8-5-2022 11:00 AM

Machine Learning and Operations Research for Intelligence Engines in Future Networks

Ali Chouman, *The University of Western Ontario*

Supervisor: Shami, Abdallah, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Electrical and Computer Engineering

Abstract

The evolution of fifth-generation (5G) and Beyond mobile technologies is spurred by the rapid demands and high-end requirements of next-generation mobile networks. It is imperative that advanced intelligence and machine learning techniques address these dynamic requirements by supporting network operations in terms of maintenance, servicing, and performance. The Third Generation Partnership Project (3GPP) has outlined a Network Data Analytics Function (NWDAF) for 5G Core (5GC) networks that should provide predictive network maintenance and improve network performance in these dynamic networks, and which must leverage the capabilities of artificial intelligence, machine learning, and advanced data analytics methods to satisfy its specification requirements. The work presented in this thesis surveys the current trends and future outlooks for 5G Core networks, in addition to presenting the capabilities of an implemented NWDAF, in emulated 5G environments, towards addressing a scaling optimization problem for Network Functions (NFs) in the 5G Control Plane. The insights from the NWDAF and its support in analytical and optimization problems justify its use as more than a network monitoring and data aggregation tool, but as an intelligence engine that will drive 5G and Beyond networks to satisfy user demand and improve consumer experience altogether.

Keywords: Future Networks, 5G, 5G Core, 6G, NWDAF, Intelligent Networks, Network Function Virtualization, Software-Defined Networking, Machine Learning, Predictive Maintenance

Summary for Lay Audience

Modern networks continue to brace for an increasing number of mobile devices such as smartphones, cars, and smart home appliances/accessories. Mobile network providers are faced with the difficult mission of preparing future networks to accommodate user demand in addition to providing the utmost experience and achieving quality requirements. Enabling technologies allow these networks to efficiently minimize costs of operations and maintenance, while at the same time, improving network performance; however, service providers must also consider the issues of their integration, including placement within the network, reliability of provided services, and the guarantee for high-importance applications' needs to be constantly met. Modern and future networks are exploring new inter-network functionalities that are focused on data analytics and new services tailored towards advanced operations and maintenance. This is what is referred to as intelligent networking: it elucidates the ability of the network to recognize events of congested network traffic or points of failure and formulates decisions for the network (predictive maintenance) to accommodate these new requirements, through instantiation of new network function instances for example. These intelligent networking techniques leverage the use of machine learning, artificial intelligence, and advanced data analytics techniques to aid networks in their operations for the purpose of improving overall performance as well as user experience. The methodology of intelligent networking, as mentioned, can be categorized into a subfield of mathematics known as operations research. The work presented here demonstrates the use of network optimization models and statistical decision analysis to improve network capabilities through simulation: specifically, a scaling optimization problem is addressed, involving how many instances of a network function are required to best serve the network and/or end users. The goal of this research is to justify the use of advanced intelligence as working engines in future networks for the purpose of improving performance and satisfying customer demands as required.

Statement of Co-Authorship

The following manuscript has been accepted for publication (IWCMC 2022) and is included in the body of this thesis:

- A. Chouman, D.M. Manias, and A. Shami. Toward Supporting Intelligence in 5G/6G Core networks: NWDAF Implementation and Initial Analysis. In 2022 International Wireless Communications and Mobile Computing Conference (IWCMC), pages 324–329, 2022.

The following manuscript has been submitted (GlobeCom 2022) and is included in the body of the thesis:

- A. Chouman, D.M. Manias, and A. Shami. A Reliable AMF Scaling and Load Balancing Framework for 5G Core Networks.

The following manuscript is pending submission and will be submitted for review, and is included in the body of the thesis:

- A. Chouman, D.M. Manias, and A. Shami. Network Data Analytics in Future Networks: Trends, Outlooks, and Future Directions for the 5G Core and Beyond Networks.

The following co-authors are recognized for their academic and technical contributions to these works:

- Dr. Abdallah Shami has contributed to the works presented in Chapters 2, 3, and 4 with his technical expertise, academic guidance and professorship, and his opinion and perspective.
- Dimitrios M. Manias has contributed to the works presented in Chapters 2, 3, and 4 with his technical expertise, research, mentorship and guidance, and his opinion and perspective.

Acknowledgements

I would like to preface this thesis with some words of acknowledgement and gratitude for those that have helped and supported me throughout the composition of this work and in my studies.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Abdallah Shami. Since the start of my Master's degree, he has instilled a motivation and fascination in me for this field of study and the research that has been conducted. Through his guidance and mentorship, I have applied my critical thinking skills to problems that I have never encountered before. Dr. Shami's professional insights and considerations, along with his warm demeanor, has allowed me to find my strengths in academia and I look forward to many future successes and collaborations with him.

I am also thankful and grateful for my colleague, companion, and comrade, Dimitrios. The pandemic enforced an abhorrent distance between the lab and me, but I had the pleasure of meeting Dimitri and my fellow lab members in my second year. I appreciate Dimitri's astounding knowledge and approaches to complex research problems, as well as brightening my mood and supporting me at the most difficult of times. His professional attitude is exemplary, and his mentorship to me, both in academia and in life, is truly exceptional. I raise my head with glee looking back at this year, and humbly invite him to join me in many more successes in our work and in our lives.

I must extend my sincere thanks to the other members of our lab, including Ibrahim. This lab has welcomed me with open arms, as I recall when Ibrahim first answered the door and I settled down at my new desk. He demonstrated to me an unrelenting will to help me when I complained of errors and issues and grievances that he had experienced in the past. Ibrahim always sought to extend his knowledge to me and I appreciate even the small things he does to make the lab an exciting place to work in. He has always brought a kindness to the lab atmosphere, even when the days were gloomy or boring. I look forward to both our future successes.

I'd like to recognize the ECE staff and faculty as they have paved the road for me to continue in academia and not only extend our work to the world, but to gain intriguing insights into similar research abroad. To the professors, coordinators, office staff, advisors, and so many more, I must inform them all that my jubilation from writing this thesis is not without due recompense.

To Adam, Ali, and Jad, my immense gratitude and heartfelt appreciation are given. They have lent me their ears when I needed to assess how to travel on the road that has led me here. Their advice on how to continue moving forward will forever leave its mark on my academic career. To them and their families, I reach out my hand in joy as we will surely continue down this road together. Here's to all our future successes.

In Guha, Harsh, and Jai, I confide, that they accompany me in celebrating this academic achievement. It's what pushes us to the basketball court that also drives my passion in academic and professional work: enjoying what you do with the ones around you that always have and always will support you. Let this thesis be a testament to our dedication to our own individual work and to each other.

To Ali, Majd, and Andrei, my software programming and design skills would not have improved without your help. Their pursuit for excellence bolstered by endeavours in undergraduate studies as well as graduate research, and for that they have my solemn acknowledgements. May we continue to see success in all our career paths, no matter where the roads may take us.

From the support of Ahmed and Manur, my aptitude for learning was bolstered and there was never a dull moment. Their relentless help and advice for my future endeavours serves as a lesson to those who discredit the value of friends inside and outside the education and career setting. I also supplement this work with their names to show that there is more waiting in store for all of us.

To Tamara, I write these words as a reminder to what she stands for and what she has always wanted: the very best for me. This thesis is an omen of what is to come in our lives.

Through her unending generosity, kind words, and support, I cherish the values that she instilled in me and that she embodies in her own education and career. This pursuit in graduate studies, let alone my undergraduate degree, would not have been possible without her and her exemplary attitude towards others and towards life, and for whom those critical thinking skills are a speciality. Allow me to engrave these heartfelt words as a timeless reminder of how far I've come due to her endless support.

To Ahmad and Yazan, the finish line is approaching and my sincere appreciation is expressed. There were times I doubted if I should continue down this road, but they were adamant in assuring me that I had their full support no matter what. And so it goes, this road had its memorable times: in particular, those times were accompanied by four pieces of scrumptious, spicy fried chicken with a side of fries and a pop. Their professional endeavours in the workforce has inspired me to take on the most difficult of challenges in my studies, and I hope to continue to make them proud and admirable. Here's to our never-ending successes as a most formidable squad.

To my family and friends back home, I definitely assured, that there would be enough gratitude to spare because they are the driving life-force and the fuel that keeps me going. This document is an accomplishment we share together and no matter how far the world may separate us, we will band and unite together in our work and our celebrations. They remind me that we mustn't forget our place in the world, and how we can make a change no matter what we do. How ironic I chose this field of study: it has permitted us to stay in touch no matter where life takes us.

To my brother Zaki, I strive to properly convey the utmost gratitude. My overbearance as an older brother may seem an immense arrogance that he perceives, but he should most definitely know that my academic career and my entire future has an ulterior motive: to leave behind a wake of opportunity that will inspire him to accomplish more than he is already capable of. Let this thesis be a message to him that I wish to see the same successes from him, no matter what he does. May the academic community never take lightly understanding the support I've received from him just from living; may these words breathe more life into him.

Last, and certainly not least, this journey would not have been possible without my parents, Hassan and Najat. I struggle to write these words because the love and support they have given me to get to where I am is a debt I can never repay. My very first bit of professional help stems from their hands. My mother would help me with math homework coming into high school, until it got out of hand with more and more complex problems. My father trusted in my critical thinking and my use of pen and paper; rather, he delivered unto me memorable experiences in life that allowed me to apply my thought into practice, to take apart and put back together, to see the world as it is, unhinged as it may be. Seeing the completion of my graduate studies come to fruition has been a long-held dream of my parents, and I hope this stands to show them how much I have accomplished for myself, but also to make them proud. May they never forget how much I am indebted to them.

Allow me to express my warmest gratitude to all who have helped me get to where I am today. To those I haven't mentioned by name, they are no less important, and this work will serve as an eternal reminder to those who have always wanted the best in me.

To the world, let this be the procession of one keen individual and his supporters; to my faithful supporters, to our professional successes, and to our life-long happiness.

Here I am.

Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	Third Generation Partnership Project
5G	Fifth-Generation
5G+	Fifth-Generation and Beyond
5GC	Fifth-Generation Core
AF	Application Function
AI	Artificial Intelligence
AKA	Authentication and Key Agreement
AMF	Access and Mobility Management Function
AnLF	Analytics Logical Function
API	Application Programming Interface
AR	Augmented Reality
AUSF	Authentication Server Function
BSF	Binding Support Function
CAPEX	Capital Expenditure
CP	Control Plane
CUPS	Control and User Plane Separation
DL	Deep Learning
DN	Data Network
DRL	Deep Reinforcement Learning
eMBB	Enhanced Mobile Broadband
EPC	Evolved Packet Core
FL	Federated Learning
GTP	GPRS Tunneling Protocol
HTTP	Hypertext Transfer Protocol
IMSI	International Mobile Subscriber Identity
IoT	Internet-of-Things
IP	Internet Protocol
ITS	Intelligent Transportation System
KPI	Key Performance Indicator
LTE	Long-Term Evolution
MANO	Management and Orchestration
MEC	Multi-access Edge Computing

MILP	Mixed-Integer Linear Programming
ML	Machine Learning
MME	Mobility Management Entity
mMTC	Massive Machine-Type Communication
MTLF	Model Training Logical Function
NF	Network Function
NFV	Network Function Virtualization
NGAP	NG Application Protocol
NR	New Radio
NRF	Network Repository Function
NSA	Non-standalone
NSP	Network Service Provider
NSSF	Network Slice Selection Function
NWDAF	Network Data Analytics Function
OPEX	Operation Expenditure
PCF	Policy Control Function
PDU	Protocol Data Unit
PFCP	Packet Forwarding Control Protocol
QoS	Quality-of-Service
RAN	Radio Access Network
REST	Representational State Transfer
RL	Reinforcement Learning
SA	Standalone
SBA	Service-Based Architecture
SBI	Service-Based Interfaces
SDN	Software-Defined Networking
SIM	Subscriber Identification Module
SLA	Service Level Agreement
SMF	Session Management Function
SQN	Sequence Number
SSL	Secure Socket Layer
TCP	Transmission Control Protocol
UDM	Unified Data Management
UDR	Unified Data Repository
UDSF	Unstructured Data Storage Function
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communication
VNF	Virtual Network Function
VR	Virtual Reality
ZSM	Zero-Touch Network Service Management

Chapter 1

Introduction

The advent of fifth-generation (5G) mobile technology is situated to meet the rapid demands and high-end requirements of next-generation mobile networks. The continuous growth of the number of wireless devices, data usage, and expected Quality of Service (QoS) has influenced the evolution of cellular networks and spurred the development of state-of-the-art solutions to address it [1]. 5G networks will focus on enhancing consumer experience through uninterrupted communication services and device connectivity, along with connected intelligent transportation systems and a low-cost communication network-operator-centric infrastructure [2]. In addition, 5G networks are anticipated to realize features such as zero-latency (low latency in the order of 1 ms), high-speed transmission rates (in the order of Gigabits per second), 10-100 times higher data rate than 4G, 1000 times higher mobile data volume per area, and 99.999% availability [3]. These envisioned 5G networks, with their apparent and revolutionary advantages compared to 4G networks, require novel and demanding technologies, architectures, and methodologies, such as Network Function Virtualization (NFV) and Software-Defined Networks (SDN) [4]. Rather than a sheer enhancement of the 4G architecture (*i.e.,* additional capacity), these 5G networks will consolidate the conceptualization, visualization, and redesign of networking system architectures *en masse*.

The International Telecommunication Union (ITU) classifies 5G networking into three categories based on industrial and consumer demand: Ultra-Reliable Low Latency Communications (URLLC), Enhanced Mobile Broadband (eMBB), and Massive Machine-Type Communications (mMTC). The importance of URLLC is that it focuses on connections with ultra-low latency, where the data rate is not expected to be very high, but it must offer high mobility. Typical applications of URLLC involve mission-critical applications, such as remote medical assistance. In contrast, eMBB focuses on a higher data rate for larger payload applications, such as high-speed internet gaming, virtual reality (VR), and augmented reality (AR). Relative to URLLC, mMTC focuses on IoT connectivity (large number of devices), but with low reliability. In particular, it focuses on long-range communication with asynchronous access, which is intended for applications such as embedded, low-power devices [5].

In order to satisfy strict QoS requirements in URLLC, edge communications solutions have been provided by researchers to bring resources closer to UE devices [6]. An issue such as end-to-end delay and reliability can possibly be solved with scheduling method optimization in communication. In wide-area communication, the issue of precise and reliable communication between controllers and slaves could be mitigated by mobility forecasting methods to improve QoS [7]. The focus on core networks and core operations in modern networks is important for blurring the distinction that previously existed in wireless networks. Core resources are made closer to end users through edge computing, such as those located near base stations in Radio Access Networks (RANs); however much core functionality may be supported at the edge, it is not part of the core and may have its own set of issues. As well, core integration in edge computing is an important issue as it may increase the risk of compromising previously non-sensitive equipment and will impact integrity and confidentiality of future networks [8].

1.1 Research Contributions

The work outlined in the subsequent chapters introduces several research contributions, listed as the following:

- Chapter 2:
  - Provides an extensive analysis on the structure and operations of the 5G Core, the benefits and impact of 5G networks in modern applications, comparing and contrasting 5G architectures, detailing the advent and future paths for the Network Data Analytics Function (NWDAF), and the associated challenges with regards to maintaining QoS and user demand
  - Highlights the contributions of enabling technologies in 5G and Beyond networks to assess their impact on service providers' integration and provisioning of network microservices
  - Summarizes the role of artificial intelligence (AI) in the NWDAF with a focus on the benefits of AI in network data analytics and the major challenges preventing widespread adoption and implementation in future networks
  - Discusses the applications, benefits, and challenges of current and emerging 5G-enabled use cases
- Chapter 3:
  - Elucidates a preliminary analysis into the types of core network function data that can be collected by the NWDAF
  - Presents insights which can be drawn using the collected core network data
  - Discusses how the NWDAF can be used to influence Management and Orchestration (MANO) activities, such as core network function placement
  - Demonstrates an outlook for the state of future networks, the expected limitations of 5G, and the motivation sparking the initial discussion of 6G networks.
- Chapter 4:
  - Formulates a mixed-integer linear programming model for the AMF scaling optimization problem

- Develops a load-balancing mechanism for the AMF load-balancing module
- Displays the use of a functional NWDAF and 5G Core prototype to generate data for the scaling and load-balancing problems

Chapter 2

Network Data Analytics in Future Networks: Trends, Outlooks, and Future Directions for the 5G Core and Beyond Networks

The structure of this chapter is as follows:

Section 2.1 provides an architectural overview of the 5G Core and its Network Functions (NFs), including the NWDAF. Section 2.2 discusses the evolution from 4G to 5G to 6G and Beyond networks. Section 2.3 discusses the 5G Core in relation to enabling technologies and the challenges of implementing 5G microservices, along with outlining their performance and operational requirements. Section 2.4 discusses the role of AI in the NWDAF in order to realize zero-touch, fully automated networks, and presents the benefits, challenges and limitations, as well as the emerging topic of advanced intelligence. Finally, section 2.5 concludes the chapter.

2.1 Architecture of the 5G Core

The 5G network system is comprised of both the 5G Core (5GC), the Radio Access Network (RAN), and the User Equipment (UE). The 5G system is designed to support services and data connectivity, which would enable deployment using enabling technologies such as Network Function Virtualization (NFV) and Software Defined Networking (SDN). The need for these novel techniques is increasing due to the multitude of microservices offered by the service-based architecture of the 5G network [9].

The 5GC architecture definition, in accordance with 3GPP, uses a SBA framework, where the architectural elements are defined in terms of NFs rather than by traditional network entities. All the network functions communicate with one another via common interfaces or reference points. So, through this common interface, a network function provides services to other authorized network functions as necessary [10]. The 5G Core is composed of network functions with their individually associated microservices and responsibilities, including: the User Plane Function (UPF) which handles the user data, the Application Function (AF) which handles the applications, the external Data Network (DN), and other NFs (AMF, NRF, etc.) [11].

The service-based architecture (SBA) for 5G networks is beneficial for decoupling network functionalities to prioritize flexible service provisioning. A service can be defined as different capabilities, which are loosely coupled within the 5G network, operating independently of one another. Some important SBA framework sequences in 5G include service discovery, authorization, and registration. The service-based interfaces (SBI) in the 5G Core connect different network functions (NFs) to the Network Repository Function (NRF), which maintains NF profiles of available NF instances and supports NF discovery. The NRF stores the availability of the different NF services that can be conducted on each NF instance when they report to the NRF. In this context, one advantage of the 5G SBA is that it allows individual services to be deployed on demand. As well, each service can be configured and updated independently with minimal impact on other services. Some security risks with this architecture may involve some user privacy information transferred from one NF to another. Between NFs, an attacker can eavesdrop on security information where it lacks integrity protection. Consequently, the security context can be tampered with by attackers to make UE devices and NFs use different security context information than previously assigned. For example, the Access and Mobility Function (AMF) for one 5G Core operator may obtain subscriber data from the Unified Data Management (UDM) for another operator, which can lead to user data integrity breaches [12].

The 5G network SDN architecture is notably comprised of three layers: the control layer, the infrastructure layer, and the application layer. The layers can be expressed as the control plane, user plane, and the application plane, respectively. The control plane is responsible for linking the infrastructure layer and the application layer by open communication interfaces. The infrastructure layer contains forwarding elements, such as routers, access points, and switches, that can be categorized as the data plane. By design, the application layer satisfies user requirements related to consumer/business applications, which provision network resources and services. Using this architecture, some examples of SDN applications include cloud computing, load balancing, and network virtualization [13].

As in 4G Evolved Packet Core (EPC), the 5G Core infrastructure separates network functions between the Control Plane and the User Plane by the Control and User Plane Separation (CUPS) design principle. CUPS allows for independent scalability for each network function and for flexibility in centralized or edge deployments [12]. In the 5G standalone (SA) architecture, the 5G NR cells and the 5G Core network operate alone such that the NR cells are used for both the control and user planes. Contrastingly, the 5G non-standalone (NSA) architecture combines the NR radio cells using dual-connectivity to provide radio access. Based on the network operator, the core network could, then, either be 5G Core or EPC. The NSA architecture requires close integration with the LTE RAN in 4G, but can provide 5G Core (5GC) functionality to customers' needs without combining resources with the current EPC (as in 4G) [12]. Regarding the RAN, the SA architecture supports simple management and handover between 4G and 5G, but its disadvantage is that it will not be able to support the existing LTE

RAN deployment if NR cells are used. The NSA architecture can support the existing LTE deployment, but the disadvantage is that the LTE and NR cells must be closely integrated and the end-user experience may be degraded. Regarding the core network, the EPC architecture supports its current EPC deployment, but does not provide optional cloud support. In contrast, the 5G Core easily supports cloud-native multi-access functionality; however, an entirely new deployment is essential [14].

Network slicing is an important feature in the 5G Core, which enables a large variety of services with diverse performance requirements by network virtualization. The network can be typically viewed as an encapsulated slice and its services are bundled with proprietary hardware supported by telecommunications equipment providers [15]. With the network virtualization technology in 5G, open-networking software can be deployed flexibly on commodity hardware to offer a multi-slice 5G core architecture where each slice can offer a different set of network services [16]. For example, by provisioning the User Plane Functions (UPFs) with different QoS requirements, the performance of such a multi-slice system can be compared with that of a single-slice architecture under the same resource assignment. Furthermore, the proposed system achieves better performance by slicing one UPF into three with proper resource allocation [9].

In the context of UE operation, network slicing involves grouping devices, into a slice, with similar performance requirements, such as delay, throughput, and transmission rate. However, from the networking perspective, network slicing can be viewed as dividing physical networks into multiple, isolated virtual networks. This network slicing architecture consists of three layers: the service instance layer, the network service instance layer, and the resource layer. The service instance layer contains the consumer/business services, each operating on an individual service instance. The network service instance layer outlines the network characteristics which are defined by a network slice blueprint, a complete description of the structure and configuration of a network slice instance, and required by a service instance. Finally, the resource layer can be expressed as the underlying network infrastructure of NFs operating under the network service instance layer [13].

Network slicing in the RAN is intended for flexible resource management and sharing. This slicing implementation is best approached by a software-defined RAN. Resource sharing between different slices is accomplished by controller scheduling and allocation within the RAN. The controller allocates resources to a network slice according to service requests in response to events such as an increase in traffic load. For example, RAN slicing requires the decomposition of NFs in order to determine the hardware specifications and requirements for each functionality [6].

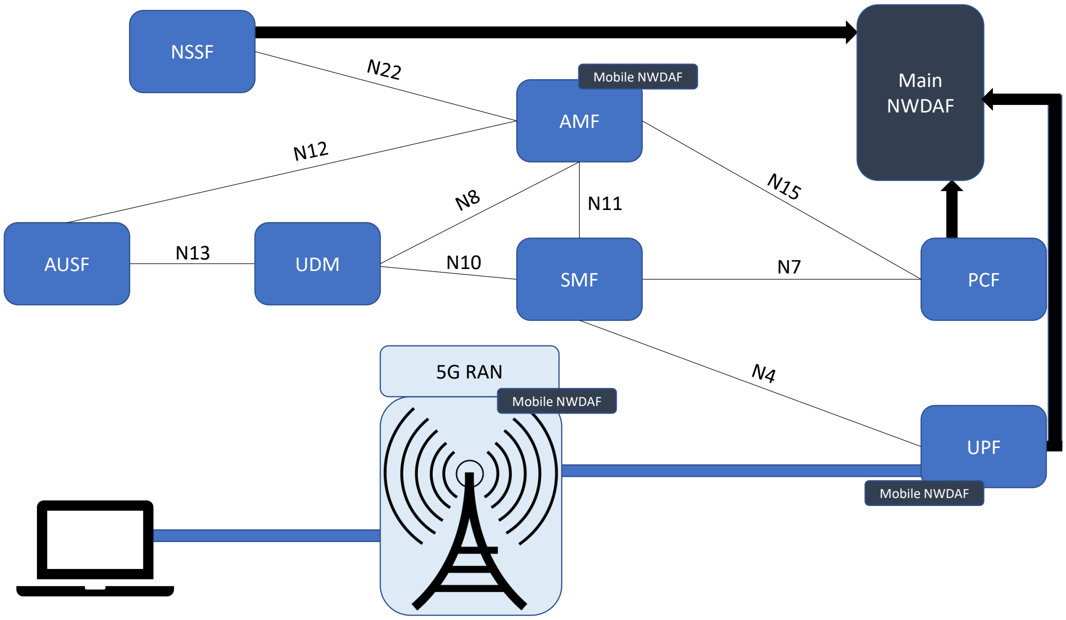

Figure 2.1: 5G Core Architecture with Distributed NWDAF

2.1.1 Network Functions

The following subsections detail the core components, and any standard names for NF-to-NF interaction interfaces outlined by the 3GPP, of a 5G core network:

2.1.1.1 Access and Mobility Management Function

The Access and Mobility Management Function (AMF) is involved with most signalling call flows in a 5G core network and supports encrypted connections between UE devices, handling their registrations, authentications, and radio cell transfers in the network. As well, it also supports activating UE devices in idle mode. The AMF interacts with the radio network through the N2 interface and with UE devices through the N1 interface. The AMF's connections to all other NFs are managed through service-based interfaces. Compared to its EPC equivalent, the Mobile Management Entity (MME), the AMF does not handle session management, but rather forward session-management related messages for the UE devices to the Session Management Function (SMF). The AMF does allow UE devices to be authenticated, but it does not handle the authentication; it orders this service to be performed by the Authentication Server Function (AUSF) [17].

2.1.1.2 Session Management Function

The Session Management Function (SMF) manages the end user device sessions in the network. In particular, the SMF is involved with the instantiation, modification, and release of a given session and IP address allocation for each session [17]. As well as communicating with other NFs through the service-based interface, the SMF is responsible for the selection and control of different User Plane Functions (UPFs) across the network through the N4 interface. The SMF's control of the UPF(s) traffic steering and enforcement, and their associated configurations. Finally, the SMF interacts with the Policy Control Function (PCF) for policy control of user sessions [18].

2.1.1.3 User Plane Function

The User Plane Function (UPF) is primarily concerned with processing and forwarding user data. Most of its interactions are with the SMF since the UPF's functionalities are controlled there [17]. The UPF will connect to external IP networks and acts as a bridge point for devices connecting to other external networks through the 5G Core. For example, an IP packet with the destination address of a given UE device will always be routed from the Internet to a specific UPF, regardless of UE mobility. With the forwarded user data, the UPF generates traffic usage reports (these may also include device charging data) and sends these to the SMF. The UPF also performs a packet inspection, which analyzes user data packets for the aforementioned traffic usage data reporting, but it can also be used as an input to guide policy decisions. Some examples of user policies include traffic redirection, enforcing, and applying data rate limits. Finally, the UPF can mark packets with Quality-of-Service (QoS) priorities and schemes in order for the radio network to handle packet priorities during network traffic congestion [19].

2.1.1.4 Unified Data Management

The Unified Data Management (UDM) Function is front-end functional interface for user subscription data stored in the Unified Data Repository (UDR) and executes AMF-requested functions [17]. It also generates authentication data for UE device attachment and can authorize user access based on their subscription data. An example of different access rules can be between home and roaming subscribers. If there is more than one instance of the AMF and/or SMF in the network, the UDM tracks which instances serve a specific device [20].

2.1.1.5 Unified Data Repository

The Unified Data Repository (UDR) acts as a database for subscription data, network policies, and user policies. The UDR acts as a central point for data storage and access for the UDM, PCF, and NRF, all of which will use these important data for their inter-NF services [17].

2.1.1.6 Authentication Server Function

The Authentication Server Function (AUSF) has limited, albeit important, functionality to the 5G Core. The AUSF provides an NF service that authenticates a UE device, using authentication credentials that are created by the UDM. In addition, the AUSF generates secure and encrypted messages to provide roaming-specific information and other associated parameters to UE devices [17].

2.1.1.7 Network Repository Function

The Network Repository Function (NRF) is a new key component of the 5G service-based architecture. The NRF actively maintains an updated repository of available elements in a 5G operator network. As well, the NRF records NF statuses, statuses of their services, and when these services are instantiated, scaled, and terminated. The NRF reduces the burden of consumer NFs on processing this data, as it prioritizes NF discovery service results based on location, capacity, network load, and priority. Registration, subscription, and discovery are key services provided by the NRF, but it can also support network traffic logging, tracing, monitoring, and visibility, making it an important source for network data analytics [21].

2.1.2 5G Network Data Analytics Function

Given its prevalence in AI-assisted applications of the 5G Core network and advanced data analytics applications in modern networks, the Network Data Analytics Function (NWDAF) is presented here in its own section of focus in the scope of this chapter. As an underlying function solely responsible for data analytics and network learning, the NWDAF represents operator-managed network analytics as a logical function. The NWDAF provides slice-specific network data analytics to any given NF. As well, the NWDAF provides network analytics information to NFs on a network slice instance level and it is not required to be aware of the current subscribers using the slice. The function also notifies NFs with slice-specific network status analytic information for any that are subscribed to it. NFs may also collect network status analytic information directly from the NWDAF [22].

In the 5G Core, both the Policy Control Function (PCF) and the Network Slice Selection Function (NSSF) are consumers of network analytics. The PCF may use that data in its policy decisions, and the NSSF may use the load-level information provided by the NWDAF for slice selection [10].

The NWDAF may be comprised of the following logical functions: the Analytics logical function (AnLF) and the Model Training logical function (MTLF). The AnLF performs inferencing (predictions based on analytics consumers' requests) on derived analytics information and statistics. The MTLF trains machine learning (ML) models on analytics information, which is either statistical information of historic events or predictive information for the future [23].

Industrial NWDAF implementations provide closed-loop automation for third-party NFs and solutions inside the 5G Core. In particular, these NWDAFs are intended for continuous monitoring of every NF, network slice, and UE device and use a variety of KPIs to measure network performance. The real-time KPIs can be used to automate network issue resolution, while ML/AI predictive analytics can be used to predict future network issues. Predictive analytics may also provide anomaly detection to be used for automating mitigation [24].

Figure 2.1 illustrates the 5G Core architecture with a distributed NWDAF as outlined throughout the section. As seen, the main reference points for all NFs are displayed, along with edge placements for NWDAF data collection. The NWDAF was originally defined as a centralized network function for data aggregation and analytics [23], but in order to reduce network resource usage and prevent overloading, the NWDAF is distributed, as shown in Figure 2.1, and is structured by local models communicating with the main NWDAF. This design is particularly useful for federated machine learning techniques, which trains algorithms across multiple distributed and decentralized nodes [25].

The following subsections outline the services offered by the NWDAF in accordance with the 3GPP standard and specifications as of June 2022 [23]:

2.1.2.1 Analytics Subscription

The *Nnwdaf_AnalyticsSubscription* service enables consumers of services to subscribe to or unsubscribe to notifications from the NWDAF. It can also transfer these subscriptions between different NWDAF instances or implementations. An important example of subscribing to such notifications is binding to network congestion events specific to a network slice. The types of observed network events include slice load level information, network slice instance load level information, NF load, network performance, UE mobility, UE communication, user data congestion, and QoS sustainability in the network [26].

2.1.2.2 Analytics Information

The *Nnwdaf_AnalyticsInfo* service enables NF service consumers to request and retrieve analytics information from the NWDAF. The NWDAF begins the data collection process to gather the necessary body parameters and data fields needed for an analytics information request, then exports this to its own data repository. Afterwards, the NWDAF employs an analyzer, or an algorithm to form behavioural patterns of the data, and the processed data is provided to the consumer. Some examples of use cases where this service is employed include automated policy control (in conjunction with the PCF) and automated network slice selection (in conjunction with the NSSF). It can also be used to track UE access and mobility for the purpose of scaling and policy decision-making [24].

2.1.2.3 Data Management

The *Nnwdaf_DataManagement* service allows analytics data consumers to subscribe/unsubscribe to and be notified about data exposed by the NWDAF or fetch the subscribed data. As well, it enables the NF services consumer to request the generation of bulked data for event IDs and analytics IDs from NFs, and retrieve the requested data [23].

2.1.2.4 ML Model Provisioning

The *Nnwdaf_MLModelProvision* service permits the service consumer to receive a notification whenever an ML model is available, provided with specific parameters in the subscription request. When the subscription is accepted by the NWDAF containing a MTLF, the AnLF receives from the NWDAF an identifier (a Subscription Correlation ID), which allows further operations, such as modification or deletion, of the given subscription. The modification of an ML model subscription can be enforced by the NWDAF based on operator policy and configuration [23].

2.1.2.5 ML Model Information

The *Nnwdaf_MLModelInfo* service allows analytics data consumers to request and retrieve analytics information from the NWDAF pertaining to specific ML model information from the MTLF. The consumer, in their request, specifies ML Model Filter information to gather detailed info by S-NSSAIs, service areas, and Analytics IDs [23].

2.2 Evolution from 4G to 5G to 6G and Beyond

The 5G system, as a next-generation network, is developed based on the successful experiences and technologies of the previous 4G generation. For 5G, the evolutionary challenge is avoiding the limitations of the previous system. Some obstacles include the limitations of the 4G architecture, home network control, malicious attacks on 4G and 4G RAN security, and user data integrity breaches [27].

One major limitation of the 4G architecture is the security protection measures which have been revised since their implementation in 3G networks. The Authentication and Key Agreement (AKA) protocol in 4G networks, which is symmetric and key-based, improves on signaling overhead and computational resource efficiency when compared to public key-based mechanisms [28]; however, it has been shown that new privacy threats and known attacks on the 4G AKA protocol can be prevalent in upcoming 5G AKA protocols. A UE device that connects to its serving network carries a Universal SIM, or USIM, which is necessary for symmetric encryption and mutual authentication. The USIM stores the International Mobile Subscriber Identity (IMSI), the secret symmetric key between the UE and the network, K_{IMSI}, and a 48-bit counter for replay protection known as a Sequence Number (SQN). The identity request procedure, which is unprotected and broadcast over-the-air, is subject to ”IMSI-catcher” attacks, which allows an attacker to track subscribers in particular geographic areas [29]. Accordingly, the 3GPP modified the identity request phase of the AKA protocol to strengthen the privacy protection requirement of 5G networks [30].

For 6G networks and beyond, user data protection is crucial to AI-assisted network operations as it is paramount for future networks to prevent data integrity breaches. User data integrity breaches are comprised of internal attacks involving resource access or external attacks through security protocols. In 4G networks, temporary identifiers, which are encrypted and always updated to prevent tracking, are used to protect against subscription identification leaking, where a user identity can be captured by an attacker during transmission; this is ideal for passive attacks, but not for active attacks. If the temporary identity is foreign or abnormal, the user must contact the network using the permanent identity to avoid permanently locking the user. The mechanism can be exploited by active attacks that trace users through the aforementioned IMSI interception [31]. 6G and Beyond networks must be compliant with Internet-of-Things (IoT) standards and so, machine learning models are tailored to ensure data integrity and improve end-to-end communications security [32].

2.3 5G Core and Enabling Technologies

The 5G Core (5GC) is designed to be “cloud native”, where NFV (Network Function Virtualization) is leveraged to create network slices. A 5G Core slice is composed of a collection of 5G Core VNFs that are chained together to support a specific use case [16]. One of the major characteristics of 5G Core, CUPS, decouples a 5G system into two parts, deploys the Control Plane (CP) as a common slice, and configures User Plane (UP) into multiple customized slices, each with different bandwidth requirements. Network slicing involves the instantiation of several separate logical mobile networks hosted atop the same physical infrastructure [10].

A network slice consists of a group of logical NFs that are independent and perform their relevant tasks, enabling a slice to deliver services according to different service level agreements (SLAs) [33].

5GC holds a key role in realizing the full potential of 5G services. 5G NSA (non-standalone) deployment leaning on legacy LTE network and EPC (Evolved Packet Core) allows for a quick launch of 5G services, but also hinders the realization of 5G's full potential [34]. In 5GC, a cloud native design is introduced to enable flexible scaling and upgrades. The fundamental concept of a cloud native 5GC is defined as "stateless microservices deployed in a container-based architecture". A Network Function (NF) is comprised of small service units called NF services and store their state information in a central database called Unstructured Data Storage Function (UDSF), which turns the network function stateless itself. Stateless NFs can be scaled with ease and specific NFs can be isolated in case of failures, which makes an uninterrupted service possible [35]. Each micro-service runs in a container and is independently scalable and reusable. These design characteristics enable the flexible launch of new services, faster time-to-market, and offers enhanced scalability [36]. As a result, the 5GC functions can be quickly created, deployed, and scaled, using automated lifecycle management. With the introduction of 5GC and the standalone network, end-to-end network slicing allows a network to suspend and resume from an inactive state, so as to allow a UE (User Equipment) device to return to a connected state as soon as possible from an inactive state. Accordingly, this leads to significant reduction in RRC signaling, and therefore, latency and battery consumption are reduced as well [37].

Network Function Virtualization (NFV) as proposed by ETSI in 2012 [38, 39] defines the decoupling of Network Functions (NFs) from their underlying hardware, and the creation of Virtualized NFs (VNFs) executed as software-based applications on commercial equipment such as datacenter servers. The motivation behind switching to an NFV-enabled environment was inspired by the increasing network connectivity demand along with the various potential benefits experienced by Network Service Providers (NSPs). From a network Management and Orchestration (MANO) perspective, NFV can provide numerous benefits, including the reduction of Capital and Operation Expenditures, enhanced scalability, reduced product development cycle and time to market for new technologies, as well as enhanced flexibility [40]. The introduction of 5G+ networks has raised several challenges regarding the MANO activities of VNFs. These challenges can be classified into three main categories: Performance Requirements, Operational Requirements, and Practical Challenges.

2.3.1 Performance Requirements

5G networks require increasingly stringent performance requirements to support the unprecedented growth of network traffic and the number of connected devices [41, 42]. These new requirements pose significant challenges to NSPs in terms of VNF MANO. Firstly, given the requirement of Ultra-Reliable Low-Latency Communication (URLLC), NSPs are required to push resources to the extremities of the network in the form of lightweight points of presence leveraging Multi-access Edge Computing (MEC) [43, 44]. This requirement complicates VNF MANO as the lightweight points of presence have limited resources available for hosting VNFs, meaning that priority should be given to critical services. Additionally, emerging use cases such as Intelligent Transportation Systems (ITSs) and the Industrial/Internet of Things (I/IoT) which leverage MEC resources to collect and relay large amounts of data, further complicates VNF MANO as communication efficiency needs to be taken into consideration to ensure QoS preservation. Regarding QoS guarantees, 5G networks will require five nines (99.999%) of availability, which translates to less than 6 minutes of downtime each year [45, 46]. Given the critical services (*e.g.*, emergency, financial) as well as the emerging ITS use case, adhering to this requirement is paramount to preserving public safety. To this end, NSPs need to explore resilient, and reliable VNF MANO solutions such as redundant instance placement and robust optimization as proposed in [47, 48] to ensure their networks can attain these levels of availability. Additionally, NSPs need to be proactively sensing the network for adverse conditions and be ready to prevent any perceived fault or failure from materializing by performing dynamic corrective VNF MANO operations, including migration and scaling.

2.3.2 Operational Requirements

Traditional networks have considered a tradeoff between scalability, efficiency, and reliability where the best-case scenario was a network possessing two of the mentioned three attributes [49]. Considering the profound impact an event such as the COVID-19 pandemic had on NSP operations, widespread disruptions and changes in user behaviour can increase in frequency and severity as user dependence on communication networks increases [50]. Given this increasing dependence and demand on networks, settling for a fraction of these attributes is no longer feasible from an operational standpoint. The impact of network transformation, motivated by next-generation networking technologies and use cases, has established the fact that the complexity of future networks and systems has greatly surpassed the human capacity for manual management [51, 52]. As such, NSPs are tasked with developing methods that reduce the complexity of the network while simultaneously removing the manual element of MANO through network automation. Increasing levels of automation have several benefits, including time and cost savings, rapid service deployment, as well as enabling humans to redirect their focus to more complicated tasks that require a more profound understanding and analysis [51, 53]. To this end, Zero-Touch Network Service Management (ZSM) has been proposed as an architectural solution to achieve full network automation [54, 52].

Before discussing ZSM, it is important to consider two critical attributes of reliable networks, namely robustness and resilience. Robustness defines the network's ability to survive a given failure, whereas resilience considers the network's ability to recover from a failure [55]. While both of these attributes are critical, they are still considered reactive in the sense that they focus on surviving an error and re-establishing an adequate level of performance. While ZSM incorporates both robustness and resilience, its main focus is to sense the network and prevent any adverse conditions from materializing. This is accomplished through the four major pillars of ZSM, namely, self-configuration, self-monitoring, self-healing, and self-optimization

[51]. By leveraging ML/AI and advanced analytics, ZSM enables the sensing and prediction of adverse network events (*e.g.*, outages and demand spikes) and the execution of proactive corrective measures (in real-time), which mitigate the event before end users are affected [50, 54]. It should be noted that a challenge arises since healing actions may lead to unintended consequences by impacting other network functionalities and services; this challenge highlights the importance of full network automation as the autonomous orchestrator requires a universal understanding of the network in order to select the appropriate action without deteriorating the performance of other network elements [54]. By implementing ZSM, NSPs aren't required to make a tradeoff between the scalability, efficiency, and reliability of their networks since the network will be entirely intent-driven [49]. In practice, ZSM has profound effects on the MANO of NFV in 5G. The rapid technological advances have significantly increased the complexity of MANO tasks to the point where they are too critical and time-sensitive to allow any manual interaction. Some examples of such tasks, specifically, 5G usage scenarios for ZSM, are outlined below as identified by the ETSI technical specification group on ZSM [54].

2.3.2.1 Network Slice Lifecycle Management

The lifecycle management of a network slice is a critical and complicated process, especially when dealing with network slices with non-standard characteristics. This complexity is further augmented by highly dynamic demands which require constant scaling actions. In order for ZSM to handle network slice MANO, it must correctly identify and manage all VNFs and resources related to the slice. Furthermore, it should have the ability to automatically analyze a slice's requirement to determine which VNFs and corresponding resources are required. Additionally, in order to ensure constant performance, ZSM should be able to scale VNFs while adhering to resource limits and NSP objectives. In terms of performance continuity, ZSM should be able to perform real-time corrective action to reallocate network resources and network slice reconfiguration without experiencing downtime.

Isolation management is another important aspect of the network slice lifecycle. Ideally, each instance in a given network slice should be protected against interference from other slice instances. In practice, total isolation is not feasible and instead, instance interdependency is leveraged to restrict the interference to an acceptable threshold. The aspect of ZSM relating to isolation presents itself as constant monitoring and analysis of the underlying infrastructure status. Through this monitoring, any adverse conditions leading to a violation of the acceptable threshold can be countered such that performance deterioration is mitigated and performance recovery is achieved. This can be accomplished through slice VNF reconfiguration and resource allocation. The ZSM architecture uses network slice monitoring to gain insights into the performance of the slice as a whole and of individual instances. Through the collection of data (*i.e.,* KPI, fault, *etc.*) an autonomous agent can perform real-time diagnosis of network issues and determine an appropriate solution; however, as previously mentioned, this is a reactive technique as it applies to an observed adverse network condition. One of the main targets of ZSM is the notion of preventative and predictive maintenance, which aims at predicting when a certain network instance will fail before it does. This predictive capability will allow NSPs to perform planned maintenance instead of reactive maintenance and maintain service continuity and performance.

2.3.3 Practical Challenges

One of the most researched challenges in NFV MANO is the VNF placement problem which determines the optimal placement of VNFs on network nodes to deliver a service. This problem has been previously defined as NP-hard [56] and has traditionally been formulated as an optimization problem. Due to the complexity of the problem itself, optimal solutions are often considered practically infeasible as the time required to determine them scales poorly with the size of the network. To mitigate this limitation, near-optimal heuristic solutions have been proposed to achieve a feasible solution in acceptable time. However, as 5G networks take shape, the complexity of both the network itself as well as the new and emerging use cases coupled with the increasingly stringent performance requirements suggest the need for the performance of the optimal solution combined with a lower execution time. This is reflected through the requirement of real-time and dynamic operational provisioning.

One of the major challenges that 5G networks face is the high-speed and bandwidth demands from large applications and in the field of Internet of Things (IoT). Large-scale industrial applications and autonomous cars can consume vast amounts of data in just few minutes, so the 5G low-latency transmission and connectivity will add to this data throughput. Cloud infrastructure support will be needed to support fast data reads and writes with low-latency compute and storage architectures on cloud. As well, the 5G architecture needs to be defined and constructed in such a fashion that big data is collected for analytics support that may already exist for distributed network and application intelligence use-cases. 5G network data can raise numerous security concerns amidst any applications today, so it is important to safeguard user privacy or company data without any compromise. Building a secure and robust infrastructure from systems to applications is critical in 5G architecture and design [34].

The adoption of NFV is another challenge to the nature of the 5G network and its associated requirements. NFV requires implementing layers that are typically deployed on a provided cloud, IaaS (Infrastructure-as-a-Service) and / or Kubernetes – a container orchestration platform. These techniques, however, are known for being difficult to operate, especially in distributed environments with hundreds of nodes. Also, many use cases require certain extensions, such as SDN, to be enabled which adds further difficulty to the process [57].

Contrary to early generations of NFV technology, 5G brings in specific requirements for VNFs onboarding and orchestration. Instead of running legacy monolithic software blocks in a virtualized environment, 5G VNFs are designed with the intent to be fully cloud-native. This means that they have to be re-designed based on the aforementioned microservice architecture. Those microservices will run inside VMs or containers with high availability and scalability, which introduces another layer of complexity over the already complex cloud environment [58].

Table 2.1 outlines the various challenges discussed throughout this section regarding 5G

Table 2.1: Challenges of Implementing 5G Microservices

Category	Challenge
Performance Requirements	MEC-enabled services Resource-constrained network edge QoS Preservation High Availability Network Sensing
Operational Requirements	Network Automation Complexity Reduction Robustness Zero-Service Network Management: **-** Self-Configuration **-** Self-Monitoring **-** Self-Optimization Network Slice Lifecycle Management Multi-Vendor Network Interoperability
Practical Challenges	Dynamic Network Service Provisioning IoT and Big Data Applications User and Data Privacy

microservice implementation.

2.4 NWDAF and Zero-Touch Full Automation of 5G Network and Service Management

The full automation of 5G networks and services requires careful consideration before it can be realized. Major challenges such as multi-vendor networks, URLLC, and edge networks add significant intricacies to the problem. Firstly, when dealing with multi-vendor networks distributed across various domains, certain aspects of the network slices will need to be managed by vertical industries through third-party interfaces. As such, a ZSM agent should be able to safely and securely manage and monitor these interactions. Additionally, such an agent will need to consider the management of multiple simultaneous requirements (*i.e.*, ULLC, URLLC, *etc.*) encouraged by the constantly developing technological landscape of 5G networks. To this end, the autonomous management of edge networks and services becomes increasingly complex. An agent should be able to perform high-risk actions such as automatic software deployment and live updates while ensuring that the instance's performance suffers no degradation. Additionally, an agent should be able to rollback any erroneous updates that lead to problematic performance and quickly restore service. When dealing with the unprecedented demand for service connectivity and the rapid increase in the number of connected devices, a ZSM agent should also be capable of flexible and elastic VNF provisioning. To this end, there needs to be an automatic flow of data and information between instances without human intervention to make optimal MANO decisions. Additionally, demand forecasting should be used to influence planning decisions based on the predicted traffic and network load. Another critical consideration is the rapid coordination of deployed VNFs to adhere to customer requirements regarding time to market.

2.4.1 Artificial Intelligence in the 5G Core

The widespread use of AI in 5G networks and systems is one of the defining characteristics of this paradigm-shifting technology. According to reported statistics, by 2025, it is projected that the telecom industry will invest USD 36.7B in AI through software and hardware investments as well as AI services [59]. In order to prepare for this AI revolution, telecom operators internationally need to begin strategic planning for the development, adoption, and integration of AI into their networks and practices. It is estimated that the majority of major network operators have already initiated the planning and integration phase of AI to improve their network management and operation [59]. The envisioned AI network ecosystem will consist of agents being fed data related to the network, including network measurements and statistics, resource utilization, traffic patterns, and alarms and will conduct inferencing to provide network automation through management and orchestration tasks such as resource optimization and system reconfiguration [59]. The following section will outline the various benefits and challenges associated with the adoption of AI in 5G networks, outline some of the key intelligence technologies being considered, and discuss some of the emerging use cases and applications currently being considered and implemented.

In recent years, AI has garnered significant attention across various industries. With the ability to automatically extract information from complex data and systems, conduct inferencing, and provide the user with a decision, the benefits of AI are numerous. For the purposes of this chapter, the benefits of AI, specifically applied to 5G networks, will be categorized and classified in terms of operational and business benefits.

2.4.1.1 Operational Benefits

The operational benefits of AI in 5G networks consider the added value the AI system provides in terms of the management and orchestration of networks. As previously mentioned, the paradigm of ZSM, enabled by AI, provides the major benefit of reliable and robust network automation; as such, one of the main benefits of AI is the ability to take proactive and predictive measures to ensure the optimization of network performance [60]. Some methods of network performance optimization include the reduction of power consumption through enhanced algorithmic performance, the maximization of throughput through optimal traffic routing and infrastructure placement, as well as the ability to support an increasingly dense number of users [60]. As demonstrated, the plethora of potential operational benefits of AI makes it an appealing and necessary technology for the feasible realization of stringent 5G network performance requirements.

2.4.1.2 Business Benefits

The use of AI in 5G networks also presents NSPs with various alluring business benefits. Firstly, through the optimization of network operations, NSPs experience a reduction in their capital and operational expenses [60]. Furthermore, AI enables the improvement of QoS and QoE, leading to better service delivered to the end-user coupled with a reduction in the number of SLA violations incurred. Ultimately, these two benefits suggest that NSPs will be able to maximize their revenue through expense reduction, while also delivering superior service to end-users, leading to improved customer satisfaction. Finally, another key major business advantage is the creation of new revenue streams realized through advanced user behaviour understanding and the ability to provide new services and use cases to customers such as connected vehicles and Industry 4.0 [60, 61]. Coupled with the operational benefits, the added business-related benefits of AI provide NSPs with the flexibility to explore new services and revenue streams that will diversify and strengthen their operations.

Table 2.2 summarizes the various operational and business benefits discussed throughout this section.

Table 2.2: Benefits of AI in 5G Networks

Category	**Benefit**
Operational	Network Automation Proactive Management Preventative Maintenance Optimized Power Consumption Performance Optimization Improved Network Strength
Business	CAPEX / OPEX Reduction QoS / QoE Improvements SLA Violation Reduction Improved Customer Satisfaction Creation of New Revenue Streams Enhanced User Behaviour Understanding

2.4.2 Challenges and Limitations of AI in 5G and Beyond

Despite the various benefits, AI adoption in 5G and Beyond networks can offer, there are still several existing challenges and limitations which must be addressed. These challenges and limitations are classified into four distinct categories: data, AI lifecycle management, privacy and optics, and operational considerations.

2.4.2.1 Data

One of the main challenges plaguing AI implementations across all fields relates to the availability of high-quality data. In networking applications, the effect of this challenge is amplified due to the lack of publicly available data sets caused by data privacy concerns and proprietary confidentiality [61]. Aside from inadequate data availability, the collection process for data in 5G networks is increasingly difficult as the increasing complexity and number of users classify the generated data as Big Data [60]. In order to effectively collect the required data to build and train AI agents, data collection interfaces need to be deployed throughout the network and constantly be monitored [62]. Once collected, the data will need to be processed, structured, and stored to enable multiple stakeholders to access the relevant data for their needs. Due to the volume and velocity at which the data is generated, this is not a trivial task. Additionally, the storage and transfer of such large amounts of data at the resource-constrained network edge are infeasible as it utilizes the limited and valuable storage, processing, and communication resources. In order to adequately address the aforementioned data challenges, significant work must be put into improving data quality and availability, as well as developing distributed and decentralized intelligence agents that do not require the transfer of data to centralized locations [63].

2.4.2.2 AI Lifecycle Management

One of the added complexities when adopting and actively implementing AI is the added lifecycle management tasks which differ from traditional software lifecycle management. In terms of AI lifecycle management, the first step that must be considered is the selection of a model. This task is especially complex given the various considerations which must be made. Firstly, it should be noted that the well-known ‘no free lunch theorem’ states that if averaged across every data instance, all ML will perform the same. This being said, the goal of an AI implementation should be to determine the best performing model for a given set of data [60]. Additionally, operational constraints such as training time, model complexity, inference time, and acceptable performance complicate the selection process further. Another challenge is determining the optimal set of hyperparameters to improve model performance. As model complexity increases, so too does the effort required to determine the optimal hyperparameters as the hyperparameter state-space greatly increases. To this end, techniques such as metaheuristic evolutionary algorithms have been proposed due to their improved convergence time [64]. A final challenge relating to the lifecycle management of AI is the presence of model drift, where changes in the deployed environment create a gap between what the model was built for and where it is being applied. All types of model drift eventually lead to performance degradation and can significantly impact the decision-making process of AI systems. As demonstrated, the lifecycle management of AI is quite complex and therefore requires additional considerations when implementing to ensure the lasting performance of the system.

2.4.2.3 Privacy and Optics

The next set of AI-related challenges considers the privacy and optics of such systems. Given the services provided through networks, especially critical services, including emergency and financial, much of the generated data is considered highly sensitive and has increased privacy measures. Considering the next-generation use cases and systems, including connected and autonomous vehicles, system and data privacy is paramount to ensuring the safety of the public; however, privacy doesn't only affect the data itself, but also the types of models that can be used and the safeguards put into place. An example of a challenge requiring a safeguard would be the prevention of data reconstruction where malicious users can extract fragments of data used to train the model [65]. Additionally, security measures need to be put into place to prevent malicious entities from accessing and tampering with a model or its training data, leading to performance degradation or more significantly manipulated decision-making. Another critical consideration when implementing AI is the notion of interpretability which pertains to a user's ability to explain a decision reached by the autonomous agent and the factors leading to said decision. It is a well-known fact that as model complexity increases, so too does the capacity for inference which ultimately leads to improved performance; however, this is at the cost of interpretability. The less interpretable a model is, the more difficult it is to adopt due to a lack of trust [61]. As such, when using more complex models, human involvement through performance monitoring and intervention will be required to build trust in the newly adopted AI technology.

2.4.2.4 Operational Considerations

There are additional challenges relating to the practical implementation of AI in 5G networks that must be considered. Firstly, a critical hindrance to the widespread adoption of AI exists because of standard fragmentation [61]. There are currently various efforts in both standardization and open-source development; however, these efforts are often developed in isolation and either overlap or are not interoperable, making it increasingly complex and difficult for positive industry reception and adoption. Additionally, with various vendors developing their own AI services and platforms, NSPs face vendor lock-in and interoperability concerns. This is counterintuitive considering that one of the goals of 5G networks and NFV is to avoid vendor lock-in and improve flexibility. Moving forward, efforts need to be made to consolidate fragmented standards and create a unified 5G AI standardization team that will ensure ubiquitous standards across all NSPs internationally.

Table 2.3 summarizes the current challenges preventing widespread AI adoption in 5G networks presented throughout this section regarding data, AI lifecycle management, privacy and optics, as well as operation considerations.

2.4.3 Advanced Intelligence in the NWDAF

Advanced intelligence techniques are required in 5G and Beyond networks to address the aforementioned limitations of conventional ML techniques provided by the NWDAF. Manias and Shami have identified two such techniques: reinforcement and federated learning, which are expected to be an integral part of NFV MANO and next-generation 5G use cases [63]. Boasting

Table 2.3: Challenges of AI Adoption in 5G Networks

Category	**Challenge/Limitation**
Data	High Quality Data Availability Complex System Data Collection Big Data Processing and Storage Data Structuring Data Accessibility Resource-Constrained Network Edge
AI Lifecycle Management	Model Selection Model Operational Constraints: **-** Training Time **-** Complexity **-** Inference Time **-** Acceptable Performance Hyperparameter Optimization Model Drift
Privacy & Optics	Data Privacy Model Safeguards Model Training Tampering Model Interpretability Public Trust
Operational Considerations	Standard Fragmentation Interoperability Vendor Lock-In

benefits such as domain adaptability and distributed intelligence, these methods are essential for the establishment of intelligence and automation in future networks. The following section will outline these techniques and discuss their potential benefits.

2.4.3.1 Federated Learning

Federated Learning (FL) is a distributed and decentralized intelligence technique proposed in 2017 [66]. The main actors in this technique are the federated nodes and the aggregation agent. Initially, the aggregation agent sends a global model to each federated node that actively collects and processes its own data. The nodes train the global model and develop a local model using their collected data. Once trained, the nodes determine the differences between the initial global model and their current local model and send an update to the aggregation agent. As its name suggests, the aggregation agent is responsible for collecting the updates from all federated nodes and aggregating them, based on a predefined aggregation scheme, in order to develop a new global model. Once developed, the new global model is passed to the federated nodes, continuing the training process. There are several benefits associated with the FL training process. Firstly, since each node collects and processes its own data and sends an update to the aggregation agent, no local data is transferred, improving both privacy and communication efficiency [67]. Additionally, since all model updates are aggregated, all nodes can benefit from insights from other federated nodes without possessing or accessing their data. This can improve data availability for nodes that collect fewer data as well as create a robust system capable of learning various concepts observed in different regions of the system. Finally, FL is a highly scalable and resilient system that can easily tolerate node or aggregation agent outages improving the overall system reliability.

2.4.3.2 Reinforcement Learning

Reinforcement Learning (RL) is a method of experiential learning that uses agent-environment interactions to develop an optimal policy [68]. Each time the agent selects a specific action, it is rewarded based on the results of that action – the better the action, the better the reward. Through numerous training iterations, the agent attempts to maximize the reward obtained in an effort to learn the optimal policy. Additionally, RL can be combined with neural networks to develop Deep RL methods (DRL) capable of addressing large state and action spaces in increasingly complex systems [69]. RL and DRL have several key benefits making them critical to 5G networks and systems. Firstly, the ability to learn optimal policy through experience is advantageous in high complexity systems where access to data and a full system representation are not possible. Additionally, since an RL agent is constantly rewarded for interactions with its environment, it inherently possesses the ability to adapt its policy to address drifts in the domain. Furthermore, given the appropriate simulation environment, RL agents can be trained on anticipated future conditions and domains to prevent adverse conditions such as performance degradation from materializing proactively.

2.5 Conclusion

As demonstrated throughout this chapter, 5G networks will have a profound impact on our daily lives. Through the architectural evolution and transition from 4G to 5G, the emerging technologies will enhance next-generation network performance and provide a plethora of novel use cases. As with any technological revolution, this paradigm-shifting network has its own associated challenges, which must be addressed to ensure a feasible and seamless implementation. One of the proposed solutions to these challenges is increasing levels of network automation through the use of intelligence techniques with the vision of having fully autonomous zero-touch network service management. This automation will also enable advances in use cases such as healthcare, which benefit from the ability to attain sub-millisecond latency, high availability, and improved reliability. Intelligence engines, such as the NWDAF, are rapidly progressing future networks to achieve fully automated network management with the focus on network performance and improving user experience.

Chapter 3

Towards Supporting Intelligence in 5G and Beyond Networks: NWDAF Implementation and Initial Analysis

3.1 Introduction

The telecommunications industry has sparked a dramatic transition to novel and improved high-speed wireless communications architectures in industry and society. The design and operation of 5G and Beyond (5G+) networks is a tightly woven cooperation of developments in both the 5G Core and 5G radio networks that has led the charge for fast-paced development in the communications industry. The 5G+ concept has become a critical tool in the introduction and development of Industry 4.0, a paradigm shift of modern wireless communications systems to true, digital economies [17].

The 5G architecture is comprised of the 5G Core (5GC) network, the new Radio Access Network (RAN), and its newly supported New Radio (NR). The Third Generation Partnership Project (3GPP) outlined the design of the 5G Core to implicitly and explicitly support new architectural features, such as a service-based architecture (SBA), consistent user experience, improved Quality-of-Service (QoS), enhanced machine-to-machine communication services, adaption to cloud-native technologies, and edge computing access. 5G defines three service grades, where each strata defines its own special requirements to adhere to customers' business models: Ultra-Reliable Low Latency Communications (URLLC), Massive Machine-Type Communications (mMTC), and Enhanced Mobile Broadband (eMBB) [70][71].

The use of AI in 5G+ networks is one of the defining characteristics of this paradigm-shifting technology. According to reported statistics, by 2025, it is projected that the telecom industry will invest USD 36.7B in AI through software and hardware investments as well as AI services [59]. The operational benefits of AI in 5G+ networks consider the added value the AI system provides in terms of the management and orchestration of networks [72][73]. One of the main benefits of AI is the ability to take proactive and predictive measures to ensure the optimization of network performance. Some methods of network performance optimization include the reduction of power consumption through enhanced algorithmic performance, the maximization of throughput through optimal traffic routing and infrastructure placement, as well as the ability to support an increasingly dense number of users [60][74].

The envisioned AI-enabled network will consist of intelligent agents being fed data related to the network, including network measurements and statistics, resource utilization, and traffic patterns and conducting inferencing to provide network automation through MANO tasks such as resource optimization and system reconfiguration [59]. However, one of the main challenges plaguing AI implementations across all fields relates to the availability of high-quality data. In order to effectively collect the required data to build and train AI agents, data collection interfaces need to be deployed throughout the network and constantly be monitored [62]. To this end, the Network Data Analytics Function (NWDAF) has been proposed by 3GPP as a solution to this problem to be directly implemented in the 5G+ core network as a key network function.

The work described in this chapter addresses the practical development of the NWDAF and considers its integration into an operational 5G core implemented using open-source software, including Open5GS [75] and UERANSIM [76].

The structure of this chapter is as follows: Section 3.2 considers relevant background information on the 5G Core and the NWDAF for moving towards supporting network intelligence. Section 3.3 presents a use case highlighting key insights obtained from the analysis of NWDAF-collected 5G Core network data and its application to the MANO of 5G+ networks. Section 3.4 discusses the vision and requirements of 6G networks, as well as the expected limitations of 5G networks motivating their initial conceptualization and development. Finally, Section 3.5 concludes the chapter and discusses opportunities for future work.

3.2 5G Core and the NWDAF

3.2.1 5G Core

The 5G Core is composed of various Network Functions (NFs) with their individually associated microservices and responsibilities. The 3GPP intended for the 5G Core to bring about a mindset shift from evolving architectures into standalone, access-independent structures. For example, the 5G Core, by design principle, does not provide backwards compatibility for any previous generations of RANs (*e.g.,* GSM, LTE). Instead, the 5GC consists of a new set of interfaces that are intended for core network-radio network interactions. In terms of the 5G RAN specifications, the 3GPP defined two architectural variants which combine the LTE and the 5G NR: the non stand-alone architecture (NSA) and the stand-alone architecture (SA). The key difference is that the NSA aims to maximize the reusability of 4G architectures by relying on LTE radio access for signaling between UE devices and the network. Specifically, it consolidates an enhanced EPC network to support 5G in the more recent deployments [17].

At the core of 5GC, NFs provide the functionality for establishing sessions and forwarding data to and from mobile User Equipment (UE) devices. Some key NFs and their operations are detailed to provide a brief summary of the 5G Core functionalities. The Access and Mobility Management Function (AMF) interacts with the UE devices and the RAN, and is involved in most 5G signalling calls. As well, the AMF supports activation for devices in idle mode. The Session Management Function (SMF) manages end user device sessions, including their establishment, modification, release, and IP address allocation. The SMF also interacts with other NFs to select and control different User Plane Function (UPF) instances over the network. This control allows it to configure traffic steering and enforcement in UPFs for individual sessions. The UPF processes and forwards user data and is controlled by the SMF. In addition, the UPF connects to external IP networks to act as anchor points, hiding mobility. The Unified Data Management Function (UDM) accesses user subscription data stored in the Unified Data Repository (UDR), a database containing network/user policies and associated data. Finally, the Authentication Server Function provides authentication services for a specific device, utilizing credentials from the UDM [12].

As an underlying function solely responsible for data analytics and network learning, the NWDAF represents operator-managed network analytics as a logical function [70]. The NWDAF provides slice-specific network data analytics to any given NF. As well, the NWDAF provides network analytics information to NFs on a network slice instance level. The function also notifies NFs with slice-specific network status analytic information for any that are subscribed to it. NFs may also collect network status analytic information directly from the NWDAF. In the 5G Core, both the Policy Control Function (PCF) and the Network Slice Selection Function (NSSF) are consumers of network analytics. The PCF may use that data in its policy decisions, and the NSSF may use the load-level information provided by the NWDAF for slice selection.

3.2.2 NWDAF

The NWDAF architecture is designed to aid policy and decision-making for NFs in the control plane and supports some important services for a given NF service consumer. Industrial NWDAF solutions typically have an N23 interface and an N34 interface as reference points to the PCF and the NSSF, respectively. As well, 5G edge computing use cases allow the NWDAF to aid the SMF in routing decisions. As the central point of network analytics, the NWDAF enables operators to capture non-SBI data in addition to SBI data as standalone 5G deployments become more prevalent [24].

As of June 2022, the NWDAF provides five different NF services: *AnalyticsSubscription*, *AnalyticsInfo*, *DataManagement*, *MLModelProvision*, and *MLModelInfo*. The *AnalyticsSubscription* service notifies the NF consumer instance of all analytics subscribed to the specific NWDAF service. The *AnalyticsInfo* service enables the NF consumer to request and retrieve network data analytics from the NWDAF. As well, it enables the NWDAF to request analytics context transfers from another NWDAF if necessary. The *DataManagement* service allows an NF consumer to subscribe to receive data or historical analytics (interpreted as data); if the data is already defined in the NWDAF, the subscription is updated. The *MLModelProvision* service enables an NF consumer to receive notifications when an ML model, matching subscription parameters, becomes available. Finally, the *MLModelInfo* service enables an NF consumer to request and retrieve ML model information from the NWDAF [77].

Industrial NWDAF implementations provide closed-loop automation for third-party NFs and solutions inside the 5G Core. In particular, these NWDAFs are intended for continuous monitoring of every NF, network slice, and UE device and use a variety of KPIs to measure network performance. The real-time KPIs can be used to automate network issue resolution, while predictive analytics can be used to predict those network issues in the future. Predictive analytics may also provide anomaly detection to be used for automating mitigation [24].

3.3 Case Study: NWDAF Implementation and Analysis

The following case study will explore the various insights and conclusions drawn from network-generated data from a 5G Core Network. The analysis conducted in this case study is an example of how the NWDAF can leverage data to provide meaningful insights to enhance the MANO of core network functions. Specifically, this case study will analyze control packets generated during the instantiation of the network core. Through these control packets, various statistics such as the size and number of packets per protocol will be displayed. Additionally, an in-depth exploration of the Binding Support Function (BSF) and its interaction with the Network Repository Function (NRF) will be discussed. Using both analyses, a recommendation can be made regarding the placement of the BSF in relation to the NRF.

The collected data for this case study was generated through Open5GS, an open-source project providing network functionalities for building private 5G networks [75]. The 5G standalone implementation was used for the system model leveraging both the Service-Based Architecture (SBA) and following the Control and User Plane Separation (CUPS) scheme, as described by 5G network standardization efforts led by the 3GPP [77]. UERANSIM, an open-source state-of-the-art 5G UE and RAN implementation, was used to complete full operation of the 5G Core with connected devices [76]. Figure 3.1 outlines the various core network functions which were operational during the data collection phase. Additionally, this figure illustrates how the proposed NWDAF fits in the 5G Core with its associated interfaces (depicted in green). In this figure, the reference point architecture, presented by solid lines, illustrates the point-to-point interaction between core network functions, whereas the SBA is illustrated by the dashed lines. Through the SBA, the NWDAF is able to collect data and statistics about all other authorized core network functions without having an explicit point-to-point reference defined. The data collection phase ran for 138 minutes and, as previously mentioned, exclusively considered the control signalling between the various network functions in the control plane, not including any GPRS Tunneling Protocol (GTP) traffic from the UE and the RAN. The data for this case study is publically available [78].

The first result presented in Fig. 3.2 considers the total number of packets associated with each observed protocol throughout the duration of the data collection phase. As seen through this figure, the overwhelming majority of packets utilize the TCP protocol, something which is expected considering the NFs communicate with each other through REST APIs leveraging the HTTP/2 protocol. RESTful SBA procedures can be categorized into Service Registration, Service Discovery, and Session Establishment. It should be noted that the three NGAP pro-

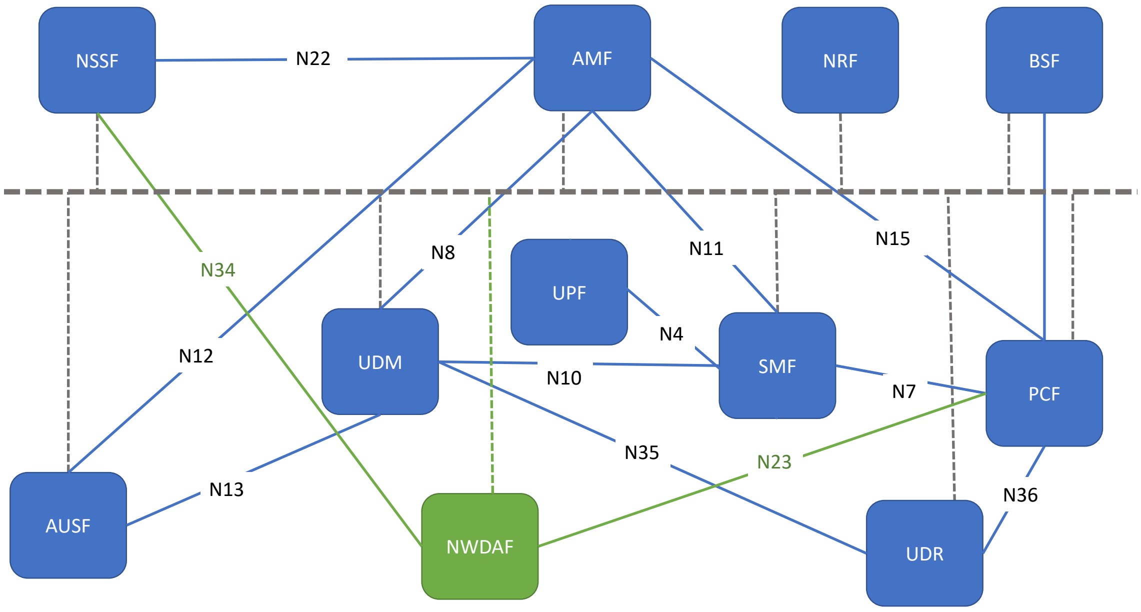

Figure 3.1: 5G Core Service Based Architecture Representation

tocols have been introduced in 5G and are used in communications between the gNB and the Access and Mobility Function.

In addition to the NGAP protocols, which are prevalent in UE registration and de-registration, the Packet Forwarding Control Protocol (PFCP) is paramount to formalizing the interactions between 5G Core NFs, specifically between the SMF and the UPF through the N4 interface. Albeit infrequent in the generated network traffic when compared to other protocols, PFCP is used in signalling procedures in the Control Plane for network attachment and in the User Plane for IPv4/IPv6 packet forwarding with the wireless RAN and the PDU [79].

The next stage in this analysis considers the average size of each protocol's packets along with statistics such as the standard deviation and maximum packet size as presented in Fig. 3.3. Through this figure, it can be seen that the largest packet sizes are attributed to the SSL protocols. However, given the volume of SSL packets presented in the previous results, these packets are infrequent. Considering both presented results, it is evident that the focus of this analysis should be on TCP packet signalling as they have the greatest volume and significant size compared to the other protocols.

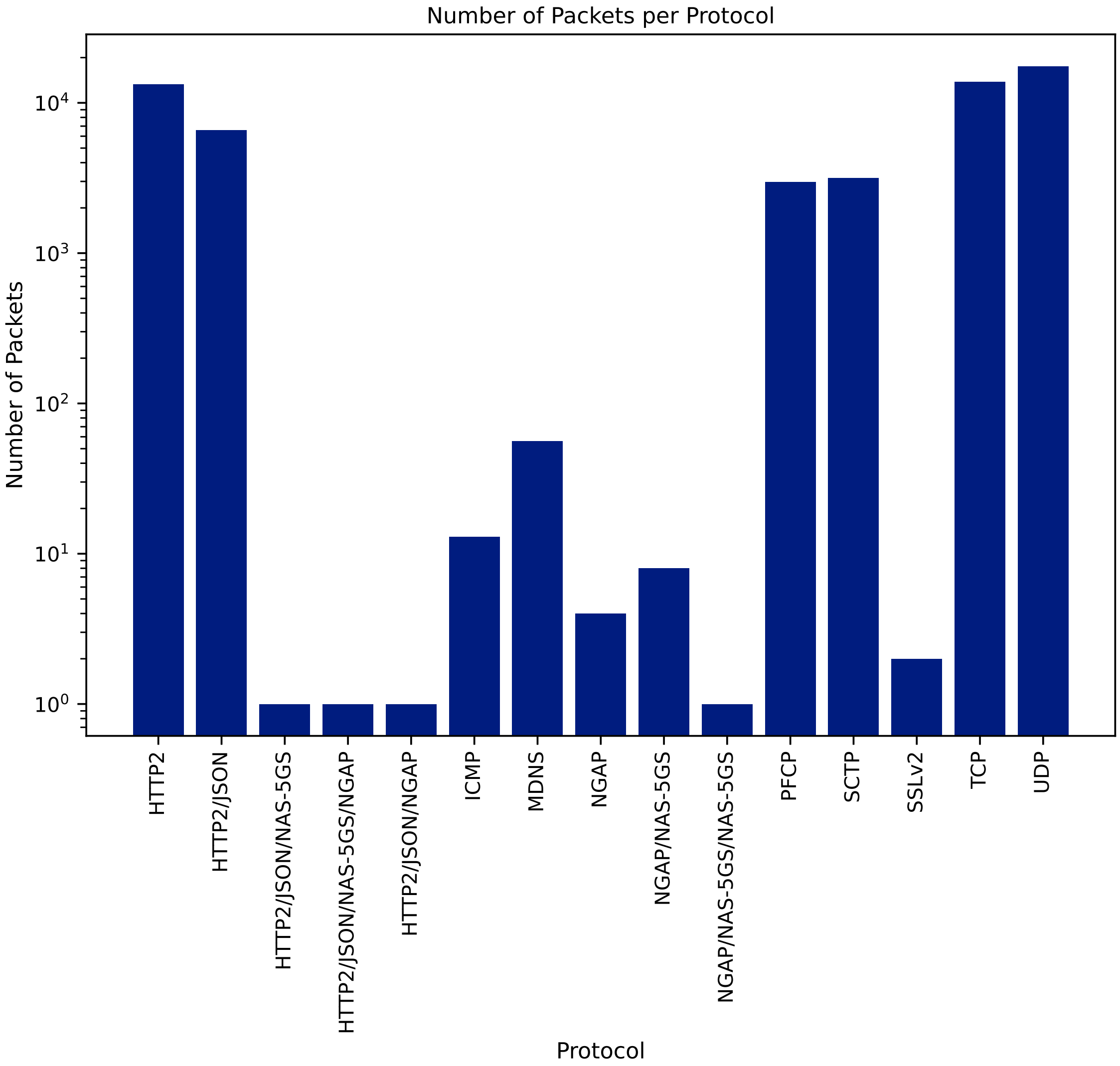

Figure 3.2: Number of Packets Collected in Core NF-NF Interaction per Protocol

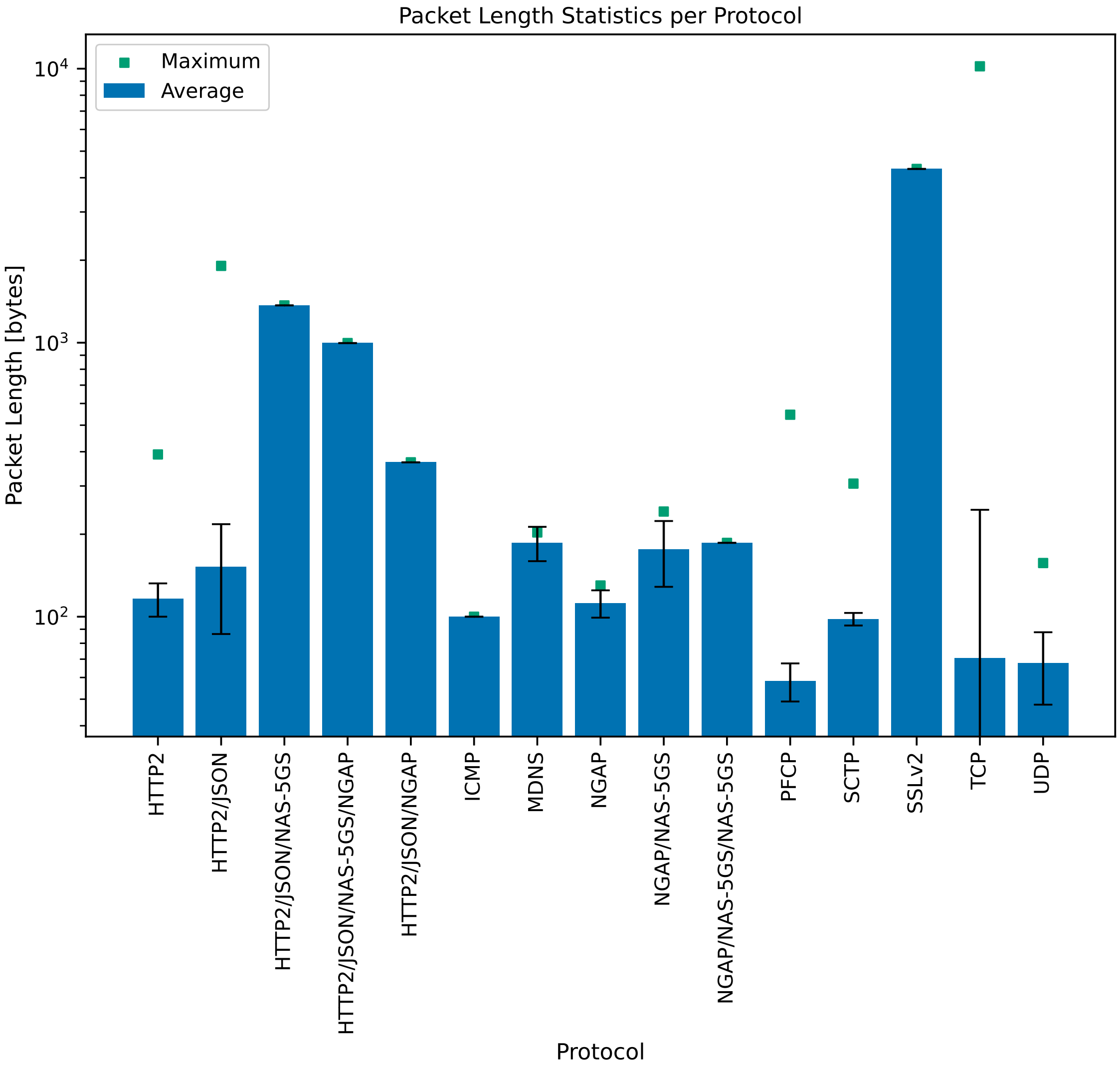

Figure 3.3: Packet Length Maximum, Average, and Error Statistics per Protocol

The following results pertain to the interaction between the BSF and the NRF. This interaction was selected to further explore the trends in TCP control packets and provide a meaningful recommendation based on the volume and frequency of data exchanged between these NFs. Figure 3.4 considers all TCP control packets exchanged between these two functions and compares the size of the packet to the time at which it was sent, effectively providing bi-direction link throughput for this interface. This figure shows a clear spike in packet size near the beginning, followed by a constant packet size for the remainder of the data collection stage. The zoomed-in portion of the graph shows that the packets are transmitted periodically with minor variations due to signalling processes.

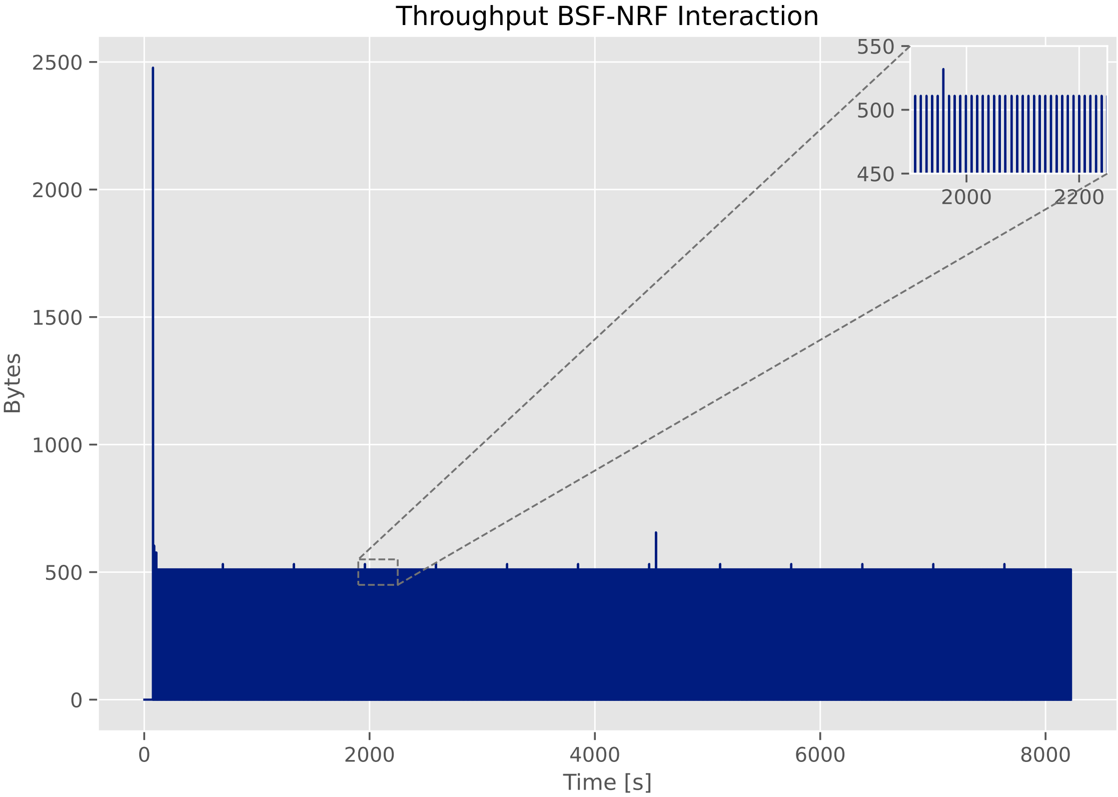

Figure 3.4: Bytes per Second Exchanged between BSF and NRF

To further explore these results and translate them to observable 5G NF events, a more intuitive analysis is done regarding the one-way communication of the BSF with the NRF as seen in Fig. 3.5. In this figure, the TCP packets with the BSF as the source and the NRF as the destination are displayed, and text annotations have been used to highlight key NF events. As seen in this figure, the initial spike in packet size is attributed to the BSF registering with the NRF. In HTTP/2-based communications, this interaction corresponds to a request/response fetch in the SBI, as opposed to a subscribe/notify callback (*e.g.,* the SMF subscribing to the NRF for notifications when other NFs go down). During this registration process, the BSF is required to send all its functional information to the NRF, resulting in the increased packet size. The second major NF event is the BSF heartbeat which occurs every 10 seconds and makes the NRF aware of any changes in its status (*i.e.,* registration, load). As illustrated in the zoomed-in portion of the graph, two packet size values emerge; the greater packet size is associated with the PATCH request used to perform the heartbeat, whereas the lesser packet size is associated with the acknowledgement of received information from the NRF in response. It is important to note that the PATCH request, partially updates the network resource, compared to a PUT request, which completely replaces the resource addressed by the URI with the JSON-formatted payload of the request.

Equivalently, Fig. 3.6 presents the one-way communication of the NRF with the BRF. As labelled through the annotation, the first major spike corresponds to the response sent when the NF has been registered (*NF_REGISTERED*) and the profile has been created. As expected, when compared with the initial request seen in Fig. 3.5, the response is significantly smaller. Furthermore, when looking at the zoomed-in portion of the graph outlining the response to the BSF heartbeat, there are once again two distinct packet sizes that emerge. The smaller of the two sizes corresponds to a simple acknowledgement, whereas the larger size corresponds to the response of the heartbeat. As outlined in the NRF schema, if no significant change has been made to the status of the function, the response to the heartbeat is a packet with an empty body; however, if there were to be a significant change to the function status, such as the signal value *NF_DEREGISTERED*, this response's body would contain the latest updated information.

Given the results presented in this case study, the NWDAF could be tasked with recommending a placement decision for another instance of the BSF function. An industrial NWDAF

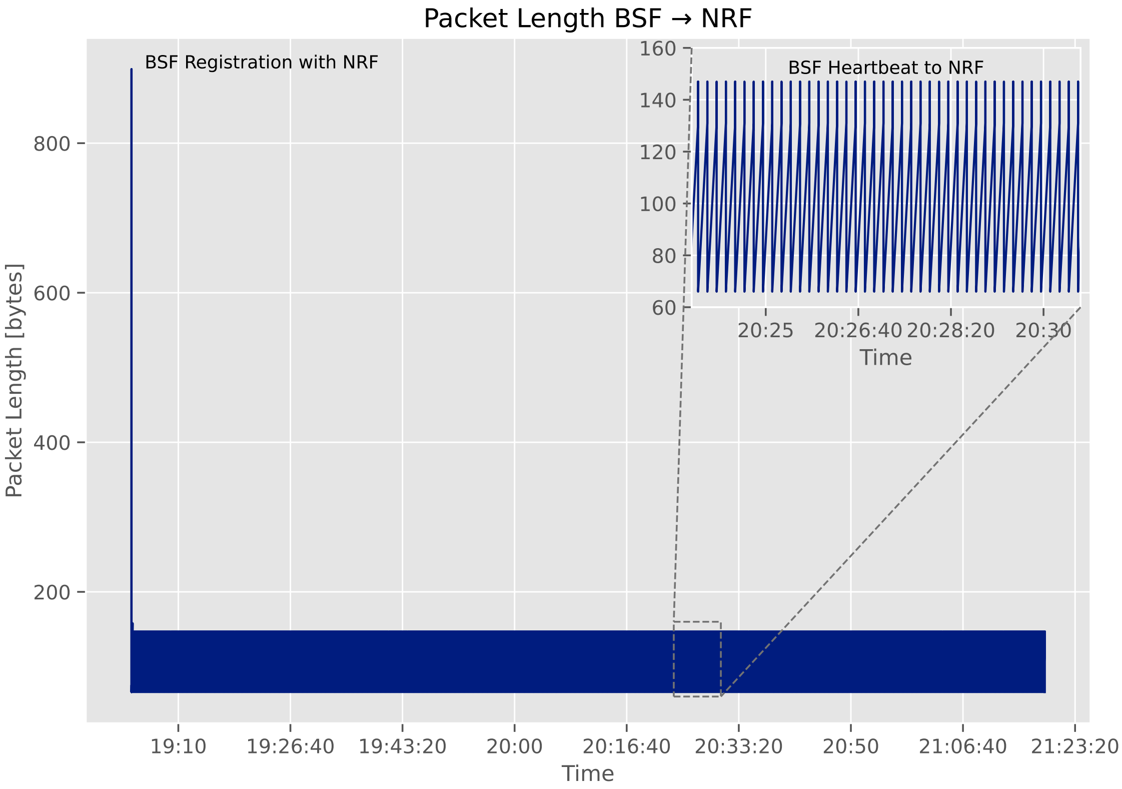

Figure 3.5: Length of Packets Sent from BSF to NRF

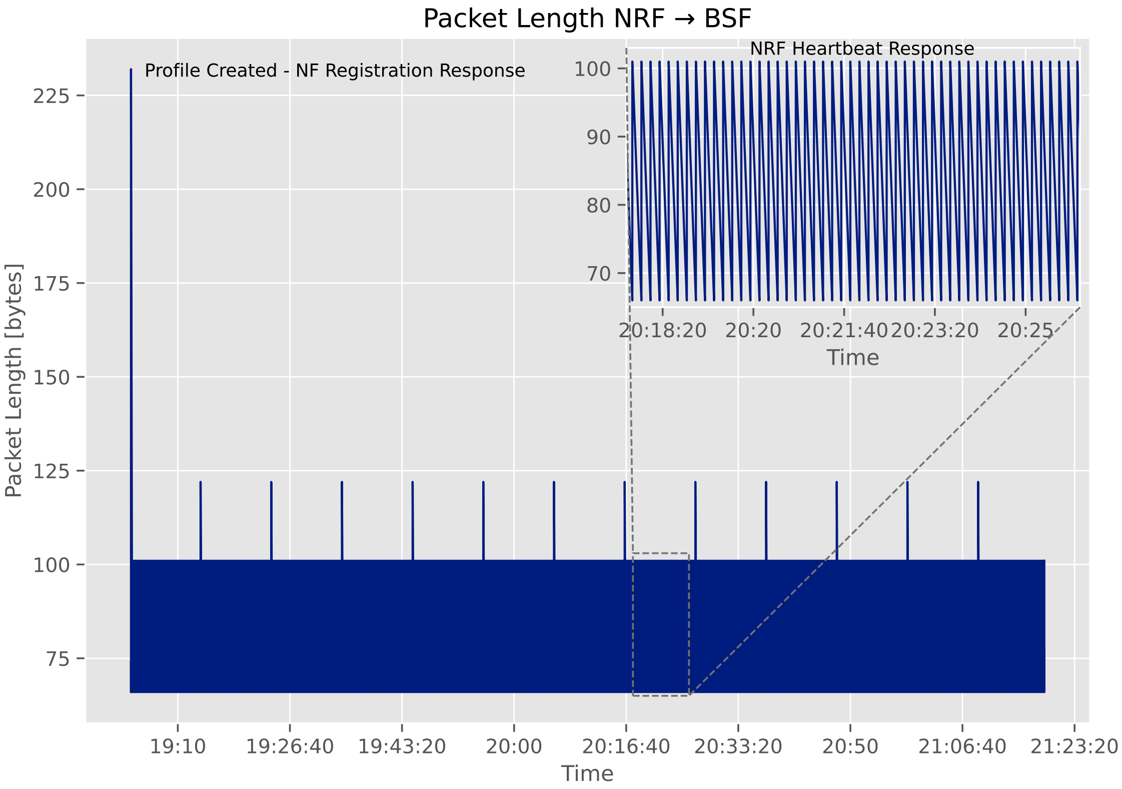

Figure 3.6: Length of Packets Sent from NRF to BSF

solution utilizes similar policy decision-making in the context of the PCF and the NSSF [24]. The PCF should take input from the NWDAF to allocate resources and steer traffic policies for dynamic network slices, and the NSSF should gather load-level information from the NWDAF for the purpose of slice selection. As illustrated, the initial registration process with the NRF results in the packet size spike, whereas the remainder of its interaction with the NRF is a set of periodic heartbeats of much smaller packet size. For this reason, the co-location of these network functions is likely not required as the amount of control information exchanged between them is limited. Future work with the BSF will explore its interaction with the PCF as it is responsible for communicating with the PCF to partially update binding information for PDU sessions which are set to binding level endpoint *NF_SET*.

3.4 Outlook for 6G Networks

As 6G networks take shape, AI will be deeply integrated into the network, more than just through a core network function. As intelligence gets distributed through the system, so do the privacy and security risks associated with it. These risks can range from data poisoning at edge nodes to system-wide model drift, each with its own intricacies and nuances which must be addressed. Additionally, with more data distributed at the edge, data privacy is paramount to ensuring public safety considering critical services such as emergency, finance, and transportation will be in jeopardy of being compromised. As such, it will become increasingly important to consider model maintenance as an integral part of the ML/AI life cycle to ensure future networks' safe and secure performance.

6G networks must fully realize the revolutionizing Industry 4.0 that started with 5G networks. In particular, it is the digital transformation of physical manufacturing systems and IoT services. IoT-based diagnostics will enhance maintenance and operation of machine communications, prioritizing cost-effectiveness and flexibility. In Industry 4.0, automation requires reliable and synchronous communication systems that 6G is situated to address through the aforementioned disruptive technologies [80].

With the advent of low-power requirements for IoT devices, AI model training could consider new specifications recommended by NWDAF data, based on federated learning (*e.g.,* learning at edge devices). As well, AI use cases guided by the NWDAF data analytics must address the lack of bounding performance in 6G networks. In contrast to the previous challenge, system design must consider worst-case scenario network events while providing a minimum acceptable QoS/performance guarantee; however, due to non-linear characteristics of such related problems, it may be infeasible for AI approaches regardless of their effectiveness in real-time inferences [81].

3.5 Conclusion

The adoption and integration of intelligence in 5G networks has the potential to revolutionize our current networking practices. Perhaps the greatest potential lies in the Network Data and Analytics Function (NWDAF) proposed by 3GPP. This function will collect a plethora of information and statistics on the operation of the network ranging from high-level data such as slice level information to very specific data related to a single NF. This chapter presented a case study that outlined an analysis of NWDAF-collected core network function data from an Open5GS and UERANSIM implementation. An initial analysis into this data and the potential insights that can be drawn from it were illustrated. In this case study, 5G Core function control messages were considered; specifically, the interaction between the Network Repository Function (NRF) and the Binding Support Function (BSF) was examined.

Future work in this area will consider the impact of the NWDAF on 5G networks and continue to explore data generated from 5G Core network functions. As mentioned in Section 3.3, a study on the interaction between the BSF and the Policy and Control Function (PCF) will be a focus. Finally, the development of advanced analytics models will be considered using the generated data for use cases such as proactive network management and forecasting.

Chapter 4

A Reliable AMF Scaling and Load Balancing Framework for 5G Core Networks

4.1 Introduction

Mobile data traffic in 4G systems is rapidly increasing and in order for 5G communications systems to accommodate these user servicing demands, increasing system capacity and provisioning network resources are necessary. Scalability techniques and strategies in the 5G Core (5GC) architecture offer more effective opportunities since the architecture splits the 4G Mobility Management Entity (MME) into the AMF, as well as the Unified Data Management (UDM), and the Session Management Function (SMF) [82]. The AMF, in particular, presents an issue with addressing ever-increasing demand for handling User Equipment (UE) sessions, and 5G networks must adopt improving technologies if there are to retain their competitiveness in the mobile communications market. Network Function Virtualization (NFV) is considered a significant contributor to the realization of 5G networks since it decouples 5G Network Functions (NFs) from associated hardware and deploys them as virtualized NFs (VNFs), which are platform-independent. Multi-access edge computing (MEC) is another promising technology in 5G networks for addressing low-latency requirements while alleviating network load [83]. These technologies have to employ effective solutions and strategies in order to realize their requirements that are hinging on 5G network performance. Mixed-integer linear programming (MILP) problems can formulate scaling strategies for the AMF by minimizing variables, such as cost or migration time, while satisfying variable constraints, such as user throughput or data rate requirements. Load-balancing mechanisms, as well, can alleviate the loads of user demand on each scaled AMF instance, depending on the algorithm and strategy employed to reduce latency in the Control Plane (CP) [84]. Therefore, this chapter proposes both an integer programming formulation of the AMF scaling problem to minimize the number of instances and a load-balancing module to address the influx of varying numbers of UE requests for sessions to the AMF.

The structure of this chapter is as follows: Section 4.2 presents related works in the field of network optimization operations research. Section 4.3 outlines the methodology followed, including the problem formulations, a description of the 5G Core system prototype, and details regarding the implementation. Section 4.4 presents and analyzes the obtained results from both problems. Finally, Section 4.5 concludes the chapter and discusses future work.

4.2 Related Work

The research directions of 5G network infrastructure optimization techniques can focus on request models or control signalling as indicators for NFs to be placed, replaced, or instantiated. Pertaining to mobility management requests for mobile users in the network, the AMF is considered the most important NF as it is also considered a bottleneck for performance optimization and deployment in the 5GC [85]. Typically, optimization problems regarding the AMF involve placement decision-making; however, network flow models, such as the previously mentioned request models, can be modelled with integer variables in MILP to consider scaling instances [86]. MILP models can achieve optimal solutions in many scaling scenarios, but as the network size increases, the solution may not easily be managed or directed. Heuristic algorithms have been used to achieve near-optimal solutions in faster times, relative to MILP solutions [87]. Linear programming models have been shown to provide a trade-off between replications and migrations of NFs, since replications are beneficial to Quality-of-Service (QoS) requirements, but require additional resources, and vice versa for migrations [88]. In similar models, additional heuristics frameworks have been proposed to provide another optimizable objective for maximizing throughput [89].

Load balancing mechanisms for NFs are important because the algorithms they employ must be adequate, for any 5G architecture, in establishing an optimal load management scheme [90]. Based on UE mobility, when it is necessary to increase system capacity for mobile data traffic, UE devices can be categorized into different priority schemes when load-balancing, based on operational speeds and mean throughput [91]. UE devices can be further grouped into service-customized network slices in the 5G framework and have also been considered in other VNF placement problems [92]. Regarding dynamic allocation of network slice bandwidth for these UE devices, traffic prediction can be achieved with long short-term memory (LSTM) networks and the problem formulation can be modelled with a fractional knapsack optimization problem [93]. Both scaling and placement (location and number of instances, respectively) can be considered in a single problem, such as the UPF Placement (UPFP) Problem [94]. Session establishment requests from UE devices to the AMF have been used as signalling procedures for finding solutions to the scaling problem and performance evaluation [95].

4.3 Methodology

The following section will outline the methodology followed in this work, including the optimization problem formulation, the load balancing module, the 5GC prototype, and the implementation parameters.

4.3.1 Optimization Problem Formulation

The variables used in the optimization problem formulation are defined as follows. α denotes the number of AMFs required. C denotes the set of classes. u denotes the number of UEs per class. w_c denotes the weight assigned to each class. r_{AMF}^{β} denotes the β resource requirement of the AMF. C_s^{β} denotes the β resource capacity of the hosting server. N_u denotes the maximum allowable number of UEs per AMF. N_{gnb} denotes the maximum number of UEs per gNB. The optimization problem is formulated as follows:

Objective:

$$minimize(\alpha) \tag{4.1}$$

Constraints:

$$\sum_{c \in C} w_c \times \sum_{u \in c} u \leq N_u \times \alpha \tag{4.2}$$

$$w_c \geq 1 \; \forall \; c \in C \tag{4.3}$$

$$r_{AMF}^{RAM} \times \alpha \leq C_s^{RAM} \tag{4.4}$$

$$r_{AMF}^{CPU} \times \alpha \leq C_s^{CPU} \tag{4.5}$$

$$\sum_{c \in C} \sum_{u \in c} u \leq N_{gnb} \tag{4.6}$$

The objective of this optimization model, as expressed through Eq. 4.1 is to minimize the number of AMF instances required to support a projected number of users. The users are split into classes, which directly translate to QoS priority and slice selection policies (*i.e.,* class 3 represents high QoS requirements, which can be attributed to intensive AR/VR application, whereas class 1 represents low QoS requirements, with applications that do not require stringent performance guarantees). UE classes have been grouped in broader categories, such as Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communications (MMC), and Ultra-Reliable Low-Latency Communications (URLLC) [93]; however, there is a greater importance to specify user demands and impacts on QoS requirements, justifying the user stratification mentioned above. In an effort to improve system reliability and resilience, the class weights w_c are introduced to act as an over-provisioning mechanism protecting against rapidly increasing UE device densities by scaling the projected number of users per class. Equation 4.2 incorporates the weight term and determines the weight-adjusted number of users projected; in order to meet this demand, the number of AMF instances α multiplied by the UE capacity per AMF N_u must exceed the weighted user projection. Equation 4.3 ensures that the weight of any given class cannot be less than one. This constraint is critical as class weight values below one would result in AMF under-provisioning. The resource constraints presented in Eq. 4.4 and 4.5 ensure that the host server has enough RAM and CPU resource capacity to support the required number of AMFs. Finally, Eq. 4.6 ensures that the number of users does not exceed the maximum tolerable number per gNB.

4.3.2 Load Balancing Module

The proposed load balancing module is used to distribute incoming user requests to an AMF set (group of associated AMFs) based on their relative capacity as outlined by 3GPP, and ETSI [96]. For each batch of incoming requests, the current capacity of each AMF is retrieved. This value is part of the AMF schema as outlined in 3GPP and indicates the relative processing capacity of an AMF compared to other AMFs in the AMF set [97]. For the purposes of this work, this parameter is scaled to a value on the range [0,1] and is calculated through Eq. 4.7, where $AMF^{x}_{capacity}$ denotes the capacity of AMF x and the AMF set is denoted by AMF. These relative capacities are then used as probabilities (since their summation is equal to one) for an incoming request being assigned to an AMF. For example, if AMF y has a greater capacity than AMF z, then it is more probable that a UE will connect to AMF y.

$$\frac{AMF^{x}_{capacity}}{\sum\limits_{x \in AMF} AMF^{x}_{capacity}} \ \forall\ x \in AMF \tag{4.7}$$

4.3.3 Prototype

Emulated environments provide real-like scenario data that the optimization and load-balancing problems can apply to. Hence, this section considers a working emulation of the 5GC standalone architecture, which consists of a single server that runs a virtual management service, hosting individual Linux virtual machines (VMs) for each NF in the 5GC or entity that interacts with the 5GC. The 5GC prototype utilizes Open5GS, which is a C-language open-source implementation of 5GC. Open5GS is also Release-16 compliant, or in accordance with the 3GPP release specifications [75]. The 5GC NFs provided by Open5GS include the following: the AMF, in addition to the Network Repository Function (NRF), Session Management Function (SMF), Authentication Server Function (AUSF), Unified Data Management (UDM), Unified Data Repository (UDR), Policy Control Function (PCF), Network Slice Selection Function (NSSF), Binding Support Function (BSF), and User Plane Function (UPF). In order to setup a RAN and multiple UE devices to interact with the 5GC, UERANSIM is used, which is an open-source 5G SA UE and RAN C++ implementation [76]. A single gNB connects to the fully operational 5GC and a single host uses multiple network interfaces to emulate different UE devices interacting with the RAN and the User Plane Function. All NFs are executed as Linux executable programs in each VM.

The system prototype contains our implementation of the NWDAF which uses network monitoring and data collection techniques to synthesize all service operations of the NWDAF, according to its 3GPP specification. As a Type 1 Hypervisor for the server's Virtual Machine Management Service, Hyper-V (used for the VMs) allows port mirroring to designate VMs as sources and destinations in terms of network traffic, where source VMs will duplicate all network traffic contained within the private network and forward their copies to a single host as the destination, acting as the central point for NWDAF analytics and operations. Apache Kafka is used to monitor and pipeline the captured network data and stream it to a MongoDB instance to aggregate historical data, as well as provide support for current state network monitoring for future policy decision-making in the 5GC. The Open5GS implementation in the private network does not use Network Address Translation (NAT), and port forwarding rules are configured such that UE device traffic can be encapsulated in the GPRS Tunneling Protocol (GTP) when communicating with the UPF. Thereafter, the UPF routes Internet connectivity to UE devices for any applications that the UE devices run. As multiple UE devices run sample applications which connect through the 5GC, newly-generated packets are processed and then transformed into schema-validated NWDAF events, which can constitute a readily available dataset for any algorithm or optimization problem, as well as in the context of load-balancing. Closed-loop automation, therefore, is a capability, from the prototype in its entirety, to maximize the potential of the NWDAF and its impact on maintenance and operation-specific decisions in the 5GC.

4.3.4 Implementation Details

The prototype developed above was used to determine the number of UEs supported by an AMF and a gNB, the resource requirements of an AMF instance, as well as the hosting server capacity. For the experiments conducted in this chapter, the maximum number of UEs supported by an AMF is 1024, the maximum number of UEs supported by a gNB is 2e6, the resource requirements of an AMF instance are 6.8 GB RAM and 1 vCPU, and the server resource capacities are 64 GB RAM and 12 vCPUs. The results section will outline various experiments conducted by varying the distribution of classes to which the UEs belong, the class weight parameters, and the number of UEs. Table 4.1 outlines the various parameters considered. Three classes were assumed in this work where class 3 exhibits the greatest QoS requirement and class 1 exhibits the lowest. It should be noted that the parameters with an alias value in the table correspond to results presented in the subsequent section.

Table 4.1: Implementation Parameters

Variable	Value	Alias
Class Distribution	*[c1, c2, c3]*	
	[0.33, 0.33, 0.33]	4
	[0.6, 0.2, 0,2]	5
	[0.8, 0.1, 0.1]	6
	[0.2, 0.2, 0.6]	2
	[0.1, 0.1, 0.8]	0
	[0.2, 0.6, 0.2]	3
	[0.1, 0.8, 0.1]	1
Class Weights	*[w1, w2, w3]*	
	[1, 1, 1]	0
	[1, 1.1, 1.2]	1
	[1, 1.2, 1.4]	2
	[1, 1.3, 1.5]	3
Number of UEs	[1, 2, 3, 4, 5, 6, 7] x 1000	

4.4 Results

The results presented in this section are separated into two sections, optimization model testing and load balancing module testing. For the optimization model testing, the model was run using all permutations of the parameters listed in Table 4.1. The load balancing module testing uses the results of the optimization model to determine the minimum number of AMFs required. A Poisson distribution of varying λ values is used to simulate the inter-arrival time of the requests. The λ values considered are from the set [1, 3, 5], where a λ value of 1 results in a lower interarrival time than that of the distribution where λ is 5. The various lambda values are used to simulate different types of events that could increase the number of UEs at different rates (*i.e.,* crowd events). The load balancing module runs using these request interarrival times; it selects an AMF for the incoming request and updates the AMFs relative capacity accordingly. Two different permutations of this experiment are conducted, the first where all AMFs begin at full capacity and the second where initial AMF capacities are randomly generated on the interval [0.68, 1].

4.4.1 Optimization Problem

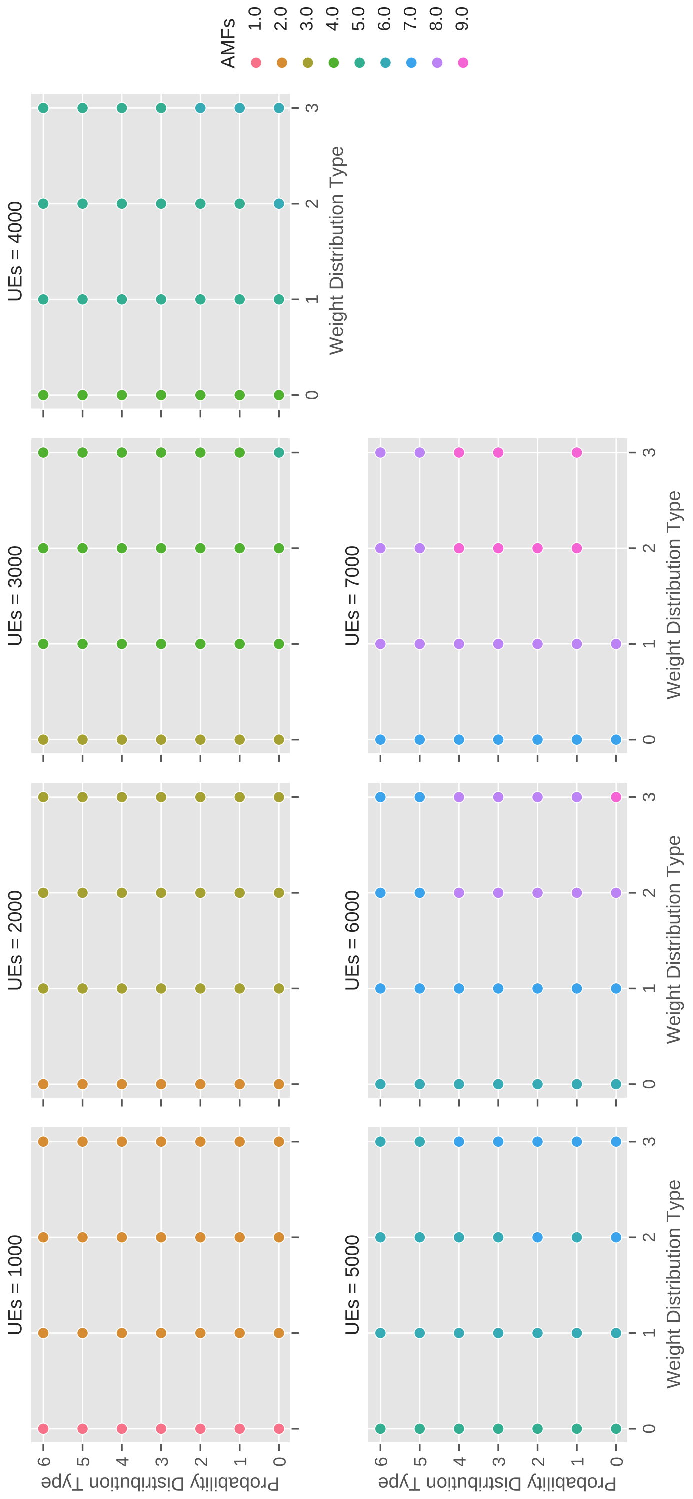

Figure 4.1: Number of AMF Instances Required per Probability and Weight Distribution Parameters

The results of the optimization problem formulation testing are presented in Fig. 4.1. Each graph in this figure represents a different number of UEs requiring AMF allocation. The x-axis of the graphs is the weight distribution used, and the y-axis is the class probability distribution used. It should be noted that for brevity, the values displayed along the axes are aliases for the weights and distributions as outlined in Table 4.1. As seen in this figure, as the number of UEs increases, so does the number of AMFs required to support those UEs; however, the weight and class distribution considered also impact the minimum number of AMFs required.

The case where the number of UEs is 7000 is a prime example of how the weight and class distribution can affect the optimal solution. As seen in this figure, there are three depicted optimal values, 7, 8, and 9 AMFs and two cases where the solution was infeasible (class distribution 0 with weight distributions 2 and 3). Regarding the weights, as seen in Table 4.1, weight distribution with alias 0 considers an equal weighting between all classes and the subsequent weights become increasingly biased towards the second and third classes associated with higher QoS requirements. As such, it is evident that as the weight distribution shifts from 0 to 3, the system's resiliency increases through the class weights controlling the conservativeness of the solution. These weights allow for increased flexibility for the network service providers as they can efficiently manage the level of over-provisioning for resilience in their system.

Intuitively, the class distribution affects the solution when considering class weighting. For the case of 7000 UEs, when the weight distribution is type two, there are three obtained solutions based on the class distribution types considered, 8, 9, and infeasible solution. Consulting Table 4.1 it is seen that distributions 5 and 6 result in the lowest number of UEs belonging to classes 2 and 3; as such, the impact of increased weight on these classes is diminished, resulting in the smallest number of AMFs. Conversely, class distribution 0 results in the greatest number of UEs belonging to class 3; the compounding impact of the number of users in class three and the increased weight of class three requires over-provisioning levels that cannot be accommodated given the current server capacities and therefore, results in an infeasible solution. The remaining distributions fall somewhere between the previously discussed boundary cases, resulting in 9 AMFs required.

The results presented in this section illustrate the optimal solution bound when exploring an increasing number of UEs, variable class distributions, and increasing class weights. This analysis has illustrated that by manipulating the class weights, service providers can increase the resilience of their system to protect against unforeseen events causing an influx of users and adverse network conditions.

4.4.2 Load Balancing Module

Figure 4.2 displays the load-balancing experiments with three different λ values for the Poisson distributions of the UE request arrival times and the capacities of all AMF instances are equal (the maximum supported capacity at 1.0, based on its configuration scale in Open5GS) [75]. Each plot shows the changing AMF capacities over time as they converge towards full capacity and underneath the plot, a bar graph demonstrates the number of times each AMF was selected for servicing the incoming UE requests. Comparing the three capacity plots, the different λ values of the Poisson distributions show longer times for the AMFs to converge to full capacity when λ is larger. The corresponding bar graphs show an almost equal distribution of AMF selections, which is expected for equal AMF capacities.

Figure 4.3 shows the same load-balancing experiments as before, with varying Poisson distributions, but the AMF instances' capacities are configured at random. Both the plots, at time $t = 0$, and the bar graphs show the starting capacities of each AMF. For example, the left-most plot demonstrates that *AMF 8* has the largest capacity, while *AMF 0* has the smallest capacity: the same information is conveyed based on the total number of times, that the corresponding AMF is selected, in the bar graph. Similar to Figure 4.2, this figure shows how the AMF instances converge to full capacities slower, with larger λ values. As well, the AMF capacity curves converge at similar rates, with equidistant descents beginning approximately at $t = 10^2$ seconds. This point is important for evaluating relative AMF capacities in the Load-Balancing service mechanism for any AMF; if all AMF capacities are analyzed at a given

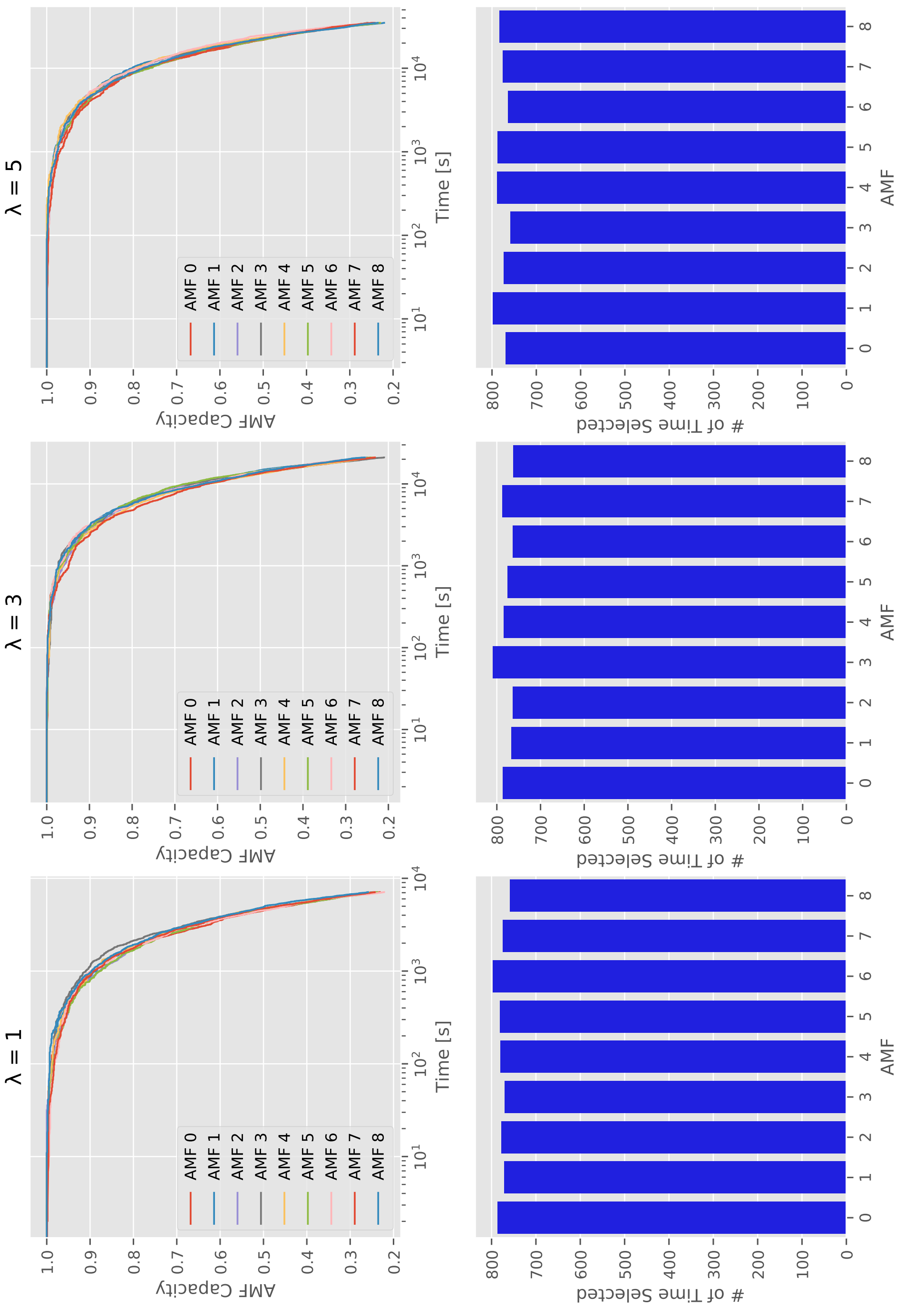

Figure 4.2: Number of Times Selected per AMF Instance and AMF Capacity (Equal Initial Capacity) over Time for Varying UE Request Distributions

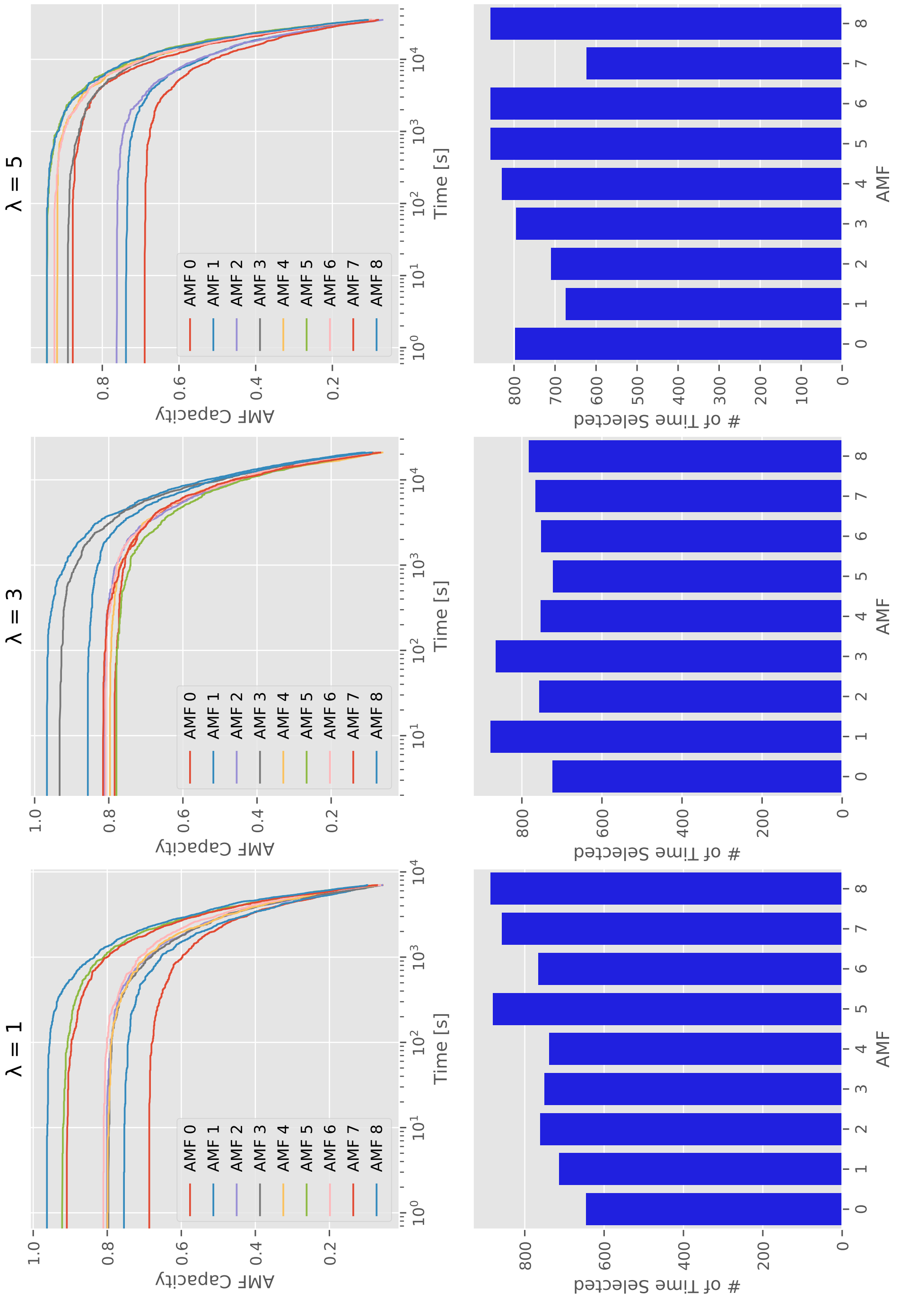

Figure 4.3: Number of Times Selected per AMF Instance and AMF Capacity (Random Initial Capacity) over Time for Varying UE Request Distributions

point in time for each plot, the initial capacities ranking, of all AMFs' capacities from largest to smallest, remains consistent as requests are completed from start to finish. AMF instances use a weight factor with the Load-Balancing service to denote the relative capacity of its own instance with respect to other instances, and this insight is important to network maintenance because weight factors are not changed frequently (typically on a monthly basis) [97].

4.5 Conclusion

The work presented in this chapter proposes a reliable AMF scaling and load balancing framework for 5G core networks. The minimum number of AMFs required to support a forecasted number of UEs is determined through an optimization problem formulation. The solution to the optimization problem is then leveraged to instantiate the appropriate number of AMF instances and perform load balancing through relative AMF capacity. The implementation phase of this work includes the development of a 5G core prototype, which is used to obtain capacity and utilization values for various elements of the core network used in the optimization problem formulation. Future work in this field will consider a traffic forecasting module such that an end-to-end proactive service pipeline can be used to provision AMF instances in advance of anticipated large-scale events and ensure a resilient and reliable network with performance and service continuity.

Chapter 5

Conclusion

5G and Beyond networks will have a profound impact on daily life in our society, as the insightful analysis on mobile networks, along with the evolution of advancing technologies and emerging use cases, has shown. The transition from 4G to 5G and to 6G and beyond has spurred a rapid architectural evolution of modern networks, and new enabling technologies are enhancing and improving the performance of next-generation networks. With any technological revolution, there are also a plethora of challenges; intelligent networking techniques have been shown to address these challenges with the aim of ensuring a feasible and seamless implementation of these modern, intelligent networks. As well, increasing levels of intelligent automation in these networks pave the way for the transition to fully autonomous zero-touch network service management, which is essential for use cases such as healthcare, which demand sub-millisecond latencies, utmost reliability, and high availability.

The NWDAF proposed by the 3GPP has the potential to revolutionize the adoption and integration of intelligence in 5G and Beyond networks. This NF has been shown to collect a large variety and volume of network data and statistics involving operations and maintenance, including high-level data related to network slices and NF service operation data within the 5G Core. It is both useful to network providers for monitoring network performance and to the network itself, acting upon NWDAF data by predictive maintenance operations and realizing complete network automation. NWDAFs can be centralized and/or distributed across different tracking areas and regions of operation to, furthermore, expose the underlying network data. A practical and operational implementation of the 5G Core and the NWDAF were presented and shown to provide key insights on core network function data, in order to draw conclusions on future decision-making within the network. In addition, advanced analytics models were constructed using the generated data to illustrate both the capabilities and the importance of proactive network management and forecasting.

Operations research has been conducted with practical, generated network data to find optimal or near-optimal solutions to decision-making problems in the 5G Core. In particular, a scaling framework and load-balancing mechanism were proposed for the AMF and its number of instances within the 5G Core. The optimization problem determines the minimum number of AMFs required to accommodate and support a forecasted number of UEs within the network. The problem solution has been demonstrated to aid in instantiation of AMF instances and perform load balancing with accordance to relative AMF capacities. The insights from this approach to the optimization problem can be extended to traffic forecasting and user prediction for network service operations.

Future works, considering the challenges that have been addressed in each chapter, will focus on the limitations of the current NWDAF prototype, as well as refining the optimization problems to better suit industrial needs and applying machine learning techniques to model-ready training datasets prepared by the NWDAF. Efficient operation is at the forefront of future 5G deployment and maintenance considerations, and mobility is a key metric/KPI that the NWDAF must follow and learn about within the network. Control plane NFs can predict this mobility, provided that the analytics operations are efficient and intelligent (*i.e.,* leads to making proactive decisions). Data required for mobility prediction can be streamed by the AMF to the NWDAF via the data collection API and can identify aperiodicity in mobility patterns. On the user plane side, a UPF area prediction service can be implemented based on UE location, capacity/availability, and distance [98].

While 5G is currently being developed and NSPs are beginning initial deployment, and integration, discussion of 6G networks is already on the horizon. The motivation behind this discussion is the rapidly increasing number of connected devices, a trend that is expected to continue for the foreseeable future as I/IoT frameworks continue to develop and expand. To this end, the outlook for future research in 5G and Beyond networks will have to address the current issues and limitations of modern operational networks as they will not be able to meet the demands of future use cases, which are projected to require transfer rates in the order of Tb/s, latency on the order of microseconds, as well as increased connection density due to a multitude of deployed sensors. The various use cases of the NWDAF will aid in addressing the AI-related challenges in 6G networks.

To address the expected limitations and shortcomings of 5G networks, research has begun into methods of expanding the capacity and capability of future networks to ensure that future demands can be met. Some proposed lines of research include the exploration of new frequency bands (THz) as well as distributed and federated intelligence throughout the network. As user behaviour and habits change and evolve, networking practices must also follow suit. 6G presents a revolutionary opportunity to scale up the presence of intelligence in networks and ultimately enable a plethora of future use cases and applications; however, this is not a trivial task.

With the ever-increasing demands of users and applications alike, and the capabilities of rapidly developing intelligence engines such as the NWDAF, next-generation mobile networks are positioned to satisfy the strict service requirements for future networks.

Bibliography

[1] Nisha Panwar, Shantanu Sharma, and Awadhesh Kumar Singh. A survey on 5g: The next generation of mobile communication. *Physical Communication*, 18:64–84, 2016.

[2] M Rohini, N Selvakumar, G Suganya, and D Shanthi. Survey on machine learning in 5g. *International Journal of Engineering Research and Technology*, 9:569–576, 2020.

[3] Gabriel Brown et al. Ultra-reliable low-latency 5g for industrial automation. *Technol. Rep. Qualcomm*, 2:52065394, 2018.

[4] Erik Dahlman, Gunnar Mildh, Stefan Parkvall, Janne Peisa, Joachim Sachs, Yngve Selén, and Johan Sköld. 5g wireless access: requirements and realization. *IEEE Communications Magazine*, 52(12):42–47, 2014.

[5] Murtaza Ahmed Siddiqi, Heejung Yu, and Jingon Joung. 5g ultra-reliable low-latency communication implementation challenges and operational issues with iot devices. *Electronics*, 8(9):981, 2019.

[6] Xin Li, Mohammed Samaka, H Anthony Chan, Deval Bhamare, Lav Gupta, Chengcheng Guo, and Raj Jain. Network slicing for 5g: Challenges and opportunities. *IEEE Internet Computing*, 21(5):20–27, 2017.

[7] Gábor Soós, Dániel Ficzere, Pál Varga, and Zsolt Szalay. Practical 5g kpi measurement results on a non-standalone architecture. In *Noms 2020-2020 IEEE/IFIP network operations and management symposium*, pages 1–5. IEEE, 2020.

[8] Hisham A Kholidy, Andrew Karam, James L Sidoran, and Mohammad A Rahman. 5g core security in edge networks: A vulnerability assessment approach. In *2021 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2021.

[9] Cheng-Chin Tsai, Fuchun Joseph Lin, and Hiroshige Tanaka. Evaluation of 5g core slicing on user plane function. *Communications and Network*, 13(03):79–92, 2021.

[10] Peter Rost, Albert Banchs, Ignacio Berberana, Markus Breitbach, Mark Doll, Heinz Droste, Christian Mannweiler, Miguel A Puente, Konstantinos Samdanis, and Bessem Sayadi. Mobile network architecture evolution toward 5g. *IEEE Communications Magazine*, 54(5):84–91, 2016.

[11] Nokia. “5g core (5gc)”. https://www.nokia.com/networks/portfolio/5g-core/ (accessed Jan. 9, 2022).

[12] Gabrial Brown. Service-based architecture for 5g core networks. *A Heavy Reading white paper produced for Huawei Technologies Co. Ltd. Online verfügbar unter: https://www. huawei. com/en/press-events/news/2017/11/HeavyReading-WhitePaper-5G-Core-Network, letzter Zugriff am*, 1:2018, 2017.

[13] Sassan Ahmadi. *5G NR: Architecture, technology, implementation, and operation of 3GPP new radio standards*. Academic Press, 2019.

[14] Guangyi Liu, Yuhong Huang, Zhuo Chen, Liang Liu, Qixing Wang, and Na Li. 5g deployment: Standalone vs. non-standalone from the operator perspective. *IEEE Communications Magazine*, 58(11):83–89, 2020.

[15] Yongwan Park. 5g vision and requirements of 5g forum, korea. In *ITU-R WP5D Workshop*, 2014.

[16] Alcardo Alex Barakabitze, Arslan Ahmad, Rashid Mijumbi, and Andrew Hines. 5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges. *Computer Networks*, 167:106984, 2020.

[17] Stefan Rommer, Peter Hedman, Magnus Olsson, Lars Frid, Shabnam Sultana, and Catherine Mulligan. *5G Core Networks: Powering Digitalization*. Academic Press, 2019.

[18] Junseok Kim, Dongmyoung Kim, and Sunghyun Choi. 3gpp sa2 architecture and functions for 5g mobile communication system. *ICT Express*, 3(1):1–8, 2017.

[19] Andy Sutton. 5g network architecture. *J. Inst. Telecommun. Professionals*, 12(1):9–15, 2018.

[20] Endri Goshi, Michael Jarschel, Rastin Pries, Mu He, and Wolfgang Kellerer. Investigating inter-nf dependencies in cloud-native 5g core networks. In *2021 17th International Conference on Network and Service Management (CNSM)*, pages 370–374. IEEE, 2021.

[21] Georg Mayer. Restful apis for the 5g service based architecture. *Journal of ICT Standardization*, pages 101–116, 2018.

[22] Danish Sattar and Ashraf Matrawy. Optimal slice allocation in 5g core networks. *IEEE Networking Letters*, 1(2):48–51, 2019.

[23] 3GPP. Architecture enhancements for 5G System (5GS) to support network data analytics services. Technical Specification (TS) 23.288, 3rd Generation Partnership Project (3GPP), 2022. Version 17.5.0.

[24] Network data analytics function (nwdaf), Oct 2021.

[25] Youbin Jeon, Hyeonjae Jeong, Sangwon Seo, Taeyun Kim, Haneul Ko, and Sangheon Pack. A distributed nwdaf architecture for federated learning in 5g. In *2022 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–2. IEEE, 2022.

[26] Γ*εώργιος* Χ*ρήστος* Τ*ζιάβας*. Implementation of the network data analytics function (nwdaf) of the 5g core architecture according to 3gpp standard. 2022.

[27] Ping Zhang, Xiaoli Yang, Jianqiao Chen, and Yuzhen Huang. A survey of testing for 5g: Solutions, opportunities, and challenges. *China Communications*, 16(1):69–85, 2019.

[28] Shunliang Zhang, Yongming Wang, and Weihua Zhou. Towards secure 5g networks: A survey. *Computer Networks*, 162:106871, 2019.

[29] Ravishankar Borgaonkar, Lucca Hirschi, Shinjo Park, and Altaf Shaik. New privacy threat on 3g, 4g, and upcoming 5g aka protocols. *Proceedings on Privacy Enhancing Technologies*, 2019(3):108–127, 2019.

[30] 3GPP. Study on the security aspects of the next generation system. Technical Specification (TS) 33.899, 3rd Generation Partnership Project (3GPP), 2017. Version 1.3.0.

[31] Ijaz Ahmad, Tanesh Kumar, Madhusanka Liyanage, Jude Okwuibe, Mika Ylianttila, and Andrei Gurtov. Overview of 5g security challenges and solutions. *IEEE Communications Standards Magazine*, 2(1):36–43, 2018.

[32] Minghao Wang, Tianqing Zhu, Tao Zhang, Jun Zhang, Shui Yu, and Wanlei Zhou. Security and privacy in 6g networks: New areas and new challenges. *Digital Communications and Networks*, 6(3):281–291, 2020.

[33] Hasna Fourati, Rihab Maaloul, and Lamia Chaari. A survey of 5g network systems: challenges and machine learning approaches. *International Journal of Machine Learning and Cybernetics*, 12(2):385–431, 2021.

[34] Telecom Infra Project. "near-real-time ric: Enabling ai/ml-driven extreme automation and granular control of open ran". https://telecominfraproject.com/near-real-time-ric-enabling-ai-ml-driven-extreme-automation-and-granular-control-of-open-ran/ (accessed Jan. 9, 2022).

[35] Imtiaz Parvez, Ali Rahmati, Ismail Guvenc, Arif I Sarwat, and Huaiyu Dai. A survey on low latency towards 5g: Ran, core network and caching solutions. *IEEE Communications Surveys & Tutorials*, 20(4):3098–3130, 2018.

[36] Federica Rinaldi, Alessandro Raschella, and Sara Pizzi. 5g nr system design: a concise survey of key features and capabilities. *Wireless Networks*, pages 1–16, 2021.

[37] Konstantinos Samdanis and Tarik Taleb. The road beyond 5g: A vision and insight of the key technologies. *IEEE Network*, 34(2):135–141, 2020.

[38] *Network Functions Virtualisation (NFV): Architectural Framework*,etsi gs nfv 002 v1.2.1. December 2014. [Online]. Available: https://www.etsi.org/deliver/etsi$_{g}s/nfv/001_{0}99/002/01.02.01_{6}0/gs_{n}fv002v010201p.pdf$.

[39] *Network Functions Virtualisation (NFV); Use Cases*etsi gr nfv 001 v1.2.1. May 2017. [Online]. Available: https://www.etsi.org/deliver/etsi$_{g}r/NFV/001_{0}99/001/01.02.01_{6}0/gr_{N}FV001v010201p.pdf$.

[40] Hassan Hawilo, Abdallah Shami, Maysam Mirahmadi, and Rasool Asal. Nfv: state of the art, challenges, and implementation in next generation mobile networks (vepc). *IEEE Network*, 28(6):18–26, 2014.

[41] Cisco. Cisco annual internet report (2018–2023). White Paper, 2020. Accessed: Oct. 14, 2021. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf.

[42] Ericsson mobility report, ericsson, 2021. [Online]. Available: https://www.ericsson.com/4a03c2/assets/local/mobility-report/documents/2021/june-2021-ericsson-mobility-report.pdf.

[43] ETSI. Mec in 5g networks. White Paper, 2018, [Online]. Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi$_{w}p28_{m}ec_{i}n_{5}G_{F}INAL.pdf$.

[44] *Mobile Edge Computing (MEC);Framework and Reference Architecture*, etsi gs mec 003 v1.1.1. March 2016. [Online]. Available: https://www.etsi.org/deliver/etsi$_{g}s/mec/001_{0}99/003/01.01.01_{6}0/gs_{m}ec003v010101p.pdf$.

[45] Dimitrios Michael Manias, Hassan Hawilo, and Abdallah Shami. A machine learning-based migration strategy for virtual network function instances. In *Proceedings of the Future Technologies Conference*, pages 563–577. Springer, 2020.

[46] Hassan Hawilo, Manar Jammal, and Abdallah Shami. Orchestrating network function virtualization platform: Migration or re-instantiation? In *2017 IEEE 6th International Conference on Cloud Networking (CloudNet)*, pages 1–6. IEEE, 2017.

[47] Dimitrios Michael Manias, Manar Jammal, Hassan Hawilo, Abdallah Shami, Parisa Heidari, Adel Larabi, and Richard Brunner. Machine learning for performance-aware virtual network function placement. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2019.

[48] Dimitrios Michael Manias, Hassan Hawilo, Manar Jammal, and Abdallah Shami. Depth-optimized delay-aware tree (do-dat) for virtual network function placement. *IEEE Networking Letters*, 2(3):149–153, 2020.

[49] Bikash Koley. The zero touch network. In *International Conference on Network and Service Management*, 2016.

[50] Matt A. V. Chaban. “how self-healing networks help keep the digital world stable and secure”. https://www.ibm.com/blogs/industries/self-healing-networks-stable-secure-digital-5g-world/ (accessed Jan. 9, 2022).

[51] Elena Fersman. “zero-touch is coming”. https://www.ericsson.com/en/blog/2019/2/zero-touch-network-is-coming (accessed Jan. 9, 2022).

[52] ETSI. “zero touch network & service management (zsm)”. https://www.etsi.org/technologies/zero-touch-network-service-management (accessed Jan. 9, 2022).

[53] Michael Martinsson. “dynamic orchestration, 5g and ai-powered self-healing networks”. https://www.ericsson.com/en/blog/2019/3/dynamic-orchestration-5g-and-ai-powered-self-healing-networks (accessed Jan. 9, 2022).

[54] *Zero-touch network and Service Management (ZSM); Requirements based on documented scenarios* etsi gs zsm 001 v1.1.1. October 2019. [Online]. Available: https://www.etsi.org/deliver/etsi$_g s/ZSM/001_0 99/001/01.01.01_6 0/gs_Z SM001v010101p.pdf$.

[55] Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. Self-healing networks: redundancy and structure. *PloS one*, 9(2):e87986, 2014.

[56] Xin Li and Chen Qian. The virtual network function placement problem. In *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 69–70. IEEE, 2015.

[57] Wang Miao, Geyong Min, Yulei Wu, Haojun Huang, Zhiwei Zhao, Haozhe Wang, and Chunbo Luo. Stochastic performance analysis of network function virtualization in future internet. *IEEE Journal on Selected Areas in Communications*, 37(3):613–626, 2019.

[58] Bo Han, Vijay Gopalakrishnan, Lusheng Ji, and Seungjoon Lee. Network function virtualization: Challenges and opportunities for innovations. *IEEE Communications Magazine*, 53(2):90–97, 2015.

[59] Huisik Hong, Leifeng Ruan, and Tong Zhang. “ai in the 5g network: Six questions you weren’t supposed to ask”. White Paper, 2021. [Online]. Available:https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/ai-in-the-5g-network-white-paper.pdf (accessed Jan. 9, 2022).

[60] DeepSig AI. “how artificial intelligence improves 5g wireless capabilities”. https://www.deepsig.ai/how-artificial-intelligence-improves-5g-wireless-capabilities (accessed Jan. 9, 2022).

[61] Ericsson. “accelerating the adoption of ai in programmable 5g networks”. White Paper, 2021. [Online]. Available: https://www.ericsson.com/4a3998/assets/local/reports-papers/white-papers/08172020-accelerating-the-adoption-of-ai-in-programmable-5g-networks-whitepaper.pdf (accessed Jan. 9, 2022).

[62] 5G PPP Technology Board. “ai and ml – enablers for beyond 5g networks version 1.0”. https://5g-ppp.eu/wp-content/uploads/2021/05/AI-MLforNetworks-v1-0.pdf (accessed Jan. 9, 2022).

[63] Dimitrios Michael Manias and Abdallah Shami. The need for advanced intelligence in nfv management and orchestration. *IEEE Network*, 35(1):365–371, 2020.

[64] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[65] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053*, 2020.

[66] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[67] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

[68] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[69] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[70] *5G; 5G System; Network Data Analytics Services; Stage 3 (3GPP TS 29.520 version 15.0.0 Release 15)*etsi ts 129 520 v15.0.0. https://www.etsi.org/deliver/etsi$_{t}s/129500_{1}29599/129520/15.00.00_{6}0/ts_{1}29520v150000p.pdf$.

[71] Abdallah Moubayed, Abdallah Shami, Parisa Heidari, Adel Larabi, and Richard Brunner. Edge-enabled v2x service placement for intelligent transportation systems. *IEEE Transactions on Mobile Computing*, 20(4):1380–1392, 2021.

[72] Dimitrios Michael Manias and Abdallah Shami. The need for advanced intelligence in nfv management and orchestration. *IEEE Network*, 35(1):365–371, 2021.

[73] Dimitrios Michael Manias, Manar Jammal, Hassan Hawilo, Abdallah Shami, Parisa Heidari, Adel Larabi, and Richard Brunner. Machine learning for performance-aware virtual network function placement. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2019.

[74] Hassan Hawilo, Abdallah Shami, Maysam Mirahmadi, and Rasool Asal. Nfv: state of the art, challenges, and implementation in next generation mobile networks (vepc). *IEEE Network*, 28(6):18–26, 2014.

[75] Open5GS. “open5gs: Open source project of 5gc and epc (release-16)”. https://open5gs.org/ (accessed Feb. 25, 2022).

[76] Aligungr. Aligungr/ueransim: Open source 5g ue and ran (gnodeb) implementation.

[77] *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Release 16 Description; Summary of Rel-16 Work Items (Release 16)* 3gpp tr 21.916 v16.1.0. January 2022. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3493.

[78] https://github.com/Western-OC2-Lab/5G-Core-Networks-Datasets.

[79] Ramy Mohamed, Sofiane Zemouri, and Christos Verikoukis. Performance evaluation and comparison between sa and nsa 5g networks in indoor environment. In *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, pages 112–116. IEEE, 2021.

[80] Zhengquan Zhang, Yue Xiao, Zheng Ma, Ming Xiao, Zhiguo Ding, Xianfu Lei, George K Karagiannidis, and Pingzhi Fan. 6g wireless networks: Vision, requirements, architecture, and key technologies. *IEEE Vehicular Technology Magazine*, 14(3):28–41, 2019.

[81] Rubayet Shafin, Lingjia Liu, Vikram Chandrasekhar, Hao Chen, Jeffrey Reed, and Jianzhong Charlie Zhang. Artificial intelligence-enabled cellular networks: A critical path to beyond-5g and 6g. *IEEE Wireless Communications*, 27(2):212–217, 2020.

[82] Imad Alawe, Yassine Hadjadj-Aoul, Adlen Ksentini, Philippe Bertin, César Viho, and Davy Darche. Smart scaling of the 5g core network: an rnn-based approach. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2018.

[83] Davit Harutyunyan, Rasoul Behravesh, and Nina Slamnik-Kriještorac. Cost-efficient placement and scaling of 5g core network and mec-enabled application vnfs. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 241–249. IEEE, 2021.

[84] Van-Giang Nguyen, Karl-Johan Grinnemo, Javid Taheri, and Anna Brunstrom. Adaptive and latency-aware load balancing for control plane traffic in the 4g/5g core. In *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 365–370. IEEE, 2021.

[85] Wenzhe Pang, Chenglin Zhao, et al. Amf optimal placement based on deep reinforcement learning in heterogeneous radio access network. 2020.

[86] Juan Pablo Vielma. Mixed integer linear programming formulation techniques. *SIAM Review*, 57(1):3–57, 2015.

[87] Rasoul Behravesh, Davit Harutyunyan, Estefanía Coronado, and Roberto Riggio. Time-sensitive mobile user association and sfc placement in mec-enabled 5g networks. *IEEE Transactions on Network and Service Management*, 18(3):3006–3020, 2021.

[88] Francisco Carpio, Admela Jukan, and Rastin Pries. Balancing the migration of virtual network functions with replications in data centers. In *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, pages 1–8. IEEE, 2018.

[89] Meitian Huang, Weifa Liang, Yu Ma, and Song Guo. Throughput maximization of delay-sensitive request admissions via virtualized network function placements and migrations. In *2018 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2018.

[90] Farah Chahlaoui and Hamza Dahmouni. A taxonomy of load balancing mechanisms in centralized and distributed sdn architectures. *SN Computer Science*, 1(5):1–16, 2020.

[91] Takashi Seyama, Shinya Kumagai, Teppei Oyama, Daisuke Jitsukawa, Takashi Dateki, Koji Matsuyama, Hiroyuki Seki, and Morihiko Minowa. Robust scheduler prioritizing ues with time-variant channels in small-delay slots from channel estimation timing for 5g large-scale mu-mimo. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–5. IEEE, 2018.

[92] Qixia Zhang, Fangming Liu, and Chaobing Zeng. Adaptive interference-aware vnf placement for service-customized 5g network slices. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2449–2457. IEEE, 2019.

[93] Suchao Xiao and Wen Chen. Dynamic allocation of 5g transport network slice bandwidth based on lstm traffic prediction. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 735–739. IEEE, 2018.

[94] Irian Leyva-Pupo, Cristina Cervelló-Pastor, and Alejandro Llorens-Carrodeguas. Optimal placement of user plane functions in 5g networks. In *International Conference on Wired/Wireless Internet Communication*, pages 105–117. Springer, 2019.

[95] Carlos Hernan Tobar Arteaga, Armando Ordoñez, and Oscar Mauricio Caicedo Rendon. Scalability and performance analysis in 5g core network slicing. *IEEE Access*, 8:142086–142100, 2020.

[96] *5G;System Architecture for the 5G System* ETSI TS 123 501 v15.3.0. September 2018. [Online]. Available: https://www.etsi.org/deliver/etsi$_ts/123500_123599/123501/15.03.00_60/ts_123501v150300p.pdf$3.

[97] *NG-RAN; NG Application Protocol (NGAP)* 3GPP TS 38.413 v16.9.0. April 2022. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3223.

[98] Jaeseong Jeong, Dinand Roeland, Jesper Derehag, Åke Ai Johansson, Venkatesh Umaashankar, Gordon Sun, and Göran Eriksson. Mobility prediction for 5g core networks. *IEEE Communications Standards Magazine*, 5(1):56–61, 2021.

Curriculum Vitae

Name:	Ali Chouman
Post-Secondary Education and Degrees:	University of Western Ontario London, ON, Canada 2020 Bachelor of Engineering Science
Honours and Awards:	Ontario Graduate Scholarship 2022
Related Work Experience:	Teaching Assistant The University of Western Ontario 2021-2022

Publications:

A. Chouman, D.M. Manias, and A. Shami. Toward Supporting Intelligence in 5G/6G Core networks: NWDAF Implementation and Initial Analysis. In 2022 International Wireless Communications and Mobile Computing Conference (IWCMC), pages 324–329, 2022.

A. Chouman, D.M. Manias, and A. Shami. A Reliable AMF Scaling and Load Balancing Framework for 5G Core Networks. (awaiting review, GlobeCom)

A. Chouman, D.M. Manias, and A. Shami. Network Data Analytics in Future Networks: Trends, Outlooks, and Future Directions for the 5G Core and Beyond Networks. (pending submission)

D.M. Manias, A. Chouman, and A. Shami. An NWDAF Approach to 5G Core Network Signaling Traffic: Analysis and Characterization. (awaiting review, GlobeCom)

D.M. Manias, A. Chouman, S. Primak and A. Shami. Deep Learning for 5G Wireless Channel Estimation in Cognitive Networks. (awaiting review, CC-ECE)

D.M. Manias, A. Chouman, and A. Shami. A Model Drift Detection and Adaptation Framework for 5G Core Networks. (awaiting review, MeditCom 2022)

D.M. Manias, A. Chouman, and A. Shami. Model Drift in Dynamic Network Environments. (awaiting review, IEEE Communications)