robots
- Good 5G coverage requirements in indoor environment due to mobility of robots

Figure 11 depicts a typical industrial setting which employs mobile robots.

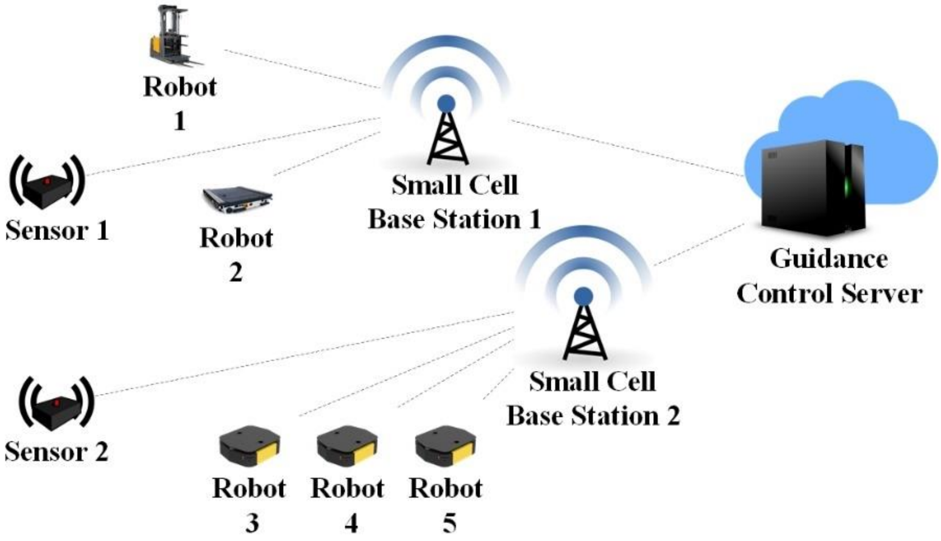


Figure 11. Mobile robots in an industrial environment.

2.5 Network Slicing

Network slicing refers to the methodology of sub-dividing a network into logically isolated sub-networks. NFV and SDN are the key techniques used here and the resulting network slices have their own capability to cater certain communication requirements. As an example, a network slice may serve to cater low latency communications with reliability guarantees and can appear to applications as a single network abstracted form underlying communication technologies [36]. Figure 12 illustrates the concept of network slicing where the physical network is sub-divided into two slices called slice 1 and slice 2.

Industry 4.0 will feature connections of all three generic types of services envisioned in 5G wireless systems i.e. eMBB, URLLC and mMTC [13]. It is difficult to optimize the network simultaneously to cater all these three types of services because of the differences of the requirement of each service. This issue can be handled by network slicing, where different slices can be created to cater each service, and the resource optimization can be done within the slice so that the slice will function optimally to cater for the desired service. As an example, the URLLC slice may be used to deliver critical control instructions to a machine in case of an emergency situation while the eMBB slice can be used to support the communication between a mobile robot and internet for a firmware upgrade.

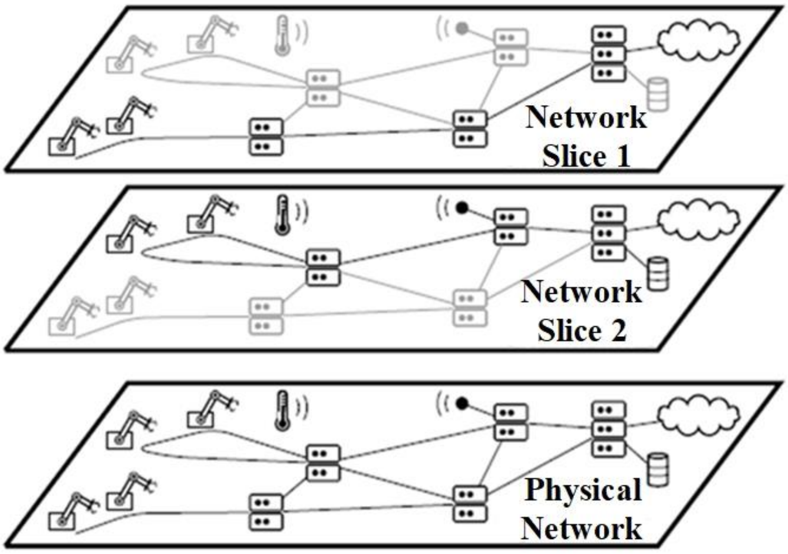


Figure 12. Dividing the physical network into two slices.

Several challenges can be identified during the implementation of network slicing in Industry 4.0 environments [37]. First, creating an ultra-reliable and low latency communications is challenging since it is difficult to accurately model the queuing delays. Second, Industry 4.0 networks are heterogeneous and equipped with many legacy protocols, which makes the end-to-end analysis complicated.

3 PROPOSED ARCHITECTURE

The communication service provider who offers services to an Industry 4.0 environment should be able to provide a reliable communication. Based on the characteristics of each Industry 4.0 use case and the communication requirements raised in their day-to-day operation, the mandatory NFs and the resources needed can be identified. These elements then can be used to derive the final network architecture for the uO. In this chapter, we consider each Industry 4.0 use case we discussed under literature review i.e. AR, massive wireless sensor networks and mobile robots and analyse their day-to-day operation in detail. Then, mandatory network functions needed in their day-to-day operation are identified and those network functions are combined to derive the architecture for each use case. Then we consider a factory where all the use cases are in operation. Since the resource requirements of the three use cases are different from each other, we propose the final architecture equipped with logical networks to cater the communication of the three use cases. We use network slicing concept to implement logical networks to provide services for each use case.

3.1 uO Architecture for AR Communication

In a typical AR use case, AR devices take the video and transmit the frames to the image processing server. Image processing server does the processing of the frames, generate augmentations, apply those augmentations on top of the picture frames and sends those frames back to AR devices. AR devices then display those augmented frames as a video. According to 3GPP TR 22.804, 5G system supporting this communication should be able to provide an E2E latency of less than 10 ms for one-way communication with a 99.9% success of frame delivery as mentioned in Table 4.

To support the stringent latency requirements, the appropriate implementation is to have a 5G network to cater the needs of AR use case. 5G network can be deployed by an MNO where there is indoor access connectivity for AR devices via small cell base stations. This is then connected to the MNO core network using fiber links. In this case, the communication is supported by MNO core network which is not located at the factory premises itself. In contrast, a local 5G network deployed by a uO covering the factory could also be used to address the needs of AR use case. The uO concept provides flexibility over the selection of architectural components and the location where the core network is hosted, unlike the traditional MNO offering. For a stringent latency requirement, the desirable implementation is to have a uO to support the AR communications and core network of uO should be located within the factory premises itself, but not mandatory. These deployment models are depicted in Figure 13 and Figure 14 respectively.

Figure 13 illustrates how a locally deployed 5G operator can support the AR communication. Multiple AR devices connect to the nearest Next Generation NodeB (gNB) and those gNBs are connected to locally deployed uO core network via routers. Image processing server is also connected via 5G gNB and has connectivity to the core network. All necessary network functions are implemented within the local core network so that the low latency communications can be achieved.

On the other hand, MNO deployment illustrated in Figure 14 clearly shows that packets from AR devices must transmit through the MNO network to reach its core network and then back to the factory premises to contact the image processing server. In terms of low latency communication, uO deployment is more beneficial.

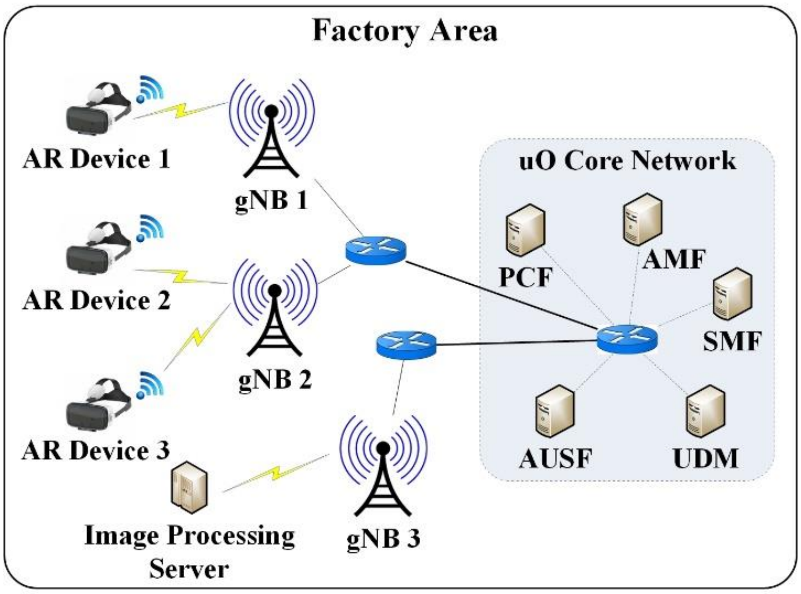


Figure 13. AR communication catered by uO deployed within factory.

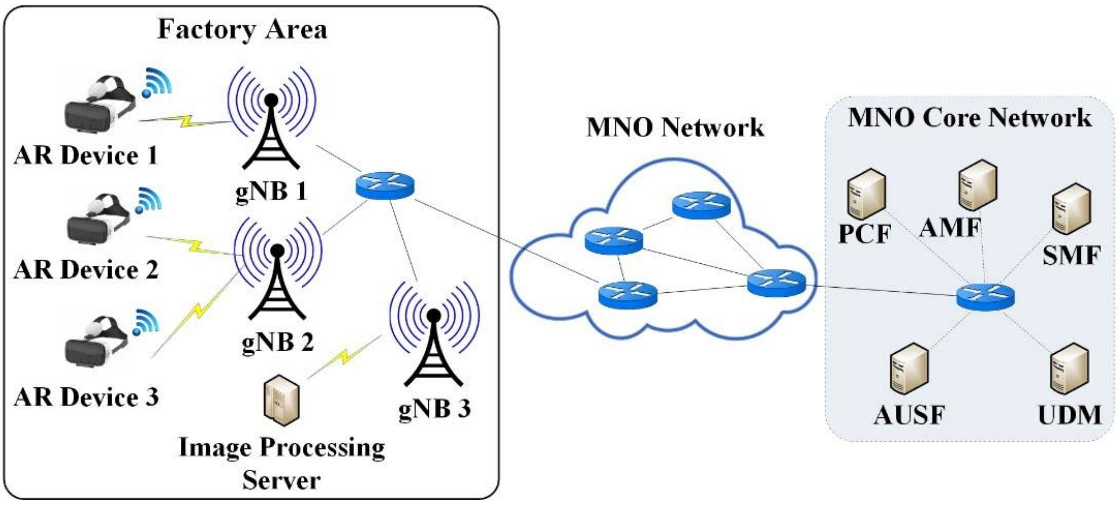


Figure 14. AR communication catered by MNO.

To decide the architectural components needed to cater the AR use case, it is necessary to identify what kind of communications occurring during a typical AR scenario. Generally, AR use case requires the 5G system facilitate the following three steps of communications.

- Registering the AR devices into the network
  This is the process where the network identifies the AR devices as elements in the 5G network. All the messages transferred between AR devices, access network and core network functions must be considered when deciding the architectural components for the use case.
- Establishing data session between the AR device and image processing server
  Once the devices are identified as network elements in the 5G system, AR device should be able to establish a data session with the image processing sever to

successfully transfer data from the devices to the server and receive the augmented images from the server to display. All the messages required for session establishment must be considered in order to identify any other network functions needed to cater this communication

- Data transfer between AR devices and image processing server
  Once the data session is established, the actual data transfer happens between the AR devices and the image processing server. Network functions needed in this user plane operation must also be considered when defining the architecture.

All the communications are happening between the AR device, gNB and core network functions. We define the registration procedure for AR device based on 3GPP specifications [38]. Figure 15 illustrates the message sequence between the entities in the architecture.

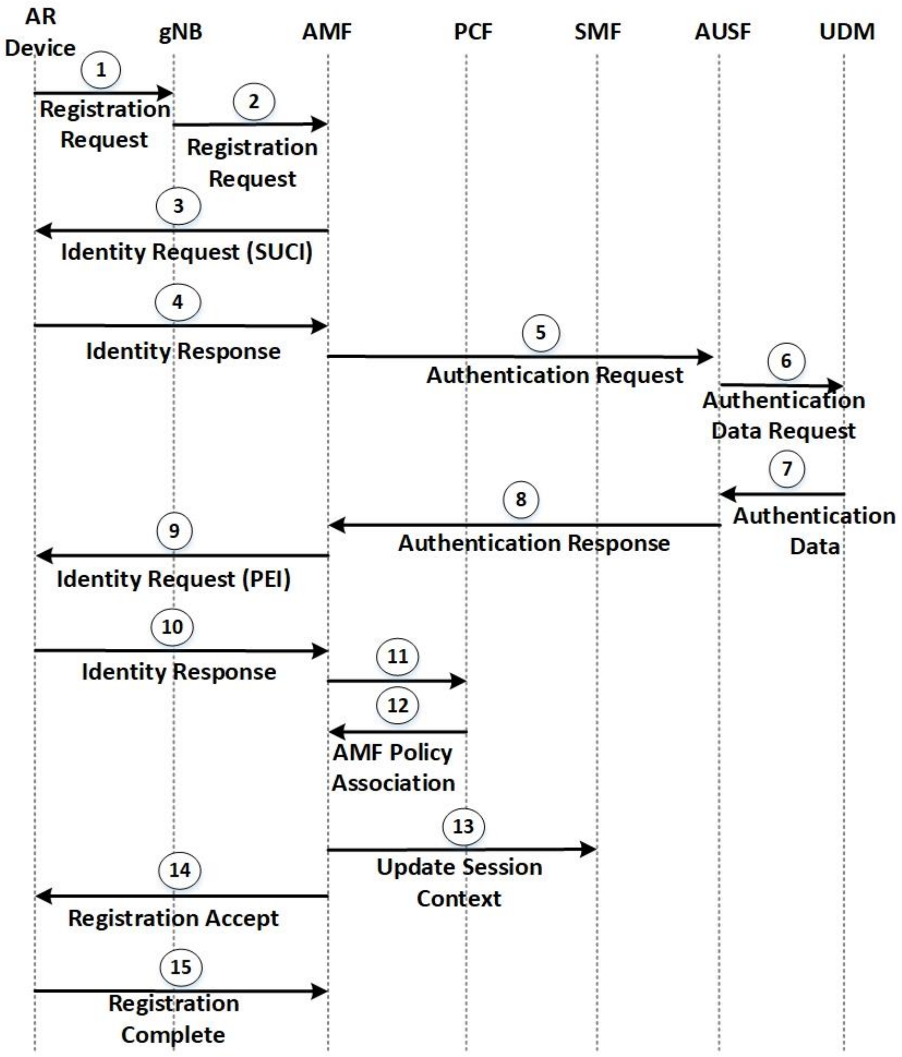


Figure 15. Message sequence chart for AR device registration procedure.

AR device initiates the registration process by sending registration request to gNB. gNB forwards the request to Access and Mobility Management Function (AMF). After that, AMF sends the identity request message to AR device and AR device responds to that by sending the identity response message. Once the identity response message is received from AR device, AMF contacts Authentication Server Function (AUSF) for the device authentication. AUSF facilities the authentication after contacting the Unified Data Management (UDM). AUSF sends the authentication data request to UDM and retrieve the authentication data. Once the authentication data is received from UDM, AUSF sends the authentication response to AMF. AMF then sends the identity request message to AR device and AR device sends the identity response messages to verify the identities. After the identity verification, AMF then works with Policy Control function (PCF) for the policy association for the AR device. Once the policy association is done, AMF sends an update to Session Management Function (SMF) informing the session context. AMF also sends the registration accept message to the AR device and the device then sends registration complete message. Once the AMF receives the registration complete message, the device registration process in completed.

After completing the registration process, AR device has to establish a data session with the image processing server to enable continuous data transfer as the 2nd step of the communication. We define the PDU session establishment procedure between AR device and the image processing server based on 3GPP specifications [38]. Figure 16 illustrates the message sequence required for PDU session establishment process.

Here, AR device initiates the process by sending PDU session establishment request to AMF via gNB. AMF then sends a request for a new session creation to SMF. In the next step, SMF registers with UDM, subsequently UDM stores data related to the session. After that SMF sends the Response to AMF regarding the session creation. Then, PDU session authentication/authorization process occurs by exchanging messages between AR device, gNB, AMF, SMF, UPF and Server. Once this step is completed, SMF works with PCF for policy association for the session by exchanging messages between SMF and PCF. Then SMF sends the session establishment/modification request to UPF and UPF sends the respective response back to SMF. Message transfer from SMF to AMF allows AMF to know which access towards the AR device to use. AMF then sends the PDU session ID information to gNB so that gNB can work with AR device for the gNB specific resource setup. After that gNB sends the acknowledgement for the PDU session request to AMF. Based on that, AMF sends request regarding PDU session update to SMF and SMF then requests UPF for session modification. Once SMF received the response from UPF regarding the session modification, SMF finally sends the response for PDU Session update to AMF concluding the PDU session establishment process.

After completing the PDU session establishment process, AR device can successfully send a continuous data stream to image processing server and display the augmentations received from the server on its screen to facilitate the workers. For this communication, only the user plane network functions are involved from the core network side. Therefore, the involved entities are AR device, gNB, UPF and the image processing server. Based on the three communication steps described above, 5G network functions needed to cater the AR use case can be identified to derive the core network of the uO architecture. Network functions which were not used in any of the communications can be ruled out, making the architecture more specific to the use case. This is an efficient arrangement because the resource requirement is also less since some of the network functions in the generic 5G architecture is not even necessary to implement inside the core network.

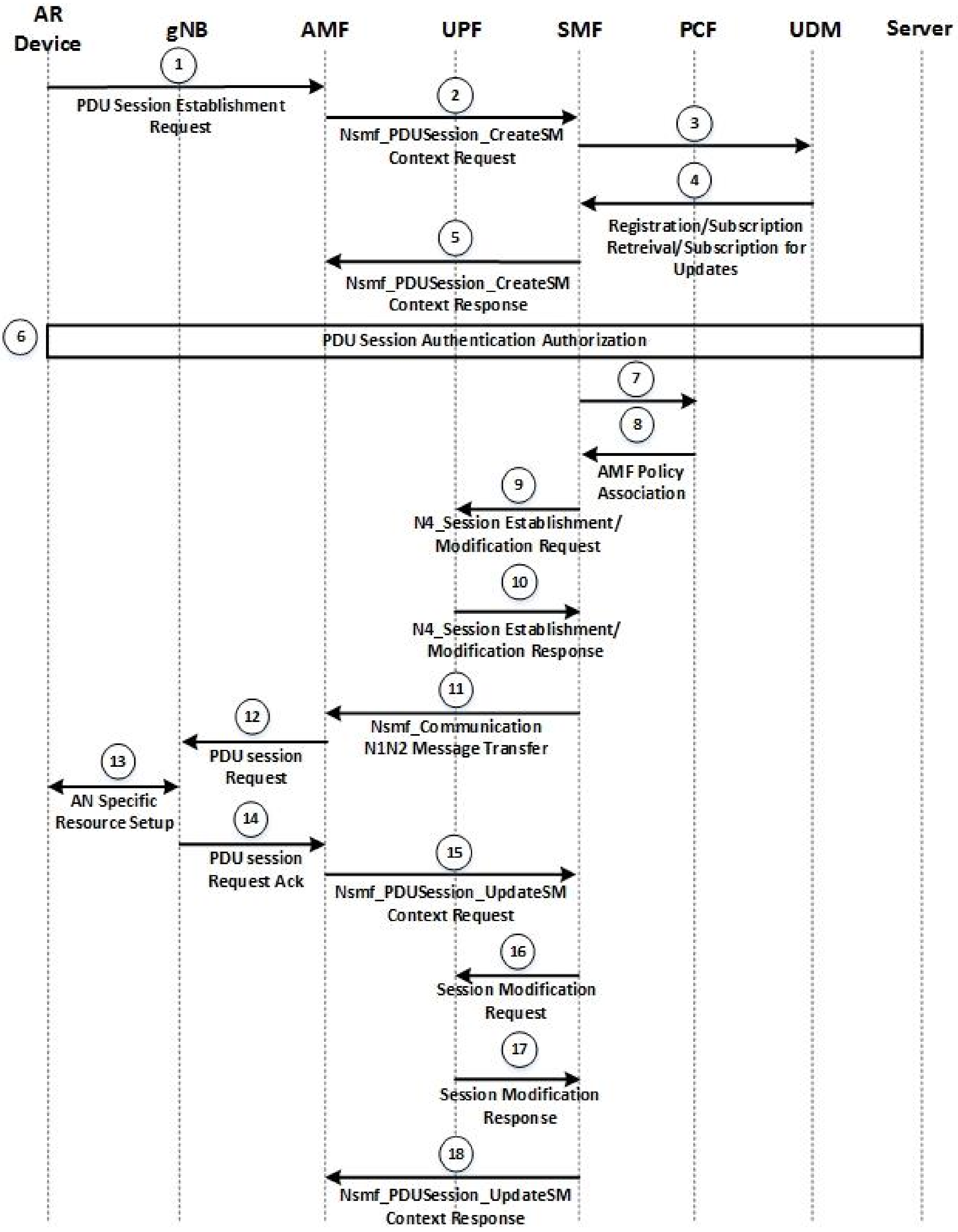

AR Device
gNB
AMF
UPF
SMF
PCF
UDM
Server
1
PDU Session Establishment Request
2
Nsmf_PDUSession_CreateSM Context Request
3
4
Registration/Subscription Retreival/Subscription for Updates
5
Nsmf_PDUSession_CreateSM Context Response
6
PDU Session Authentication Authorization
7
8
AMF Policy Association
9
N4_Session Establishment/ Modification Request
10
N4_Session Establishment/ Modification Response
11
Nsmf_Communication N1N2 Message Transfer
12
PDU session Request
13
AN Specific Resource Setup
14
PDU session Request Ack
15
Nsmf_PDUSession_UpdateSM Context Request
16
Session Modification Request
17
Session Modification Response
18
Nsmf_PDUSession_UpdateSM Context Response

Figure 16. Message sequence chart for session establishment between AR device and server.

Using the network functions described in the above three steps, the derived architecture to cater the communications of AR use case is illustrated in Figure 17. Network functions towards the right of the image is not used but illustrated only for the sake of clarity.

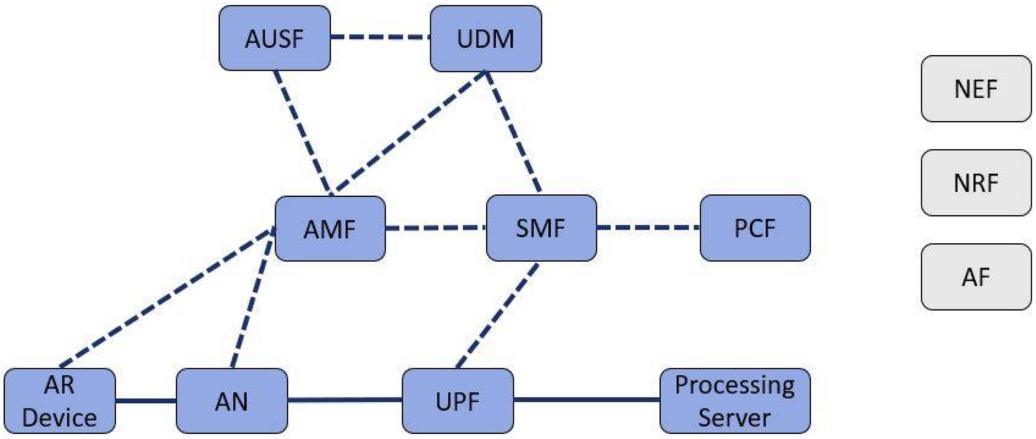


Figure 17. Architectural components of uO to cater AR use case.

Network functions that are not used in the architecture are Network Exposure Function (NEF) which handles masking of network and user sensitive information to external Application Function’s (AF) according to the network policy, Network Repository Function (NRF) and AF.

3.2 uO Architecture for Massive Wireless Sensor Networks

As explained early under the use cases, the objective of having wireless sensor networks to monitor a factory environment is to detect any anomalies, decide appropriate action to be taken after considering the severity of the malfunction. These actions can make sure that the factory operation can continue without interruption or prevent disasters which affect the safety of the factory environment and the workers.

When designing a communication system for a wireless sensor network in a future smart factory, stringent latency is required. Table 5 outlines the latency requirements of each Industry 4.0 sensor network communication scenarios. Future 5G systems should be designed to cater these latency requirements. In a 5G connected wireless sensor network, all the sensors, machines, centrally located management servers and controlling instances should be considered as elements of the 5G system.

As discussed before for the AR use case, the communication solution can be deployed either by an MNO or by a locally established uO within the factory premises. However, to make sure that the low latency requirements are achieved, the preferable deployment option would be to establish the 5G connectivity using a uO located inside the factory premises. In this case all the core network components of the uO are also located inside the factory, minimizing the transport delay of communications to achieve low latency requirements.

Figure 18 depicts how a locally deployed 5G operator can support the communication of the wireless sensor network. Multiple machines (actuators) and sensors connect to the nearest gNB and those gNBs are connected to locally deployed uO core network via routers. Alarm management server which analyses all the data from sensors is also connected via gNB and has connectivity to the core network. Server analyses the severity of the malfunction and sends the data to controlling server. Controlling server executes emergency action on the relevant machine. All necessary network functions are implemented within the local core network so that the low latency communications can be achieved.

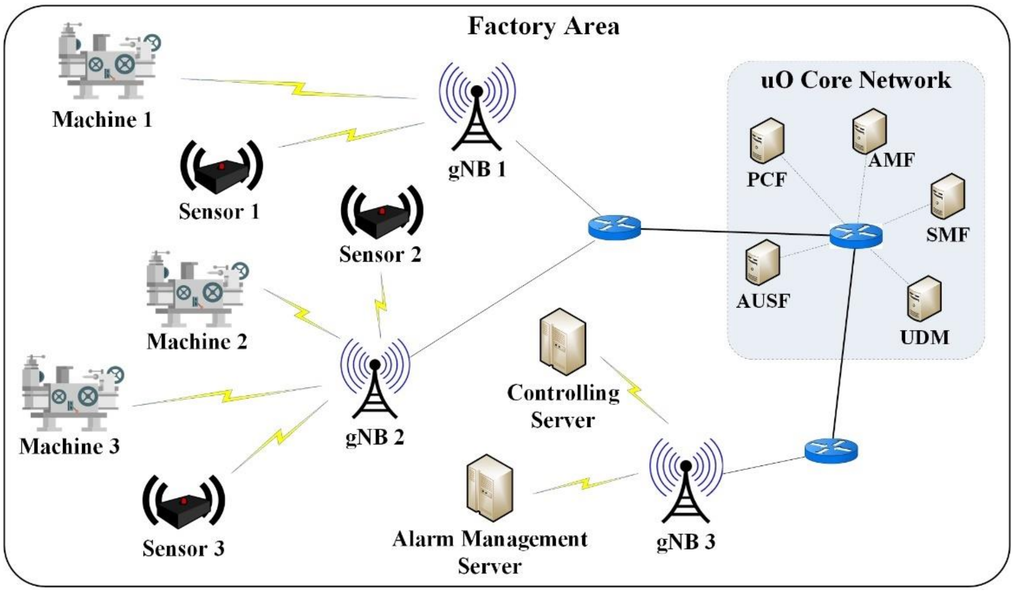


Figure 18. Wireless sensor network covered by 5G uO deployed within factory.

If the sensor network communication is catered by an MNO, the core network will be located outside the factory premises. A deployment model by an MNO is depicted in Figure 19 where the sensor network communication happens via MNO networks and its core network.

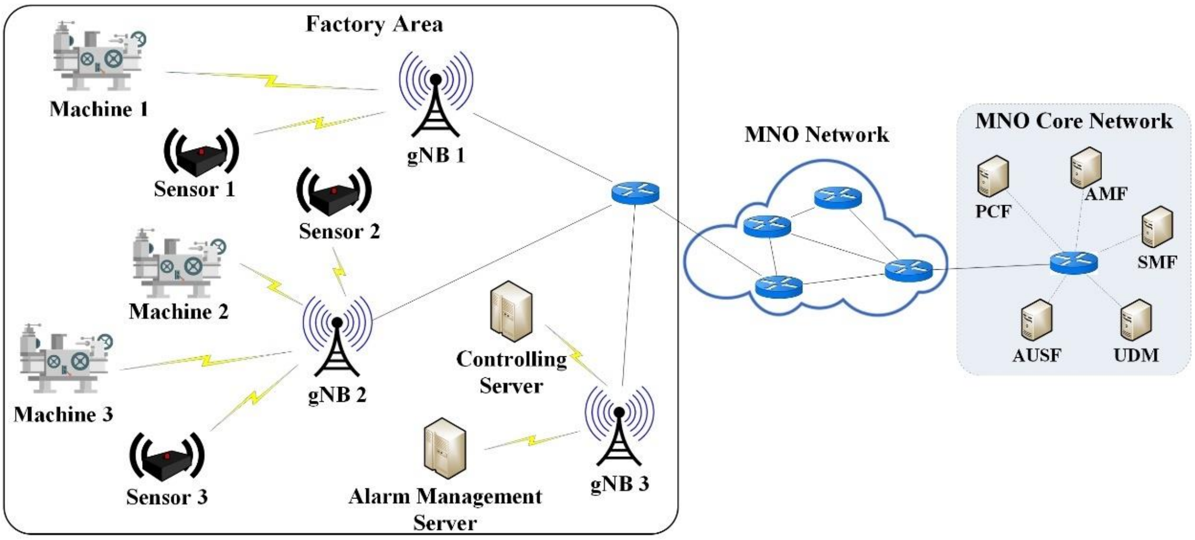


Figure 19. Sensor network communication catered by MNO.

Similar to AR use case, to decide the architectural components needed in industrial wireless sensor networks, it is necessary to identify the communications occurring in a typical wireless sensor networks scenario. Generally, following steps of communications should be supported by 5G system.

- Registering the sensors, actuators and servers in to the 5G network. After this step, everything in the network is identified as 5G network elements.

- Establishing data connectivity from the sensor devices to the alarm management server. Once this is done, sensors can send their sensed data to the alarm management server periodically or incident basis.
- Once an anomaly is detected an analysed by the management server, a data session should be established between the management server and the controlling instance.
- After this is done, management server can send data to the controlling server.
- Control server creates a session with the respective actuator on which the controlling action should be taken.
- Send data to the actuator from the control server to take the appropriate action.

We define the message sequence for the sensor network communication scenario with the aid of 3GPP specifications. Figure 20 illustrates the message sequence diagram for a typical sensor network communication. All the sensors, machines, alarm management server and controlling server has to be registered in the 5G network. This registration process employs the comprehensive message sequence illustrated in Figure 15, where the terminal devices in this case being the sensors, machines or servers. Then the sensor detects and abnormal behaviour and communicates it to the alarm management server. For this to happen, the sensor requires a PDU session to be established with the alarm management server. This happens according to the message sequence illustrated in Figure 16 and the initiator of the sequence is the sensor. Once the PDU session is established, sensor transfers the data to the alarm management server. The server then analyses the severity of the anomaly and establishes a PDU connection with the controlling server, again using the message sequence illustrated in Figure 16. Once this is done, alarm management server successfully transfers the data to the controlling server. Controlling server decide which action to take on which machine by processing the data. Then the controlling server establishes a PDU session with the relevant machine. Next, controlling server transfers the data to the relevant machine completing the communication process. Relevant action will be executed on the machine after this process, for example emergency stop.

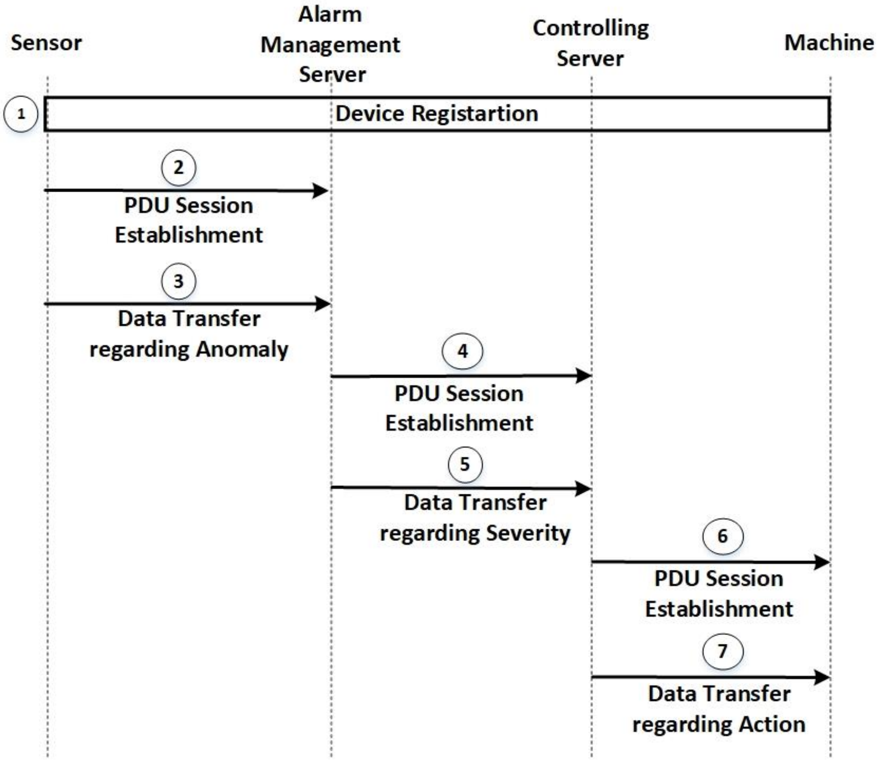

Sensor
Alarm Management Server
Controlling Server
Machine
1
Device Registartion
2
PDU Session Establishment
3
Data Transfer regarding Anomaly
4
PDU Session Establishment
5
Data Transfer regarding Severity
6
PDU Session Establishment
7
Data Transfer regarding Action

Figure 20. Message sequence chart for emergency actioning based on sensor detection.

Using the message sequence depicted in Figure 20 we designed the uO architecture for the sensor network communication. For the device registration process AN, AMF, PCF, AUSF and UDM network functions are needed. For the PDU session establishment process AN, AMF, UPF, SMF, PCF and UDM network functions are needed. For the data transfer process AN and UPF network functions are used. Therefore, the architecture should comprise all the above mentioned network functions to support the communication. Hence, we can use the network architecture depicted in Figure 21 to support the sensor network communication, which is the same architecture we used for AR communication.

However, compared with AR use case the amount of user plane data transfer differs in wireless sensor network use case. As an example, in AR use case, continuous video transfer requires high data rate but for sensor data transfers, much higher data rate is not required. So, the capabilities needed in UPF network function may differ for the two use cases. In other words, the user plane network functions can have less resources in sensor network use case. Network functions towards the right of the image is not used but illustrated only for the sake of clarity as in previous AR use case.

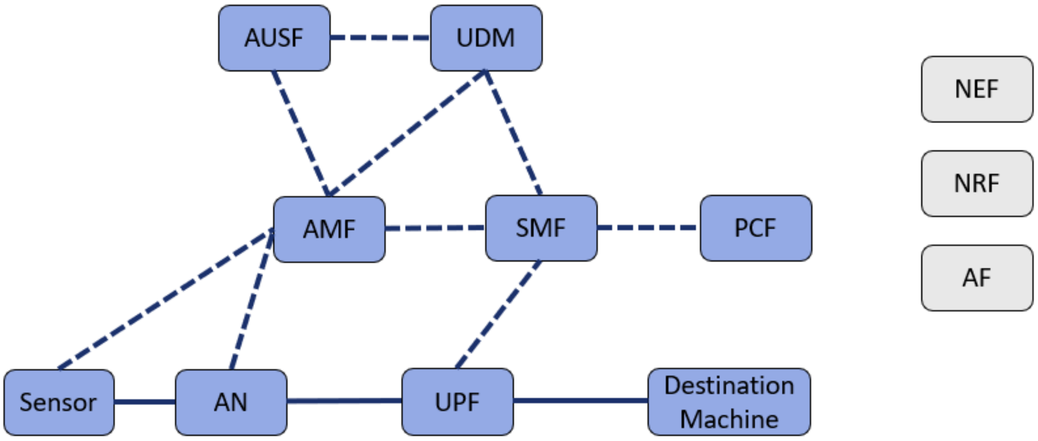


Figure 21. Architectural components of uO to cater sensor network use case.

3.3 uO Architecture for Mobile Robots

With the increase of mobile robots in future factories, the robots themselves must have the ability to take decisions on their own by analyzing the data they receive from the sensors in the environment, the data they receive from the other robots and the data they receive from the guidance control system. As in the previous AR and sensor network use cases, the 5G network catering the mobile robot communication can be deployed either by a local 5G operator or by an MNO. These two deployment models are depicted in Figure 22 and Figure 23 respectively.

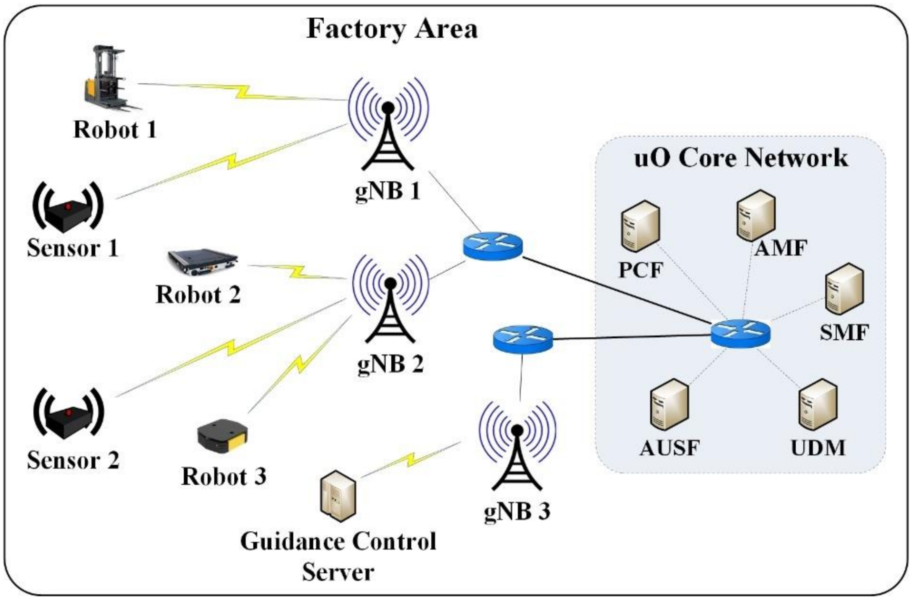


Figure 22. Mobility of robots use case covered by 5G uO.

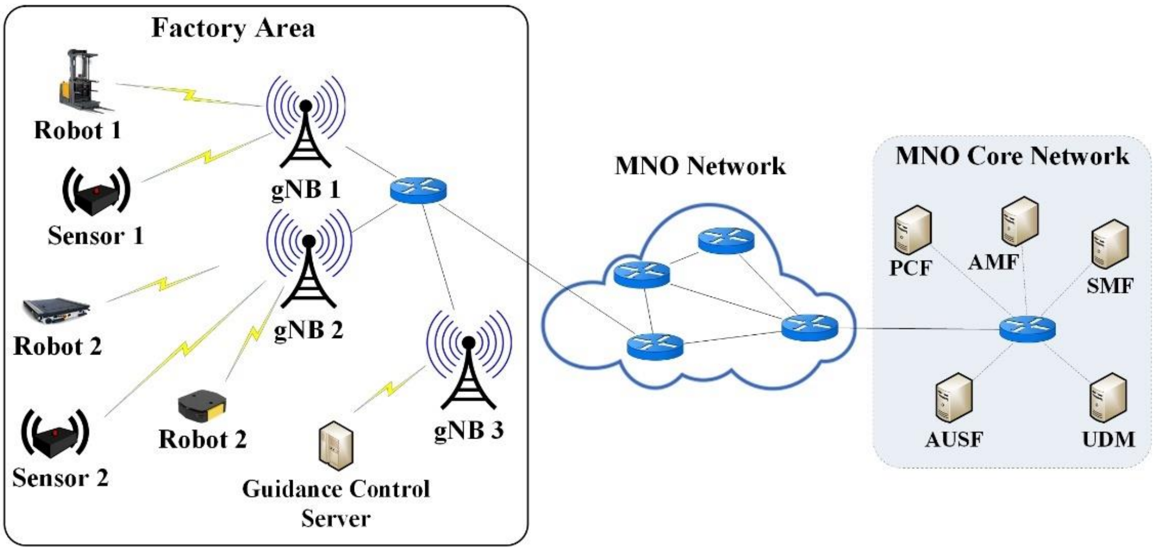


Figure 23. Mobility of robots use case covered by MNO.

In Figure 22, the environment sensors and the mobile robots are connected to the nearest gNB and those gNBs are connected to the locally deployed uO core network via routers. The guidance control server is also connected via gNB and has connectivity to the core network. Guidance control server processes the data from mobile robots for control and management purposes. The server can command a mobile robot to a stop in case of any emergency.

In the case of MNO implementing the network to cater the communication needs of mobile robots, the core network will be located outside the factory premises. Therefore, this deployment will not support stringent the latency requirements as uO does. To make sure that the low latency requirements are met, the preferable deployment model would be the uO approach.

Following communication scenarios will be seen in Industry 4.0 mobile robots use case.

- Registering the environment sensors, mobile robots and server to the 5G network. After this step everything in the network is identified as 5G network elements.
- Establishing data connectivity between mobile robots to the guidance control server. Once this is done, robots can send their data to the guidance control server either periodically or incident basis.
- Robots can establish connections between other robots and the sensors.
- Once all the connections are established robots can perform their regular operation.
- If the guidance control server needs to take any controlling or management action on the robots, it can send the data via the session which is already established.
- Robots perform frequent handovers due to their mobility.
- After moving to a different location, if it is necessary, the robot must register in the network again and the communication scenarios repeat from step 1.

Regular operation of mobile robots requires registration in the 5G network and become the network elements of 5G, establishing PDU session between the relevant devices and the data transfer process between the devices. We already identified the network functions needed for these operations under previous two use cases using the message transfers depicted in Figure 15 and Figure 16. Hence, in this case also, it is mandatory to have AN, AMF, PCF, AUSF, UDM, UPF and SMF.

However, the main difference with mobile robots is that they frequently move from one place to another. With a high density of mobile robots, assuring the seamless mobility is one of the challenges to the 5G system connecting the robots. In addition to that proper 5G coverage has to be there throughout the entire factory environment for the functioning of robots. Therefore, the handover process has to be done seamlessly even if there are a higher number of robots. Also, we need to consider the network functions needed additionally for the handover process and we need to include them to derive the architecture that supports mobile robot operation. We consider the message sequence defined by 3GPP depicted in Figure 24, to identify the network functions for the handover process [38].

The handover process execution is initiated by the source gNB forwarding the data to the target gNB. Target gNB then sends path switch request to AMF. Once AMF receives the path switch request, AMF sends a session update request to SMF and then SMF sends the session modification request to UPF. UPF then sends the session modification response to back to SMF. In the meantime, UPF sends the end marker packets to source gNB and it also sends the downlink packets to mobile robot via target gNB. After that SMF sends the PDU session update response to AMF, and AMF then acknowledges the path switch request to target gNB. Then the target gNB sends release resources message to source gNB completing the handover process.

To cater the communication of the handover process, we need AN, AMF, SMF and UPF network functions. Therefore, the network architecture comprising all the network functions needed to cater the mobile robot communication can be depicted as in Figure 25. Network functions towards the right of the image is not used but illustrated only for the sake of clarity.

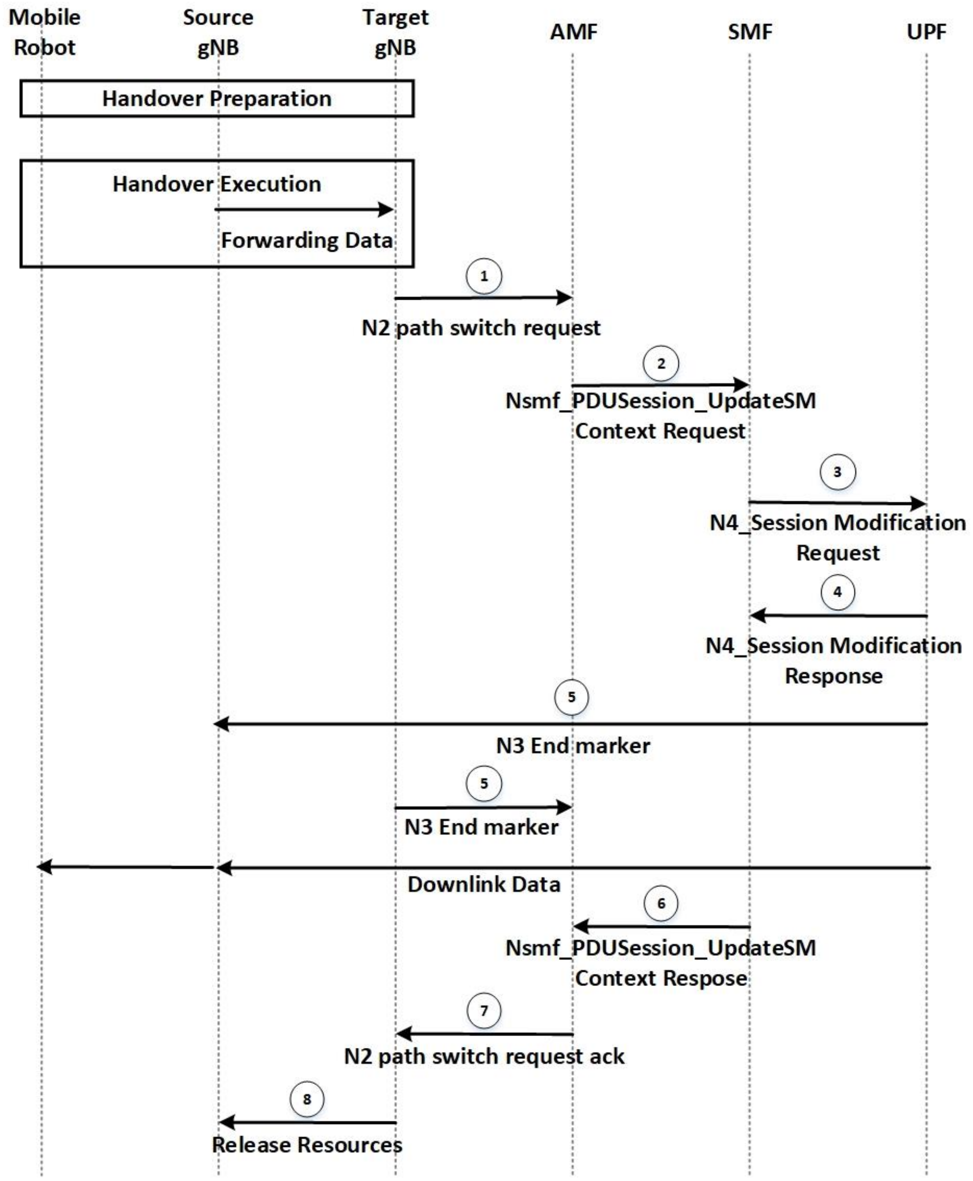


Figure 24. Message sequence chart for mobile robot handover procedure.

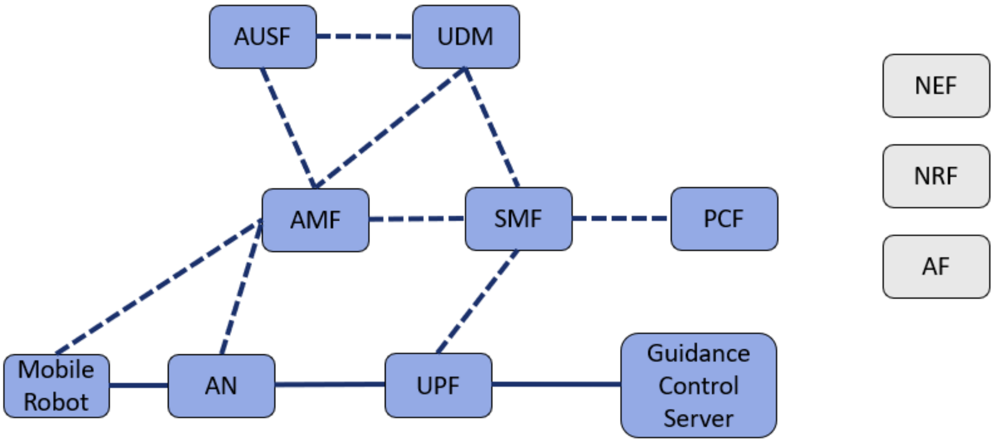


Figure 25. Architectural components of uO to cater mobile robots use case.

3.4 Proposed uO Architecture

In the previous sections, we discussed about three architecture options for the three use cases we selected. However, it was noticed that for certain use cases like AR which includes transferring video from AR devices to the image processing server, the user plane network functions such as UPF must have higher resources. On the other hand, the sensor network communication may not transfer that much data around the network. Therefore, to optimally serve each use case and efficiently achieve the required low latency requirements, establishing logical networks for each use case is a preferable option. Having logical networks provides the flexibility in resource management in network functions for each use case separately. It also provides isolation among the use cases. This is achieved by implementing network slices to serve the communication requirements of each use case.

3.4.1 Use of Network Slicing

Network slicing proposes a way to create logical networks on a common infrastructure to enable different types of communication services [39], [40]. When uO uses three network slices to cater the three use cases, uO must create network slices before any actual communication happens over a selected slice. 3GPP introduces three network slice management functions for creating and managing network slices [41] [42].

- Communication Service management Function (CSMF)
  Responsible for translating communication service related requirement to network slice related requirements.
- Network Slice Management Function (NSMF)
  Derive network slice subnet related requirements from network slice related requirements and, responsible for management and orchestration of NSI.
- Network Slice Subnet Management Function (NSSMF)
  Responsible for management and orchestration of Network Slice Subnet Instances (NSSI).

An illustration of using different network slice instances to cater the three use cases is given in Figure 26. In the figure, there are two access network slice subnet instances namely AN NSSI 1 and AN NSSI 2 and three core network slice subnet instances CN NSSI 1, CN NSSI 2 and CN NSSI 3. The logic of creating the network slice subnet instances is based on the requirement of the services, based on available resources with the operator or based on operator policies.

Figure 26 depicts that the AR use case is supported by NSI A which is created using CN NSSI 1 and AN NSSI 1. On the other hand, sensor network use case is catered by NSI B which is a combination of CN NSSI 2 and AN NSSI 2. Hence, logically isolated resources are used to serve the two use cases. It can also be seen that the NSI C and NSI B shares AN NSSI 2 but uses different CN NSSIs. This is an illustration of the concept of re-using the NSSI. This is obviously depending on the real resource requirements of the use cases but Figure 26 illustrates how different network slices can be formed and how they can be used to server different use cases.

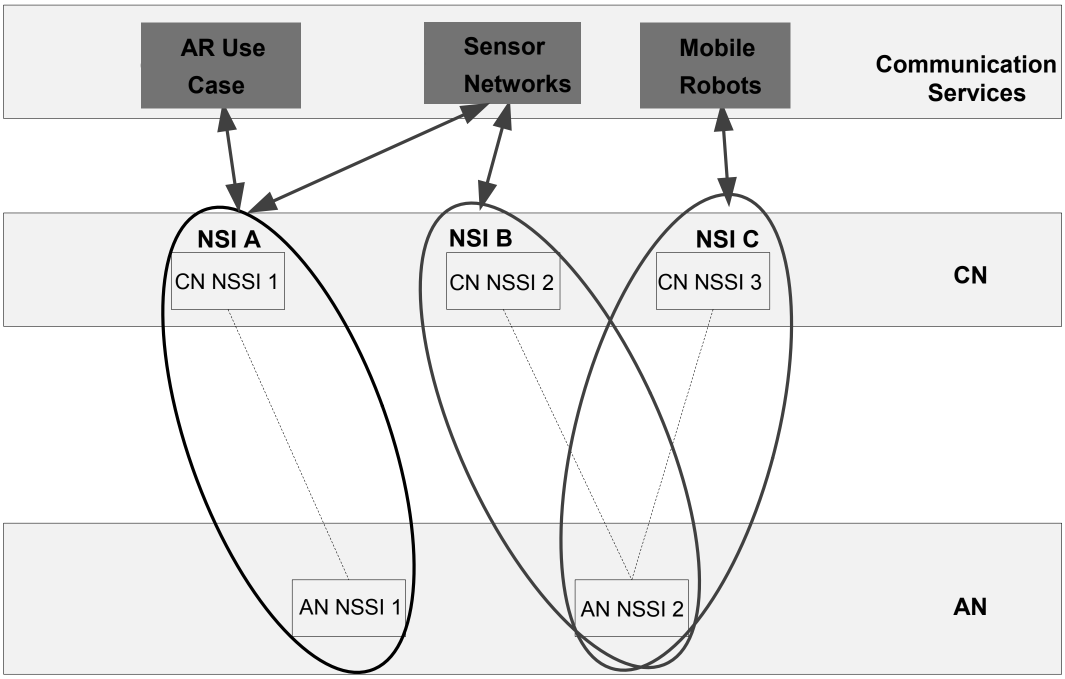


Figure 26. Service provisioning via different NSI [41].

3.4.2 uO Architecture

Therefore, to simultaneously cater AR, sensor network and mobile robots use cases, uO needs to create the NSIs before the communication begins. To create the network slices, three network slice management functions CSMF, NSMF and NSSMF should also be there in the uO architecture. These three network functions allow the communication service requirements to be translated to network slice requirements, and then those network slice requirements to be translated to network slice subnet requirements and ultimately the network slice subnets can be created. These network slice subnets are then used to create the network slice instances.

Apart from that, the best fitting slice must be selected before the communication begins. This is done by Network Slice Selection Function (NSSF), which is an obligatory element in the uO architecture that supports communication over multiple network slices. Combining these concepts, final architectural components to support AR, wireless sensor network and mobile robots use cases can be derived. Figure 27 represents the derived uO architectural components.

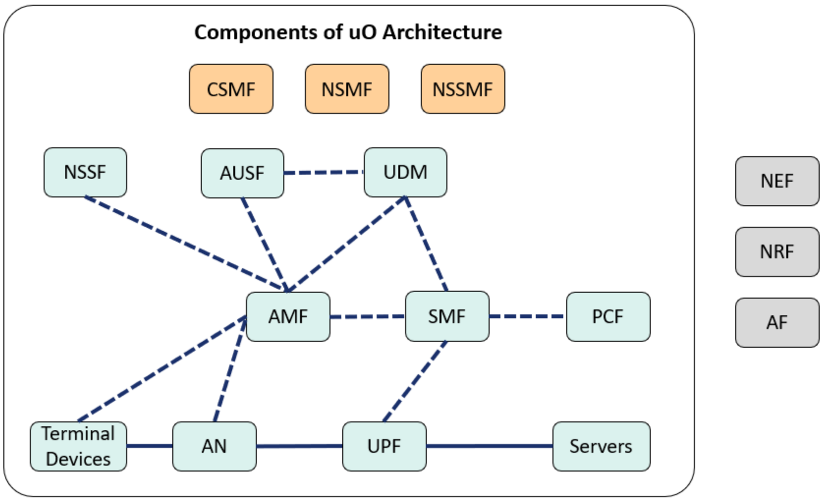


Figure 27. Proposed uO architectural components.

Even though Figure 27 illustrated the complete architectural components, it does not illustrate the concept of network slicing embedded in the architecture. The final uO architecture comprised with three network slices to cater communications of three uses is shown in Figure 28. The three slice management functions which translate the communication requirements to network slice subnet requirements and NSSF which selects the best fitting slice and their interaction to the core network is clearly depicted in Figure 28.

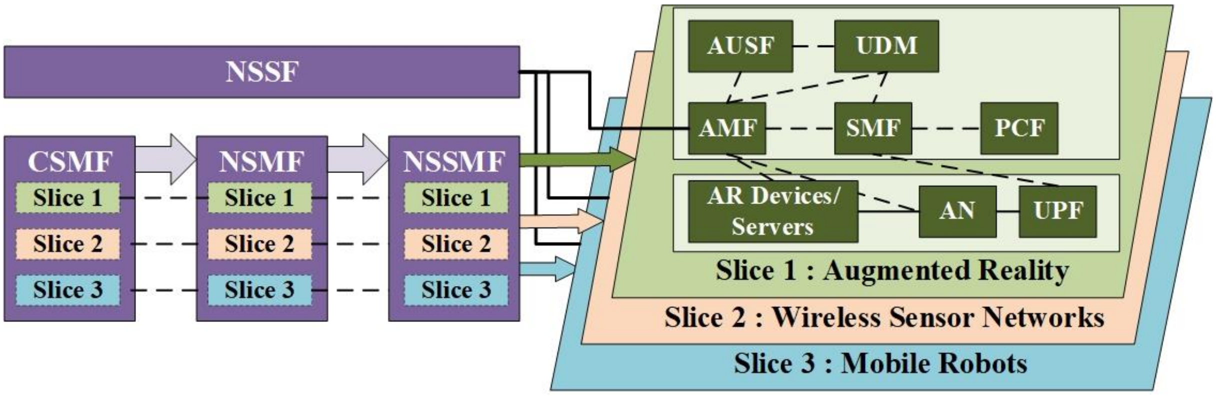


Figure 28. Proposed uO architecture.

4 EXPERIMENTAL SETUP

In this chapter, we explain the experimental setup in detail. First it discusses the deployment models used. The thesis considers two deployment models, the uO based model and the MNO based model. In the uO based model, we consider that the factory is served by a uO deployed local 5G network. In the MNO based model, the factory is covered by a 5G network implemented by MNO. Next subsection focuses on different simulation platforms we used to conduct simulations. Each simulation platform is explained briefly. Common simulation parameters are presented next along with their sources. Latency is considered as the output performance measure. Finally, this chapter presents the details of the real world implementation of a uO network using the Kuha base station and the Cumucore EPC.

4.1 Proposed Deployment Models

uO based deployment model is depicted in Figure 29 and the MNO based model is depicted in Figure 30. In the uO based deployment model, all the terminal devices such as machines, sensors, mobile robots, AR devices are connected to the gNB via 5G connectivity. In addition to that servers such as image processing server, guidance control server, alarm management server connects to another gNB. Those gNBs are connected to the uO core network via routers. Factory owns all the terminal devices and the servers. uO core network is located within the factory premises.

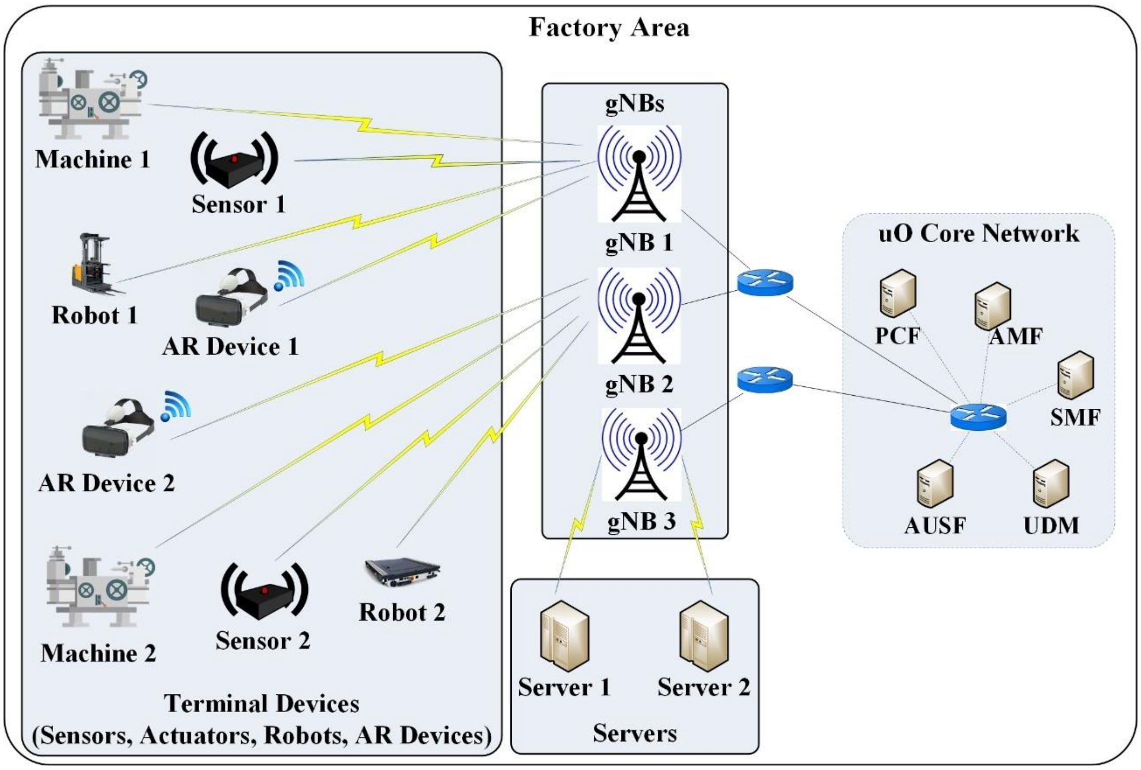


Figure 29. uO deployed 5G network to serve the factory.

In the MNO base deployment model all the terminal devices connect to gNBs via 5G connectivity, and those gNBs then connect to a gateway router. This gateway router establishes the connectivity to the MNO network which then connects to the MNO core network. Factory owns all the terminal devices and the servers. In this case, the core network is located outside the factory premises.

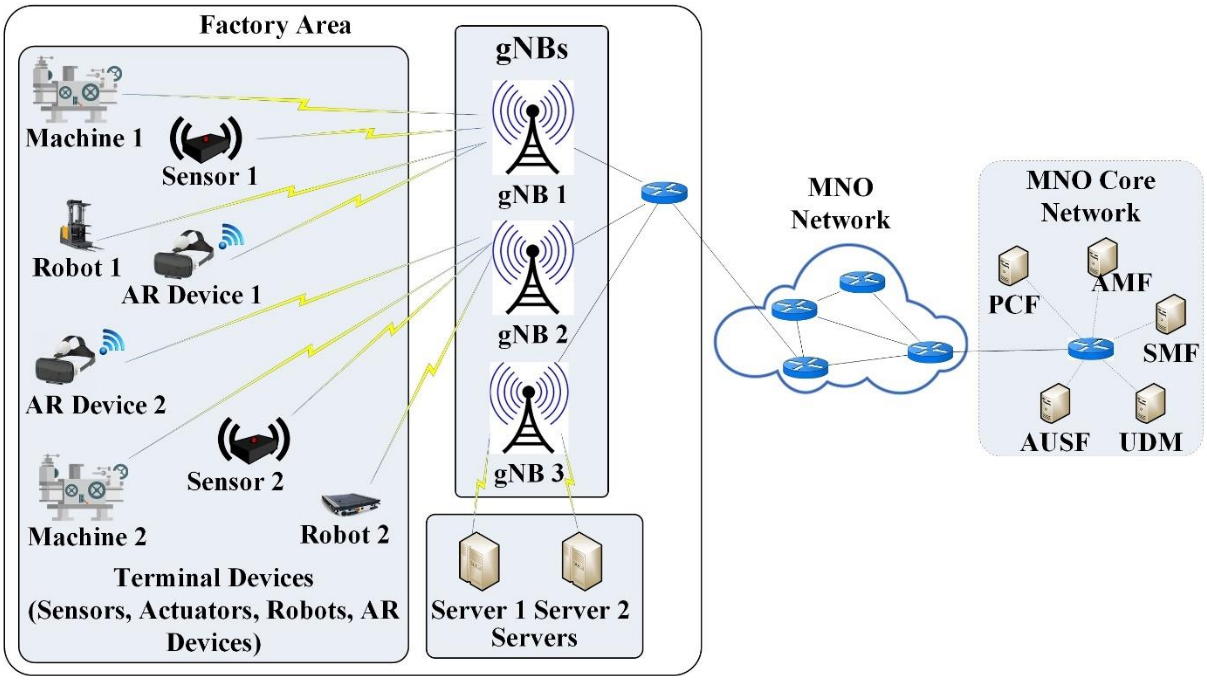


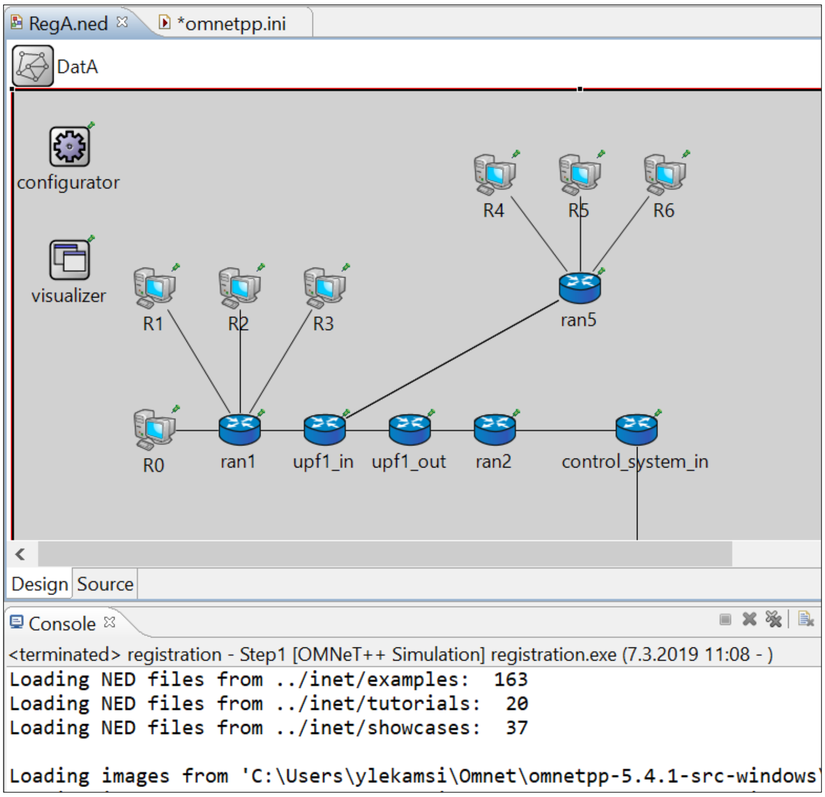
Figure 30. MNO deployed 5G network to serve the factory.

In this model, we consider the MNO is simultaneously serving total of N such factories having those Industry 4.0 use cases as depicted in Figure 31. Each factory is assumed to be having similar network setup and similar requirements. Without loss of generality, we have assumed that uO's processing power is $1/N$ of the possessing power of an MNO. It validates the fact that MNO's resources are equally divided among N uOs in each factory.

Figure 31. MNO's service for N factories having Industry 4.0 use cases.

4.2 Simulation Platforms

As the simulation platform, we used Omnet++ framework to generate the results. Omnet++ is a modular, component based C++ simulation library and framework [43]. INET Framework is an open-source model library for the OMNeT++ simulation environment [44]. INET provides the protocols for communication networks such as TCP, UDP, IPV4, OSPF, wired and wireless link layer protocols and several application layer protocols. A screenshot of Omnet++ is depicted in Figure 32.



<terminated> registration - Step1 [OMNeT++ Simulation] registration.exe (7.3.2019 11:08 -)

```
Loading NED files from ../inet/examples:  163
Loading NED files from ../inet/tutorials:  20
Loading NED files from ../inet/showcases:  37

Loading images from 'C:\Users\ylekamsi\Omnet\omnetpp-5.4.1-src-windows
```

Figure 32. Omnet++ simulation environment.

4.3 Simulations and General Parameters

The objective of the simulations is to measure the latency advantage that the factory can achieve by implementing the communications using 5G uO rather than implementing it with an MNO. We model below three communication procedures for AR use case. We discussed these procedures in detail in Chapter 3.1.

- Registration procedure of terminal devices (AR devices, image processing server) into the 5G network. In this procedure, we model all the information flows from the AR device to AN, communication between the AN and the 5G core and back, and the information flow between the network functions which is needed for the registration process. After this procedure, all the devices are identified as 5G network elements.
- PDU session establishment procedure illustrates how AR device can establish a data session between the AR server before any actual data transfer happens. All messages in session establishment process have to be considered.

- Data transfer process from AR device to the image processing server and back to the AR device. AR data stream travels through UPF of MNO/uO core via the 5G AN and then routes back to the image processing server inside the factory. Server processes the data, apply augmentations on the video frames and sends those augmented frames via UPF back to the AR device.

For the massive wireless sensor network case we simulate the following series of communications which corresponds to an emergency action taken in the case of abnormal behaviour detected by sensor data. Similar to the previous use case, all the information flows needed to model this scenario has been considered.

- PDU Session establishment process from sensor node to alarm management server
- Transferring data to the alarm management server from sensor node
- Alarm management server detects the abnormal behaviour and establishes a PDU session with controlling server
- Alarm management server sends data to controlling server
- Controlling sever analyses data and establish PDU session with the relevant terminal node
- Controlling server sends data to terminal node which includes the emergency action to be taken

For the mobile robots use case, handover process is also a key process because there will be lot of handovers due to robot mobility. Therefore, we first analyse the handover process. However, the critical communication is the data transfer between mobile devices and the controlling server. Therefore, we also model the data transfer process between terminal nodes. For each case, we take the latency data to analyse the benefits of having uO deployed 5G network rather than having MNO deployed 5G network for these Industry 4.0 use cases.

Parameters common for all the experiments are outlined in Table 6. Parameters that vary in each experiment are mentioned at the appropriate experiment. Acces network latency is based on the 3GPP study on next generation access technologies [45]. We assumed that the access networks deployed by both uO and MNO have similar properties and therefore having the same latency. We take all the backhaul connections as a fiber connections and the latency of the fiber backhaul is selected based on a study of 5G backhaul challenges [46]. Latency of image processing is based on 3GPP study on communication for automation [8]. As the performance measuring parameter, we take the E2E latency. For each experiment, we measure the E2E latency to produce the comparison results.

Table 6. General simulation parameters

Parameter	Value
Latency between AR device and AN	0.5 ms [45]
Latency between AN and core network	0.05 ms per km [46]
Image processing server delay	30 ms [8]
Distance to uO core network	500 m
Number of factories served by MNO (N)	10

4.4 Implementation of a uO setup with Kuha Base Station

A real-world implementation of a local uO has been carried out using a Kuha base station and Cumucore EPC [47] [48]. The implementation setup supports maximum of ten users connecting to Kuha base station and the packet core network is implemented by Cumucore. Kuha mobile network is a low-cost minimum effort solution to extend mobile networks. Low cost is one of the main benefits of Kuha mobile network because there are no infrastructure costs and the power is also provided by the end user. It is a plug and play device and in case of any troubleshooting or management activity, it can be remotely taken care by Kuha with the help of an internet connectivity.

The implementation contains one Kuha base station which can act as a local uO along with the operation of Cumucore EPC. This network enables direct data communication between the ten devices attached to the base station if they have IP addresses configured. If there is any internet connectivity requirement, the EPC can be connected to external network which provides the internet connection. In our implementation, we used a 4G LTE router having a wired connection to Cumucore EPC and then the router was wirelessly connected to 5G test network (5GTN) in University of Oulu via Band 7. With this setup, each device can access internet via 5GTN. Implementation of the network setup is depicted in Figure 33.

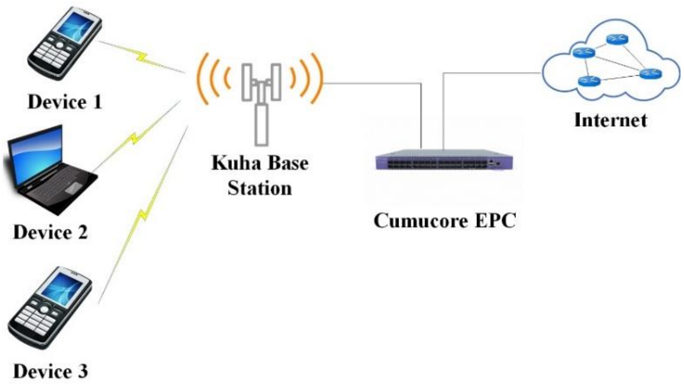


Figure 33. uO Implementation with Kuha base station and Cumucore EPC.

Using Kuha based uO implementation, we performed experiments and measured the E2E latency and the throughput of the communication between devices. First, two Subscriber Identity Module (SIM) cards provided with Kuha base station were used to setup a local network. DHCP protocol was enabled in Cumucore EPC so that the devices will automatically acquire IP addresses. One device was used as an iperf server and other one as iperf client. By sending an iperf command from the client to server, we measured the throughput. We used ping command to measure the latency. This is similar to a locally deployed uO setup where the core network is also placed local. The implementation is depicted in Figure 34.

As the second experiment, we used commercial SIM cards available in the market to perform the same experiment. In this case, the devices connect to the internet and the communication is supported by the MNO core network. We carry out the experiment using DNA, Telia and Elisa. Latency and throughput measurements were taken to compare the performance of uO implemented communication service vs. MNO implemented communication service.

During this experiment, we faced few technical difficulties and we were able to overcome some of them. Cumucore EPC provides IPV4 addresses to its clients. Using those IPV4 addresses, clients can communicate with each other. Hence, both iperf and ping commands were successfully executed for Kuha uO network. But for the MNO case, the IPV4 addresses provided by the MNO are not global IP addresses. Hence it was not possible to execute either iperf or ping command using those IP addresses. The solution was to use IPV6 addresses in both mobile phones when they are connected to DNA network. Cumucore EPC does not provide IPV6 addresses. But there should not be any performance difference in using IPV4 addresses for Kuha experiments and IPV6 addresses for MNO experiments. The next problem was that Telia does not provide IPV6 addresses therefore the experiment was not carried out for Telia at this moment. Even though Elisa provides IPV6 addresses both iperf and ping commands are not executing via Elisa network. The reason could be the operator policy to block ICMP traffic. Hence, we carried out the experiments for Kuha network and only for DNA network and compared the results.

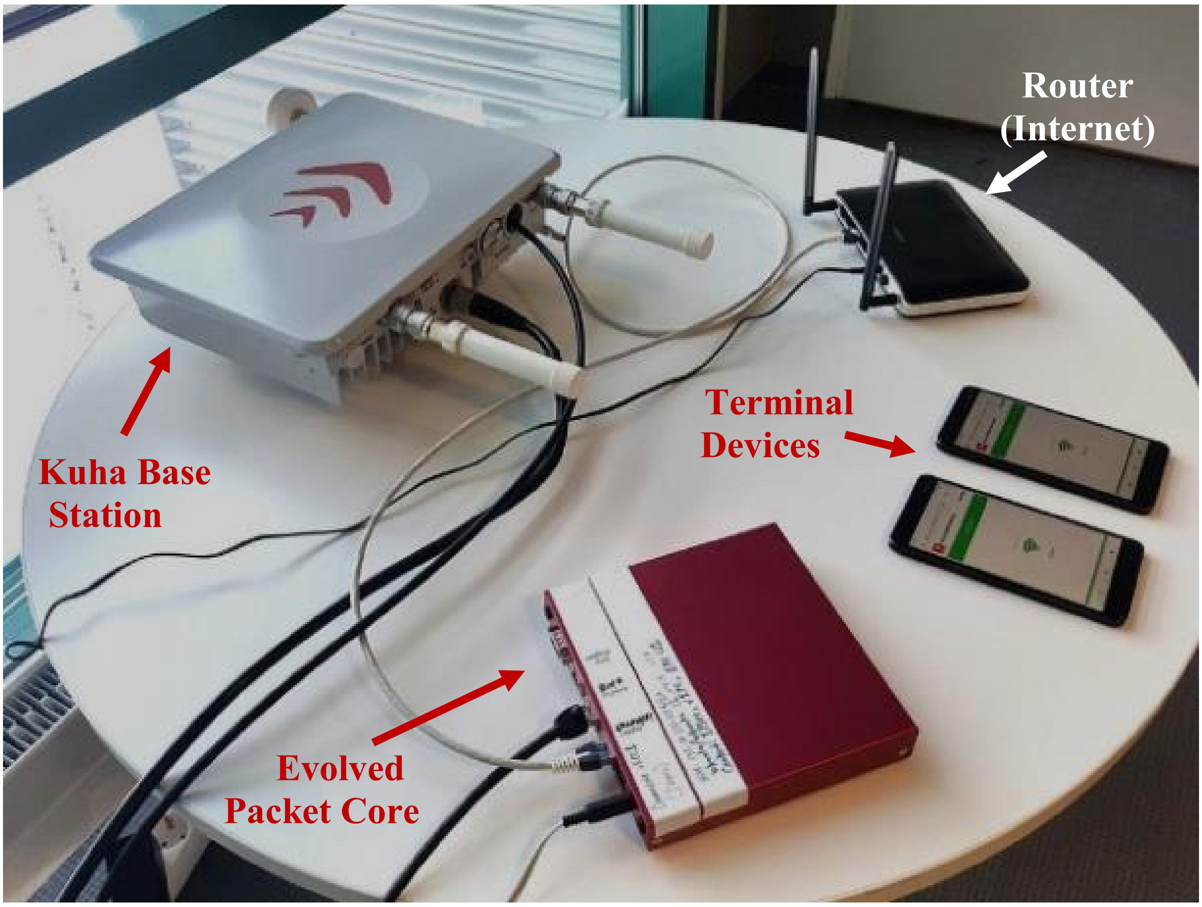


Figure 34. uO network implementation with Kuha base station and two mobile devices.

5 RESULTS ANALYSIS

This chapter provides the analysis of simulation results for each use case. All three use cases AR, massive wireless sensor networks and mobile robots were considered for the analysis. Each use case involves a set of communication steps like registration, session establishment and data transfer. Each step is analysed for the AR use case and the most obvious steps are not analysed for the remaining use cases. The performance measurement is the latency and we analyse the latency with respect to two independent parameters called distance to core network and the network function processing delay. Each analysis consists of a comparison between the uO based deployment model and the MNO based deployment model. We consider that the three use cases are served by three network slices, therefore the NF resources being used by the three use cases are in different logical networks. This enables us to conduct independent analysis for each use case. We also discuss how MNO can establish the core network at a certain distance from the factory location, by increasing the core network resources.

Results analysis also consists of a comparison of three uses cases considering the most critical communication steps and which parameter plays a dominate role deciding the latency, whether it is the distance to core network or the network function processing delay. Finally, it presents the results of the experiments conducted using the Kuha base station setup.

5.1 AR Use Case

5.1.1 AR Device Registration Process

First, we simulate the AR device registration process for MNO deployed model and compare it with uO deployed model for latency measurements. We vary the MNO core network distance from 500 m to 500 km in 50 km intervals. uO core network distance is kept constant at 500 m justifying the fact the uO core network is located inside the factory premises. We use the parameters presented in Table 6 and we take the network function processing delay of uO as 1 ms and network function processing delay of MNO as 0.1 ms. This is because generally, MNO has more resources than uO which will enable faster processing in MNO core network than uO. Result of this experiment is illustrated in Figure 35. Latency of uO is illustrated in Figure 35 for the purpose of comparing the results with the MNO.

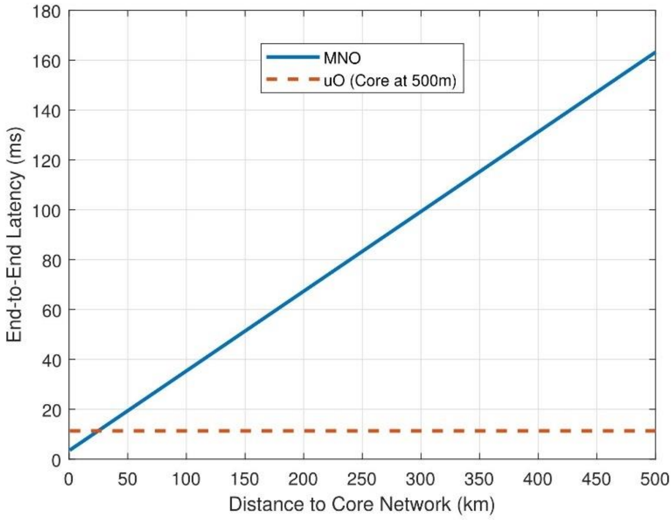


Figure 35. E2E latency of AR device registration process with respect to distance.

It is seen from Figure 35 that the E2E latency shows a linear increase with respect to the distance to the core network. Increment is approximately 0.32 ms per 1 km. AR device registration process has multiple round-trips communications towards the core network and back to the device. This communication depends on the distance to core network and when the distance is higher, latency will also be higher. Even though we have taken the processing times at the network functions of MNO to be lower than the uO deployment model, it is difficult to achieve a low latency as uO, because the distance to core network contributes as a dominating factor to increase latency. Figure 35 also reveals that, for MNO to achieve the same latency given by the uO deployment, the core network of MNO should be located closer than 18.21 km. This is not always a viable implementation because usually the factories are most likely to be in diverse geographic areas.

As the second experiment, we examine the latency of the same process with respect to network function processing delay. During the registration process, the registration message must be processed by core network and access network functions multiple times until the registration process is completed. Therefore, having efficient network functions with higher resources can minimize the overall latency of the communication. For this experiment, we keep the uO core network distance as 500 m as the previous case and we also fix the MNO core network distance at 250 km, by taking the average of the distance range we considered for the previous experiment. Network function processing delay is varied from 1 µs to 1000 µs and the resulting E2E latency graph is depicted in Figure 36.

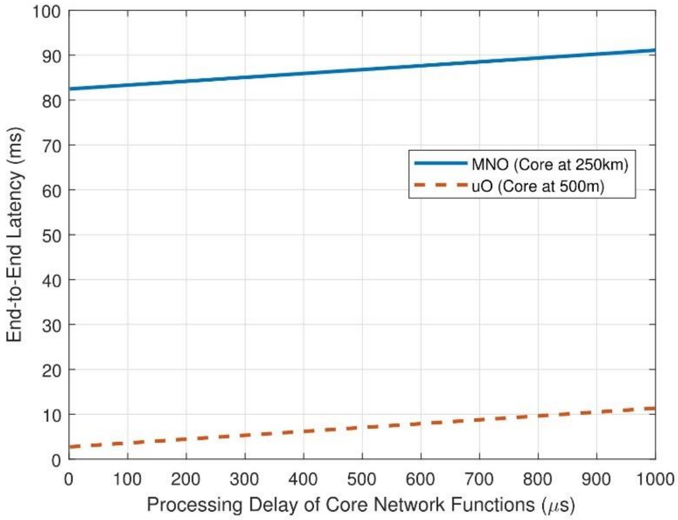


Figure 36. E2E latency of AR device registration process with respect to NF processing delay.

E2E latency exhibits a linear variation with respect to NF processing delay for both uO and MNO. Latency increase is approximately 8.63 µs when NF processing delay is increased by 1 µs. Unlike in the previous experiment, it is seen that uO always perform better than the MNO. This is because of the distance to core network being always higher in the MNO case than the uO case.

In the real deployments, uO only serves for the AR communication needs in the factory because it is providing a tailored service. Therefore, the communication over each slice is optimized for the use case served by that slice. On the other hand, MNO must serve the ten factories simultaneously. Moreover, it has to provide the service to its regular subscribers such as mobile broadband users. This places higher load on MNO core network than the uO core network. Therefore, there is a high probability of having high NF processing delay in MNO case compared with uO case.

According to the message transfer procedure, E2E latency for the registration process of AR devices can be expressed as follows,

$$L_{reg} = k_1.T_{access} + k_2.T_{backhaul} + k_3.T_{NF} \tag{1}$$

where L_{reg} is the E2E latency of the AR device registration process, T_{access} is the delay from AR device to access network, $T_{backhaul}$ is the delay from AN to core network, T_{NF} is the network function processing delay. k_1 and k_2 are the number of times where the registration message travels via the access channel and the fiber backhaul respectively. k_3 being the number of times which the registration message is processed through a network function. Since $T_{backhaul}$ directly proportional to the distance between access network and core network, we can express $T_{backhaul}$ as follows,

$$T_{backhaul} = k_4.D_{backhaul} \tag{2}$$

where $D_{backhaul}$ is the distance between access network and core network, and k_4 is a constant based on the properties of the fiber backhaul. Using equations (1) and (2), for a given L_{reg}, the distance between access network and the core network can be derived as follows,

$$D_{backhaul} = (L_{reg} - k_1.T_{access} - k_3.T_{NF})/(k_2.k_4) \tag{3}$$

Moreover, we can say that the network function processing delay varies based on two parameters, i.e. operator resources and network traffic load. The individual relationships between the NF processing delay can be expressed as follows.

NF processing delay is directly proportional to the network traffic load and inversely proportional to operator resources, yielding relationship (4).

$$T_{NF} \propto Network\ Load/Operator\ Resources \tag{4}$$

Therefore, we consider following four different cases of resource availability at MNO compared with uO.

- Case 1 - MNO resources = uO resources
- Case 2 - MNO resources = 10 × uO resources
- Case 3 - MNO resources = 100 × uO resources
- Case 4 - MNO resources = 1000 × uO resources

In these four cases also, we consider that the uO is serving only one factory and the MNO is simultaneously serving 10 similar factories. We take the network function processing delay of uO as 1 ms. For MNO, it varies based on the resource availability. For example, under case 1, NF processing delay for MNO will be 10 ms because its resources are shared among the ten factories. Therefore, under case 1, it is obvious that the MNO cannot achieve the same low latency as uO achieves. In case 2, where MNO has ten times the resources as uO, NF processing delay of uO and MNO will be the same. Hence, both MNO and uO have the similar performances in terms of latency. Therefore, MNO will also show the same latency as uO when its core network is 500 m away from the factory.

This behavior changes when there are higher resource levels with MNO. When the resource level is high at MNO, it can to locate the core network in a distant location and get the same latency as the uO, but not too far from the factory. For example, under case 3, when MNO has 100 times more resources than uO, MNO can establish the core network at 18.21 km away from the factory premises and still achieve the low latency requirement achieved by uO. Furthermore, when the MNO resource level is increased from 100x to 1000x of uO, MNO can establish the core network at 20.52 km away from the factory, somewhat far than case 3. We also see that the distant advantage from 10x to 100x is higher than the distance advantage from 100x to 1000x. This shows that even the MNO increases its resource to achieve insignificant NF processing delays, the advantage is diminishing. This is because the distance to core network is more dominant than the NF processing delay for these experiments. These results are outlined in Table 7.

Table 7. Distance to MNO core when uO T_{NF} = 1 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	10 ms	-
2	1 ms	500 m
3	0.1 ms	18.21 km
4	0.01 ms	20.52 km

We carry out two more scenarios with less uO resources to observe how MNO can establish the core network for those cases. Therefore, we take NF processing delay of uO as 10 ms and 100 ms for these scenarios. This is because operator resource is inversely proportional to NF processing delay. The same procedure has followed, and the obtained results are presented in Table 8 and table 9 respectively.

Table 8. Distance to MNO core when uO T_{NF} = 10 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	100 ms	-
2	10 ms	500 m
3	1 ms	231.92 km
4	0.1 ms	255.07 km

Table 9. Distance to MNO core when uO T_{NF} = 100 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	1000 ms	-
2	100 ms	500 m
3	10 ms	2314.78 km
4	1 ms	2546.21 km

We can see that if the uO performance is low MNO has the ability to establish the core network at larger distances to achieve the same E2E latency as uO achieves. However, by definition, uO provides a case specific, localized service. Therefore, uO performance is optimized to cater the use case. In addition to that, parameters such as NF processing delay lies in the range of µs instead of ms. Hence, we consider these two use cases are unrealistic and no further consideration is given.

5.1.2 PDU Session Establishment Process

Next step is the simulation of PDU session establishment process. AR device must establish a session with the image processing server before the continuous data transfer occurs between AR device and the server. As we did before, we vary the MNO core network distance from 500 m to 500 km in 50 km intervals. uO core network distance is kept constant at 500 m justifying the fact the uO core network is located inside the factory premises. We use the parameters presented in Table 6 and we take the network function processing delay of uO as 1 ms and network function processing delay of MNO as 0.1 ms. Result of this experiment is illustrated in Figure 37 along with the latency of uO is for the comparisons.

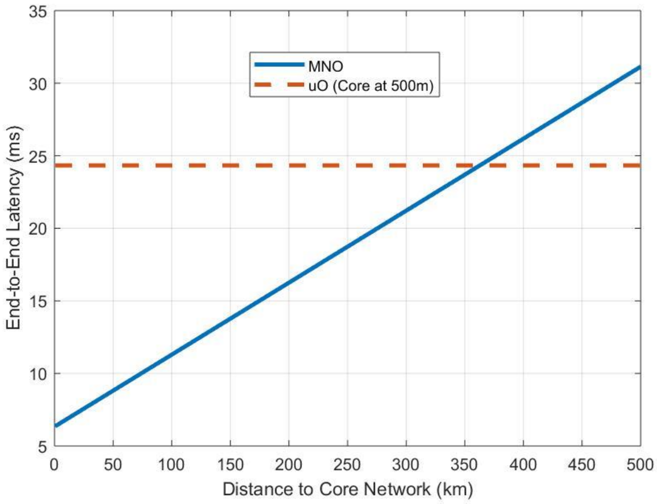


Figure 37. E2E latency of AR session establishment process with respect to distance.

Figure 37 shows that the E2E latency exhibits a linear increase with respect to the distance to the core network. Increment is approximately 49.61 µs per 1 km. Even though session establishment process has multiple round-trip communications towards the core network and back to the device, there are lot of communications and processing inside the core network also. Therefore, the dominant factor is the network function processing delay rather than the distance to core network. In this case, MNO can achieve low latencies than uO for a range of core network distance which are more realizable. In fact, if the core network distance of MNO is less than 362.84 km, MNO can have low latency than uO for this communication step.

We do not simulate the E2E latency of the PDU session establishment process with respect to network function processing delay because from Figure 36, we noticed that it will depend on the distance to core network. Since the core network distance of uO is 500 m and the core network distance of MNO is 250 km, uO always performs better in terms of low latency.

5.1.3 E2E Data Transfer Process

We also simulate the end-to-end data transfer process from AR device to image processing server and back to the AR device. This communication should happen within 50 ms to avoid cyber-sickness [8]. Therefore, the objective of our simulation is to identify the distance to core network for the MNO deployment, in which the desired latency can be achieved. We consider the NF processing delay of uO to be 1 ms and for MNO to be 0.1 ms. Figure 38 illustrates the results of the simulation. For the purpose of comparison, 50 ms latency curve and the uO latency when its core network is located 500 m away from the factory is also depicted on the same graph.

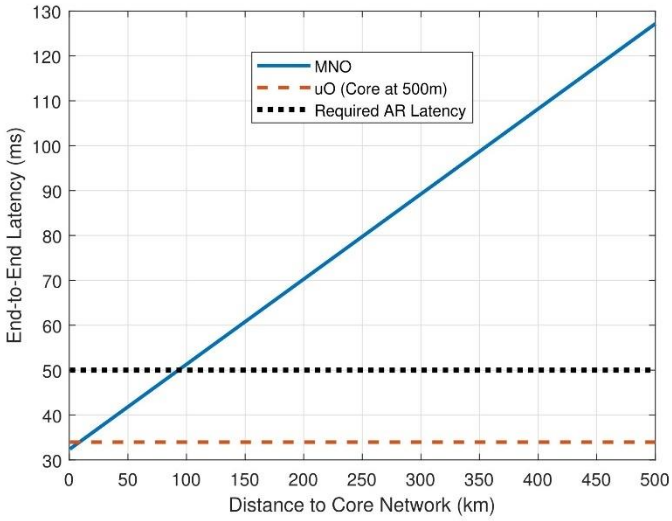


Figure 38. E2E latency of data transfer between AR device and server.

From the results, we can see that the core network of MNO should be located closer than 92 km from the factory premises to achieve the desired 50 ms E2E latency [8]. However, in the real scenarios, factories are in diverse geographical areas and this requirement is difficult to be satisfied. Result also shows that the uO can have far better performance compared with MNO, even though the NF processing delay of uO is 10 times higher than the MNO, making uO deployment the more favorable option.

As we did for the registration process, we can express the E2E latency of data transfer process as,

$$L_{dat} = k_1.T_{access} + k_2.T_{backhaul} + k_3.T_{NF} + T_{server} \tag{5}$$

where L_{dat} is the E2E latency of the data transfer process, T_{server} is the aggregated delay at the image processing server takes to process the image, apply augmentations and send back to AR device. Therefore, for a given L_{dat} is we can derive the distance to MNO core network as,

$$D_{backhaul} = (L_{dat} - k_1.T_{access} - k_3.T_{NF} - T_{server})/(k_2.k_4) \tag{6}$$

As we explained in Chapter 5.1.1, here also we consider the four different cases where MNO has different resource levels compared with uO. For each case, we observe how far MNO can establish its core network away from the factory premises, while achieving the same E2E latency provided by uO. Results of this experiment are outlined in Table 10.

Table 10. Distance to MNO core when uO T_{NF} = 1 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	10 ms	-
2	1 ms	500 m
3	0.1 ms	9.5 km
4	0.01 ms	10.4 km

Similar to what we have seen in registration process, if both uO and MNO have the same resources, it is not possible for MNO to cater the E2E latency supported by uO. When MNO resources is 10x higher than of uO, MNO provides similar E2E latency as uO with a core network distance of 500 m. This is because MNO is serving 10 factories simultaneously. Even with much higher resource levels, MNO can get only a slight advantage for core network distance. By increasing the resources from 10x to 100x, MNO achieves only a 9.5 km advantage. By increasing the resources from 100x to 1000x a further 1 km advantage is acquired. This makes MNO the non-favorable deployment option.

We consider a second scenario with uO network function processing delay at 10 ms. We obtain E2E latency for the data transfer process as carried out in the previous experiment and the result we obtained is 52.1 ms. Based on [8], the latency requirement is 50 ms, therefore this scenario is not realistic, and we do not analyze the distance to core network for this scenario.

5.2 Massive Wireless Sensor Networks

We first analyze the E2E latency of the wireless sensor network communication with respect to the distance to core network. We keep the core network distance of uO as 500 m and we vary the core network distance of MNO from 500 m to 500 km in 50 km intervals.

The whole communication process includes multiple session establishments and data transfer processes. Objective of either uO or MNO is to minimize the E2E latency. We take that the uO operation is tailored to cater the sensor network communication and optimized to achieve minimum E2E latency. We first take the network function processing delay of uO is 1 ms. In this experiment also, we discuss about four cases where we assume different resource levels at MNO, as we did in previous experiment. In case 1, MNO and uO have equal resource. In cases 2,3,4 MNO has 10x, 100x and 1000x resources of uO respectively.

Since we know that the MNO is catering for 10 factories simultaneously, under case 1, MNO cannot have better latency performance than uO. This is because when MNO and uO have the same resources, T_{NF} for MNO is 10 times the T_{NF} of uO. Under case 2, MNO can produce the same performance as uO, since MNO resources are divided among ten factories and T_{NF} is the same for both uO and MNO. In case 3 and case 4, MNO can have better performance, but the performance may vary with the distance to the core network. Figure 39 depicts the latency performance of MNO under case 3. In this case, we take T_{NF} of MNO as 0.1 ms since MNO resources are divided among ten factories and increased resources corresponds to decreased T_{NF}. Latency achieved by uO is also depicted in the figure for the comparisons.

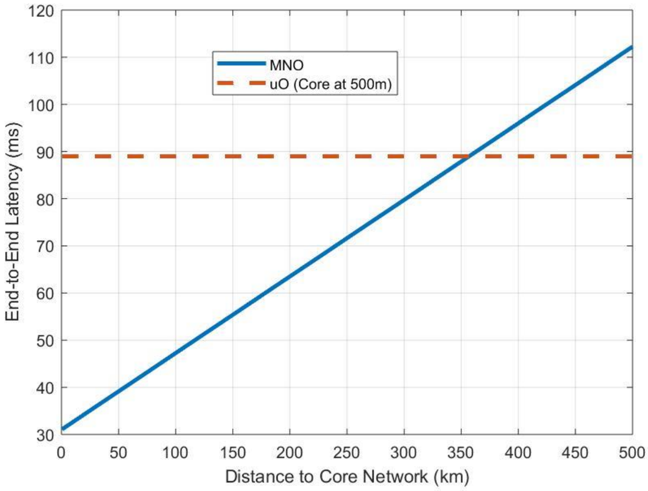


Figure 39. E2E latency of sensor network communication with respect to distance.

Figure 39 shows that the latency increases linearly with respect to the distance to core network. Increase in latency is 0.16 ms per km. Unlike in AR use case, MNO can achieve better latencies even for higher core network distances. It also shows that when the core network distance is higher than a threshold, MNO cannot support the latency provided by the uO. MNO can have better performance only if the core network is located closer than 356.87 km.

According to the message transfer sequence of the sensor network communication, E2E latency for the process can be expressed as follows,

$$L_{sensor} = k_1.T_{access} + k_2.T_{backhaul} + k_3.T_{NF} + T_{alarm} + T_{control} \tag{7}$$

where L_{sensor} is the E2E latency of the sensor network communication process, T_{alarm} is the processing delay of alarm management server and $T_{control}$ is the processing delay at the controlling server. k_1 and k_2 are the number of times where the relevant messages travel via the access channel and the fiber backhaul respectively. k_3 being the number of times which the messages are processed through a network function. Since $T_{backhaul}$ directly proportional to the distance between access network and core network as mentioned in equation (2), combining it with equation (7), for a given L_{sensor}, the distance between access network and the core network can be derived as follows,

$$D_{backhaul} = (L_{sensor} - k_1.T_{access} - k_3.T_{NF} - T_{alarm} - T_{ctrl})/(k_2.k_4) \tag{8}$$

We then analyse how MNO can locate its core network when it has 1000x resources than uO. We use equation (8) to derive it because we already know the latency provided by the uO is 88.99 ms from data depicted in Figure 39. For MNO to achieve the latency provided by uO, when it has 1000x resources of uO, it can move the core network up to 374.9 km. These results are summarized in Table 11.

Table 11. Distance to MNO core when uO T_{NF} = 0.1 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	10 ms	-
2	1 ms	500 m
3	0.1 ms	356.87 km
4	0.01 ms	374.9 km

Next, we consider the same experiment assuming that the uO has more resources. This means that T_{NF} for uO is even lower. We take it as 0.1 ms and we consider the four cases. In this case when the uO core network is at 500 m, we get the latency as 31.09 ms. Table 12 summarizes the results.

Table 12. Distance to MNO core when uO T_{NF} = 0.1 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	1 ms	-
2	0.1 ms	500 m
3	0.01 ms	36.14 km
4	0.001 ms	64.69 km

We consider the uO operation is tailored to provide the sensor network communication therefore we take the core network of uO is optimized with required resources for network functions. Therefore, we can say that if the uO is very efficient in communications, MNO cannot achieve the same latency because the core network of uO has to be there very close to the factory. This is not possible in all cases due to the geographical locations of the factory. However, if the performance of uO is less, MNO can achieve lower latencies for sensor network communication. This is because in this case the core network processing is more dominant than the distance factor.

Then we analyze the latency against the network function processing delay because in the first experiment, we saw that T_{NF} plays the dominating role is this communication. Having efficient network functions with higher resources can minimize the overall latency. For this experiment, we keep the uO core network distance as 500 m as the previous case and we also fix the MNO core network distance at 250 km, by taking the average of the distance range we considered for the previous experiment. Network function processing delay is varied from 1 µs to 1000 µs and the resulting E2E latency graph is depicted in Figure 40.

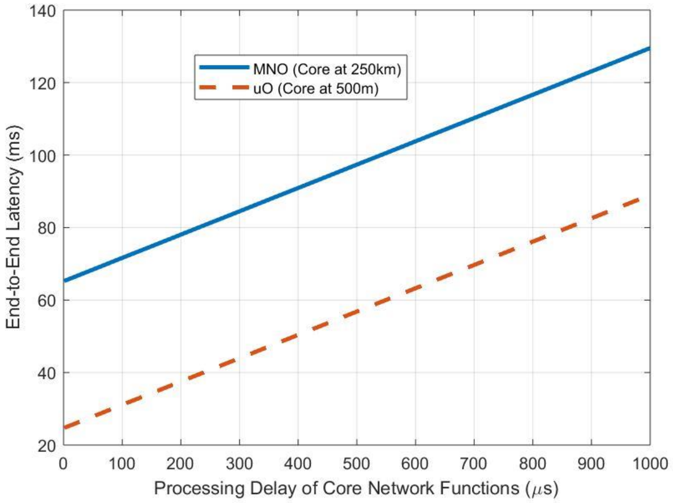


Figure 40. E2E latency of sensor network communication vs. NF processing delay.

E2E latency shows a linear variation with respect to NF processing delay for both uO and MNO. Latency increase is approximately 64.32 µs when NF processing delay is increased by 1 µs. Unlike in the previous experiment, it is seen that uO always perform better than the MNO in this case. This is because of the distance to core network being always higher in the MNO case than the uO case.

In the real deployments, uO operation is optimized to cater the needs of the sensor network communication because uO always provides a tailored service. On the other hand, MNO must serve the ten factories, and in the meantime, it has to provide the service to its other subscribers as well such as mobile broadband traffic. This places higher load on MNO core network than the uO core network. Therefore, there is a high probability of having high NF processing delay in MNO case compared with uO case.

5.3 Mobile Robots

As discussed in chapter 3.3 Mobile robots use case includes the registration of mobile robots into the 5G system, PDU session establishment between two mobile robots, a mobile robot and the guidance control server and sometimes between the robots and the environment sensors. It also includes the data transfer procedure between devices. The registration process is similar to what we have discussed under the AR use case and the PDU session establishment process is included in massive wireless sensor network communications. Data transfer process we discussed in AR communication in detail. The other significant communication component associated with mobile robots is the handover procedure. If the number of mobile robots is high, then the number of handovers will also be high. It also depends on the speed of the robots. Therefore, more emphasis has been put to analyse the handover process in mobile robots' communication.

5.3.1 Handover Procedure

First, we simulate the handover process for MNO deployed model and compare it with uO deployed model for latency measurements. Here also, we vary the MNO core network distance from 500 m to 500 km in 50 km intervals. uO core network distance is kept constant at 500 m justifying the fact that the uO core network is located inside the factory premises. We use the parameters presented in Table 6 and we take the network function processing delay of uO as 1 ms and network function processing delay of MNO as 0.1 ms. This is because MNO has more resources than uO which will enable faster processing than uO. Result of this experiment is illustrated in Figure 41. Latency of uO is also illustrated in Figure 41 for the purpose of comparing the results with the MNO results.

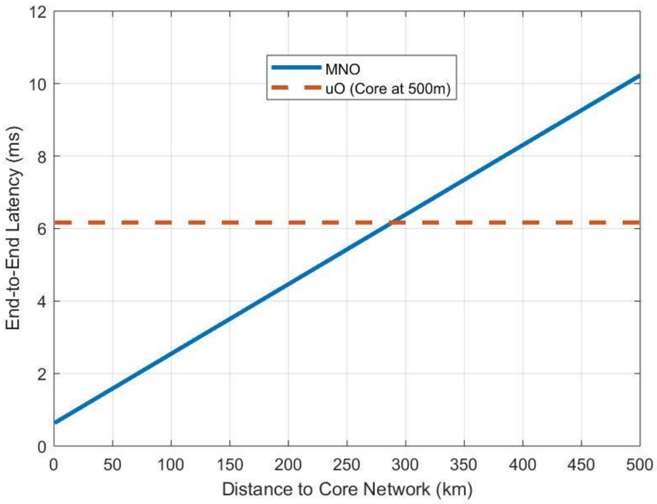


Figure 41. E2E latency of robot handover process with respect to distance.

It is seen from Figure 41 that the E2E latency shows a linear increase with respect to the distance to the core network. Increment is approximately 19.19 μs per 1 km. Handover process has no multiple round-trips between core network and the mobile robot. Therefore, with higher resources at the core network MNO can have better performance than uO even with its core network at higher distances. When the core network distance is greater than 288.64 km, latency of the handover process at MNO becomes higher than the latency given by uO.

As the second experiment, we monitor the latency with respect to the network function processing delay. During the handover process, messages must be processed by core network and access network functions multiple times until the handover process is completed. Therefore, having efficient network functions with higher resources can minimize the overall latency of the handover process. For this experiment, we keep the uO core network distance as 500 m and we also fix the MNO core network distance at 250 km, by taking the average of the distance range we considered for the previous experiment. Network function processing delay is varied from 1 μs to 1000 μs and the resulting E2E latency graph is depicted in Figure 42.

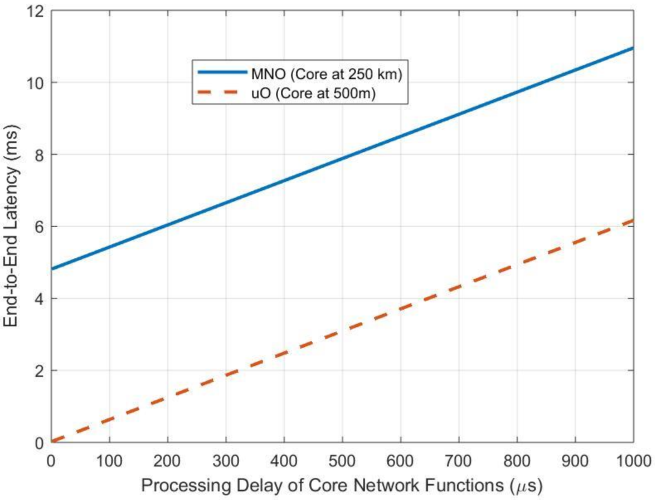


Figure 42. E2E latency of robot handover process vs. NF processing delay.

E2E latency shows a linear variation with respect to NF processing delay for both uO and MNO. Latency increase is approximately 6.15 μs when NF processing delay is increased by 1 μs. Unlike in the previous experiment, it is seen that uO always perform better than the MNO in this case. This is because of the distance to core network being always higher in the MNO case than the uO case.

In the real deployment, uO only serves for the mobile robot communication since uO provides a tailored service through a different network slice. Therefore, the communication is optimized for the use case. On the other hand, MNO must serve 10 similar factories simultaneously. Moreover, MNO has to provide the service to its regular subscribers, for example internet access of broadband users. This places higher load on MNO core network than the uO core network. Therefore, there is a high probability of having high NF processing delay in MNO case compared with uO case.

5.3.2 E2E Data Transfer

E2E data transfer between robots, between robot and the guidance control server are the frequent data transfers occur in this use case. We consider data transfer between two devices irrespective of their specific roles. We consider that the core network of uO is located 500 m from the factory and we vary the distance to MNO core network from 500 m to 500 km. We keep the T_{NF} from uO as 1 ms and T_{NF} for MNO as 0.1 ms. Figure 43 illustrates the results of the simulation. For the comparison, latency curve of uO also depicted in the same graph.

We can see that uO can achieve better performance in almost all the range we considered. In fact, only if the MNO core network is closer than 9.32 km, MNO can have a better latency value. Therefore, considering both the handover and the data transfer processes, uO is the better deployment option to achieve low latency communication.

Here also, we can express the E2E latency of data transfer instance as,

$$L_{dat\_r} = k_1.T_{access} + k_2.T_{backhaul} + k_3.T_{NF} \tag{9}$$

where L_{dat_r} is the E2E latency of the data transfer between two mobile robots, a mobile robot and the guidance control server, a mobile robot and an environment sensor node.

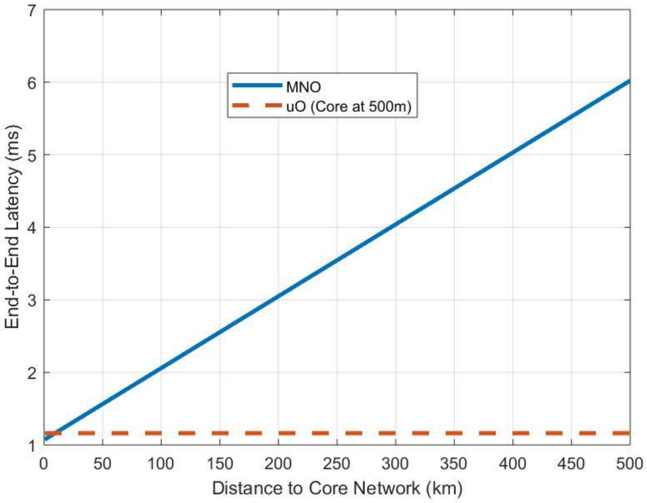


Figure 43. E2E latency of data transfer between robots and other devices.

Therefore, for a given L_{dat_r} we can derive the distance to MNO core network as,

$$D_{backhaul} = (L_{dat\_r} - k_1.T_{access} - k_3.T_{NF})/(k_2.k_4) \tag{10}$$

As we explained in Chapter 5.1.1, here also we consider the four different cases where MNO has different resource levels compared with uO. For each case, we observe how far MNO can establish its core network away from the factory premises, while achieving the same E2E latency provided by uO. Results of this experiment are outlined in Table 13.

Table 13. Distance to MNO core when uO T_{NF} = 1 ms

Case	NF Proc. Delay of MNO	$D_{backhaul}$
1	10 ms	-
2	1 ms	500 m
3	0.1 ms	6.2 km
4	0.01 ms	15.2 km

In this case also, at much higher resource levels, MNO can get only a slight advantage for core network distance. By increasing the resources from 10x to 100x, MNO achieves only a 6.2 km advantage. By increasing the resources from 100x to 1000x a further 9 km advantage is acquired. This makes MNO the non-favorable deployment option to cater the communication.

5.4 Performance Measurements of Kuha Base Station Setup

Using the uO setup implemented with Kuha base station and the Cumucore EPC, we conduct two experiments. One to measure latency and the other to measure throughput. For the experiments, we used two android smartphones attached to Kuha base station and the Cumucore EPC provides dynamic IP addresses to those devices. We use android applications to measure the latency and throughput. We carry out the same experiment 10 times and get the average value as the final value. The applications use IP addresses to communicate with the mobile devices attached to Kuha base station. To compare the performance results, we carry out the same experiment for DNA operator. We use the same devices and perform the experiments at the same locations to ensure a fair comparison of the performance results. We also consider different times of the day from 0800 hours to 1800 hours and take measurements in every two-hour intervals. Table 14 compares the latency measurements for Kuha based uO network and DNA network. Latency is always lower for Kuha uO network and does not show much variation. This is because the core network is closely located. On the other hand, latency of DNA is higher and has significant variations. This is due to the core network distance and the different traffic congestions in DNA network at different times.

Table 14. Comparison of latency between Kuha based uO setup and DNA

Time of the Day	**Kuha uO Latency (ms)**	**DNA Latency (ms)**
0800	87.68	245.49
1000	81.54	256.68
1200	80.41	102.74
1400	86.31	247.14
1600	82.39	212.34
1800	80.66	126.49

Table 15 compares the throughput measurements for Kuha based uO network and the DNA network. Throughput is always higher for Kuha uO network than DNA. This is because the only traffic in Kuha uO network is the traffic originated by the mobile phones used for testing. On the other hand, throughput of DNA is lower in all the cases. Reason could be the traffic congestions at DNA network. Both the latency measurements and the throughput measurements can be used for future research activities.

Table 15. Comparison of throughput between Kuha based uO setup and DNA

Time of the Day	**Kuha uO Throughput (Mbps)**	**DNA Throughput (Mbps)**
0800	15.3	1.35
1000	16.6	1.35
1200	12.7	1.34
1400	13.6	1.34
1600	13.4	1.25
1800	13.9	1.34

5.5 Summary of Results of Three Use Cases

Our experiments were based on the latency performance measurement and we discussed two main factors which cause the latency in uO and MNO cases. We discussed the experiments for three use cases namely AR, massive wireless sensor networks and mobile robots. Even though there are multiple communication steps involved in each use case, we can identify the critical communication step(s) which the operator would want to minimize the latency. Also, we can identity the factor that manly contributes to the latency, whether it is the distance to the core network or whether it is the network function processing delay. If the communication has multiple round trips between the core network and the terminal devices, distance to core network will affect more to decide the final latency. If the communication has fewer round trips to devices, but more processing at the core network, then the network function processing delay will be the dominant factor which decides the latency. Based on that we can decide whether it is uO or MNO would be the best deployment option. Table 16 summarizes this information.

Table 16. Summary of critical communications of each Use Case

Use Case	Critical communication Step(s)	Dominating Factor
AR	Continuous data transfer from AR device to image processing server and back	Distance to core network
Sensor Network	End-to-end data transfer from sensor to destination machine	Network function processing delay
Mobile Robots	Data transfer and Handover	Distance to core network

Based on the data in Table 16, we can identify the preferable deployment option to cater for each use case. Since the AR use case depends on the distance to core network and the fact that uO always has the ability to implement the core network within the factory premises, uO will be the preferable deployment option. For the sensor network use case, both uO and MNO can cater the communication given that their core networks are fast and has low latencies in core network processing. But, uO has more control over the core network resources and MNOs core network delay may vary on the traffic from external factors such as regular broadband users, making MNO less favourable. For the mobile robots use case, both MNO and uO can efficiently server for fast handovers but for the data communication uO provides low latencies. Therefore, uO will be the preferable option for mobile robots.

Therefore, for a factory environment comprising all three use cases, the results highlight that the uO deployment option is better for low latency communications.

6 DISCUSSION

This section critically analyses and compares the work that has been carried in this thesis with other works in the same research field. The contribution of the thesis has been analysed against the similar work which is already available, if any. Next, a critical analysis is presented considering the thesis objectives and the real outcomes of the thesis focusing on whether the thesis objectives are met and to what level that has been addressed. Finally, the discussion focuses on the future research direction. This also includes the ideas that on possible approaches those have emerged during the thesis work.

6.1 Comparative analysis with Similar Work

This thesis defines a system architecture for a local 5G operator providing a tailored service. It proposes an architecture for a uO operating in a future industrial environment and having three common use cases named augmented reality, massive wireless sensor networks and mobile robots. Even though the uO concept has been there for a while, there has not been much work carried out to analyze possible architecture options for uO. Based on the uO concept, it can operate individually to cater the communications of a specific use case or it can collaboratively work with an MNO to provide a set of specific services. A possible system overview uO network is suggested in [49], however it does not demonstrate in depth details related to uO network architecture other than mentioning that the architecture would comprise 5G gNBs and it would provide services to different verticals such as industry, health and education. The possibility of creating network slices to serve the needs of different verticals like media, industry and health is explained in [1]. It also explains the fact that the service optimized local architectures could cater these communications. This work provides an overview of the idea of slice utilization of uO in each vertical. The fact that the uO architecture must be developed in conjunction with different technologies such as SDN and NFV has been highlighted in [50]. This considers uO as a relatively small in scale operator and holds limited resources to provide particular and necessary services to a certain number of users. Here also the network architecture has not been analyzed in detail.

To bridge that gap of unavailability of a properly defined uO architecture, this research proposes a network architecture for a uO which provides case specific and localized communication services. In this thesis, communication steps of each use case have been analyzed in detail. Message sequences occur between the core network functions and access network clearly studied referring to 3GPP specifications. Based on the network function definitions and the functional descriptions release by 3GPP, this thesis proposes the conceptual uO architecture. To facilitate for slicing, the thesis also considers the three slice management functions called CSMF, NSMF and NSSMF as part of the proposed architecture. Considering the role of NSSF which helps to identify the best fitting slice for a specific communication, the thesis includes NSSF also as part of the architecture therefore, the work that that been carried out in this thesis investigates an area which has not been analyzed in greater detail previously.

6.2 Evaluation on Meeting the Thesis Objectives

The objective of the thesis is to define a network architecture for a uO, providing case specific and localized services. Since uOs are specialized to provide tailored services, the system architecture of a uO and its deployment may also depend on the environment and the use case. Therefore, the scope of the thesis is confined to defining the architecture to cater the communication needs of three selected use cases called augmented reality, massive wireless sensor networks and mobile robots in a future industrial environment.

To realize the objective of defining the architecture, thesis uses the 5G network functions defined by 3GPP since uO is also a 5G operator. During the design of the architecture, a detailed analysis has been carried out for the communication steps occur in each use case. The communication steps and the message transfer between the core network functions, access network and terminal devices has been comprehensively investigated for each use case. Based on the message sequence, the architectural components were identified and those were then used to define the architecture for each use case. For a factory employing all three use cases, the final architecture has been derived in this thesis with the use of network slicing, which enables creating logical networks on the same physical infrastructure.

Once the conceptual architecture is defined which caters the three use cases, simulations were conducted to demonstrate the benefits of the uO based deployment over the MNO deployed model. The performance metric considered for comparisons is the latency measurement for each communication and the variables considered were the distance to core network and the network function processing delay which is inversely proportional to operator resources. Critical communications steps of each use case has been analyzed and the benefits of having uO architecture is discussed as follows.

If the communication includes multiple round trips to core network and the access network, then the deciding factor for the latency will be the distance to core network. For such scenarios, the thesis proves that the uO deployment is always better because uO can deploy the core network inside the factory premises, but for MNO it is highly unlikely. If the communication includes more processing between the network functions, then the deciding factor is network function processing delay. For this case, an MNO with higher resources in the core network may also provide better performance. However, MNO needs to handle other forms of traffic such as the broadband traffic of users therefore optimizing for a specific use case is challenging. On the other hand, since uO is providing a case specific service, it can always be optimized its resources to cater the communication with latency guarantees. This way, the thesis argues that uO can have guaranteed better performance. Therefore, as the outcome of the thesis, it produces an architecture for a uO, which provides efficient communication services than MNO deployment. The outcome of the thesis can be used in deploying uO 5G networks in industrial environments in the future.

6.3 Future Research Directions

The thesis defines a uO architecture for a factory environment assuming the factory is covered only by uO. Core network of uO which includes the network functions needed for the communications, is established inside the factory premises to achieve low latency. A different approach for the deployment will be a hybrid architecture where both MNO and uO collaboratively implement the communication service to cater the industry use cases. One approach is that the network functions which are latency critical can be implemented by uO similar to the architecture defined in this thesis and the other network functions which are not latency critical can be implemented by MNO. This way uO needs less resources in its own network which will be an advantage to the uO. On the other hand, implementation, maintenance and troubleshooting will be more complex for uO and MNO. How and what resources will be implemented by uO and MNO will be considered as a different research problem.

MEC brings the computational resources towards the edge of the network. In the MNO deployed model, MNO can bring the latency critical network functions to the edge of the network and implement the communication service for the factory use cases. Since the processing is carried out at the edge of the network, this implementation will be more appropriate for use cases having multiple round trips to the core network. Because the distance is made shorter by implementing this, MNO can achieve lower latencies than the conventional method of establishing the core network in a central location. The resources to be implemented on the edge and the scaling of the resources will differ based on the use case.

To cater the three use cases, this thesis proposes three logical networks created using network slicing. But the specific detail of how the network slices should be created is not discussed in this thesis. This can be addressed as a different research problem related to network slicing in uO networks. For example, augmented reality use case needs continuous transfer of video streams within the network and it will require comparatively higher network resources. On the other hand, massive wireless sensor networks may not require transferring large volumes of data. Therefore, the resource allocation for a given network function in the augmented reality slice could be higher and for the sensor network slice it could be lower.

As performance measurement, the thesis considers only the latency values. The research can be extended to take other measurements such as throughput because throughput is important in high data rate applications such as augmented reality. On the other hand, distance to core network and the operator resources are considered as independent variables. Other parameters such as network traffic that leads to congestion can also be considered as a variable and can be analysed together with the above variables. This approach will result in more comprehensive outputs and will provide more insights for the actual deployments.

The thesis focuses on future industrial environment and three most probable use cases. This is because the architecture depends on the specific service to be implemented. More environments such as hospitals and universities can also be considered for uO deployments. The use cases will be different with the chosen environment and communication processes have to be carefully examined to define architecture for these use cases.

7 SUMMARY

The goal of this research work is to define a network architecture for a uO providing case specific, localized communication services. The uO is assumed to be providing services for a factory environment having industry 4.0 standards, running the augmented reality, massive wireless sensor networks and mobile robots use cases. Since uO itself is a 5G network service provider, the thesis defines the network architecture using the 5G network functions defined by 3GPP. To derive the architectural components for the three industry use cases, the thesis analyzes the communication steps occur in each use case to gain understanding of the adequate network functions required in the architecture. For each communication, the message transfer sequences between core network functions, message transfer between core network and the access network, message transfers between the access network and the terminal 5G devices have been comprehensively analyzed. The analysis results in identifying the mandatory network functions which form the network architecture for each use case. For a factory environment having all the three use cases simultaneously, the final uO architecture has been derived by with the use network slicing, assuming that each use case is served by a different network slice. The final architecture includes the network functions which are mandatory to cater the use case communications, three network slice management functions (CSMF, NSMF, NSSMF) which are required to create and manage network slices for each use case and the NSSF which chooses the best fitting slice for each use case communications.

To demonstrate the benefits of having a uO to serve the industry use cases rather than having an MNO deployed network, the thesis conducts simulations based on two models. In the first model, the factory is entirely served by a uO, all the terminal 5G devices and the servers are owned by the factory and located within the factory premises, and the core network of the uO is also established within the factory premises. In the second model, the factory is served by an MNO, all the servers and terminal devices are owned by the factory and located within the factory, but the core network of MNO is located outside the factory. It is also assumed that the MNO is serving 10 such factories simultaneously, making the MNO model more realistic. E2E latency has been considered as the key performance measurement. Distance to the core network and the operator resources which is inversely proportional to the network function processing delay have been considered as the variable parameters which affect the E2E latency. Simulations have been carried out for each communication step of the use cases to analyze how E2E latency varies with the two parameters.

The results of the simulations yielded below outcomes. If the communication has multiple round trips to the core network and back to the terminal devices, distance to core network will be the dominant factor deciding the E2E latency. Because uO can establish the core network very close to the factory premises, uO will be the preferred deployment option for such use cases. Since MNO is providing communication services to multiple factories, establishing the core network in close proximity is not realistic, making MNO the non-favorable choice. For example, in the augmented reality use case, to achieve the given round trip latency of 50 ms given in the 3GPP study, MNO has to establish the core network at least at 92 km from the factory location. However, MNO can increase its network resources to achieve the same E2E latency by minimizing the network function processing delay and increasing the core network distance. Although the advantage in distance that MNO can achieve is not significant. Even from 10x to 100x increase of network resources MNO can get only 9.5 km advantage to move the core network away from the factory. By increasing resources from 100x to 1000x, MNO can only get a further 1 km advantage which is not very useful. Mobile robots use case exhibits the same characteristics because it is also depending on the distance to core network. Therefore, uO will be the preferred deployment option for mobile robots use case too.

If the communications have lesser round trips between the terminal devices and core network, and if there are more processing inside the core network, the dominant parameter contributing the E2E latency is the network function processing delay. Typically, MNOs have higher resources than uOs, therefore MNO can cater use cases of this kind more efficiently than the previous cases. On the other hand, MNOs must simultaneously serve for other traffic forms such as the broadband traffic of users, therefore increasing the resources for a specific use case such as Industry 4.0 communication will be challenging. Since uO is providing a tailored service, uO operation can be optimized for the use case by increasing the resources easier than MNO, making uO also a good option for such deployments. Therefore, we argue that both MNO and uO are preferred options to cater these communications, but uO has a slight advantage because uO, by definition providing a tailored service, can optimize its resources easily based on the factory requirements than an MNO. The second use case, massive wireless sensor network communication is an example scenario for this.

For an industrial environment having all three use cases, it is obvious that the uO will be the preferable deployment option because of its close proximity of core network location and the higher flexibility in resource optimization.

8 REFERENCES

[1] Matinmikko-Blue, Marja and Yrjoelae, Seppo and Latva-aho, Matti, “Micro operators for ultra-dense network deployment with network slicing and spectrum micro licensing,” IEEE 87th Vehicular Technology Conference (VTC Spring), pp. 1–6, 2018.

[2] Andrews, Jeffrey G and Buzzi, Stefano and Choi, Wan and Hanly, Stephen V and Lozano, Angel and Soong, Anthony CK and Zhang, Jianzhong Charlie, “What will 5G be?,” IEEE Journal on selected areas in communications, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] European Commission, “Strategic Spectrum Roadmap towards 5G for Europe: RSPG Second Opinion on 5G Networks,” Radio SpectrumPolicy Group (RSPG), RSPG18-005,2018.

[4] Ahokangas, P., Matinmikko, M., Yrjo¨la¨, S., Okkonen, H., & Casey, T. (2013). ‘‘Simple rules’’ for mobile network operators’ strategic choices in future cognitive spectrum sharing networks. IEEE Wireless Communications, 20(2), 20–26.

[5] M. Matinmikko, M. Latva-Aho, P. Ahokangas, S. Yrj¨ol¨a, and T. Koivum¨aki, “Micro Operators to boost Local Service Delivery in 5G,” Wireless Personal Communications, vol. 95, no. 1, pp. 69–82, 2017.

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

[7] 3GPP, “System Architecture for the 5G System,” Technical Specification, June 2018.

[8] 3GPP, “Study on Communication for Automation in Vertical Domains (CAV),” Technical Report, July 2018.

[9] 3GPP, “Procedures for the 5G System,” Technical Specification, June 2018.

[10] Y. Siriwardhana, P. Porambage, M. Liyanage, J. S. Walia, M. Matinmikko-Blue and M. Ylianttila,” Micro-Operator driven Local 5G Network Architecture for Industrial Internet”, to be appeared in Proc. of IEEE Wireless Communications and Networking Conference (WCNC), Marrakech, Morocco, April 2019.

[11] Soldani, David and Manzalini, Antonio, “Horizon 2020 and beyond: on the 5G operating system for a true digital society,” IEEE Vehicular Technology Magazine, vol. 10, no. 1, pp. 32–42, 2015.

[12] Iwamura, Mikio, “NGMN view on 5G architecture,” IEEE 81st Vehicular Technology Conference (VTC Spring), pp. 1–5, 2015.

[13] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, “5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1201–221, 2017.

[14] R ITU, “IMT Vision Framework and Overall Objectives of the Future Development of IMT for 2020 and beyond,” Rec. ITU-R M.2083, Sept. 2015.

[15] Madhusanka Liyanage, Ijaz Ahmed, Ahmed Bux Abro, Andrei Gurtov, Mika Ylianttila, A comprehensive Guide to 5G Security, Wiley, 2018.

[16] A. Osseiran et al., "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project", IEEE Commun. Mag., vol. 52, no. 5, pp. 26-35, May 2014.

[17] Hossain, Ekram and Hasan, Monowar, “5G Cellular: Key Enabling Technologies and Research Challenges,” 2015.

[18] Akyildiz, Ian F and Nie, Shuai and Lin, Shih-Chun and Chandrasekaran, Manoj, “5G roadmap: 10 key enabling technologies,” Computer Networks, vol. 106, pp. 17–48, 2016.

[19] Madhusanka Liyanage, Mika Ylianttila, Andrei Gurtov, Software Defined Mobile Networks (SDMN): Beyond LTE Network Architecture , Wiley, 2015.

[20] Gupta, Akhil and Jha, Rakesh Kumar, “A survey of 5G network: Architecture and emerging technologies,” IEEE access, vol. 3, pp. 1206–1232, 2015.

[21] Sezer, Sakir and Scott-Hayward, Sandra and Chouhan, Pushpinder Kaur and Fraser, Barbara and Lake, David and Finnegan, Jim and Viljoen, Niel and Miller, Marc and Rao, Navneet, “Are we ready for SDN? Implementation challenges for software-defined networks,” IEEE Communications Magazine, vol. 51, no. 7, pp. 36–43, 2013.

[22] Pi, Zhouyue and Khan, Farooq, “An introduction to millimeter-wave mobile broadband systems,” IEEE Communications Magazine, vol. 49, no. 6, pp. 101–107, 2011.

[23] Lu, Lu and Li, Geoffrey Ye and Swindlehurst, A Lee and Ashikhmin, Alexei and Zhang, Rui, “An overview of massive MIMO: Benefits and challenges,” IEEE journal of selected topics in signal processing, vol. 8, no. 5, pp. 742–758, 2014.

[24] N. Bhushan et al., "Network densification: The dominant theme for wireless evolution into 5G", IEEE Commun. Mag., vol. 52, no. 2, pp. 82-89, Feb. 2014.

[25] I. Hwang, B. Song, S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks", IEEE Commun. Mag., vol. 51, no. 6, pp. 20-27, Jun. 2013.

[26] Kamel, Mahmoud and Hamouda, Walaa and Youssef, Amr, " Ultra-dense networks: A survey", IEEE Communications Surveys & Tutorials, vol. 18, no. 4, pp. 2522-2545, 2016.

[27] Mao, Yuyi and You, Changsheng and Zhang, Jun and Huang, Kaibin and Letaief, Khaled B, “A survey on mobile edge computing: The communication perspective,” IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322–2358, 2017.

[28] Porambage, Pawani and Okwuibe, Jude and Liyanage, Madhusanka and Ylianttila, Mika and Taleb, Tarik, “Survey on multi-access edge computing for internet of things realization,” IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2961–2991, 2018.

[29] Ahokangas, Petri and Moqaddamerad, Sara and Matinmikko, Marja and Abouzeid, Alhussein and Atkova, Irina and Gomes, Julius Francis and Iivari, Marika, “Future micro operators business models in 5G,” The Business & Management Review, vol. 7, no. 5, pp. 143, 2016.

[30] Matinmikko, Marja and Latva-aho, Matti and Ahokangas, Petri and Seppänen, Veikko, “On regulations for 5G: Micro licensing for locally operated networks,” Telecommunications Policy, vol. 42, no. 8, pp. 622–635, 2018.

[31] M. Wollschlaeger, T. Sauter, and J. Jasperneite, “The future of Industrial Communication: Automation Networks in the Era of the Internet of Things and Industry 4.0,” IEEE Industrial Electronics Magazine, vol. 11, no. 1, pp. 17–27, 2017.

[32] Lasi, Heiner and Fettke, Peter and Kemper, Hans-Georg and Feld, Thomas and Hoffmann, Michael, “Industry 4.0,” Business & information systems engineering, vol. 6, no. 4, pp. 239–242, 2014.

[33] Gilchrist, Alasdair, “Introducing Industry 4.0,” Industry 4.0, pp. 195–215, 2016.

[34] A. Varghese and D. Tandur, “Wireless Requirements and Challenges in Industry 4.0,” in Contemporary Computing and Informatics (IC3I), 2014 International Conference on. IEEE, 2014, pp. 634–638.

[35] X. Li, D. Li, J. Wan, A. V. Vasilakos, C.-F. Lai, and S. Wang, “A Review of Industrial Wireless Networks in the Context of Industry 4.0,” Wireless Networks, vol. 23, no. 1, pp. 23–41, 2017.

[36] Network Slicing in Industry 4.0 Applications: Abstraction Methods and End-to-End Analysis.

[37] Network Slicing for Ultra-Reliable Low Latency Communication in Industry 4.0 Scenarios.

[38] 3GPP, “Procedures for the 5G System,” Technical Specification, June 2018.

[39] NGMN Alliance, “Description of Network Slicing Concept,” NGMN 5G P, vol. 1, 2016.

[40] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, “Network Slicing based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges,” IEEE Communications Magazine, vol. 55, no. 8, pp. 138–145, 2017.

[41] 3GPP, “Study on Management and Orchestration of Network Slicing for Next Generation Network,” Technical Report, June 2018.

[42] Chang, Chia-Yu and Nikaein, Navid and Arouk, Osama and Katsalis, Kostas and Ksentini, Adlen and Turletti, Thierry and Samdanis, Konstantinos, “Slice Orchestration for Multi-Service Disaggregated Ultra-Dense RANs,” IEEE Communications Magazine, vol. 56, no. 8, pp. 70–77, 2018.

[43] “Omnet++ Discrete Event Simulator.” [Online]. Available: https://www.omnetpp.org/

[44] "INET Framework." [Online]. Available: https://inet.omnetpp.org/

[45] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies," Technical Report, September 2018.

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

[47] "Kuha Mobile Network." [Online]. Available: https://www.kuha.io/

[48] "Cumucore." [Online]. Available: https://cumucore.com/

[49] Prasad, Athul and Li, Zexian and Holtmanns, Silke and Uusitalo, Mikko A, "5G micro-operator networks—A key enabler for new verticals and markets," 2017 25th Telecommunication Forum (TELFOR), pp. 1–4, 2017.

[50] Tseng, Chia-Wei and Huang, Yu-Kai and Tseng, Fan-Hsun and Yang, Yao-Tsung and Liu, Chien-Chang and Chou, Li-Der, "Micro Operator Design Pattern in 5G SDN/NFV Network," Wireless Communications and Mobile Computing, 2018.