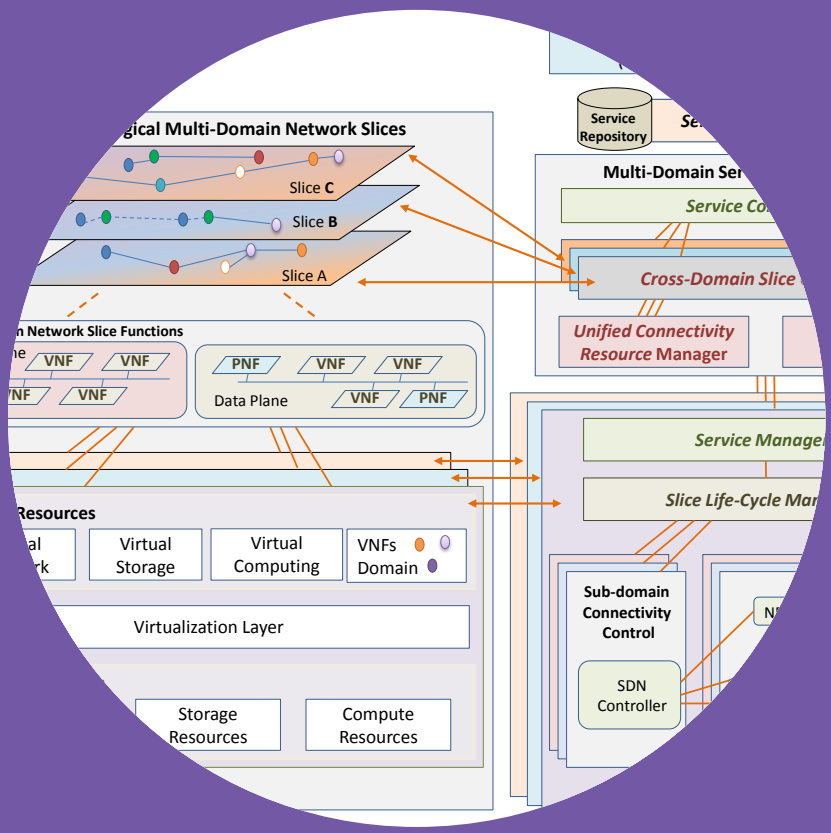
Department of Communications and Networking

Towards a Service-aware E2E Network Slice Orchestration and Resource Provisioning across Multiple Domains

Ibrahim Adewale Afolabi



Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 68/2022

Towards a Service-aware E2E Network Slice Orchestration and Resource Provisioning across Multiple Domains

Ibrahim Adewale Afolabi

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall AS2 of the school on 02 June 2022 at 12PM

Aalto University
School of Electrical Engineering
Department of Communications and Networking

Supervising professor
Professor Jukka MJ Manner, Aalto University, Finland

Thesis advisor
Dr. Miloud Bagaa, Aalto University, Finland

Preliminary examiners
Professor Ari Pouttu, University of Oulu, Finland
Dr. Yulei Wu, University of Exeter, United Kingdom

Opponent
Dr. Tao Chen, VTT Research, Finland

Aalto University publication series
DOCTORAL THESES 68/2022

ISBN 978-952-64-0802-6 (printed)
ISBN 978-952-64-0803-3 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0803-3

Unigrafia Oy
Helsinki 2022

Finland

Publication orders (printed book):
Department of Communications and Networking (Comnet)
P.O.Box 15500, 00076 Aalto, Finland
Maarintie 8, Espoo

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Ibrahim Adewale Afolabi

Name of the doctoral thesis
Towards a Service-aware E2E Network Slice Orchestration and Resource Provisioning across Multiple Domains

Publisher School of Electrical Engineering

Unit Department of Communications and Networking

Series Aalto University publication series DOCTORAL THESES 68/2022

Field of research Networking Technology

Date of the defence 2 June 2022

Permission for public defence granted (date) 18 March 2022 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

The 5th generation mobile network will rely fundamentally on network slicing to provide the required support to enable different vertical industry use cases, such as Ultra Reliable Low Latency Communication (uRLLC), Enhanced/extreme Mobile BroadBand (e/xMBB) and Massive Machine Type Communication (mMTC). In order to deliver these variant 5G use cases to end users from the same physical infrastructure, slice providers have to ensure the customization and optimization of End-to-end (E2E) mobile network slices to truly reflect the type of services that will be running on them. This dissertation is a compilation of research outcomes from our investigations into the practical implementations of various infrastructures and theoretical architectural designs. We also propose a set of algorithms to support the orchestration of service-aware network slices to enable different 5G and beyond use cases.

Network function virtualization and software-defined networking are fundamental enabling principles in network slice orchestration. The orchestration of network slices that cuts across the different technological domains of the mobile network (i.e., RAN, Transport, and Core network) as well as the administrative domains (i.e., different cloud providers, MNO clouds and other private/public clouds) is called E2E network slicing. Orchestrating E2E network slices to deliver different 5G services involves a number of requirements and constraints at different technological domains of the mobile network. Fulfilling these divergent but distinctive network requirements for each slice type while provisioning virtual resources from multiple administrative domains will require a holistic framework and intricate architecture capable of considering all of the possible slicing ramifications and the necessary system support. Traffic isolation, dynamic resource provisioning, and dynamic re/configuration are among the most notable challenges of network slicing.

The aim of the studies presented in this dissertation is to make advances towards overcoming the specific challenges of network slicing. These challenges have been approached from the perspectives of a sound theoretical analysis, the implementation of a practical framework, the design and development of novel architecture, and the development of efficient schemes. We started by conducting a comprehensive survey to provide a holistic understanding of the concept of network slicing, its design principles, and enabling technologies. Then, we implemented a framework for the E2E slicing of the mobile network and 5G services. We then proposed a novel architecture for facilitating the orchestration of network slices from multiple administrative domains. Finally, we developed a scheme with an enabling architecture for the dynamic

Keywords Network Slicing, Network Softwarization, Slice Orchestration, NFV, SDN, 5G, 6G

ISBN (printed) 978-952-64-0802-6 **ISBN (pdf)** 978-952-64-0803-3

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2022

Pages 192 **urn** http://urn.fi/URN:ISBN:978-952-64-0803-3

Preface

In the name of Allah, the most beneficent, the most merciful

I would like to start by giving absolute praise and glory to God, the Almighty, for keeping me safe and sound throughout this interesting journey of my doctoral degree. I consider myself blessed and fortunate to witness the successful completion of this dissertation.

My most profound appreciation goes to my parents and family for always being there for me. Their unwavering support to keep me motivated and inspired has helped me to remain focused despite the many challenges and distractions along the way. To my wife and kids, who are my bedrock and a steady source of joy and balance, I would like to express my deepest gratitude for your understanding, advice, and patience.

I owe a great deal of gratitude to my first supervisor, Prof. Tarik Taleb, who provided me with the opportunity to pursue this doctoral degree. I would equally like to show my heartfelt appreciation and genuine acknowledgement to my second supervisor, Prof. Jukka Manner for his timely interventions, availability and valuable advice. In the same vein, I would like to immensely thank Dr. Miloud Bagaa, my advisor, for his quality time, advice, support and encouragement. My sincere appreciation also extends to the pre-examiners, Prof. Ari Pouttu and Dr. Yulei Wu, and opponent, Dr. Tao Chen for their noteworthy remarks.

To my colleagues and friends in COMNET, I say a massive thank you to you for all the beautiful moments we shared together. You will forever remain in my heart.

Espoo, April 20, 2022
Ibrahim Afolabi

Table of Contents

List of Publications

This dissertation consists of an overview of the following publications, which are referred to in the text by their roman numerals.

I I. Afolabi, M. Bagaa, W. Boumezer and T. Taleb, "Toward a Real Deployment of Network Services Orchestration and Configuration Convergence Framework for 5G Network Slices," *IEEE Network Magazine*, vol. 35, no. 1, pp. 242 - 250, doi. 10.1109/MNET.011.2000146, Jan/Feb 2021.

II I. Afolabi, J.P. Garzon, M. Bagaa, T. Taleb, and P. Ameigeiras, "Dynamic Resource Provisioning of a Scalable E2E Network Slicing Orchestration System," *IEEE Transactions on Mobile Computing*, vol. 19, no. 11, pp. 2594 - 2608, doi. 10.1109/TMC.2019.2930059, Nov 2020.

III T. Taleb, I. Afolabi, K. Samdanis and F. Z. Yousaf, "On Multi-domain Network Slicing Orchestration Architecture and Federated Resource Control," *IEEE Network Magazine*, vol. 33, no. 5, pp. 242 - 252, doi. 10.1109/MNET.2018.1800267, Sep 2019.

IV I. Afolabi, T. Taleb, P. A. Frangoudis, M. Bagaa and A. Ksentini, "Network Slicing-Based Customization of 5G Mobile Services," *IEEE Network Magazine*, vol. 33, no. 5, pp. 134-141, doi. 10.1109/MNET.00-1.1800072, Sep 2019.

V T. Taleb, I. Afolabi, and M. Bagaa, "Orchestrating 5G Network Slices to Support Industrial Internet and to Shape Next-Generation Smart Factories," *IEEE Network Magazine*, vol. 33, no. 4, pp. 146 – 154, doi. 10.1109/MNET.2018.1800129, July 2019.

VI I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies and Solutions," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 3, pp. 2429-2453, doi. 10.1109/COMST.2018.2815638, Mar 2018.

Author's Contributions

Publication I: "Toward a Real Deployment of Network Services Orchestration and Configuration Convergence Framework for 5G Network Slices"

The author prepared the content of the manuscript, provided the literature review, and wrote the paper. Prof. Taleb supervised the work during the research process and contributed to the manuscript by making revisions and enhancing the quality of the paper. The author and other authors participated in the development of the framework and contributed towards the system setup and generation of meaningful results.

Publication II: "Dynamic Resource Provisioning of a Scalable E2E Network Slicing Orchestration System"

The original idea of a scalable network slicing orchestration system was introduced by Prof. Taleb. The author defined the research questions, the problem formulation and developed the very first version of the system model and introduced the initial algorithm. The author also developed the federated NSOS architecture, which was enhanced by Prof. Taleb. The queuing model and auto-scaling solutions were initiated and defined by the author, and enhanced by Dr. Garzon. Similarly, the system simulation was supervised by Dr. Bagaa and Dr. Garzon respectively. Dr. Garzon also provided useful suggestions in analysing the results, while Dr Ameigeiras supported in the overall enhancement of the research work by providing useful important comments through revision.

Publication III: "On Multi-domain Network Slicing Orchestration Architecture and Federated Resource Control"

Prof. Taleb and Dr. Samdanis co-developed the idea of the paper and provided the preliminary draft. The author enhanced the manuscript significantly through iterations of revision, addition of important comparative studies that are relevant to the subject matter and redesign of the proposed architectures in order to better convey the intended purpose and substantiate the overall construct of the research work. Dr. Zarrar introduced the slice creation procedures and techniques and support in the overall revision of the paper. Prof. Taleb also improved the overall quality of the paper by suggesting valuable insights at the end of the revision cycle.

Publication IV: "Network Slicing-Based Customization of 5G Mobile Services"

The idea of the research work was introduced by Prof. Taleb and enhanced by Dr. Bagaa. The author developed the research work, did the literature review, conducted the test and evaluated the system. The developed work led to the implementation of an end-to-end (E2E) slice orchestration and management framework known as E2E Slicer. The E2E Slicer tool was jointly developed by the author, Dr. Frangoudis and Dr. Bagaa. The tool was further enhanced and modularized by the author. Prof. Ksentini provided very useful suggestions and pieces of advice for the enhancement of the research work and support in the revision of the work.

Publication V: "Orchestrating 5G Network Slices to Support Industrial Internet and to Shape Next-Generation Smart Factories"

The idea of this research work was conceived by Prof. Taleb and elaborated upon by Dr. Bagaa but developed by the author. The author in collaboration with Dr. Bagaa under a close supervision of Prof. Taleb also wrote the paper and developed the proposed architecture of the framework. The use cases were fine-tuned by Prof. Taleb after iterations from the author. The author did the system setup, generated the results and did the final analysis of the work.

Publication VI: "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies and Solutions"

The idea of the research paper was incubated by the author with supervision from Prof. Taleb. The author prepared and developed the first version of the paper with most of the literature review and largely wrote the paper. Very valuable suggestions and additions were provided by Dr. Samdanis and Prof. Ksentini. Mr Flinck also supported in revising the paper and providing useful insights. Prof. Taleb supervised the structure of the work and its development. He also provided the final revision in order to enhance the quality of the paper.

List of Abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth Generation wireless network technology
AF	Application Function
AMF	Access and Mobility management Function
API	Application Programming Interface
AUSF	Authentication Server Function
CAPEX	CAPital EXpenditure
CDN	Content Delivery Network
cRAN	cloud Radio Access Network
Cloud SP	Cloud Service Provider
CN	Core Network
ComNet	department of Communications and Networking
CP	Control Plane
CriC	Critial Communications
CSMF	Communication Service Management Function
CSP	Communication Service Provider
DASA	Dynamic Auto-Scaling Algorithm
DC	Data Center
DCN	Dedicated Core Network
DRP	Dynamic Resource Provisioning
DSO	Domain-Specific Orchestrator
DN	Data Network
DP	Data Plane
E2E	End-to-End
EC	Edge Cloud
eMBB	Enhanced Mobile Broadband

eNB	evolved Node B
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
eV2X	Enhanced Vehicular to Everything
GO	Global Orchestrator
HSS	Home Subscriber System
InP	Infrastructure Provider
IoT	Internet of Things
ISP	Internet Service Provider
IT	Information Technology
LTE	Long Term Evolution
MAC	Medium Access Control
MEC	Multi-access Edge Computing
mIoT	Massive Internet of Things
mMTC	Massive Machine Type Communication
MME	Mobility Management Entity
MNO	Mobile Network Operator
MVA	Mean Value Analysis
MVNO	Mobile Virtual Network Operator
NBI	NorthBound Interface
NEF	Network Exposure Function
NFFG	Network Function Forwarding Graph
NFV	Network Function Virtualization
NFVI	Network Functions Virtualization Infrastructure
NFVO	Network Functions Virtualization Orchestrator
NFV MANO	Network Functions Virtualization Management and Orchestration
NSB	Network Slice Broker
NSI	Network Slice Instance
NSMF	Network Slice Management Function
NSOS	Network Slicing Orchestration System

NSSMF	Network Slice Subnet Management Function
OPEX	OPerational EXpenditure
OTT	Over-The-Top
PGW	Packet data network GateWay
PHY	Physical
PLMN-id	Public Land Mobile Network Identifier
PM	Physical Machine
PNF	Physical Network Function
QNA	Queuing Network Analyzer
QoE	Quality of Experience
QoS	Quality of Service
RAE	Resource Awareness Engine
RAN	Radio Access Network
RRB	Radio Resource Blocks
SAE	System Awareness Engine
SDN	Software Defined Networking
SFC	Service Function Chaining
SGW	Serving GateWay
SIG	Slice Information Graph
SLA	Service Level Agreement
SMARTER	new Services and Markets Technology Enablers
SMF	Session Management Function
SOR	Slice Orchestration Request
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management
uRLLC	Ultra Reliable Low Latency Communication
vEPC	virtual Evolved Packet Core
VIM	Virtual Infrastructure Manager
VM	Virtual Machine
VNE	Virtual Network Embedding
VNF	Virtual Network Function
xMBB	Extreme Mobile Broadband

List of used Latin Symbols

a_k, b_{ik}	Coefficients of the set of linear equations to estimate the SCVs of the aggregated arrival process at each queue k
$C(m, \rho)$	The Erlang's C formula
c_{ak}^2	Squared coefficient of variation of the aggregated arrival process at queue k
c_{0k}^2	Squared coefficient of variation of the external arrival process at queue k
c_{sk}^2	Squared coefficient of variation of service processes at queue k
c_{si}^2	Squared coefficient of variation of service processes at queue i
K	Number of G/G/m queues
k, i	Network nodes indexes
m_i	Number of servers of node i
P	Steady state transition probability matrix
p_{ki}	Probability of a packet leaving a node k to node i
p_{0k}	Probability that a packet leaves the network
q_{0k}	The proportion of arrivals to node k from its external arrival process
q_{ik}	The proportion of arrivals to node k from node i
T	The overall mean response time
$Tmax$	Time threshold for which T has to be kept
T_k	Mean system response time of node k
V_k	The average number of visits to node Q_k
W_k	Mean waiting time of node k
$W_k^{M/M/m}$	The mean waiting time for an M/M/m queue

List of used Greek Symbols

β	The Kraemer and Langebach-Belz approximation
λ_{0k}	Mean arrival rate of the external arrival process at queue k
μ_k	Mean service rate of each server at queue k
ρ_k	The utilization of the node k
$C(m, \rho)$	The Erlang's C formula
ω_k, x_i, γ_k	Auxiliary variables when a_k and b_{ik} are computed

List of Figures

1. Introduction

The global network coverage of 4G technology has steadily increased over the years, with coverage extended to approximately 80 percent of the total global population at the end of 2019 and 90 percent expected to be covered by 2025 [1]. There is no doubt that the global adoption and deployment of the 4G mobile network technology has been very successful and is expected to continue in this vein in the coming years. The increase in the global acceptance and deployment of the 4G mobile network technology and its capabilities has stimulated the creation of a wide range of services and applications.

These services are continually evolving and pushing the limits of the 4G mobile network in terms of both capacity and capabilities. In particular, requirements such as network latency, throughput, reliability, allowed jitter, allowed packet loss and availability have been out to the test. Each of these requirements serves as a significant determinant of the quality of service (QoS) expected by the service and the quality of experience observed by the application owners. Due to the steadily increasing innovative development of a wide variety of services, driven mostly by constant competition and the need to serve both existing and new market verticals better than before, different variations of these requirements have had to be composed to enable different services. These variations in requirements have sometimes been noticeably distinctive and diverse. Over time, it is becoming increasingly evident that operating this array of services on the same network technology is not sustainable. That is, the centralized one-fits-all nature of the 4G architecture is becoming outdated [2–4].

Consequently, another architecture had to evolve both to serve the improved old services better and to accommodate the identified set of new verticals [5–9], which have enabled and provided a wide range of business opportunities. To this end, the development of the 5th generation mobile

network, dubbed 5G, is inevitable. 5G technology is designed to address the numerous challenges that 4G technology could not properly tackle, especially from an architectural perspective. It also intends to accommodate new innovative services with requirements beyond the capacity and capability of the 4G network technology. Specifically, 5G technology promises to enable services with strict requirements [11] such as low latency over the air (1ms) and high reliability, a massive data rate of 10Gbps, and in some cases up to 20Gbps. It has a 100 Mbps data rate broadband, the energy efficiency is improved by a factor of 100 over the 4G network, and the spectral efficiency is three times higher than the 4G network [11].

Not only will the 5G network technology support the provisioning of services with these sets of requirements, it is actually designed to do so while enabling the operation of these services in parallel. Bearing in mind that the required composition of some of the identified and aggregated 5G services, as per the 3GPP classifications, such as the Massive Internet of Things (MIoT) [6], Critical Communications (CriC) [7], Enhanced Mobile Broadband (eMBB) [8] and Enhanced Vehicular to Everything (eV2X) [12] are remarkably diverse and sometimes opposing [13–16]. Therefore, to efficiently utilize an infrastructure capable of powering these systems, a new concept known as *Network slicing* came to life.

Network slicing is an enabling technology concept first introduced by the Next Generation Mobile Network (NGMN) body [18]. The aim of network slicing is to promote the efficient utilization of network infrastructure and value creation for vertical industries. The network slicing concept has introduced an entirely new paradigm to network service creation, configuration, and management. The high-level architecture of this networking technology, as depicted in Figure 1.1 [18], allows multiple instances of services to run in parallel on shared physical infrastructure. In some cases, the sharing may also extend beyond the physical resources to some common functionalities needed to control and manage the services. Moreover, some of these services may sometimes demand a diverse set of business requirements. This approach is possible because the system leverages other essential and fundamental enabling technologies, such as *network virtualization, network softwarization* and *software defined networking*.

The high-level architecture of the network slicing concept, presented in Figure 1.1, has motivated several research works [3, 20–22], including research presented in this dissertation. This motivation arises from the need to properly investigate and understand how the orchestration and

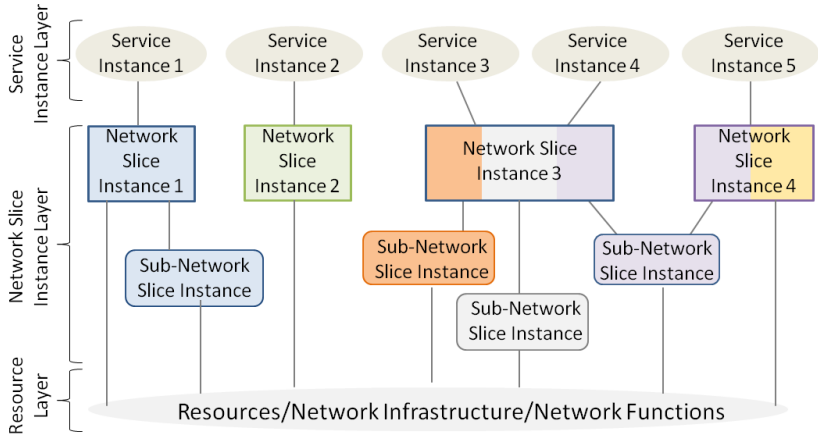


Figure 1.1. NGMN Network Slicing Concept [19].

life-cycle management of network slices can be seamlessly achieved to customize services using slicing principles based on the requirements of the relevant emerging businesses. Among the many advantages of network slicing, a significant advantage is the creation of value for vertical business segments, over-the-top application providers, and other third-party players. The inception of network slicing has also enabled the network ecosystem to accommodate service brokers who do not have their physical network infrastructure by offering physical and virtual resources, such as the radio, networking, computing, and storage. These resources are used to provide customized network services, each with unique and well-defined QoS requirements and service treatment while running in parallel.

Based on the principles of *Automation, Customization, Elasticity, End-to-end, Isolation, Hierarchical abstraction* and *Programmability* enabled by network slicing, the 5G technology will be able to power the numerous promising use cases identified in the various 3GPP next-generation SMARTER reports [6–8, 12]. Some of these use cases were also described and enumerated by the NGMN in their vision for the 5G white paper as follows: *Broadband access everywhere, Broadband access in dense areas, Extreme real-time communication, High user mobility, Lifeline communication, Massive internet of things, Multi-connection, Ultra reliable communication, etc.* A better-aggregated list of these sets of 5G use cases is presented in Figure 1.2, where, in addition to the above-mentioned use cases, there are others, such as *Service integration and migration, and Lightweight communication*, each of which has a different set of service characteristics. For example, while on one hand, *Massive internet*

of things use case is characterized with massive low-cost, long-range, low-power broadband machine type communication (MTC) for smart wearable sensor networks, on the other hand, *Extreme real time communication* is simply characterized with ultra-low latency communication for tactile internet services.

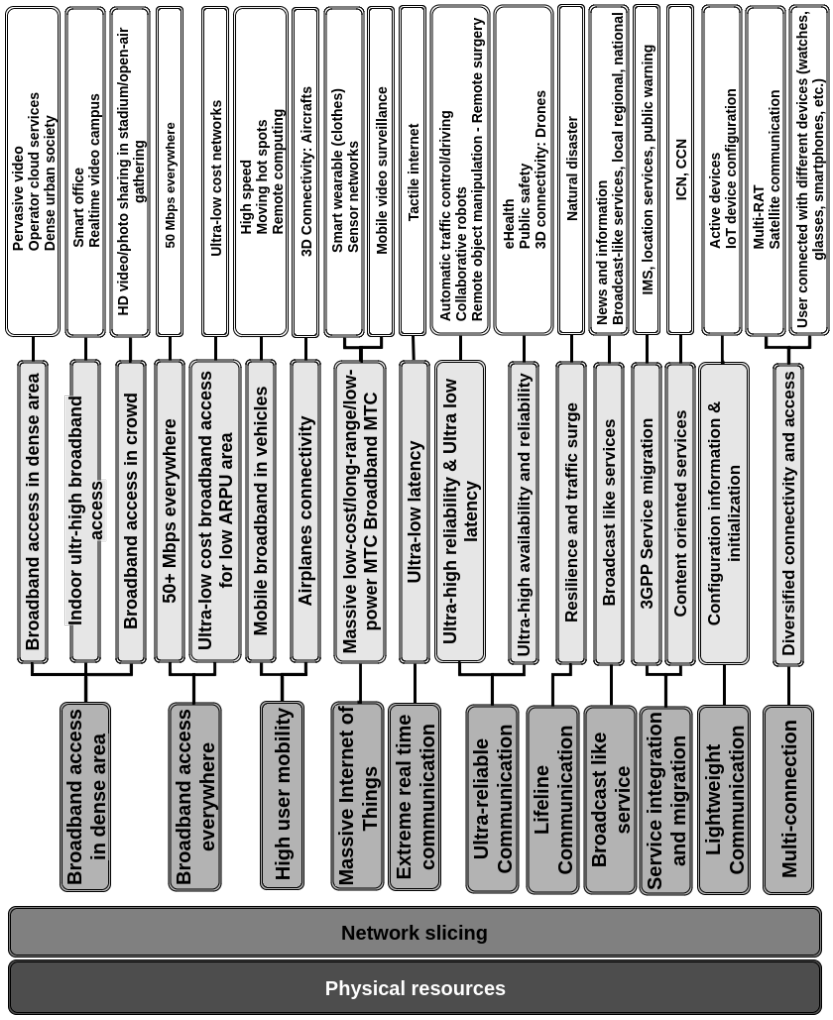


Figure 1.2. 5G slicing use cases [3].

1.1 Scope and objectives

Network slicing or slice orchestration, is a networking approach that allows for the parallel deployment, operation, and life-cycle management of logically separated virtual networks on top of a shared physical infrastructure. Each deployed virtual network and resource is fine-tuned and

customized to deliver a particular 5G service. The use of the network slicing approach for the orchestration of different 5G services, with sometimes diverse functional and operational needs, requires a robust orchestration solution capable of taking into account all of the variant parameters that could affect the proper differentiation and efficient operation of network slices. These service differentiations are specified and described through standard slice templates, e.g., generic network slice template (GNST) defined by GSMA [10], based on the slice types, ensuring that virtual and physical resources are optimally allocated.

Different network slices require a different combination of resources based on different QoS constraints. They may also impose additional needs specific to enabling certain use cases based on users demands. For instance, an extended reality (XR) slice, which is a type of e/xMBB slice, may also specify the use of a particular brand of CPU and GPU, and a particular virtualization technology, e.g., cloud native containers such as kubernetes, to better support its performance and expected user experience. As a result, while the requirements of different slice types may vary significantly, the specifications of slices of the same type may be non-uniform.

Furthermore, besides using heterogeneous and distributed resources cutting across different technologies for the orchestration of network slices, some slices may also require coverage across multiple administrative domains. This particular requirement has introduced a new paradigm to the approach of network slice orchestration: mainly that this need might come on-demand (i.e., the original slice is expanded to cover users in a nearby city/region). Accommodating this level of flexibility in slice creation and management will also open a new ecosystem of network players [17] (e.g., network infrastructure owners, vertical application providers, MVNO, service brokers, OTT, MNO) and propose new functionalities for the efficient management of slices across multiple technological and administrative domains. In this way, with a clear distinction in the roles and services offered by the different players, not only will a new market be created, but this approach will also drive the efficient utilization and lower operation expenditure (OPEX) of the network infrastructure and bring about better value for capital investment (CAPEX).

Besides the fact that the orchestration of network slices and meeting their ever-changing operational demands to address their use case scenarios is essential, another critical aspect of slicing is providing adequate

resources to support the Network Slice Orchestration Systems (NSOS). This aspect is critical for NSOS in order to increase its productivity and efficiency.

In this respect, this dissertation seeks to provide satisfactory answers to the following relevant research questions, each addressed across its different chapters as presented in Table 1.1:

1. What are the network slicing enabling technologies and how can we facilitate network slice orchestration across the RAN, Core and Transport network parts?
2. How can we orchestrate customized network slices in a service-aware fashion?
3. How can we dynamically provision virtual network resources to network slice orchestration systems in order to improve their operational efficiency?
4. How can we create an efficient platform that converges the existing ETSI MANO compliant orchestration framework for the orchestration and management of network slices across multiple administrative cloud domains?
5. How can we apply an efficient platform such as the one described above to enable the use case of Industry 4.0?

In this dissertation, we have considered and presented all of the slicing enabling technologies, aspects of network slice orchestration, implementation of slicing framework, proposition of techniques and algorithms for predictive resource provisioning for NSOS, and evaluation of the achieved results, simulations, and experiments from the entire system setup. The details of these studies are discussed in the works presented in the list of Publications. The relationships between this dissertation, the research questions, and the Publications are aggregated in Table 1.1.

1.2 Contribution of the dissertation

The research work presented in this dissertation is aimed at advancing the state of the art concerning network slicing, focusing on service-oriented mobile network slice orchestration and vertical industry service provisioning. This work intends to deepen the available knowledge regarding the different aspects of network slice orchestration and life-cycle

Table 1.1. Relationship between research questions, publications, and chapters in this dissertation.

Question	Main theme	Publication	Chapter/Section
Q1	Network Slice Orchestration and composition	VI	Chapter 2
Q2 and Q5	Orchestration of Network Slices for the customization of 5G services	IV and V	Chapter 3
Q3	Dynamic resource provisioning for network slicing orchestration system	II	Chapter 4
Q4	Multi-domain slice orchestration over federated resources	I and III	Chapter 5

management of network slice instances, and for efficient management of the orchestration systems. It provides an essential and elaborate work overview of leading academic research works. It introduces robust network slicing architectures, carefully developed to support the 5G slice orchestration systems and beyond. The proposed architectures are robust for the orchestration of network slices in enabling vertical industry applications and services from multiple technological and administrative cloud domains. The findings presented in this dissertation are the outcomes of our research. They are compiled for the primary purpose of advancing the field of research in the direction of network slicing in the following ways:

- **Comprehensive survey of the principles, enabling technologies and solutions for network slicing, Publication VI.** This survey, in our humble opinion, is the first of its kind on this particular subject. It presents a detailed discussion on the principles and enabling technologies of network softwarization and slicing by introducing diverse 5G service use cases and relevant ongoing research works in their early stages. It also overviews various existing network slicing architectures and their different building blocks. The study captures the perspectives of different standardization bodies on network slicing as an enabler of the 5G technology, most significantly, the 3GPP and ITU-T. Furthermore, the study examines the different stages of network slicing from de/composition slice requests, orchestration, de/commissioning to life-cycle management. Finally, we present the likely future challenges in network slicing, and the potential for this dissertation to serve as the foundation for other studies.

- **An end-to-end dynamic network slice orchestration frame-**

work (E2E Slicer), Publication IV. This framework is developed to have a complete software tool for the dynamic orchestration of network slices while taking into account the diverse resource requirement of different vertical applications. In order to investigate and propose a robust framework for the orchestration of network slices that are service-aware in an E2E manner, there is a need to develop one. This need is what has necessitated the implementation of the E2E Slicer, which takes cognizance of the variations in the resources requirements of the different types of network slices, such as eMBB, mMTC, uRLLC, etc, from the RAN network to the core network part of the 4G/5G/B5G systems. This system can allocate the appropriate amount of network resources based on the slice type definitions available in the service catalog for different slices and maintain it throughout the life-cycle of the network slice.

- **Dynamic Auto-Scaling Algorithm for network slicing orchestration system, Publication II.** The work introduces a hierarchical architecture for the orchestration of network slices while incorporating dedicated entities per domain to efficiently manage the different segments of the mobile network, which spans across the access, and transport to the core network part. To maximize the productivity of the system and enable it to scale autonomously and adapt to changes in slice orchestration requests, we also introduced a scheme called DASA, which includes both proactive and reactive virtual resource provisioning capabilities. This scheme ensures that the system respects a given average overall time needed to process any slice orchestration request and scale autonomously when necessary to satisfy the set objective. This virtual resource dimensioning algorithm is based on a queuing model developed for the NSOS.

- **Multi-domain slice orchestration architecture and federated resource control, Publication III.** This work introduces a sophisticated management and orchestration architecture for an efficient orchestration and life-cycle management of multi-domain network slices that share federated resources across multiple administrative domains. The proposed architecture seamlessly integrates the capabilities of the SDN and NFV technologies into the 3GPP network slice management approach. This concept is developed to facilitate the coherent and smooth composition of network slices using stan-

dard models and consistent constructs.

- **Convergence framework for the orchestration and configuration of 5G network slices over multiple domains, Publication I.** The earlier work has inspired and stimulated the implementation of a robust and straightforward network slicing framework for the orchestration and automated configuration of network slices across multiple administrative domains. This framework provides a common platform for unifying virtual resources managed and controlled by totally independent network service orchestrators. It then harmonizes their operation as a single entity through a federated slice template/blueprint that reflects the different placement of the individual virtual network functions (VNF)s that compose the resultant multi-domain slice instance. The framework takes advantage of existing widely explored VNF and physical network function (PNF)s orchestrators using standard open APIs to implement this network orchestration convergence platform.

- **5G network slice orchestration for supporting industrial internet, Publication V.** The objective of this work was to address the need of the next generation smart factories by enabling a paradigm shift towards Industry 4.0. It is designed to enable the industrial internet of things use-case of 5G technology by enabling seamless interconnectivity between different factory sites through network slicing methodology. This study exposes the exact needs and use of network slicing in promoting inter-factory cooperation and provides the necessary flexibility in supporting the operational requirements of Industry 4.0 over heterogeneous domains and resources, and proposes an efficient slicing architecture specific to the industrial environment.

2. Network Slice Orchestration and Composition

In this chapter, we provide a high-level overview of the orchestration of network slices and their composition. We also present insight into the leading enabling technologies of network slice orchestration and discuss how the identified technologies have shaped the orchestration of network slices. Later on in the chapter, we focus our attention on the different mobile network segments and sub-networks and their participation in the orchestration of E2E mobile network slices. This Chapter captures expressly the contribution of **Publication VI** as presented earlier in Table 1.1.

2.1 Network slicing concept

Network slices are essentially composed of carefully planned virtual resources that are reserved, instantiated, and configured from the different segments of a complete network solution, all running on a set of hardware infrastructure. These concept and procedures are possible thanks to advancements in the programmability of network infrastructure. The network infrastructure programmability and virtualization principles have been the primary focus of the standardization efforts by the European Telecommunications Standards Institute (ETSI). This concept seeks to virtualize everything about network functions, including their resources and operational management; that is, the network functions virtualization (NFV) concept. The basis for this concept is detailed in the NFV architectural framework in ETSI's pioneering document [26]. This document provides the details of the NFV framework and presents the essential architecture to support its foundation. The NFV foundation architecture has evolved over the years with extensions to accommodate the additional building blocks needed to enable different customized 5G ser-

vices, as presented in [27–29]. However, it should be noted that the ETSI architectural framework was initially defined with basically four main core components, namely:

- NFV management and orchestration,
- network functions virtualization infrastructure (NFVI),
- virtual network functions (VNFs), and
- the operational support system/business support system (OSS/BSS).

2.2 Key network slice orchestration enabling technologies

The key enabling technologies for the orchestration of Network slices are network virtualization, network softwarization, and software-defined networking. These enabling technologies are defined and discussed below.

Network virtualization as an enabling technology has garnered massive attention over the past decade or so. This can be largely attributed to the enormous potential it offers as an essential enabling technological component of network slice orchestration. Simply put, network virtualization is a concept that focuses on and promotes a software-based representation of network functions and resources that are normally implemented on integrated hardware components and are partly integrated in software applications for both user and control plane functionalities. Hence, the idea of network function virtualization comes to life. It is fundamental to the softwarization of networking platforms and their operational resources. As a result, this concept has given rise to the idea of network softwarization.

Network softwarization, on the other hand, is an umbrella concept that encompasses the planning, design, architecture, development, deployment, and efficient management of networks and their components. All of these processes are carried out fundamentally using software programmability properties and principles, as detailed in [23]. These principles and properties enable the network to be more flexible, agile, adaptive, and dynamically reconfigurable based on timely changes in network requirements, which directly reflects network behavior. These changes are often introduced due to functional, operational, and cost optimization during the life-cycle management of the network.

Furthermore, the concept of software-defined networking (SDN) is to separate the control plane functionalities from the data plane functionalities of networking components and hardware equipment. This way, the network will have a centralized global logical control view while keeping the data plane as light-weight as possible [24, 25]. This principle has introduced more intelligence to the network by making it agile, scalable, service-oriented, and dynamically adaptable to reflect changes in networking requirements. By so doing, SDN technology has enabled the network to evolve with a simplified management view and has led to innovative networking services.

2.3 Network slice orchestration

According to the 3GPP [30], network slicing is a technology that allows mobile network operators to create customized networks with optimized solutions in terms of isolation, functionality, and performance to fulfill the diverse requirements of different emerging market scenarios. Similarly, ITU-T [31] defines network slicing as logically isolated network partitions composed of multiple virtual resources and equipped with a programmable interface for the management and control of both the data and the control plane. Therefore, by marrying these two interpretations of network slicing together, a more comprehensive definition of network slicing is produced. Thus, network slicing can now be defined as a concept that facilitates the creation and operation of multiple logically fully/partially separated networks orchestrated on top of a shared physical infrastructure that brings together a mutually beneficial existence of stakeholders to create added-values in terms of business innovation while integrating physical and virtual cloud resources into an open software-oriented multi-tenant network environment.

2.4 Network slice composition

A fully-fledged mobile network slice is composed of variant virtual network functions (VNFs), physical network functions (PNFs), a dedicated or shared virtual network, as well as cloud resources, customized value-added services in the form of software or hardware in the RAN, with transport and core networks homogeneously connected to serve a predefined service requirement. In this light, a network slice is composed of

carefully chained and strategically placed PNFs and VNFs, which are the smallest elementary units of its components. The process of connecting or chaining the different functional units of a slice to deliver a particular network service following specific activation order within a network slice is known as *service function chaining* [32]. Moreover, the strategic positioning or placement of VNFs and PNFs [33, 34] based on specific known service characteristics and user behaviors is known as *Virtual network embedding* [35]. All of these different concepts usually come into place in the composition of a network slice. These concepts are then reflected in the slice using the slice blueprint or template, often present in the slice service descriptor.

A VNF is typically hosted in a virtual machine (VM). All the essential and necessary virtual resources for computation, networking, and storage are commonly provided for the smooth operation of deployed function in the VM. The VM naturally emulates a physical machine, provided with the needed operating system and other software drivers needed to function correctly as a complete machine. Depending on the underlying virtualization technology, VMs could be deployed from a physical machine using either full virtualization or para-virtualization hypervisor technology [36].

Containers have already gained a tremendous amount of traction over the years to run virtual functions [37]. Containers are a light-weight form of VMs that often share the same kernel space as other containers deployed on the same host machine. The level of isolation between traditional containers is usually limited to the applications' scope running on each. Typical examples of containers are the LXC/LXD, Docker containers, and the Kubernetes [38, 39].

2.5 Network slice subnets

According to the vision of NGMN for 5G network [18], the instantiation of an E2E network slice is envisioned to cover 3 major layers, namely, the *Service instance*, *Network instance* and *Resource layer* as presented in Figure 1.1. In the network slice instance layer, a network slice could belong to multiple sub-network slice instances, while multiple network slice instances could also share the same sub-network slice instance. These slice sub-network instances could either be a control plane or data plane function in the literal perception, and these form the basis for the composition

of the slice(s).

This slice composition approach implies that a mobile network slice would be composed of sub-network/subnet slice instances from the RAN, Transport, and Core network segments of the 4G/5G network. Each of these slice subnets would have a dedicated management and operational control function, a subset of the network slice management function (NSMF) that is committed to monitoring the operations of the instance of each slice subnet. This function is known as the network slice subnet management function, as defined by the 3GPP [40].

2.5.1 RAN network slice subnets

RAN network slicing began as network sharing, which is not new in the literature. The 3GPP has proposed different scenarios for which the RAN network segments are shared between network operators and other mobile virtual network operators [54]. To enhance the sharing scenarios and procedures further, 3GPP has defined the concept of master operator-network manager (MO-NM) and the sharing operator-network manager (SO-NM) in [55]. This functional component manages the sharing of network resources between the host mobile network operator, referred to as the master, and the tenant network operator referred to as the sharing operator. In this case, the RAN sharing is typically done based on a long-term fixed contract between the two parties, whereby network resources and capacity are allocated to and released from the participating sharing operators.

RAN slicing seeks to take this process a step forward by allowing a dynamic RAN resources allocation and admission control between multiple MVNOs, OTT and industry verticals, and the host network operator in an on-demand fashion with a specific duration validity, while keeping to a predefined QoS SLA. The RAN resources and capacity are generally determined and configured across the protocol layers responsible for specifying the network capacity and scheduling of physical resources, the PHY, and MAC scheduling layers. The physical radio resource block (PRB) is the elementary unit responsible for the network bandwidth, and the frequency of scheduling is responsible for the network latency. These two main factors are the essential components that determine the performance of network slices at the RAN [56]. It is possible to achieve the RAN resource capacity allocation and isolation either statically or dynamically [57].

Typically, the RAN slice will constitute the RAN sub-network compo-

nent of the slice, which in this case, will be a subset of the RAN physical network functions and resources. As a result, the RAN subnet is expected to be a combination of VNF(s) and PNF(s). The PNF composition of the RAN subnet are due to the involvement of the physical functional components, such as the radio heads and other physical functions running in the RAN cabinet, which do not support virtualized implementations yet.

2.5.2 Core network slice subnets

The core network slice sub-network/subnet provides the core network functions that constitute the network slice. This core subnet is also called the core network sub-slice, and is typically comprised of virtualized implementations of the core network elements. The ETSI NFV standards have enabled the virtual implementation of most network services that offer support for virtual environment deployment.

This functional implementation approach for the deployment of network functions over NFV infrastructure is expected to bring the necessary flexibility, programmability, elasticity, and customizability to meet the requirements of the network slice instance at the core network sub-slice level. The elasticity characterization aspect of the sub-slice will allow the network slice to scale on-demand, thereby promptly facilitating the resource needs and performance optimization for the network slices. The ETSI MANO standard has a robust definition to support these functionalities in its relevant architecture framework using standard API definitions.

2.5.3 Transport network slice subnets

The transport network slice sub-network/subnet forms the transport segment of the mobile network sub-slice. This sub-slice is enabled thanks to technologies such as virtual LANs and SDN. They form an essential part of the complete network slice.

Moreover, the transport network slice subnet is not only traditionally useful in connecting the RAN sub-slice to the core sub-slice, but is also absolutely essential nowadays to facilitate the creation of an appropriate transport path for steering the traffic of the connected RAN and core network sub-slices to the intended external packet data network(s) using standard networking technology.

The transport network is undoubtedly one of the most challenging seg-

ment of the mobile network to slice. This challenge is as a result of the fact that the communication traffics for the different slice types only travel through the same physical transport medium with almost no possibility to practically enforce strict traffic isolation as in the case of virtualized environment. Although, traffic belonging to different slices can be treated differently through the use of different layer 2 and 3 technologies, nonetheless, ensuring absolute isolation of different slices across the fronthaul/midhaul and backhaul segment of the network still remains a major challenge.

To overcome this challenge, various techniques are rapidly gaining momentum, most notably, the concept of Time-Sensitive Networking (TSN), which allows traffic prioritization, scheduling, regulation, path reservation and selection in a deterministic network environment [58, 59] for enabling transport slicing support on the physical infrastructure. In addition, different best practices are being proposed in the literature, which include the use of different technology, mostly as overlay networks [60] that will run on the physical spine & leaf switches as well as routers that are running on the network. These technologies may include the use of L2/L3 VPNs, MPLS tunnels, IP tunnels, overlay transport virtualization (OTV), segment routing, different L2 VLANs and so on, depending on the transport subnet profile of a particular network slice [61,62] and technologies supported by the underlying infrastructure.

2.6 Summary

In this chapter, we have presented the concept of network slicing and discussed the different enabling technologies needed for the orchestration of network slices. In addition, we have introduced the reader to the principle and proper definition of network slice orchestration, most importantly from the approach of two renowned standardization bodies, namely, the 3GPP and ITU-T. Furthermore, we have provided a comprehensive discussion on the composition of a complete mobile network slice. Finally, we examined the dynamics of network slice subnetworks based on the vision of NGMN for the 5G and beyond networks, which is primarily composed of the RAN, Core and Transport network slice subnets as detailed in **Publication VI**. Through this chapter and the **Publication VI** from which it has been extracted, we have managed to give a comprehensive foundation of the network slicing principles and its enabling technologies to the

reader. In the following chapter, we will highlight and discuss the orchestration of network slices to support the customization of 5G and beyond services based on the contributions of **Publications IV** and **V**.

3. Orchestration of Network Slices for the Customization of 5G Services

This chapter is focused on how to orchestrate network slices in order to provide the necessary 5G services in an E2E manner. This challenge is considered from two perspectives: 1. the architectural perspective and 2. the development of the framework, which are the main crux of **Publication IV**. In the latter part of this chapter, we present the design of a dedicated slicing framework which enables specific Industry 4.0 use cases that can not be facilitated using the proposed generic slicing solution as detailed in **Publication V**.

3.1 E2E network slice orchestration framework

The ETSI NFV MANO framework has paved the way for orchestrating different network services, including network slices precisely orchestrated and tailored to satisfy the needs of particular users/applications. The customization of network slices to meet the specific needs and functional requirements of particular use cases has been an essential aspect of network slicing and its evolution. These slicing needs and peculiar resource requirements can practically be fulfilled through the use of a robust slicing architecture, which extends the functions and elements of the traditional ETSI NFV MANO framework such as the one presented in Figure 3.1.

In the 5G system technology, much of the focus is on enabling new and innovative services that will provide new business opportunities by being able to meet the diverse requirements of those businesses, and to introduce an advanced ecosystem of market players with frequently changing dynamics. These 5G services are intended to have a significant impact in society, and enhance many different aspects of human lives. Also, 5G promises to bring better efficiency and system optimization to the next generation of mobile network technology. Network slicing is recognized

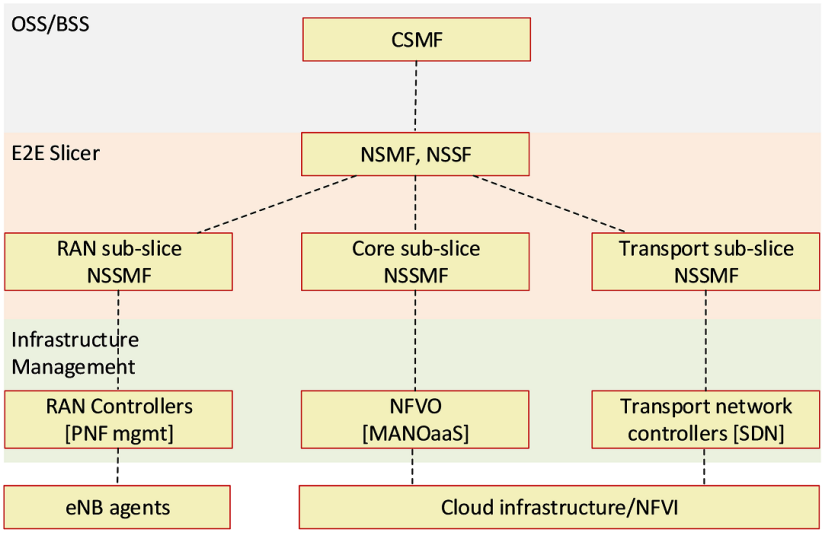


Figure 3.1. High-level view of our E2E network slicing architecture, Publication IV [66].

as the fundamental enabler of these numerous promises. Network slicing allows the parallel operation of services with distinctive functional and operational requirements, such as eHealth, automobiles, smart manufacturing, media, entertainment and so on, over a shared physical infrastructure.

The orchestration of network slices intended to facilitate any of these listed vertical 5G services has to be customized to support users' needs. This customization will run through the different segments of the mobile network involved in the network slice and reflect that segment in the form of a slice orchestration template/blueprint as captured in our E2E network slicing architecture presented in Figure 3.1. This slice customization will be capable of considering essential parameters, such as service contents per group of users, location, age group, ethnicity, mobility pattern, preferred language, size of users, and so on, to produce the QoS parameters that will be reflected in the network slice.

3.2 Definition of slice components, templates and types

In order to facilitate a homogeneous orchestration of slices tailored towards actualizing a specific service need, there is a need to meticulously

define the following concepts in a standardized fashion:

- the different components of the network slice,
- the slice templates/blueprints, and
- the different slice types or slice service types (SST).

3.2.1 Slice components

Generally, a network slice is actualized from functions running mainly on virtual resources as VNFs and sometimes PNFs. The VNFs could be instantiated either as VMs or containers. To realize a fully functional slice, these VNFs are interconnected/stitched together with optimal placement [33] and clear QoS definition based on a specific set of requirements to fulfill a particular business scenario. As a result, different business use cases may require an entirely different set of VNFs and QoS requirements. Notwithstanding, due to the flexibility and energy efficiency that are integral parts of the 5G technology, VNF sharing, and reuse will be practiced where necessary to orchestrate network slices [30]. A 5G slice will consist of virtualized functional components that enable the RAN, the Core, and transport network parts of a mobile network.

3.2.2 Slice templates

The slice template provides a clear definition of specifics, based on which a network slice is characterized. These definitions and descriptions of network slices will then be translated into detailed virtual resources and QoS tailored to meet the requirements of the network slice. Slice templates are functional blueprints that define the instantiation, configuration, operational management and placement of network slice based on the different specific slice parameters. The slice blueprints specify essential parameters such as a list of VNFs, virtualization technology type, VNF Forwarding Graph (VNFFG), VNF placement/instantiation locations, level of isolation, guaranteed security level, validity period (lifespan), etc. They also reflect the different supported slice types.

3.2.3 Slice types

The 3GPP primarily classifies network slices into the Extreme/Enhanced Mobile Broadband (x/eMBB), Ultra Reliable Low Latency Communication (uRLLC), Massive Machine Type Communication (mMTC), and Vehicle

to Everything (V2X) communication type of slice. Each class of network slice is defined by a set of specific functional and operational requirements that are generally identified by the VNF types and the necessary virtual resources (e.g., CPU, storage type, network types, etc.) used in their orchestration. It also defines other system requirements, such as system throughput, latency, permitted jitter, system availability, link quality, and so on.

3.3 5G vertical service use cases

In the preceding sections, the concept and fundamentals of network slicing and its composition have been discussed. Similarly, its orchestration and management have both been presented and explained, especially from the perspective of a high-level architecture based on compliance with the evolving standardization efforts of the 3GPP specifications. Now, having discussed the potentials of network slicing, including its numerous benefits, types, composition, orchestration and management, with a focus on the practicalities of its implementation. It is then imperative to demonstrate the feasibility of the proposed architecture in Figure 3.1 by implementing a proof of concept of it. This proof of concept was specifically implemented based on the theoretical presentation of the system that was discussed in the previous sections. As a result, by following our proposed slicing architecture, we have designed and developed a network slicing framework, called E2E Slicer, which encompasses the functionalities of the components presented in Figure 3.1.

To test the operational effectiveness of the developed E2E Slicer, we have introduced two different slicing use cases as presented in Figure 3.2. Figure 3.2 shows the framework we developed, which was implemented in compliance with the standards and taxonomy of the 3GPP [40]. Our focus was on the orchestration of two different types of slices (red and green slices) based on their requirements. One slice is orchestrated with a set of high quality (HQ) VNFs (red solid thicker line with VNFs on the blue cloud) to fulfill its E2E user need for video streaming (depicted with the Youtube logo). The other one is orchestrated with a set of low quality (LQ) VNFs (green solid line with VNFs on the green cloud) to satisfy the user requirements for smart factory service delivery (depicted with the green industrial building logo).

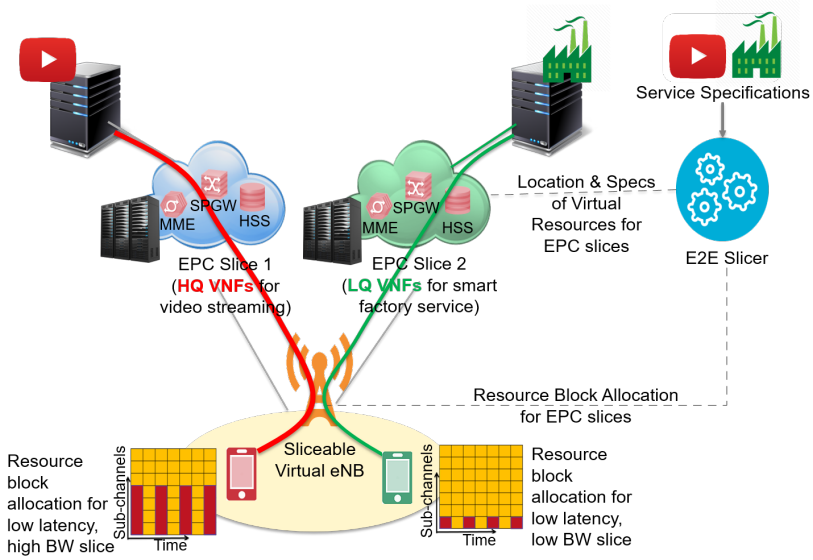


Figure 3.2. Customized slices for 5G vertical use cases enabled by the network slicing architecture, Publication IV [66].

As for the video streaming slice, which is an eMBB type of network slice, the RAN sub-slice also has a suitable amount of radio resource blocks (shown as a matrix of yellow and red squares in Figure 3.2) allocated and scheduled for low latency (depicted by the space between the red squares) and high bandwidth (depicted by the height of the red squares) access communication from the Sliceable Virtual eNB. These are carefully orchestrated to satisfy the need of the video streaming application to deliver an excellent quality of experience to the end-users.

The smart factory service slice, on the other hand, was orchestrated following its set of service requirements. Based on the needs of the smart factory application which is intended to be deployed on the slice, appropriate types of vEPC VNFs are instantiated and connected with the right amount of RAN sub-slice resources from the Sliceable Virtual eNB as depicted in Figure 3.2. This slice type requires low latency and a low bandwidth type of communication at the RAN to ensure the timely delivery of the factory service, necessitating a normal vEPC core sub-slice VNFs without any additional service enhancement.

3.4 Distinctive features of the E2E Slicer framework

In this section, we discuss the set of supporting features implemented in our framework to facilitate its effective operation, which we have dubbed

"the E2E Slicer" [66]. This set of supporting features is implemented as software modules in the framework. The software modules are designed to operate according to the 3GPP standards specifications, which enable the seamless orchestration of network slices. Each network slice is customized to fulfill the requirements of a specific 5G service type, as described in the use cases above. Specifically, the implemented distinctive features are functional modules related to the network slice management function (NSMF), Network slice selection function (NSSF), and network slice subnet management function (NSSMF), as shown in Figure 3.1. These software packages are needed to facilitate our E2E network slicing paradigm that cuts across all the mobile network technology from the RAN, Core to the Transport. The functional features are listed and discussed below.

3.4.1 RAN slicing

Our RAN slicing procedure is implemented using similar principles as the FlexRAN [68], whereby we deploy a centralized RAN resource controller in a server to apply resource allocation policies at the MAC layer of the involved eNBs. Resource allocation, management and control are effected through agents running on each eNB. These agents provide interface for the exposure of RAN-level status information and resource availability to the RAN resource controller using their north-bound protocol API. The RAN sub-slice NSSMF then enforces the resource allocation request by ensuring that adequate RAN resources are shared in the right proportion amongst all the simultaneously operating slices. The RAN resources are designed to reflect the requirements of the service-specific slice (such as a high network throughput demand for xMBB or an extremely low latency demand for uRLLC slice), often provided during the slice instantiation phase.

3.4.2 Slice-dedicated core network support

For RAN sub-slices to maintain an association with core network subslices, it is essential that a particular type of interface is implemented that allows an eNB instance to maintain connection with multiple MME instances. This interface is known as the S1-Flex interface. The S1-Flex interface is primarily defined for load balancing, whereby an eNB could connect to a pool of MMEs in an LTE network.The S1-Flex interface was

implemented for the following reasons:

- to enable an instance of an eNB to create and handle eNB-MME association dynamically with different slice-dedicated core network sub-slices, since an eNB may host multiple RAN sub-slices.
- to enable a RAN sub-slice to also load-balance between multiple instances of the same dedicated core network sub-slices. This could be a necessity for the orchestration of a highly available and reliable E2E network slice type.

3.4.3 Network slice selection function

Our E2E Slicer also incorporates and deploys a network slice selection function (NSSF), which is enabled using a set of functional software modules. The NSSF is mainly manifested upon an attach request from a UE. The eNB queries the E2E Slicer through a REST API endpoint to select the right MME of the responsible core network sub-slice to direct the request of a UE based on the identified public land mobile network (PLMN) ID. While the selection of an MME is typically the primary role of the eNB in LTE networks, the responsibility of selecting and managing an MME has now been transferred to the E2E Slicer. This way, the framework introduces more flexibility to the system by facilitating the implementation of different per-slice core network sub-slice selection policies and other optimization algorithms in the form of software plugins at the E2E Slicer. However, the potential downside to this flexibility is the additional attachment latency that could be introduced due to the additional eNB-E2E Slicer association and communication thereafter.

3.4.4 Slice instantiation and management

The slice instantiation and run-time management functionalities are provided to the CSMF through the northbound REST protocol of the E2E Slicer. This interface allows a slice owner to request the instantiation of dedicated network slices and manage them appropriately from the OSS/BSS. Using the northbound interface of the E2E Slicer, a slice owner can specify a slice type from the slice catalog and provide additional slice requirements, such as slice reliability and availability targets, the number of expected UEs connections, slice availability areas, potential traffic characteristics, and slice lifespan, etc. Based on these specifications, the NSMF

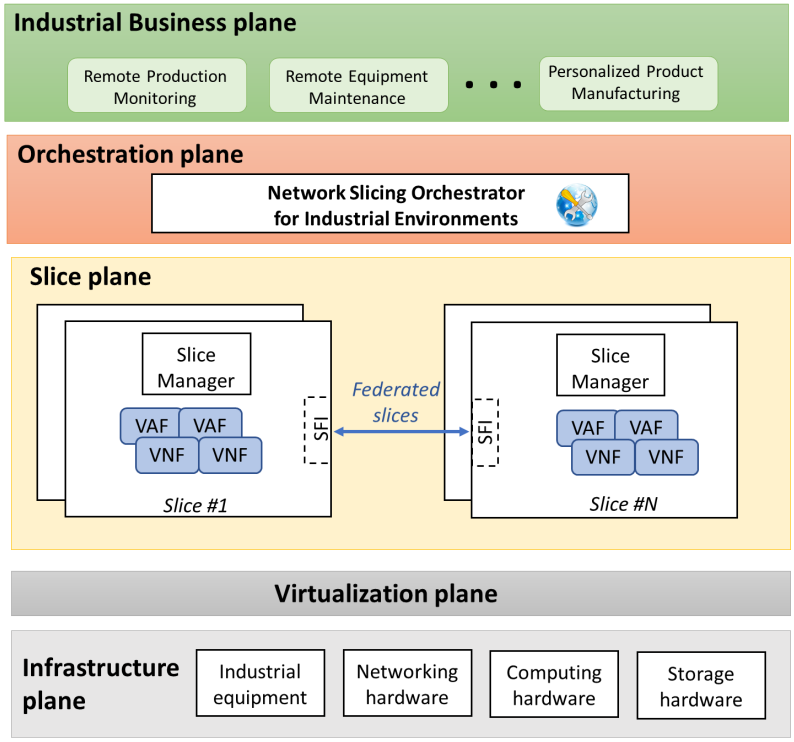


Figure 3.3. Framework architecture for the orchestration of industrial network slices, Publication V [41].

will then compose a customized slice structure and delegate the instantiation and management of each sub-slice to the appropriate NSSMF, as depicted in Figure 3.1.

3.5 Dedicated slicing architecture for enabling Industry 4.0

Network slicing is a fundamental enabler of the 5G technology [40]. It facilitates the provisioning of optimized virtual networks. Such networks are often tailored to address the communication service needs of most vertical industries. As the momentum towards enabling the Industry 4.0 industrial Internet increases, it is gradually becoming more evident that overcoming the challenges of its communication services will likely require intrinsic and dedicated architectural solutions. These solution may not be optimally provisioned over generic network slicing orchestration frameworks.

Therefore, significant research effort is required to determine a method to ameliorate the network slicing requirements for enabling some Indus-

try 4.0 use cases using a dedicated slicing architecture [41], as presented in Figure 3.3. In this section, we present the design of our dedicated slicing architecture for enabling Industry 4.0Therefore, significant research effort is required to determine a method to ameliorate the network slicing requirements for enabling some Industry 4.0 use cases using a dedicated slicing architecture [41], as presented in Figure 3.3. In this section, we present the design of our dedicated slicing architecture for enabling Industry 4.0 and the different u

and the different use cases for which it can be deployed. This slicing platform is designed to incorporate the following fundamental characteristics:

- advanced network reconfigurability to support dynamic production processes.
- seamless federation of edge cloud resources over multiple technology and administrative domains.
- enhanced radio communications to efficiently support multiple wireless industrial-based requirement.
- industrial equipment slicing for extending manageability for smart manufacturing devices while integrating re-useability of production patterns.
- control without ownership with encapsulation of the complexity of underlying industrial infrastructure.

This architecture is designed to extend the network slicing paradigm to enhance the functionalities and accommodate industrial use cases over heterogeneous domains. Most importantly, this architecture will leverage advanced software-based network programmability for 5G technology, the cloud and edge computing approach, and dynamic on-demand reconfigurability [42] processes in order to enable the much-needed flexibility in the next generation industrial networks and facilitate the inter-factory cooperation and transformation of manufacturing processes.

The framework presented in Figure 3.3, shows an advanced slicing architecture consisting of five major fundamental planes. The *Industrial Business plane*, *Orchestration plane*, *Slice plane*, *virtualization plane* and *Infrastructure plane*. Each of these planes plays a critical role in actualizing the aforementioned integrated fundamental concepts towards en-

abling efficient orchestration of network slices over heterogeneous multiple physical industrial infrastructure.

The *Industrial Business plane* provides the interactive interface for preparing and submitting the diverse requirements for the orchestration of the different industrial slices from the participating industrial stakeholders. Different solutions, such as the enterprise resource planning systems, can be integrated into the orchestration system to facilitate the product manufacturing processes of the industrial environment.

The *Orchestration plane* is intended to facilitate the orchestration of industrial network slices over heterogeneous industrial resources and domains with functionalities to support VNF and VAF instantiated close to the extreme administrative domain edge, especially for industrial types of machinery using suitable virtualization technology. Using the specified QoS slice requirements and traffic pattern, the orchestration plane is able to facilitate network reconfigurability to enhance timely changes in the industrial processes. This framework can also seamlessly interact with other essential network components to enable dynamic network traffic steering and management. With supporting functions to interact with ETSI-based NFV MANO deployments flawlessly, virtual network resources can be pre-provisioned for slices and enable coherent interoperability between the slice components.

The *Slice plane* is a logical representation of the orchestrated slices, reflecting the interactions between different VNFs and VAFs for enabling a particular industrial use case. The framework is designed to leverage both the cloud and edge computing paradigms to orchestrate slices with embedded functions, i.e., VAFs tailored to fulfill a specific need, such as reliability and real-time service procedures. Besides, industrial network slices are also integrated with slice manager functions to provide smooth in-network procedures among the elements and granular life-cycle management of an industrial slice. This way, the slices can better support E2E processes and enhance slice control. Slices orchestrated over federated industrial domains are equipped with slice federation interface (SFI) for proper interworking among slices in terms of encryption and decryption of sensitive information.

The *virtualization plane* is intended to provide the virtualization support for hosting variant network and application functions referred to as VNFs and VAFs, respectively. This layer will explore the use of different hypervisor technologies, including novel proprietary solutions in

providing the needed abstractions for virtualizing the identified physical resources and considering the additional industrial equipment with their actuation and sensing capabilities.

The *Infrastructure plane* encompasses the entire physical resources from which virtualized industrial processes can be enabled. This plane consists of the standard physical resources from data centers, such as Networking hardware, computing hardware, and storage hardware. It also accommodates the industrial heavy machinery equipment with numerous sensors and actuators on board. Since these resources are federated, they may likely belong to different administrative entities, including public and private domains.

3.5.1 Industry 4.0 use case example

The Industry 4.0 use case occupies a major consideration in the 5G network development architecture support scenario. Moreover, this use case cuts across many sectors, which include but are not limited to smart factories/manufacturing, industrial internet of things, energy, automotive, agriculture & food production, etc. It is noteworthy that most of these identified Industry 4.0 sectors can not be optimally supported using a generic 5G network slicing architecture due to their special requirement in terms of access to features and functionalities that are only available on on-site integrated industrial equipment. Most of these industrial equipment are often controlled via sensors and actuators, which are not readily virtualized infrastructures.

Hence, the need arises for a dedicated architecture as presented in Figure 3.3. So, for example, this dedicated architecture can be deployed to enable an optimal slicing scenario for any of the above-mentioned Industry 4.0 sectors. Take for instance, in the energy sector, Company A that owns a crude oil refinery can take order from a partner company B, where company B wanted to remotely monitor the refining of its crude oil. This process is called remote production monitoring, which is based on certain agreed stipulated SLA and established level of trust. Such a request can be submitted via the *Industrial Business plane*. The request will define different slice requirement profiles which will also include different security levels as well as access to certain industrial equipment located at the refinery.

The *Orchestration plane* will compose the requested slice based on its specified profile and equip it will all the necessary VAF and VNFs includ-

ing different APIs that are needed to effectively communicate and interact with the on-site large industrial equipment (including temporary storage tanks, distillation chambers, controls for closing & opening of valves, etc.) and execute its tasks optimally. This slice composition will be visible through the *Slice plane* once it has been orchestrated and configured for company B to use it. At this point, all the needed credentials and right of access will then be granted to company B to remotely monitor or control the refining of its crude oil to meet its own business need. Such business need might be for the purpose of ensuring the high standard of the refined end product, i.e., quality assurance. It should be noted that all the operations of the slices will be guided by the terms and conditions specified in the agreed SLA, which will also include necessary penalties for the violation of such requirements.

3.6 Summary

In this chapter, we have highlighted our contributions towards the extension of the ETSI MANO framework to enable the orchestration of E2E network slices, with the addition of essential functional blocks. We also introduced and defined slicing terminologies such as, slice components, template and types. Based on our introduced architecture in Figure 3.1, we also presented an implementation of an E2E Slicer, which can be used to orchestrate an E2E network slice that is customized with the requirements of 5G and beyond services, while highlighting its distinctive features. Using this approach, we also introduced a number of 5G vertical use cases and discussed how they can be realized based on our implementation architecture shown in Figure 3.2, whose details are presented in **Publication IV**. Finally, we proposed a dedicated slicing architecture to address certain identified needs of Industry 4.0 use case, which may not be efficiently addressed by a generic slicing framework. The details of this contribution are available in **Publication V**. In the following chapter, we will discuss the concept of dynamic resource provisioning for network slicing orchestration system as detailed in **Publication II**.

4. Dynamic Resource Provisioning for Network Slicing Orchestration System

The primary purpose of this chapter is to highlight and present our efforts to extend the operational functionalities of a network slice orchestration system with some level of intelligence and operational efficiency through the proposition of a smart algorithm. This intelligent algorithm allows the system to scale while automatically adhering to a certain orchestration time threshold. The contributions presented and discussed in this chapter are based on the outcome of **Publication II**.

4.1 Dynamically scalable network slicing orchestration system

Scalability and efficient resource utilization is often a common challenge in most service orchestration framework [43]. Developing a flexible system that can dynamically scale based on the size of slice request and service requirements can not be overemphasized. The primary purpose of this work is to develop a robust system for the orchestration of network slices that maximizes all of its available resources and dynamically scales to satisfy slice orchestration requests within a predefined maximum orchestration time.

This research has led to the design of a high-level scalable and hierarchical architecture for the orchestration of network slices. The architecture has a dedicated global orchestrator (GO) and multiple domain-specific orchestrators (DSO). It also has supporting functional blocks, i.e, resource awareness engine (RAE) and system awareness engine (SAE) for the timely monitoring and control of resources available and consumed respectively by the system, as shown in the high-level queuing model depicted in Figure 4.1. Each DSO is assumed to have control over a complete mobile network solution and is capable of orchestrating mobile network slices.

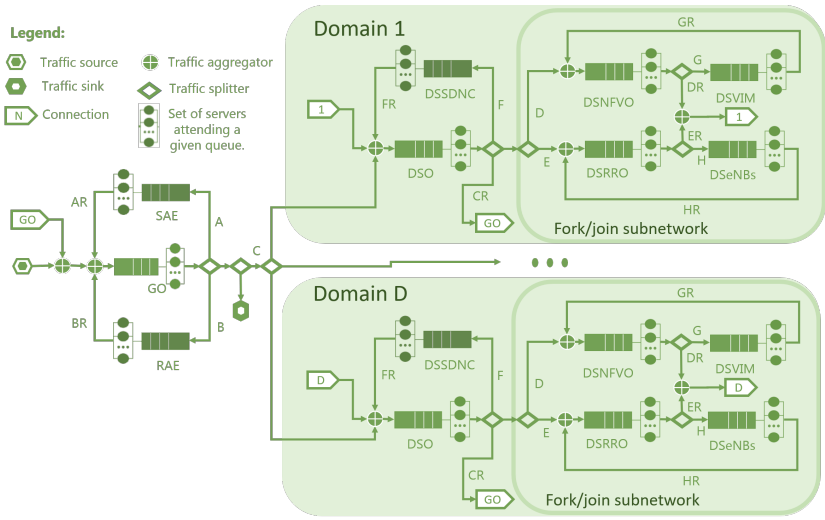


Figure 4.1. Queuing model of the Network Slicing Orchestration System, Publication II [71].

As presented in Figure 4.1, the components operating under the control of each DSO are:

- domain-specific network function virtualization orchestrator (DSNFVO) - the NFVO element for each domain that is responsible for orchestrating different network services.
- domain-specific radio resource orchestrator (DSRRO) - the network element that is responsible for orchestrating the radio access network resources. It controls and configures the eNB resources running in each domain.
- domains-specific virtualized infrastructure manager (DSVIM) - as the name implies, it manages the virtualized resources running in each domain.
- domain-specific eNBs (DSeNB) - these are the list of eNBs that are available and registered for the orchestration of network slices in each domain.
- domain-specific SDN controller (DSSDNC) - the SDN controller that is responsible for interconnecting the different slice subnetwork components.

Based on the queuing model depicted in this Figure, we assume that there are a number of Domains 1 to D, each with its own DSO component managing its own domain-specific elements (DSNFVO, DSRRO, DSVIM, DSeNB and DSSDNC) with access to a list of server resources.

This system considers all the needed system resources, both used and available, that could be activated for satisfying slice orchestration requests in a network slicing orchestration system (NSOS). This framework takes advantage of our novel dynamic auto-scaling algorithm (DASA), which automatically reacts to changes in the workload of the system while satisfying network slice orchestration (NSO) requests. Using our proposed algorithm, we can keep the slice orchestration response and processing time within a set target. This system is developed based on queuing theory, and uses the queuing network analyzer method to estimate a reasonable mean response time for the orchestration of VNFs.

4.2 System model and problem formulation

This section presents a high-level representation of the queuing system model of the NSOS shown in Figure 4.1, and formulates a problem statement in order to provide a comprehensive overview and understanding of the operations of the system components. We have modelled each component of the system as a set of entities $E = \{ GO, SAE, RAE, DSO_d, DSNFVO_d, DSVIM_d, DSSDNC_d, DSRRO_d, DSeNBs_d \}$, which interact with each other. Each entity $e \in E$ could have multiple instances deployed on an individual Virtural Machine (VM), which may belong to different domains denoted with subindex $d \in [1, 2, .., D]$ which is associated with a specific geographical location. Normally, an instance i of a given entity $e \in E$ will execute a task, which will consume virtual resources (e.g., CPU, RAM, network bandwidth), to process incoming packets initiated by the slice orchestration request (SOR). For simplification in our model, we assume the CPU is the resource that has to be optimally provisioned for each of the entities in the NSOS.

An essential objective of this work is to propose a dynamic resource provisioning (DRP) solution for our NSOS, which will enable the system to scale automatically depending on the current and predicted future workload, in order to cope with a set of performance requirements. For simplicity, we have considered a critical requirement in the scope of this work, which is the mean response time of the NSOS T needed to serve a SOR that has to be kept under a threshold T_{max}. The details of these notations, i.e., T, T_{max}, etc, is presented in Table 4.1 herein after. Our DRP solution is made up of modules from which the resource dimensioning module plays an integral role. The resource dimensioning module determines how

much resources to provision for a given workload [44]. The resource dimensioning problem for the NSOS has been formulated in the following way:

$$minimize\left(F(m)=\sum_{e\in E}\sum_{i}m_i^{(e)}\right) \tag{4.1a}$$

Constraints :

$$C1: \quad T(m)\leq T_{max}, \tag{4.1b}$$

$$C2: \quad m_i^{(e)}\leq m_{max}^{(e)} \quad \forall \quad k\in[1,K]\cap\mathbb{N} \tag{4.1c}$$

We have considered the number of physical CPU cores $m_i^{(e)}$ allocated to each instance i of a given entity $e \in E$ as the decision variable. Hence, the main purpose of objective (4.1a) is to minimize the amount of computational resources (expressed in number of CPU cores) to be allocated to the NSOS, which in turn will minimize the energy consumed by the system, and reduce its OPEX. Constraint (4.1b) guarantees that the actual mean response time of the NSOS to serve a SOR is below a mean delay threshold T_{max}. Constraint (4.1c) limits the maximum number of physical cores allocated to a single entity instance. As is the case in the virtual environment, each physical machine has a maximum number of physical CPU cores, and they are shared among several VMs. Constraint (4.1c) facilitates the bin packing problem in the context of network embedding.

The resource dimensioning problem earlier formulated can be solved using a brute force algorithm, which is an exhaustive search method. Using the brute force algorithm, we could determine the initial processing CPU instances allocated to each NSOS entity by means of the system stability condition. The system stability condition $m_e = \lceil \lambda_e \oslash \mu_e \rceil$, where m_e, λ_e, and μ_e are vectors containing the number of processing instances allocated, the aggregated arrival rate, and the packet processing rate per processing instance for each entity respectively. So, assuming there are N_e different entities in the system, where $M_0 = \sum_{i=1}^{N_e} m_e(i)$ processing instances are allocated in the initial assignment to fulfill the stability condition. Hence, the brute force algorithm will iterate until the mean response time of the NSOS is below the maximum delay threshold, i.e., $T < T_{max}$. At each iteration, the total number of processing CPU instances M allocated to the system is incremented by one. After this step, the algorithm will check every combination to allocate the $M - M_0$ processing instances among the entities of the NSOS. The result of the check will enable the

algorithm to choose the allocation that achieves the minimum mean response time of the system. Despite the simplicity of the exhaustive search method, it requires

$$N_{checks} = \sum_{M=M_0}^{M^*} \cdot N_e M - M_0 = \sum_{M=M_0}^{M^*} \frac{(N_e + M - M_0 - 1)!}{(M - M_0)!(N_e - 1)!}$$

evaluations of the mean response time of the NSOS, which is impractical.

4.3 Estimation of the mean response time of the NSOS using queuing theory

Our model uses an open network of G/G/m queues to estimate the performance of the NSOS system. The G/G/m queue in Kendall's notation represents a queuing node with m servers, arbitrary arrival and service processes, FCFS (First-Come, First-Served) discipline, and infinite capacity and calling population. Each queue stands for an instance i of the entity of the set $E \in \{\ GO,\ SAE,\ RAE,\ DSO_d,\ DSNFVO_d,\ DSVIM_d,\ DSSDNC_d,\ DSRRO_d,\ DSeNBs_d\ \}$, where the subindex $d \in [1, 2, .., D]$ is included to specify the domain, which may be associated with a specific geographical region the entity belongs to. It should also be noted that the three solid horizontal dots in Figure 4.1 stand for domains with indexes from 2 to $D - 1$. In modelling the system, we have imagined a 3GPP-like service-based architecture where each components of the 5G core is considered an instantiated service, which can be horizontally scaled out based on the system load.

As described in the beginning of the chapter, the GO orchestrates the slices of different domains, using dedicated DSO, DSNFVO, DSVIM, DSSDNC, DSRRO, and DSeNBs entities, which are in charge of orchestrating and managing the slices of a given geographical region. We have modelled the system such that entities from different domains are treated separately from one another. In a queue, there may be multiple $m_i^{(e)}$ servers that represent processing instances allocated to the corresponding instance i of the entity $e \in E$. These processing instances (i.e., virtual CPUs running on physical CPU cores) process one or several messages to serve every incoming SOR. It is assumed that each entity has a sentry that evaluates and distributes the workload (messages) among its in-

Table 4.1. Model parameters, Publication II [71].

Notation	Description
K	number of G/G/m queues
P	steady state transition probability matrix
k, i	network nodes indexes
p_{ki}	probability of a packet leaving a node k to node i
p_{0k}	probability that a packet leaves the network
λ_{0k}	mean arrival rate of the external arrival process at queue k
c_{0k}^2	squared coefficient of variation of the external arrival process at queue k
μ_k	mean service rate of each server at queue k
c_{sk}^2	squared coefficient of variation of service processes at queue k
c_{ak}^2	squared coefficient of variation of the aggregated arrival process at queue k
a_k, b_{ik}	Coefficients of the set of linear equations to estimate the SCVs of the aggregated arrival process at each queue k.
ω_k, x_i, γ_k	Auxiliary variables when a_k and b_{ik} are computed.
q_{0k}	the proportion of arrivals to node k from its external arrival process
q_{ik}	the proportion of arrivals to node k from node i
m_i	number of servers of node i
c_{si}^2	squared coefficient of variation of service processes at queue i
ρ_k	the utilization of the node k
T_k	Mean system response time of node k
W_k	Mean waiting time of node k
β	The Kraemer and Langebach-Belz approximation
$W_k^{M/M/m}$	the mean waiting time for an M/M/m queue
$C(m, \rho)$	the Erlang's C formula
T	the overall mean response time
T_{max}	time threshold for which T has to be kept
V_k	the average number of visits to node Q_k

stances according to their processing capacity.

4.3.1 Response time estimation

The mean response time of the system is the most critical performance metric considered in this work, and it is also the essential factor that determines the auto-scaling process of the NSOS. This means that the computational resources of the system will be scaled up or down in order to guarantee a given mean response time to serve a SOR depending on the projected peak in the workload of the system based on the past system´s events. Therefore, the mean response time of the system can be derived based on the aforementioned proposed queuing model.

In order to estimate the mean response time of the system, we denote the number of instances for each entity e of the set E by K_e . Then, we have $K = \sum_{e \in E} K_e$ instances in the system and, thus, K queues in the queuing model. To simplify the notation in this analysis, we map each entity instance to an integer index $k \in [1, K]$. To estimate the mean response time of the system, the following input parameters are required: i) the steady state transition probabilities matrix $P = [p_{ki}]$, where p_{ki} denotes the probability of a packet leaving the node k to the next node i or leaves the network with probability $p_{0k} = 1 - \sum_i p_{ki}$; ii) the mean and squared coefficient of variation (SCV) of the external arrival processes at queue k, λ_{0k} and c_{0k}^2; and iii) the mean and SCV of the service processes at queue k, μ_k and c_{sk}^2.

Internal flows parameters estimation

The mean arrival rate to each queue k, λ_k, can be computed by solving the flow balance equations:

$$\lambda_k = \lambda_{0k} + \sum_{i=1}^{K} \lambda_i \cdot p_{ik} \tag{4.2}$$

The QNA method helps to estimate the SCV of the aggregated arrival process to each queue c_{ak}^2 from a set of linear equations. In this way, we are only required to monitor the external arrival processes to the system, which is the arrivals of new SORs to the GO in our case, to estimate the second order moment of the arrival process at each entity instance. Specifically, we can estimate the SCV of the aggregated arrival process to each entity instance by solving the following set of linear equations:

$$c_{ak}^2 = a_k + \sum_{i=1}^{K} c_{ai}^2 b_{ik}, \qquad 1 \leq k \leq K \tag{4.3}$$

$$a_k = 1 + \omega_k \Bigg\{ (q_{0k} c_{0k}^2 - 1) + \sum_{i=1}^{K} q_{ik}[(1 - p_{ik}) + p_{ik}\rho_i^2 x_i] \Bigg\} \tag{4.4}$$

$$b_{ik} = \omega_k q_{ik} p_{ik} (1 - \rho_i^2) \tag{4.5}$$

$$x_i = 1 + m_i^{-0.5}(max\{c_{si}^2, 0.2\} - 1) \tag{4.6}$$

$$\omega_k = \left(1 + 4(1 - \rho_k)^2(\gamma_k - 1)\right)^{-1} \tag{4.7}$$

$$\gamma_k = \left(\sum_{i=0}^{K} q_{ik}^2 \right)^{-1} \tag{4.8}$$

In order to simplify the computation of the c_{ak}^2, the QNA method employs approximations. Specifically, it uses a convex combination of the asymptotic value of the SCV $(c_{ak}^2)_A$ and the SCV of an exponential distribution ($c_{exp}^2 = 1$), i.e., $c_{ak}^2 = \alpha_k (c_{ak}^2)_A + (1 - \alpha_k)$. The asymptotic value can be found as $(c_{ak}^2)_A = \sum_{i=1}^{K} q_{ik} c_{ik}^2$, where q_{ik} is the proportion of arrivals to Q_k that came from Q_i. That is, $q_{ik} = \frac{\lambda_i \nu_i p_{ik}}{\lambda_k}$. And α_k is a function of the queuing node utilization $\rho_k = \frac{\lambda_k}{\mu_k m_k}$ and the arrival rates. This approximation yields the above set of linear equations, which may be solved to get $c_{ak}^2, \ \forall \ \{k \in \mathbb{N} | 1 \leq k \leq K\}$. Last, in the above equations $q_{0k} = \lambda_{0k}/\lambda_k$.

Mean response time computation per node

Once the λ_k and c_{ak}^2 for the aggregated arrival process to each node k are estimated, we can compute the mean queuing waiting time for each node k, W_k.

If the node k has only one server (or one processing instance allocated, $m_i^{(e)} = m_k = 1$), W_k can be estimated as:

$$W_k = \frac{\rho_k \cdot (c_{ak}^2 + c_{sk}^2) \cdot \beta}{2 \cdot \mu_k (1 - \rho_k)} \tag{4.9}$$

with

$$\beta = \begin{cases} exp(-\frac{2 \cdot (1 - \rho_k) \cdot (1 - c_{ak}^2)^2}{3 \cdot \rho_k \cdot (c_{ak}^2 + c_{sk}^2)}) & c_{ak}^2 < 1 \\ \beta = 1 & c_{ak}^2 \geq 1 \end{cases} \tag{4.10}$$

If, by contrast, the node k is a GI/G/m queue ($m_i^{(e)} = m_k = m$), W_k can

be estimated as:

$$W_k = 0.5 \cdot \left(c_{ai}^2 + c_{si}^2\right) \cdot W_k^{M/M/m} \tag{4.11}$$

where $W_k^{M/M/m}$ is the mean waiting time for a M/M/m queue, which can be computed as:

$$W_k^{M/M/m} = \frac{C(m_k, \frac{\lambda_k}{\mu_k})}{m_k \mu_k - \lambda_k} \tag{4.12}$$

and $C(m, \rho)$ represents Erlang's C formula, which has the following expression:

$$C(m, \rho) = \frac{\left(\frac{(m \cdot \rho)^m}{m!}\right) \cdot \left(\frac{1}{1-\rho}\right)}{\sum_{k=0}^{m-1} \frac{(m \cdot \rho)^k}{k!} + \left(\frac{(m \cdot \rho)^m}{m!}\right) \cdot \left(\frac{1}{1-\rho}\right)} \tag{4.13}$$

Global Response Time Computation

Let V_k denote the visit ratio for the node k (Q_k), which is defined as the average number of visits to node Q_k by a packet during its lifetime in the network. That is, $V_k = \lambda_k / (\sum_{k=1}^{K} \lambda_{0k})$. And let $\mathcal{K}_e \subset \{1, ..., K\}$ with $|\mathcal{K}_e| = K_e$ be the subset of indexes associated with the instances of the entity e. Then, we can compute the mean response time for each entity, T_e, as:

$$T_e = \sum_{k \in \mathcal{K}_e} (W_k + \frac{1}{\mu_k}) \cdot V_k \tag{4.14}$$

As stated earlier, the $DSNFVO_d$, $DSVIM_d$, $DSRRO_d$, and $DSeNBs_d$ entities of the domain d are modeled as a fork/join subnetwork of queues with two parallel branches. The mean response time of the fork/join subnetwork of the domain d, $T_d^{(FJS)}$ will be given by:

$$\begin{aligned} T_d^{(FJS)} = max(&T_{DSNFVO_d} + T_{DSVIM_d}, \\ &T_{DSRRO_d} + T_{DSeNBs_d}) \end{aligned} \tag{4.15}$$

Finally, the overall mean response time T can be estimated as:

$$T = T_{GO} + T_{SAE} + T_{RAE} + \sum_{d=1}^{D} \left(T_{DSO_d} + T_d^{(FJS)}\right) \tag{4.16}$$

4.3.2 Transtion Probabilities for the Slices Orchestration System

As stated earlier, the steady-state transition probabilities are input parameters which are needed to estimate the mean response time of the system. These can be derived directly from the system operation.

Let V_e denote the visit ratio of the entity $e \in E$ which is defined as the average number of visits to entity e by a SOR during its lifetime in the NSOS.

That is,

$$V_e = \lambda_e / \sum_e \lambda_{0e} = \lambda_e / (\lambda_{0GO}) \tag{4.17}$$

The visit ratios and the transition probabilities are related through (4.2) (flow balance equations):

$$V_e = \frac{\lambda_{0e}}{\sum_{e \in E} \lambda_{0e}} + \sum_{es \in E} V_{es} \cdot p_e^{es} \tag{4.18}$$

where p_e^{es} denotes the transition probability from entity $es \in E$ to entity $e \in E$. The visit ratios of the GO, RAE, and SAE entities are given by the number of messages they have to process for every slice orchestration request. Then, $V_{GO} = 3$, $V_{SAE} = 1$, and $V_{RAE} = 0$. Let α_d denote the percentage of the total incoming slices orchestration requests to the system which are addressed to the domain d. Then, the visit ratios of the entities DSO_d, $DSSDNC_d$ $DSNFVO_d$, $DSVIM_d$, $DSRRO_d$, and $DSeNBs_d$ will be α_d times the number of messages these entities have to process for every incoming slice orchestration request, i.e., $V_{DSO_d} = alpha_d \cdot 3$, $V_{DSNFVO_d} = alpha_d \cdot 2$, $V_{DSVIM_d} = alpha_d$, $V_{DSRRO_d} = alpha_d \cdot 2$, $V_{DSeNBs_d} = alpha_d$, and $V_{DSSDNC} = \alpha_d$.

Additionally, the sum of the transition probabilities for a given entity e are normalized to unity:

$$p_{0e} + \sum_{ed \in E} p_{ed}^e = 1 \tag{4.19}$$

In our case, we can compute the transition probabilities between entities for the slices orchestration system model using (4.18) and (4.19). Specifically, we get

$$p_{SAE}^{GO} = \frac{1}{3}; \quad p_{RAE}^{GO} = 0; \quad p_{DSO_d}^{GO} = \alpha_d \cdot \frac{1}{3}; \tag{4.20}$$

$$p_{GO}^{SAE} = 1; \quad p_{GO}^{RAE} = 1; \tag{4.21}$$

$$p_{DSNFVO_d}^{DSO_d} = \frac{1}{3}; \quad p_{DSSDNC_d}^{DSO_d} = \frac{1}{3}; \quad p_{GO}^{DSO_d} = \frac{1}{3}; \tag{4.22}$$

$$p_{DSO_d}^{DSNFVO_d} = 0.5; \quad p_{DSVIM_d}^{DSNFVO_d} = 0.5; \tag{4.23}$$

$$p_{DSNFVO_d}^{DSVIM_d} = 1; \quad p_{DSO_d}^{DSSDNC_d} = 1; \tag{4.24}$$

Finally, assuming that the workload is distributed among the instances of a given entity e according to its processing capacity, the transition probability $p_{e2_j}^{e1_i}$ from the instances i of the entity $e1 \in E$ to the instance j of the entity $e2 \in E$ can be simply computed as

$$p_{e2_j}^{e1_i} = \frac{m_j^{(e2)}}{\sum_l m_l^{(e2)}} \cdot p_{e2}^{e1} \tag{4.25}$$

4.4 Dynamic auto-scaling algorithm using QNA

DASA is a smart technique integrated into our NSOS to enable it to scale autonomously based on the current system-wide workload and the running server nodes. DASA employs the QNA technique for the dimensioning of the total available computational, network, and memory and storage virtual resources across all of the running servers. DASA also considers the projected near-future workload to keep the mean response time T of the system below a target time threshold T_{max}. In order for the system to maintain a target time threshold T_{max} while orchestrating network slices, the solution relies on a few essential building blocks. such as the *workload predictor*, *Dimensioning algorithm*, *Reactive provisioning*, *Scaling of the network service* and *Request policing* to facilitate a smooth operation of the system autonomously. All of these building blocks work in harmony and play a significant role in realizing this novel solution, as depicted in Figure 4.2.

4.4.1 Workload predictor

This component is responsible for predicting the next peak period in the orchestration demand for the NSOS based on the input received from the *workload monitoring* agent. These inputs consist of statistical data regarding the peak traffic workload that arrives at the NSOS *e.g.*, mean arrival rate $\lambda_{0GO,prev}$ and squared coefficient of variation (SCV) of the packet inter-arrival times $c^2_{0GO,prev}$ during the previous Δt time. Based on these inputs, it then produces as output, the estimated values of the mean arrival rate λ_{0GO} and the SCV of the packet inter-arrival times c^2_{0GO} of the peak traffic demand for the next period of Δt time. This component can also be developed using machine learning algorithms.

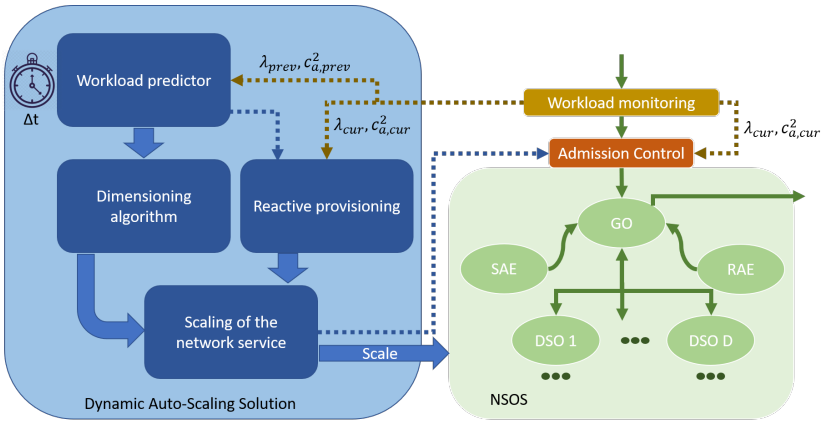


Figure 4.2. Dynamic Auto-Scaling Solution for the NSOS, Publication II [71].

4.4.2 Dimensioning algorithm

This component is another fundamental unit of the framework responsible for sizing the virtualization resources of the NSOS with a particular focus on the virtual CPUs. It sizes these resources to ensure that there is an optimal amount of computational resources for both the mean and SCV of the external slice orchestration arrival request at the global orchestrator λ_{0GO} and c^2_{0GO} respectively so that $T \leq T_{max}$ always during the next orchestration round of length Δt. In this way, the amount of computational resources consumed by the NSOS is minimized, thereby also minimizing energy consumption [45] and operational costs.

In order to achieve these objectives, we employed the heuristic method described in Algorithm 1. The method takes as input the target mean response time T_{max}, the maximum number of available physical CPU cores M_{max}, the mean and SCV of the external arrival process provided by the predictor (i.e., λ_{0GO} and c^2_{0GO}), and the mean service rate μ_e and the SCV of the service times c^2_{es} per entities in the set E.

The main objective of Algorithm 1 is to continuously find the minimum number of processing instances to be allocated to the network so that $T < T_{max}$ by iterating until either $T \leq T_{max}$ or $M \geq M_{max}$. At each round of iteration , it searches for the entity $e^* \in E$ that best contributes to reducing T when an additional processing instance is allocated to such entity. After identifying such entity e^*, the algorithm will then trigger the allocation of an additional processing instance to it. This way, the system ensures that $T < T_{max}$ at every orchestration cycle.

Algorithm 1 Dimensioning Algorithm

Input:

T_{max}: Mean response time.
M_{max}: Maximum number of available physical CPU cores.
λ_{0GO}: Mean of the external arrival process at the GO.
c^2_{0GO}: Mean SCV of the external arrival process at the GO.
μ_e: Mean service rate per entity.
c^2_{es}: Mean SCV per entity of the service times.

Output:

Required number of instances (or virtualization containers) I per entity and the processing instances to be allocated to each entity m.

1: **Initialization**

Compute λ_e using eq (4.17);
$m \leftarrow \lceil \lambda_e \oslash \mu_e \rceil$;
$M_0 \leftarrow \sum_{e \in E} m(e)$;
$M \leftarrow min\{M_0,\, M_{max}\}$;
$K \leftarrow \lceil m \oslash m_{max} \rceil$;
Compose the network of queues and compute P using eq (4.20)-(4.25);
Compute internal flows parameters λ and ${c_a}^2$ using eq (4.2)-(4.8);
T estimation given the initial stability conditions using eq (4.9)-(4.16);

2: **while** $T > T_{max}$ or $M < M_{max}$ **do**
3: $m_{aux} \leftarrow m + 1_{N \times 1}$;
4: $K_{aux} \leftarrow \lceil m \oslash m_{max} \rceil$;
5: Recompose the network of queues for m_{aux} and I_{aux}; and compute P using eq (4.20)-(4.25);
6: Recompute λ and ${c_a}^2$ using eq (4.2)-(4.8);
7: Estimate the queuing waiting times vector W_{aux} using eq (4.9)-(4.13) and considering the above input parameters;
8: $e^* \leftarrow {}_{k \in [1,K] \cap \mathbb{N}}(W(e) - W_{prev}(e))$;
9: $m(e^*) \leftarrow m(e^*) + 1$;
10: $K(e^*) \leftarrow \left\lceil m(e^*)/m^{(e)}_{max} \right\rceil$;
11: $T \leftarrow T - W_{prev}(e^*) + W(e^*)$;
12: $M \leftarrow M + 1$;
13: **end while**

4.4.3 Scaling of the network service

This component plays a crucial role in the actual allocation and release of network processing resources. It takes input as the required processing instance per entity m provided by either the dimensioning algorithm or the reactive provisioning module. It continually monitors the number of resources allocated to the NSOS. This way, it can quickly determine the usable size, or the size of the reservable resource, based on inputs from both the dimensioning algorithm or reactive provisioning modules.

4.4.4 Reactive provisioning

As the name suggests, this module periodically receives a statistical report of the traffic demand based on λ_{cur} and $c^2_{a,cur}$ as measured by the workload monitoring agent. Based on this report, the reactive provisioning unit can issue an asynchronous resource scaling request when an unexpected surge in the workload is detected.

4.4.5 Request policing

This unit is an integral part of the admission control agent, enabling the system to accept or decline excess SORs when the system is working at full capacity or during a temporary system overload before the trigger of the reactive provisioning is fully effected. This module uses essential information provided by the workload monitoring agent based on the current traffic demand λ_{cur} and $c^2_{a,cur}$ retrieved from the SORs register.

4.4.6 Performance evaluation

To evaluate the performance of our proposed algorithm based on the developed queueing model, we utilized a queueing simulator for NSOS using the MATLAB Simulink environment. The configuration of the main parameters is presented in Table 4.2. The workload predictor uses 10 neurons as the default value in the hidden layer of the neural network. The RMSE value is obtained experimentally via simulation. The service rate of each processing instance and SCV are $10000\ packets/second$ and 0.65 respectively, and are derived from real measurements of service times for a VNF that has a similar operation to the entities of the slice orchestration system. Our general assumption is that the overall system workload is equally distributed among the different domain-specific

Table 4.2. Parameters Configuration, Publication II [71].

External arrival process at the GO & Workload Predictor	
External arrival process at the GO	Modulated Poisson process whose arrival rate versus time is given in Figure 4.3
Workload predictor	Focused time delay neural network with 10 neurons
Root mean squared error (RMSE) of the workload predictor	361.5 SORs per second
Service processes	
Service rate of each processing instance	10000 packets per second
SCV of the processing instance service time	0.65
QoS requirements	
$\overline{T}_{max}$	$2\ ms$

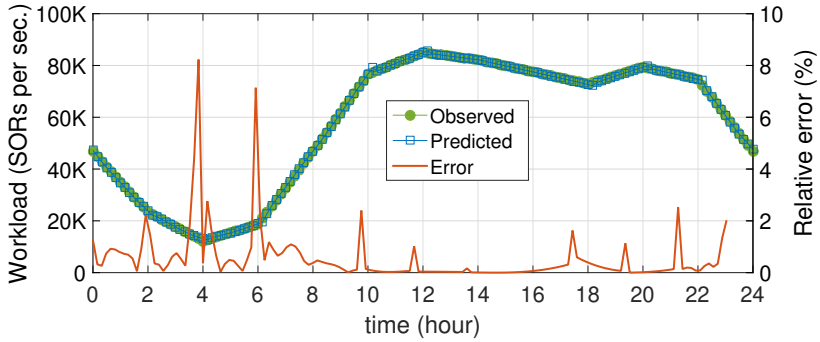


Figure 4.3. The temporal distribution of the workload profile, Publication II [71].

orchestrators (DSO). The instance of each entity is simulated as a first come first serve [46] (FCFS) queue with one or several generic servers, with the same service time distribution assumed for each of them. The mean and SCV for the service time for each instance is $100\ \mu s$ (equivalent to a service rate equal to $10000\ packets/second$) and 0.65 respectively.

In our simulation, we assumed that every SOR arrives at the global orchestrator (GO) according to a Poisson distribution. We modulated the aggregated arrival rate also according to the temporal distribution presented in [69] as explicitly depicted in Figure 4.3 for the considered workload profile.

Moreover, the implementation of our DRP solution operates based on the request of our workload predictor for which we have chosen the Levenberg-Marquardt [47] as a training algorithm. The relative prediction error of

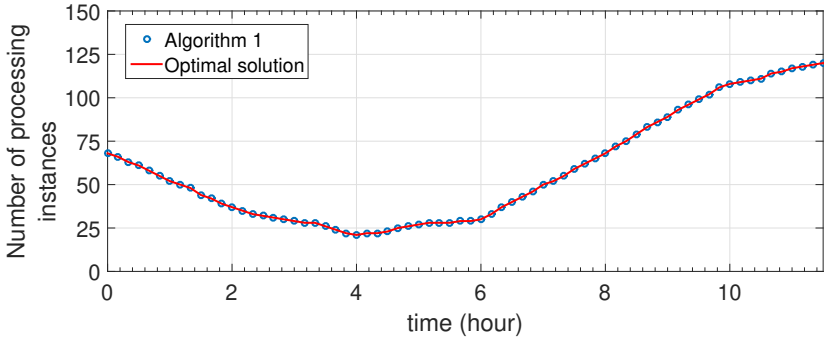


Figure 4.4. Comparison of the required computational resources estimated by our proposed dimensioning algorithm and the optimal solution, Publication II [71].

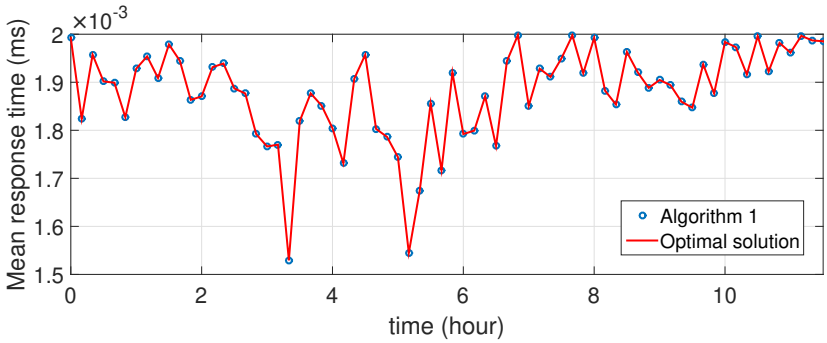


Figure 4.5. Comparison of the overall mean response time of the system achieved by our algorithm and the optimal solution, Publication II [71]

the workload predictor module is shown in Figure 4.3. The system provisions periodically every ten minutes. As the request policy agent, we used a single token bucket at the GO, whose generation rate depends largely on the resources allocated to the NSOS. Similarly, the system triggers the reactive provisioning module when $(\lambda_{0GO} - \lambda_{0GO,pred})/\lambda_{0GO} \geq 0.05$ or $(\lambda_{0GO} - \lambda_{0GO,pred})/\lambda_{0GO} \leq -0.5$. We set the VM instantiation time to 82 seconds, as presented in [70]. As mentioned above, the target mean response time of the system to serve a SOR was set to $T_{max} = 2\,ms$.

We evaluated the system based on the presented parameters. Because finding the optimal solution is usually computationally intensive, we only considered one DSO entity for the first twelve hours of the day. Figure 4.4 shows the result of the comparison between our algorithm and the optimal solution to estimate the number of computational resources. A comparison between our algorithm and the optimal solution to estimate the achieved mean response can be made from Figure 4.5. From this figure, it is clear that our algorithm achieves the optimal solution in all the considered test cases.

4.5 Summary

This chapter introduced and discussed our contribution to the concept of dynamically scalable network slicing orchestration system. To better understand our approach, we introduced a system model and formulated a resource dimensioning problem around it. Basically, the model must ensure that the system fulfills a critical requirement, which is the mean response time for which a SOR must be processed. Using queuing theory, we estimated the mean response time that the system must respect, and introduced a dynamic auto-scaling algorithm to ensure that the system adheres strictly to this crucial requirement using QNA. The details of this contribution can be found in **Publication II**. The next chapter introduces the reader to an entirely different concept in network slicing, which is based on **Publications I and III**. The concept of multi-domain slice orchestration over federated resources running on multi-site infrastructure is the crux of discussion in the following chapter.

5. Multi-domain Slice Orchestration over Federated Resources

This chapter essentially introduces the reader to a new paradigm in network slicing, which is the aspect of orchestration of network slices across federated multi-administrative resources, which is the main contribution of **Publication III**. Basically, our aim is to exchange our views and understanding of this concept and propose relevant suggestions that could have practical contributions on the evolution of the standard architecture propositions. Through our research findings in **Publication III**, we have managed to contribute and make relevant propositions to this important slicing topic. Similarly, **Publication I** contributes a more practical approach towards the implementation of a framework that takes into consideration our contributions in **Publication III**.

5.1 Composition of a multi-administrative domain network slice

The ETSI NFV architecture is primarily designed for the orchestration of network services, considering only a single domain. However, this architecture has evolved into a solution for the orchestration of mobile network slices after integrating slice-specific components to extend its scope and functional supports [48]. Nonetheless, many mobile network operators and network infrastructure owners need a sophisticated network slicing architecture to support network slices orchestration across multiple administrative domains while being agnostic [49] of underpinning virtualization and infrastructure technologies. Such architecture will evolve the traditional network slice management and orchestration architecture through standard models and constructs to compose a fully-fledged network slice as depicted in Figure 5.1 over federated resources.

To this end, we have proposed the futuristic multi-domain network slicing architecture presented in Figure 5.2. This is an extension of the tra-

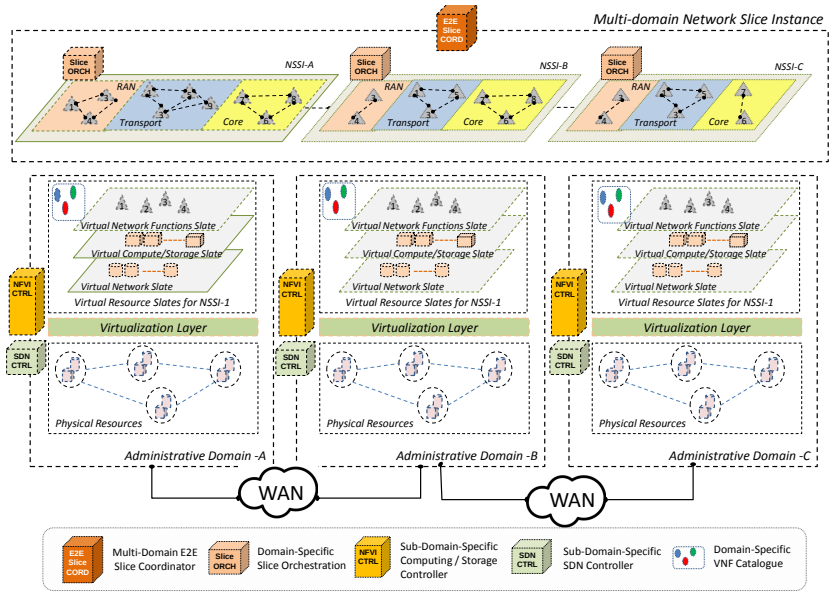


Figure 5.1. Federated Network Slice Across Multiple Administrative Domains, Publication III [72].

ditional ETSI MANO architecture: both the software-defined networking (SDN) and network function virtualization (NFV) principles are integrated into the essential 3GPP network slice management standard definitions. Our proposed architecture is broadly divided into four significant strata: the Multi-domain service conductor stratum, Domain-specific fully-fledged orchestration stratum, Sub-domain management and orchestration and connectivity stratum, and a logical multi-domain slice instance stratum. We have provided a detailed description of each stratum, including the essential operational specifics for the instantiation and management of the eventual multi-domain network slices.

Apart from the integration of the SDN technology and principles into the proposed architecture, we have also introduced a few but crucial elements (a bulk of which are specifically present in the multi-domain service conductor stratum) into the architecture that makes it stand out from others and truly enables it to support a coherent orchestration and life-cycle management of network slices across multiple administrative domains from federated network resources. Also, novel elements are integrated into the fully-fledged orchestration stratum. We have proposed a step-by-step procedure for the creation and modification of network slice orchestration requests, respectively, with specific details [72].

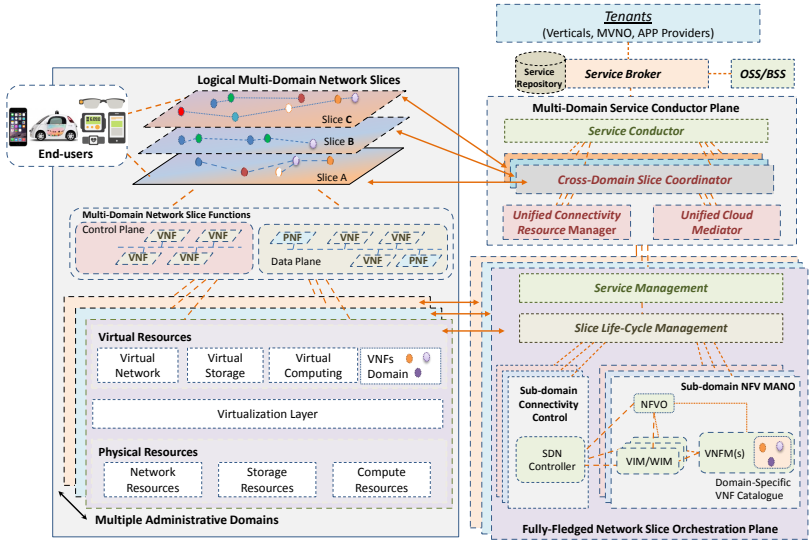


Figure 5.2. Multi-domain network slice orchestration architecture, Publication III [72].

5.2 Specific functional enabling components of the architecture

Most of the cutting-edge enabling building blocks of our proposed framework are located on the multi-domain service conductor plane and on the fully-fledged orchestration plane. In this section, we provide details of these proposed components and explain the contribution of each entity to the overall functionality of the federated architecture.

5.2.1 Service conductor

This critical building block is primarily responsible for decomposing the network slice request into units that will be addressed to different administrative domains. It also extracts the right combination of domain resources from the request for providing the appropriate connectivity across the concerned domains. For meticulous management and control over the instantiated multi-domain slices, it also instantiates a multi-domain slice coordinator, which performs a fine-grained life-cycle management procedure and provides the necessary credentials and access to rightful vertical owners. This element is also responsible for carrying out updates/upgrades based on a request during performance degradation or policy update to specific services across the federated domains.

5.2.2 Cross-domain slice coordinator

This vital element provides an indispensable set of functions to the multi-domain network slice instance. It is fundamentally responsible for monitoring, managing, and controlling the underlying resources related to a multi-domain network slice instance and ensures that trusted connectivity exists across the corresponding administrative domains. It also plays a mediation role among federated administrative domains regarding resource management and allocation within specific domains to update/re-adjust resources, in computing, in storage, and in networking which compensate for any potential performance degradation.

5.2.3 Unified cloud mediator

The cross-domain slice coordinator seamlessly handles resource allocation and control within specific domains using the unified cloud mediator component. This component explicitly handles the high-level translation of the federated cloud resources' performance capability and metrics description.

5.2.4 Unified connectivity resource manager

The unified connectivity resource manager is primarily in charge of translating the connectivity aspect of a network slice orchestration request and maps it onto the different indicated specific domains. The connectivity resources will then be negotiated and activated across the participating administrative domains.

5.2.5 Service management function

This functional block plays a pivotal role in the instantiation of the different VNFs indicated in the cross-domain slice orchestration request. It is primarily responsible for decomposing a received slice request into the RAN, core, and other value-added service VNFs. Based on its assessment of the request and the identified set of functions, it will determine the necessary logical links and return the needed service and performance capability information based on the current state of the underlying resources to the cross-domain slice coordinator.

Table 5.1. Network Slicing Orchestration Architectures and their Offered Support, Publication III [72].

Orchestration Architectures	Multi Admin. Domains	Multi Tech. Domains	RAN Orch.	Broker; AC/Neg.	Service Chain & SDN	Service Mang.	Federated LCM	3^{rd} Party Cntl. / Orch.	Programability	Recursive Virtualiz.	Unif. Connct. Mgmt.	Unif. Cloud Med.
3GPP TS28.530 [74]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗
SDN TR-526 [75]	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗
ITU-T Y.3011 [76]	✓	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✗
5G-EX [77]	✓	✓	✗	✗	✗	✓	✓	✗	✓	✗	✗	✓
5G!Pagoda [78]	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✓	✓
5G-Norma [79]	✗	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗
NFV-MANO MD [80]	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓
Proposed MDO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Supported feature ✓ Unsupported feature ✗

5.2.6 Slice life-cycle management function

This functional element directly interacts with the domain-specific orchestrators. The request from the service management function identifies a suitable slice template from which a logical network graph will be produced. This graph will, in turn, be used to compose the necessary low-level network resources such as the compute, storage, memory, and networking unit, including other requirements such as the desired network topology and policies. Once all of these are determined, the instantiation, configuration, run-time management, performance monitoring, and modification-related operation of an NSSI considering a specific administrative domain is the fundamental responsibility of the slice life-cycle management function [50].

5.2.7 Sub-domain SDN controller

This building block is significant in enabling E2E connectivity and seamless stitching [51] among the different VNFs and PNFs that make up the NSSI. It provides both domain-specific connectivity among the different VNF instances and remote service chaining between different cloud environments, which could also be enabled via physical switches and routers. It then provides the slice life-cycle management function with network resource view and monitoring reports used to monitor conformity with the service level agreement (SLA).

The enabling functional components discussed above are intended to facilitate and enhance the overall functionality of the proposed framework by providing the features highlighted in the Table5.1. As presented in

Table5.1, the highlighted set of features are crucial to the overall effective and efficient operation of a complete framework developed for the orchestration of network slices from federated resources. Some of these features are not readily available in all of the other presented orchestration systems compared to our proposed multi-domain orchestration system, thereby revealing the advantages of adopting our multi-domain slice orchestration approach.

5.3 Orchestration convergence framework for multi-domain slicing

The concept of network slicing, especially from a multi administrative and multi technological domain perspectives, has been identified as an essential aspect needed to catapult the generation of more revenue for the different network slicing ecosystem players. Most significantly, the MNOs and infrastructure owners [52] with a sizable amount of infrastructure distributed across different essential regions, as well as various vertical application providers with a significant user spread can generate more revenue through the practice of multi domain network slicing. These slicing ecosystem players will need a cost-effective means of orchestrating and managing network slices for optimal revenue generation. As a result, not only is it essential to design a functional multi-domain orchestration framework that takes into account the different contributions of these market players and translate them into revenues, but it is equally important to leverage the already achieved gains of the different existing ETSI MANO standard compliant orchestration platforms for providing the orchestration of slices stretching over multiple domains. To this effect, it is necessary to develop a framework for seamless interworking between different existing NFV MANO platforms by harnessing their functionalities through a unified interface for the orchestration of large scale network slices from federated and heterogeneous resources running across multiple administrative domains. This unified interface is developed with formal abstractions and Open APIs, i.e., Swagger [53] that enables it to act as a convergence platform through it is possible to simultaneously interact with each of the registered NFV orchestrators in a distributed manner. A platform such as this will enable network operators to orchestrate and configure different network slices directly from the infrastructure of registered credible slice providers and their orchestration solutions from truly distributed resources and a set of domain-specific orchestrators.

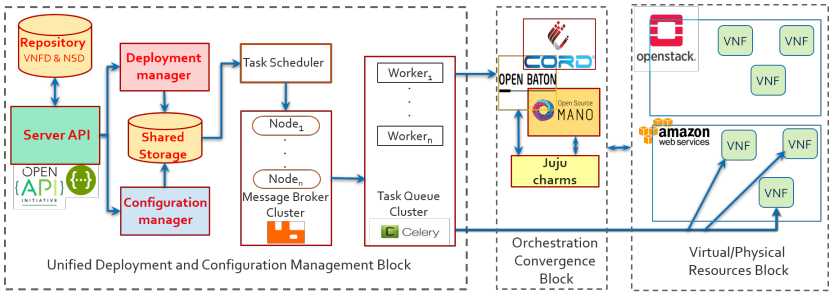


Figure 5.3. Implementation architecture of the framework, Publication I [73].

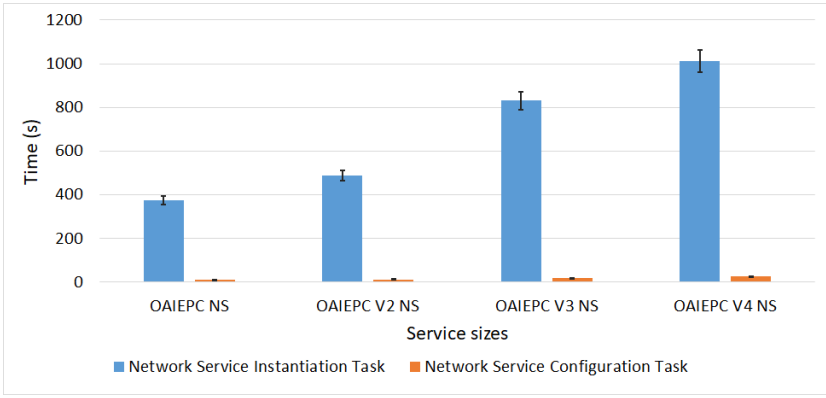
We have developed this solution using the architecture shown in Figure 5.3. The architecture is built upon three main blocks: the *Unified Deployment and Configuration Management Block* (UDCMB), the *Orchestration Convergence Block* (OCB) and the *V/PRB*. The UDCMB is made up of eight essential components, the *Server API*, *VNFD & NSD Repository*, *Deployment manager*, *Shared storage*, *Configuration manager*, *Task scheduler*, *Message broker cluster* and *Task queue cluster*. The OCB is the block that registers and coordinates between different NFV orchestrators. The V/PRB are the resources both physical and virtual that are used for the orchestration of network slices through their respective virtualized infrastructure managers (VIMs).

This system presents a practical example of a type of federated multi-domain network slicing concept introduced and described earlier on in Figure 5.2 above. Although this implementation architecture does not capture all of the slicing concept presented in Figure 5.2 in its totality, it can serve as a starting point towards achieving a genuinely multi-domain slice orchestration system over distributed resources.

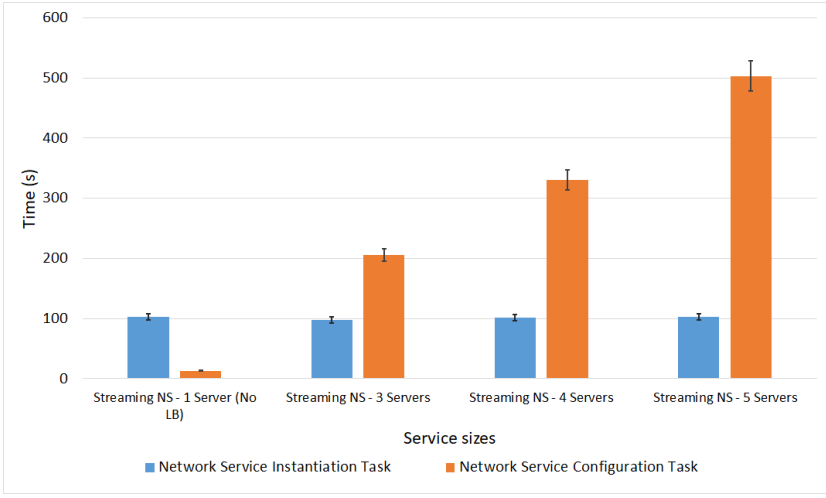
5.4 Performance evaluation

We have carried out the evaluation of our implemented framework based on the orchestration of two different network services, namely; (1) the mobile network service and (2) the video streaming service. Whereas the mobile network service is composed of the mobility management entity (MME), serving-packet data network gateway (S-PGW) and the Home Subscriber Server (HSS) of the Open Air Interface (OAI) mobile network solution, the streaming service is simply composed of the streamer and load-balancer of the Nginx software solution. Different sizes of each of

these network services were orchestrated across multiple NFVOs with particular emphasis on two important factors, namely, the *network service instantiation and configuration* duration in seconds respectively as shown in Figures 5.4(a) and 5.4(b) respectively.



(a) Ave. time (sec) of instantiation and configuration of different compositions of a complete OAI EPC network service.



(b) Ave. time (sec) of instantiation and configuration of different compositions of a video streaming service including a load balancer.

Figure 5.4. Experimental results of the convergence framework in terms of the instantiation and configuration of different network service types, Publication I [73].

Figure 5.4(a) presents the system evaluation result for both the *instantiation* (in blue) and *configuration* (in orange) of the OAI mobile network service with respect to time in seconds. We have orchestrated different sizes of the OAI mobile network service in an increasing order of sizes multiple times and measured the average time it took to instantiate and

configure each. So, for instance, while the instantiation of the VNFs instances of the *OAIEPC V4 NS* took a little above 1000sec, their configuration took only about 27sec. Similar trend is observed for all the different orchestrated OAIEPC sizes.

Conversely, however, for the orchestration of the streaming service, which was also carried out in a similar increasing order as in the case of the mobile network service, the configuration of the composing VNF instances took more time than their instantiation. For example, to instantiate the *Streaming NS-5 Servers*, it took about 100sec while their configuration took about 500sec as presented in Figure 5.4(b).

Although, the results might seem somewhat conflicting with respect to the instantiation time being noticeably more in Figure 5.4(a) than in Figure 5.4(b) and vice versa for the configuration time. However, it has actually further confirmed our conclusions in [66] (Publication IV), where it was revealed that the orchestration of VNF instances from uncached VNF images would generally take longer time than that of cached images, and the duration of VNF instantiation largely depends on its image size. Similarly, the duration of configuration for a VNF instance basically depends on both the number of parameters and files to be updated in the instance.

So, for the case of the network service, not only are the individual images (MME, S-PGW and HSS) different, they are also larger than that of the streaming service, which is basically, an Nginx software configured to operate as a streamer and a load-balancer. Thus, the duration of the instantiation for the streaming service stays fairly the same at about 100sec for all the variation of sizes considered.

5.5 Summary

In this chapter, we have discussed in details the orchestration and composition of network slices across multiple administrative domains. Specifically, we have also proposed our vision of the composition of a multi-domain network slice, that is orchestrated from federated network resources. Moreover, in order to enable the orchestration of the fully-fledged federated network slice, we also presented a robust multi-site/domain network slice orchestration architecture empowered with specific functional enabling building blocks. We also showed how our proposed architecture can extend and improve the current state of art with respect to existing and ongoing standardization efforts as detailed in Table 5.1 as detailed in

Publication III. Finally, following the principles of our proposed federated slice orchestrator, we presented our implementation of an orchestration and configuration convergence framework for the purpose of orchestration of network slices from ETSI compliant NFVOs that are distributed across multiple administrative cloud domains based on the contributions of Publication I. Finally, we also evaluated the system in terms of both the duration of instantiation and configuration of different network services.

6. Discussion, Conclusion, and Future Research

In this section, we summarize the main contributions of the research works discussed in this dissertation, the challenges addressed, and the requirements considered, especially in terms of the implementation of our E2E Slicer and Multi-domain network slice orchestration and configuration convergence framework. Finally, we discuss some of the future challenges and propose some research directions.

6.1 Discussion

The main focal area of the research works presented in this dissertation is the design and development of our E2E slicing framework for the orchestration of network slices in a service-aware fashion. This framework is called E2E because it is also tailored towards slicing the RAN segment of the mobile network. It could otherwise be called RAN slicing [81]. RAN slicing has been identified by many research works including [81] as one of the most challenging aspects of network slicing. Besides RAN slicing, which is necessary to create RAN sub-slices, we are also able to isolate slice traffic in the transport network part through the use of VLANs enabled over Open Virtual Switch (OVS).

Translating the high-level slice requirements based on the slice types (e/xMBB, mMTC, uRLLC, etc) into granular network resources to reflect the expected measurable QoS parameters (in terms of latency, bandwidth, jitter, availability, virtualization technology, isolation level, etc.) for a smooth operation of the network slice is quite challenging. In addition to these diverse requirements, especially at the RAN segment of the network, another crucial aspect is dynamically determining the adequate virtual resources (CPU, RAM, storage (SSD/HDD)) needed to power up the core sub-slice of the E2E network slices on the fly.

In addition to the mentioned non-trivial challenges, another essential aspect considered in this research work is the concept of flexible resource provisioning for the network slice orchestration system, with particular focus on the service orchestrator. This concept is essential to keep the orchestration system stable, operating efficiently, and delivering timely slice orchestration requests. The dynamic resource provisioning scheme will allow the system to be elastic by scaling in/out or up/down to adapt to the slice orchestration request workload.

Even though network slicing will form the foundation of the 5G technology, it has also been observed that due to the particular nature of some industrial use cases, orchestrating slices from a generic slice orchestration platform may not be the most effective. The reason is that the generic solution does not accommodate some fundamental resource support. As a result, to enable this set of industrial use cases, there is a need to design a customized slicing platform equipped with the needed resources and mechanisms available only on the identified sliceable industrial machinery. Particularly challenging in this aspect is the virtualization and abstraction of the integrated actuation and sensing functions that are readily available on this large industrial equipment and robots alike.

Furthermore, designing a feasible and robust architecture for the orchestration of network slices from federated multi-domain resources while taking resources from multiple administrative domains into account is considered both essential and challenging. Such an architecture is expected to enhance and extend the functionalities of existing network slice orchestration and management solutions through standard models and constructs that will enable them to be agnostic of the underlying virtualization and network infrastructure technologies. The slicing architecture is expected to integrate NFV and SDN concepts with the 3GPP slicing approach to deliver a system that will facilitate the composition of well-stitched network slices over federated resources with integrated life cycle manager [82].

This dissertation is intended to provide different levels of acceptable answers to the challenges mentioned earlier and possibly to others that are not mentioned. Also, it is targeted towards proposing and perhaps developing a wholesome network slice orchestration framework by functioning as proof of concept. Such a framework is designed to manage and control resources over multiple domains in a resource agnostic fashion while also considering the specific needs of the different services running on the

orchestrated network slices. To proffer solutions to these challenges, the contents of the chapters of this dissertation have contributed to the state of the art in the following ways.

In Chapter 2, we introduced the concept of network slice orchestration and presented the essential components of a complete E2E network slice. We discussed the different enabling technologies and concepts involved in the orchestration of network slices. In this chapter, the different slice sub-networks are presented, namely the RAN subnet, core subnet, and the transport subnet. The chapter also presented a strong background and definitions to facilitate a proper understanding of the network slicing concept, especially from the perspective of E2E in mobile networks. Moreover, different standardization organizations' activities towards enabling the orchestration of slices for fostering a better conceptualization by stakeholders in both academia and the industry are presented. This chapter therefore provides the research community with a strong resource for drawing on the knowledge included, the shared vision and the explanation of the slicing concept.

To address the identified challenges towards enabling a framework for the orchestration of network slices aimed for the customization of 5G services, a high-level E2E network slicing architecture was presented in Chapter 3. This architecture has been designed to facilitate the orchestration of 5G services in a service-aware manner based on the use case architecture presented in Section 3.2. This architecture also considers the service requirements and constraints of the RAN network by providing support for RAN slicing. This way, the developed framework can indeed be E2E by reflecting the service needs from the virtual core and RAN. All of these procedures are then processed using proposed slice templates.

In order to complement the proper operation of the developed framework, a set of distinctive features are implemented in line with the 3GPP standard specifications, including the RAN slicing procedure and slice-dedicated core network support, Network slice selection, and Slice instantiation and management function. These slice orchestration auxiliary functions are developed to provide an adequate operational and management foundation for our slicing framework.

Based on the success and progress derived from the framework discussed above, which is intended as a general-purpose E2E slicing system, we realized that a similar approach could also be extended towards enabling the slice orchestration from the industrial infrastructure. This

specific solution is based on the understanding that industrial equipment tends to have essential resources, specifically actuation and sensor hardware that are not readily available to regular customer of the shell (COTS) server infrastructures. As a result, there may be a need for a dedicated industrial-specific slicing architecture to capture some specific industrial use case concepts. This is explained in Section 3.5.

In order to reinforce the overall operation of a slice orchestration system and make it useful and efficient in terms of optimally serving slice orchestration requests, it is necessary to develop algorithms for predicting the orchestration request workload, which will enable the system scale elastically to cope with timely orchestration demands. This challenge is another essential aspect of slicing that has been identified within the scope of this research work. To prepare to meet this challenge, we have in Chapter 4, introduced our studies on the dynamic resource provisioning for network slicing orchestration systems (NSOS).

Through this study, we have proposed a dynamic resources auto-scaling solution for NSOS. The NSOS has been modeled as a global orchestrator that manages multiple domain-specific orchestrators, each managing virtual resources from an independent set of physical servers. The dynamic auto-scaling algorithm (DASA) uses the queuing network analyzer (QNA) technique to dimension the total available network resources, most notably and as an example, the computational resources across all the running servers based on the projected system workload. The result of this system dimensioning procedure will then be used to determine when the NSOS should dynamically scale to cope with a target response time T for the orchestration of network slices.

Chapter 5 presents a fine-tuned, novel, and modern architecture for the orchestration of multi-domain network slices over federated resources by taking a critical look at seamless connectivity between slice components and providing adequate management support for operating each orchestrated network slice. In this work, we provide detailed insight into a federated network slice structure across multiple administrative and technological domains. We also present a homogeneous front for harnessing and integrating developments in the different enabling technologies and standardization efforts, especially that of the 3GPP with a more business-oriented approach by designing an innovative multi-domain network slice orchestration architecture.

Finally, in a bid to leverage the existing NFVO solutions and some-

what harmonize distributed orchestration resources, whereby different resource infrastructures are managed by different NFVO, we have developed the slice orchestration convergence framework presented in Section 5.2. The solution is intended to provide a unified platform for the orchestration network slices from different NFVOs registered to operate on the platform in an NFVO agnostic manner. In a standard way, this solution will support the orchestration of multi-domain network slices, particularly from distributed resources managed and controlled by ETSI compliant NFVOs.

6.2 Conclusion

The research work summarized in this dissertation is intended to advance the knowledge and the state of the art regarding network slice orchestration and dynamic resource provisioning particularly in the areas of orchestration of E2E network slices that are service-driven or service-aware. In this vein, this dissertation compiles research works seeking satisfactory answers to the 5 questions raised in the opening chapter which are:

1. What are the network slicing enabling technologies and how can we facilitate network slice orchestration across the RAN, Core and Transport network parts?
2. How can we orchestrate customized network slices in a service-aware fashion?
3. How can we dynamically provision virtual network resources to network slice orchestration systems in order to improve their operational efficiency?
4. How can we create an efficient platform that converges the existing ETSI MANO compliant orchestration framework for the orchestration and management of network slices across multiple administrative cloud domains?
5. How can we apply an efficient platform such as the one described above to enable the use case of Industry 4.0?

Chapter 2 was developed to find an adequate answer to question 1 based on the outcome of Publications VI. Chapter 2 includes an outline and explanation of the different network slicing enabling technologies, and provides an overview of the various standardization authorities seeking to

standardize efforts from different research communities towards the realization of a uniform standard slicing framework for promoting 5G services. In the same light, we have highlighted the different types of slices and how they are composed and constructed using a slice template (a blueprint for determining network slices' composition tailored to a specific slice type). Also, we discussed the concept of E2E network slicing while considering three major technological domains of the mobile network, RAN, Transport, and Core networks, respectively.

In Chapter 3, we have attempted to provide a comprehensive answer to questions 2 and 5 as in Table 1.1 based on the outcomes of Publications IV and V. The chapter started by defining a suitable architecture in line with the approach of 3GPP for the orchestration of E2E network slices customized to deliver a particular 5G use case. This slicing architecture is designed to enable network slicing across the different interacting mobile network technology domains while considering the different slice requirements of each technology domain. To bring more clarity to the E2E slicing architecture operation, we also introduced a 5G vertical use case architecture. We discussed how different 5G services could be customized and delivered through the slices that are orchestrated using our framework. Furthermore, according to the 3GPP slicing concept for the smooth operation of the framework, we have developed and discussed some of our distinctive slicing features in detail. Finally, we have identified and extended the slicing concept beyond the traditional COTS servers to realize a dedicated novel slicing framework for enabling specific industrial use case scenarios, which may lead to more businesses and drive income generation for the manufacturing industry by proposing a dedicated novel slicing framework.

Chapter 4 is dedicated to providing an appropriate answer to question 3, with a strong focus on guaranteeing the efficient and effective operation of an NSOS by making the system elastic, dynamic and adaptive by accommodating an increase in workload that may emanate from orchestration requests. In this regard, we have developed a dynamic auto-scaling algorithm that provides dimensions for the entire system based on the QNA technique and triggers the NSOS to scale dynamically, primarily based on the system workload using extracted intelligence from the previous slicing requests based on the feedback generated from the workload monitoring agents. The DASA algorithm carries out dynamic resource provisioning either from the feedback from a reactive resource provisioning module or

directly from the workload predictor as captured in detail in Publication II.

Finally, in Chapter 5, we provided the details of our attempt to present a convincing answer to question 4 of Table 1.1, specifically, through the contributions of Publications I and III. This chapter proposed a well-structured federated network slice instance architecture that reflects the composition of slice components across both the technological and administrative domains in a very novel and realistic fashion. To facilitate the orchestration of such a comprehensive multi-domain slice instance, an innovative and holistic multi-domain slice orchestration system architecture is required. For this reason, we have also developed a universal futuristic multi-domain network slicing architecture with integrated features proposed to cater to the system and the operational requirements of the different 5G ecosystem stakeholders towards the provisioning of optimized and customized network slices for 5G verticals and their end-users. This novel system architecture then serves as an essential stimulant towards implementing our network services' orchestration and configuration convergence framework, which was concisely presented and briefly discussed in Section 5.2.

In all, the results from the many research works carried out and presented in this dissertation have provided some beneficial findings, which can be summarized as follows:

1. E2E network slicing is indeed feasible, but traffic isolation is quite challenging, most remarkably from the RAN and Transport networks. While different levels of isolation can be enforced at the RAN depending on the implemented type of RAN slicing (slicing of frequency spectrum or enforcing slicing policy at the MAC scheduler), there has to be a trade-off between traffic isolation and network bandwidth.

2. Different network services are often orchestrated from different service images and may also require different kinds of configurations. While some service images may be light-weight depending on the provided functions and supported features, others may be heavy in size, which will directly impact the speed of instantiating such a network service. Also, cached images [83, 84] tend to be instantiated more quickly than uncached ones.

3. While it may take some services a shorter time to be instantiated,

the same may not be claimed for their configuration duration. This result implies that there is no direct correlation between the duration of service instantiation and configuration.

4. The introduction of network slice orchestrators, especially for those including the RAN slicing, has equally introduced an additional delay overhead to the network in terms of network slice selection [85] for UE attachment. A function that was not needed in traditional mobile network setup, since the attachment of UEs is expressly handled by the eNodeB that selects a suitable MME [86] from a pool based on their *relative capacity* indicator for a network where load balancing is implemented and deployed. Now, the inception of network slices has necessitated the use of NSSF to carry out this functionality; hence, further delay overhead has been introduced to the network.

5. E2E network delay is a cumulative constraint [87] that would require the optimization of procedures from all of the different technological domains involved in the slicing process. While hardware acceleration [88] and the over-dimensioning of resources could ameliorate this challenge from some network segments (RAN and Core), the network links used in the transport network may be a source of fundamental limitation.

6.3 Future research

Network slicing will play a critical role in enabling not just the 5G technology but also beyond 5G networks. Hence, it is essential to have a comprehensive understanding of the nitty-gritty of this enabling technology in order to provide the right optimization techniques towards handling the needs of most vertical services and societal and business drivers [89]. It is anticipated that the 5G technology may be unable to adequately provide the necessary support to enable certain vertical services in its entirety. Most notably, the services are those within the category of the uRLLC and other time-sensitive and mission-critical ones with extremely stringent constraints. Therefore, enabling these services and many more like them will require extensive research work on different aspects of the network, considering the behavioral predictions of the system and analytics [90] where necessary. The research works conducted and reported

within the scope of this dissertation have laid a good foundation for understanding the essential requirements of service-oriented network slices and their orchestration systems.

As we continually optimize the network to deliver services with a consistent reduction in the E2E delay, enforcement of service traffic isolation and drive towards more cost-efficient system solutions, these critical factors may not always be possible simultaneously from the perspective of network slicing. Depending on the need for different network services, there might be a need for a trade-off. Slice owners may have to pay more to have certain slice types orchestrated and tick all the boxes for fulfilling its requirements, but some might still be missing. For example, a slice orchestration might be cost-efficient with a reduced E2E delay but not well isolated. In contrast, others may be well isolated with reduced delay but not cost-efficient, and so on. However, it should be possible to find a favorable balance between all of the critical requirements of these network slices. In this regard, machine learning techniques [91] and edge intelligence [92] may play a crucial role in finding the right balance.

For the machine learning concepts that could be integrated on the slice orchestration system to make it operate smarter, different candidate approaches are available. For example concepts such as deep [93], reinforcement [94] or transfer [95] learning could be deployed on the control plane of a network slice orchestration system to form part of its algorithmic and logical building block. By so doing, different important metrics could be monitored, measured and analysed on both the orchestration system and the deployed network slice in order to make the operation, administration and maintenance (OAM) of the network orchestration system more intelligent in handling the fundamental requirements of different network slices and continuously optimizing any possible performance degradation that might arise from resource over consumption/over subscription.

As a particularly indispensable part of a 5G network slice, which is responsible for carrying the slice traffic from one network segment to another, the **transport network slice subnet** plays a fundamental role in ensuring that network slices adhere to different stringent service requirements, most importantly in terms of latency and throughput. To this end, as part of an evolving standardization effort by different SDOs, the Internet Engineering Task Force (IETF) in particular, has been evolving standard specifications such as **deterministic networking** [63], which is similar and compatible with the IEEE802.1 TSN technology as discussed

in Section 2.5.3. While the deterministic networking specification is developed to operate over both layer 2 and 3 protocols in general, the IETF has also evolved a slice-specific specification known as the **transport-slice-assurance-policy**, which is specific for 5G slice information model definition [64].

Essentially, the **transport-slice-assurance-policy** slice information model is intended to encapsulate information related to the slices quality assurance at the transport network based on resource monitoring, analytics and continuous optimization. These abstracted information include the slice assurance type, time interval and frequency [64] of telemetry data transfer to the transport network controller [62, 65], which will in turn aggregate the data as key metrics for further processing. This component of the slice information model, when present will enable the slice orchestrator to continuously monitor, enforce and improve the agreed service requirements [65] of the network slice when a degradation is observed.

Finally, slice infrastructure providers and mobile network operators will always seek to maximize their profit by developing a slice orchestration system that can continuously optimize the response or execution time for the orchestration of network slices [96]. This way, the network operator can adequately estimate the number of resources needed to serve the slicing request from their customers and the amount of potential revenue they can generate and plan accordingly. In this way, their system will be able to scale autonomously and reduce the rate at which slice orchestration requests will be churned away. The research effort required to achieve this will be considerable. To provide a clear path forward towards beyond 5G and possibly to the generation of a 6G network, many aspects of the network require a critical and thorough investigation.

References

[1] Ericsson Mobility Report, [Online; accessed 08-September-2020] URL: https://www.ericsson.com/en/mobility-report/reports/june-2020/network-coverage

[2] H. Cao, S. Wu, F. Tian and L. Yang, "Efficient Virtualized Resources Allocation in Network Virtualization Environment: A Service Oriented Perspective," 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 2019, pp. 1-6, doi: 10.1109/WCSP.2019.8927846.

[3] I. Afolabi, A. Ksentini, M. Bagaa, T. Taleb, M. Corici, and A. Nakao, "Towards 5G Network Slicing over Multiple-Domains," in IEICE Trans. on Communications, Vol. E100.B, No. 11, Nov. 2017, pp. 1992-2006.

[4] I. Afolabi, M. Bagaa, T. Taleb, and H. Flinck, "End-to-End Network Slicing Enabled Through Network Function Virtualization," in Proc. IEEE CSCN'17, Helsinki, Finland, Sep. 2017.

[5] 3GPP TR 22.891, Feasibility Study on New Services and Markets Technology Enablers, Release 14, Sep. 2016.

[6] 3GPP TR 22.861, Feasibility Study on New Services and Markets Technology Enablers for Massive Internet of Things, Release 14, Sep. 2016.

[7] 3GPP TR 22.862, Feasibility Study on New Services and Markets Technology Enablers for Critical Communications, Release 14, Oct. 2016.

[8] 3GPP TR 22.863, Feasibility Study on New Services and Markets Technology Enablers - Enhanced Mobile Broadband, Release 14, Sep. 2016.

[9] 3GPP TR 22.864, Feasibility Study on New Services and Markets Technology Enablers - Network Operation, Release 15, Sep. 2016.

[10] GSM Association, Generic network slice template, version 3.0, May, 2020.

[11] ITU - RM.2083-0, IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and beyond, Sep. 2015

[12] 3GPP TR 22.864, Study on enhancement of 3GPP support for 5G V2X services, Release 15, Dec. 2018.

[13] R. A. Addad, D.L.C. Dutra, M. Bagaa, T. Taleb, H. Flinck, and M. Namane "Benchmarking the ONOS Intent interfaces to ease 5G service management," in IEEE Globecom'18, Abu Dhabi, UAE, Dec. 2018.

[14] H. Hellaoui, A. Chelli, M. Bagaa, and T. Taleb, "Towards Mitigating the Impact of UAVs on Cellular Communications," in IEEE Globecom'18, Abu Dhabi, UAE, Dec. 2018.

[15] M. Bagaa, T. Taleb, A. Chelli, and H. Hellaoui, "Constraint hubs deployment for efficient machine-type communications," IEEE Trans. Wireless Communications, vol. 17, no. 12, pp. 7936–7951, December 2018.

[16] Y. Khettab, M. Bagaa, D. Dutra, T. Taleb, and N. Toumi, "Virtual Security as a Service for 5G Verticals," in Proc. IEEE WCNC 2018, Barcelona, Spain, Apr. 2018.

[17] J. Ordonez-Lucena, C. Tranoris and J. Rodrigues, "Modeling Network Slice as a Service in a Multi-Vendor 5G Experimentation Ecosystem," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/ICCWorkshops49005.2020.9145225.

[18] Description of Network Slicing Concept, [Online; accessed 10-September-2020]
URL: https://www.ngmn.org/publications/description-of-network-slicing-concept.html

[19] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing & Softwarization: A Survey on Principles, Enabling Technologies & Solutions," in IEEE Communications Surveys & Tutorials, 3rd Quarter, Vol. 20, No. 3, Mar. 2018, pp. 2429-2453.

[20] A. Laghrissi, T. Taleb, and M. Bagaa, “Conformal mapping for optimal network slice planning based on canonical domains,” IEEE Journal on Selected Areas in Communications, vol. 36, no. 3, pp. 519–528, March 2018.

[21] I. Farris, T. Taleb, M. Bagaa, and H. Flinck, “Optimizing Service Replication for Mobile Delay-sensitive Applications in 5G Edge Network,” in Proc. IEEE ICC 2017, Paris, France, May 2017.

[22] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, “On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Architecture Orchestration,” in IEEE Communications Surveys Tutorials J., Vol. 19, No. 3, May 2017. pp. 1657 - 1681.

[23] Anwer Al-Dulaimi; Xianbin Wang; Chih-Lin I, “Network Softwarization View of 5G Networks,” in 5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management , IEEE, 2018, pp.499-518, doi: 10.1002/9781119333142.ch13.

[24] W. Xia, Y. Wen, C. H. Foh, D. Niyato and H. Xie, “A Survey on Software-Defined Networking,” in IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pp. 27-51, Firstquarter 2015, doi: 10.1109/COMST.2014.2330903.

[25] B. A. A. Nunes, M. Mendonca, X. Nguyen, K. Obraczka and T. Turletti, “A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks,” in IEEE Communications Surveys & Tutorials, vol. 16, no. 3, pp. 1617-1634, Third Quarter 2014, doi: 10.1109/SURV.2014.012214.00180.

[26] ETSI GS NFV-SWA 001 : “Network Functions Virtualisation (NFV); Virtual Network Functions Architecture,” version 1.1.1, Dec. 2014.

[27] Ilias Benkacem, M. Bagaa, T. Taleb, Q.N. Nguyen, T. Tsuda, and T. Sato, “Integrated ICN and CDN Slice as a Service”, in IEEE Globecom’18, Abu Dhabi, UAE, Dec. 2018.

[28] Y. Khettab, M. Bagaa, D. Dutra, T. Taleb, and N. Toumi, “Virtual Security as a Service for 5G Verticals,” in Proc. IEEE WCNC 2018, Barcelona, Spain, Apr. 2018.

[29] A. M. Zarca, J. B. Bernabé, I. Farris, Y. Khettab, T. Taleb, and A. F. Skarmeta, “Enhancing iot security through network softwarization

and virtual security appliances," Int. Journal of Network Management, vol. 28, no. 5, 2018.

[30] 3GPP TR 23.799, Study on Architecture for Next Generation System, Rel.14, Dec. 2016

[31] ITU-T Y.3011, Framework of Network Virtualization for Future Networks, Next Generation Network - Future Networks, Jan

[32] H. Hantouti, N. Benamar, and T. Taleb, "Service Function Chaining in Next-Generation networks: Challenges and Open Research Issues," in IEEE Network Magazine. (to appear)

[33] A. Laghrissi and T. Taleb, "A Survey on the Placement of Virtual Resources and Virtual Network Functions," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1409 - 1434, 2nd Quarter 2019.

[34] I. Benkacem, T. Taleb, M. Bagaa, and H. Flinck, "Optimal vnfs placement in CDN slicing over multi-cloud environment," IEEE Journal on Selected Areas in Communications, vol. 36, no. 3, pp. 616–627, Mar. 2018.

[35] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer and X. Hesselbach, "Virtual Network Embedding: A Survey," in IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 1888-1906, Fourth Quarter 2013, doi: 10.1109/SURV.2013.013013.00155.

[36] Understanding Full Virtualization, Paravirtualization and Hardware assist, [Online; accessed 10-December-2021]
URL: https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/VMware_paravirtualization.pdf

[37] R. Behravesh, E. Coronado and R. Riggio, "Performance Evaluation on Virtualization Technologies for NFV Deployment in 5G Networks," 2019 IEEE Conference on Network Softwarization (NetSoft), 2019, pp. 24-29, doi: 10.1109/NETSOFT.2019.8806664.

[38] Use containers to Build, Share and Run your applications, [Online; accessed 10-December-2021]
URL: https://www.docker.com/resources/what-container

[39] Building modern apps with Linux containers, [Online; accessed 10-December-2021]

URL: https://www.redhat.com/rhdc/managed-files/RED19-0059_Intro_to_Linux_Containers_for_Developers_Ebook_D3%20(1).pdf

[40] 3GPP TR 28.801, Study on management and orchestration of network slicing for next generation network, Release 15, Sep. 2017.

[41] G. d. Campo, S. Calatrava, G. Cañada, J. Olloqui, R. Martinez and A. Santamaria, "IoT Solution for Energy Optimization in Industry 4.0: Issues of a Real-life Implementation," 2018 Global Internet of Things Summit (GIoTS), Bilbao, 2018, pp. 1-6, doi: 10.1109/GIOTS.2018.8534537.

[42] S. K. Datta and C. Bonnet, "MEC and IoT Based Automatic Agent Reconfiguration in Industry 4.0," 2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Indore, India, 2018, pp. 1-5, doi: 10.1109/ANTS.2018.8710126.

[43] H. Halabian, "Distributed Resource Allocation Optimization in 5G Virtualized Networks," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 3, pp. 627-642, March 2019, doi: 10.1109/JSAC.2019.2894305.

[44] B. Urgaonkar, P Shenoy, A. Chandra, P. Goyal, and T. Wood, "Agile Dynamic Provisioning of Multi-tier Internet Applications," *ACM Trans. Auton. Adapt. Syst.*, vol. 3, no. 1, pp. 1–39, Mar 2008.

[45] M. H. Dahir, H. Alizadeh and D. Gözüpek, "Energy efficient virtual network embedding in federated software defined networks," 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017, pp. 1-4, doi: 10.1109/SIU.2017.7960206.

[46] S. Arnautov, P. Felber, C. Fetzer and B. Trach, "FFQ: A Fast Single-Producer/Multiple-Consumer Concurrent FIFO Queue," 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Orlando, FL, 2017, pp. 907-916, doi: 10.1109/IPDPS.2017.41.

[47] S. Basterrech, S. Mohammed, G. Rubino and M. Soliman, "Levenberg—Marquardt Training Algorithms for Random Neural Networks," in The Computer Journal, vol. 54, no. 1, pp. 125-135, Jan. 2011, doi: 10.1093/comjnl/bxp101.

[48] J. Chen, H. Cao and L. Yang, “NFV MANO Based Network Slicing Framework Description,” 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), YILAN, Taiwan, 2019, pp. 1-2, doi: 10.1109/ICCE-TW46550.2019.8991887.

[49] G. Tseliou, F. Adelantado and C. Verikoukis, “NetSliC: Base Station Agnostic Framework for Network SliCing,” in IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3820-3832, April 2019, doi: 10.1109/TVT.2019.2902320.

[50] A. Boubendir, F. Guillemin, S. Kerboeuf, B. Orlandi, F. Faucheux and J. Lafragette, “Network Slice Life-Cycle Management Towards Automation,” 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Arlington, VA, USA, 2019, pp. 709-711.

[51] S. Clayman, F. Tusa, A. Galis and L. M. Contreras, “WIM on-demand – A modular approach for managing network slices,” 2020 6th IEEE Conference on Network Softwarization (NetSoft), Ghent, Belgium, 2020, pp. 395-403, doi: 10.1109/NetSoft48620.2020.9165342.

[52] F. Esposito, D. Di Paola and I. Matta, “A general distributed approach to slice embedding with guarantees,” 2013 IFIP Networking Conference, Brooklyn, NY, 2013, pp. 1-9.

[53] API Development for Everyone, [Online; accessed 10-September-2020]
URL: https://swagger.io/

[54] 3GPP TR 22.852, Study on Radio Access Network (RAN) Sharing enhancements, Release 13, Sep. 2014.

[55] 3GPP TS 32.130, Telecommunication management; Network Sharing; Concepts and requirements, Rel.12, Jan. 2016.

[56] H. Chien, Y. Lin, C. Lai and C. Wang, “End-to-End Slicing as a Service with Computing and Communication Resource Allocation for Multi-Tenant 5G Systems,” in IEEE Wireless Communications, vol. 26, no. 5, pp. 104-112, October 2019, doi: 10.1109/MWC.2019.1800466.

[57] B. Ahmad, T. Taleb, A. Vajda, and M. Bagaa, “Dynamic Cloud Resource Scheduling in Virtualized 5G Mobile Systems,” in Proc. IEEE Globecom 2016, Washington, USA, Dec. 2016.

[58] S. Bhattacharjee et al., “Network Slicing for TSN-Based Transport Networks,” in IEEE Access, vol. 9, pp. 62788-62809, April 2021, doi: 10.1109/ACCESS.2021.3074802.

[59] L. Chinchilla-Romero, J. Prados-Garzon, P. Ameigeiras, P. Muñoz, J. Lopez-Soler, “5G Infrastructure Network Slicing: E2E Mean Delay Model and Effectiveness Assessment to Reduce Downtimes in Industry 4.0,” Sensors 2022, 22, 229.https://doi.org/10.3390/s22010229

[60] Data Center Overlay Technologies, [Online; accessed 10-January-2022]
URL: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-730116.html

[61] Technologies Evolution for Precision 5G Transport Network, [Online; accessed 11-January-2022]
URL: https://www.zte.com.cn/global/about/magazine/zte-technologies/2021/2-en/Special-Topic/1.html

[62] 5G transport slice control in end-to-end 5G networks, [Online; accessed 11-January-2022]
URL: https://onestore.nokia.com/asset/207124

[63] Deterministic Networking (DetNet) Controller Plane Framework draft-ietf-detnet-controller-plane-framework-01, [Online; accessed 11-January-2022] URL: https://datatracker.ietf.org/doc/draft-ietf-detnet-controller-plane-framework/

[64] 5G Transport Slice Connectivity Interface, [Online; accessed 11-January-2022]
URL: https://datatracker.ietf.org/doc/html/draft-rokui-5g-transport-slice#section-6.6.3

[65] Transport slicing in end-to-end 5G networks, [Online; accessed 11-January-2022] URL: https://onestore.nokia.com/asset/206889

[66] I. Afolabi, T. Taleb, P. A. Frangoudis, M. Bagaa and A. Ksentini, “Network Slicing-Based Customization of 5G Mobile Services,” in IEEE Network Magazine, vol. 33, no. 5, pp. 134-141, Sept.-Oct. 2019.

[67] W. Whitt, “The Queueing Network Analyzer,” in The Bell System Technical Journal, vol. 62, no. 9, pp. 2779-2815, Nov. 1983, doi: 10.1002/j.1538-7305.1983.tb03204.x.

[68] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. P. Kontovasilis, “FlexRAN: A flexible and programmable platform for software-defined radio access networks,” in Proc. ACM CoNEXT, 2016.

[69] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin, “Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks,” in *Proceedings of the 7th International Workshop on Hot Topics in Planet-scale mObile Computing and Online Social neTworking, (ser. HOTPOST ’15),* New York, NY, USA, 2015, pp. 19–24.

[70] F. Manco, C. Lupu, F. Schmidt, J. Mendes, S. Kuenzer, S. Sati, K. Yasukata, C. Raiciu, and F. Huici, “My VM is Lighter (and Safer) Than Yours,” in *Proc. of the 26th Symposium on Operating Systems Principles (SOSP ’17)*, Shanghai, China, 2017, pp. 218-233.

[71] I. Afolabi, J.P. Garzon, M. Bagaa, T. Taleb, and P. Ameigeiras, “Dynamic Resource Provisioning of a Scalable E2E Network Slicing Orchestration System,” in IEEE Transactions on Mobile Computing, Vol. 19, no. 11, Nov. 2020, pp. 2594 - 2608

[72] T. Taleb, I. Afolabi, K. Samdanis and F. Z. Yousaf, “On Multi-domain Network Slicing Orchestration Architecture & Federated Resource Control,” in IEEE Network Magazine, Vol. 33, No. 5, Sep. 2019, pp. 242 - 252.

[73] I. Afolabi, M. Bagaa, W. Boumezer and T. Taleb, “Toward a Real Deployment of Network Services Orchestration and Configuration Convergence Framework for 5G Network Slices,” in IEEE Network Magazine, vol. 35, no. 1, Jan/Feb. 2021, pp. 242 - 250.

[74] 3GPP TS 28.530, “Management of 5G networks and network slicing; Concepts, use cases and requirements”, Rel.15, Apr. 2018.

[75] ONF TR-526, Applying SDN Architecture to 5G Slicing, Issue 1, Apr. 2016.

[76] ITU-T Y.3011, Framework of Network Virtualization for Future Networks, Jan. 2012.

[77] R. Guerzoni, et al., “Analysis of end-to-end Multi-domain Management and Orchestration Frameworks for Software Defined Infras-

tructures: An Architectural Survey", Transactions on Emerging Telecommunications Technologies, Vol. 28, No. 4, Apr. 2017.

[78] I. Afolabi. A. Ksentini, M. Bagaa, T. Taleb, M. Corici, A. Nakao, "Towards 5G Network Slicing over Multiple-Domains", IEICE Transactions on Communications, Vol.100B, No.11, Nov. 2017.

[79] 5G NORMA, "D3.3: 5G NORMA network architecture - Final report", September 2020, available at http://www.it.uc3m.es/wnl/5gnorma/pdf/5g_norma_d3-3.pdf.

[80] ETSI NFV ISG, *"Network Function Virtualisation (NFV); Management and Orchestration;Report on architecture options to support multiple administrative domains,"* ETSI GR NFV-IFA 028 V3.1.1, 2018.

[81] H. D. R. Albonda and J. Pérez-Romero, "An Efficient RAN Slicing Strategy for a Heterogeneous Network With eMBB and V2X Services," in IEEE Access, vol. 7, pp. 44771-44782, 2019, doi: 10.1109/ACCESS.2019.2908306.

[82] S. Kukliński and L. Tomaszewski, "DASMO: A scalable approach to network slices management and orchestration," NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, Taipei, 2018, pp. 1-6, doi: 10.1109/NOMS.2018.8406279.

[83] Z. Zhang, W. Zhou, L. Qian, Z. Luo, S. Sun and X. Huang, "An Image Cache in IaaS Cloud," 2012 7th Open Cirrus Summit, Beijing, 2012, pp. 16-19, doi: 10.1109/OCS.2012.33.

[84] Z. Huang, S. Wu, S. Jiang and H. Jin, "FastBuild: Accelerating Docker Image Building for Efficient Development and Deployment of Container," 2019 35th Symposium on Mass Storage Systems and Technologies (MSST), Santa Clara, CA, USA, 2019, pp. 28-37, doi: 10.1109/MSST.2019.00-18.

[85] M. Shurman, J. Rawashdeh and A. Jaradat, "Slice Selection In 5G Networks: Novel Approach for Accessing Multiple Slices Simultaneously," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 113-117, doi: 10.1109/ICICS49469.2020.239560.

[86] E. Aqeeli, A. Moubayed and A. Shami, "Dynamic SON-Enabled Location Management in LTE Networks," in IEEE Transactions on

Mobile Computing, vol. 17, no. 7, pp. 1511-1523, 1 July 2018, doi: 10.1109/TMC.2017.2775221.

[87] M. Laner, J. Fabini, P. Svoboda and M. Rupp, “End-to-end Delay in Mobile Networks: Does the Traffic Pattern Matter?,” ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems, Ilmenau, Germany, 2013, pp. 1-5.

[88] L. Nobach and D. Hausheer, “Open, elastic provisioning of hardware acceleration in NFV environments,” 2015 International Conference and Workshops on Networked Systems (NetSys), Cottbus, 2015, pp. 1-5, doi: 10.1109/NetSys.2015.7089057.

[89] M. Latva-aho and K. Leppanen (ed.),“Key drivers and research challenges for 6G ubiquitous wireless intelligence” (white paper). Oulu, Finland: 6G Flagship, Sep. 2019.

[90] T. Taleb, R. L. Aguiar, I. G. B. Yahia, B. Chatras, G. Christensen, U. Chunduri, A. Clemm, X. Costa, L. Dong, J. Elmirghani, B. Yosuf, X. Foukas, A. Galis, M. Giordani, A. Gurtov, A. Hecker, C. W. Huang, C. Jacquenet, W. Kellerer, ... M. Zorzi (2020). White Paper on 6G Networking [White paper]. (6G Re-search Visions, No. 6). University of Oulu. http://urn.fi/urn:isbn:978952622684

[91] S. Ali, W. Saad, D. (Eds.). Steinbach (2020). White Paper on Machine Learning in 6G Wireless Com-munication Networks [White paper]. (6G Research Visions, No. 7). University of Oulu. http://urn.fi/urn:isbn:978952622673

[92] E. Peltonen, M. Bennis, M. Capobianco, M. Debbah, A. Ding, F. Gil-Castiñeira, M. Jurmu, T. Karvonen, M. Kelanti, A. Kliks, T. Leppänen, L. Lovén, T. Mikkonen, A. Rao, S. Samarakoon, K. Seppänen, P. Sroka, S. Tarkoma, T. Yang (2020). 6G White Paper on Edge Intelligence [White paper]. (6G Research Visions, No. 8). University of Oulu. http://urn.fi/urn:isbn:9789526226774

[93] M. Abbasi, A. Shahraki and A. Taherkordi, “Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey,” Computer Communications, 170, pp.19-41, Feb. 2021, doi: https://doi.org/10.1016/j.comcom.2021.01.021

[94] J. Tanveer, A. Haider, R. Ali and A. Kim, “An Overview of Reinforcement Learning Algorithms for Handover Management in 5G Ultra-

Dense Small Cell Networks," Applied Sciences, 12(1), pp.426, Jan. 2022, doi: https://doi.org/10.3390/app12010426

[95] H. Dhillon and A. Haque, "Towards Network Traffic Monitoring Using Deep Transfer Learning," in 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2020, doi: 10.1109/TrustCom50675.2020.00144

[96] R. Fantacci and B. Picano, "When Network Slicing Meets Prospect Theory: A Service Provider Revenue Maximization Framework," in IEEE Transactions on Vehicular Technology, vol. 69, no. 3, pp. 3179-3189, March 2020, doi: 10.1109/TVT.2019.2963462.

Network function virtualization and software-defined networking are fundamental enabling principles in network slice orchestration. The orchestration of network slices that cuts across the different technological domains of the mobile network (i.e., RAN, Transport, and Core network) as well as the administrative domains (i.e., different cloud providers, MNO clouds and other private/public clouds) is called E2E network slicing. Orchestrating E2E network slices to deliver different 5G services involves a number of requirements and constraints at different technological domains of the mobile network. Fulfilling these divergent but distinctive network requirements for each slice type while provisioning virtual resources from multiple administrative domains will require a holistic framework and intricate architecture capable of considering all of the possible slicing ramifications and the necessary system support. Traffic isolation, dynamic resource provisioning, and dynamic re/configuration are among the most notable challenges of network slicing.

Aalto-DT 68/2022

ISBN 978-952-64-0802-6 (printed)
ISBN 978-952-64-0803-3 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Communications and Networking
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES