NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
SCHOOL OF SCIENCE
DEPARTMENT OF PHYSICS

SECTION OF ELECTRONIC PHYSICS AND SYSTEMS

Design of Cloud-Native Communications Networks in Mobile Environments

DOCTORAL DISSERTATION

by

Alexandros-Ioannis Manolopoulos

Athens, December 2024.

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
SCHOOL OF SCIENCE
DEPARTMENT OF PHYSICS
SECTION OF ELECTRONIC PHYSICS AND SYSTEMS

Design of Cloud-Native Communications Networks in Mobile Environments

DOCTORAL DISSERTATION

by

Alexandros-Ioannis Manolopoulos

Supervisory Committee: Anna Tzanakaki, Chair
Dimitra Simeonidou, Committee Member
Dionysios Reisis, Committee Member

Doctoral Committee:

.....	.....
Anna Tzanakaki Associate Professor NKUA	Dionysios Reisis Professor NKUA
.....	.....
Dimitra Simeonidou Professor University of Bristol	Markos Anastasopoulos Associate Professor NKUA
.....	.....
Ektoras Nistazakis Professor NKUA	Athanasia Alonistioti Professor NKUA
	.....
	Alexandros Kaloxylos Professor University of Peloponnese

Athens, December 2024

DECLARATION OF AUTHORSHIP

I, Alexandros-Ioannis MANOLOPOULOS, declare that this thesis titled, “Design of Cloud-Deployed Communication Networks in Mobile Environments” and the work presented in it are my own. I confirm that:

- This research was conducted during my candidature for a research degree at this University, either wholly or mainly.
- In cases where any section of this thesis has previously been submitted for a degree or qualification at this University or any other institution, it has been explicitly declared.
- Where I have consulted the published work of others, this is always clearly attributed.
- The source of any quoted material from the work of others is always provided. Aside from such quotes, this thesis represents solely my own original work.
- I have acknowledged all significant sources of assistance.
- If this thesis is based on collaborative work between myself and others, I have clearly delineated the contributions made by others and myself.

Signed:

Date:

ΠΕΡΙΛΗΨΗ

Η άφιξη των δικτύων 5ης γενιάς κινητών επικοινωνιών (5G) και πέραν αυτών (B5G) θα διαδραματίσει καθοριστικό ρόλο στην αντιμετώπιση θεμελιωδών προκλήσεων που σχετίζονται με τη βιώσιμη κοινωνική μεταμόρφωση και την οικονομική ανάπτυξη. Τα δίκτυα αυτά απαιτούν σχεδιασμό που θα υποστηρίζει νέες δυνατότητες υπηρεσιών, παρέχοντας ευρεία συνδεσιμότητα για μεγάλο αριθμό συσκευών με αυξημένα επίπεδα κινητικότητας σε ετερογενή περιβάλλοντα, καθώς και υπηρεσίες κρίσιμης σημασίας με αυστηρές απαιτήσεις απόδοσης, οι οποίες θα ικανοποιούνται με οικονομικά και ενεργειακά αποδοτικό τρόπο. Ωστόσο, για την υποστήριξη αυτών των απαιτητικών και ποικιλόμορφων υπηρεσιών απαιτείται η μετάβαση από τις υπάρχουσες κλειστές, εξειδικευμένες υποδομές σε ανοιχτά και ευέλικτα οικοσυστήματα τα οποία θα υιοθετούν ελαστικά αρχιτεκτονικά μοντέλα, επιτρέποντας τη σύγκλιση και την ενσωμάτωση διαφόρων τεχνολογιών πληροφορικής και δικτύων, όπως τα υπολογιστικά νέφη (cloud), η εικονικοποίηση (virtualization) και η αποσυγκέντρωση (disaggregation) των υπολογιστικών, δικτυακών και αποθηκευτικών πόρων.

Προς αυτή την κατεύθυνση, η παρούσα διδακτορική διατριβή επιδιώκει να αξιοποιήσει διάφορα αρχιτεκτονικά μοντέλα και τεχνολογίες που υιοθετούνται από τα δίκτυα 5G/B5G για τον σχεδιασμό και την υλοποίηση συστημάτων που θα παρέχουν αποδοτικές υπηρεσίες σε περιβάλλοντα με ποικίλα επίπεδα κινητικότητας. Συγκεκριμένα, υποστηρίζει την ιδέα μιας καθολικής πλατφόρμας 5G, φιλοξενούμενης σε υποδομές νεφών, με ενοποιημενούς φυσικούς πόρους. Αυτή η πλατφόρμα θα ενσωματώσει διάφορα μαθηματικά μοντέλα, πρωτόκολλα και αλγορίθμους με στόχο τη βελτιστοποίηση της συνολικής απόδοσης του συστήματος.

Η διατριβή επικεντρώνεται σε τρεις διακριτές συνεισφορές, οι οποίες υλοποιούνται και επικυρώνονται πειραματικά μέσω ενός συνόλου εργαλείων ανοικτού κώδικα. Η πρώτη συνεισφορά εξετάζει τις αρχιτεκτονικές επιλογές που παρέχονται από τα δίκτυα 5G/B5G και προτείνει αποτελεσματικές στρατηγικές κατανομής πόρων για τη βέλτιστη παροχή υπηρεσιών, λαμβάνοντας υπόψη τις απαιτήσεις κινητικότητας των χρηστών. Εξετάζει τη διαχείριση πόρων σε διάφορα τμήματα του δικτύου, όπως το Ραδιοδίκτυο Πρόσβασης (RAN), το Δίκτυο Πυρήνα (CN) και το Δίκτυο Δεδομένων (DN). Οι πόροι αυτοί εικονικοποιούνται και αποσυγκεντρώνονται για να ενισχυθεί η ευελιξία στην τοποθέτηση των Δικτυακών Λειτουργιών (NFs) μέσα στο δίκτυο. Για την υποστήριξη της κινητικότητας των χρηστών, γίνεται χρήση της δυνατότητας ζωντανής μεταφοράς Εικονικών Μηχανών (live VM migration), που επιτρέπει τη δυναμική μετακίνηση κρίσιμων λειτουργιών 5G, διασφαλίζοντας την απρόσκοπτη παροχή υπηρεσιών. Επίσης, αναπτύσσεται ένα πολυεπίπεδο πλαίσιο βελτιστοποίησης για τη βέλτιστη κατανομή δικτυακών και υπολογιστικών πόρων, ελαχιστοποιώντας τα λειτουργικά κόστη. Η ανάλυση βασίζεται σε στατιστικά πραγματικής κινητικότητας και μετρήσεις κατανάλωσης πόρων από μια εργαστηριακή υποδομή 5G σε cloud.

Η δεύτερη συνεισφορά αφορά την ανάπτυξη ενός πολυτεχνολογικού ραδιοδικτύου πρόσβασης, που ενσωματώνει τεχνολογίες 3GPP και non-3GPP, με σκοπό την υποστήριξη μιας ευρείας γκάμας κινητών τερματικών. Χρήστες με διαφορετικά μοτίβα κινητικότητας συνδέονται στο δίκτυο, ζητώντας τη δημιουργία εικονικών δικτύων υπό την μορφή επί μέρους τεμαχίων της υπαρκτής δικτυακής υποδομής δικτύου (network slices) προκειμένου να εξυπηρετηθούν. Όταν οι συσκευές συνδέονται μέσω του δικτύου

πρόσβασης 3GPP, τα τεμάχια δικτύου δημιουργούνται κατά μήκος της διαδρομής που διασυνδέει τον σταθμό βάσης (gNodeB-gNB), τους κόμβους Λειτουργίας Επιπέδου Χρήστη (UPFs) και το DN. Αντίθετα, για τις συσκευές που εξυπηρετούνται μέσω των non-3GPP δικτύων, η σύνδεση με το DN επιτυγχάνεται μέσω της Λειτουργίας Διασύνδεσης non-3GPP (N3IWF), η οποία διασφαλίζει την ασφαλή και απρόσκοπτη ενσωμάτωση των non-3GPP συσκευών στην αρχιτεκτονική του δικτύου. Για την αντιμετώπιση του προβλήματος βέλτιστης επιλογής δικτύου, προτείνεται μια πολιτική κατανομής πόρων που στοχεύει στη ταυτόχρονη ελαχιστοποίηση των ποσοστών απόρριψης για τους αργούς χρήστες και των ποσοστών διακοπής για τους γρήγορους χρήστες. Αυτή η προσέγγιση αξιολογείται θεωρητικά μέσω ενός δισδιάστατου μοντέλου Markov, καθώς και πρακτικά μέσω υλοποίησης στην πλατφόρμα.

Τέλος, στην τρίτη συνεισφορά αναπτύσσεται ένα πλαίσιο Διαχείρισης και Ενορχήστρωσης (MANO), το οποίο έχει σχεδιαστεί ειδικά για τις λειτουργίες ενορχήστρωσης δικτύων 5G. Το πλαίσιο αυτό δίνει έμφαση στη διαχείριση του κύκλου ζωής (LCM) των 5G λειτουργιών, με στόχο τη δημιουργία ενός λειτουργικού περιβάλλοντος με ελάχιστη ανθρώπινη παρέμβαση, σε ευθυγράμμιση με το παράδειγμα της Διαχείρισης Υπηρεσιών Δικτύων Μηδενικής Παρέμβασης (ZSM). Στον πυρήνα του πλαισίου αυτού ενσωματώνεται μια μονάδα Τεχνητής Νοημοσύνης (AI) και Μηχανικής Μάθησης (ML) με δυνατότητες παρακολούθησης του δικτύου, η οποία λαμβάνει αποφάσεις σε διάφορα επίπεδα της υποδομής. Αυτή η πλατφόρμα βελτιστοποιεί την κατανομή και την ενορχήστρωση τόσο των δικτυακών όσο και των υπολογιστικών πόρων, ενισχύοντας τη συνολική αποδοτικότητα και προσαρμοστικότητα του συστήματος. Το προτεινόμενο πλαίσιο επικυρώνεται μέσω δύο περιπτώσεων χρήσης. Η πρώτη παρουσιάζει την δυναμική ανάπτυξη τεμαχίων δικτύου πάνω σε μια πολυλειτουργική πλατφόρμα 5G εργαστηριακού περιβάλλοντος βασισμένη σε λογισμικό. Η δεύτερη επιδεικνύει έναν προληπτικό μηχανισμό πρόβλεψης για τον απαιτούμενο αριθμό κόμβων UPF. Ο μηχανισμός αυτός επιτρέπει στο σύστημα να προβλέπει μελλοντικές απαιτήσεις κατανάλωσης πόρων για ένα τεμάχιο δικτύου και να προσαρμόζει δυναμικά την κατανομή πόρων ώστε να καλύπτει αποτελεσματικά αυτές τις ανάγκες του.

KEYWORDS

Δίκτυα 5ης γενιάς, Διαχείριση Κινητικότητας, Εικονικοποίηση Δικτυακών Λειτουργιών, Δίκτυα Μηδενικής Παρέμβασης, Υλοποίηση Δικτύων σε Υπολογιστικά Νέφη

ABSTRACT

The advent of the 5th generation (5G) and beyond mobile networks will play a key role in addressing fundamental challenges related to a sustainable societal transformation and economic growth. These networks require a design that will support new service capabilities, enabling connectivity for a large number of devices, ubiquitous access with increased levels of mobility in heterogeneous environments and mission critical services with stringent performance requirements that need to be met in a cost-effective and energy-efficient way. However, to support these demanding and diverse services a paradigm shift is necessary that will enable migration from closed purposely developed infrastructures into open and elastic ecosystems. This can be achieved adopting flexible architectural models, and facilitating convergence of compute and network technologies. To this end, this thesis aims to exploit various architectural enhancements and technologies adopted from 5G/Beyond 5G (B5G) networks to design and deploy systems that can efficiently provide services in environments with varying levels of mobility. More specifically, it supports the idea of a universal 5G platform hosted in a cloud infrastructure that comprises interconnected compute, network and storage components. This platform is enhanced with suitable mathematical models, protocols and algorithms that are adapted and implemented with the aim to optimize the overall system performance.

The thesis includes three distinct contributions that are implemented and experimentally validated through the use of an open-source toolset. The first contribution examines the architectural options provided by 5G/B5G networks and proposes efficient resource allocation strategies to enable optimal service delivery while addressing user mobility requirements. It introduces optimal resource management across various network segments, including the Radio Access Network (RAN), Core Network (CN), and Data Network (DN). These resources are virtualized and disaggregated to enhance flexibility in Network Function (NF) placement. To accommodate user mobility, this approach incorporates the feature of live Virtual Machine (VM) migration, which allows VMs hosting critical 5G functionalities to relocate dynamically as users move, ensuring uninterrupted service delivery. Furthermore, a multistage optimization framework is developed to achieve optimal allocation of both network and computational resources while minimizing operational costs. The analysis is supported by real-world mobility statistics and lab-based profiling measurements obtained from a 5G cloud testbed. The findings highlight trade-offs between latency and infrastructure costs, identifying optimal operational points for various scenarios.

The second contribution focuses on the development of a multi-access connectivity system that integrates both 3GPP and non-3GPP access network technologies to support a diverse range of mobile User Equipment (UE). In this system, UEs with varying mobility patterns connect to the network, requesting the establishment of end-to-end (E2E) slices with the DN. When UEs connect via the 3GPP access network, their E2E slices are formed along paths that interconnect gNodeBs (gNBs), User Plane Functions (UPFs), and DN nodes. Conversely, for UEs opting for non-3GPP networks, connectivity to the DN is achieved through the Non-3GPP Interworking Function (N3IWF), which ensures secure communication and seamless integration into the network architecture. To address the optimal network selection problem, a resource allocation policy is proposed that aims to jointly minimize blocking rates for slow-moving UEs and dropping rates for fast-moving UEs. This approach is analytically

evaluated using a two-dimensional Markov chain model. Additionally, the system's performance is validated through the cloud-based 5G testbed, providing key insights into E2E performance and resource consumption.

Finally, in the third contribution a Management and Orchestration (MANO) framework is developed which is specifically designed for the orchestration operations of 5G networks. This framework emphasizes LifeCycle Management (LCM) of 5G components to establish an operational environment with minimal human intervention or manual configuration, aligned with the Zero-touch network (ZTN) Service Management (ZSM) paradigm. At the core of this framework an Artificial Intelligence (AI)/Machine Learning (ML) block with monitoring capabilities is present, performing decision-making across various layers of the infrastructure. This platform optimizes the allocation and orchestration of both networking and edge/cloud computing virtual resources, enhancing the overall efficiency and adaptability of the system. The proposed framework is validated through two use-cases. The first demonstrates the dynamic deployment of network slices on a softwarized, multi-operator, lab-based 5G platform. The second showcases a proactive User Plane Function (UPF) provisioning mechanism, which enables the system to predict future compute and network demands of a slice and dynamically adjust resource allocation to meet those demands effectively.

KEYWORDS

5G, B5G, Mobility Management, NFV, ZSM, Cloud Mobile Networks

PUBLICATIONS

This thesis is based on the following papers:

I. A. -I. Manolopoulos, M. P. Anastasopoulos, V. -M. Alevizaki and A. Tzanakaki, "Optimal Service Provisioning in Mobile 5G and Beyond Systems," in *IEEE Transactions on Services Computing*, vol. 16, no. 4, pp. 2841-2854, 1 July-Aug. 2023, doi: 10.1109/TSC.2022.3225011.

II. A. -I. Manolopoulos, V. M. Alevizaki, M. Anastasopoulos, and A. Tzanakaki, "Demonstration of Multi-Access 6G Networks with User Mobility Considerations," in *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, October 2024.

III. A. -I. Manolopoulos, V. -M. Alevizaki, M. Anastasopoulos and A. Tzanakaki, "An AI-Assisted Framework for Lifecycle Management of Beyond 5G Services," in *IEEE Access*, doi: 10.1109/ACCESS.2024.3507359.

IV. A. Tzanakaki, A. -I. Manolopoulos, V. -M. Alevizaki and M. Anastasopoulos, "Optimized and dynamic resource provisioning in AI assisted 6G networks," *2023 IEEE Future Networks World Forum (FNWF)*, Baltimore, MD, USA, 2023, pp. 1-6, doi: 10.1109/FNWF58287.2023.10520569.

V. P. Georgiadis, M. Anastasopoulos, A. -. Manolopoulos, V. -. Alevizaki, N. Nikaein, and A. Tzanakaki, "Demonstration of Energy Efficient Optimization in Beyond 5G Systems supported by Optical Transport Networks," in Optical Fiber Communication Conference (OFC) 2023, Technical Digest Series (Optica Publishing Group, 2023), paper W4F.4.

VI. V. -M. Alevizaki, M. Anastasopoulos, A. -I. Manolopoulos and A. Tzanakaki, "Distributed Service Provisioning for Disaggregated 6G Network Infrastructures," in *IEEE Transactions on Network and Service Management*, vol. 20, no. 1, pp. 120-137, March 2023, doi: 10.1109/TNSM.2022.3211097.

VII. A. Tzanakaki, A. Manolopoulos, M. Anastasopoulos, and D. Simenidou, "Optical Networking in Support of User Plane Functions in 5G Systems and Beyond," in Photonics in Switching and Computing 2021, W. Bogaerts, K. Morito, S. Ben Yoo, M. Fiorentino, K. Ishii, and B. Offrein, eds., OSA Technical Digest (Optica Publishing Group, 2021), paper W2B.3.

VIII. V. M. Alevizaki, A. I. Manolopoulos, M. Anastasopoulos and A. Tzanakaki, "Dynamic User Plane Function Allocation in 5G Networks enabled by Optical Network Nodes," *2021 European Conference on Optical Communication (ECOC)*, Bordeaux, France, 2021, pp. 1-4, doi: 10.1109/ECOC52684.2021.9606154.

IX. M. Anastasopoulos, A. Pelekanou, A. Manolopoulos, A. Tzanakaki and D. Simeonidou, "Optical Networks in Support of Open-RAN in 5G Systems and Beyond," *2021 European Conference on Optical Communication (ECOC)*, Bordeaux, France, 2021, pp. 1-4, doi: 10.1109/ECOC52684.2021.9605967.

X. A. Tzanakaki, M. Anastasopoulos, A. Manolopoulos and D. Simeonidou, "Mobility aware Dynamic Resource management in 5G Systems and Beyond," *2021 International Conference on Optical Network Design and Modeling (ONDM)*, Gothenburg, Sweden, 2021, pp. 1-3, doi: 10.23919/ONDM51796.2021.9492515.

ACKNOWLEDGMENTS

The completion of this thesis marks the closure of a significant chapter that started for me five years ago. Several people have played an important role in this journey, so I feel the need to mention them.

First and foremost, I would like to express my deep gratitude to my supervisor, Associate Professor Anna Tzanakaki, initially for offering me the opportunity to carry my PhD, and subsequently for the motivation, guidance and the support she provided to me throughout this journey.

I would also like to extend my gratitude to Associate Professor Markos Anastasopoulos, whose comments and remarks were crucial in the development of my work.

Furthermore, I would like to sincerely thank my colleague Dr. Victoria Alevizaki for the excellent collaboration, discussions and advices from her side. Similarly, I would like to thank the other members of the team, Petros Georgiadis and Ilias Floudas, for their collaboration.

Finally, I would like to thank my family for their everlasting love and continuous support in each stage of my life!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ACRONYM	DESCRIPTION
3GPP	3rd Generation Partnership Project
4G	Fourth Generation Telecommunication systems
5G	Fifth Generation Mobile Networks
5GC	5G Core
5GMM	5G Mobility Management
5G-RG	5G Residential Gateway
5GS	5G Systems
5GSM	5G Session Management
ADAES	Application Data Analytics Enablement Service
AF	Application Function
AI	Artificial Intelligence
AKA	Authentication Key Agreement
AMF	Access and Mobility Management Function
AN	Access Network
ANLF	Analytics Logical Function
AP	Access Point
API	Application Programmable Interface
AR	Augmented Reality
AS	Application Server
ASICS	Application Specific Integrated Circuits
AUSF	Authentication Server Function
BAR	Buffer Action Rules
BBU	Baseband Unit
BH	Backhaul
BS	Base Station
B-UPF	Branching-UPF
CC	Central Cloud
CLI	Command Line Interface
CMM	Centralized Mobility Management
CN	Core Network
CNN	Convolutional Neural Network
COTS	Commercial-Off-The-Shelf
CP	Control Plane
CPU	Central Processing Unit
CU	Centralized Unit
CUPS	Control User Plane Separation
DC	Data Center
DFE	Digital Front End
DHCP	Dynamic Host Configuration Protocol
DMM	Distributed Mobility Management
DN	Data Network
DNS	Domain Name Server
DRB	Data Radio Bearer
DSCP	Differentiated Services Code Point
DU	Distributed Unit
E2E	End-to-end
EAP	Extensible Authentication Protocol
EGMF	Exposure Governance Management Function
EMBB	Enhanced Mobile Broadband
EPC	Enhaned Packet Core
EPS	Enhanced Packet System

ETSI	European Telecommunications Standards Institute
FAPI	Femto API
FAR	Forwarding Rules
FH	Fronhaul
FL	Federated Learning
FN-RG	Fixed Network Residential Gateway
FPGA	Field Programmable Gate Array
FWA	Fixed Wireless Access
GRE	Generic Encapsulation Protocol
GTP-U	GPRS Tunneling Protocol-User Plane
HO	Handover
HTTP	Hypertext Transfer Protocol
IAAS	Infrastructure as a Service
ICT	Information Communication Technology
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IKE	Internet Key Exchange
ILP	Integer Linear Programming
IM	Information Model
IOT	Internet of Things
IP	Internet Protocol
IPAM	IP Address Management
IPSEC	IP Security
ITU	International Telecommunication Union
I-UPF	Intermediate-UPF
KPI	Key Performance Indicator
LADN	Local Area Data Network
LADN	Local Area Data Network
LBO	Local Breakout
LCM	Lifecycle Management
LDAP	Lightweight Directory Access Protocol
LPP	Lightweight Presentation Protocol
LR	Linear Regression
LSTM	Long Short-Term Memory
LTE	Long Term Evolution
LXC	Linux Container
MAC	Medium Access Control
MANO	Management and Orchestration
MC	Markov Chain
MDA	Management Data Analytics
MDP	Markov Decision Process
MEC	Multi-access Edge Computing
MICO	Mobile Initiated Connection Only
MILP	Mixed Integer Linear Programming
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MMTC	massive Machine Type Communications
MMWAVE	Millimeter Wave
MNO	Mobile Network Operator
MPLS	Multiprotocol Label Switching
MTLF	Model Training Logical Function
N3IWF	Non-3GPP Interworking Function
N5C	Non-5G Capable
N5CW	Non-5G Capable over WLAN
NAAS	Network as a Service

NAS	Non-Access Stratum
NBI	Northbound Interface
NEAR-RT-RIC	Near-Real Time-RIC
NEF	Network Exposure Function
NF	Network Function
NFV	Network Function Virtualization
NFVI	NFV Infrastructure
NFVO	NFV Orchestrator
NG	Next Generation
NGAP	NG Application Protocol
NGMN	Next Generation Mobile Network
NGMN	Next Generation Mobile Network
NN	Neural Network
NON-RT-RIC	Non-Real Time RIC
NPN	Non-Public Network
NR	New Radio
NRF	Network Repository Function
NSA	Non-Standalone
NSI	Network Slice Instance
NSSF	Network Slice Selection Function
NSSI	Network Slice Subnet Instance
NST	Network Slice Template
NTP	Network Time Protocol
NWDAF	Network Data Analytics Function
O/D	Origin/Destination
OAI	OpenAirInterface
OAM	Operations, Administration, Maintenance
ONAP	Open Network Automation Platform
O-RAN	Open-RAN
OS	Operating System
OSM	Open Source MANO
PAAS	Platform as a Service
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PDR	Packet Detection Rules
PDU	Packet Data Unit
PFCP	Packet Forwarding Control Protocol
PLR	Packet Loss Ratio
PNF	Physical Network Function
PON	Passive Optical Network
PSA-UPF	PDU Session Anchor-UPF
QER	QoS Enforcement Rules
QFI	QoS Flow Identifier
QOE	Quality of Experience
QOS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RBAC	Role-Based Access Control
RFC	Request For Comments
RFSP	RAT Frequency Selection Priority
RIC	RAN Intelligent Controller
RL	Resource Layer
RLC	Radio Link Control
RNN	Recurrent Neural Network

RQI	Reflective QoS Indicator
RRC	Radio Resource Control
RRM	Radio Resource Management
RU	Remote Unit
SA	Security Association
SAAS	Software as a Service
SBA	Service-Based Architecture
SBI	Service-Based Interfaces
SCTP	Stream Control Transmission Protocol
SD	Service Descriptor
SDAP	Service Data Adaptation Protocol
SDF	Service Data Flow
SDN	Software Defined Networking
SDO	Standard Developing Organization
SEAL	Service Enabler Architecture Layer
SI	Service Instance
SI	Service Instance
SIL	Service Instance Layer
SLA	Service Level Agreement
SMF	Session Management Function
SMO	Service Management Orchestration
SMS	Short Message Service
SMSF	SMS Function
S-NSSAI	Single-Network Slice Selection Assistance Information
SOR	Steering of Roaming
SSH	Secure Shell
SST	Service Slice Type
TCP	Transmission Control Protocol
TEID	Tunnel Endpoint Identifier
TLS	Transport Layer Security
TNAN	Trusted Non-3GPP Access Network
TNAP	Trusted Non-3GPP Access Point
TNGF	Trusted Non-3GPP Gateway Function
TSDB	Time Series Database
TWAN	Trusted WLAN Access Network
TWIF	Trusted WLAN Interworking Function
UC	Use Case
UDM	Unidied Data Management
UDP	User Datagram Protocol
UDR	Unified Data Repository
UE	User Equipment
UI	User Interface
UP	User Plane
UPF	User Plane Function
URLLC	Ultra Reliable Low Latency Communications
URR	Usage Report Rules
URSP	UE Route Selection Policy
VBBU	Virtualized BBU
VCA	VNF Configuration and Abstraction
VCA	VNF Configuration Abstraction
VIM	Virtual Infrastructure Manager
VLD	Virtual Link Descriptor
VM	Virtual Machine
VMM	Virtual Machine Manager
VNF	Virtual Network Function

VNFD	VNF Descriptor
VR	Virtual Reality
VSAT	Very Short Aperture Link
W-5GAB	Wireline 5G Access Network
W-5GCAB	Wireline 5G Cable Access Network
W-AGF	Wireline Access Gateway Function
WDM	Wavelength Division Multiplexing
WLAN	Wireless Local Area Network
WLAN	Wireless Local Area Network
YAML	YAML ain't Markup Language
ZSM	ZTN Service Management
ZTN	Zero Touch Network

Chapter 1. INTRODUCTION

1.1. Motivation and Problem Statement

Digital technologies have been identified as key in addressing fundamental challenges associated with societal and economic objectives, such as improved quality of living for citizens, sustainable development and economic growth. In this context, the 5th generation (5G) and beyond (B5G) infrastructures will play a fundamental role in bringing these technologies to society, transforming everyday life in the way services are provided, and businesses are run. However, this transformation requires new service capabilities that networks need to support including: i) connectivity for a growing number of very diverse devices, ii) ubiquitous access with varying degrees of mobility in heterogeneous environments and, iii) mission critical services, supporting highly variable performance attributes in a cost and energy-efficient manner.

Figure 1. 1: Vison for 5G Systems [1]

In contrast to 4G technologies, such as Long Term Evolution (LTE), which primarily focused on improving the performance of previous technology generations (2G and 3G), 5G systems (5GS) introduce significant architectural changes that target to revolutionize mobile communication systems and their capabilities. More specifically, 5G encompasses a wide variety of mobile broadband services, applications, and user-related wearable devices, enabling an unprecedented level of human and machine connectivity [1], as shown in Figure 1. 1. Typical examples of this next-level interaction include Augmented Reality (AR) and Virtual Reality (VR) applications, which are expected to redefine entertainment and professional environments. Internet of Things (IoT) applications are also expected to significantly influence everyday life, supporting a variety of services and use cases including smart homes and cities, cloud-based offices,

and immersive entertainment [2]. Furthermore, 5G aligns seamlessly with the 4th Industrial Revolution, a transformative era characterized by the fusion of physical, digital, and biological spheres. It expands the notion of "customer" by driving advancements across vertical sectors, including transportation, industries, healthcare, agriculture, entertainment, education, finance, and environmental sustainability. In the transportation domain, for example, 5G enables innovations such as collaborative driving, enhanced road user protection mechanisms, and improved efficiency in railroad systems. It also supports mining and construction industries by allowing remote control of vehicles and machinery in hazardous or hard-to-reach areas. In addition, the healthcare sector can be significantly benefited through the advent of wirelessly enabled smart pharmaceuticals and the potential for remote surgeries using haptic feedback technologies. Similarly, agriculture can leverage IoT and big data analytics facilitated by 5G technologies to track, monitor, and optimize farming operations, leading to higher-quality products, increased yields, and reduced waste. Overall, it can be adopted to improve energy efficiency, optimize transportation systems, enhance environmental monitoring, and implement advanced waste management techniques in smart city environments.

The standardization process for 5G networks was developed through the cooperation of several bodies and organizations such as the 3rd Generation Partnership Project (3GPP), the Internet Engineering Task Force (IETF) and European Telecommunications Standards Institute (ETSI). In 2018, 3GPP completed the Release 15 [3] which specified the first full set of standards for 5G standalone networks. This release provided the fundamental architecture of 5G and addressed a set of use-cases summarized below [4]:

- **Enhanced Mobile Broadband (eMBB):** This class of use cases represents an evolution of the data access services provided by 4G with significantly improved user experience and performance. eMBB focuses on delivering seamless, high-speed connectivity anytime and anywhere, with noticeable improvements in user data rates. Use cases (UCs) within eMBB range from high-density hotspots, requiring exceptional traffic capacity and low user mobility, to wide-area coverage scenarios with medium to high mobility. Examples include multimedia streaming, cloud-based services, and enhanced VR and AR applications.

- **Ultra-Reliable and Low-Latency Communications (URLLC):** This class targets use cases with stringent requirements for mobility (up to 100 km/h), latency (as low as 1 ms), reliability, and availability (exceeding 99.999% or "five nines"). These capabilities enable 5G to support critical applications beyond the traditional Information and Communication Technology (ICT) sector, such as wireless control in Industry 4.0 manufacturing processes, remote medical operations, smart grid automation, and transportation safety systems. URLLC is pivotal for applications requiring real-time responsiveness and ultra-reliable performance.

- **Massive Machine-Type Communications (mMTC):** mMTC focuses on services involving the transmission of non-delay-sensitive data from a vast number of connected devices, typically at low data rates. The devices used for mMTC applications are designed to be cost-effective and energy-efficient, ensuring extended battery life. Prominent use cases include smart agriculture, smart metering, and logistics. For example, in agricultural settings, low-power sensors can be deployed over large areas to monitor ground humidity, soil

fertility, and other environmental conditions, enabling better resource management and productivity.

3GPP Release 16 [5], often referred to as "5G Phase 2," focuses on advancing the foundational features of the 5G system. It targets improvements across several key areas, including enhanced coverage, increased capacity, reduced latency, optimized power efficiency, improved mobility support, higher reliability, and simplified deployment. These enhancements aim to refine and expand the capabilities of 5G networks to better address diverse use cases and operational challenges. Finally, Release 17 [6] represents the third milestone in the 3GPP roadmap, aiming to further enhance the performance of 5G systems. It extends 5G's capabilities to support new devices and deployment scenarios while continuing to improve the performance and efficiency of existing 5G features.

5G is based on a flexible and modular architecture, designed to offer multi-tenancy capabilities, combined with new spectrum bands, higher peak throughput and degree of mobility, increased spectral and energy efficiency, in order to support services associated with a large variety of ecosystems. Therefore, it manages to overcome limitations of the architectural approaches adopted by previous generations, that commonly have been targeting a very specific set of requirements and a single business ecosystem. In contrast, 5G adopts a network architecture that relies on Network Function Virtualization (NFV) and Software Defined Networking (SDN) [7]. With NFV, traditional black-box network elements are converted to Network Functions (NFs) that can be deployed on Virtual Machines (VMs), containers or general purpose servers, enabling faster service delivery and cost reduction. By decoupling network functions from dedicated hardware, NFV enhances flexibility and resource allocation granularity, allowing resources to be dynamically provisioned without the need for proprietary hardware or software. This independence enables greater adaptability to evolving technologies and service requirements. SDN is a network architectural approach that decouples control and forwarding functions of the network. In this architecture, a logically centralized entity, known as the controller, manages multiple forwarding devices. This separation enables networks to become programmable and easily manageable, facilitating efficient resource allocation, improved scalability in distributed data centers, and device virtualization, all of which are essential for effective resource sharing. However, the communication between the controller and the physical devices introduces latency considerations. Therefore, careful attention must be given to the placement and size of the controller to minimize total end-to-end latency.

In this direction, 5G comes with significant transformations in both the RAN and Core part of the network:

- The 5G Radio Access Network (RAN), also known as 5G New Radio (5G NR) or Next-Generation RAN (NG-RAN), is designed to support a heterogeneous set of interfaces and ensure compatibility with the previous technologies. The 5G-RAN architecture allows a variety of deployment options spanning from a monolithic to a fully disaggregated solution facilitating the allocation of the baseband processing functions at different resources and locations in accordance to well defined and standardized functional split options [9]. In particular, the Open-RAN (O-RAN) architecture [10] defines RAN resource disaggregation into the Remote Unit (RU), the Distributed Unit (DU) and the Centralized Unit (CU), providing benefits such as reduced operational costs, improved scalability, and enhanced flexibility. Complementing this, 5G integrates multiple Radio Access Technology (RAT) connectivity, enabling user connectivity through traditional

RAN or non-3GPP access technologies such as WiFi. This integration supports higher throughput, increased reliability and different levels of mobility [13], further enhancing the adaptability and performance of the 5G network.

- The 5G Core Network (CN) comprises several NFs implemented as microservices interconnected over a Service-Based Architecture (SBA) [8]. Each of these softwarized NFs can be independently hosted over general purpose hardware. Furthermore, 5G CN enables the separation of the Control Plane (CP) from the User Plane (UP). CP functions are responsible for management and decision-making operations, while the UP is responsible for handling actual user data. Since these operations may have different levels of requirements, for example in terms of latency, their separation offers the ability to place them in different locations in the network.

Building on the advanced capabilities of the 5G architecture, network slicing emerges as a crucial mechanism to address the diverse and demanding requirements of modern applications and industries. Network slicing provides the ability to divide the physical network infrastructure into multiple, isolated logical networks, each designed to support specific use cases and/or services by interconnecting subsets of radio, core and transport network segments. In this context, Mobile Network Operators (MNOs) can use network slicing to provide customized services with slices dedicated to specific functionalities and requirements, i.e. mobility related requirements [11]. To enable the dynamic and efficient management of network slicing and other 5G capabilities, the 5G Management and Orchestration (MANO) framework serves as a cornerstone of the architecture. MANO oversees and coordinates resource allocation across the network slices, ensuring that the underlying infrastructure can meet the diverse requirements for various network service types [12]. By integrating SDN and NFV, it provides the flexibility to adapt to varying demands in real-time. MANO also monitors the slice performance, ensuring consistent quality of service (QoS) and adherence to service-level agreements (SLAs).

Finally, Artificial Intelligence (AI) /Machine Learning (ML) techniques have been increasingly adopted the past years by telecommunications industry to enhance network optimization, strengthen security, improve QoS etc. In the context of Beyond 5G (B5G)/6G networks, which are expected to generate and process vast amounts of data, these techniques will be widely adopted for converting raw data into actionable knowledge. This knowledge can drive automation, enhance service Lifecycle Management (LCM), and reshape business models and opportunities in the telecommunications sector. A typical example of AI/ML applications in mobile networks include network traffic forecasting, where AI/ML techniques have demonstrated improved performance. For example, Neural Networks (NNs) have emerged as a powerful tool for time series data prediction due to their ability to capture and model complex, non-linear relationships within the data. These advancements are crucial in ensuring that next-generation networks remain efficient, resilient, and capable of meeting evolving demands.

However, while the 5G vision introduces an ecosystem that enables openness, flexibility and elasticity for vertical industries and services, it also poses some limitations that should be taken into account. For example, the explosive growth in the number of users, as well as the support of extremely heterogeneous/diverse applications and use cases, introduce very complex user mobility patterns that may abruptly change the resources needed to ensure smooth and efficient network operation. So, mobility management becomes a crucial topic for these systems that must be carefully considered in network

design, planning and deployment. Several technologies and tools are included in the standardization process of 5GS that can assist in mobility management, but most of the current implementations treat these technologies in a segmented way. To this end, this thesis proposes and implements a unified framework that integrates 5G/B5G advanced technologies, ensuring seamless interoperability between various components, while meeting the stringent performance and scalability requirements of network applications. By doing so, it facilitates the development and evaluation of scenarios involving user mobility and dynamic service demands, offering practical insights into the design and management of virtualized mobile networks.

1.2. Thesis Focus and Contributions

This thesis addresses the complex challenge of integrating advanced 5G/B5G technologies into a unified framework that can efficiently provide services in environments with varying levels of mobility. Specifically, it explores the paradigm of virtualization and softwarization of network functionalities, with emphasis on their deployment in cloud infrastructures and data centers (DCs). By adopting resource disaggregation, this work enables flexible allocation and placement of resources across diverse locations to meet specific service requirements. Central to this research is the design and practical implementation of a universal 5G platform, hosted in a private cloud infrastructure within a lab testbed environment. The lab testbed allows a real-world evaluation of these technologies, offering valuable insights on how they can support a range of scenarios involving user mobility and dynamic service demands. Additionally, the developed framework integrates various mathematical models, protocols, and algorithms focused on optimizing system performance. Through these contributions, this research provides a comprehensive approach to deploying and managing virtualized mobile networks, aligned with the evolving needs of Next-Generation (NG) connectivity. Bellow, we discuss the contributions related to this thesis.

The first contribution explores the architectural options offered by 5G/B5G networks and proposes effective resource allocation strategies to support optimal service delivery, taking into consideration user mobility requirements. Optimal resource management is proposed for various network segments such as the RAN, CN and Data Network (DN), where resources are virtualized and disaggregated in order to enable flexibility in the placement of NFs. To address user mobility, the option of live VM migration is adopted where VMs hosting critical 5G functionalities can be migrated to different locations as the user moves, without causing any service disruptions. Moreover, a multistage optimization framework is purposely developed to enable optimal resource allocation of both network and compute resources minimizing the network operational costs. The analysis is based on real mobility statistics and lab-based profiling measurements extracted from a 5G cloud testbed. This work can be found at [14] where the observed trade-offs between latency and infrastructure-related costs determine the optimal operational points for different scenarios.

The second contribution demonstrates the development of a multi-technology RAN integrating 3GPP and non-3GPP access networks to support a diverse set of mobile User Equipments (UEs). UEs with various mobility patterns are attached to the network requesting the establishment of end-to-end slices with the DN. In case where UEs are attached to the 3GPP access network, the corresponding slices are established over paths formed by interconnecting the gNodeB (gNB), User Plane Functions (UPFs) and

the DN. However, in case where UEs select a non-3GPP network, their connectivity with the DN is achieved through the Non-3GPP Interworking Function, that ensures secure communication and facilitates smooth integration with the overall network architecture. To solve the optimal network selection problem, a resource allocation policy is proposed that tries to jointly minimize the blocking and dropping rates for slow and fast, respectively, moving UEs. The overall process is theoretically evaluated using a two-dimensional Markov Chain. Finally, the performance of the system is validated by a practical deployment in our 5G cloud testbed, providing insights into E2E performance and resource consumption. This work can be found at [15].

The third contribution aims at the development of a MANO framework that specifically targets orchestration operations of 5G networks. This framework focuses on the lifecycle management of the 5G components, in order to achieve an operational environment with minimal human intervention or manual configuration following the parading of Zero-touch network Service Management (ZSM). Within this ecosystem, an AI/ML module has comprehensive monitoring capabilities and influences decisions across various layers or aspects of the infrastructure. This includes optimizing the allocation and orchestration of both networking and edge/cloud computing virtual resources within the infrastructure. This work can be found in [16], where the validity of the proposed framework is demonstrated by two use cases. The first one focuses on the dynamic deployment of network slices on top of a softwarized multi-operator lab-based 5G platform. The second part of the implementation concentrates on the demonstration of a proactive UPF provisioning mechanism, which ensures that the system can predict future compute and network demands of a slice in order to adapt the resources accordingly.

1.3. Thesis Outline

The thesis is structured as follows:

Chapter 2 provides an overview of the 5G system architecture and involved technologies. First, we present the 5G functional architecture, followed by an extensive review of RAN, and CN segments including related protocols and interfaces. Then, we discuss various innovative technologies adopted by 5G/B5G networks, such as network slicing, MANO, AI/ML tools and key mobility management aspects.

Chapter 3 focuses on the 5G Cloud Platform. First, we discuss some background preliminaries regarding cloud computing and virtualization concepts. Then, a description of the used hardware and software is provided as well as the functional blocks of the infrastructure (private cloud platform, monitoring/profiling platform, MANO platform). Next, we present various 5G deployment options supported from the testbed, designed for specific application scenarios. Finally, we obtain some profiling results related to resource consumption (compute, network, energy).

Chapter 4 proposes suitable resource allocation schemes to enable optimal service provisioning with user mobility considerations. The proposed schemes are evaluated through a purposely developed multistage optimization framework. This evaluation framework facilitates optimal placement of network and compute resources minimizing the associated infrastructure operational costs. The results have been produced using real user mobility statistics as well as lab-based profiling measurements of the 5G infrastructure and provide a detailed trade-off analysis between latency and

infrastructure related costs indicating optimal operational points for different scenarios.

Chapter 5 concentrates on a multi-access technology system where user connectivity is offered by 3GPP and non-3GPP Access Networks (ANs) simultaneously. In this setup, users are directed to a suitable access network according to their mobility demands and the status of the network in terms of available resources. To solve the problem of optimal network selection we propose a resource allocation policy based on queuing theory.

Chapter 6 proposes a MANO framework that specifically targets the orchestration operations of B5G networks. In this framework, network slices can be dynamically deployed based on network descriptors that map the requirements for UP and CP core elements. Two use cases are presented in this direction. The first focuses on dynamic network slice deployment on top of a softwarized multi-operator environment. The second presents a mechanism that proactively provisions UPF nodes without human intervention.

Chapter 7 provides a conclusion of the thesis and a discussion related to future work.

References

[1] Marsch, P. and Bulakci, O. and Queseth, O. and Boldi, M., *5G System Design: Architectural and Functional Considerations and Long Term Research,* ISBN: 9781119425120, Wiley, 2018. Available: https://books.google.gr/books?id=QFxTDwAAQBAJ

[2] Rao, S.K., Prasad, R., "Impact of 5G Technologies on Industry 4.0.," *Wireless Personal Communications 100,* pp. 145–159, 2018. Available: https://doi.org/10.1007/s11277-018-5615-7

[3] 3GPP, *System Architecture for the 5G System (5GS); Stage 2,* 3GPP TS 23.501, Release 15, Dec. 2017. [Online]. Available: https://www.3gpp.org/

[4] Liu, G., & Jiang, D. (2016). 5G: Vision and requirements for mobile communication system towards year 2020. *Chinese Journal of Engineering*, *2016*(1), 5974586.

[5] 3GPP, *System Architecture for the 5G System (5GS); Stage 2,* 3GPP TS 23.501, Release 16, Mar. 2020. [Online]. Available: https://www.3gpp.org/

[6] 3GPP, *System Architecture for the 5G System (5GS); Stage 2,* 3GPP TS 23.501, Release 17, Jun. 2022. [Online]. Available: https://www.3gpp.org/

[7] Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., Ramos-Munoz, J. J., Lorca, J., & Folgueira, J. (2017). Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine*, *55*(5), 80-87.

[8] Rommer, S., Hedman, P., Olsson, M., Frid, L., Sultana, S., & Mulligan, C. (2019). *5G core networks: powering digitalization*. Academic Press.

[9] L. M. P. Larsen, A. Checko and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146-172, Firstquarter 2019, doi: 10.1109/COMST.2018.2868805.

[10] M. Polese, L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376-1411, Secondquarter 2023, doi: 10.1109/COMST.2023.3239220.

[11] R. Wen, G. Feng, J. Zhou and S. Qin, "Mobility Management for Network Slicing Based 5G Networks," *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, Chongqing, China, 2018, pp. 291-296, doi: 10.1109/ICCT.2018.8600026.

[12] H. -M. Chen, Y. -F. Lu, S. -Y. Chen, C. -J. Chang and Z. -X. Zheng, "Design of an NFV MANO Architecture for 5G Private Network with 5G CN Cloud-Edge Collaborative Mechanism," *2022 8th International Conference on Applied System Innovation (ICASI)*, Nantou, Taiwan, 2022, pp. 92-95, doi: 10.1109/ICASI55125.2022.9774446.

[13] S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard and D. Chandramouli, "5G multi-RAT multi-connectivity architecture," *2016 IEEE International Conference on Communications Workshops (ICC)*, Kuala Lumpur, Malaysia, 2016, pp. 180-186, doi: 10.1109/ICCW.2016.7503785.

[14] A. -I. Manolopoulos, M. P. Anastasopoulos, V. -M. Alevizaki and A. Tzanakaki, "Optimal Service Provisioning in Mobile 5G and Beyond Systems," in *IEEE Transactions on Services Computing*, vol. 16, no. 4, pp. 2841-2854, 1 July-Aug. 2023, doi: 10.1109/TSC.2022.3225011.

[15] A. I. Manolopoulos, V. M. Alevizaki, M. Anastasopoulos, and A. Tzanakaki, "Demonstration of Multi-Access 6G Networks with User Mobility Considerations," in *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, October 2024.

[16] A. I. Manolopoulos, V. M. Alevizaki, M. Anastasopoulos, and A. Tzanakaki, “An AI-Assisted Framework for Lifecycle Management of Beyond 5G Services” in *IEEE Access*, submitted: November 2024.

Chapter 2. 5G SYSTEM ARCHITECTURE & TECHNOLOGIES

2.1. Chapter Introduction

5G networks are now in a commercial roll-out in many countries across the globe [1]. 5G marks a significant leap in the telecommunications sector, aiming to reshape the global connectivity landscape. Unlike previous generations, 5G introduces a highly versatile and dynamic ecosystem designed to support a broad range of applications including various industries. This ecosystem includes mobile network operators, equipment vendors, cloud providers, application developers, and end-users, all interconnected through a flexible, high-performance network architecture [2]. In this context, the vision for 5G networks is to provide faster, more reliable, and scalable communications, while enabling new business models and digital services. Furthermore, the architectural design discards the traditional monolithic network model, adopting a sophisticated, flexible and service-driven approach that strongly relies on concepts such as virtualization, softwarization and edge computing.

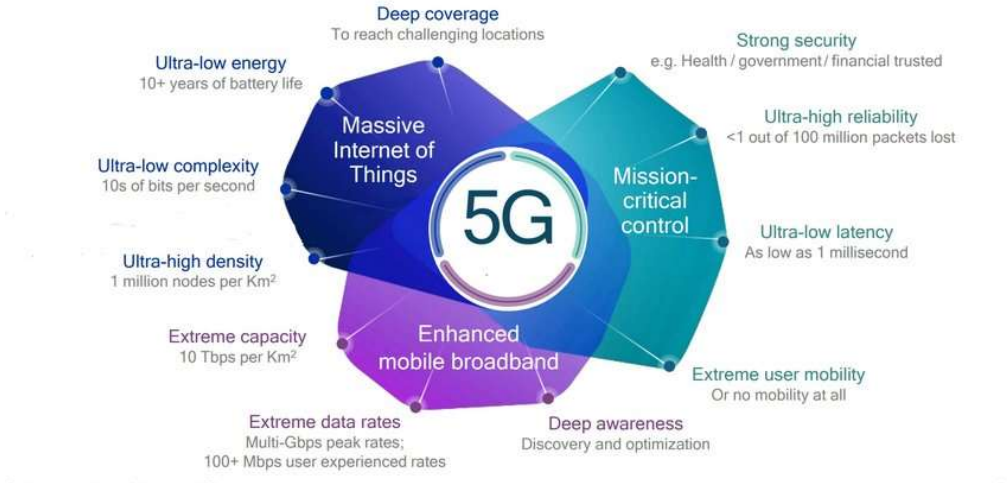


Figure 2. 1: 5G System Use Cases and related KPIs [4]

To enable the vertical applications (eMBB, mMTC, URLLC) defined by the ecosystem, 5G networks are built around several Key Performance Indicators (KPIs) that define their capabilities and performance. These include data rates, latency, energy efficiency, reliability and coverage aspects among others [4], as illustrated in Figure 2. 1.

The development and deployment of 5G networks are driven by standardization bodies and global collaborative efforts. 3GPP [3] is the leading organization responsible for developing technical specifications for 5G, covering aspects such as radio access networks, core networks, and security. The International Telecommunication Union (ITU) [5] sets global spectrum regulations, ensuring that 5G can operate across a variety

of frequency bands. In parallel, organizations like ETSI [6] and the O-RAN Alliance [7] play critical roles in promoting network virtualization, edge computing, and open, interoperable radio access networks.

The rest of the chapter is structured as follows. First, we present an overview of the overall network architecture. Then, each block and component of the network architecture is described in detail. Finally, some of the most innovative technologies and concepts adopted by 5G networks are identified.

2.2. Network Architecture Overview

The overall E2E architecture for 5GS introduces several components and a variety of deployment options, as illustrated in Figure 2. 2. The architecture is built upon the principles of softwarization and virtualization, reflecting the vision for 5GS. The RAN segment relies on the concept of resource disaggregation which involves the separation of the functionalities from the underlying components and placing them at discrete geographical locations. This design enables to potentially split the Baseband Unit (BBU) processing chain into three logical entities: The Remote Unit (RU) the Distributed Unit (DU) and the Central Unit (CU) [8]. Depending on the scenario under consideration, these units can either be co-located or separated and placed at different locations of the network. While the monolithic setup resembles the LTE architecture and offers backward compatibility with the Enhanced Packet Core (EPC), disaggregated deployments allow further optimization for the RAN segment, for example by enabling the split of the CP from the UP. Furthermore, various advanced technologies for the radio segment are introduced, such millimeter wave (mmWave) communications, massive Multiple-Input Multiple- Output (MIMO), as well as the installation of dedicated antennas, such as macro or small cells. Macrocells are high-power base stations that provide wide-area coverage and form the backbone of cellular networks. In contrast, small cells are low-power, short-range base stations designed to extend coverage and improve network capacity [9]. 5G networks depend on a dense deployment of small cells to deliver high-speed, low-latency connectivity, especially in areas where macro cells struggle, such as indoors or densely populated urban zones. Small cells help reduce network congestion, improve performance, and support real-time applications like gaming and AR/VR due to their lower latency. They also consume less power, enhancing network efficiency and reducing costs for operators. However, deploying small cells is expensive and requires significant infrastructure investments, ongoing maintenance, and careful management to avoid interference with macro cells. Finding suitable deployment locations is another challenge, as it often requires access to public or private property and approval from local authorities. Despite these hurdles, small cells will play an increasingly important role as 5G evolves, ensuring high-quality connectivity for end-users.

Apart from the traditional RAN connectivity, the architecture supports convergence of 5G systems with non-3GPP access networks [10], such as WiFi. This integration provides effective solutions that aim to relieve data traffic congestion and address

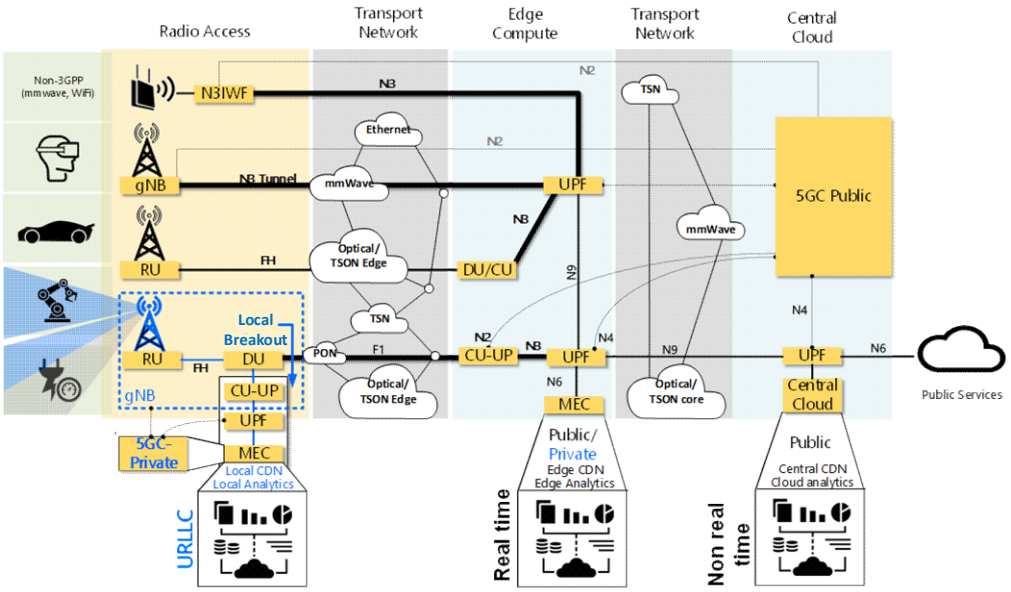


Figure 2. 2: Overall Network Architecture

capacity and coverage aspects that are crucial in an environment that needs to handle the massive growth of IoT devices as well as industrial communication.

Additionally, a high-capacity, flexible optical transport network combined with mmWave links offers a promising solution for 5G backhaul (BH) and fronthaul (FH) connectivity [11]. While mmWave links are easier and cheaper to deploy than optical fibers, they provide lower capacity and are more suitable for backhauling small cell sites with modest data rates. To meet the high-capacity demands and flexibility for traffic aggregation, hybrid solutions, such as passive optical networks (PONs) exploiting wavelength division multiplexing (WDM) and dynamic, elastic optical networks, are ideal [12].

The 5G CN design is based on two fundamentals: the separation of the UP from the CP (CUPS), and the adoption of a SBA for the involved core entities. CUPS allows the placement of the UPF, which is responsible for handling user traffic, in any location of the network. The UPF can thus operate closer to the user site in specific scenarios such as delay sensitive applications and specific QoS requirements, or complex mobility patterns. At the same time, CP functionalities can be located at a central location. SBA refers to the design of the functionalities as small, independent services that communicate via well-defined Application Programming Interfaces (APIs). This approach enables each service to scale or update independently without disrupting other parts of the application. 5G microservices allow the creation of logical architectures tailored to the specific performance and functional requirements of various use cases [13]. The core idea is to break down traditionally monolithic NFs, which were often tied to physical network elements in previous cellular systems, into modular components with sufficient granularity. This modularization of 5G network functions separates those associated with the AN from those of the core CN, reducing the dependency between the 5G core and the access network. This separation enables new forms of connectivity beyond traditional cellular radio.

To support service processing needs and application requirements, the network architecture provides compute and storage resources, utilizing SDN and NFV. These resources are divided into two key classes:

- **Centralized Data Center (Central Cloud - CC):** High-performance servers and storage at a centralized location handle core functions like authentication, billing, and network management.
- **Multi-Access Edge Computing (MEC):** Smaller data centers located near the network edge, closer to end-users, reduce latency and jitter. MEC supports low-latency applications like AR/VR, autonomous vehicles, and industrial automation, and can host virtualized baseband units (vBBUs) for RAN functions. Fast, low-latency FH connections between MEC and RUs are essential, often requiring speeds of 1-10 Gbps per antenna. Efficient use of these connections, like multiplexing, and minimizing Ethernet switching, are critical to maintaining performance.

In this context, with the complex and sophisticated network architecture on the one hand, and the extreme growth of end devices on the other hand, efficient mobility management becomes necessary in order to ensure a smooth operation and performance for the ecosystem. To this end, the adoption of various tools and techniques such as Network Slicing, as well as the insertion of automation and AI/ML tools in B5G networks can provide effective solutions for handling user mobility.

2.2.1. 5G Core Network

The 5G Core (5GC) is designed to support all the advanced requirements and capabilities of 5G networks. 5GC follows a Service-based architecture for the interconnection of NFs where interfaces are based on standardized protocols that allow for coordinated interaction between the NFs [15].

Furthermore, it strongly relies on virtualization and softwarization which decouples the functionalities from the underlying hardware/infrastructure. This way, the 5GC can benefit from the advantages that cloudified and cloud-native deployments offer.

Another fundamental aspect of 5GC is the separations of the UP from the CP. This way, the UPF which is the NF that handles user data, can be placed closer to the users in order to support time-critical services and applications, whereas the CP NFs operate from a CC position.

Moreover, the combination of all the above mentioned technologies enables the realization of more complex deployment options, e.g. topologies with increased UPF nodes. Overall, design enables flexibility, scalability and efficiency in the delivery of 5G services and applications. The main elements in the 5GC, as illustrated in Figure 2. 3, are:

AMF (Access and Mobility Management Function): It handles mobility and UE connectivity aspects and interacts with other NFs to provide seamless mobility across different access networks. It is involved in most of the signaling call flows in the network and allows device registration and authentication through encrypted signaling.

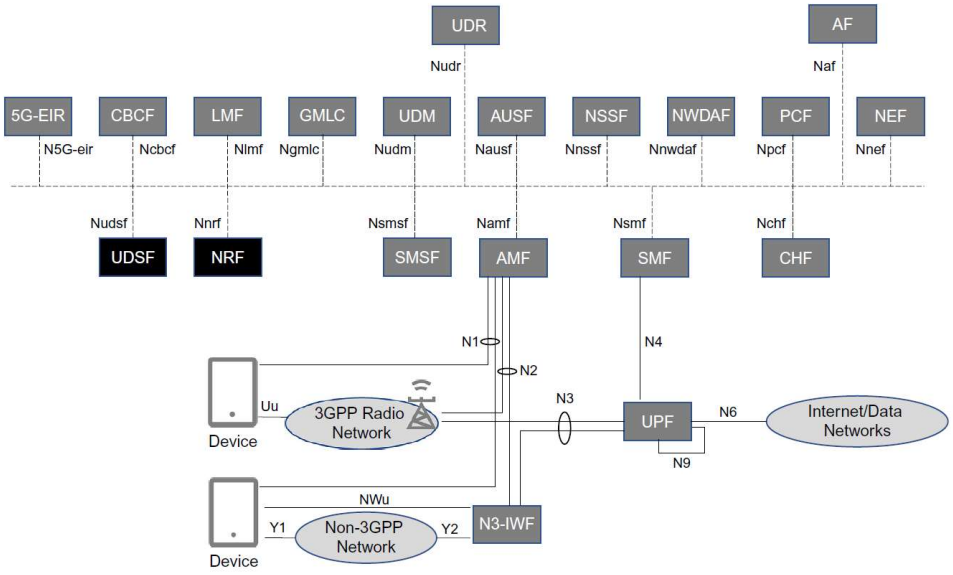


Figure 2. 3: 5G Core Network Architecture [15]

SMF (Session Management Function): SMF is responsible for the establishment, modification and release of the Packet Data Unit (PDU) session, i.e. the path between the UE and the Data Network. It also performs data plane management actions such as IP address allocation, traffic routing. Additionally, it interacts with the Policy Control Function (PCF) to enforce policies onto the PDU sessions.

UPF: The main task of the UPF is to process and forward all user data, based on rules that it retrieves through its interaction with the SMF. It also acts as a stable anchor point with external Internet Protocol (IP) networks, thus hiding device mobility. Packets with a destination address that belongs to a specific device are always routable from the internet to a specific UPF even when the UE is moving to another cell. Moreover, different types of processing are performed from the UPF in the forwarded data. Other functionalities include traffic usage reports for the SMF and charging reports that are destined for other NFs, packet inspection used for policy decisions and traffic reporting, traffic redirection, data rate limitations etc. Moreover, it applies QoS rules such as prioritization of packets in scenarios where the network is congested. Even though the UPF is defined as one element, the standard outlines several functional roles that the UPF can perform along the user plane path. These roles are:

- Branching Point (B-UPF): This is a specialized UPF responsible for classifying and directing uplink traffic, enabling traffic steering or branching to different network functions or paths.
- PDU Session Anchor (PSA-UPF): The PSA is the node that terminates the N6 interface, providing the connection to the DN.
- Intermediate (I-UPF): It is a UPF positioned on the user plane path between the (R)AN and the PSA. Its role is to forward traffic between the (R)AN and the PSA.

PCF : It handles network policy decisions, including policy rules, QoS management, and UE access control. The PCF interacts with other NFs such as the SMF and AMF to

enforce policy decisions and to ensure that network resources are used according to predefined policies. For the SMF, it provides QoS and charging control for Service Data Flows (SDFs) and it performs policy control and Packet Data Unit (PDU) session event reporting. For the AMF, it supports access and mobility policy control, managing the RAT Frequency Selection Priority (RFSP) index and service area restrictions. The PCF also interacts with the UEs through the AMF providing policy information, such as discovery, network slice selection and session continuity mode selection.

UDM (Unified Data Management): The UDM is used to handle subscription and authentication data, usually through interacting with the AMF and SMF where it provides user information in order to ensure service continuity and personalized service delivery. Additionally, in case where more than one SMF and AMF nodes exist in the network, it keeps track of which instance serves each device.

UDR (Unified Data Repository): It is a typical database which stores subscription data, as well as various types of network and user policies. These data are usually retrieved from other NFs such as the UDM, AMF and SMF.

NSSF (Network Slice Selection Function): 5G Systems introduce the concept of Network Slicing, where the physical network can be divided in multiple isolated logical networks that are used for specific purposes and different use cases (eMBB, URLLC, mMTC). The NSSF is responsible for handling all the information and parameters related to each network slice, as well as for directing the AMF to the appropriate slice based on the UE's subscription and service requirements.

AUSF (Authentication Server Function): The functionality of the AUSF is related to the authentication of devices, utilizing the credentials provided by the UDM. It supports various authentication methods, including 5G Authentication and Key Agreement (5G-AKA) and Extensible Authentication Protocol (EAP)-based schemes. Additionally, it is responsible for generating cryptographical material to ensure secure updated of roaming information and other parameters related to the UE.

NEF (Network Exposure Function): The NEF is responsible for providing a secure interface that exposes specific events and capabilities from the 5G system to either the NFs of the network or to a third-party network. These capabilities include policy control, service quality, analytics, UE location, reachability, roaming status and more. The NEF basically acts as a mediator, ensuring that external requests are authorized and compliant with network policies. Applications that are authorized by the network can use the NEF to satisfy various requests.

NRF (Network Repository Function): It acts as a directory service that keeps the profiles of all NFs of the 5G Core allowing them to discover and communicate with each other. This way, the NRF facilitates dynamic service orchestration and load balancing. When a specific NF needs to interact with another, it goes through the repository in order to obtain its profile. Newly deployed or altered NFs will have to report their status to the NRF, in order for them to be reachable from other NFs.

AF (Application Function): It represents the application layer in the 5G architecture, interfacing with the core network to request specific QoS or policy enforcement for trusted applications that are located either inside or outside the operator's network. The AF is often used by applications requiring network resources for optimized service delivery. AFs can interact with the 5GC NFs either directly, or through the NEF.

N3IWF (Non-3GPP Interworking Function): Introduced in 3GPP Rel-15, it is a critical component designed to enable seamless integration between non-3GPP access technologies (such as Wi-Fi) and the 5G CN. Its most important functionality is to provide secure connectivity for untrusted ANs.

2.2.1.1. Protocols & Interfaces

NG Application Protocol (NGAP): NGAP is designed to be used over the N2 interface between the RAN and AMF. It supports all the necessary mechanisms to handle RAN-AMF procedures as well as transparent transport procedures between the UE and AMF or other 5GCN functions. Moreover, NGAP can be applied both to 3GPP and non-3GPP access technologies integrated with the Core network. According to TS 38.413 [16], NGAP supports the following capabilities:

- procedures to establish, maintain and release NG-RAN part of PDU sessions.
- procedures to perform intra-RAT handover and inter-RAT handover.
- the separation of each UE on the protocol level for user specific signaling management.
- the transfer of NAS signaling messages between UE and AMF.
- mechanisms for resource reservation for packet data streams.

Additionally, NGAP consists of set of Elementary Procedures/Signaling messages.

Packet Forwarding Control Protocol (PFCP): PFCP is used between the SMF and UPF [17] over the N4 interface to perform controlling actions on the UPF, so it basically implements CUPS. PFCP runs over the User Datagram Protocol (UDP) protocol, and it consists of node related (PFCP Association setup/Update/Release etc.) or session related (PFCP Session Establishment/Modification/Deletion) messages.

GPRS Tunneling Protocol for User Plane (GTP-U): It uses a tunnel mechanism to carry UP data over the N3 (UE-UPF) and N9 (UPF-UPF) interface [18]. It operates over the UDP protocol. A GTP-U tunnel is used between two nodes to separate traffic into different communication flows. Each tunnel is uniquely identified by a local Tunnel Endpoint Identifier (TEID), an IP address and a UDP port.

5G Non-Access-Stratum (5G NAS): It refers to the primary CP protocols between the UE and the core network. NAS has several key functions that are essential to 5G operations[15]. These include managing UE registration and mobility, which entails core access control tasks such as connection management, authentication, NAS security handling, and both UE identification and configuration. Additionally, NAS supports session management procedures necessary to establish and maintain PDU Session connectivity, ensuring QoS for the User Plane between the UE and the DN. NAS also serves as a general transport layer for messages exchanged between the UE and the AMF, even for messages not strictly defined within the NAS protocol. These messages may include Short Message Service (SMS), location services using the Lightweight Presentation Protocol (LPP) protocol, UDM data like Steering of Roaming (SOR) messages, and policy information related to UE-specific rules UE like Route Selection Policy (URSP). To support these functions, NAS comprises two main protocols: the 5GS Mobility Management (5GMM) protocol and the 5GS Session Management (5GSM) protocol [19]. The 5GMM protocol operates between the UE and the AMF, serving as

the fundamental NAS protocol for managing UE registrations, mobility, security, and message transport. It not only transports the 5GSM protocol but also facilitates general NAS message delivery, such as communication between the UE and the PCF or SMS Function (SMSF). The 5GSM protocol, on the other hand, is responsible for managing PDU Session connectivity and operates between the UE and the SMF through the AMF. This protocol is carried over the 5GMM protocol, and the interaction among these elements. Furthermore, NAS in 5G systems is applicable over both 3GPP and non-3GPP access types. This is a key differentiation from Enhanced Packet System) (EPS)/4G systems, where it was designed exclusively for 3GPP access.

Stream Control Transmission Protocol (SCTP): SCTP, introduced in IETF Request For Comments (RFC) 2960 in 2000 [20] is a transport protocol operating at the same level in the protocol stack as UDP and Transmission Control Protocol (TCP), but unlike them it offers enhanced functionality and is more resilient to network failures. SCTP is commonly used over the N2 interface ensuring reliable delivery of signaling messages between the AMF and the AN.

Generic Routing Encapsulation (GRE) Protocol: The GRE is designed to tunnel one network layer protocol over another, providing a flexible encapsulation mechanism that allows various protocols (such as IP or Multiprotocol Label Switching (MPLS)) to be carried over other network layer protocols. This differs from many other tunneling mechanisms, which often require specific protocols at one or both ends. In general, tunneling encapsulates one network protocol—referred to as the payload protocol—within another, known as the delivery protocol. While encapsulation is fundamental to protocol stacks, where higher-layer protocols are encapsulated in lower layers, tunneling is distinct as it typically involves encapsulating one layer-3 protocol (e.g., IP) within a different layer-3 protocol or within another instance of the same protocol. The GRE operation begins with encapsulating a packet from protocol A (the payload) in a GRE packet, which is then encapsulated again within protocol B (the delivery protocol) for transport to the destination. The receiver decapsulates the packet to recover the original payload. In 5G systems, GRE is used primarily to tunnel PDUs between the UE and the N3IWF. GRE also enables the inclusion of the QoS Flow Identifier (QFI) value and Reflective QoS Indicator (RQI) for reflective QoS in its header, with both carried in the GRE key field alongside the encapsulated PDU.

IPsec: IPsec provides security services for both IPv4 and IPv6. It operates at the IP layer, offers protection of traffic running above the IP layer, and it can also be used to protect the IP header information on the IP layer. 5GS uses IPsec to secure communication on several interfaces, in some cases between nodes in the core network and in other cases between the UE and the core network. For example, IPsec is used to protect traffic in the core network. IPsec is also used between the UE and the N3IWF to protect NAS signaling and User Plane traffic.

Extensible Authentication Protocol (EAP): It is a flexible authentication framework that facilitates authentication of User UE in a network. Originally introduced by IETF [21] for the Point-to-Point Protocol (PPP), EAP has since expanded for use in other contexts, including Internet Key Exchange (IKE)v2 and wireless LANs. Rather than being an authentication method itself, EAP provides a structure that supports various specific authentication protocols, known as EAP methods (EAP-AKA, EAP-Transport Layer Security (TLS) etc.). In 5GS EAP-AKA' is widely used for authentication over both 3GPP and non-3GPP access networks.

2.2.2. 5G Radio Access Network (RAN)

With the increased complexity that is brought by new cellular network technologies and concepts [22] such as MIMO, mmwave and sub-terahertz communications, network-based sensing and ML-based digital signal processing , as well as the heterogeneous nature of the environments where these networks will operate, arises the need for new approaches that will enable openness and flexibility regarding the Radio Access Networks. Several research and standardization bodies are working towards this direction, promoting the O-RAN as the RAN paradigm of the future. In contrast with the current all-in-one blackbox approaches, O-RAN deployments rely on the disaggregation and virtualization of software-based network components, as well as interoperability across multiple vendors. Such systems are based on cloud-native principles and increase flexibility, reconfigurability and resiliency for the RAN. Moreover, the introduction of open and interoperable interfaces and protocols enables the integration of intelligent and data-driven control while at the same time opening the RAN ecosystem for the smaller vendors [23]. The architectural O-RAN approach in comparison with the traditional black box base station is depicted in Figure 2. 4.

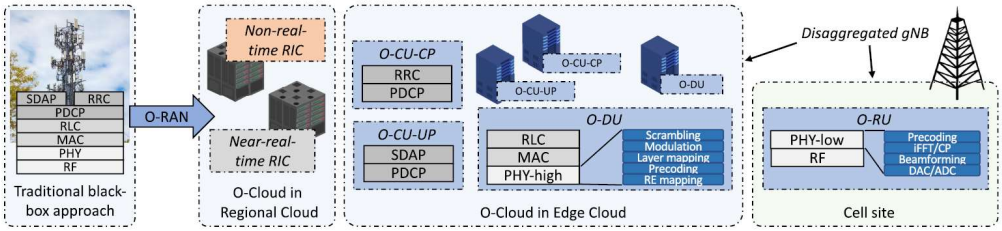


Figure 2. 4: O-RAN Architecture. Evolution from the trafitional monolithic approach to the Disaggregated Functional (RU/DU/CU) Split [23]

The leading organization to standardize and enhance the O-RAN vision is called O-RAN Alliance [7] created in 2018 and it identifies four main principles for the O-RAN architecture:

1. ***Disaggregation of the base station:*** The 3GPP NR 7.2 split is adopted and extended for the 5G base stations called gNBs, where its functionalities are disaggregated into [24]:

    - A CU which mainly handles the higher layers of the protocol stack, namely the Radio Resource Control (RRC), Packet Data Convergence Protocol (PDCP) and Service Data Adaptation Protocol (SDAP). Furthermore, it is possible to decouple the CP from the UP in the CU, allowing different functionalities to be executed at different sites of the network and/or in different hardware components.

    - A DU which features the RLC, MAC and higher Physical layer procedures. Both the CU and the DU can be hosted in white box servers located at the Edge of the network

    - A RU which includes the RF antennas and Field Programmable Gate Arrays (FPGAs) along with Application-Specific Integrated Circuits (ASICS) that handle the lower Physical layer procedures

2. ***Virtualization:*** Additional components are introduced in the O-RAN architecture that aim to manage and optimize the performance of network

infrastructures and operations, such as edge systems and virtualization platforms [25]. O-RAN refers to a hybrid cloud platform called O-Cloud which is defined as a set of compute resources and virtualization infrastructure that form datacenters designed to host the O-RAN functionalities. This way, the software components can be decoupled from the hardware, hardware capabilities for the O-RAN infrastructure can be properly defined and harmonized, hardware sharing among different tenants is enabled as well as automated deployment and instantiation of the RAN functionalities. Moreover, virtualization will ease scaling compute resources up or down depending on the requirements, thus increasing the overall energy efficiency of 5G/B5G/6G systems.

3. ***RAN Intelligent Controllers (RICs):*** The O-RAN architecture also includes a set of programmable components that can perform optimization procedures based on closed-loop control as well as RAN orchestration. Through the addition of two logical software-based controllers, the element of intelligence is introduced to future RAN networks, which enables the optimization and orchestration of various functionalities through data-driven, closed-loop control. The RICs are centralized and abstracted components, thus they can gather and exploit various data from different parts of the network in order to perform Management and Control tasks. The two RICs are:

    - The near-Real Time RIC (near-RT RIC): This controller is placed close to the network, usually at the edge, to perform control loops and handle events that require a response time between 10 ms and 1 sec. The near-RT RIC is deployed in the form of various microservice-based applications called xApps that are used to perform radio resource management through defined service models and interfaces. Specifically, the near-RT RIC presents a termination point for the interfaces O1, A1 and E2.

    - The non-Real Time RIC (non-RT RIC): The non-RT RIC integrates with the network orchestrator, operating from within the Service Management and Orchestration (SMO) framework and it manages events with a response time higher than 1 second. This component includes a set of interfaces that interact with different parts of the network and collect data that it feeds to AI/ML algorithms in order to perform network monitoring and control procedures.

4. ***Open Interfaces and protocols:*** The addition of new, open interfaces to the O-RAN model is an enabler for the gNB disaggregation and at the same time they enhance openness by exposing data analytics and telemetry to the RICs. This way, the monolithic black box approach can be overcome, leading the way to automation, virtualization and deployment optimization of the RAN segment. Furthermore, O-RAN aims to break vendor lock-in by increasing the level of access to the equipment from the operators [26]. Some of the interfaces introduced in the O-RAN ecosystem, are:

    - E2, connecting the near-RT RIC with the RAN nodes

    - O1, connecting to all RAN components for management and orchestration of the network functionalities

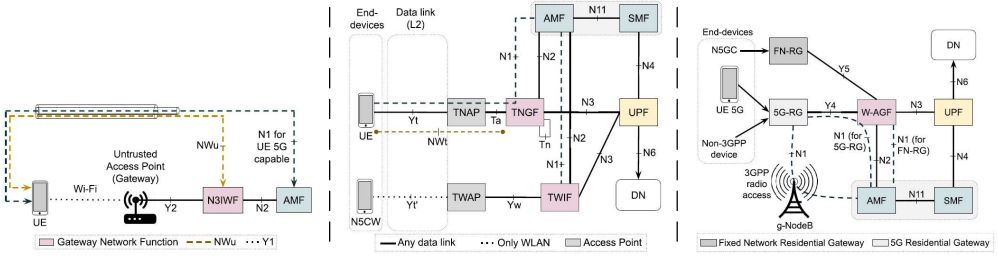
    - O2, connecting the non-RT RIC and SMO with the O-Cloud.

- A1, connecting the two RICs

2.2.3. Integration with Non-3GPP Access Networks

The integration of 3GPP and non-3GPP access networks is a key aspect to the adoption of 5G Non-Public Networks (NPNs) in vertical domains that utilize diverse access technologies, such as the pervasive use of Wi-Fi [27]. Achieving intelligent coordination between these networks is essential to minimize data congestion, improve capacity, and extend coverage, addressing the demands of emerging use cases driven by the rapid expansion of IoT devices and industrial communications. Additionally, this integration allows end devices that connect via non-3GPP networks and lack native 5G capabilities—such as legacy and IoT devices—to benefit from 5G-enabled scenarios. These include: (i) eMBB for increased bandwidth, (ii) mMTC for high connection density, and (iii) URLLC for reduced end-to-end latency. Seamless interworking between 5G access and other industrial technologies reduces operational costs and accelerates the broader adoption of 5G networks, particularly in the early stages of deployment.

To enable UE connectivity through non-3GPP access networks, three types of access are defined: (i) untrusted, (ii) trusted, and (iii) wireline [32]. A brief overview of each type is provided below:



a) b) c)

Figure 2. 5: Connectivity options for non-3GPP access networks with the 5G Core: a) untrusted, b) trusted, c) wireline [27]

Untrusted Networks: Untrusted access refers to situations where the MNO does not trust the security provided by the non-3GPP access network [29]. As a result, traffic must be securely transported in a way that meets the MNO's security requirements. The primary component facilitating untrusted access is the Non-3GPP Interworking Function (N3IWF). Introduced in 3GPP Release 15 [30], N3IWF serves as a gateway for communication between the UE and the 5GCN. Figure 2. 5 a) illustrates how untrusted non-3GPP networks, such as Wireless Local Area Network (WLAN) or Wi-Fi, integrate with the 5G core. It also shows the use of encrypted IP Security (IPSec) tunnels, known as NWu, which secures traffic from the untrusted network to the 5G core while isolating non-3GPP and 3GPP data flows. This model can be extended by integrating other access technologies with the 5GC, such as LiFi [31]. This way, the heterogeneity of 5GS can be further enhanced, i.e with the support of handover among different networks as a built-in feature.

Trusted Networks: Trusted access, standardized in 3GPP Release 16 [28], establishes another relationship between the non-3GPP access network and the 5GCN compared to the untrusted access scenario. While the 3GPP standard does not explicitly define the level of trust [33], the behavior of trusted access resembles that of 3GPP access. In a trusted network, the operator has full control over the Trusted Non-3GPP Access Point (TNAP) and the radio link, allowing the operator to manage encryption or rely on the security provided by the non-3GPP network. TNAP allows UE to connect to the trusted access network using non-3GPP wireless or wired technologies. The Trusted Non-3GPP Gateway Function (TNGF) enables UEs to connect to the 5GCN through the trusted access network by exposing the N2 and N3 interfaces. One implementation of a Trusted Non-3GPP Access Network (TNAN) is the Trusted WLAN Access Network (TWAN), which specifically supports WLAN. TWAN includes components such as the Trusted WLAN Access Point (TWAP) and the Trusted WLAN Interworking Function (TWIF), which facilitate secure connections to the 5GCN for devices in a WLAN that lack 5G capabilities. These devices, referred to as Non-5G Capable over WLAN (N5CW), rely on the TWIF for NAS signaling via the N1 reference point. Figure 2. 5 b) illustrates two trusted access options: (i) connecting to the 5GCN via a generic solution using TNAP and TNGF, and (ii) connecting N5CW devices over WLAN using TWAP and TWIF.

Wireline: This type of access was also introduced in 3GPP Release 16 [28], specifying two types of Wireline 5G Access Networks (W-5GAN): (i) Wireline 5G Broadband Access Network (W-5GBAN) and (ii) Wireline 5G Cable Access Network (W-5GCAN), as defined by the Broadband Forum (BBF) [34] and CableLabs. The Wireline Access Gateway Function (W-AGF) acts as a gateway and connects these wireline access networks to the 5G Core. The 5G Residential Gateway (5G-RG) acts as a UE and handles NAS signaling with the 5GCN. In contrast, the Fixed Network Residential Gateway (FN-RG) is a legacy gateway used in existing wireline networks which are not able to support 5G functionalities such as N1 signaling. The W-AGF links the N2 and N3 interfaces to the 5GCN and can handle the N1 interface signaling for devices connected through FN-RG, i.e. Non-5G Capable (N5GC) devices. Additionally, W-AGF is in charge of data traffic relay between residential gateways and the UPF [35]. As shown in Figure 2. 5 c), the 5G-RG can connect to the 5GCN either via (i) W-AGF or (ii) gNodeB, enabling Fixed Wireless Access (FWA). A UE connected to the 5G-RG can simultaneously access the 3GPP RAN and the non-3GPP network via W-AGF, using separate N1 interface instances for data traffic.

2.3. Enabling Technologies

2.3.1. Network Slicing

In the 5G era, the traditional one-size-fits-all network model is adapted in order to support the various types of network deployments, use cases, applications and subscriber types with their diverse and, in many cases, contradictory requirements [15]. This level of heterogeneity cannot be achieved only by introducing sophisticated technologies, but also architectural solutions need to be adopted. The term network slicing refers to the division of a physical network into several logical and isolated networks, where each one meets the requirements for a targeted type of application or user. This architectural approach becomes feasible by leveraging certain technologies and concepts such as virtualization that disaggregates the software from the underlying hardware, SDN [36] which simplifies network management and introduces

programmability and open network access, as well as cloud computing. This way the physical network resources can be distributed dynamically and efficiently on a per slice basis according to the changing customer needs. It is important to note that the customer here is not necessarily an end-user, but in can refer to a business entity with specific demands, a service provider e.g. a streaming platform, or even the network operator itself. A slice can extend across multiple domains, including the RAN, core networks deployed on distributed cloud infrastructure, and transport networks that support flexible placement of virtual functions. The core principle of 5G network slicing is to provide only the necessary functions customized to handle the specific traffic requirements of the use case. An illustration of network slicing is provided in Figure 2. 6, with the physical infrastructure at the bottom comprising access, transport and cloud computing technologies. This infrastructure is shared between three logical networks, resembling the 5G use cases (eMBB, mMTC, URLLC).

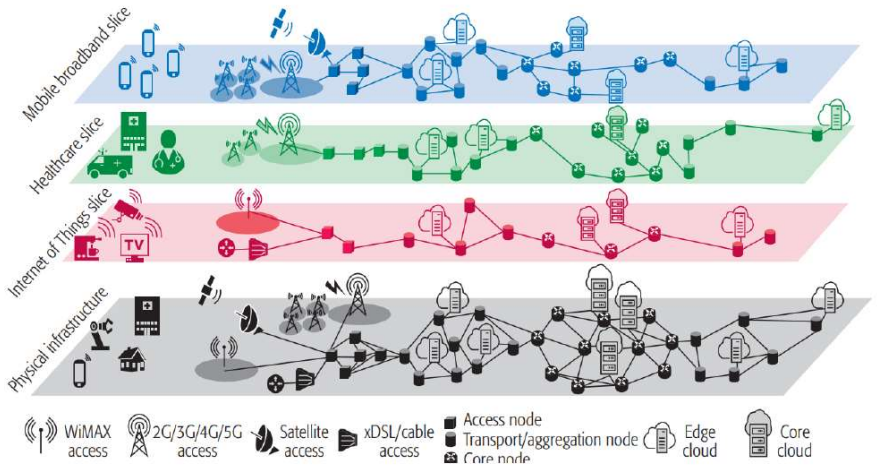


Figure 2. 6: Network Slicing Examples on top of a common physical infrastructure [37]

Network slices offer not only specialized capabilities designed for particular services but also the flexibility to adapt to evolving needs. One of the distinguishing features of network slicing, compared to traditional network sharing, is its ability to offer a virtualized end-to-end environment that can be opened to third parties.

According to the definition provided in [38], network slicing is organized into three layers:

- **Service Instance Layer (SIL)**: Represents end-user or business services, each represented by a service instance.
- **Network Slice Instance (NSI) Layer** : Encompasses the network slice instances, which provide the necessary network features for the corresponding service instance.
- **Resource Layer (RL)**: Supplies the physical or virtual resources and network functions needed to create a network slice instance.

Depending on the implementation, a network slice can be characterized as hard, if it is completely isolated, or soft if it shares network resources (for example SMF) with other slices. To application providers or distinct vertical sectors without a physical network infrastructure, network slicing can provide radio, cloud, and networking services. By

tailoring network operation to customers' needs based on the type of service, it enables service differentiation [39].

The specific characteristics and requirements of a network slice are depicted through two attributes, the Service Slice Type (SST) and the Service Descriptor (SD) feature [40]. The SST is a high-level identifier that represents a specific type of service that the network slice is intended to provide. It provides an abstract representation of the service, such as eMBB, URLLC and mMTC. The SD, on the other hand, is a detailed specification that defines the functional and non-functional requirements of the service, including network functions, network topology, performance metrics, and security policies. It provides a complete description of the slice, which can be used to configure and provision the network slice. The SD is typically composed of several parts, including a slice subnet, network function requirements, performance requirements, security requirements, and charging and billing requirements. The slice subnet refers to the IP address subnet used by the slice, while the functional requirements dictate the types of network functions required. Performance requirements are the non-functional requirements like maximum latency and throughput and security requirements refer to policies like authentication and encryption.

The Next Generation Mobile Network (NGMN) divides a slice concept in three basic layers, namely the 5G SIL, the 5G NSI Layer and the 5G RL [41]. The SIL provides the services instances (SIs) that are going to be supported. The NSI describes the network requirements for these services and can be shared across multiple SIs. The NSI may comprise network slice subnet instances (NSSIs), which can be dedicated or shared among several NSIs. Finally, the RL consists of physical and logical resources that are allocated to the slice.

For a unified view of the 5G-slicing concept, 3GPP has united multiple Standard Developing Organizations (SDOs) to define the concept, use cases, requirements and solutions for MANO of network slices. The NSI concept is redefined as “a set of network functions and the resources for these network functions which are arranged and configured, forming a complete logical network to meet certain network characteristic” [42]. The NSI may be comprised of network slice subnet instances (NSSIs), which can be dedicated or shared among several NSIs. In relation to NFV, an NFV service can be regarded as a resource-centric view of an NSI. Finally, 3GPP established a 3-step MANO lifecycle of 5G slices, that comprises of:

- Instantiation, configuration and activation. During this step, all the resources required for the NSI are created and configured. Any other actions that are needed to be performed in order for the NSI to be operative and active are also performed in this step.
- Run-time. During this step, the NSI is operational, and can be supervised and monitored. Furthermore, run-time actions, e.g. scaling, can be also performed.
- Decommissioning. During the decommissioning phase the NSI is deactivated and the resources that were allocated to the NSI are released.

2.3.2. Management and Orchestration

A core prerequisite for 5G systems has been the support of flexible and configurable network architectures, so that they can adapt to any use case and service requirements. To meet the requirements of diverse use cases, 5G systems embrace technologies such

as SDN and NFV to enable dynamic deployment of network functionalities, replacing the need for extensive network reconfiguration. Services will no longer be deployed, configured and managed on a node-by-node basis, but in an integrated and coordinated way. This approach is related to the removal of individual device configuration in favor of a more robust management mechanism that can offer network-wide service design, configuration, deployment, and monitoring. Such a process requires implicit autonomic control over all systems, resources, and services as well as inherent intelligence.

To manage and configure every component of the entire network service simultaneously, higher-level abstractions and automated methods are required [43]. Abstraction enables representation of entities through chosen attributes, that are shared by similar resources and that are handled and controlled, while concealing or summarizing characteristics unrelated to the selection criteria. The administration of resources can be generalized and made simpler by the abstraction, removing the initial impediments caused by manufacturing differences, in particular the technology used to create them, or the resources' physical realization. This opens up the potential for implementation of advanced services across all network domains and offers significant prospects for achieving the required degree of automation and targeted KPIs.

The 5G MANO framework plays a vital role towards this direction [44]. It targets a robust and centralized management of service creation and to do that, it encompasses a suite of technologies, standards, and frameworks designed to streamline the lifecycle management of Virtualized Network Functions (VNFs) and network services, enabling operators to deliver innovative services with agility, efficiency, and scalability.

Central to the 5G MANO paradigm is the concept of NFV, which forms the cornerstone of virtualized network infrastructure in 5G networks. A key advantage of the NFV architecture is the ability to deploy network functions on virtual machines or containers running on standard computing servers, eliminating the need for proprietary, vendor-specific hardware. This flexibility allows the system to dynamically adjust to changing NF requirements based on the current system load. The classic NFV architecture is shown in Figure 2. Within NFV, there are three primary elements [45][46][47][48]:

- The VNFs that perform a different NF in isolation. A VNF may consist of several internal components, e.g VMs. For example, a VNF that consists of two VMs, one for the data storage and a second for the main functions. The number of VNFs and their arrangement within the virtual infrastructure are typical for each Network Service (NS) we want to deploy.

- The NFV Infrastructure (NFVI) that consists of general-purpose computing hardware and software and could be geo-distributed over several locations. It offers the virtualization layer that is responsible for abstracting and delivering the physical resources (compute, storage, network) to support the VNFs. VNFs can be located anywhere in the network, as long as the location of each one in the network is known so that they can communicate properly with each other.

- The NFV MANO that interacts with the VNF and NFVI blocks to manage all virtualization-specific management duties within the NFV framework.

All infrastructure resources and VNFs can be managed and coordinated using the NFV MANO. In more detail, it offers virtual machines for the VNFs, sets them up and controls the management of physical resources for the VMs and the lifecycle of the VNFs [49]. NFV MANO is composed of virtualized infrastructure managers (VIMs), VNF managers and NFV orchestrators. VIMs control and manage the interactions of a VNF

with its computing, storage and network resources, and ensures their virtualization. The VNF manager is responsible for managing the entire VNF lifecycle. It is responsible to initialize, query, update and terminate VNF instances. Finally, NFV orchestrators are responsible for orchestrating and managing new network services into a virtual framework, which includes instantiation, policy management, performance measurement and monitoring. Together, these blocks are responsible for deploying and connecting functions and services when they are needed throughout the network.

2.3.3. Intelligence and Automation

B5G systems are expected to build highly heterogeneous networks with diverse service requirements and increased operational complexity. Considering that, AI/ML tools can enhance network and service performance optimization while at the same time minimize costs by introducing automation [50]. AI/ML leverages the ability to learn without explicit programming and can assist in autonomous decision making by facilitating analytics. MNOs can exploit AI/ML for planning, optimizing and operating their networks. Therefore, reducing operational costs is a short-term objective, with the ultimate goal being to generate new revenue streams by leveraging B5G services in combination with big data. AI/ML plays a crucial role in enhancing customer experience and creating innovative services. MNOs should explore value generation through new applications and platforms that offer services driven by data analytics and AI/ML [51]. Providing AI/ML services to premium subscribers and third-party applications, such as for ensuring desired performance levels, can enable MNOs to extend their offerings beyond connectivity and increase revenue [52]. For example, autonomous driving applications can benefit from AI/ML services by gaining insights into future network conditions, allowing them to adjust vehicular automation levels based on predicted network performance. Similarly, mixed-reality applications can use AI/ML services to proactively synchronize distributed sources by leveraging resource flexibility. In practice, mobile networks can adopt AI/ML services across various network segments. In network management and orchestration, the use of AI/ML aims on improving network resource allocation, assuring network performance, as well as analyzing failures [53]. The RAN segment relies on real-time or near real-time data to make predictions and analysis of user access and radio conditions. For the Core Network, AI/ML focuses on control plane operations such as mobility, security and network performance assurance. Finally, the Application layer focuses on optimizations regarding QoS/Quality of Experience (QoE), negotiating policies and synchronizing distributed application sources. This integration allows MNOs to provide smarter, more adaptable services, unlocking new potential for innovation and differentiation.

Across the network segments, various functions are defined to enable adoption of AI/ML capabilities:

- ***The Network Data Analytics Function (NWDAF)*** [54][55] leverages user data to provide analytics on various functionalities such as mobility, communication patterns, traffic, and abnormal behavior. NWDAF interacts with other network functions, applications, and the 5G core to collect data such as user location, QoS, and traffic volume [56][57], enabling predictive and real-time analytics. It also uses data from applications via the NEF to assess and optimize QoE and network performance. NWDAF is divided into two components: the Analytics Logical Function (AnLF) for inference and analytics, and the Model Training Logical Function (MTLF) for training AI/ML models. Federated Learning (FL) can also be used to distribute model training across

network elements, ensuring continuous model improvement and performance accuracy.

- ***The Management Data Analytics (MDA)*** [58][59] enables analytics within the 3GPP MANO plane, focusing on tasks like resource optimization, feasibility checks, user performance analysis, and root cause detection. Unlike NWDAF, which specifies a function, MDA introduces a service (MDAS) that offers flexibility in deployment across management functions. MDA collects Operations, Administration, Maintenance (OAM) performance data [60], KPIs [61], trace data, and other configuration information to generate insights, with more complex analytics often involving interaction with NWDAF for data such as Quality of Experience (QoE). MDA can provide recommendations for network issues, such as optimal configuration changes or connectivity strategies. It may also expose analytics to third parties using the Exposure Governance Management Function (EGMF). Unlike NWDAF's focus on statistics and predictions, MDA provides both analytics and actionable recommendations.

- ***The Application Data Analytics Enablement Service (ADAES)*** [62] provides application-specific analytics, including predictions and statistics, for vertical or edge applications, offering insights into service parameters. ADAE clients in UE can supply application data to the ADAE server, which may also collect additional data from the 5G core and management plane via network exposure. ADAES uses a Service Enabler Architecture Layer (SEAL) to offer analytics related to applications, 5G core, and network management to vertical applications. ADAES supports various functions, including location and group management, identity management, and network resource management. The architecture consists of a data collection function and a repository for storing historical data. ADAES provides several value-added capabilities [63], such as performance analytics for applications and sessions, edge load analytics to assist scaling decisions, UE-to-UE session performance predictions, and insights on application performance within network slices. It also includes analytics on location accuracy and service API usage to optimize services and network configurations proactively.

- For the ***RAN segment***, O-RAN [7] defines two types of RICs (as described in 2.2.2): a non-real-time RIC for tasks over 1 second and a near-real-time RIC for tasks under 1 second, both enabling RAN optimization and control. These RICs support AI/ML-based micro-services, such as model training, inference, and data collection. Non-real-time RICs handle AI/ML model deployment and configuration, while near-real-time RICs manage real-time control functions. This architecture promotes flexibility in RAN deployment across cloud platforms and supports various AI/ML applications to optimize network performance.

2.4. Mobility Management

The principles of Mobility Management in 5GS build on those from previous 3GPP systems like LTE, but with some key distinctions [15]. As in previous generations, mobility is fundamental to 5GS, ensuring the network can locate users for notifications, allow user-initiated communication, and maintain connectivity and active sessions as users move across or within access technologies. These functions are enabled by

establishing and maintaining connections between the UE and the network through specific mobility management procedures. Mobility Management also supports UE identification, security, and facilitates broader communication between the UE and the 5GC. The 5GC network aims to serve as a converged core compatible with any access technology, while offering flexible mobility options to accommodate diverse use cases. For instance, a stationary device operating in a factory will require completely different mobility procedures from a moving vehicle.

Architectural enhancements like SDN, NFV, Network Slicing, and CC/MEC can enhance mobility management in 5GS. However, the dramatic increase in connected devices, traffic, and cell densification demands optimized mobility management to prevent handover failures, reduce ping-pong effects, and optimize resource utilization [64]. The traditional Centralized Mobility Management (CMM) which was adopted in LTE, is therefore less suitable for 5GS. Instead, Distributed Mobility Management (DMM) offers a solution by anchoring traffic closer to the user's point of attachment, contributing to network flattening [65]. This approach enables a device to connect to multiple mobility anchors, improving packet routing efficiency as the UE moves across attachment points. DMM introduces advanced control and data plane management features, supporting diverse scenarios, protocol enhancements, and potential user and network performance improvements. For example, Some DMM implementations focus on managing mobile video traffic more effectively, while others provide a comprehensive overview of standards and developments in DMM.

The handover (HO) process in 5G networks divides into XN-based and N2-based handovers, depending on the managing node [66]. The XN-based HO is coordinated through the X2 interface between source and target gNBs in different cells and can occur only if both gNBs connect via the same AMF through the N2 interface, known as intra-AMF handover. In contrast, the N2-based HO is used when no X2 connection exists between source and target gNBs, usually due to broader network distribution, and is managed through the CN. This HO process comprises two phases: preparation and implementation. During preparation, the gNB and SMF/UPF are configured to ensure quick, uninterrupted execution of the handover.

5GS also introduces vertical handovers due the adoption of multi-RAT connectivity. Vertical handovers involve transitions between 3GPP access networks, such as 5G to LTE, or between a 3GPP and a non-3GPP access network like Wi-Fi. Connecting a device over an untrusted WiFi network implies security concerns, due to its common use of password-based authorization and potential lack of payload encryption. These security concerns make Wi-Fi less dependable for accessing critical mobile network services.

For fixed wireless access scenarios, a full set of mobility procedures may not be necessary, enabling selective addition or removal of procedures as part of a "mobility-related service." This flexibility allows varying degrees of mobility support, with minimal signaling for stationary UEs, beyond periodic registration updates. In line with this flexibility, 5GS includes optional mobility management features introduced by 3GPP. Service Area Restriction allows mobility with session continuity to be managed at the UE level in specific areas. Local Area Data Network (LADN) enables session-level control of mobility within specified areas, while the Mobile Initiated Connection Only (MICO) feature provides optional paging capabilities within the mobility service.

5GMM procedures are categorized into three types, based on purpose and initiation criteria. Common procedures are available when the UE is in the CM-CONNECTED state, while specific procedures allow only one UE-initiated process per access type at

any given time. Connection management procedures establish secure signaling between the UE and the network, request data resource reservations, or both. These updates reflect the 5GS's flexibility in addressing varied mobility requirements and use cases across industries, establishing it as a versatile and responsive network infrastructure.

2.5. Summary

5G networks aim to support a wide variety of services with extremely stringent and diverse requirements, compared to the previous mobile network technologies. This chapter explores various aspects of the 5GS. At the beginning the chapter describes the 5G functional architecture, designed to support applications related with different use cases such as eMBB, URLLC and mMTC. To this end, the main architectural components, 5GC and 5G RAN, are examined. The NG-RAN provides radio access to end-users, while the 5G Core manages connectivity and control of 5G services. Following this overview, the chapter investigate various enabling technologies adopted by 5GS, like network slicing, MANO and AI/ML. Network slicing enables the division of the physical infrastructure into multiple virtual networks, allowing each network slice to be independently configured in order to support a variety of unique use cases and applications. MANO enables the management and orchestration and provides MNOs the ability to easily perform operations in the network like reconfiguration, scaling, monitoring etc. Moreover, AI/ML tools are useful in various aspects such as network optimization, automation and predictive maintenance. Finally, the chapter provides a brief overview of Mobility Management in 5G systems.

References

[1] O. Liberg *et al.*, "Introducing 5G Advanced," in *IEEE Communications Standards Magazine*, vol. 8, no. 1, pp. 52-57, March 2024, doi: 10.1109/MCOMSTD.0003.2200059.

[2] 5g-PPP (2022). 5G PPP TB & 5G IA Verticals TF, "Empowering Vertical Industries through 5G Networks - Current Status and Future Trends, Version 1.0. [online]. Available: https://5g-ppp.eu/wp-content/uploads/2020/09/5GPPP-VerticalsWhitePaper-2020-Final.pdf

[3] 3rd Generation Partnership Project (3GPP) [online]. Available: https://www.3gpp.org/

[4] S. Idris, U. Mohammed, J. Sanusi and S. Thomas, "Visible Light Communication: A potential 5G and beyond Communication Technology," *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, Abuja, Nigeria, 2019, pp. 1-6, doi: 10.1109/ICECCO48375.2019.9043201.

[5] International Telecommunication Union (ITU) [onlne]. Available: https://www.itu.int/

[6] European Telecommunications Standards Institute (ETSI) [online]. Available: https://www.etsi.org/

[7] O-RAN Aliance [online]. Available: https://www.o-ran.org/

[8] L. M. P. Larsen, A. Checko and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146-172, Firstquarter 2019, doi: 10.1109/COMST.2018.2868805.

[9] [27] M. Kamel, W. Hamouda and A. Youssef, "Ultra-Dense Networks: A Survey", in IEEE Communications Surveys & Tutorials, vol. 18, no. 4, pp. 2522-2545, 2016. Available: https://doi.org/10.1109/comst.2016.2571730

[10] A. Kunz and A. Salkintzis, "Non-3GPP Access Security in 5G," in *Journal of ICT Standardization*, vol. 8, no. 1, pp. 41-56, 2020, doi: 10.13052/jicts2245-800X.814.

[11] A. Tzanakaki et al., "Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services," in IEEE Communications Magazine, vol. 55, no. 10, pp. 184-192, Oct. 2017. Available: https://doi.org/10.1109/MCOM.2017.1600643

[12] A. Tzanakaki, A. Manolopoulos, M. Anastasopoulos, and D. Simenidou, "Optical Networking in Support of User Plane Functions in 5G Systems and Beyond," in Photonics in Switching and Computing 2021, pp. W2B.3, January 2021. Available: https://doi.org/10.1364/PSC.2021.W2B.3

[13] M. Satyanarayanan et al., "An open ecosystem for mobile-cloud convergence," in *IEEE Communications Magazine*, vol. 53, no. 3, pp. 63-70, March 2015. Available: https://doi.org/10.1109/MCOM.2015.7060484

[14] H2020 Project 5G-COMPLETE, Deliverable D2.3: "Final report on 5G-COMPLETE network architecture, interfaces". [online]. Available: https://5gcomplete.eu/wp-content/uploads/2023/05/D2_3.pdf

[15] Rommer, S., Hedman, P., Olsson, M., Frid, L., Sultana, S., & Mulligan, C. (2019). *5G core networks: powering digitalization*. Academic Press.

[16] 5G; NG-RAN; NG Application Protocol (NGAP) 3GPP TS 38.413 version 15.3.0 Release 15 [online]

Available:https://www.etsi.org/deliver/etsi_ts/138400_138499/138413/15.03.00_60/ts_138413v150300p.pdf

[17] LTE; 5G; Interface between the Control Plane and the User Plane nodes. 3GPP TS 29.244 version 16.4.0 Release 16 [online]. Available:https://www.etsi.org/deliver/etsi_ts/129200_129299/129244/16.04.00_60/ts_129244v160400p.pdf

[18] Universal Mobile Telecommunications System (UMTS); LTE; 5G; General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U) 3GPP TS 29.281 version 16.0.0 Release 16 [online]. Available: https://www.etsi.org/deliver/etsi_ts/129200_129299/129281/16.00.00_60/ts_129281v160000p.pdf

[19] 5G; Non-Access-Stratum (NAS) protocol for 5G System (5GS); Stage 3 3GPP TS 24.501 version 15.2.1 Release 15 [online]. Available: https://www.etsi.org/deliver/etsi_ts/124500_124599/124501/15.02.01_60/ts_124501v150201p.pdf

[20] RFC 2960 - Stream Control Transmission Protocol [online]. Available: https://datatracker.ietf.org/doc/html/rfc2960

[21] IETF - Extensible Authentication Protocol (EAP) [online]. Available: https://datatracker.ietf.org/doc/html/rfc3748

[22] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies," in *IEEE Communications Magazine*, vol. 58, no. 3, pp. 55-61, March 2020, doi: 10.1109/MCOM.001.1900411.

[23] M. Polese, L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376-1411, Secondquarter 2023, doi: 10.1109/COMST.2023.3239220.

[24] NG-RAN; Architecture Description, Version 17.0.0, 3GPP Standard (TS) 38.401, Apr. 2022. [Online]. Available: https://www.3gpp.org/ DynaReport/38401.htm

[25] O-RAN Working Group 1, "O-RAN architecture description 5.00," O-RAN, Alfter, Germany, document O-RAN.WG1.O-RANArchitecture- Description-v05.00 Technical Specification, Jul. 2021.

[26] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in O-RAN for data-driven NextG cellular networks," IEEE Commun. Mag., vol. 59, no. 10, pp. 21–27, Oct. 2021.

[27] M. T. Lemes, A. M. Alberti, C. B. Both, A. C. De Oliveira Júnior and K. V. Cardoso, "A Tutorial on Trusted and Untrusted Non-3GPP Accesses in 5G Systems—First Steps Toward a Unified Communications Infrastructure," in *IEEE Access*, vol. 10, pp. 116662-116685, 2022, doi: 10.1109/ACCESS.2022.3219829.

[28] Technical Specification Group Services and System Aspects; Release 16 Description; Summary of Rel-16 Work Items (Release 16), 3rd Gen- eration Partnership Project-3GPP, Sophia Antipolis Cedex, France, document TR21.916, V0.1.0, 3GPP, Sep. 2019. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/21_series/21.916/.

[29] Cardoso, K. V., Both, C. B., Prade, L. R., Macedo, C. J., & Lopes, V. H. L. (2020). A softwarized perspective of the 5G networks. *arXiv preprint arXiv:2006.10409*.

[30] Technical Specification Group Services and System Aspects; Release 15 Description; Summary of Rel-15 Work Items (Release 15), 3rd Generation Partnership Project-3GPP, Sophia AntipolisCedex, France, document TR21.915, V15.0.0, 3GPP, Oct. 2019.

[Online]. Available: https://www.etsi.org/deliver/etsi_tr/121900_121999/121915/15.00.00_60/tr_121915v150000p.pdf

[31] V. Jungnickel, M. Hinrichs, K. Bober, C. Kottke, A. Corici, M. Emmelmann, J. Rufo, P.-B. Bok, D. Behnke, and M. Riege, ``Enhance lighting for the Internet of Things," in Proc. Global LIFI Congr. (GLC), 2019, pp. 1-6.

[32] Access to the 3GPP 5G Core Network Via Non-3GPP Access Networks 3rd Generation Partnership Project-3GPP, Sophia Antipolis Cedex, France, document ETSI TS 124.502, V16.7.0, 3GPP, Apr. 2021. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/124500_124599/124502/16.07.00_60/ts_124502v160700p.pdf

[33] J. T. Penttinen, 5G Second Phase Explained: The 3GPP Release 16 Enhancements. Hoboken, NJ, USA: Wiley, 2021, p. 219.

[34] Broadband Forum.Web Site. Broadband Forum. Accessed: Aug. 13, 2021. [Online]. Available: https://www.broadband-forum.org

[35] Y. S. Tao Wan and M. Pala. (Oct. 2019). Authentication in 5G Wireline and Wireless Convergence. SCTE-ISBE. [Online]. Available:https://www.nctatechnicalpapers.com/Paper/2019/2019- authentication-in-5g-wireline-and-wireless-convergence-2

[36] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," in *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94-100, May 2017, doi: 10.1109/MCOM.2017.1600951.

[37] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," in *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80-87, May 2017, doi: 10.1109/MCOM.2017.1600935.

[38] NGMN 5G Initiative Team, “A Deliverable by the NGMN Alliance: NGMN 5G White Paper”. [Online]. Available: https://www.ngmn.org/

[39] M. Jiang, M. Condoluci and T. Mahmoodi, "Network slicing management & prioritization in 5G mobile systems," European Wireless 2016; 22th European Wireless Conference, Oulu, Finland, 2016, pp. 1-6.

[40] I. Afolabi, M. Bagaa, T. Taleb, H. Flinck, “End-to-End network slicing enabled through network function virtualization.”, 2017 IEEE Conference on Standards for Communications and Networking (CSCN), Helsinki, Finland, 2017, pp. 30-35.

[41] NGMN 5G Initiative Team, “A Deliverable by the NGMN Alliance: NGMN 5G White Paper”. [Online]. Available: https://www.ngmn.org/ wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf

[42] Study on management and orchestration of network slicing for next generation network (Release 15), Technical Specification Group Services and System Aspects, Telecommunication management, 3GPP TR 28.801 V15.1.0, January 2018

[43] Management and orchestration; 5G Network Resource Model (NRM) (Release 16), 3GPP TS 28.541 5G, March 2020. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3400.

[44] Management and Orchestration; Provisioning, 3GPP TS 28.531 January 2020. [Online].

[45] Akyildiz, Ian & Nie, Shuai & Lin, Shih-Chun & Chandrasekaran, Manoj, "5G Roadmap: 10 Key Enabling Technologies," in Computer Networks, Vol 106, 2016. Available: https://doi.org/10.1016/j.comnet.2016.06.010.

[46] Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV, ETSI GS NFV 003 V1.2.1 (2014). [Online].

[47] Network Functions Virtualisation (NFV); Architectural Framework, ETSI GS NFV 002 V1.2.1 (2014a). [Online].

[48] Network Functions Virtualisation (NFV); Management and Orchestration, ETSI GS NFV-MAN 001 V1.1.1 (2014b). [Online].

[49] K. Abbas, T. A. Khan, M. Afaq and W. -C. Song, "Network Slice Lifecycle Management for 5G Mobile Networks: An Intent-Based Networking Approach," in IEEE Access, vol. 9, pp. 80128-80146, 2021, doi: 10.1109/ACCESS.2021.3084834.

[50] Taleb, T., Benzaïd, C., Addad, R. A., & Samdanis, K. (2023). AI/ML for beyond 5G systems: Concepts, technology enablers & solutions. *Computer Networks*, *237*, 110044.

[51] GSMA, The mobile economy, 2023 [online]. Available: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/wp-content/uploads/2023/03/270223-The-Mobile-Economy-2023.pdf

[52] G. Frisiani, J. Jubas, T. Lajous, P. Nattermann, A future for Mobile Operators: The Keys to Successful Reinvention, McKinsey & Company, Telecommunications, 2017.

[53] R. Addad, et al., Network slice mobility in next generation mobile systems: Challenges and potential solutions, IEEE Netw. 34 (1) (2020) 84–93.

[54] 3GPP TR 23.791, Study of enablers for network automation for 5G, 2019, Rel.16.

[55] 3GPP TS 23.288, Architecture enhancements for 5G System (5GS) to support network data analytics services, 2020, Rel.16.

[56] 3GPP TS 23.501, System architecture for the 5G System (5GS), 2023, Rel.15.

[57] 3GPP TS 23.502, Procedures for the 5G System (5GS), 2023, Rel.15.

[58] 3GPP TR 28.809, Study on enhancement of Management Data Analytics, 2020, Rel.16.

[59] 3GPP TS 28.104, Management and orchestration; management data analytics, 2022, Rel.17.

[60] 3GPP TS 28.552, Management and orchestration; 5G Performance Measurements, 2023, Rel.15.

[61] 3GPP TS 28.554, Management and orchestration; 5G end to end Key Performance Indicators (KPI), 2023, Rel.15.

[62] 3GPP TS 23.436, Functional architecture and information flows for application data analytics enablement service, 2023, Rel.18.

[63] E. Pateromichelakis, D. Dimopoulos, A. Salkintzis, NetAPPs enabling application-layer analytics for vertical IOT industry, IEEE Internet Things Mag. 5 (4) (2022) 130–135.

[64] I. Shayea, M. Ergen, M. Hadri Azmi, S. Aldirmaz Çolak, R. Nordin and Y. I. Daradkeh, "Key Challenges, Drivers and Solutions for Mobility Management in 5G Networks: A

Survey," in *IEEE Access*, vol. 8, pp. 172534-172552, 2020, doi: 10.1109/ACCESS.2020.3023802.

[65] Siddiqui, M. U. A., Qamar, F., Tayyab, M., Hindia, M. N., Nguyen, Q. N., & Hassan, R. (2022). Mobility management issues and solutions in 5G-and-beyond networks: A comprehensive review. *Electronics*, *11*(9), 1366.

[66] E. Gures, I. Shayea, A. Alhammadi, M. Ergen and H. Mohamad, "A Comprehensive Survey on Mobility Management in 5G Heterogeneous Networks: Architectures, Challenges and Solutions," in *IEEE Access*, vol. 8, pp. 195883-195913, 2020, doi: 10.1109/ACCESS.2020.3030762.

Chapter 3. 5G SYSTEM PROFILING, MONITORING AND CLOUD PLATFORM

3.1. Chapter Introduction

In recent years, Cloud Computing defined as [1] has emerged enabling on-demand availability of computing resources such as processing, storage and networking and eliminating the need for active management by the user. The Cloud computing paradigm is adopted from various sectors hosting different types of applications in business, education, entertainment, art and social life. Cloud infrastructures are typically deployed in large DCs that can either be located on a single centralized location or distributed over multiple places. Cloud computing systems are generally considered to have the following characteristics [2]:

- On-Demand Self-Service: Consumers can independently provision computing capabilities, like server time and network storage, as needed and on-demand, without requiring human intervention from the service provider.
- Broad Network Access: Cloud capabilities are accessible over the network and available through standard mechanisms, supporting various client platforms, including mobile devices, tablets, laptops, and workstations.
- Resource Pooling: Providers pool computing resources to serve multiple consumers using a multi-tenant model, with physical and virtual resources dynamically allocated and reallocated according to consumer demand.
- Rapid Elasticity: Cloud resources can be quickly scaled up or down to meet demand, often automatically, giving the appearance of unlimited availability to consumers and enabling resources to be acquired in any quantity as needed.
- Measured Service: Cloud systems automatically manage and optimize resource usage through monitoring appropriate to the service type (e.g., storage, processing, bandwidth, active user accounts). This allows monitoring, control, and transparency of usage for both the provider and consumer, although for some organizations, increased usage may impact profitability differently than capital investments might.

These characteristics enable cloud computing to offer flexible, scalable, and transparent services that aim to satisfy a wide range of user needs.

Recently, with the digital transformation of mobile communications and the adoption of various technologies such as NFV that act as enablers, mobile networks are also migrating to the Cloud. Traditional mobile networks face several limitations related to hardware, connectivity, resource consumption etc., while cloud computing delivers robust computing capabilities as services via virtualization and service-oriented techniques to reduce costs and improve performance [3]. Mobile cloud computing enables powerful, complex computations on resource-limited mobile devices, extending

device functionality as well as other notable benefits in terms reliability, scalability, data management etc.

This chapter focuses on the 5G cloud testbed comprising a private cloud platform that provides a hosting environment for 5G/B5G mobile network and application deployments. Several added components are also presented such as s monitoring platform and MANO platform. Through this testbed, the resulting implementations cover several use case scenarios in terms of NF placement and provided services. Finally, the performance of each scenario is evaluated through extensive profiling.

The rest of the chapter is structured as follows: in subsection 3.2 we discuss the relevant background, followed by the overview of the environment in subsection 3.3. Subsection 3.4 includes a description of the available deployment options and subsection 3.5 we present the system profiling framework along with some results. Finally, in subsection 3.6 we provide a summary of the chapter.

3.2. Background Preliminaries

3.2.1. Cloud Computing

Cloud computing is a modern computing paradigm that delivers resources such as processing power, storage, networking, and software as services over the Internet. These resources are provided remotely, enabling users to access them from virtually anywhere. Users can access cloud applications through web browsers, thin clients, or mobile devices, while all the data and software are stored on remote servers, which also handle intensive processing tasks. This approach allows businesses and enterprises to operate more efficiently by offloading the management and maintenance of computing resources, turning computing into a service rather than a product [4]. By sharing computing hardware, cloud computing reduces idle time and increases the productivity of machines. As computing shifts from being a resource to a utility, cloud computing introduces new billing models based on time and usage. Some of the key features

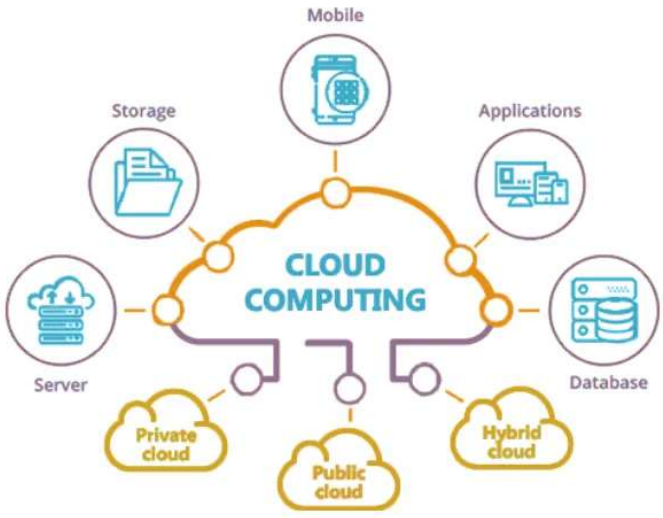


Figure 3. 1: Cloud Computing Services and Deployment Models [5]

include on-demand availability, easy provisioning, dynamic scaling, and virtually infinite scalability. An overview of cloud computing functionalities is in illustrated in Figure 3. 1 and as can be seen at the bottom, three deployment models can be distinguished: Public, Private an Hybrid. The **Public Cloud** dominates the current market share and provides cloud hosting services that are accessible to the public. It's open to all users, from individuals to multinational corporations. However, multiple customers may share the same physical hardware at the data center, as resources are pooled. The **Private Cloud** offers a similar model, but with dedicated, isolated resources reserved for each customer. This includes dedicated hardware and, in some cases, even specific server placements within a data center. It's an ideal option for larger organizations focused on security and control. The **Hybrid Cloud** combines both public and private cloud solutions, allowing organizations to balance cost and security. Businesses often choose to run non-sensitive or less critical operations on a cost-effective public cloud, while keeping sensitive customer data or proprietary information on a secure private cloud. This approach leverages the strengths of both environments.

Cloud services are typically categorized into three main models [6] (Figure 3. 2):

- **Infrastructure as a Service (IaaS)**: Provides fundamental cloud resources such as virtual machines, block storage, firewalls, load balancers, and networking.
- **Platform as a Service (PaaS)**: Offers a complete platform, including the operating system, programming environments, databases, and web servers, allowing developers to build and deploy applications.
- **Software as a Service (SaaS)**: Delivers software applications over the cloud, which users can access through devices like computers, mobile devices, or web browsers, without the need for local installation or maintenance.

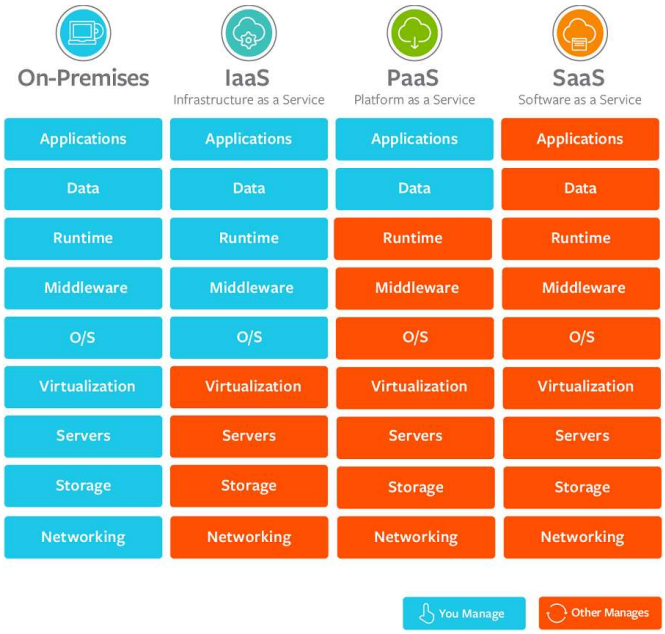


Figure 3. 2: Cloud Computing Service Models [7]

3.2.2. Virtualization

Virtualization is a powerful technology that has transformed many aspects of modern system architectures, with advantages that are still being revealed as a consequence of the massive deployment of commodity hardware and software systems. Virtualization relies in the use of an encapsulating software layer called hypervisor which provides the surrounding environment for virtual instances to be deployed over the physical resources. This environment provides the same behavior, inputs and outputs that would be expected from an actual host. Decoupling the physical from the logical state enables the deployment of multiple virtual instances on a single set of hardware resources. Virtualization offers many benefits in modern computing systems and security is one of these by presenting some inherent attributes such as isolation. At the same time, the increased architectural complexity, sophisticated software and network channels etc, provides a wide attack surface and leads to multiple vulnerabilities and threats. To provide a secure and robust virtualized environment, all the security considerations must be identified, to obtain the means and countermeasures needed to overcome them.

3.2.2.1. System Virtualization

According to Popek's and Goldberg's definition [8], system virtualization refers to the use of an encapsulating software layer that provides an underlying and surrounding environment for an operating system, which offers (almost) the same behavior as would be expected form physical hardware. This software layer is called a Virtual Machine Monitor (VMM) or Hypervisor and its main purpose is to decouple the software from the hardware state, while at the same time providing an environment to the software that looks like the host system. The resulting VMs are used to install operating systems. Since a VM is not dependent on the state of the hardware on top of which it operates, multiple VMs can be deployed on a single physical host.

3.2.2.2. Virtualizable Architectures

The requirement for a classically virtualizable architecture is defined by two properties regarding Central Processing Unit (CPU) instructions, i.e. privilege level and sensitivity level. An architecture is virtualizable if it can trap all the sensitive instructions and calling the VMM. To be fully virtualizable, the set of sensitive instructions for this computer must be a subset of the privileged instructions. Otherwise, if some sensitive instructions are not capable of being trapped, then the architecture is referred as partially virtualizable.

3.2.2.3. VMM Implementation Types and Key Components

VMMs are typically implemented with extra features that enable VM management and virtual hardware provisioning to VMs. VMM implementations are distinguished in two types [9]: type-I hypervisor refers to a VMM that directly operates the hardware resources. The host system is running alongside the guest VMs and both need to cope with the hypervisor. Typical examples of type-I implementations are the Xen and the KVM hypervisor. Type-II hypervisors run as applications of the host system and they cannot access the hardware directly, instead they rely on the host Operating System

(OS) routines to access the hardware resources. Well-known examples of type-II hypervisors are QEMU and Oracle VirtualBox.

For a virtualization platform to operate, a set of entry points, control channels and network channels exist. Control channels allow the configuration and management of the VMM and its child VMs. Such operations include changing of settings, control of the operational status of the VMs and the hypervisor, as well as facilitation of screen and keyboard access. Additionally, VM control channels are used for a set of control and automation actions such as (OS shutdown, file transfers and software running in guest OS. Control channels offer elevated levels of access, thus strong precautions must be taken to avoid compromise. Networking is also an important part of virtualization. Consequently, a set of network communication data flows and channels exist that interconnect all the virtualization layers.

3.2.2.4. OS-Level virtualization

OS-level virtualization, also known as containerization, is a form of virtualization technology that enables multiple isolated user-space instances, often called containers, to run on a single OS kernel. Unlike traditional virtualization methods, which rely on hypervisors to manage multiple operating systems on a single physical machine, OS-level virtualization operates at the OS level, leveraging the host OS's kernel to partition resources and provide isolation between containers. Each container shares the host OS's kernel but maintains its own filesystem, processes, and network interfaces, allowing for lightweight and efficient deployment of applications. This approach offers benefits such as improved resource utilization, faster startup times, and simplified management compared to traditional virtual machines. Popular containerization platforms like Docker and Linux Containers (LXC) utilize OS-level virtualization to package and deploy software applications across diverse computing environments.

3.2.2.5. Unikernel Virtualization

Unikernel virtualization is a specialized approach to application deployment and execution that focuses on creating lightweight, single-purpose virtual machine instances tailored to run a specific application. Unlike traditional virtualization methods that involve running complete operating systems on virtual machines, unikernels are highly optimized, minimalistic environments that include only the necessary components to support a particular application. Unikernels are typically compiled as a single, self-contained image that integrates the application code, libraries, and a minimal operating system kernel, stripping away unnecessary functionalities found in traditional operating systems. This streamlined design results in reduced resource overhead and improved performance compared to traditional virtual machines or containers. Unikernel virtualization is particularly suitable for use cases where performance, security, and resource efficiency are critical, such as in embedded systems, IoT devices, and cloud-native applications. By eliminating unnecessary layers of abstraction and reducing the attack surface, unikernels offer a compelling alternative for deploying lightweight and secure applications in modern computing environments.

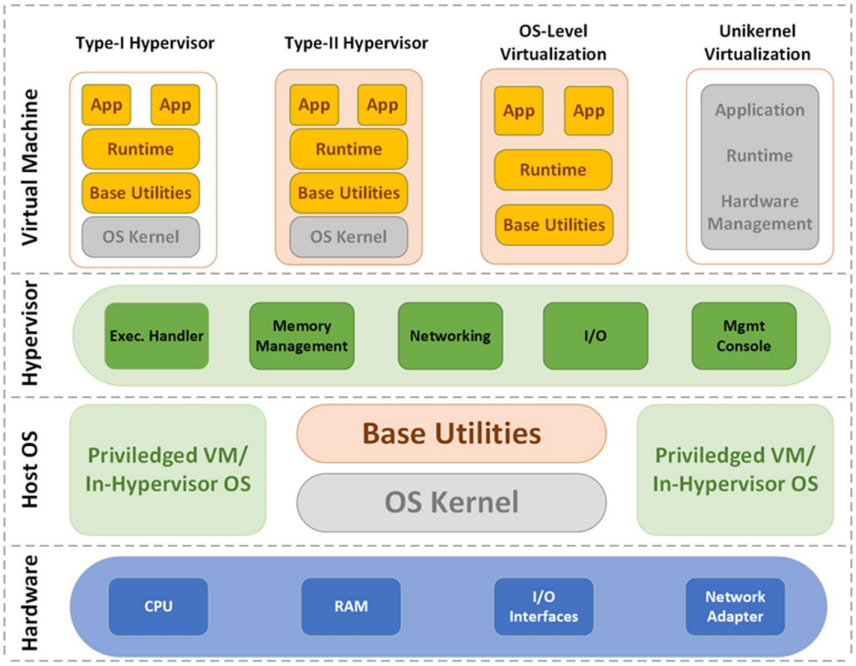


Figure 3. 3: Virtualization Architecture (adapted from [10])

3.3. Environment Overview

3.3.1. Software

3.3.1.1. Cloud Platform Tools

Openstack [11] is a highly scalable, open-source cloud operating system, developed through a global collaboration of developers and cloud computing experts. It serves as a widely adopted platform for building both public and private clouds. Openstack consists of a series of interrelated projects that deliver various components essential for cloud infrastructure, enabling the management of large pools of compute, storage and networking resources across a data center. These resources are centrally managed through OpenStack's web-based dashboard, which provides administrators with full control over the environment. At the same time, it empowers users to provision and manage resources through an intuitive web interface. The main openstack components [13] as illustrated in Figure 3. 4 are:

- Compute (Nova): Nova serves as the cloud computing fabric controller. It is responsible for deploying and managing large numbers of VMs and instances to handle various computing tasks in the cloud.
- Object Storage (Swift): Swift is a scalable, redundant storage system designed for objects and files. Data is written to multiple disk drives spread across servers in the data center. Swift ensures data replication and integrity across the cluster.
- Block Storage (Cinder): Cinder provides block-level storage devices for use with OpenStack compute instances. It manages the creation, attachment, and

detachment of block devices to servers, allowing persistent storage, similar to traditional disk storage systems.

- Networking (Neutron): Neutron offers robust networking capabilities, enabling the management of networks and IP addresses in an efficient and flexible manner. It handles the provisioning and management of network connectivity for OpenStack resources.
- Dashboard (Horizon): Horizon is the graphical interface that allows administrators and users to access, provision, and automate cloud resources through an intuitive web-based platform.
- Identity Service (Keystone): Keystone is the central directory for managing users and their access to OpenStack services. It provides a unified authentication system and can integrate with existing directory services such as Lightweight Directory Access Protocol (LDAP).
- Image Service (Glance): Glance provides services for discovering, registering, and delivering disk and server images. These images can be used as templates when deploying new virtual machine instances.
- Telemetry (Ceilometer): Ceilometer collects usage data across the cloud. This data can be used for billing purposes, tracking each user's consumption of cloud resources.
- Orchestration (Heat): Heat allows developers to define the resources needed for a cloud application using a template file. This service automates the provisioning and management of infrastructure resources based on the application's requirements.

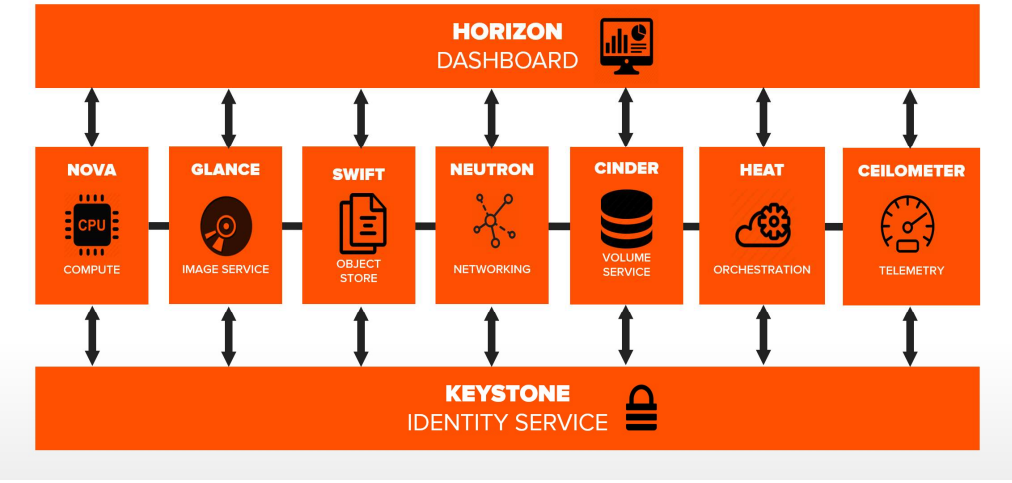


Figure 3. 4: Openstack main components [12]

One of the possible ways to deploy an Openstack cluster, is through MAAS (Metal as a Service) and Juju. **MAAS** [14] is a cloud platform for managing bare metal servers and virtual machines, enabling scalable automation, reconfiguration, and reliability for networks, machines, and OS images—all from a single point of control. Data centers face complex operational demands like uptime, reliability, security, energy efficiency, and hardware management. MAAS is specifically designed to address these challenges,

making it ideal for data center operators and administrators across various industries. This open-source solution, supported by Canonical, allows data centers to automate the management of physical servers for efficient on-premises operation. MAAS treats physical servers similarly to virtual machines, turning bare metal into an elastic, cloud-like resource pool. Rather than managing each server individually, administrators can provision machines automatically from a pooled resource. When a machine is no longer needed, it is released back into the pool for other uses.

Key Features:

- Web User Interface (UI) and Full API/Command Line Interface (CLI) Support: For easy interaction and management.
- Multi-OS Installation Support: Including Ubuntu, CentOS, Windows, Red Hat Enterprise Linux, SUSE, and VMware ESXi.
- IP Address Management (IPAM): Manages IP addresses across networks.
- High Availability and Role-based Access Control (RBAC): Ensures system reliability and secure, role-based access.
- IPv4 and IPv6 Support: For modern and legacy networks.
- Network Management: Includes Dynamic Host Configuration Protocol (DHCP), Domain Name Server (DNS), DHCP relay, Virtual Local Area Network (VLAN), and fabric support.
- Time Management: Integrated Network Time Protocol (NTP) for infrastructure-wide synchronization.
- Hardware Management: Automated inventory of components, hardware testing, and support for composable hardware.

MAAS integrates all necessary tools for provisioning and managing physical infrastructure, making it Canonical's recommended solution for physical provisioning systems. It provides a streamlined experience, allowing data center operators to efficiently manage a large number of machines with the flexibility and ease of cloud-like resources.

Juju [15] is an open-source application management framework designed to simplify and streamline the deployment and management of applications across hybrid cloud environments. It enables users to shift from traditional configuration management to application-focused management through small, reusable units called charms. Juju minimizes the operational overhead by automating tasks such as deployment, configuration, scaling, integration, and day-to-day management of applications. It supports public and private cloud services, bare-metal servers, and local containerized deployments, allowing administrators to manage diverse infrastructure types with consistency and efficiency. Juju's application-centric approach helps teams orchestrate and optimize applications across complex cloud landscapes, enhancing agility and operational control.

3.3.1.2. 5G Platform Tools

free5GC [16] is an open-source project written in go, which develops a complete 5GC platform fully compliant with the 3GPP Release 15 standards. It is designed for research, experimentation, and learning purposes, providing a modular and scalable architecture

to simulate, test, and validate 5G core functionalities in non-commercial settings. The project implements a 5G core network platform that is compliant with the 3GPP standards and supports all the vital NFs such as AMF, SMF, PCF, UPF, N3IWF, NSSF, AUSF, NRF, UDM and UDR. Additionally, it can be deployed in multiple virtualized environments, either virtual machine or container-based such as Docker and Kubernetes. The ecosystem can also be integrated with other RAN open-source projects, thus providing an end-to-end 5G network testing environment.

OpenAirInterface (OAI) [17] is an open-source project focused on developing 3GPP-compliant software implementations of wireless communication networks, particularly for 4G LTE and 5G technologies. It provides a fully functional, end-to-end software stack for both RAN and CN components, enabling the development, testing, and experimentation of mobile networks. With OAI, it is possible to build a full mobile network that can operate on off-the-shelf hardware, allowing for end-to-end testing of communication systems. It is based on a modular and flexible architecture where the CN and RAN segments are developed as separate projects, allowing developers to integrate OAI with other systems or test individual parts of the network. This makes suitable for prototyping and research.

3.3.1.3. Monitoring Platfom Tools

Prometheus [18] is an open-source toolkit for system monitoring and alerting that has become popular among various organizations, backed by a vibrant community of developers and users. At its core, it collects and stores metrics as time series data. Metrics, in simple terms, are numerical measurements that help users assess the performance of their applications. In Prometheus, each metric is recorded with a timestamp and can include optional key-value pairs known as labels. Prometheus is designed to scrape metrics from both instrumented jobs and intermediary push gateways. The data collected is stored locally, and Prometheus can apply rules to this data to aggregate it, generate new time series, or create alerts. Visualization of the metrics is often done using tools like Grafana. The Prometheus ecosystem consists of several components [19], including the core Prometheus server, client libraries for integrating with application code, a push gateway for short-lived jobs, exporters that convert service metrics into a Prometheus-compatible format, and an AlertManager for managing alert notifications. Some of its standout features include a multi-dimensional data model, which allows metrics to be identified by both name and labels, as well as PromQL, a flexible query language that enables users to effectively analyze their metrics. Importantly, each Prometheus server operates autonomously, without reliance on distributed storage. In summary, Prometheus serves as a powerful monitoring solution and is particularly effective for recording numeric time series data, making it suitable for both machine-centric monitoring and highly dynamic service-oriented architectures. Its support for multi-dimensional data collection and querying is advantageous in microservice environments.

Grafana [20] is an open-source platform that enables users to query, visualize, and analyze metrics, logs, and traces from various data sources [cite Grafana]. It provides powerful tools for converting time-series database (TSDB) data into interactive visualizations and supports integrations with numerous data sources, such as Prometheus, InfluxDB, and Graphite, as well as SQL databases, cloud services, and other APIs. This flexibility allows users to monitor system performance, identify anomalies, and correlate events effectively. Grafana's templating and dashboard variables also enable the creation of customizable and reusable dashboards tailored to

specific use cases, offering insights into infrastructure health and application behavior. One of the key strengths of Grafana is its extensive visualization capabilities. The platform offers a variety of panels—building blocks for visualizations—that users can configure to display different types of data, such as graphs, heatmaps, and pie charts [21]. Users can build comprehensive dashboards by arranging these panels to showcase a range of metrics from multiple data sources, giving them full control over the presentation of their data. For example, panels can be configured to show CPU usage data from Prometheus, visualized through gauges or histograms. Moreover, Grafana supports plugins, allowing users to create or add custom panels, enhancing the platform's adaptability and the potential for community-driven development. Grafana's alerting system is another core feature, which allows users to set up notifications through various channels like email, Slack, PagerDuty, and more. This system ensures that users are immediately informed of any abnormalities or issues that require attention, reducing downtime and maintaining system health. Alerts are set up by configuring alert rules, which trigger notifications when conditions are met. Grafana also offers an annotation feature, enabling users to leave notes directly on visualizations, marking critical points or events. This is useful for contextualizing data, marking significant incidents, or providing guidance for further investigation. Overall, Grafana offers a robust and flexible platform for monitoring and visualizing metrics and logs from diverse sources. Its user-friendly interface, alerting mechanisms, and community-driven ecosystem make it an essential tool for anyone managing infrastructure and applications.

3.3.1.4. Management & Orchestration Tools

Open Source MANO (OSM) [22],developed by ETSI, serves as an orchestrator for end-to-end network services, leveraging open-source tools and development practices. It aligns with the ETSI NFV architecture, using a modular approach that enables scalability and easy replacement of modules when needed. OSM simplifies network service management by abstracting technical complexities, offering an ETSI NFV-aligned Information Model (IM) that automates and monitors the lifecycle of network functions, services, and slices across various infrastructures. OSM also provides a unified northbound interface (NBI) based on the NFV SOL005 standard for comprehensive control over system and service operations. It extends the concept of a Network Service (NS) to encompass virtual, physical, or hybrid network functions, treating all components equally and facilitating transport links between different sites on demand. Within the NFV-MANO architecture, OSM performs both NFV Manager and VNF Orchestrator roles, supporting Networks as a Service (NaaS). It offers two NaaS service objects: the NS, composed of VNFs, and the NSI, which aggregates multiple NSs as a single entity. OSM manages the entire lifecycle of VNFs through its LCM module and VNF Configuration and Abstraction (VCA) layer. The LCM module handles VNF/NS lifecycle events like instantiation, scaling, and upgrades, while the VCA layer provides a unified interface to the compute resources, regardless of the underlying virtualization technology. Together, they ensure efficient resource allocation and VNF management throughout the lifecycle.

3.3.2. Hardware

The physical infrastructure includes several hardware components as illustrated in Figure 3. 5. Specifically, on the right part of the figure can be seen the hardware components which are used for the 5G cloud testbed, comprising compute, network and storage resources and Software-Defined Radio (SDR) boards used for the RAN segment. The specifications of each compoment of the cloud infrastructure can be found in Table 3. 1.

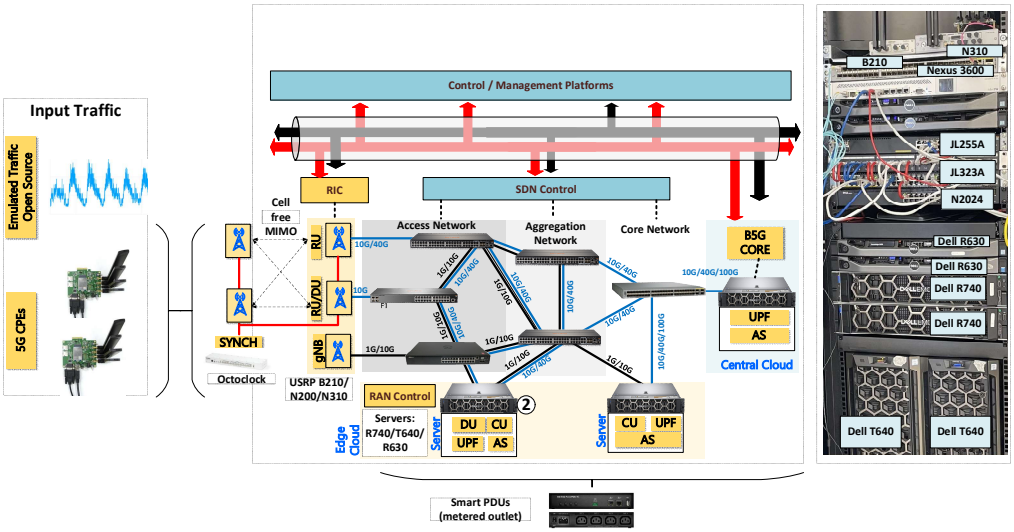


Figure 3. 5: Testbed Infrastructure

Table 3. 1: Specifications of Cloud testbed Hardware Components

Component	Model	CPU	RAM	Disk	Networking	OS
Server (2)	PowerEdge R740	Intel(R) Xeon(R) Silver 4110 CPU 16 Cores, 2.1 GHz	64 Gb	1.4 TB	10G	Ubuntu 20.04
Server (2)	PowerEdge T640	Intel(R) Xeon(R) Silver 4110 CPU 16 Cores, 2.1 GHz	64 Gb	1.8 TB	10G	Ubuntu 20.04
Server	PowerEdge R630	Intel(R) Xeon(R) CPU E5-2673 v3 48 Cores, 2.4 GHz	128 Gb	1.8 TB	10G	Ubuntu 20.04
Server	PowerEdge R630	Intel(R) Xeon(R) CPU E5-2630L v3 32 Cores, 1.8 GHz	64 Gb	1.8 TB	10G	Ubuntu 20.04
Component	**Model**	**Specifications**				
Router	Mikrotik CCR1009-7G-1C-1S+	**Ethernet Ports:** 7 (1 Combo Port), **CPU:** TLR4-00980 (9 Cores 1.2 GHz), **RAM:** 2 GB, **OS:** RouterOS				
Switch	Aruba 2930F	**Throughput:** 1G, **Ethernet Ports:** 24, **SFP Ports:** 4				
Switch	Aruba 2930M	**Throughput:** 1G, **Ethernet Ports:** 48, **SFP Ports:** 4				

Switch	Dell N2024	**Throughput:** 1G, **Ethernet Ports:** 24, **SFP Ports:** 4
Radio Board (3)	Ettus B210	**Frequency:** 70 MHz – 6 GHz, **MIMO**: 2x2, **Connectivity:** USB 3.0
Radio Board	Ettus N310	**Frequency:** 10 MHz – 6 GHz, **MIMO**: 4x4, **Connectivity:** Ethernet
5G Module	Quectel RM500Q	**Frequency:** Sub-6 GHz, Modes: LTE-A, 5G, **Connectivity:** USB 3.1
Energy Metering	Netio PowerPD U 4C	**Interfaces:** 2x LAN Ethernet, **Outputs:** 4, **Measurements:** Current [A], Consumption [kWh], Power [W], True Power Factor

3.3.2.1. 5G Cloud-Testbed

The 5G Cloud testbed architecture is illustrated in Figure 3. 6. At the bottom can be seen the physical infrastructure, which is pooled into a cloud cluster through Openstack platform and provides the virtual compute, network, memory and storage resources. The physical infrastructure comprising servers and networking components as described in section 3.3.2, is first clustered into a bare metal cloud through MAAS. This pool is used as a backing cloud to host the openstack microservices in LXD containers. The preparation of the environment and the installation of openstack microservices is executed through juju.

The virtualized resources provide the environment where 5G VNFs are deployed, typically in virtual machines. Physical hosts are connected to energy metering devices and all energy-related measurements are also stored in the monitoring platform. Finally, the testbed incudes a MANO platform which is based on OSM.

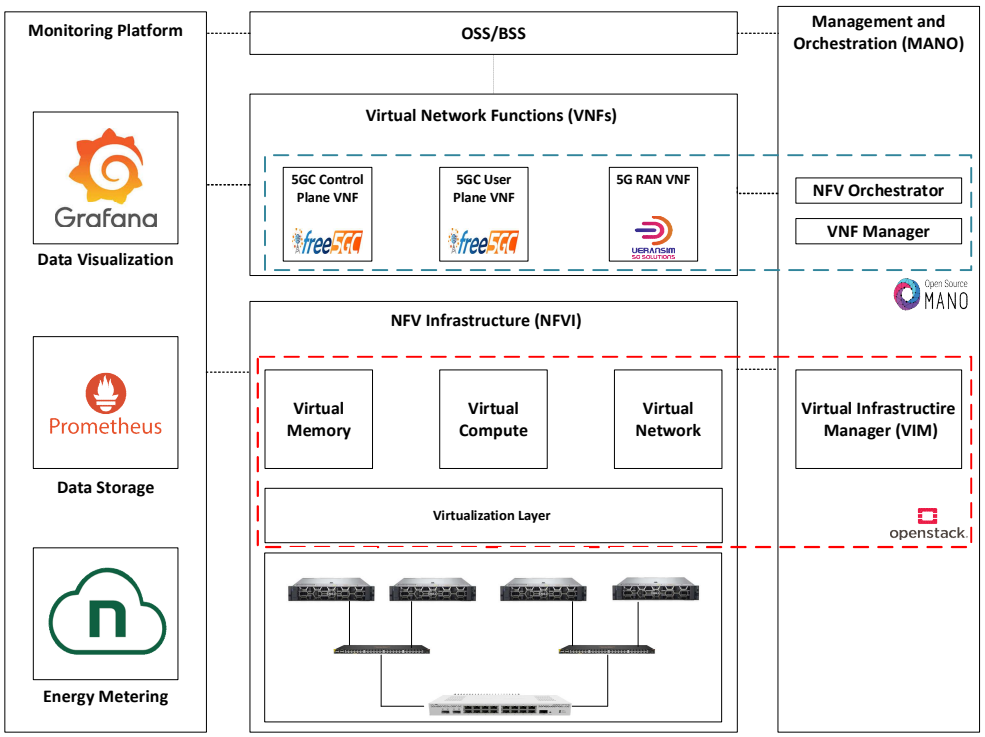


Figure 3. 6: 5G Cloud Testbed Architecture

Openstack Cluster: The cloud platform focuses on hosting 5G CN VNFs so the environment has been configured in order to ensure smooth and seamless VNF operation. First, we define the resources that can be allocated in a newly deployed virtual instance, called openstack flavors. Five different flavors are defined, the characteristics of which can be found in Table 3. 2. Additionally, several images are uploaded with various OSs to be used for the deployment of the VMs.

Table 3. 2: Openstack flavor Characteristics

Flavor Size	vCPU	RAM (GB)	Root Disk
Tiny	1	1	10
Small	1	2	20
Medium	2	4	40
Large	4	8	80
XLarge	8	16	160

Then, the networking environment must be configured. A virtual network is implemented comprising of a virtual router that offers connectivity between the internal subnets and external networks. As described in *Section II*, 5GC is based on a SBA where external interfaces communicate over point-to-point connections. For this reason, a separate subnet is created in our implementation for each point-to-point interface. This increases isolation and makes monitoring of each functionality easier. A virtual instance, depending on the NF it hosts will have interfaces on the relevant subnets and each subnet will be used only for communication of the protocol associated with this

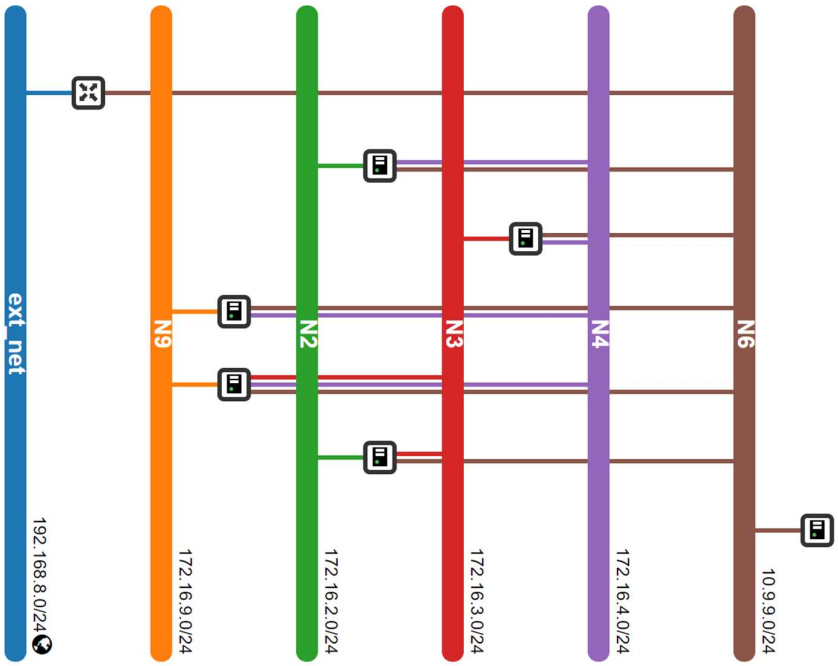


Figure 3. 7: Openstack networking configuration for the 5G platform

subnet. For example, a UPF node will have an interface on the N4 subnet for communication with the SMF over PFCP, and interfaces on N3, N6, N9 for data traffic over GTP-U. An illustration of the networking can be found in Figure 3. 7.

Openstack offers a framework to handle security aspects, called Security Groups where the administrator can activate/deactivate network ports, protocols etc. Consequently, for our 5G implementations we define a security group where all 5G protocols and their related ports are activated. A screenshot of the security groups table is provided in

Displaying 26 items

Direction	Ether Type	IP Protocol	Port Range	Remote IP Prefix	Remote Security Group	Description	Actions
Egress	IPv4	Any	Any	0.0.0.0/0	-	-	Delete Rule
Egress	IPv4	TCP	2000	0.0.0.0/0	-	-	Delete Rule
Egress	IPv4	TCP	3386	0.0.0.0/0	-	for GTP	Delete Rule
Egress	IPv4	TCP	5000	0.0.0.0/0	-	for webconsole	Delete Rule
Egress	IPv4	TCP	5201 - 5210	-	free5gc	for iperf connections	Delete Rule
Egress	IPv4	TCP	8888	0.0.0.0/0	-	jupyter notebook	Delete Rule
Egress	IPv4	TCP	9100	0.0.0.0/0	-	node exporter	Delete Rule
Egress	IPv4	TCP	20000	0.0.0.0/0	-	-	Delete Rule

Figure 3. 8: Openstack Security Groups configuration for 5G implementations

Figure 3. 8.

When the preparation of the environment is complete, openstack is ready for deployment of the instances. In Figure 3. 9, a screenshot from the Openstack Instances tab is provided, that shows information regarding each istance like its name, size, network configuration, power state etc.

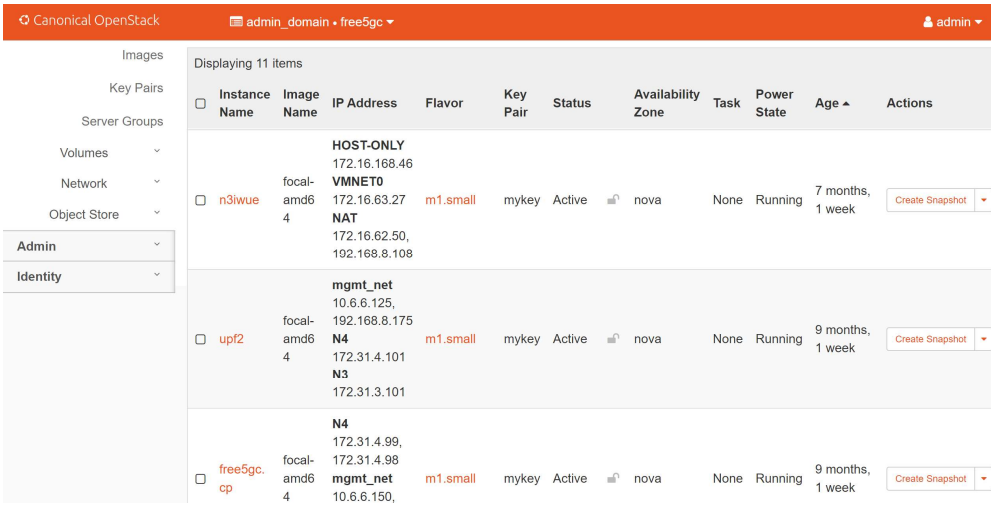


Figure 3. 9: Openstack 5G instances

Monitoring Platform: The monitoring platform is able to provide real-time monitoring at the infrastructure layer, the virtualization layer as well as the VNFs. In all hosts either physical or virtual, a network agent called “node exporter” is installed

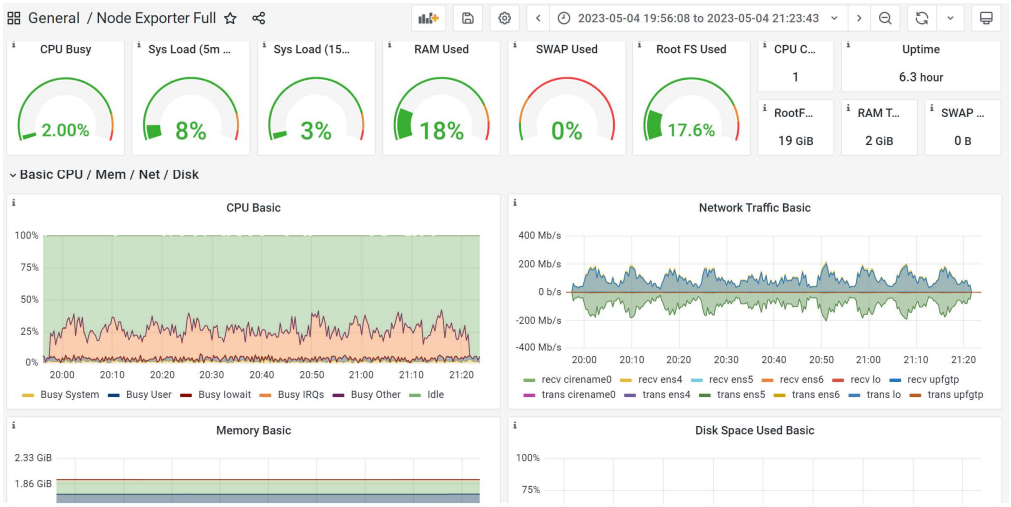


Figure 3. 10: Grafana Monitoring template

which gathers compute, networking, memory and storage metrics and exposes them at a certain port. From there, the statistics are gathered from Prometheus and stored in its database. Through simple queries, the data can be visualized through Grafana. Figure 3. 10 shows a screenshots from the Grafana visualization template that includes various resource consumption statistics.

Finally, the monitoring platform includes power consumption metrics provided by energy metering devices that are attached to the physical infrastructure. With all the different types of data gathered in the Prometheus database, a comprehensive networking benchmark can be performed in the cluster. The relative Grafana screenshot in Figure 3. 11 depicts load for varying number of UPF nodes with the same throughput passing from each node.

Figure 3. 11: Grafana power consumption screenshot

3.4. Deployment Options

This subsection provides a detailed description of the deployment options offered from the 5G cloud testbed. First, the main configurations are described that are either based on monolithic deployments or incorporate simple concepts (i.e. CP/UP split). Then, we present sophisticated implementations designed to meet the requirements of specific use cases and scenarios.

3.4.1. Main Configurations

The main deployment options for the RAN and Core segments are shown in Figure 3. 12. For the 5GC, the possible configurations are shown on top of the figure based on their level of isolation, where orange color indicated CP functionalities and light red color is associated with UP functionalities. The available deployment options (based on free5GC) are:

- Collocated: Where all 5GC related NFs are deployed in a single virtual instance that has a single interface for external network interactions. N1, N2 and N3 communications are carried over the same interface and all the other SBIs are realized on the internal loopback network.
- CUPS: Where the CP functions are split from the UP and deployed in separate hosts. The CP host has interfaces on the N2 and N4 subnet, while the UP host is connected to N3 and N4 subnets.
- Customizable: This type of deployments supports simultaneous operation of multiple UPF nodes that can be assigned with different roles (branching point, intermediate of PSA). Depending on their role the UPF nodes will have deployed

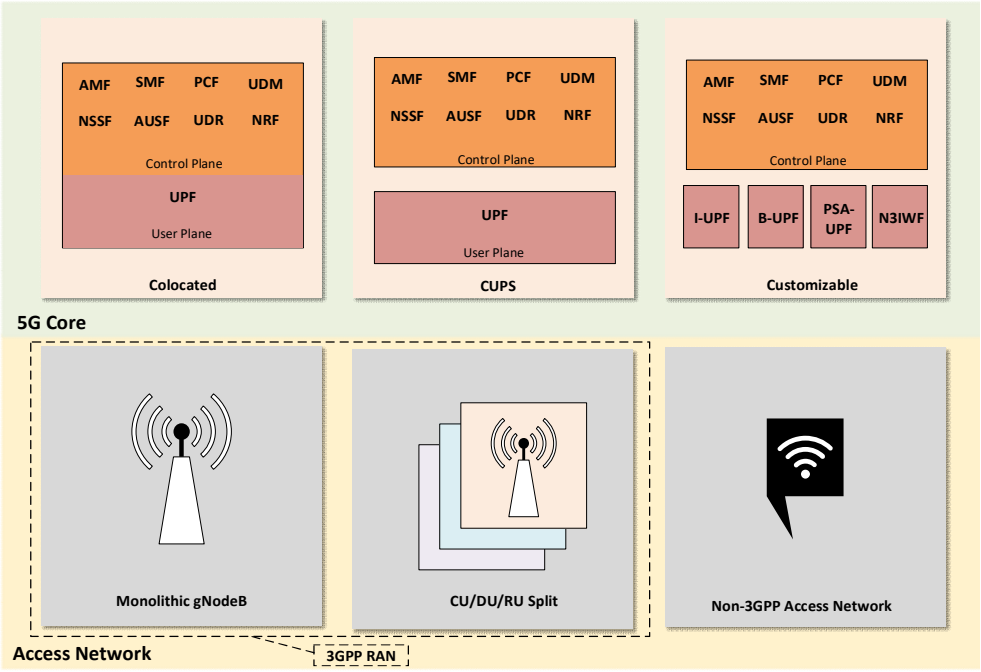


Figure 3. 12: 5G Main Deployment options. Core: Colocated, CUPS, Customizable. RAN: Monolithic, Functional Split, Non-3GPP Access Network

interfaces on N3, N6 and/or N9 subnets. Additionally, the N3IWF functionality is introduced to offer connectivity with untrusted non-3GPP access networks. VNF isolation can be extended to CP functions as well, for example in a case of "hard" slice deployment where each slice requires its own separate SMF NF. Scenarios of customized deployments are presented in the following subsections.

The access side of the network is shown at the bottom of Figure 3. 12. The left part inside the dashed box shows the deployment options based on 3GP 5G RAN. The implementations are based on OAI 5G RAN and end-to-end connectivity is offered through OAI 5G CN. This implementation supports the following deployment options:

- Monolithic gNB, where all the protocol stack and processing functions are collocated in the same site.
- CU/DU/RU split, where the RU handles the Digital Front End (DFE) parts of the physical layer, the DU is responsible for High-PHY(Femto Application Platform Interface (FAPI)), Medium Access Control (MAC), Radio Link Control (RLC) and RRC protocols as well as support for the F1 interface, and the CU contains both CP and UP functionality and support for PDCP, GTP-U, RRC and S1AP protocols over the interfaces S1, F1 and E2.

Moreover, access connectivity is offered from the testbed through non-3GPP access networks (right-bottom of Figure 3. 12). In this case, a customized 5GC with N3IWF enabled is required, which is then connected to the non-3GPP (Nwu) interface. Nwu may be bridged with different types of ANs, such as WiFi, LiFi, LoRaWAN etc.

3.4.2. Dedicated Network Slice Configuration

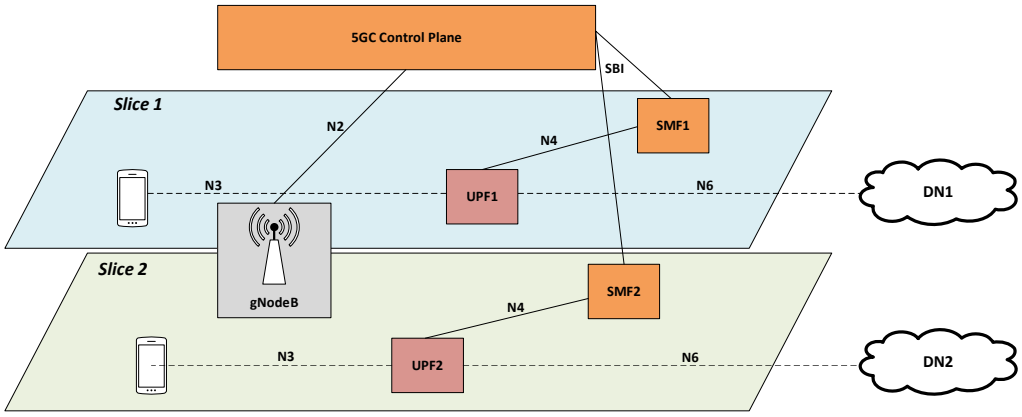


Figure 3. 13: Configuration of two "hard" network slices over a shared CP

This configuration involves and implementation where multiple "hard" network slices. A hard slice is one that does not share resources with other slices. In this case, each slice is configured with dedicated UPF and SMF functions over a common 5C CP which does not handle any critical slicing accept, so the slices can be defined as hard. An illustration of the topology is shown in Figure 3. 13 where two network slices can be seen, one marked with blue and one with green color. Network slices are generally differentiated by their SST and SD values (as discussed in section 2.3.1). Five virtual instances are used in total for the deployment of the CN: two UPF nodes, two SMF nodes and on node hosting CP elements. Two users are connected on the same emulated gNB which is

deployed on a separate machine. Each slice is differentiated by its SST an SD values and user data are routed through the appropriate UPF nodes base on these characteristics. Figure 3. 14 show captures of two packet captures from wireshark, one from each user of the deployed scenario. The NS parameters are located in the NGAP part of the IP packet (PDUSessionResourceSetupRequest), inside the Single-Network Slice Selection Assistance Information (S-NSSAI) field.

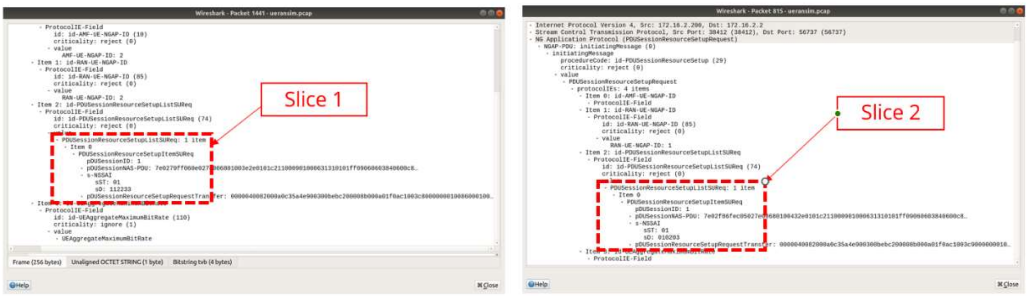


Figure 3. 15: Wireshark packet traces identifying the two configured slices

3.4.3. Uplink Classifier Configuration

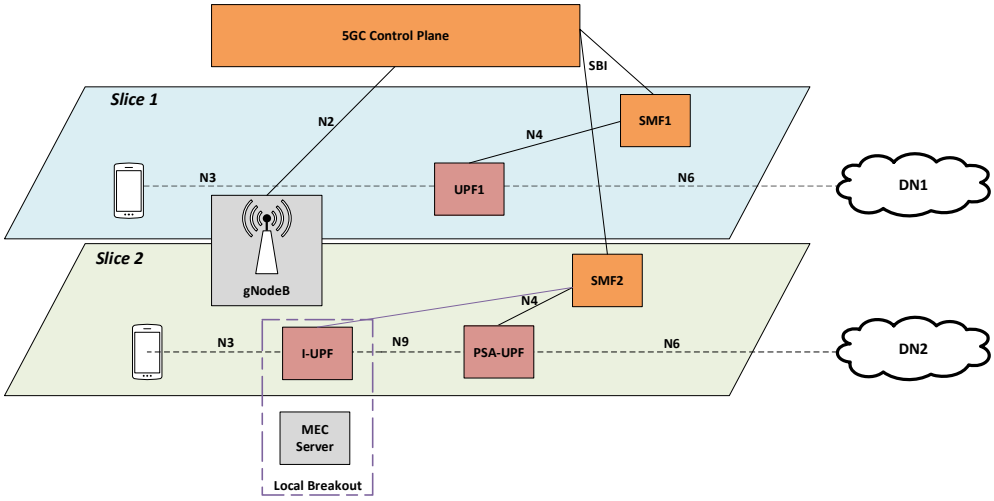


Figure 3. 14: Uplink Classifier Configuration in a multislice environment

The Uplink Classifier (UL/CL) is an innovative component in modern mobile networks, especially for QoS management and packet prioritization aspects. Its main function is to identify and classify different types of uplink traffic based on various attributes such as packet contents, source or destination IP addresses, port numbers and specific protocol types. This type of implementation can provide a variety of benefits in the network in the context of resource utilization, QoS enhancement and local traffic optimization. The UL/CL can be implemented as an extension of the dedicated slice configuration or a stand-alone deployment. Figure 3. 14 presents the former scenario which consists of:

- Six virtual instances for the CN: SMF1, UPF1, SMF2, I-UPF, PSA-UPF

- One instance for the MEC server which is reachable through a Local Breakout (LBO)
- Emulated RAN with two connected users

Figure 3. 16: Network Traffic of UL/CL configuration for various captured from Grafana: UE traffic, UPF1, I-UPF, PSA-UPF

In this scenario, UE1 traffic is directed through UPF1 which is handled from SMF1 and altogether they form an isolated slice. UE2, can be served either through the DN2 or from the MEC server, depending on its QoS characteristics in terms of latency. This is illustrated in Figure 3. 16 which shows a data traffic snapshot from the monitoring/visualization platform. On the top left part the traffic generated from the two UEs can be seen. UPF1 (UE1) traffic is shown on the top right snapshot. UE2 creates two traffic streams, one is a delay constrained video streaming application (served through the MEC) and the second is a file transfer. As seen from the bottom snapshots of Figure 3. 16, the traffic related to video streaming is terminated in the I-UPF while the file transfer passes through PSA-UPF in order to reach be served from the central DN2.

3.4.4. Multiaccess Connectivity Configuration

The last scenario offered from the testbed involves the configuration of multiaccess connectivity. In this deployment, a UE can be connected from a standard 3GPP RAN base station or through a non-3GPP AN such as WiFi. When connected to the 5G RAN, data traffic will follow the route through the gNB, to UPF1 and finally to the DN. In case of connectivity through a non3GPP AN the formed route will go through the Access Point (AP)-N3IWF-UPF-DN. In this setup which is illustrated in Figure 3. 17, the ANs are deployed in two separate hosts and the 5G CN is distributed on 3 virtual hosts:

- One host for the CP NFs
- One host for the UPF
- One host for the N3IWF

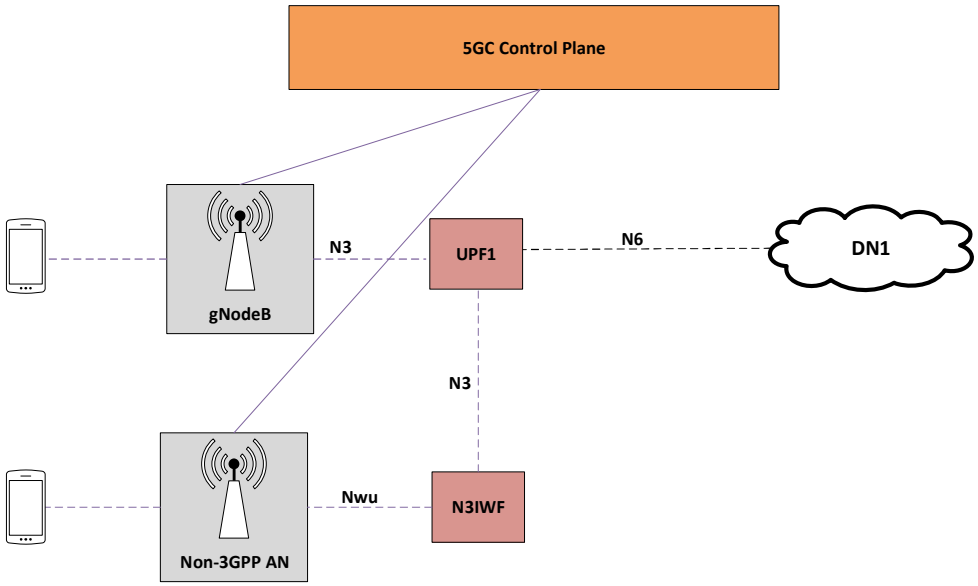


Figure 3. 17: Multiaccess connectivity configuration

3.4.5. Automated VNF Deployment

After manually deploying a 5G system such as the ones described above, OSM MANO platform can be introduced in order to enable the automation of the deployment process. This is performed by mapping the components of each topology to the appropriate VNFs. This is done by executing the following steps:

- First, for each topology the required VNFs must be defined. For example, if the CN is composed of a CP and a UP VNF, two VNF Descriptors (VNFDs) must be created, one for each VNF. The VNFDs define the specifications (in terms of resources) of the VMs that will host the 5G functionalities and the charms that dictate the actions that need to be performed for their proper configuration.
- Following this, a Network Service Descriptor (NSD), that defines the networking configuration needed for this VNF to operate is created. For example, for a UPF node a NSD defines connectivity to N3, N6 and N9 subnets.
- Finally, network slices are defined. The NSSIs that comprise each slice are investigated, along with their network connections and sharing capabilities. In OSM all this necessary information about the slice is stored in a Network Slice Template (NST).

NSTs are created, configurations must be applied to enable appropriate slice deployment. These configurations are delivered to the OSM via Juju charms and are categorized as follows:

- Day 1 actions (initial-config-primitive in the VNFD): Actions automatically executed during slice instantiation such as ssh access configurations, VM IP addressing, manipulation of configuration files, UPF role definition etc.

- Day 2 actions (config-primitive in the VNFD): Actions that can be dynamically executed during the ongoing deployment of the slice. For example Day 2 actions can be used for starting or stopping an NF (UPF, SMF etc).

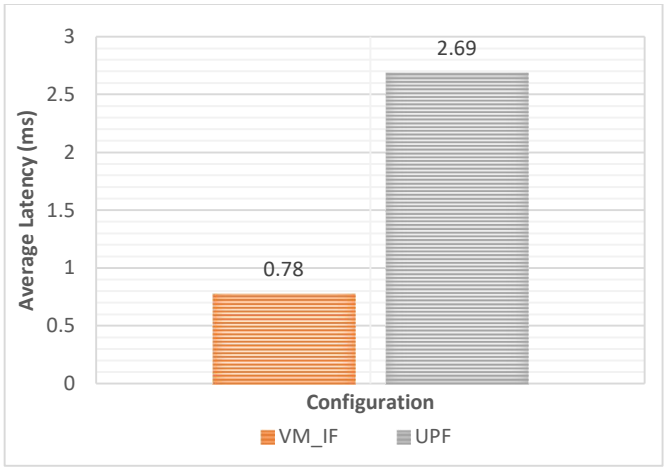


Figure 3. 18: Average latency. 5G System and Ethernet

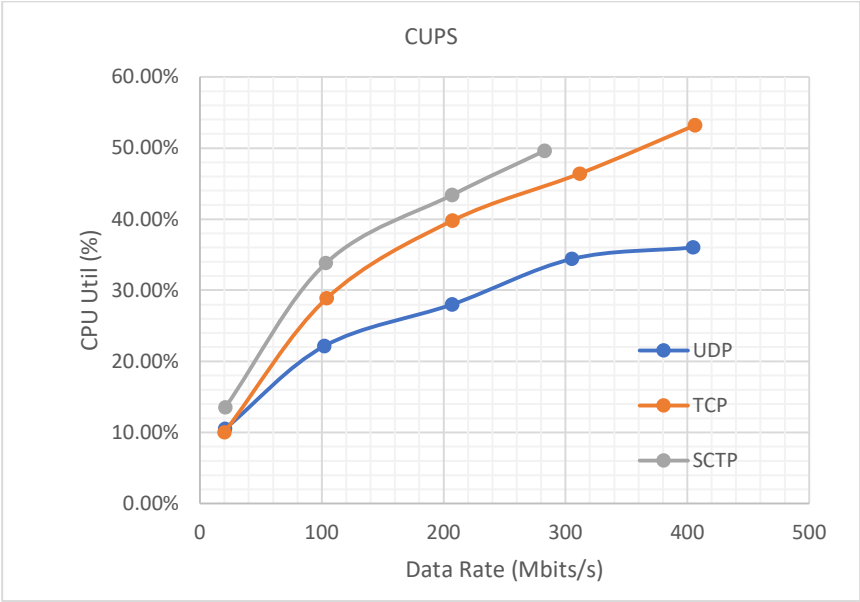


Figure 3. 19: Compute and network benchmark of UPF node by threee different protocols

3.5. System Profiling

One of the important capabilities offered from the cloud platform is the ability to correlate different types of statistics. This can be realized by executing specific scenarios and using the monitoring platform to quantify their impact in terms of resource consumption. For example, by creating network traffic through iperf connections from

a UE to an iperf server, network traffic statistics can be mapped to CPU utilization of UPF node. In this direction, Figure 3. 19 illustrates the CPU consumption of a VM hosting a UPF node for increasing data rates, where network traffic is carried over different transport protocols, namely UDP, TCP and SCTP. This logic can be extended to CP functions as well. Additionally, through extensive profiling other networking parameters can be defined as well, such as latency and jitter. Figure 3. 18 shows network latency performance of the 5G network in comparison with one of the external interfaces of the same VM. The measurements where gathered based on the ping command and it can be seen that 5G introduces some overhead. This is attributed to the use of the GTP-U protocol, and it is possibly related to the 5G platform software as well. Furthermore, in Figure 3. 21 the relationship between jitter and data rate is presented. In both figures, the purple color depicts the VM external interface while the grey color is indicates data passing through the 5G platform. In Figure 3. 20, the average load and average current, respectively, are shown for increasing number of benchmarked CPU cores, on three different servers of the testbed. The measurements were gathered by the energy metering devices while benchmarking the resources of various UPF nodes placed in the same physical host.

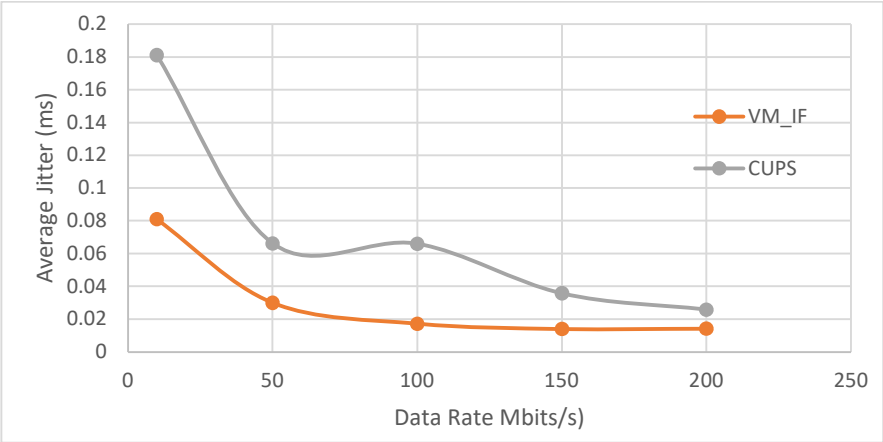


Figure 3. 21: Average jitter vs Data Rate. Comparison between GTP-U interface and external interace

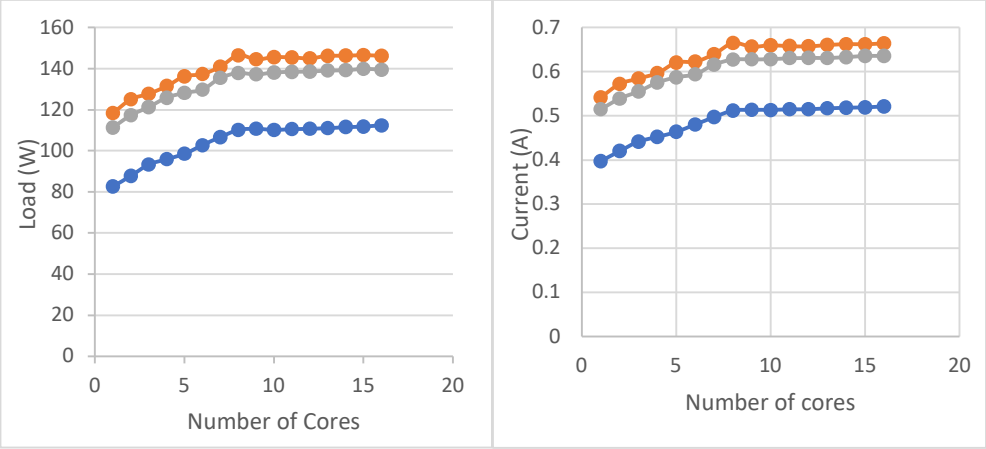


Figure 3. 20: Power consumption benchmark. Load, Current

3.6. Summary

In this chapter we presented the 5G Cloud testbed environment that was purposely developed to support the activities reported in this thesis. The testbed includes several components such as a private cloud platform, a monitoring and system profiling platform, a 5G platform and a MANO framework. First, an overview of cloud computing and related concepts is provided and then each of the components is presented. In this environment, a variety of 5G deployment options were implemented, targeting different scenarios, applications and use cases. Apart from a manual configuration, the instantiation of all 5G elements is automated through OSM software. Finally, we presented some system profiling results, in terms of compute, network and energy consumption.

References

[1] Peñalvo, F. J., Sharma, A., Chhabra, A., Singh, S. K., Kumar, S., Arya, V., & Gaurav, A. (2022). Mobile Cloud Computing and Sustainable Development: Opportunities, Challenges, and Future Directions. *International Journal of Cloud Applications and Computing (IJCAC), 12*(1), 1-20. https://doi.org/10.4018/IJCAC.312583

[2] Cloud Computing [online]. Available: https://en.wikipedia.org/wiki/Cloud_computing#Service_models

[3] Noor, T. H., Zeadally, S., Alfazi, A., & Sheng, Q. Z. (2018). Mobile cloud computing: Challenges and future research directions. *Journal of Network and Computer Applications*, *115*, 70-85.

[4] Introduction to Openstack [online]. Available: https://www.alhadeeth.bh/elibrary/assets/2012-aims-openstack-handouts-7e6xyh97.pdf

[5] A primer on Cloud Computng [onine]. Available: https://medium.com/@colinbaird_51123/a-primer-on-cloud-computing-9a34e90303c8

[6] Marinescu, D. C. (2022). *Cloud computing: theory and practice*. Morgan Kaufmann.

[7] Nutanix, "Understanding cloud computing service delivery models," *The Forecast by Nutanix*. [Online]. Available: https://www.nutanix.com/theforecastbynutanix/technology/understanding-cloud-computing-service-delivery-models. [Accessed: Oct. 15, 2024].

[8] Popek, G. J., & Goldberg, R. P. (1974). Formal requirements for virtualizable third generation architectures. *Communications of the ACM*, *17*(7), 412-421.

[9] Pearce, M., Zeadally, S., & Hunt, R. (2013). Virtualization: Issues, security threats, and solutions. *ACM Computing Surveys (CSUR)*, *45*(2), 1-39.

[10] Compastié, M., Badonnel, R., Festor, O., & He, R. (2020). From virtualization security issues to cloud protection opportunities: An in-depth analysis of system virtualization models. *Computers & Security*, *97*, 101905.

[11] Openstack [online]. Available: https://www.openstack.org/

[12] Pure Storage, "OpenStack overview: Modular collection," *Pure Storage OpenStack Documentation*. [Online]. Available: https://pure-storage-openstack-docs.readthedocs.io/en/zed_a/openstack-overview/section_modular-collection.html. [Accessed: Oct. 15, 2024].

[13] Rosado, T., & Bernardino, J. (2014, July). An overview of openstack architecture. In *Proceedings of the 18th international database engineering & applications symposium* (pp. 366-367).

[14] MAAS [online]. Available: https://maas.io/docs

[15] Juju [online]. Available: https://juju.is/

[16] Free5GC [online]. Available: https://free5gc.org/

[17] Openairinterface [online]. Available: https://openairinterface.org/

[18] Prometheus [online]. Available: https://prometheus.io/docs/introduction/overview/

[19] What is Prometheus – Use Cases [online]. Available: https://medium.com/@MetricFire/what-is-prometheus-use-cases-8613f3910ceb

[20] Grafana [online]. Available: https://grafana.com/docs/grafana/latest/introduction/

[21] What is Grafana [online]. Available: https://medium.com/@MetricFire/what-is-grafana-8de44d241765

[22] Open Source MANO [online]. Available: https://osm.etsi.org/

Chapter 4. OPTIMAL SERVICE PROVISIONING IN MOBILE 5G AND BEYOND SYSTEMS

4.1. Chapter Introduction

The rapid evolution of 5G/B5G networks presents a significant challenge in optimizing the design and deployment of the various network functions under mobility considerations. To address this challenge, the present chapter focuses on the joint optimization of the various 5G network segments. The main building blocks of a 5G system, including the 5G-RAN, the 5G-CORE segments and the DN hosting the end-users' applications are depicted in Figure 4. 1. In this environment, user services are established through end-to-end connections between the UE and the DN. However, in order to optimize these connections, the specificities and design principles of each building block of the 5G system should be taken into account.

As discussed in Chapter 2, 5G-RAN system relies on the disaggregation of base station node in three logical units, the RU, DU and CU respectively. Based on the placement of those elements, the gNB can be deployed as "monolithic", with all three units are collocated at the same site or as "disaggregated", where the RAN function and protocols are split across different sites. While the first option offers backward compatibility with the LTE EPC, the second enables the separation of the CU-CP and CU-UP, allowing further optimization for the location of different RAN entities.

The DU mainly hosts some or all the lower layer-2 protocols and physical layer processing such as RLC, MAC, and PHY, while the CU functions include non-real-time upper layer processing (i.e. RLC, PDCP, and SDAP) as well as core network functions that have been moved to the edge of the network in order to enable MEC services. Additionally, in an effort to minimize the transport capacity and latency requirements between RU and DU(s), some of the PHY processing can be moved to the RUs. In order to enable connectivity in the proposed architecture, new logical interfaces are introduced in support of the gNB functionality and, specifically, the FH interface for the interconnection of the RU with the DU and the F1 interface to support connectivity between CUs and DUs. Furthermore, significant effort has been put in the design of the 5G RAN segment. This involves virtualization of the DU, CU elements and identification of the optimal location where compute nodes can be placed [2] .

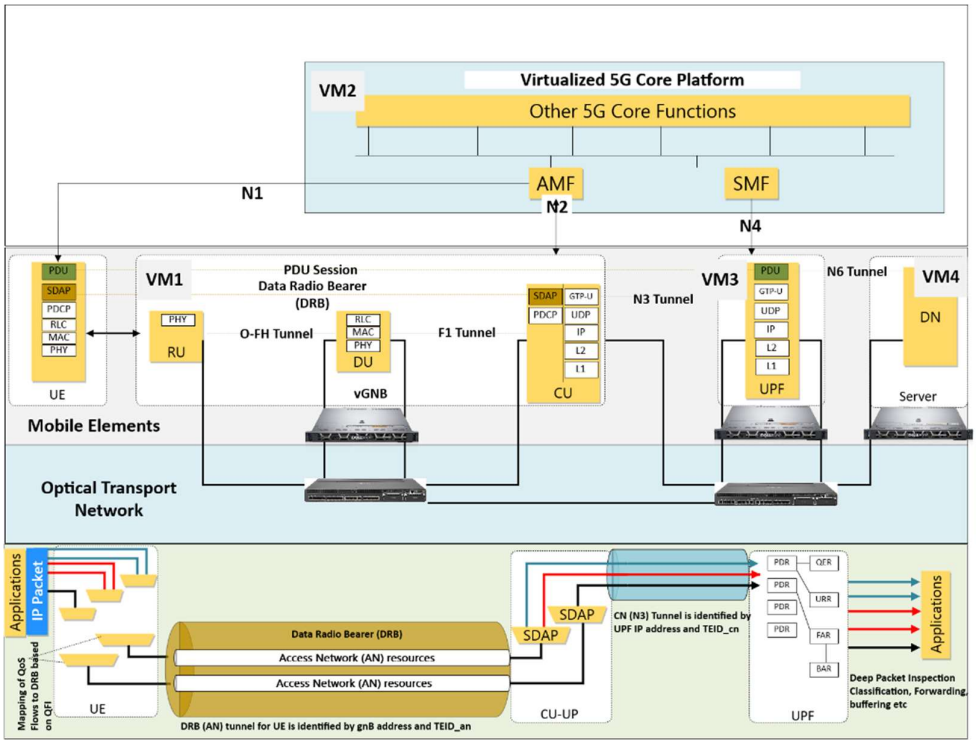


Figure 4. 1: 5G system architectural approach

In addition to the 5G-RAN design, provisioning of end-to-end services over the 5G-RAN and 5G-CORE elements is also critical. The required connectivity between UEs and DNs involves interconnection of several network segments at the user plane including:

- the UE and the physical gNBs or virtualized gNBs over the radio access technology
- the gNBs and the UPF over the N3 interface
- several UPFs with different roles via the N9 interface
- the UPF connecting the external DN/Cloud Server over the N6 interface.

From the above it is clear that a big part of user plane functionality in 5G Systems is handled by the UPF that must be designed to support challenging 5G services with very tight performance requirements. Part of the UPF's functionality is to set the data path between the UE and the DN and as such it is responsible for the PDU (i.e. the end-to-end user plane connectivity between the UE and a specific DN) session establishment and the maintenance of the UE connectivity under user mobility. UPF functionalities are controlled by the SMF through the PFCP interconnecting the 5G network with external IP networks, while it acts as an anchor point for the UEs. PFCP has a key role in 5G systems as it is used to define specific traffic management policies and install rules associated with Packet Detection Rules (PDR), Packet Forwarding Rules (FAR), QoS Enforcement Rules (QER), Usage Report Rule (URR) and Buffer Action Rule (BAR) at the UPF.

Based on the information that is included at the interfaces and the information that it receives from the SMF, the UPF can take several actions including:

- Mapping of traffic to the appropriate tunnels based on the QFI information. This requires UPFs to be able to perform Deep Packet Inspection and identify the necessary values in the GTP-U header, associate QFIs with the appropriate Differentiated Services Code Point (DSCP) codes in the external IP network and perform the relevant protocol adaptations (encapsulations/decapsulations) at line rate
- Steering of packets to the appropriate output port and take the necessary packet forwarding actions.
- Packet counting for charging and policy control purposes.
- Deep packet inspection for security and anomaly detection purposes.
- Buffering and queuing management for traffic service differentiation and assurance of end-to-end delays.

The above actions introduce scalability issues due to the large number of packet detection rules required to support policies (i.e., end to end services) subject to limited network resources (e.g., memory in UPF-compliant net-work interfaces).

Finally, user applications are implemented in the form of virtual functions hosted in VMs (or Containers) running on top of physical edge compute nodes. Hosting these functions in DNs placed close to the end-users (i.e. distributed at different edge locations), in accordance to the MEC architecture allows to satisfy strict 5G service requirements. This architectural approach integrated with 5G systems is expected to bring significant advantages for operators and vertical industries supporting ultra-low latency and highly reliable services, reducing the volume of traffic requiring backhaul and fronthaul services and enabling new 5G deployment options (i.e., private and public 5G systems sharing the same physical entities).

However, integration of MEC with 5G systems brings new issues and challenges that should be resolved. On the one hand, edge nodes usually have limited capabilities and are responsible to provide services targeting small geographical areas. On the other hand, mobile users such as smartphones and intelligent vehicles, tend to frequently move in between those small coverage areas [3]. Therefore, a main issue that is introduced is how network and compute resources are allocated when a user leaves the area of coverage of a MEC node and enters the area served by another MEC node. Under such scenarios a decision has to be made defining whether the service that is provided to the user will have to be migrated to the next Edge node or not [4]. Another challenging aspect is associated with the reservation of sufficient resources across all elements of the 5G system (RAN, Core and transport network providing connectivity between these) to support mobility. As users move from one gNB to another, PDU sessions with the required QoS Flow Identifier should be established. This requires reservation of specific resources to set up appropriate Data Radio Bearer (DRB) tunnels between the UE and the gNBs and N3 GTP-U tunnels between the gNB and the UPFs.

In addition to the PDU sessions, for services requiring access to a specific DN (i.e., MEC server) N6 tunnels between the UPFs and the MEC should be established and maintained for the whole duration of the connection of the mobile user. Therefore, a critical decision that should be taken by the SMF is when and over which elements these sessions should be established to ensure service continuity.

To precisely address this challenge, the present chapter focuses on the design of a B5G system by jointly optimizing the 5G-RAN, 5G-CORE and DN segments under mobility

considerations. This involves identification of the number of elements required and the location where CU, DU and UPF elements can be placed. These decisions are taken with the objective to minimize the operational expenditure of the network, subject to available compute and network resources and given the service related traffic demand. In addition to this, we examine if a VM supporting the operation of specific user needs to migrate to a new server or not based on a tradeoff analysis considering latency and network costs. Our results show that optimal operational points can be identified in cases where the requirements of all stakeholders involved (users, operators) are met.

The main contributions of our work [5] are summarized as follows:

- We formulate a mathematical problem that optimizes the deployment of 5G-NR systems under mobility considerations. The objective is to identify the location where the virtualized functions of the 5G system are hosted and the location of the VMs where user services are executed, in order to minimize the cost of the resulting network configuration.
- We carry out the analysis using as input to our mathematical model realistic measurements extracted from an actual 5G standalone system that has been deployed in a private cloud environment. In this system, detailed profiling has been carried out to evaluate several parameters affecting the performance of the system including the traffic and overhead generated during VM migrations, the impact of wireless network traffic on RAN and Core components, the correlation between wireless access network traffic and mobility.
- We evaluate the performance of the system under different 5G deployment strategies covering the RAN, the Core network and the application function. Each strategy has a specific optimization objective that is related to the minimization of the compute and network costs, minimization of overheads during migration of the application server and balancing of compute resources utilization. Our results show that the strategy considering mobility in the placement of the network functions demonstrates the best performance, while network and cost savings increase with UE mobility.

The rest of the chapter is organized as follows. Section 4.2 presents related work across the various network segments. In Section 4.3, a brief overview of the key components and functionalities of a 5G system is presented with emphasis on mobility management. The main parameters that are used to model the system are also introduced. The mathematical model used to optimize the design of the 5G system is presented in Section 4.4. Section 4.5 provides a description of the 5G standalone experimental platform used to measure and validate the proposed modeling framework. The relevant numerical results are presented in Section 4.6, while Section 4.7 concludes the reported work.

4.2. Related Work

4.2.1.1. Optimal RAN design

Optimal placement of the 5G RAN components (RU, DU, CU) is of key importance in the design of networks that aim to support applications with increased latency constraints and high data rates [6]. Towards this direction, in [2], the optimized placement of DU/CU nodes is formed as an Integer Linear Programming (ILP) model with a 3-layer RAN architecture, and the authors make some interesting observations regarding Baseband Unit (BBU) consolidation. The problem of dynamic selection of the Functional Split for the gNBs is formulated as a mixed-integer non-linear program in [7], which is then solved with heuristics, thus allowing its solution during runtime. Other parameters such as FH network topology and UE concentration are also evaluated. An optimization model targeting minimization of energy consumption and handovers (Apt-RAN) is presented in [8]. The study is based on real gNB energy measurements and proposes a lightweight polynomial time heuristic algorithm. Finally, an ILP-based BBU placement model, which aims at minimizing network cost, and a heuristic algorithm for large scale network scenarios are proposed in [9].

4.2.1.2. Optimal Core Network design

In addition to the RAN elements, a careful and optimal placement of the CN elements must be taken into consideration in 5G systems. This mainly concerns placement of the UPF node(s) involving actual user data traffic, as well as other elements such as the SMF, AMF etc. Relevant research work focusing on this area has already been reported in the literature. For example, in [10], two Mixed Linear Integer Programming (MILP) models are proposed addressing UPF placement in 5G networks, aiming to optimize a number of parameters such as latency and reliability and to minimize the number of required UPF nodes. The same authors are considering a dynamic UPF placement reconfiguration under user mobility in [11]. An evolutionary-based meta-heuristic algorithm is proposed in [12] for the placement of 5G NFs in relation with Network Slices. The authors claim that they achieve a near-optimum solution for their imposed constraints. Finally, a set of solutions for the placement of VNFs considering user mobility in 5G cloud environments is proposed in [13].

4.2.1.3. Optimal Application Server Placement

MEC network optimization has attracted great attention over the recent years and the relevant work is classified in two main categories: (i) network design optimization and (ii) network provisioning optimization. Regarding the former, ILP and heuristic techniques are used in [14] and [15] with the aim to minimize installation costs and average delay, respectively. Regarding network provisioning in MEC, the majority of relevant research focuses on optimizing service placement and migration [16]-[23]. Specifically, [15] aims to reduce the VM migration-induced overhead by designing two optimization algorithms, one that calculates the weight of certain moving trajectories and a heuristic that predicts uncertain moving trajectories. [17] proposes a dynamic virtual edge node placement scheme for Mobile Edge Cloud systems that implements Long Short-Term Memory (LSTM), aiming to optimize service quality, as well as reduce placement costs.

Considering the nature of the provisioned services, minimizing latency is of great importance and has been studied thoroughly in the literature [17],[19],[20],[21],[24]. In [17] a mobility-aware service placement framework is designed, that optimizes the latency/migration cost trade-off based on Lyapunov optimization. [19] formulates a Markov Decision Process (MDP) framework and presents a Deep Q-Learning algorithm with the aim to meet latency requirements and at the same time minimize system's costs. A predictive service placement algorithm with the aim to minimize the long-term time-average service delay is proposed in [20] that uses the two-timescale Lyapunov optimization method to incorporate user-mobility prediction. [24] proposes a mobility aware edge placement algorithm to minimize latency and deployment cost. In [21], network utility maximization theory is used for Mobility-Aware Service Provisioning to minimize latency.

Other related literature concentrates on optimizing the migration decision, that is, developing mechanisms to decide whether to migrate or not, considering the trade-off between migration gains and costs. [22] develops a service migration decision algorithm in MEC based networks, which is modeled as MDP and the migration decision strategy is solved by a Q-Learning algorithm. [23] solves a complex cost function for optimizing Service Migration with Reinforcement Learning.

4.3. System Model

4.3.1. 5G System Architecture Preliminaries

We consider a 3GPPP/O-RAN compliant architecture [1] relying on an IEEE 1914.1-2019 transport network solution [26], to interconnect a variety of general and specific purpose compute/storage elements. These elements combined can support a variety of 5G-NR deployment options spanning from the monolithic gNB deployment, where CU, DU, RU and their corresponding RAN functions and protocols are colocated within the same site to the disaggregated gNB deployment. Based on the split options adopted and the scale of the network, different RAN deployment options may be used including:

- independent RU, DU, and CU entities
- integrated CU and DU entities connected to RU
- integrated RU and DU entities connected to CU
- integrated RU, DU, and CU entities.

From the core network perspective, depending on the latency requirements of the mobile services, mobile core entities (e.g., UPF) can be placed at different locations of the 5G network. For services such as URLLC or critical mMTC, with very low latency and high reliability requirements (see 3GPP TS 36.211 [26]), mobile core entities can be located closer to the end-user i.e. network edge. In this context, RUs are located close to the DU and the DU is integrated with the CU. On the other hand, less latency sensitive services can be handled at the core site in accordance to the general architecture illustrated in Figure 4. 1. In this context, UPF is responsible for the establishment of connections between the UEs and the UPF (PDU sessions as shown in Figure 4. 1), the maintenance of the UE connectivity under mobility, and mapping of PDU sessions to the external network interconnecting the UPF with the DN [27].

It should be noted that in order to minimize deployment costs in 5G-NR systems, 5G RAN and 5G Core elements may be hosted at the same physical machines with the end-user applications. Given that all these elements are implemented in software, they can be hosted at VMs, containers or Unikernels running on compute nodes and placed either at the network edge or deeper in the network. However, as already briefly discussed integration of the MEC architectural approach with 5G systems brings new challenges that need to be addressed. These include dynamic and optimal allocation of network and compute resources when a user leaves the area of coverage of a MEC node and enters the area served by another MEC node [6].

4.3.2. Mobility Management in 5G Networks

The establishment of end-to-end connectivity in 5G systems, requires reservation of specific resources to set up the appropriate DRB tunnels between UEs and gNBs and N3 GTP-U tunnels between the gNB and the UPFs. In addition, N6 tunnels between the UPFs and the MEC should be established and maintained for the whole duration of the mobile user connection. Therefore, SMF has to decide when and through which elements these sessions should be established to ensure service continuity.

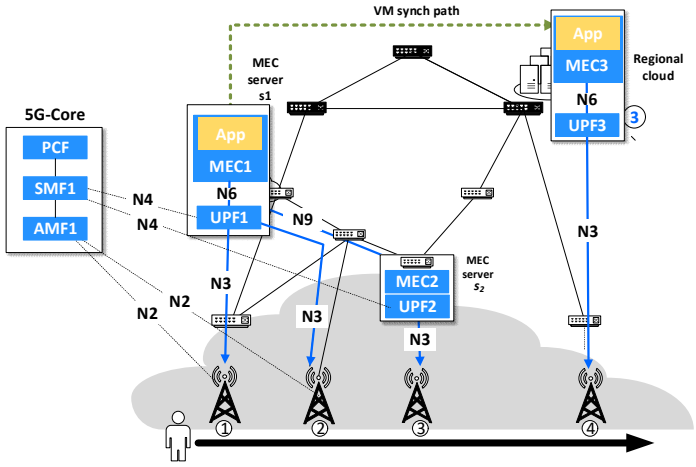


Figure 4. 2: The Joint user handover and VM migration problem to ensure service Continuity in MEC-assisted 5G environments.

To address this challenge, our study considers the adoption of joint user handover and VM migration to ensure service continuity in MEC-assisted 5G environments supporting advanced transport network connectivity. As an example of the supported functionality, consider the scenario shown in Figure 4. 2 where a mobile UE moves from a source gNB to a target gNB. This relocation triggers a handover-related signaling procedure that is implemented in 5G systems using the N2 interface. In the simplest scenario where the UE moves from gNB1 to gNB2, the handover process will trigger SMF to redirect the N3 tunnel from UPF1 to gNB2. As the UE moves from gNB2 to gNB3 a new intermediate UPF (UPF2) is inserted by the SMF. This new UPF is hosted in MEC2 and is used to provide the necessary connectivity between the gNB3 and the application (App) server through UPF1. As before, the SMF will establish an N4 session with the UPF3 in order apply the necessary rules to UPF3 and create an N9 tunnel between UPF1 and UPF3 and an N3 tunnel between gNB3 and UPF3.

In the above cases the Application Server (AS) is hosted in MEC1 and therefore, the connection through the N6 tunnel interconnecting the MEC with UPF1 remains unaltered. However, as the user moves to gNB4 the distance between the UE and the VM where the APP server is hosted increases leading to an increase in the end-to-end delay. In this case, the application will be transferred to a server that is closer to the location of the mobile user. To realize this, a path interconnecting the source (MEC1) with the target (MEC3) server should be established to enable migration of the user context from MEC1 to MEC3. This process, also known as live Service Migration can be used to move active VMs (or containers) along with their applications to appropriate servers. When considering the concept of VM migration in 5G environments it is clear that this decision should be taken jointly with the placement of the UPF. In Figure 4. 2 it is shown that once the migration has been completed a tunnel interconnecting gNB4 with MEC3 should be established through UPF3.

It is clear that to ensure service continuity for MEC-assisted 5G services a complex chain of several processes needs to be performed ensuring efficient allocation of connectivity between the UEs and the MEC nodes [11]. To successfully complete all these processes in a timely manner with reduced service disruption, several issues need to be considered during the service provisioning phase including allocation of: i) sufficient network resources for the establishment of the necessary connections between the 5G RAN and the 5G CORE elements, ii) sufficient computational resources to host not only the virtualized 5G functions (CU, DU, UPF etc.) but also end user applications and iii) availability of network resources for the interconnection of servers to perform live migration.

In response to this, a multi-stage optimization framework is developed in which a decision related to the placement of each VM to the appropriate servers is taken at each process stage. The objective of the proposed framework is to minimize network and compute costs for the provisioning of services to end-users with the required KPIs. This cost function considers the weighted average of the utilization of the network and compute elements also considering migration overheads. The analysis is based on realistic statistics for network traffic and users' mobility patterns as well as actual measurements for the VM migration process overheads.

4.3.3. Network and Mobility modeling

To study this problem, we consider a 5G network modeled as an undirected graph $G(\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ is the set of nodes and $\mathcal{E}$ the set of links. This 5G system comprises both RAN and core elements. The 5G-RAN segment comprises a set $\mathcal{R}, \mathcal{R} \subseteq \mathcal{N}$, of R gNBs used to provide connectivity services for a set $\mathcal{U}$ of U mobile users (see Figure 4. 3).

For gNBs, the concept of functional split is adopted according to which the RUs, DUs and CUs may be separated. The processing requirements of each gNB r have been calculated using the open source 5G platform OpenAirInterface (OAI) [28]. In accordance to the O-RAN standard, OAI supports split option 7.2, splitting the physical layer (PHY) into a high-PHY and a low-PHY allowing as already discussed gNB dissagregation into DUs and CUs. In this context, in our model the computational requirements denoted as $C_{rti}, i = DU, CU$ (measured in Giga Operations per Second – GOPS) can be determined for element $i, i \in \{CU, DU\}$ supporting RU r at time t.

To extract, CU and DU specific measurements OAI has been deployed in a containerized environment running as VNFs in Commercial Off-The-Shelf (COTS) MEC servers.

Apart of 5G-RAN elements, MEC servers also host core elements (i.e., UPF) which are necessary for the establishment of the required end-user services.

To establish a PDU Session, the UE must be registered with the AMF. PDU session establishment also involves communication between the UE the AMF and the SMF. At the end of the process, the SMF selects a suitable available UPF and the DN that will be used to support the requested service. Now let $\mathcal{N}_f \subseteq \mathcal{N}$ be the set of UPF nodes and $\mathcal{S}$ the set of servers. A UE connection is then established after the required resources setup to interconnect the UE with the RAN, the RAN with the UPF and the UPF with the DN. Therefore, as mentioned in the previous section, to ensure service continuity, necessary resources need to be reversed across the travelling paths of all users. To achieve this, accurate modeling of mobility patterns of the UEs is necessary.

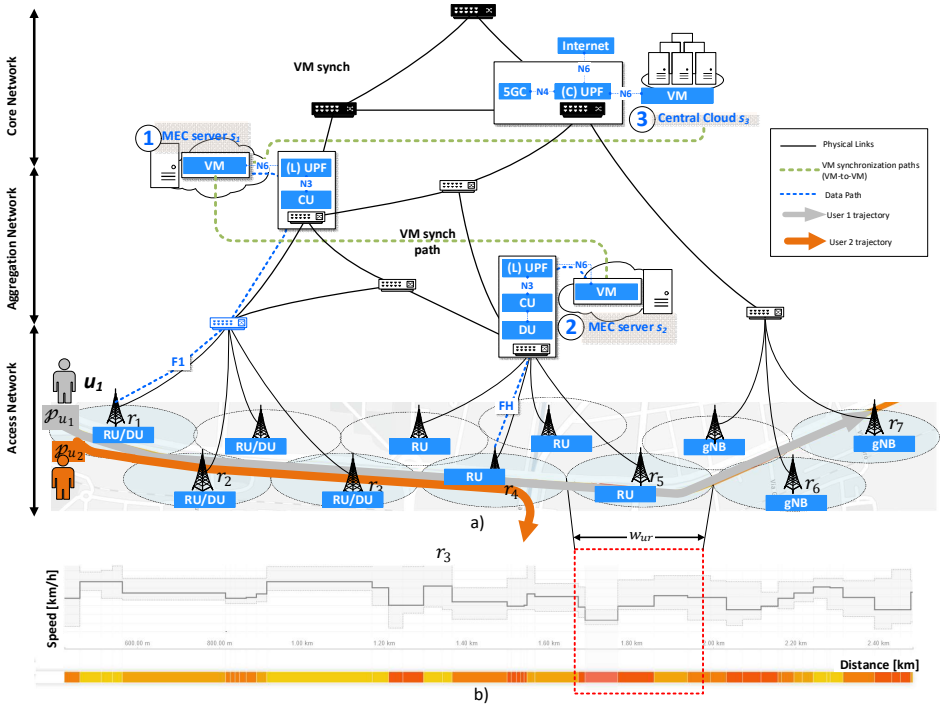


Figure 4. 3: 5G Network with mobility support.

To model user mobility, we assume that for each user u, there is a specific origin/destination (O/D) pair defined through the vector $z_u = \{z_u^{(o)}, z_u^{(d)}\}$ with $z_u^{(o)}$ being the origin and $z_u^{(d)}$ the destination locations (e.g. coordinates of UEs). The departure/arrival time for the O/D pair z_u is known in advance for every time period $t \in \mathcal{T}$, with $\mathcal{T}$ being the horizon (or number of times) where the control policy is applied. Each period t has a duration equal to $\hat{\tau}$. With regards to the timing events, we assume that the arrival of a user at the origin location $z_u^{(o)}$ occurs at time instant $\tau_u \in [t, t + \hat{\tau}]$, $t \in \mathcal{T}$. Based on the O/D pair z_u and using standard route planner software, the traveling path p_u for every user $u \in \mathcal{U}$ can be determined. Wireless connectivity across traveling path p_u is provided through a set $\mathcal{R}_{p_u} \subseteq \mathcal{R}$ of RUs. Each RU $r \in \mathcal{R}_{p_u}$ is responsible to provide network connectivity for a segment of the route p_u.

As mobile users travel across $\mathcal{p}_u$, the average time spent at every segment of the route can be readily determined based on history road traffic statistics. To model this parameter, Δ_{urt} is introduced to capture the residence time of user $u \in \mathcal{U}$ within the path segment covered by RU $r \in \mathcal{R}_{\mathcal{p}_u}$ during time period t. The corresponding normalized residence time (fraction of time spent is the area covered by RU r over the duration the time period $\hat{\tau}$) is given by $\widehat{\Delta}_{urt} = \Delta_{urt}/\hat{\tau}$.

Based on Δ_{urt} and τ_u, the time periods where each user is supported by the corresponding RUs in its trip can be determined. This is modeled through the binary coefficient ℓ_{urt} taking value equal to 1 if user u is located in the segment of the route covered by RU r at time t, 0 otherwise.

In addition, mobile users should be interconnected with the remote AS. ASs can be hosted at a set $\mathcal{S}$ of S servers that can be placed close to the edge (i.e., MEC) and/or at the metro/core CC regions. Connectivity between RUs and compute resources is provided through an optical transport network [29]. Based on the 5G deployment option adopted, this transport network can be used to support:

- The requirement of the FH protocol for the RU-DU interconnection
- the requirements of the F1 interface for DU-CU connectivity
- the requirements of the N3, N6 and N9 interfaces for CU – UPF, UPF-MEC, and UPF-UPF, respectively, connectivity.

At this point it should be mentioned that to provide services with QoS guarantees, specific KPIs (measured in terms of network throughput, processing and transport network delay) across the entire traveling path $\mathcal{p}_u$ should be satisfied. The transport/network, computational/processing and service delay requirement of services requested from user $u \in \mathcal{U}$ are denoted as h_u, π_u and d_u, respectively.

4.4. Problem Formulation

4.4.1. User plane design

Given a set of mobile users $\mathcal{U}$ requiring services with specific characteristics in terms of processing, latency and throughput, we investigate the optimal resource management problem in 5G systems under network and compute resource constraints. The objective is to design a virtualized 5G-NR system that can support end-to-end connections between the UEs and the AS (in accordance to the MEC model) ensuring service continuity under mobility considerations. This involves the creation of service chains traversing the virtualized gNBs and the UPF elements. The virtualized 5G-NR system comprises a set of DNs $\mathcal{S}$, a set of virtualized gNB $\mathcal{R}$, a set $\mathcal{N}_f$ of UPF nodes, a set of links $\mathcal{E}_{rft}$interconnecting gNB r with UPF f and, finally a set of links $\mathcal{E}_{fst}$ interconnecting UPF f with server s.

To solve this problem, we initially estimate the topology and resource requirements of the virtualized 5G-NR system using as input the traffic load, Λ_{rt}, for every RU $r \in \mathcal{R}$ at time $t, t \in \mathcal{T}$. Λ_{rt} is calculated by the summation of traffic generated by all users located within RU r and can be estimated through the following equation

$$\Lambda_{rt} = \sum_{u \in \mathcal{U}} h_u \ell_{urt}, \qquad \forall, r \in \mathcal{R}, t \in \mathcal{T} \tag{1}$$

Note that C_{rti} corresponds to the compute resources required to support UE traffic Λ_{rt}. We assume that each user can be served by a single server. To model this, the binary variable a_{urst} is introduced taking value equal to 1 if the AS belonging to user u located in the area covered by RU $r, r \in \mathcal{R}$ during time period t is processed at server s, otherwise this variable takes value equal to 0. This is mathematically described through the following equation that holds for every possible user location and every time period:

$$\sum\nolimits_{s \in \mathcal{S}} a_{urst} = \ell_{urt}, \ u \in \mathcal{U}, r \in \mathcal{R}, t \in \mathcal{T} \tag{2}$$

The establishment of end-to-end sessions between UEs and the MEC servers initially requires the establishment of a PDU session between the UE u and the UPF $f \in \mathcal{N}_f$ and then a connection between the UPF $f$and the MEC s. Now let $\mathcal{P}_{rfst}$ be the set of paths interconnecting gNB r to server s through UPF f and x_{upt} the volume of traffic from UE u allocated to path $p \in \mathcal{P}_{rfst}$. The following end-to-end session constraints should be satisfied:

$$\sum_{r \in \mathcal{R}} \sum_{f \in \mathcal{N}_f} \sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}_{rfst}} a_{urst}\, x_{upt} = h_u, \forall u \in \mathcal{U} \tag{3}$$

Introducing the binary coefficient δ_{fpt} taking value equal to 1 if UPF $f \in \mathcal{N}_f$ belongs to path $p \in \mathcal{P}_{rfst}$, 0 otherwise, the total traffic from all UEs terminated at UPF f, denoted as $\hbar_{ft}$, will be equal to:

$$\sum_{u \in \mathcal{U}} \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}_{rfst}} \delta_{fpt} x_{upt} = \hbar_{ft}, \forall f \in \mathcal{N}_f, t \in \mathcal{T} \tag{4}$$

Every path $p \in \mathcal{P}_{rfst}$ is formed by the interconnection of an N3 tunnel from gNB r to UPF f and an N6 tunnel from UPF f to server s. Now let δ_{uept} a binary coefficient taking value equal to 1 if virtual link $e \in \mathcal{E}_{rft}$ belongs to path $\mathcal{P}_{rfst}$ realizing transport network demands of user u at time t. Then, the capacity $y_{et}^{(r,f)}$ of link $e \in \mathcal{E}_{rft}$ interconnecting gNB r to UPF f will be equal to:

$$\sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}_{rfst}} \delta_{uept} x_{upt} = y_{et}^{(r,f)}, \forall f \in \mathcal{N}_f, r \in \mathcal{R}, e \in \mathcal{E}_{rft}, t \in \mathcal{T} \tag{5}$$

Following the same rationale, the capacity $y_{et}^{(f,s)}$ of link $e \in \mathcal{E}_{fst}$ can be estimated. To achieve this, the binary coefficient δ'_{ept} is used that equals 1 if link e belongs to path $p \in \mathcal{P}_{fst}$ inteconnecting UPF f with server s. $y_{et}^{(f,s)}$can be then estimated through:

$$\sum_{f \in \mathcal{N}_f} \sum_{p \in \mathcal{P}_{fst}} \delta'_{fept} \hbar_{ft} = y_{et}^{(f,s)}, \forall e \in \mathcal{E}_{fst}, f \in \mathcal{N}_f, s \in \mathcal{S}, t \in \mathcal{T} \tag{6}$$

Through the above model, the capacities for the interconnection of gNB r with server s can be determined. In the following section the RAN design problem (interconnection of RU, DU and CU and placement to the appropriate server) is presented.

4.4.2. RAN design

During the design of the 5G-RAN segment the model takes as input the aggregated traffic per RU expressed through (1). RUs are then interconnected through the O-FH interface to a set $\mathcal{D}$ of $\mathcal{D}$ DU hosted at servers $\mathcal{S}$. Introducing the binary decision variable

β_{rdt} taking values equal to 1 if RU r is connected to DU d at time t, the following RU-DU association constraint should hold:

$$\sum_{d\in\mathcal{D}} \beta_{rdt} = 1, \forall\, r \in \mathcal{R}, t \in \mathcal{T} \tag{7}$$

Equation (7) limits the number of DUs that can serve each RU to 1. Subsequently, each DU should be connected with a single CU. As before, the binary decision variable γ_{dct} is introduced taking values equal to 1 if CU c is used by DU d at time t. The following DU-CU association constraint should hold:

$$\sum_{c\in\mathcal{C}} \gamma_{dct} = 1, \forall\, d \in \mathcal{D}, t \in \mathcal{T} \tag{8}$$

Given that the BBU processing chain should be implemented in a specific order, the following constraint ensures that CU related processing activities will be performed after DU processing has been completed:

$$\beta_{rdt} \geq \gamma_{dct}, \forall r \in \mathcal{R}, d \in \mathcal{D}, c \in \mathcal{C}, t \in \mathcal{T} \tag{9}$$

The total CU, DU computational requirements for every time period can be described through the following equations:

$$C_{dt} = \sum_{r\in\mathcal{R}} C_{rtd}\, \beta_{rdt}, \quad \forall d \in \mathcal{D}, t \in \mathcal{T} \tag{10}$$

$$C_{ct} = \sum_{d\in\mathcal{D}} C_{rtc}\, \gamma_{dct}, \quad \forall c \in \mathcal{C}, t \in \mathcal{T} \tag{11}$$

CUs and DUs can be realized through virtualized resources hosted at server $\mathcal{S}$. Introducing the binary decision variables y'_{dst} and z'_{cst} taking equal to 1 if DU d and CU c, respectively, are hosted at server $s \in \mathcal{S}$, 0 otherwise, the following CU – DU server placement constraints hold:

$$\sum_{s\in\mathcal{S}} y'_{dst} = 1, \forall d \in \mathcal{D}, t \in \mathcal{T} \tag{12}$$

$$\sum_{s\in\mathcal{S}} z'_{cst} = 1, \forall c \in \mathcal{C}, t \in \mathcal{T} \tag{13}$$

The capacity for the interconnection of the RUs with the DUs and then the DUs with the CUs can be estimated adopting a similar rational with the analysis presented in the previous section. Specifically, let $\mathcal{P}_{rdt}$ be the set of paths interconnecting RU r to DU d and z_{rpt} the volume of traffic from RU r allocated to path $p \in \mathcal{P}_{rdt}$ supporting O-FH connectivity. Assuming that FH_r is used to describe the fronthaul capacity requirements of RU r at time $t \in \mathcal{T}$, the following constraints are introduced:

$$\sum_{d\in\mathcal{D}} \sum_{p\in\mathcal{P}_{rdt}} \beta_{rdt}\, z_{rpt} = FH_{rt}, \forall r \in \mathcal{R}, t \in \mathcal{T} \tag{14}$$

For the DU-CU connectivity a similar modeling approach is adopted. Assuming that $\mathcal{P}_{dct}$ is the set of paths interconnecting DU d to CU c and z'_{dpt} the volume of traffic from DU d allocated to path $p \in \mathcal{P}_{dct}$ supporting F1 connectivity with network capacity requirements$F1_{dt}$, the following equation holds:

$$\sum_{c\in\mathcal{C}} \sum_{p\in\mathcal{P}_{dct}} \gamma_{dct}\, z'_{dpt} = F1_{dt}, \forall r \in \mathcal{R}, t \in \mathcal{T} \tag{15}$$

Now let $\mathcal{E}_{rdt}$ and $\mathcal{E}_{dct}$ denote the set of links supporting the FH and F1, connections respectively, and ζ_{rept} a binary coefficient taking value equal to 1 if virtual link $e \in \mathcal{E}_{rdt}$ belongs to path $p \in \mathcal{P}_{rdt}$ realizing transport network demands of RU r at time t. Then, the capacity $y_{et}^{(r,d)}$ of link $e \in \mathcal{E}_{rdt}$ will be given through:

$$\sum_{d\in\mathcal{D}}\sum_{p\in\mathcal{P}_{rdt}}\zeta_{repf}z_{rpt}=y_{et}^{(r,d)},\forall f\in\mathcal{N}_f,r\in\mathcal{R},e\in\mathcal{E}_{rdt},t\in\mathcal{T} \quad (16)$$

Following the same rationale, the capacity $y_{et}^{(d,c)}$ of link $e\in\mathcal{E}_{fst}$ can be readily determined.

4.4.3. System Optimization

To identify the optimal location where virtualized 5G network functions need to be hosted, a multi-stage framework is developed. Stage 1 focuses on the identification of routing and function placement problem while Stage 2 decides whether a user VM should be migrated or not.

Stage 1- MEC selection for VM placement and routing decisions: During Stage 1 of the optimization process (time $t=t_0$), the 5G function and routing placement problem is solved. Assuming that $\mathcal{N}_s$ and $\mathcal{L}_e$ are parameters used to capture the server and network cost per capacity, the optimal strategies are identified by minimizing the following cost function:

$$f_1=\sum\nolimits_{s\in\mathcal{S}}\mathcal{N}_sC_{st_0}+\sum\nolimits_{e\in\mathcal{E}}\mathcal{L}_e\ C_{et_0} \quad (17)$$

In (17), C_{st_0} denotes the processing capacity used by all functions at server s. C_{et_0} is the total network capacity of link e used used for the realization of the disaggregated gNB r (interconnection of RU with DU and CU), the interconnection of gNB r with UPF f and finally the interconnection of UPF f with the AS s ($i.e.,C_{et_0}=y_{et_0}^{(r,d)}+y_{et_0}^{(d,c)}+y_{et_0}^{(r,f)}+y_{et_0}^{(f,s)}$). (17), should be minimized subject to the network and processing constraints described above. In addition to these, end-to-end delay requirements should be also satisfied. These are captured through:

$$D_{upt}\le d_u,\quad \forall\, u\in\mathcal{U},p\in\ P_{rs} \quad (18)$$

In (18), D_{upt} is used to describe the propagation delay of network flow originating from user u to server s across path $p\in P_{rs}$.

At this point it should be mentioned that as users travel across their selected paths, their distance from MEC servers may increase resulting in an increase in the end-to-end delay. To counterbalance this effect VM migration may be applied. To model this effect the binary coefficient $m_{ut[(rs)\to(r's')]}$ is introduced taking value equal to 1 if a user originally located within the range of RU r hosting its VMs at server s (in this case $a_{urst}=1$) decides to switch to server $s'\neq s$ while crossing the region covered by RU $r',r'\neq r$ ($a_{ur's't}=1$). This scenario is mathematically modeled through the following equation.

$$m_{ut[(rs)\to(r's')]}=a_{urst}\cdot a_{ur's't}\quad \forall u\in U,\quad \forall r,r'\in\mathcal{R}:r\neq r',\ \forall s,s'\in\mathcal{S}:s\neq s',\forall t\in\mathcal{T} \quad (19)$$

Given that the multiplication of the variables in (19) result in a non-linear equation, the following linearization constraints are introduced:

$$0\le m_{ut[(rs)\to(r's')]}\le 1\ \forall u\in U,\quad \forall r,r'\in\mathcal{R}:r\neq r',\ \forall s,s'\in\mathcal{S}:s\neq s',\forall t\in\mathcal{T} \quad (20)$$

$$m_{ut[(rs)\to(r's')]}\le a_{urst}\ \forall u\in U,\quad \forall r,r'\in\mathcal{R}:r\neq r',\ \forall s,s'\in\mathcal{S}:s\neq s',\forall t\in\mathcal{T} \quad (21)$$

$$m_{ut[(rs)\to(r's')]}\le a_{ur's't}\ \forall u\in U,\quad \forall r,r'\in\mathcal{R}:r\neq r',\ \forall s,s'\in\mathcal{S}:s\neq s',\forall t\in\mathcal{T} \quad (22)$$

$$m_{ut[(rs)\to(r's')]} \geq a_{urst} + a_{ur's't} - 1, \quad \forall u \in U, \forall r, r' \in \mathcal{R}: r \neq r', \forall s, s' \in \mathcal{S}: s \neq s', \forall t \in \mathcal{T} \tag{23}$$

To successfully complete the migration process, sufficient network and compute resources need to be assigned in order to keep the migration time (i.e., time needed for source and destination VMs to synchronize their states) as low as possible. The total migration time is affected by the volume of the virtualized resources that need to be transferred and the network capacity as migration involves the transfer of the entire volume of information (processes, memory, network) from one physical server to another. In addition to network migration overheads there is also a Pre-Migration Overhead which is introduced due to a set of operations that are not part of the live migration phases themselves. These include additional processes that need to take place involving selection of the destination host, initialization of the machines and reservation of the relevant resources. The Pre-Migration overhead does not depend on the size of the virtualized resources and can be considered as static overhead. Let h'_{ut} denote the network capacity needed to support the VM migration phase for user u at time period t and and ΔM_{ut} the total migration time i.e., the time period where h'_{ut} is reserved across the path interconnecting the source with the destination servers ($p \in P'_{(rs)\to(r's')}$). The normalized migration time is equal to $\hat{\Delta M}_{ut} = \Delta M_{ut} \backslash \hat{\tau}$. The network migration constraints can now be written in the following form

$$\sum_u m_{ut[(rs)\to(r's')]} \hat{\Delta M}_{ut} h'_u = C_{pt}, \quad \forall r, r' \in \mathcal{R}: r \neq r', \forall s, s' \in \mathcal{S}, \forall p \in P'_{(rs)\to(r's')}, t \in \mathcal{T} \tag{24}$$

$$\sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \sum_{r'\in\mathcal{R}\setminus\{r'\neq r\}} \sum_{s'\in\mathcal{S}\setminus\{s\neq s'\}} \sum_{p\in P'_{(rs)\to(r's')}} \delta_{pet} C_{pt} = C'_{et}, \forall e \in E, t \in \mathcal{T} \tag{25}$$

In (24), C_{pt} indicates the network capacity required to establish path p for the migration of a VM from server s to server s'. In (25), δ_{pet} is a binary coeeficient indicating whether link e belongs to path p supporting VM migration while C'_{et} is the capacity required by link e.

The server processing migration overhead is modeled through parameter π'_u. The additional capacity C'_{st} used at server s for VM migration can be then written as:

$$\sum_r \sum_s m_{ut[(rs)\to(r's')]} \hat{\Delta M}_{ut} \pi'_u = C'_{st}, \quad \forall s \in \mathcal{S}, t \in \mathcal{T} \tag{26}$$

Finally, the total processing and network capacity used for user-to-server and server-to-server connectivity should not exceed the total available capacity for server s (namely $\mathcal{C}_s$) and link e (namely $\mathcal{C}_e$) described through the following equations

$$\begin{aligned} C_{st} + C'_{st} &\leq \mathcal{C}_s \\ C_{et} + C'_{et} &\leq \mathcal{C}_e \end{aligned} \tag{27}$$

Once the first stage optimization problem has been formulated, the remaining server and network resources can be allocated to handle new resource requests and accommodate VM migration processes for mobile users.

Stage 2: VM migration and routing decisions under mobility: The second stage problem tries to identify whether a VM associated with user u that has been originally assigned at server $s_{t_0} \in \mathcal{S}$ need to be migrated to another server s_{t_1} during the next time period (i.e. time period $t = t_1$). The VM migration decision variable in this stage is $a_{ur's'_1}(a_{urst_0}; \xi_{t_1})$ which indicates, given the first stage decision a_{urst_0} if the VM of user u will migrate in the next time period to server s' after observing mobility

scenario ξ_{t_1}. Connectivity between servers s_{t_0} hosting VM user u during period t_0 and the candidate servers s_{t_1} hosting the same VM during the second stage of the problem (period t_1) is provided through a set of candidate paths $P'_{(rs)\to(r's')}$. It is clear that the decision variables of the second stage optimization problem that are responsible to forward and allocate the VM to server s_{t_1}, depend on the results of the first stage problem.

A typical example includes the set of paths $P'_{(rs)\to(r's')}$ that can be used to provide connecticity between MEC servers during the VM migration process. This set depends on the decisions taken by the first stage problem regarding the servers where a VM of user u has been originally placed. Other examples include the available capacity at the servers and network links. All this unknown information is revealed gradually as users move across a path. Hence, for each stage t_j the decision $a_{urs't_j}(a_{urs\,t_0}, \cdots, a_{urs\,t_{j-1}}; \xi_{t_2,\cdots}, \xi_{t_j})$ involves the t-th stage objective value and the goal is to minimize the expected value of the sum of these T objectives.

The optimal compute and optical network resource assignment problem in 5G environments can be solved through the minimization of the following nested cost function:

$$\min_{\mathbf{x}_0 \in \mathcal{X}_0} f_0 + \sum_{t=t_1}^{T} \mathbb{E}\left[\inf_{\mathbf{x_t} \in \mathcal{X}_t} f_t\right] \tag{28}$$

where extending (17) yields:

$$f_t = \sum_{s \in \mathcal{S}} \mathcal{N}_s(C_{st} + C'_{st}) + \sum_{e \in E} \mathcal{L}_e\,(C_{et} + C'_{et}) \tag{29}$$

(28)-(29) are minimized subject to the set of constraints described through (1)-(27).

4.5. Experimental Platform Description

To solve the problem of joint VM migration and mobility management in 5G systems, a 5G testbed has been deployed over a virtualized cloud environment allowing accurate estimation of network and compute resources consumed during the establishment of new UE connections.

These measurements are coupled with actual network traffic and user mobility statistics collected over an operational mobile network system. To quantify this cost, the 5G standalone version of OAI [27] has been deployed in the private cloud infrastructure, as mentioned above, shown in Figure 4. 4.

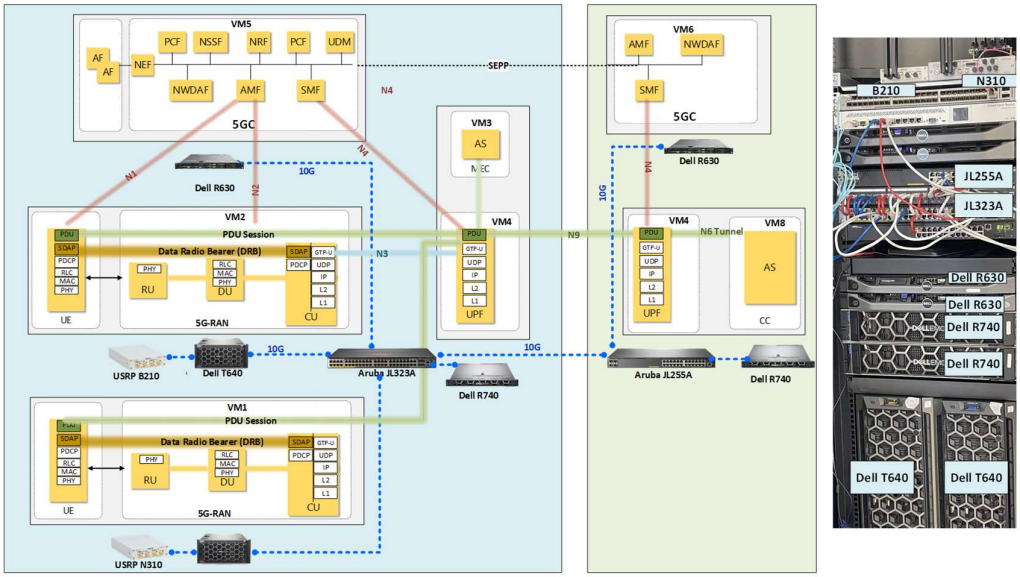


Figure 4. 4: Connectivity diagram (left) and experimental infrastructure used to host the 5G SA platform

Figure 4. 5: Traffic generated at a UE (top) traversing the UPF (middle and terminated at the AS (bottom).

In this environment, we consider the private 5G platform (shown in the blue box) comprising a set of softwarized RAN and core elements. These are hosted at physically separated servers. An AS has been also deployed at a different server. 5G UEs based on the Rel.16 RM500Q sub-6GHz module have been used to request access to the AS. The RUs have been implemented using N310 and B210 USRPs. The N310 has been configured to operate in a 2x2 MIMO, 100MHz bandwidth setting. All compute nodes are physically interconnected with an Aruba JL323A. An OAI based 5G SA Core platform has been also deployed having some basic 5G functionality (for our experimental purposes we deployed only AMF, SMF and UPF elements). These

elements are hosted in physically separated machines through an Aruba JL255A. Both switches are interconnected using a 10G point to point link. Network/compute resource utilization metrics have been collected through the monitoring platform Prometheus and have been visualized using Grafana.

Typical samples of the traffic generated by a UE, reaching the UPF in VM4 and, finally terminated at the AS in VM 3 is shown in Figure 4. 5.

Impact of VM migration of network traffic:

Figure 4. 6 a) shows an example of the network traffic generated during migration from a source MEC server to a target MEC server, measured in our experimental testbed. In this example, the AS has been configured to host a 4K streaming video server. During this live service migration process, the memory and disk state of the VM is transferred from the source host (VM8) to the destination host (VM3). Storage transfer is performed through a steady throughput, while memory transfer is done through multiple synchronization iterations. As mentioned above, a prerequisite for the success of the AS migration process is the availability of network and compute resources during the storage and memory copy phase. The availability of these resources depends on the area where the UEs move and the background network traffic. Higher background network traffic is observed in densely populated areas (e.g., city centers) where the speed of the mobile UEs is also lower.

In this set of experiments, we quantify the network cost of VM migration. As discussed above, as a user moves across the network, the distance from the application server (external DN) also increases resulting to an increase in the end-to-end latency. To keep latency below an acceptable threshold, a VM migration process may be triggered. To quantify the VM migration cost in terms of network resources, we have used Openstack's block live migration [30] which uses a pre-copy migration mechanism both for storage and for memory state transfer. The decision to use this migration mechanism was taken because it fits best our scenario, where migrations will take place between different infrastructures either at the Edge, or the Central Cloud. In Figure 4. 6 a), we can see a steady throughput that corresponds to the storage transfer, while the second

Table 4. 1: VM configurations used to host the virtualized 5GC platform

VM	Number of CPU Cores	RAM [GB]	Storage [GB]	Network Gbps
Small	1	2	20	1
Medium	2	4	40	1
Large	4	8	80	1
XLarge	8	16	160	1

part corresponds to the memory transfer iterations.

It is important to note that the total migration time depends on the network throughput as well as the size of the VM which will be migrated. In our configuration we use four different VM sizes with varying number of CPU cores, memory and storage. The details for all four configured VM sizes are provided in Table 4. 1.

Impact of wireless access traffic on core compute requirements

We also evaluate the computational requirements of the 5G core network as a function of throughput assuming that the system is hosted at VMs with different allocated computational resources. The results are also based on our 5G SA OAI platform. Figure 4. 6 b) shows the impact of the PDU session throughput on CPU resource utilization. These results are based on a set of tests that where performed using the network benchmarking tool iperf [31]. More specifically, the tests involve a UE that is connected to the 5G Core platform and generates traffic at different data rates, namely 10, 50, 100, 150 and 200Mbps over the UDP protocol. The same process was repeated with different resources allocated to the VM hosting the 5G Core (small, medium, large, xlarge). It is observed that as we allocate more resources to the Core Network, the CPU utilization is reduced. For instance, for the same data rate of 50Mbps, the small host consumes 19.9% of its CPU, while the medium host consumes 11.8%, and as the VM size increases to large and xlarge, the CPU consumption is decreased (5.6% for the large host and 4% for the xlarge). It should be clarified, that since our experiments mainly involve User Plane traffic, the Network element responsible for the CPU consumption is the UPF. This speculation was verified by repeating the test in our lab with a CP-UP split configuration.

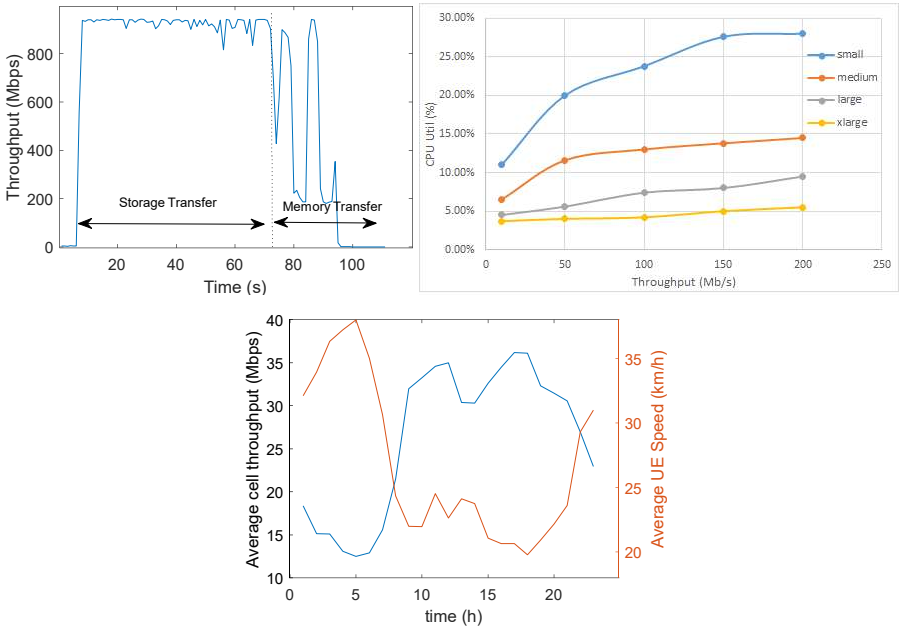


Figure 4. 6: a) Time series showing the traffic generated during VM migration from a source to a target VM. b) Impact of average UE throughput on 5GC computational resources b) Correlation between background mobile network traffic per gNB and speed per UE

From the above it is concluded that possible migrations associated with a user moving from a gNB covering a sparsely populated region to a densely populated region should be treated carefully as service disruptions may occur.

Correlation between Wireless Access traffic and UE Mobility

The interrelation between the average mobile traffic per gNB and the average speed per UE within the area covered by the gNB is shown in Figure 4. 6 c). The relevant traces have been captured from an operational mobile environment, whereas average speed statistics have been collected from GPS trackers. Mobile network and mobility traffic statistics indicate that under time periods with average UE speeds (low traffic

conditions) the average mobile network traffic is low as the number of active UEs is limited. However, for time periods and regions where the average UE speed is low (traffic jamming), mobile network traffic is high due to the large number of users that exist in the area.

4.6. Numerical Results and Discussion

A comparative analysis of different 5G network deployment options in terms of network cost (total capacity in Gbps), compute cost (number of vCores) and monetary costs is shown in Figure 4. 8 and Figure 4. 7 for a low and a high mobility scenario, respectively. For the comparisons we have considered four different optimization objectives:

1. 'Network-Aware' policy that in the optimization process considers only the network resources required for the system to operate.
2. 'Compute-Aware' policy that in the optimization process considers only the compute resources required for the system to operate.
3. 'Balancing-Aware' policy that tries to uniformly allocate all compute tasks to the servers. In this policy the average utilization per server can be reduced
4. 'The 'Migration-Aware' that tries to minimize the VM migration overhead under mobility scenarios

For the monetary costs of compute and network resources the values reported in [32] have been considered. More specifically, for network resources we have assumed a price of 0.33€/hour/Gbps, 1 while for compute resources we have assumed an on demand hourly price of €0.11 per vCore. The results have been extracted under different mobility scenarios. The MEC has been configured on a single thread per core, while central cloud supports two-threads per core.

Under the low mobility scenario, our results illustrated in Figure 4. 8 a) indicate that the policy minimizing network resources is the lowest network cost policy, as in this case all virtualized entities are placed close to the RUs. The relevant results have been extracted assuming that the end-to-end delay (measured in terms of number of hops) from the UE to the AS is 2. We also observe that the policy that minimizes migrations has higher network cost as in this case the AS is placed in more centralized locations to reduce the number where the migration process is triggered. High network costs are also observed for the policies that minimize the compute costs and balance traffic across servers. These policies aim at minimizing the number of active servers in the 5G system satisfying at the same time the required KPIs. This results in overall higher network resource requirements as longer end-to-end function chains are established.

Figure 4. 8 b) shows the derived compute cost under different optimization strategies. The compute-aware policy achieves the best performance as it optimally allocates functions to the appropriate compute nodes increasing centralization gains (i.e. a single UPF hosted in a more centralized location can support multiple RUs). As expected, the compute cost for the other placement strategies is higher as all entities of the service chain are placed close to the RUs requiring instantiation of additional 5G Core and RAN elements. A good compromise is observed in the migration-aware strategy requiring less compute resources to operate compared to the network or balancing aware policy.

[1] https://aws.amazon.com/directconnect/pricing/

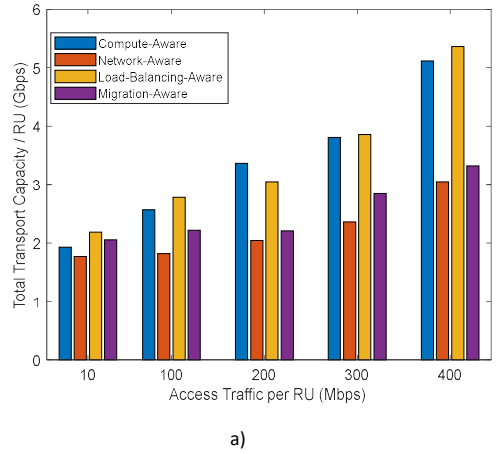


Figure 4. 7: Time: Comparative analysis of different 5G network deployment strategies in terms of a) network, b) compute and c) total costs under low mobility

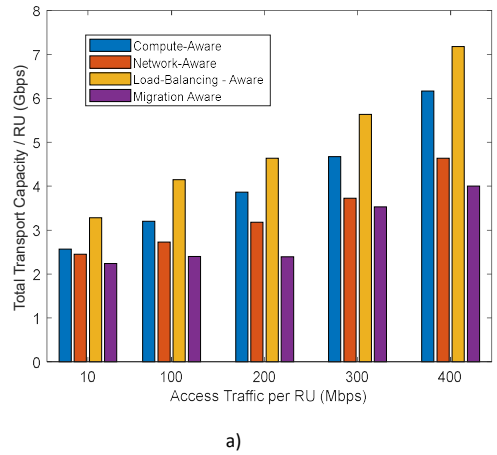


Figure 4. 8: Time Comparative analysis of different 5G network deployment strategies in terms of a) network, b) compute and c) total costs under high mobility

The combined network and compute cost in monetary values per RU is shown in Figure 4. 8 c). Given that the network connectivity costs are higher compared to compute costs, the most efficient deployment option is to place all elements close to the edge. It is also observed that under low mobility, the migration-aware strategy has almost the same deployment costs with the network-aware policy as the higher network costs in the 'migration-aware' policy are counterbalanced by the savings in compute costs.

The performance of the different deployment strategies under high mobility is shown in Figure 4. 7. Under high mobility the overall compute and network costs increase for all strategies. Under high mobility additional resources need to be allocated across all segments of the system to ensure seamless handovers. This includes additional resources for the UPFs to handle traffic steering, as well as network and compute capacity to support VM migration tasks. However, the relevant cost increase for the 'migration-aware' strategy is smaller. By predicting UE trajectories, the migration-aware policy places 5G elements to appropriate positions and also suitably sizes their capacities reducing associated overheads. The relevant results for network, compute and total costs are shown in Figure 4. 7 a), b) and c) respectively. Therefore, under high mobility we observe that the policy which minimizes network cost places ASs close to the UEs. In this case, the migration process is triggered frequently as ASs will follow the users mobility patterns: every time a user moves to a new gNB, the AS will be placed to a closely located server. The migration overhead is also very high and independent of the end-to-end service delay. On the other hand, the policy that minimizes the number of migrations results in high overheads when end-to-end delays are strict. By relaxing these constraints, the scheme predicts the future position of the UEs and optimally places the AS' in order to minimize the associated migration overhead.

The impact of wireless access traffic on the percentage of RUs that have selected to collocate all elements of their chain (DU, CU. UPF, VM) at the same physical machine (all-in-One deployment option) is shown in Figure 4. 9. We observe that as the access traffic per RU increases, the percentage of RUs that select the all-in-One deployment option increases, as the capacity of the servers and/or the transport network may not be sufficient to transfer and process all elements in the same physical machines. It is also observed that under low traffic conditions the all-in-One deployment option is selected by the majority of the RUs, while the option to split processing is beneficial when network traffic exceeds a specific threshold threshold.

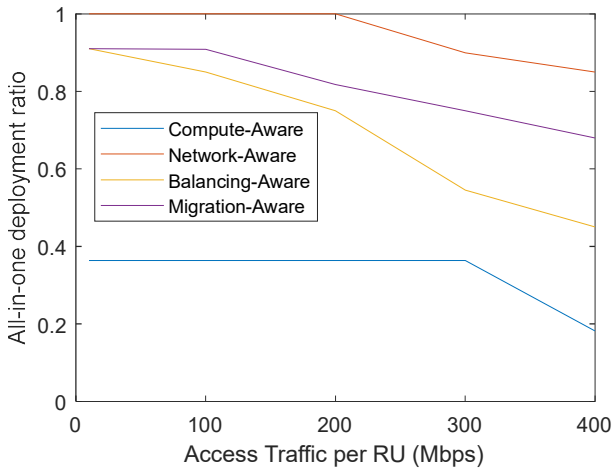


Figure 4. 9: Ratio of Access Points selecting the all-In-One deployment

4.7. Conclusion

This chapter provided an overview of 5G and beyond architectural options and proposes suitable resource allocation schemes to enable optimal service provisioning taking into consideration user mobility requirements. In this context it proposes optimal resource

management of 5G RAN and Core NFs as well as DN elements in the presence of user mobility taking also advantage of the option of Live VM Migration. Towards this direction, a multistage optimization framework has been purposely developed to enable optimal resource allocation of both network and compute resources minimizing the network operational costs. Our analysis exploits real user mobility statistics as well as lab based profiling measurements of the 5G infrastructure and discusses observed trade-offs between latency and infrastructure related costs indicating optimal operational points for different scenarios.

References

[1] O-RAN ALLIANCE, https://www.o-ran.org/

[2] Yu, H., et.al.,. DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks. *In: proc. of ONDM 2019*. Lecture Notes in Computer Science, vol 11616. Springer, Cham, 2020.

[3] S. Wang et.al., "A Survey on Service Migration in Mobile Edge Computing," *IEEE Access*, vol. 6, pp. 23511-23528, 2018,

[4] T. Taleb, et.al., "Follow-Me Cloud: When Cloud Services Follow Mobile Users," *IEEE Trans. on Cloud Computing*, vol. 7, no. 2, pp. 369-382, 1 April-June 2019,

[5] A. -I. Manolopoulos, M. P. Anastasopoulos, V. -M. Alevizaki and A. Tzanakaki, "Optimal Service Provisioning in Mobile 5G and Beyond Systems," in *IEEE Transactions on Services Computing*, vol. 16, no. 4, pp. 2841-2854, 1 July-Aug. 2023, doi: 10.1109/TSC.2022.3225011

[6] Tzanakaki *et al.*, "Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services," *IEEE Comms. Mag*, vol. 55, no. 10, pp. 184-192,.2017,

[7] A. M. Alba, S. Janardhanan and W. Kellerer, "Enabling Dynamically Centralized RAN Architectures in 5G and Beyond," *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 3509-3526, Sept. 2021,

[8] H. Gupta, M. Sharma, A. Franklin A. and B. R. Tamma, "Apt-RAN: A Flexible Split-Based 5G RAN to Minimize Energy Consumption and Handovers," *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, pp. 473-487, March 2020,

[9] M. Ahsan et.al.,, "Functional Split-Aware Optimal BBU Placement for 5G Cloud-RAN Over WDM Access/Aggregation Network," *in IEEE Systems Journal*, 2022

[10] I. Leyva-Pupo, C. Cervelló-Pastor, A. Llorens-Carrodeguas, "Optimal Placement of User Plane Functions in 5G Networks". *In: Proc. of* WWIC 2019. *LCNS*, vol 11618. Springer, 2019.

[11] I. Leyva-Pupo et al.,. Dynamic Scheduling and Optimal Reconfiguration of UPF Placement in 5G Networks. In *Proc. of MSWiM '20*, 2020 NY, USA, 103–111

[12] Shinde, S. S., Marabissi, D., & Tarchi, D. "A network operator-biased approach for multi-service network function placement in a 5G network slicing architecture". *Computer Networks*, 201, 108598, 2021

[13] T. Taleb, M. Bagaa and A. Ksentini, "User mobility-aware Virtual Network Function placement for Virtual 5G Network Infrastructure," In, Proc. of ICC, 2015, pp. 3879-3884

[14] A. Ceselli, M. Premoli and S. Secci, "Mobile Edge Cloud Network Design Optimization," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1818-1831, June 2017,

[15] K. Cao, L. Li, Y. Cui, T. Wei and S. Hu, "Exploring Placement of Heterogeneous Edge Servers for Response Time Minimization in Mobile Edge-Cloud Computing," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 494-503, Jan. 2021,

[16] F. Zhang, et.al., "Reducing the network overhead of user mobility–induced virtual machine migration in mobile edge computing". *Softw Pract Exper.*; 49: 673– 693. 2019

[17] T. Ouyang, Z. Zhou and X. Chen, "Follow Me at the Edge: Mobility-Aware Dynamic Service Placement for Mobile Edge Computing," IEEE Journal on Selected Areas in Communications, vol. 36, no. 10, pp. 2333-2345, Oct. 2018,

[18] X. Yuan, M. Sun and W. Lou, "A Dynamic Deep-Learning-Based Virtual Edge Node Placement Scheme for Edge Cloud Systems in Mobile Environment," *IEEE Transactions on Cloud Computing*, vol. 10, no. 2, pp. 1317-1328, 1 April-June 2022,

[19] Q. Yuan, J. Li, H. Zhou, T. Lin, G. Luo and X. Shen, "A Joint Service Migration and Mobility Optimization Approach for Vehicular Edge Computing," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9041-9052, Aug. 2020,

[20] H. Ma, Z. Zhou and X. Chen, "Leveraging the Power of Prediction: Predictive Service Placement for Latency-Sensitive Mobile Edge Computing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6454-6468, Oct. 2020,

[21] Y. Ma et.al., "Mobility-Aware and Delay-Sensitive Service Provisioning in Mobile Edge-Cloud Networks," *IEEE Transactions on Mobile Computing*,. 21 (1), pp. 196-210, 2022.

[22] Fan, C., & Li, L. "Service migration in mobile edge computing based on reinforcement learning". *Journal of Physics:* Vol. 1584, No. 1, p. 012058). IOP Publishing, 2020

[23] F. Brandherm, et.al, "A learning-based framework for optimizing service migration in mobile edge clouds" *In Proc. of EdgeSys* 2019..

[24] Zhao, X., Shi, Y., & Chen, S. (2020). MAESP: Mobility aware edge service placement in mobile edge networks. Computer Networks, 182, 107435.

[25] "IEEE Standard for Packet-based Fronthaul Transport Networks," in IEEE Std 1914.1-2019 , pp.1-94, 21 April 2020,

[26] Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, (3GPP TS 23.211 version 13.2.0 Rel 13).

[27] H2020 Project 5G-COMPLETE, Deliverable D2.1

[28] https://gitlab.eurecom.fr/oai/

[29] A. Tzanakaki, M. Anastasopoulos, A. Manolopoulos and D. Simeonidou, "Mobility aware Dynamic Resource management in 5G Systems and Beyond," *In Proc. of ONDM* 2021,

[30] Openstack live block migration. https://docs.openstack.org/nova/pike/admin/configuring-migrations.html

[31] Iperf network benchmarking tool. https://iperf.fr/

[32] Amazon pricing. https://aws.amazon.com/s3/pricing/

Chapter 5 DEMONSTRATION OF MULTI-ACCESS 6G NETWORKS WITH USER MOBILITY CONSIDERATIONS

5.1. Chapter Introduction

5G Networks are designed to operate in highly heterogeneous environments, supporting a wide array of services across different sectors and vertical industries, including eMBB URLLC and mMTC. These networks incorporate traditional cellular technologies such as LTE and 5G New Radio (NR), as well as non-3GPP access technologies like Wi-Fi, satellite, and other wireless networks. Seamless interworking between 5G and these other access technologies helps minimize operational costs and facilitates the widespread adoption of 5G NPNs in vertical domains that already utilize non-3GPP technologies like Wi-Fi. This integration also enables end-devices that lack native 5G capabilities, such as legacy devices and IoT equipment, to access the 5G network and benefit from the enhanced services provided by 5G.

A critical component in the 5G architecture that enables seamless integration of untrusted non-3GPP access technologies is the N3IWF [1]. The N3IWF plays a pivotal role in ensuring that devices can connect to the 5GC network via non-3GPP access networks. It acts as a bridge, providing secure and efficient interworking between non-3GPP access networks and the 5GC. The N3IWF supports key functionalities such as user authentication, data encryption, and secure tunneling, which are essential for maintaining the integrity and security of the network while facilitating seamless mobility and service continuity. In addition to N3IWF, other elements within the 5GC network, such as the AMF and the UPF, also contribute to managing connectivity and mobility across different access technologies. AMF handles connection management and mobility procedures, while UPF manages the user plane traffic, ensuring efficient data routing and low-latency communication.

Integration of non-3GPP technologies with the 5G Core have been tackled recently by a number of researchers. Most of the reported work attempts to verify the feasibility of such deployments with various AN technologies. A relevant tutorial article focusing on the convergence of non-3GPP ANs with the 5GC is presented in [2]. The authors classify the types of non-3GPP access in three categories i.e. trusted, untrusted and wireline. They also discuss several aspects such as authentication and authorization procedures as well as PDU session establishment. Finally, they present an experimental environment that interconnects the 5GC with Wi-Fi along with a basic performance evaluation. A testbed that combines elements from different 5GC open-source projects and integrate these with non-3GPP access technologies such as Wi-Fi is presented in [3]. The authors use their testbed implementation to investigate various security aspects. Authors in [4] demonstrate an implementation that integrates LiFi and 5G which is suitable for Smart Factory scenarios. They also present a multistage demonstrator set-up in order to evaluate the protocol stack enhancements that are

needed to support handover between LiFi and 5G. Finally, the authors in [5] propose the integration of a Very Short Aperture Link (VSAT) Satellite link with 5GC. They discuss various parameters that are needed to enable this integration such as the protocol stack and signaling procedures. Moreover, they present an implementation that interconnects the N3IWF with a satellite emulator.

In mobile environments where users exhibit varying mobility patterns and service requirements, efficient network selection and handover management are crucial to minimizing service interruption and resource consumption. Factors influencing RAT selection include service requirements such as bandwidth and latency, real-time network conditions like signal strength and congestion levels, cost considerations, and the need for seamless mobility management between RATs.

One of the relevant challenges that need to be addressed involves managing handovers between cells. As users move through the network, frequent handovers can lead to interruptions and degraded service quality. This challenge is magnified when considering handovers between 3GPP and non-3GPP networks. Effective cell management strategies, such as dynamic handover algorithms and intelligent traffic steering, are essential to minimize these disruptions and maintain seamless connectivity.

Various works exist in the literature address optimal multi-RAT selection in 5G/B5G systems. [6], focuses on the logical evolution of 5G networks. The authors propose a novel scheme that provides to the user the ability to select between various access technologies. Moreover, they discuss the integration of analytics as a means to enhance end-to-end performance. The authors in [7] discuss multi-connectivity features and propose a flexible architecture that offers connectivity through various technologies such as 5G, LTE and WiFi as a way to optimize throughput, mobility and resource management aspects. A solution that offers two connectivity paths, one through LiFi and one through standard 5G components is presented in [8], as well as a handover mechanism between licensed and unlicensed access technologies. The proposed solution is suitable for industrial and private applications. Finally, in [9] a model is proposed based on queuing network theory that aims at optimizing the design and planning of multiservice RANs with diverse network requirements in dense areas.

The present chapter [10] investigates access network selection strategies aiming at reducing handover interruptions by leveraging experimental data from an opensource 5G platform. Our results have been produced using real user mobility statistics as well as lab-based profiling measurements of a 5G infrastructure that supports non-3GPP access. By analyzing resource consumption for both non-3GPP and 5G networks and applying queuing theory, we can derive insights regarding end-to-end performance and optimize handover processes to ensure seamless user experiences.

The remainder of this chapter is structured as follows: in *Section 5.2* we describe the system architecture. *Section 5.3* presents the analytical framework based on queuing theory for E2E performance analysis. *Section 5.4* discusses the implementation and evaluates the proposed model. Finally, *Section 5.5* concludes the chapter.

5.2. Integration of 5G Systems and Non-3GPP Access Networks: An Overview

5G systems introduce several architectural advancements compared to previous generations (3G/4G), incorporating numerous new concepts and technologies to meet the extensive variety of applications they are designed to support. As illustrated in Figure 5. 1, the overall 5G network architecture adopts microservices and NF decomposition, allowing flexible scaling and updates without disrupting existing services. This approach ensures adaptability and efficiency across the network. To access services provided by a MNO, the UE can connect either through the 5G RAN, or through a non-3GPP AP with the help of N3IWF. In the first case, the UE connects over the air interface to the gNB, and initiates NAS signaling processing at the AMF and PDU session establishment. NFs in the SBA communicate with each other over the SBI using the Hypertext Transfer Protocol (HTTP) and secure connections via TLS, or through reference points using specific transport and application layer protocols.

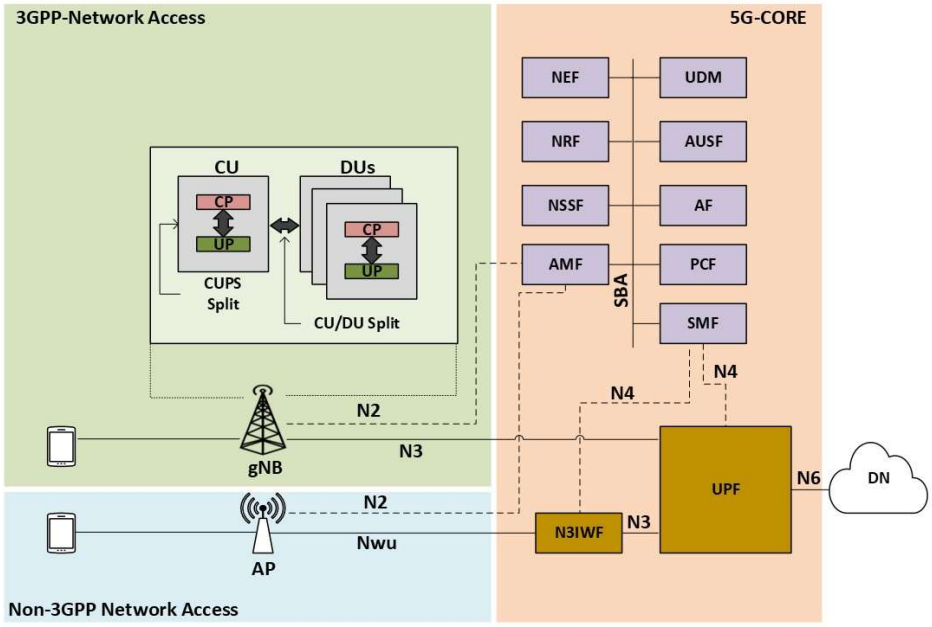


Figure 5. 1: 5G Network Topology

To support the connectivity of UE via a non-3GPP access network, 5G systems specify three types of access: trusted, wireline, and untrusted (see 2.2.3). Trusted access implies a high trust level where the operator has control over the non-3GPP network. Wireline access supports legacy devices and FWA, providing flexibility in connecting residential gateways to the 5GCN. Finally, in the untrusted access, the MNO does not trust the security of the non-3GPP network. In this case, there is a need for secure transport methods like the N3IWF. N3IWF connects to the AMF via the N2 interface CP communication and to the UPF via the N3 interface for UP data traffic. After the UE undergoes authentication and authorization, it can access the 5GCN through the untrusted non-3GPP network, performing NAS signaling via the N1 interface. Data packets between the UE and the DN are transferred securely using an IPSec tunnel between the UE and N3IWF. Additionally, GTP-U establishes a tunnel between N3IWF and UPF for user data.

Accessing 5GCN from untrusted networks involves network access discovery and selection, registration, authentication, authorization, and PDU session establishment. To register to the 5GCN, the UE first needs to select and connect to a WLAN using the WLAN protocol. Once the UE is configured with a local IP address from the selected WLAN, the UE selects the N3IWF and initiates the IPSec Security Association (SA) establishment procedure over the NWu interface using the IKEv2 protocol. After the IKEv2 SA establishment, the N3IWF starts the EAP-5G procedure with the UE which initiates the registration and authentication procedure using the NAS protocol with the AMF over N1 interface. The NAS messages are transported using the EAP-5G/IKEv2 between the N3IWF and UE over NWu interface and using the NGAP/SCTP between the N3IWF and AMF over the N2 interface. Successful authentication involves AMF, AUSF, and UDM, resulting in a shared security key for NAS and N3IWF [11].

Finally, the PDU session establishment allows both UE and the network to initiate sessions for new connections or handovers. The UE-requested PDU session process involves sending a request to the AMF via IPSec SA, which selects the SMF to create the PDU session context, authorize it, and establish IPSec child SAs for QoS profiles. Once the session is accepted, the N3IWF facilitates the secure data communication between the UE and the data network, encapsulating PDU sessions inside GRE tunnels [2][11].

5.2.1. Radio Access Network Selection

We consider a multi-technology radio access network comprising 3GPP and non-3GPP technologies as shown in Figure 5. 2 a). For this network, end-to-end slices should be established interconnecting UEs with the DNs. In case where UEs are attached to the 3GPP access network, the corresponding slices will be established over the path 2-4-5 formed by interconnecting the gNB, UPF and DN nodes. However, in case where UEs select a non-3GPP network, the relevant slices will be established over WiFi-N3IWF-UPF-DN (1-3-4-5) path. This network can be modeled as a mixed network of queues where new arriving UEs can be served either through the gNBs or the WiFi APs. In case of mobility, handovers can be managed either maintaining the same access technology (i.e., handover from 3GPP to 3GPP) or forwarding connections to a different access network (i.e., handover from 3GPP to Non-3GPP and reverse).

In order to ensure seamless service provisioning, redundant physical resources should be reserved. The amount of redundant resources increases with the speed of end-user mobility, the size of the wireless cells (mobile users associated with small cells will exhibit very frequent handovers) and the traffic model adopted. Based on their technical characteristics, the wireless access technologies adopted in this work, can address end-user mobility with different levels of effectiveness, e.g. WiFi with smaller coverage can support users with relatively low mobility whereas gNBs with broader coverage can support higher mobility levels as less frequent handovers are required.

To maximize the benefits provided by the available technologies, users with low mobility and arrival rate λ_s can exclusively use the non-3GPP access. In addition to this, slow moving UEs can also use the 3GPP network which is shared with fast moving users that arrive in the system with rate λ_F. Fast users can access only 3GPP network resources. The overall network selection scheme is illustrated in the simple two-dimensional Markov Chain (MC) shown in Figure 5. 2 b). The state space of the MC model is defined as $\mathcal{S} = \{(i,\mathrm{j}) | i \leq \mathcal{I}, j \leq \mathcal{J}\}$, where i, j correspond to the number of Non-3GPP and 3GPP users, respectively, and (i, j) is a feasible state in $\mathcal{S}$. In this scheme, we observe that if the total number of slow and fast mobile moving UEs in the network is less than a

threshold (i.e., $i + j < C_{th} \downarrow$), then access to 3GPP resources is granted to both type of users. However, above this threshold, access to 3GPP resources is permitted only to fast moving users. The main rational behind this approach is that by reserving resources in the 3GPP access for fast mobile UEs, handover frequencies are reduced. It is clear that by reducing this threshold more resources in the 3GPP access can be allocated to serve fast moving UEs reducing the probability their QoS requirements not to be met and dropped. However, this comes at the cost of increased blocking for slow moving users. To address this problem, an optimal point that minimizes the weighted sum of normalized blocking and dropping ratio for slow and fast moving UEs, respectively, can be identified.

In the following subsection, a description of the testbed used to benchmark the proposed framework is proposed.

5.3. Implementation and Evaluation

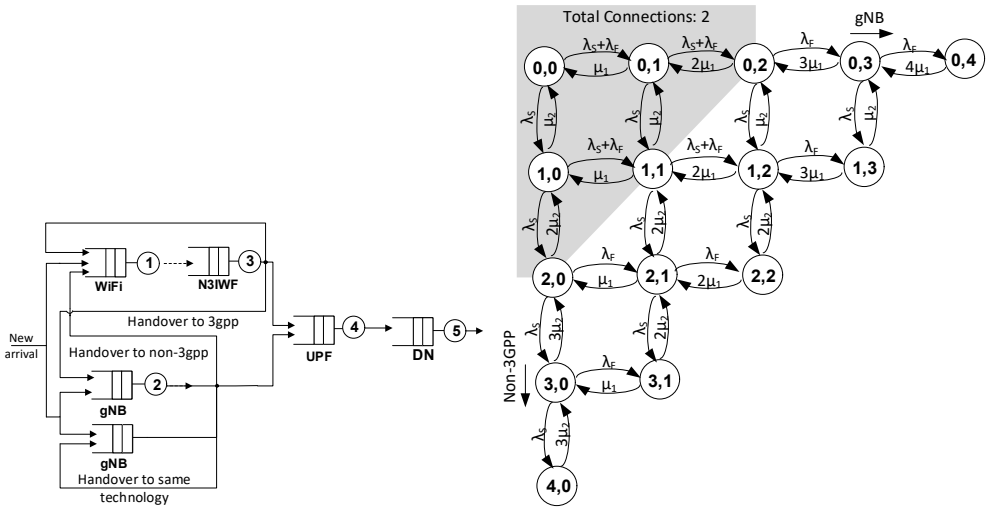


Figure 5. 2: multi-technology access modelled as an open network of queues, b) Two-dimensional Markov Chain illustrating the Access network selection process for fast- and slow-moving mobile devices. The grey array illustrates the network states where 3GPP resource

In this section we provide an overview of the implementation that was carried out in this work. The main idea was to design and deploy a 5G platform that offers multi-access UE connectivity in an environment where the 5G components are implemented as VNFs that run over a cloud infrastructure. At the same time, by extensively profiling each of the functionalities of the 5G system, we can determine critical parameters such as the downtime between switching access technologies, as well as measure the performance and consumption in terms of network, compute and memory. These parameters can then be mapped to the AN selection model in order to provide realistic values to its variables.

The lab testbed that was used for the implementation of this work, consists of the following:

A Private Cloud Platform: The cloud platform provides the underlying environment on top of which the 5G VNFs are deployed. It based on openstack and all it functionalities

run as microservices on LXD containers that are hosted the bare metal servers (Linux OS).

A Monitoring/Visualization Platform: The monitoring platform extracts resource consumption metrics from all physical and virtual components and stores them in a Prometheus database. From there, the data can either retrieved or visualized (Grafana).

5.3.1. Deployment Overview

The network topology is illustrated in Figure 5. 1 where five instances are used to host the VNFs and their interfaces. The deployment leverages the CP/UP split paradigm which enables flexibility and isolation of the NFs, while at the same time makes it possible to place the UPF node closer to the users in order to handle delay-constrained scenarios. Moreover, the network setup is in accordance with the SBA. All external interfaces are implemented in dedicated network subnets, enhancing security as well as ease of monitoring.

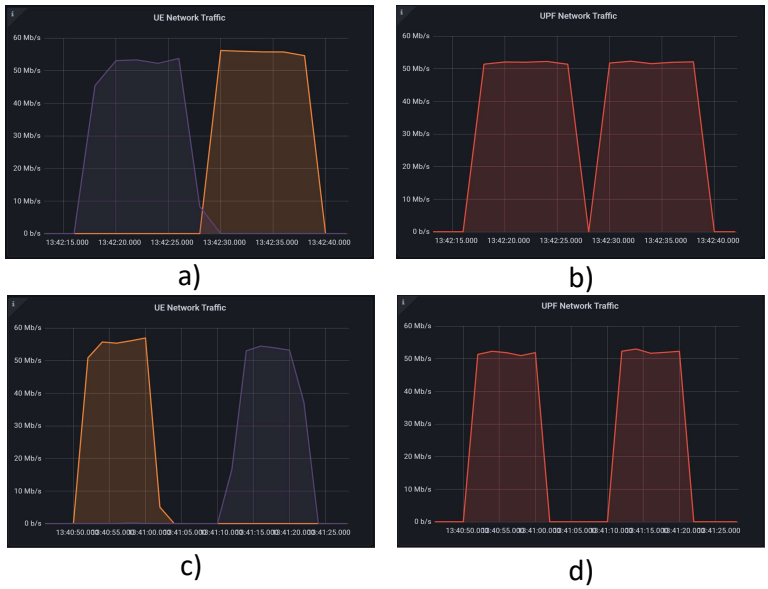


Figure 5. 3: Network Traffic for a UE that switches access network: a) from non-3gpp to 3gpp, b) UPF total, c) from 3gpp to non-3gpp, d) UPF total

More specifically, the deployment includes:

A 5GC CP VNF. For the 5G CN, we used free5gc [12] which is an open-source for the core network, written in go. Most of the CP NFs use internal SBIs except from the SMF which uses the N4 subnet for the PFCP, and the AMF which uses the N2 subnet for the NGAP.

A UPF VNF. As mentioned above, the UPF node is deployed in a separate instance following the CUPS architecture. Three interfaces are attached to the node, a N3 interface for the GTP-U, a N4 for the PFCP and a N6 interface to provide connectivity with the Data Network.

A N3IWF VNF. The N3IWF acts as gateway with the untrusted access network. For this reason, it is also implemented in a dedicated instance. It connects with the AMF via the N2 interface and with the UPF with N3. Additionally, it connected with an IPSec tunnel to the NWu node which is realized by the xfrm module.

A NWu interface. The NWu is deployed in a separate machine. Typically, this machine is interconnected with the Access Point (e.g. Wifi), where the traffic is bound with the IPSec tunnel in order to reach the 5G CN.

Emulated RAN and UE. The UE and RAN part are based on UERANSIM [13]. The node is connected with the N1, N2 and N3 subnets.

5.3.2. Evaluation

After successful deployment of all instances we generate some network traffic from the UEs connected to each access network. The network traffic is generated through iperf connections. The connections are interrupted as a UE switches between AN technologies and then get reconnected. The resulting graphs are shown in Figure 5. 3 where network traffic is captured at different parts of the network. In Figure 5. 3 a) and c) the purple color represents the traffic generated from the RAN-connected UE and the orange color indicates the non-3GPP UE. Figure 5. 3 b) and d) depicts the network traffic captures at the UPF node (red color). In a) the UE is first connected to the 3GPP AN and after a while it connects to the 5G-RAN. The process is repeated in reverse in Figure 5. 3 c). It is worth mentioning that since the AN segments of the network as well as the UEs are emulator-based, no UE mobility aspects are considered, the deployment

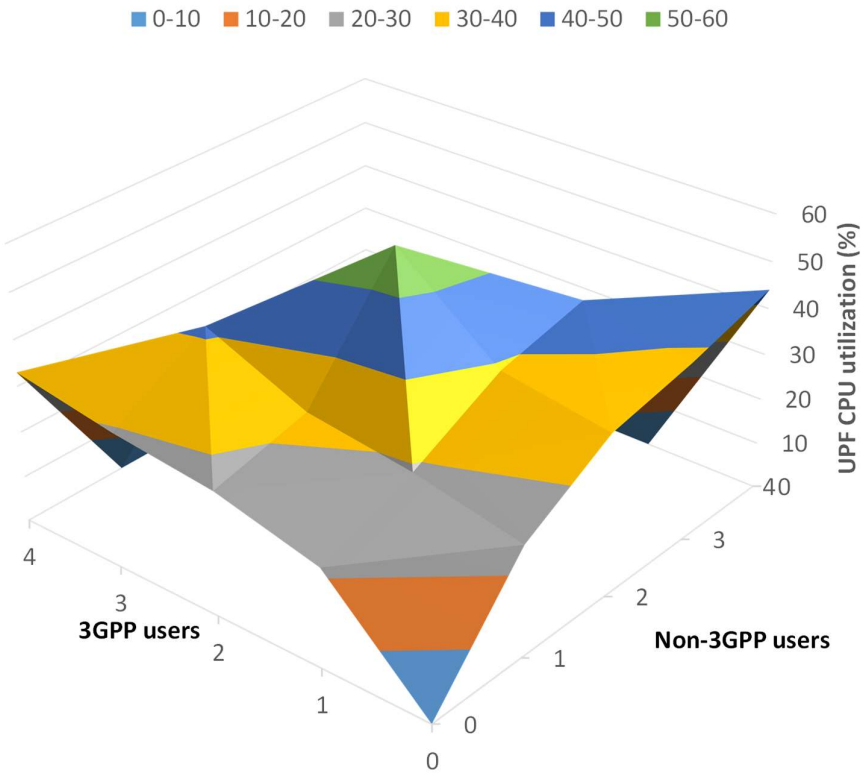


Figure 5. 4: CPU utilization for the physical machine hosting the UPF under various combinations of incoming 3GPP and Non-3GPP users

focuses on the time needed to switch between ANs. Additionally, it can be observed that reconnection from RAN to non-3GPP AN technology is more time consuming, due to security mechanisms implemented for the untrusted networks (GRE, IPSec, IKE etc.). Finally, some minor differences observed between the traffic in the UE and the UPF are attributed in the existence of background traffic.

5.4. Optimal threshold identification for the radio access network selection process

To identify the optimal threshold for the radio network access selection process described in the previous section we initially benchmark our network in order to evaluate for different arrival rates the corresponding service rates of the system. Based on the system's current configuration, the main bottleneck is associated with the VM hosting the UPF nodes used to serve both 3GPP and Non-3GPP connections. An example of the CPU utilization for the VM hosting the UPF for different combinations of incoming connections is shown in Figure 5. 4. UEs connected to the system request the establishment of a service slice with 50Mbps throughput. When only 3GPP connections are established, the utilization of the CPU reaches 33% for 4 UEs. In case where the same number of UEs are connected through the Non-3GPP access, the CPU utilization increases to 44%. The latter is explained by the existence of multiple traffic flows and the increased complexity of the implementation compared to standard single-RAN deployments. A mixture of 3GPP and Non-3GPP traffic introduces further overhead to the system leading to a CPU utilization exceeding 56% for the same number of UEs (2 UEs connected through 3GPP and 2 through non-3GPP).

Once the service rates for the corresponding system have been determined, in the next step the optimal radio access selection threshold is determined. Figure 5. 5 illustrates the normalized dropping and blocking rates for the fast and slow moving UEs, respectively. Numerical results have been extracted for a system that can host up to 10 UEs. As the capacity threshold increases, the gNB resources that can be exclusively used by the fast moving UEs are reduced leading to an increase of the number connections whose QoS

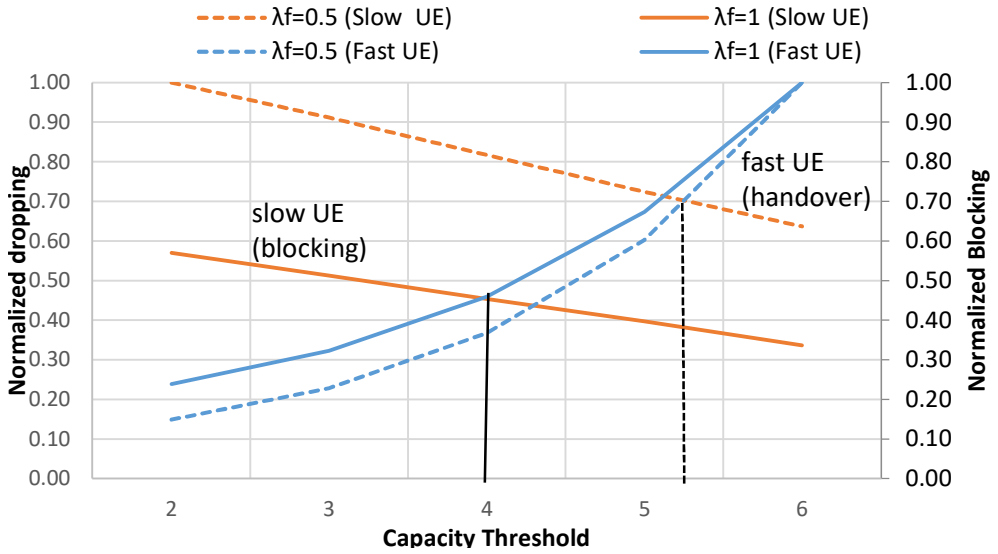


Figure 5. 5: Optimal capacity threshold for different arrival rates

requirements are not met and, consequently, are dropped. However, this policy benefits the slow moving UEs leading to a reduction of their corresponding blocking ratio. Finally, we observe that an increase of the arrival rate of the of the fast moving UEs results to a decrease of the capacity threshold used by the proposed mechanism.

5.5. Conclusion

The integration of 3GPP and non-3GPP technologies in a multi-technology radio access network is crucial for addressing the diverse demands of modern communication. Utilizing the N3IWF enables seamless connectivity and secure access to the 5G Core. This chapter explores such a network, focusing on minimizing handover interruptions through experimental data derived from an open-source 5G platform deployed in our private lab. Our findings, based on real user mobility statistics and lab profiling measurements, highlight resource consumption across both non-3GPP and 5G networks. By applying queuing theory, we derive valuable insights into E2E performance and optimize handover processes to enhance user experience.

References

[1] "3GPP TS 23.502 version 15.3.0 Release 15, '5G;Procedures for the 5G System', " [Online].

[2] M. T. Lemes, A. M. Alberti, C. B. Both, A. C. De Oliveira Júnior and K. V. Cardoso, "A Tutorial on Trusted and Untrusted Non-3GPP Accesses in 5G Systems—First Steps Toward a Unified Communications Infrastructure," in *IEEE Access*, vol. 10, pp. 116662-116685, 2022, doi: 10.1109/ACCESS.2022.3219829.

[3] Matan Broner, Sangwoo Lee, Liuyi Jin, and Radu Stoleru. 2023. Poster: Towards Multi-Radio Access in 5G Networks. In Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23). Association for Computing Machinery, New York, NY, USA, 575–576. https://doi.org/10.1145/3581791.3597373

[4] M. Müller *et al.*, "LiFi with 5G for the Smart Factory," *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, Austin, TX, USA, 2022, pp. 2310-2315, doi: 10.1109/WCNC51071.2022.9771969.

[5] M. Luglio, M. Quadrini, C. Roseti, D. Verde and F. Zampognaro, "Performance evaluation of untrusted non-3GPP Access to a 5G Core Network via satellite," *2022 International Symposium on Networks, Computers and Communications (ISNCC)*, Shenzhen, China, 2022, pp. 1-6, doi: 10.1109/ISNCC55209.2022.9851800.

[6] V. Agarwal, C. Sharma, R. Shetty, A. Jangam and R. Asati, "A Journey Towards a Converged 5G Architecture & Beyond," *2021 IEEE 4th 5G World Forum (5GWF)*, Montreal, QC, Canada, 2021, pp. 18-23, doi: 10.1109/5GWF52925.2021.00011.

[7] S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard and D. Chandramouli, "5G multi-RAT multi-connectivity architecture," *2016 IEEE International Conference on Communications Workshops (ICC)*, Kuala Lumpur, Malaysia, 2016, pp. 180-186, doi: 10.1109/ICCW.2016.7503785.

[8] T. Metin, M. Emmelmann, M. Corici, V. Jungnickel, C. Kottke and M. Müller, "Integration of Optical Wireless Communication with 5G Systems," *2020 IEEE Globecom Workshops (GC Wkshps*, Taipei, Taiwan, 2020, pp. 1-6, doi: 10.1109/GCWkshps50303.2020.9367502.

[9] Marin, A., Meo, M., Sereno, M., & Marsan, M. A. (2024). Queuing Network Models of Multiservice RANs. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, *9*(2), 1-26.

[10] A. -I. Manolopoulos, V. M. Alevizaki, M. Anastasopoulos, and A. Tzanakaki, "Demonstration of Multi-Access 6G Networks with User Mobility Considerations," in IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, October 2024.

[11] "Interworking of Untrusted Non-3GPP Networks and the 5G-Core Network." Wipro. Accessed July 15, 2024. https://www.wipro.com/network-edge-providers/untrusted-non-3gpp-access-network-interworking-with-5g-core/.

[12] Free5gc, [online]. Available: https://free5gc.org/

[13] UERANSIM, [online]. Available: https://github.com/aligungr/UERANSIM

Chapter 6 AN AI-ASSISTED FRAMEWORK FOR LIFECYCLE MANAGEMENT OF BEYOND 5G SERVICES

6.1. Chapter Introduction

The introduction of 5G technology that promises faster data rates, ultra-low latency, massive machine-type communications and increased network reliability has transformed mobile networks over the past few years. A key enabler behind these advancements is the exploitation of cloud computing to support the operation of 5G networks, introducing the notion of 5G Cloud networks [1]. 5G Cloud networks are able to offer innovative services and applications with improved performance. This is achieved by taking advantage of advanced networking technologies with increased scalability and flexibility features that cloud infrastructures inherently offer [2]. In addition, 5G technologies introduce the feature of network slicing that allows partitioning of the physical network infrastructure into multiple virtual/logical network slices [3]. Each slice can operate independently and can be configured to meet specific service requirements. These requirements can be mapped to various QoS classes corresponding to different levels of bandwidth, latency, traffic priority, security, reliability etc.

To achieve this, 5G networks adopt novel architectural concepts such as microservices, network function decomposition, as well as CUPS. The adoption of these concepts and approaches enables the 5G infrastructure to become flexible and adaptable increasing the efficiency with which resources are being utilized. In the RAN domain, network functions are separated based on their roles (control or user plane) and can be placed at different locations according to their resource requirements and delay constraints. The CN also adopts CUPS, leveraging virtualization and softwarization. The overall CN architecture follows the paradigm of the SBA involving a set of key VNFs. [4]

While 5G delivers impressive advancements in data rates, latency and connectivity, it still faces the challenge of explosive growth in data-intensive and latency-sensitive applications. To address this, B5G is emerging, that brings advancements beyond the current 5G standards, setting the path towards 6G [5]. B5G is build on 5G foundations and introduces key improvements, particularly in integrating AI and ML for intelligent network management and predictive maintenance, aiming to support a set of diverse services. These services are related to various use cases, sectors and vertical industries [6] spanning from Unmanned autonomous Vehicles and automated production lines to entertainment, Extended Reality (XR) and the IoT. Therefore, these networks are expected to affect and in some cases even reshape various aspects of every day's life as well as the means of interaction between humans and technology [7].

The heterogeneity and dynamicity of these complex environments poses new challenges in terms of management and performance optimization for these advanced systems. In

this context, AI and ML can play a key role [8]. More specifically AI/ML tools can enable intelligent automation, proactive network management and optimization in resource allocation. By analyzing the massive volumes of data generated by 5G networks, AI/ML algorithms can identify patterns, predict network behavior, detect anomalies and security threats as well as make informed real-time decisions, leading to further optimization of network performance, QoS enhancement, delivery of undisrupted services etc. Furthermore, network operators can exploit AI/ML tools to efficiently facilitate network planning and management of 5G infrastructures in a cost-effective and efficient manner[9].

It is therefore clear that B5G networks are transforming into open, flexible and efficiently shared infrastructures. In this environment, traditional NFs, softwarized according to the NFV [10] paradigm, are managed through a centralized MANO Platform [11]. MANO introduces robust and centralized management of service creation, offering network-wide service design, configuration, deployment and monitoring. Additionally, it enables automation in the arrangement and coordination of network elements as well as scaling of resources and services. This network-wide orchestration introduces the benefit of a single integration point and a centralized representation of distributed network resources, regardless of the volume of resources involved or the location of these resources. By automating, through the orchestrator, the infrastructure configuration and monitoring processes, it is possible to reduce the inherent complexity of delivering and administering sophisticated and multi-featured services.

Solutions such as OSM [12] and Open Network Automation Platform (ONAP) [13] have emerged to handle the lifecycle of the NFs, according to the standards set by the European ETSI and the Open Network Foundation (ONF) for NFV. Both platforms aim to provide a versatile approach enabling the onboarding of any application. In their typical design approach, each vendor's application is paired with its own *Operator*- a software component that encloses the application along with comprehensive instructions for tasks such as deployment, configuration, scaling, and integration on the cloud. However, this often leads to discrepancies and inaccuracies in the terminology and taxonomy associated with cloud-native principles and MANO. Furthermore, the designs and implementations of 5G elements from different vendors often exhibit static behavior, even for basic tasks such as IP address assignment and resolution. This reliance on static setups necessitates human intervention that introduces scalability challenges. Such malpractices are often justified as minor engineering issues, but they are in fact a reflection of carrying over a common practice from either legacy Physical NFs (PNFs) or outdated design architectures that do not take into consideration the nature of cloud and cloud native deployments. More specifically, in case every vendor creates a unique Operator, that is specific to the vendor applications, inconsistencies may arise, introducing increased complexity. To address this challenge, Operators are assigned to higher-level, standardized network components—such as network terminals, functions, or slices. This enables Operators to manage logical network elements in a standardized way, rather than being tied to specific applications from different vendors. This approach allows to mix and match NFs, addressing previous challenges arising from individual vendors being locked with their own MANO and OAM solution.

Going one step further, the vision for the B5G era is to further extend the levels of automation and minimize human intervention in network and service management. ZSM is introduced with the aim to enable a network that is self-configured, self-

monitored, self-healing and self-optimized [14]. To achieve this, ZSM strongly relies on tools and attributes supported by AI/ML schemes and MANO.

Contributions: Aligned with the B5G vision for zero touch system operation, this work [15] leverages some of the key technologies and concepts of 5G networks to develop a framework that automates provisioning, deployment, management and orchestration of network slices and services. This framework can be used to easily deploy, manage, modify, and delete 5G services and functions, while also autonomously performing re-configuration actions, without human intervention. The chapter also presents the complete LCM of CN components that allows 5G systems to operate in an intelligent, adaptable, and flexible multi-slice environment.

To achieve this functionality, a set of building blocks have been implemented and are able to:

- host and instantiate various 5G deployment options involving multiple network slices and sophisticated topologies.
- handle LCM operations and manage network slices, leveraging domain NFV Orchestrators (NFVOs) and controllers.
- monitor the entire 5G system collecting a variety of performance measurements for the virtualized network functions and the underlying physical entities.
- take intelligent and fully optimized decisions leveraging AI/ML algorithms. These algorithms are trained based on data collected from the monitoring system recommending optimal LCM actions.

To test the efficiency of the deployed system, a two-stage evaluation process has been adopted. The first stage tests the ability of the platform to perform a set of orchestrator-based multi-slice network deployments. These tests evaluate the efficiency of the platform to instantiate appropriate 5G topologies (deployment options) across a diverse set of operational scenarios. Once the 5G network has been deployed, the second stage tests are used to evaluate the ability of the system to perform optimal reconfigurations with zero human interventions optimizing UP traffic forwarding policies.

The rest of the chapter is structured as follows: Section 6.2 provides a brief overview and a literature review of the basic components in 5G. This includes 5G architectural aspects, mechanisms for network management with emphasis on slicing and, finally, AI-based tools and algorithms for optimal decision making. The progress and main innovations of the proposed work compared to the state of the art is also highlighted. Details of the proposed Overall System Design are presented in Section 6.3, while demonstration and experimentation results are provided in Section 6.4. Finally, Section 6.5 summarizes the conclusions of the chapter and proposes directions for future work.

6.2. Background and Related Work

6.2.1. 5G System Architecture

The 5G ecosystem brings together a set of heterogeneous applications and use cases such as eMBB, mMTC and URLLC [16][17][18]. These use cases and the related applications have extremely diverse and stringent requirements in terms of network and compute resources, QoS aspects etc. Consequently, 5G Systems introduce several

architectural advancements compared to previous generations of mobile communication systems (3G/4G) and involve the adoption of new concepts and technologies in order to meet the requirements of the applications they are expected to support.

The overall 5G network architecture is shown in Figure 6. 1. 5G networks adopt the microservices architecture [19] and network function decomposition [20], to provide a network that allows for each service to scale or update without disrupting other services in the network. These concepts have been applied in the design of RAN and CN segments.

In the RAN, the concepts of *vertical* and *horizontal* functional splits [21] is adopted in the design of both control and data planes. The CUPS in 5G is known as "vertical split" while "horizontal split" refers to the decomposition of the baseband function stack into a set of individual independent functions that can be allocated to different computational resources. The most flexible RAN solution supports splitting of the

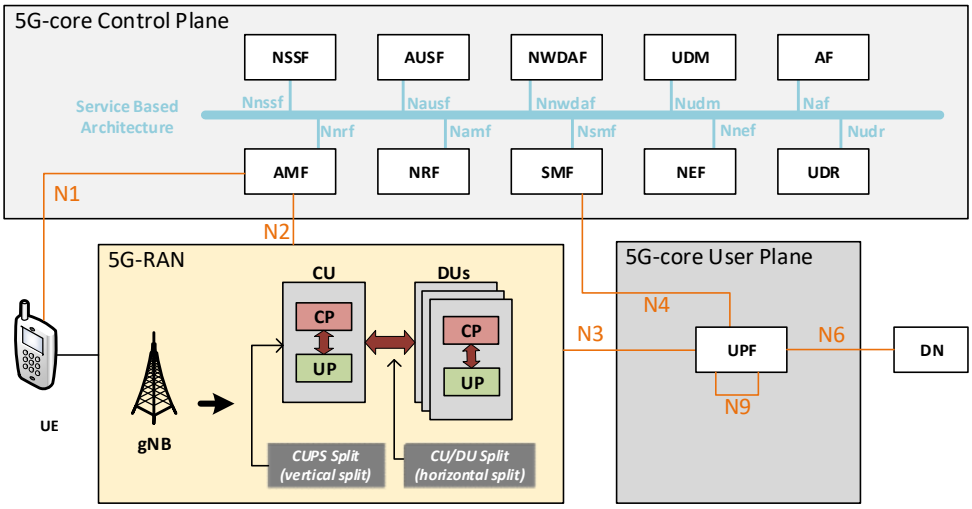


Figure 6. 1: 5G Network Architecture

baseband function stack into a set of functions that can be independently allocated to the RU, the DU and the CU according to the service requirements, aiming to maximize resource and energy efficiency as well as service performance. This architecture enables a more flexible mapping of NFs to physical network entities, depending on the use case and deployment constraints [22].

In the 5G core segment, the concept of CUPS is adopted by separating the control and user plane NFs [4]. Furthermore, it relies on virtualization and softwarization which decouples the various control functionalities from the underlying hardware/infrastructure. This way, the 5G Core can benefit from the advantages of Cloudified and cloud-native deployments. The CP of core network comprises multiple VNFs that interact through service-based interfaces (SBI) [23] and is accountable for decision-making and network management. These VNFs include the AMF to facilitate user registration and the SMF that handles user connections. The UP of the CN includes the UPF, which manages the data path and traffic policies. More specifically the UPF performs packet inspection and routing, as well as UP QoS handling. The communication of the UP elements with the CP elements is achieved through point-to–

point interfaces [24] where each interface serves specific purposes, such as carrying user plane traffic (N3, N9, N6), managing mobility and session establishment (N1, N2), or managing the UPF nodes (N4) [25].

6.2.2. 5G MANO and Network Slicing

A core prerequisite for 5G systems is the support of flexible and configurable network architectures, so that they can adapt to any use case and service requirements. To achieve this, 5G systems embrace technologies such as NFV (already briefly discussed) and SDN to enable dynamic deployment of network functionalities, replacing the need for manual, node-by-node configuration. This centralized approach is related to the migration of individual device configuration in favor of a more robust management mechanism that can offer network-wide service design, configuration, deployment, and monitoring. Such processes require implicit autonomic control over all systems, resources, and services as well as inherent intelligence.

In this context, NFV enables deployment of network functions on VMs or containers hosted on general-purpose servers, rather than relying on vendor-specific hardware. This approach allows the system to adapt dynamically to varying network requirements and optimize resource allocation based on actual end-users current and future demands. A high-level view of the NFV architecture is shown in Figure 6. 2 comprising several key elements [26][22], [27], [28] including: a) VNFs, which are software-based instances that perform specific network functions; b) the NFVI, that provides the necessary virtualized resources, such as computing, storage, and network; and c) the NFV MANO framework, that oversees the lifecycle management of VNFs and coordinates the efficient use of resources.

The MANO framework integrates various managers, such as the VIM, the VNF Manager, and the NFV Orchestrator, to ensure automated deployment and orchestration of network services. MANO has been adopted by several research studies and projects to automate network configurations in 5G networks. For example, in [29] the authors adopt MANO to trigger allocation of a set of resources and service scaling policies to meet specific Service Level Agreement (SLA) requirements. Three different scaling types are addressed, namely, application-, resource- and scaling level. Their solution was implemented in a proof-of-concept virtualized platform using in the wireless access part an IEEE802.11p network verifying the ability of MANO to automate network deployment and update service instances. Unlike [29] that relies on a monolithic IEEE802.11p RAN solution, in our work MANO is used to automate deployments in a highly complex disaggregated 3GPP Rel.15 compliant 5G network for

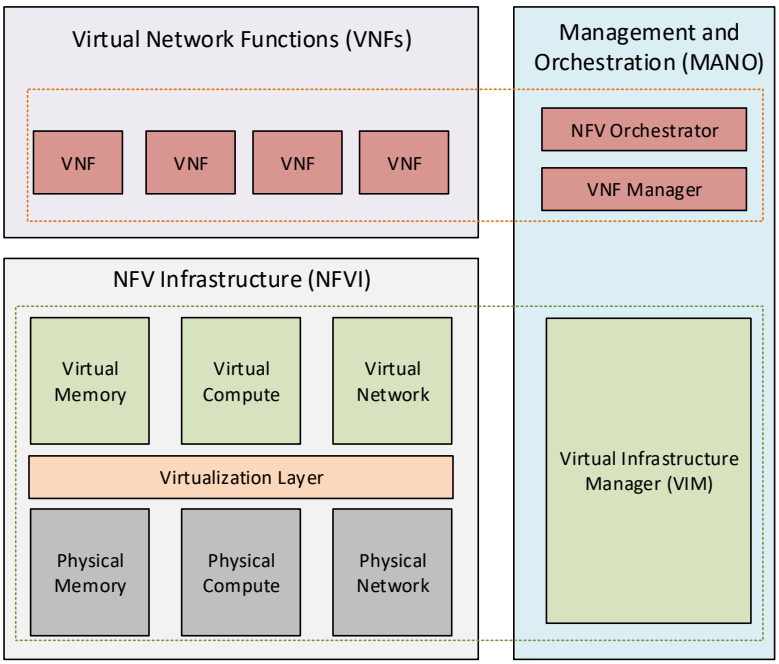


Figure 6. 2: NFV Architectural Framework.

the RAN and core segments. Similar studies have been also performed by the authors in [30], [31] where MANO is used for the management of 4G and WiFi network testbeds. The MANO concept has been also adopted by the authors in [32] to automate the network service lifecycle management with the use of knowledge management and decision support techniques Their studies have been carried out over a simulated (ns-3 based) 5G network showing the ability of MANO to provide flexibility in service deployment and decommissioning.

Network slicing is also a crucial concept in 5G networks, enabling the partitioning of a single physical network into multiple, isolated virtual networks or "slices"[33]. Each slice is customized to meet specific service requirements, such as enhanced connectivity, reliability, performance, and scalability that allow the support of differentiated services over the same infrastructure. Network slicing supports multitenancy, where each tenant can operate independently over customly configured virtualized network instances that are either fully isolated ("hard") or share certain resources ("soft"). The specific characteristics of a network slice are defined by attributes such as the SST and the SD. The SST is a high-level identifier that represents a specific type of service that the network slice is intended to provide (eMBB, URLLC or mMTC). The SD, on the other hand, is a detailed specification that defines the functional and non-functional requirements of the service, including network functions, network topology, performance metrics, and security policies.

To implement network slicing effectively, the Next Generation Mobile Network (NGMN) structures the 5G slice architecture into three distinct layers [33]. The 5G SIL provides the service instances that need support, while the 5G NSI Layer defines the network requirements and configurations for these services, potentially incorporating shared or dedicated NSSIs. Finally, the 5G RL manages the allocation of both physical and logical resources to each slice. 3GPP has established a comprehensive framework for managing network slices, encompassing a three-step lifecycle: instantiation, configuration, and activation; run-time operations; and decommissioning. During the

first phase, all necessary resources for the network slice are provisioned and configured to become operational. In the run-time phase, the slice is actively managed, monitored, and adjusted as needed to maintain performance. Finally, in the decommissioning phase, the slice is deactivated, and its resources are released, ensuring efficient resource utilization across the network [34]. MANO and Network Slicing integration has been considered by several studies to optimize performance and service delivery. The authors in [35] present a highly automated management framework for E2E network slices, designed for multi-tenant 5G networks. The proposed framework is designed to define slices from business models for network slice providers. Their solution is prototyped and experimentally validated in a large-scale 5G Non-StandAlone (NSA) infrastructure. In [36] the design of a 5G network with configured slices that offer low-latency services is being presented. The implementation includes a cloud computing platform, MANO and a 5G NSA platform, based on open source tools. The implementation includes a cloud computing platform, MANO and a 5G NSA platform, based on open source tools. Our study differs from [35] and [36] in that we use MANO for LCM of a Stand-Alone 5G Core implementation combined with predictive analytics. Finally, in [37] MANO is used to assist network slicing operations in the RAN domain without involving network reconfiguration and automation aspects

6.2.3. Artificial Intelligence/Machine Learning (AI/ML)

Over the past few years AI/ML techniques are increasingly being adopted by the telecommunications industry, in an effort to advance network optimization, security aspects, QoS etc. Specifically, in the context of B5G/6G networks that are expected to generate and handle enormous amounts of data, the adoption of AI/ML techniques aims to transform this data into knowledge. This knowledge can prove to be extremely beneficial in automation as well as service lifecycle management and reshape the relevant business models and opportunities. For example, time series predictive models such as network traffic forecasting is a field where the adoption of AI/ML techniques can provide significant improvements. NNs have become a powerful tool in time series data prediction, due to their ability to model complex, non-linear relationships within data. By capturing temporal dependencies and patterns, neural networks, such as Recurrent Neural Networks (RNNs) and their variants like LSTM networks [38], succeed in accurately predicting future values based on past observations. Several types of NNs and algorithms are well-suited for time series data forecasting apart from RNNs, such as Convolutional Neural Networks (CNNs) [39], attention mechanisms [40] and Hybrid models [41]. These networks can learn intricate sequences and trends over time, that makes them highly effective for forecasting applications. By continuously learning from new data, neural networks increase forecasting accuracy and provide robust and adaptive models that outperform traditional statistical methods in many scenarios.

AI/ML functionalities have also been introduced in different domains of 5G Systems with their relevant interfaces and functions [1]:

In *MANO*, AI/ML can enable optimization of network resource allocation, network performance, efficient analysis of failures and design of e2e network slices. Additionally, it can support root cause analysis and alarm correlation. An orchestration framework for the lifecycle management and orchestration based on closed loop optimization is presented in [43]. Specifically, apart from the Day 0 operations that include VNF onboarding, the authors introduce a zero-touch slice deployment with intelligent

decisions on optimal placement for Day 1 operations. Day 2 operations include analytics functionalities and an optimization engine. A wide set of use cases such as optimal resource allocation, dynamic VNF placement and performance optimization are supported by the proposed framework. Similarly, in [44] a framework for the management of networks with massive network slices is proposed. This MANO framework achieves automation through the use of multiple, distributed, AI-driven control loops that can work at different levels (node, slice, interslice and orchestration domain level). Finally, a MANO framework for B5G vehicular edge service which is based on closed-loop orchestration, is presented in [45].

The CN focuses on AI/ML services in the control plane, targeting specific sessions, flows, or UE. These services aim to analyze or predict users' communication behavior, assess security risks, and ensure desired network performance. For example, the authors in [46] use analytics and ML techniques for three different use cases. Firstly, an Extreme Gradiant Boosting (XGBoost) model is implemented to obtain anomalous behaviour in UPF nodes. Secondly, they consider RNN, LSTM and Linear Regression (LR) models to predict load traffic in 5G cells and thirdly, a closed-loop automation model is implemented to predict SMF resource usage and automatically instantiate SMF instances. The authors in [47] propose the adoption of AI techniques in order to optimize placement and scaling aspects in the 5G CN. They explore how AI-based scaling algorithms combined with functionality-aware placement can enable the design of network slices. A mobility prediction ML-assisted scheme which reduces the signaling-induced overhead, is proposed in [48].

Within *RAN,* AI/ML utilizes real-time or near-real-time data to predict and analyze user access and dynamic radio conditions. The aim is to optimize tasks such as scheduling, interference control, and radio resource allocation. To this end, the authors in [48] propose an intelligent Radio Resource Management (RRM) scheme that aims to handle traffic congestion. The viability of such solution is tested on a real-world dataset. Finally, a network intelligence orchestration framework is presented in [49], which is designed within the concepts of O-RAN and automatically computes the optimal set of data-driven algorithms and their execution location.

Summarizing the above, although it is widely accepted that MANO and AI can provide significant benefits in automating and optimizing B5G systems operations, the majority of the existing studies treat these concepts in a segmented way. To the best of our knowledge this is the first study that addresses jointly the following topics:

Experimentally validates MANO over an operational 5G testbed considering 3GPP Rel 15 5G components.

Uses MANO to dynamically instantiate a wide range of 5G network deployment options, responding to a variety of operational scenarios, considering the concepts of functional split both at the RAN and at the Core segments.

Integrates MANO with AI algorithms to perform real time decision making, taking actions to rescale, modify or even delete individual 5G system building blocks.

Perform extensive experimentation quantifying resource requirements and (re)-configuration times for the deployed platform.

6.3. Proposed Platform Design

The main goal of this work is to design a framework that can assist network operators to easily deploy, manage, modify, reconfigure and delete 5G services. In order to do this, a platform/environment that takes advantage of all technologies and concepts entailed in the 5G ecosystem has been designed. This section describes this environment along with the tools and concepts that have been used to support its functionalities. The detailed structure of the proposed 5G management platform is illustrated in Figure 6. 3 comprising the following building blocks: (1) the 5G platform used to host the main 5G network functions, (2) the Data Collection and Monitoring block collecting statistics for the physical and virtual elements of the system, (3) the Predictive analytics (AI/ML) block used to assist decision making and (4) the Automated-LCM block. These building blocks are integrated to serve the purposes of this work as stated above. In the following subsections each block is presented in detail.

6.3.1. 5G Cloud Platform

5G Systems need to operate in highly heterogeneous environments with stringent, varying and sometimes conflicting service requirements. For this reason, traditional all-in-one deployments adopted from the previous generations of mobile networks have been replaced with more flexible and adjustable deployment options. Following this approach, the present framework supports a variety of 5G network deployment options that can be flexibly adapted to meet the requirements of the offered services. A graphical illustration of the main deployment options that are supported by the platform is shown in Figure 6. 3. These include:

The Monolithic deployment where all the functions are collocated in the same site. This option provides a fast, easy-to-deploy solution suitable for private 5G networks since

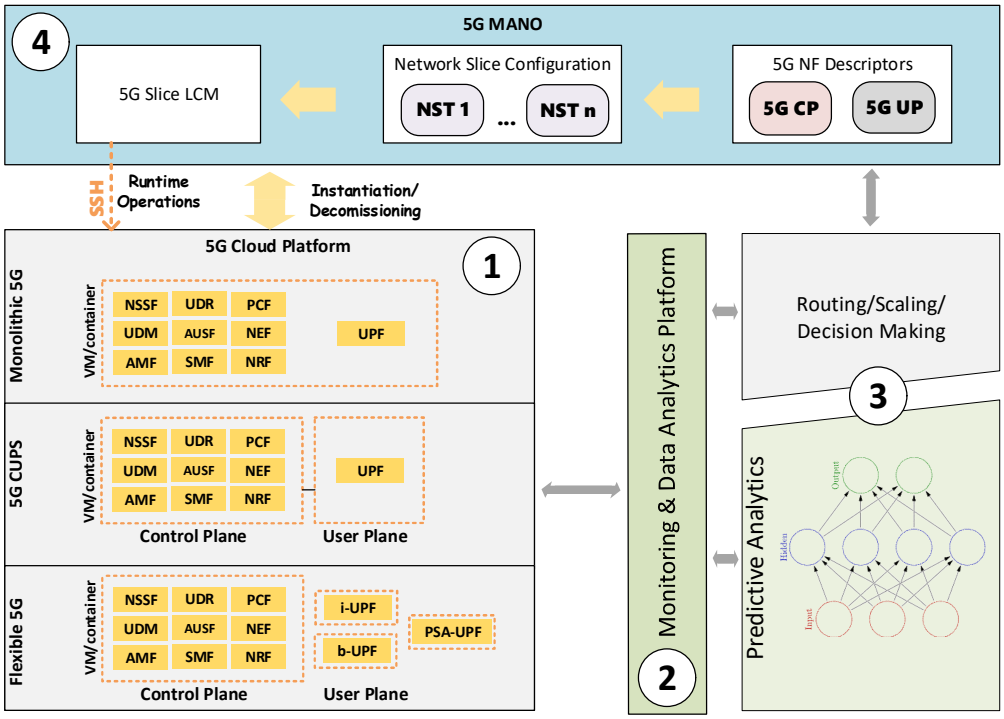


Figure 6. 3: Proposed Framework with 4 components: 1) 5G Cloud Platform supporting flexible deployment options, 2) Monitoring Platform, 3) Predictive Analytics/Decision Making 4) Automated LCM.

there are no shared elements and all NFs are fully isolated in one environment, providing enhanced security.

The CUPS deployment where the CP Functions are separated from the User Plane. Through this deployment option the UPF can be hosted in physical/virtual machines that are located close to the end-users, providing better support for delay-sensitive use cases. In this option, the CP and UP elements of the 5G platform are hosted in different machines[2].

The Flexible deployment option, where the concept of the CP/UP split is expanded to the CP elements, as well as the addition of extra UP nodes, Local Breakouts etc. Local Breakouts [50] refer to the capability of routing traffic flows directly from the edge of the network to its destination without passing through the core network, reducing latency and improving performance. This deployment option can support implementation of isolated network slices, applications with different QoS requirements, node up/down-scaling for environments supporting high mobility scenarios etc. This flexible approach supports diverse applications by isolating these through network slicing and enables services to adapt to dynamic traffic requirements, including high-mobility scenarios, through efficient scaling mechanisms.

6.3.2. Data Collection and Monitoring

In order to monitor the complex technologies and infrastructures deployed in the present study, the proposed framework includes a set of SA functions that comply with the 3GPP TS 23.501 standard [25]. By integrating monitoring tools in the framework, we can extract valuable data that map resource consumption to certain network functionalities. This enables MANO to make more informed network decisions, leading to improved performance, increased security, and reliability. For example, using data collected from network monitoring, malfunctions can be identified that combined with alerting mechanisms can trigger reconfiguration actions when needed.

In the current implementation, the monitoring platform consists of network agents that are placed in every compute and network component. These agents collect performance metric values from the hosts and expose these to certain interfaces. These values are then retrieved from the monitoring server and are stored in a database for further analysis, visualization, SLA and trending reporting. [51]. The monitoring platform is based on open-source tools, such as Prometheus and Grafana. Additionally, the devices are connected to energy metering sensors that collect energy-related data and store these to the Prometheus database. This is a valuable component, as for all system operations, such as service provisioning, network deployment and reconfiguration etc., resource consumption can be measured in compute, network and energy levels.

6.3.3. Predictive Analytics

The predictive analytics block of Figure 6. 3, interoperates with the Data Collection and Monitoring component of the platform to gather, process and analyze the retrieved data and make predictions regarding future data traffic load in the system. Data traffic forecasting is crucial for 5G networks, as it can lead to resource allocation efficiency, optimized network design and increased QoS management. By accurately predicting data traffic patterns, we can dynamically allocate resources depending on the demands

[2] The term "machine" refers to either VMs or containers, depending on the underlying virtualization layer

and avoid under/over-provisioning of resources in support of the respective functionalities.

In this direction, we have developed a forecasting model that is based on LSTM neural networks to realize this building block. The model was developed in Python using the Keras library. The model is implemented as follows:

- *Data Retrieval*: The data are retrieved from the Database through a simple Query
- *Data Preprocessing*: The data are first normalized and then split into a training set (67%) and a testing set (33%)
- *Model Training:* The model is trained through the Keras library
- *Prediction:* After completion of the model training the model is ready to make predictions

6.3.4. Automated Lifecycle Management

This block provides tools with which network operators can easily deploy, manage, reconfigure and rearrange the network. In this context, an LCM framework was developed that is able to dynamically manage network slices, leveraging domain NFVOs and controllers. The development of the framework was performed in three stages:

5G Deployment Option Selection: The 5G platform is able to instantiate a variety of deployment options spanning from the monolithic all-in-one to the fully disaggregated approach where CP and UP elements are separated. To achieve this, the automated LCM framework is able to apply 5G network configuration policies that can be used to instantiate:

- CP and UP entities that are fully separated,
- multiple UPF elements with different roles. The LCM framework is able define the number and type of UPF elements in the User Plane path. Therefore, it is able to instantiate:
    - a PDU-PSA UPF that acts as a single termination point for the PDU Session.
    - anI-UPF: This UPF is located within the path between the RAN and the PSA-UPF and is responsible for forwarding data between the RAN and the PSA-UPF .
    - a B)UPF: The B-UPF redirects uplink traffic to the appropriate UPF that ends the PDU Session and merges the downlink traffic from different PSA-UPFs to the UE.
- Network Slices with different QoS characteristics
- ASs processing users' data. AS can be hosted at various locations i.e., on traditional CCs, or on computing resources that are closer to the network edge (MEC), enabling low-latency, high-bandwidth applications in 5G networks.

5G NF Automation: The next step maps the above elements to the necessary descriptors that feed the NFVOs. In the current implementation, OSM has been used as the main

NFVO, since its Information Model (IM) is ETSI NFV-aligned and is agnostic of the underlying infrastructure, so that its models can be used across various VIM types and transport technologies. OSM uses configuration templates, called descriptors, to describe the key characteristics of managed objects (e.g VNFs or NSs) in a network. For each component, descriptors specify how it will be deployed and used, as well as how it will interact with other components. Descriptors are written in YAML, a markup language designed for data that is easy to read and understand.

Therefore, to automate the deployment process in 5G networks the required VNFs for the main 5G elements, i.e., VNFs for 5G CP and UP have been implemented. The specifications of the VMs that comprise VNFs along with their connections are exposed to OSM through the VNFD. Thus, two VNFDs, one for the 5G-CP and one for the UPF were created.

In order to be able to flexibly mix and match the two 5G planes creating multiple 5G topologies, we considered each 5G plane as a distinct Network Service in the context of OSM and two NSDs are created, one for each plane. An NSD is a higher-level abstraction that defines the structure and behavior of a network service composed of multiple VNFs. NSDs reference one or more VNFDs to specify the VNFs that compose the network service, their arrangement and the connections between them. In this work, 5G NSDs describe the network connections of the 5G VNFs internal to the private cloud. Figure 6. 4 shows a graphical illustration of the 5G VNFDs and NSDs that were developed. The created descriptors represent a generic implementation of the two 5G planes, but they can be parametrized at the instantiation of the slice, in order to be tailored to support its specific requirements.

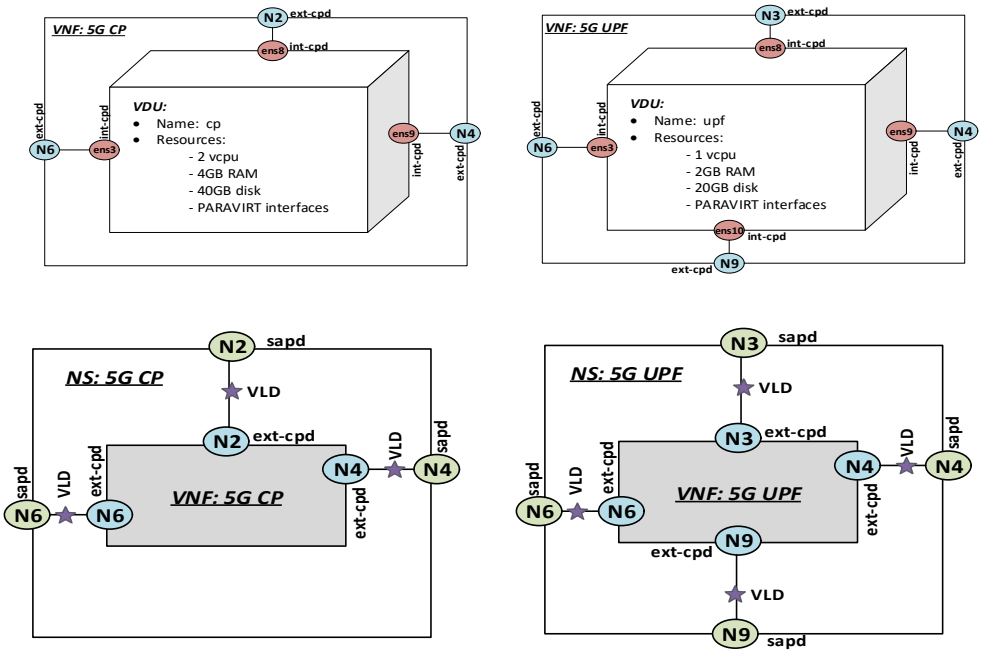


Figure 6. 4: Graphical illustration of the VNFDs and NSDs for 5G CP and UP respectively. In the VNFD we create the appropriate network interfaces in the VMs that host the 5G NFs, and in the NSDs we connect these interfaces to the appropriate networks inside the private cloud

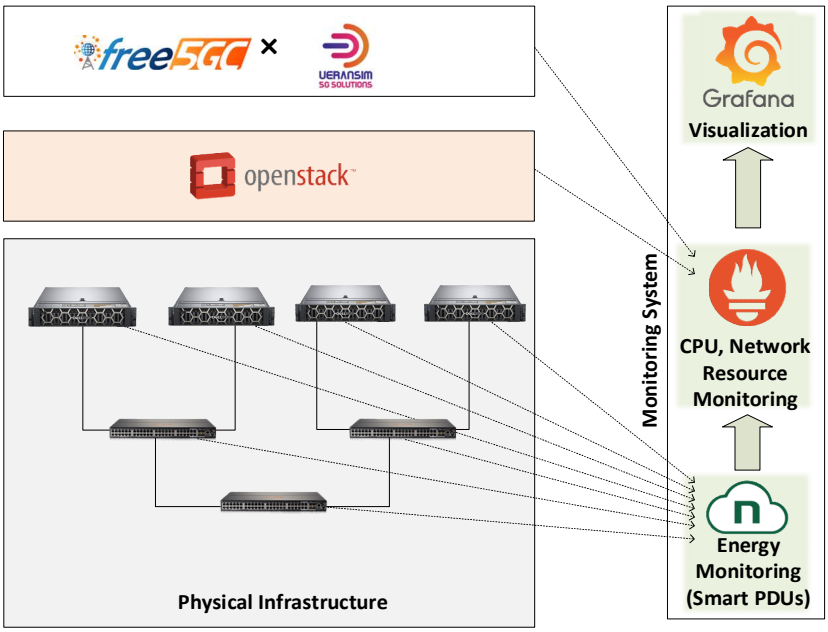


Figure 6. 5: Lab testbed overview.

Slice Deployment: In the next step, the requirements of the service slices are given as input. These are then automatically converted into configuration parameters for the generic 5G descriptors. Subsequently, the framework prepares a NST, that describes the characteristics, requirements, and behavior of the network slice. The NST comprises two main blocks, the NSSIs and the slice virtual links (VLDs). The NSSIs include the 5G CP and UP NSs and can be shared among multiple slices. The VLDs include the network connections that are required for the slice. The NST is provided to the NFVO that is responsible for the instantiation of the slice.

The LCM of the slice can be performed either by the NFVO, or directly by the LCM framework. In the first scenario, we use the VCA layer of OSM that uses juju [52] software to achieve the 3-step MANO lifecycle of 5G slices. Juju uses a set of generic scripts and metadata, called charms, that encapsulate DevOps expertise and can be adapted for various software deployments. Charms are given to OSM along with the VNFDs and configure VNFs by executing automated scripts via Secure Shell (SSH). They are written in any language executable from the command line and consist of YAML configuration files and "hooks" which manage software installation, service control, charm configuration, and interactions between charms.

Charms perform actions that are classified into:

- actions that are automatically performed at the instantiation of the slices (Day 1 actions)), and
- actions that can be dynamically performed during the deployment of the slice (Day 2 actions).

The charms that were developed for the appropriate operation of the 5G control and user planes, perform the following actions:

- Day 1:

1. Configure ssh access and IPs of the VMs

2. Manipulation of the configuration files for the proper operation of 5G core and UPF
3. Role of UPF (i-upf or psa with N3/N9 interface)
4. Load required modules (e.g. GTP-U tunnel for UPF)
5. Start UPFs

– Day 2:

- Start/Stop 5G CP
- Start/Stop UPFs

Charms can be allocated within the VMs of the VNFs (native) or, most commonly, are hosted in LXD containers within the OSM machine (proxy). When deploying a proxy charm, several time-consuming steps occur by default. The LXD container hosting the charm must be launched and configured and the charm must be installed. For time sensitive scenarios, in order to avoid delays, the developed LCM framework can bypass the need of juju charms, by directly performing the necessary actions for the slices appropriate functionality through SSH connections. The NFVO in this scenario is responsible for the allocation of the appropriate resources to the slice, and the rest of the configuration and run-time operations are performed through the LCM framework.

, and behavior of the network slice. The NST comprises two main blocks, the NSSIs and slice VLDs. The NSSIs include the 5G CP and UP NSs and can be share among multiple slices. The VLDs include the network connections that are required for the slice. The NST is provided to the NFVO that is responsible for the instantiation of the slice.

The lifecycle management of the slice can be performed either by the NFVO, or directly by the LCM framework. In the first scenario, we use the VCA layer of OSM that uses juju software to achieve the 3-step MANO lifecycle of 5G slices. Juju uses a set of generic scripts and metadata, called charms, that encapsulate DevOps expertise and can be adapted for various software deployments. Charms are given to OSM along with the VNFDs, and configure VNFs by executing automated scripts via SSH. They are written in any language executable from the command line and consist of YAML configuration files and "hooks," which manage software installation, service control, charm configuration, and interactions between charms.

Charms perform actions that are distinguished to:

actions that are automatically performed at the instantiation of the slices (Day 1 actions)), and

actions that can be dynamically performed during the deployment of the slice (Day 2 actions).

The charms that were developed for the proper operation of the 5G control and user plane, perform the following actions:

Day 1:

- Configure ssh access and IPs of the VMs
- Manipulation of the configuration files for the proper operation of 5G core and UPF

- Role of UPF (i-upf or psa with N3/N9 interface)
- Load required modules (e.g. GTPU tunnel for UPF)
- Start UPFs

Day 2:

- Start/Stop 5G CP
- Start/Stop UPFs

Charms can be allocated within the VMs of the VNFs (native) or, most commonly, are hosted in LXC containers within the OSM machine (proxy). When deploying a proxy charm, several time-consuming steps occur by default. The LXD container hosting the charm must be launched and configured and the charm must be installed. For time sensitive scenarios, in order to avoid these time delays, the developed LCM framework can bypass the need of juju charms, by directly performing the necessary actions for the proper functionality of the slices through SSH connections. The NFVO in this scenario is responsible for the allocation of the appropriate resources to the slice, and the rest configuration and run-time operations are performed through the LCM framework.

6.4. Testbed Implementation and Experimental Results

The framework along with all the relevant building blocks that was presented in Section 6.3, aims to provide a tool that MNOs can utilize to support a wide variety of use cases, services and applications. In this Section, two comprehensive implementations are presented, both carried out by the proposed framework. The first implementation involves a multi-slice network deployment, where each new slice can be instantiated on-the-fly, without any disruption to other parts of the network. The second implementation includes a real-time network configuration in terms of UPF nodes, which is based on predicted compute resource requirements provided by the analytics block of the framework. The details for the two implementations are provided in Subections 6.4.2 and 6.4.3.

6.4.1. Experimental Testbed

The lab testbed along with the related open source components used is depicted in Figure 6. 5. The physical infrastructure comprises a set of servers, optoelectronic switches, routers and physical links. All the physical resources are clustered into a (openstack) private Cloud platform that is used to deploy and host all 5G RAN & Core functions. The 5G CN is deployed through the free5gc open-source platform while for the RAN side, UERANSIM is used. The monitoring and data collection platform stores, visualizes and monitors the system resources. Metrics from all compute and network resources are collected and exposed from agents to specific ports and are then retrieved and stored in the Prometheus database. Additionally, energy consumption metrics are retrieved from energy metering devices. All the metrics are visualized through Grafana. Finally, OSM is used as a MANO platform for the automated deployment of CP and U P Core functions.

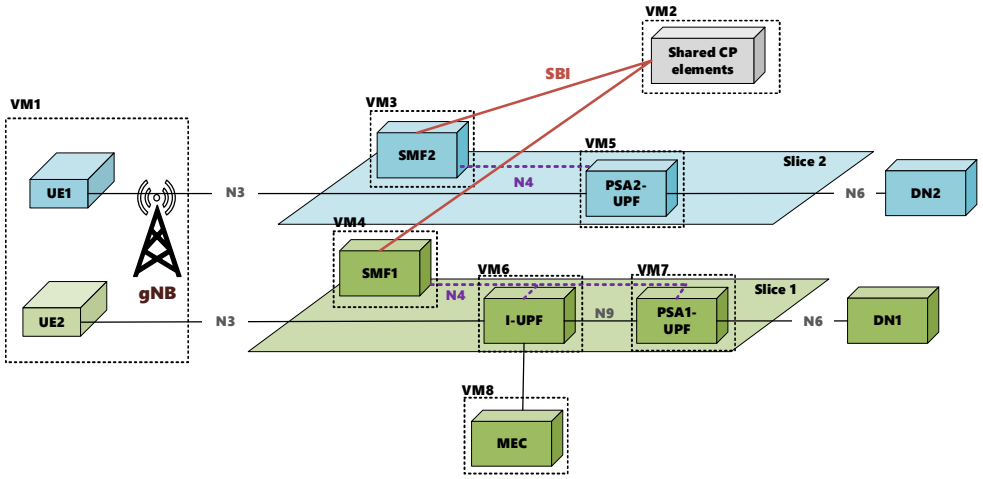


Figure 6. 6: 5G topology considered for automated LCM. The topology consists of two slices that have distinct user plane paths, and SMFs, but share the other CP 5G NFs.

6.4.2. Automated LCM of Network Slices

The first part of the implementation includes an automated deployment of a functional 5G network with two configured slices as shown in Figure 6. 6. The environment is hosted in a total of eight virtual instances. The first slice consists of an SMF node, two UPF nodes and one Local Breakout loop targeting delay sensitive applications that are served through a local MEC node. The second slice is deployed with one SMF and one UPF node. Following the SBA paradigm, dedicated subnets are created for the external interfaces (i.e. N2, N3, N4, N6, N9) and each virtual instance uses a separate virtual network interface to enable isolation and ease the monitoring procedure of each protocol. For the internal interfaces/NFs a loopback network is used.

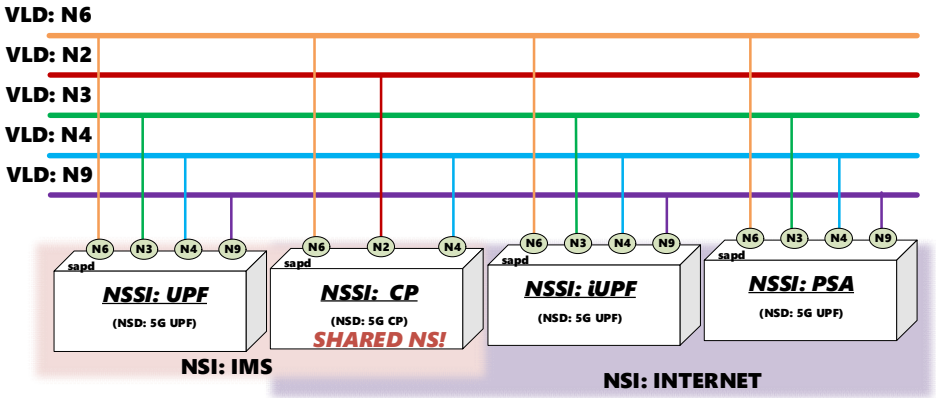


Figure 6. 7: Graphical illustration of two NSTs for two 5G slices.

The 5G core is automatically instantiated through the LCM platform. The LCM framework maps the CP and UP NFs from free5gc to two different VNFs. Their combination to create a specific slice is described in the NST. Figure 6. 7 shows the two

NSTs that were developed for the two slices. The two slices share the CP elements but have distinct UPF nodes.

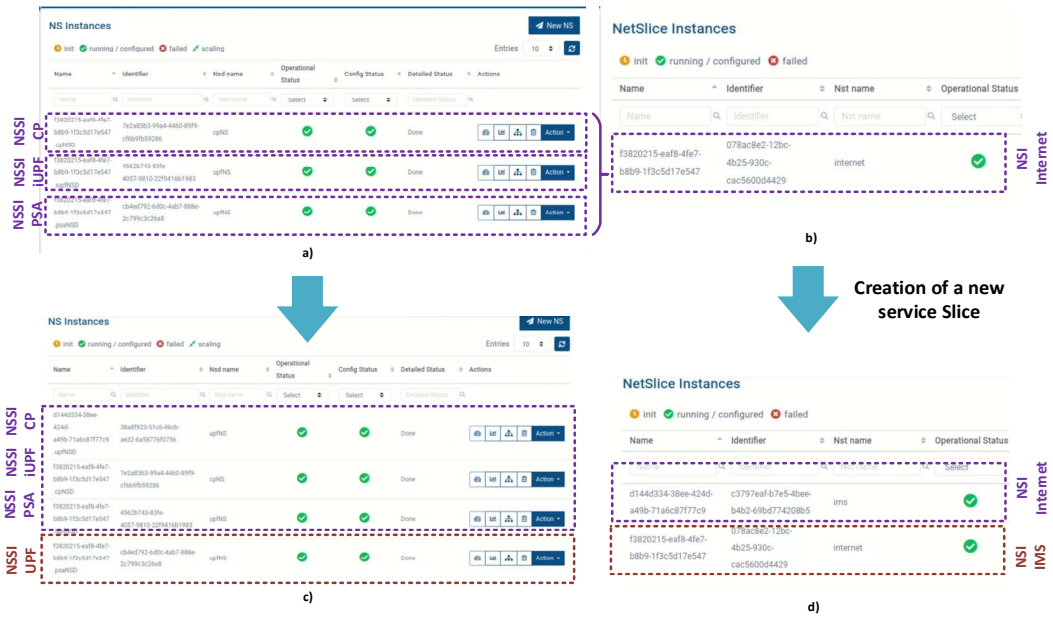


Figure 6. 8: MANO-triggered Slice instantiation: a) NS Instances: The main 5G networks elements (CP, iUPF and PSA) implementing the "Internet" Slice shown in Figure 6. 7 are created. b) The "Internet" slice is instantiated. c) Creation of a new slice (IMS) is requested. This requires a new UPF node to be added. d) The IMS slice is instantiated. The platform now supports the "Internet" and the "IMS" slices.

The deployment is shown in Figure 6. 8. First, the main 5G network elements (CP, I-UPF, and PSA-UPF) for the "Internet" slice are created, and the slice is instantiated (Figure 6. 8 a-b). A new slice, 'IMS,' is then requested. This requires the addition of a new UPF node. The orchestrator is aware of the already instantiated shared CP elements and only instantiates the new UPF node for this slice. Once the 'IMS' slice is instantiated (Figure 6. 8 c-d), the platform supports both the 'Internet' and 'IMS' slices.

In this procedure, initially OSM takes the responsibility to instantiate the slice, by allocating the necessary resources. Then the LCM framework performs a set of configuration actions to enable the required functionality of the two 5G planes. These include:

- Manipulation of the configuration files according to the input parameters for appropriate operation of the 5G core and UPF
- Configuration of the UPF role for the 5G UP (I-UPF or PSA with N3/N9 interface)
- Load of the GTPU tunnel module to the VM that hosts the 5G UP

Finally, the LCM starts the CP and UP NFs, and the slice becomes operational. It is important to note that LCM can perform this type of actions (start/stop NFs) dynamically during the deployment of the slice.

6.4.3. ZSM Mechanism for Load Traffic Data Management

The second part of the implementation presented in this work aims to employ a mechanism that follows the ZSM paradigm. The mechanism relies on all four building blocks that were presented in Section 6.3; hence, the analytics block is added. The idea is to develop a framework that optimizes resource utilization without human intervention, in a dynamic environment where data traffic load can vary significantly in time. Specifically, the framework includes:

A simple 5G CN topology based on the CP/UP split and one gNB.

City-wide mobile network traffic statistics: The second set of measurements that are used for the development of the AI/ML models include large scale mobile network traffic statistics that are available online [53]. The traffic statistics were captured from a Base Station (BS) with a varying number of connected users over a two-week period.

5G related measurements collected from our lab testbed. The network traffic statistics were mapped to 5G network and compute resources through iperf connections. For each user, an iperf connection was made with a static allocated bandwidth. The instances that host the 5G environment are continuously monitored and the resulting data are collected and stored in the Prometheus database.

A two-stage NN model implemented to proactively react to traffic fluctuations. The first stage includes an *LSTM-based forecasting algorithm* that collects and processes the monitored data to predict future load data traffic. The second stage includes a *Decision algorithm* that evaluates the predictions and triggers the instantiation of new UPF nodes to split the data traffic.

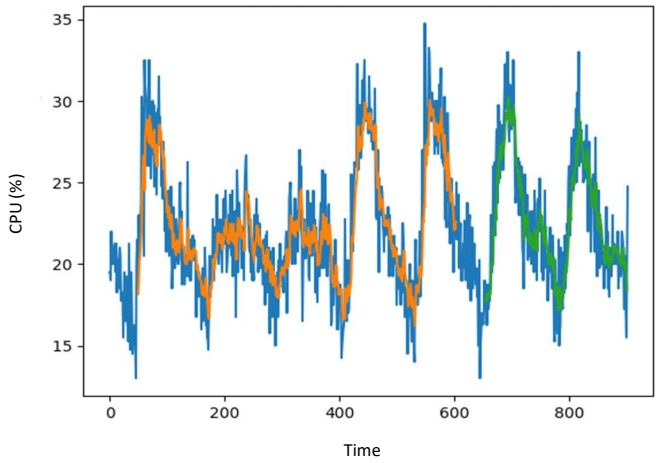


Figure 6. 9: Recorded compute resource utilization measurements for virtualized UPF. (blue line: raw data set, orange line: training dataset, green line: test dataset)

A numerical example showing the prediction of the UPF CPU load, based on the available history CPU data that are extracted from our lab's monitoring system is shown in Figure 6. 9. The LSTM input vector corresponds to the total UPF load at an arbitrary time step t, while the LSTM output vector corresponds to the total load at time step t+1. To train the LSTMs, the dataset containing history measurements of each UPF is split into two parts, the training set and the test set. The training set is used during the

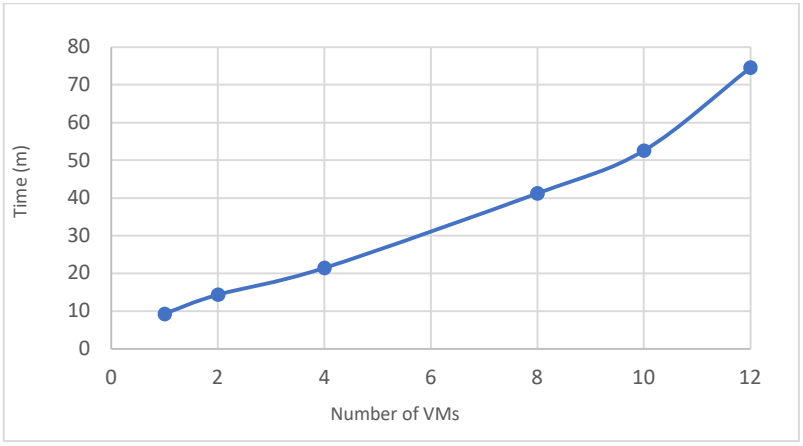
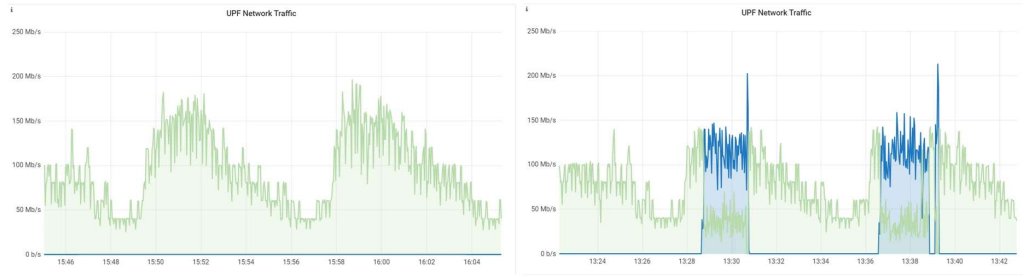


Figure 6. 10: UPF instantiation time.

training of the LSTM network, while the test set is used to validate the effectiveness of each LSTM designed.

During the predictive stage, a critical parameter that should be carefully considered in the decision-making process is the prediction horizon. The prediction horizon should be at least equal to the time needed by the system to calculate and then apply the optimal reconfiguration and resource allocation policies (i.e., modify/add/delete routing paths or compute resources). To determine the prediction horizon, we measured the time needed from OSM to instantiate a UPF instance in our private cloud. The relevant results are shown in Figure 6. 10.

The model at each time t predicts the value of required compute resources to accommodate the traffic at the next timestep. If this value exceeds a predefined threshold, the algorithm triggers instantiation of a new UPF node creating an SSH connection to the orchestrator. The threshold depends on the capacity of the UPF node (250 Mb/s in our case). The prediction horizon of the algorithm is set to 4min so that the MANO can apply appropriate configurations in time, as discussed earlier.



(a) (b)

Figure 6. 11: System Performance Evaluation a) Single-upf traffic distribution b) Two-upf traffic distribution

Once the prediction exceeds the threshold, the new UPF node is instantiated. The resulting topology has now two available UPF nodes to handle the incoming traffic which is distributed in both available instances. Figure 6. 11 provides system performance snapshots from Grafana. Figure 6. 11 a) shows the distribution of traffic without the implementation of the algorithm (single-VM) and in Figure 6. 11 b), the traffic is distributed among two nodes. The proposed framework optimizes the

utilization of resources by allocating network and compute resources when they are needed based on the traffic predictions and releasing them when load traffic decreases. The addition of the extra node leads to extended network capacity for the system and avoids stretching the available resources of a single node above its optimal levels, leading to a smooth system behavior.

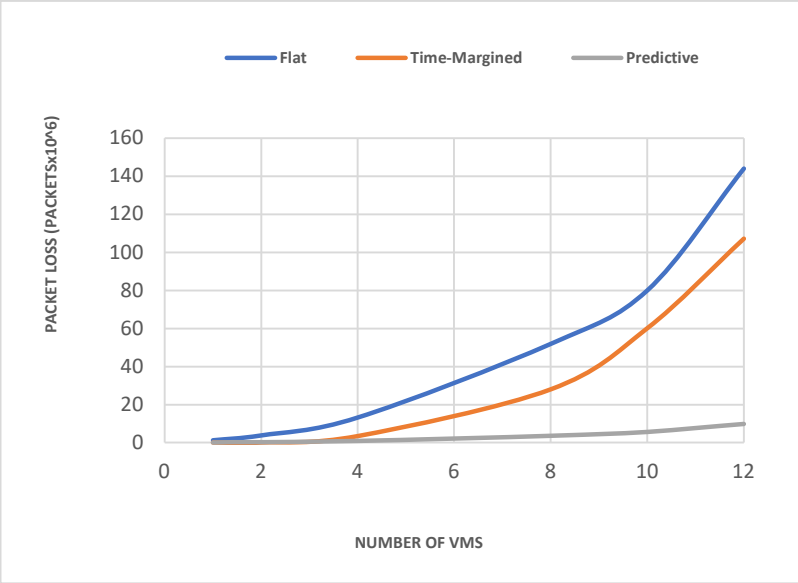


Figure 6. 12: System Packet Loss for varying number of VMs. Three different strategies have been employed for the instantiation of the VMs (flat, time-margined, predictive).

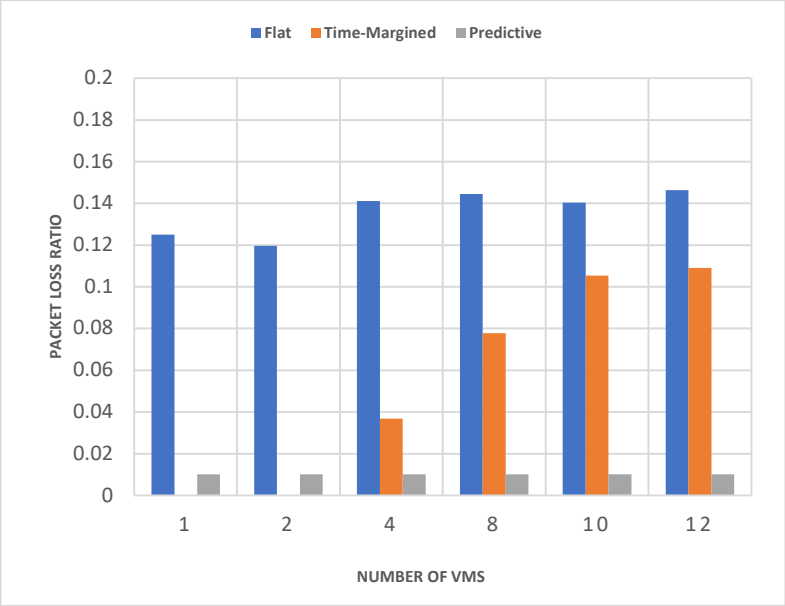


Figure 6. 13: Packet Loss Ratio for varying number of VMs for the three different approaches (flat, time-margined, predictive).

To validate the efficiency of the developed framework, we consider a scenario with multiple gNBs each served by a single UPF. Similar network patterns are assumed for all the nodes in the system. Based on the data traffic statistics extracted from the monitoring platform, a decision must be made whether the deployment of additional

VM(s) is needed or not. For this decision we define three strategies (flat, time-margined, predictive) and compare their performance in terms of packet delivery. Specifically, the flat strategy generates the deployment of a VM when network traffic reaches the threshold (that was defined before), the time-margined strategy generates the VM when the traffic levels reach the threshold minus a time margin and the predictive strategy initiates the VM deployment based on the predictions of the algorithm. The graph in Figure 6. illustrates the results of each strategy for varying number of VMs in terms of packet loss, where the flat strategy presents the highest levels of packet loss, followed by the time-margined approach. The predictive strategy achieves the best performance since packet loss is only related to the prediction error. Similarly, Figure 6. presents the Packet Loss Ratio (PLR) for each approach. It is worth mentioning that the time-margined strategy performs very well, especially for a small number of VMs but at the expense of risking an erroneous instantiation, i.e. the load traffic reaches the time-margined threshold and then decreases. In our dataset, twelve cases of erroneous instantiation were observed.

6.5. Conclusion

In this chapter, a MANO framework based on OSM has been proposed and developed, that specifically targets orchestration operations of B5G networks. We have created network descriptors for the core and the user plane network elements. Combining those descriptors, we can successfully deploy dynamic network slices. In order to test the validity and performance of the proposed framework we demonstrated two use cases. The first focuses on the dynamic deployment of network slices on top of a softwarized multi-operator 5G platform hosted in our private lab testbed. The second part of the implementation concentrated on the demonstration of a proactive UPF provisioning mechanism, ensuring that the system can detect on time the compute and network demands of a slice, that may change dynamically, and adapt to these demands accordingly. Both cases highlight the ZSM approach of our network, and its capability to directly and dynamically manage 5G elements and optimize resource utilization.

References

[1] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, “A Survey on Mobile Edge Computing: The Communication Perspective,” in IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, Fourthquarter 2017. Available: https://doi.org/10.1109.

[2] "Cloud RAN and MEC: A Perfect Pairing", Etsi.org, 2019. [Online]. Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp23_MEC_and_CRAN_ed1_FINAL.pdf.

[3] Barakabitze, Alcardo & Ahmad, Arslan & Hines, Andrew & Mijumbi, Rashid, “5G Network Slicing using SDN and NFV: A Survey of Taxonomy, Architectures and Future Challenges.”, in Computer Networks, Vol. 167, 106984, 2020. Available: https://doi.org/10.101.

[4] Rommer, S., Hedman, P., Olsson, M., Frid, L., Sultana, S., Mulligan, C., 5G Core Networks: Powering Digitalization, ISBN: 9780081030097, Elsevier Science, 2019.

[5] Imoize, Agbotiname & Adedeji, Oluwadara & Tandiya, Nistha & Shetty, Sachin. (2021). 6G Enabled Smart Infrastructure for Sustainable Society: Opportunities, Challenges, and Research Roadmap. Sensors. 21. 1-57. 10.3390/s21051709.

[6] https://5g-ppp.eu/verticals/.

[7] “3GPP, TR 22.864, "Feasibility study on new services and markets technology enablers for network operation; Stage 1",” [Online].

[8] M. E. Morocho-Cayamcela, H. Lee and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," in IEEE Access, vol. 7, pp. 137184-137206, 2019, doi: 10.1109/ACCESS.2019.2942390.

[9] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran and Q. Emad Ul Haq, "Machine Learning Techniques for 5G and Beyond," in IEEE Access, vol. 9, pp. 23472-23488, 2021, doi: 10.1109/ACCESS.2021.3051557.

[10] ETSI NFV, see http://www.etsi.org/technologies-clusters/technologies/nfv.

[11] 3GPP, “TS 28.531 Management and Orchestration; Provisioning,” January 2020.

[12] ETSI. Open Source MANO. [Online] OSM (etsi.org)

[13] The Linux Foundation. ONAP. [Online] https://www.onap.org/.

[14] Liyanage, M., Pham, Q. V., Dev, K., Bhattacharya, S., Maddikunta, P. K. R., Gadekallu, T. R., & Yenduri, G. (2022). A survey on Zero touch network and Service Management (ZSM) for 5G and beyond networks. *Journal of Network and Computer Applications*, *203*, 103362.

[15] A. -I. Manolopoulos, V. -M. Alevizaki, M. Anastasopoulos and A. Tzanakaki, "An AI-Assisted Framework for Lifecycle Management of Beyond 5G Services," in IEEE Access, doi: 10.1109/ACCESS.2024.3507359.

[16] 3GPP, TR 22.863, "Feasibility study on new services and markets technology enablers for enhanced mobile broad-band; Stage 1" [Online].

[17] 3GPP, TR 22.861, "Feasibility Study on New Services and Markets Technology Enablers for massive Internet of Things; Stage 1", [Online].

[18] “3GPP, TR 22.862, "Feasibility study on new services and markets technology enablers for critical communications; Stage 1", [Online].

[19] "Microservices.io (2020). What are microservices? [Online]. Available: https://microservices.io/".

[20] Ian F. Akyildiz, Shuai Nie, Shih-Chun Lin, Manoj Chandrasekaran, “5G roadmap: 10 key enabling technologies”, Computer Networks, Volume 106, 2016, Pages 17-48, ISSN 1389-1286. Available: https://doi.org/10.1016/j.comnet.2016.06.010.

[21] 5G; NR; NR and NG-RAN Overall description; Stage-2, ETSI TS 138 300 V16.4.0 (2021c). [Online]

[22] Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV, ETSI GS NFV 003 V1.2.1 (2014). [Online].

[23] F. T. Kuhn, F. Schnicke and P. O. Antonino, “Service-Based Architectures in Production Systems: Challenges, Solutions & Experiences,” in ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K), Ha Noi, Vietnam, 2020.

[24] View on 5G Architecture, 5G PPP Architecture Working Group, Version 3.0, 06-2019. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2019/07/5G-PPP-5G-Architecture-White-Paper_v3.0_PublicConsultation.pdf

[25] System architecture for the 5G System (5GS), 3GPP TS 23.501 version 16.6.0 Release 16. [Online].

[26] Akyildiz, Ian & Nie, Shuai & Lin, Shih-Chun & Chandrasekaran, Manoj, “5G Roadmap: 10 Key Enabling Technologies,” in Computer Networks, Vol 106, 2016. Available: https://doi.org/10.1016/j.comnet.2016.06.010.

[27] Network Functions Virtualisation (NFV); Architectural Framework, ETSI GS NFV 002 V1.2.1 (2014a). [Online].

[28] Network Functions Virtualisation (NFV); Management and Orchestration, ETSI GS NFV-MAN 001 V1.1.1 (2014b). [Online].

[29] X. Li et al., "Automated Service Provisioning and Hierarchical SLA Management in 5G Systems," in IEEE Transactions on Network and Service Management, vol. 18, no. 4, pp. 4669-4684, Dec. 2021, doi: 10.1109/TNSM.2021.3102890.

[30] Dreibholz, T. (2020). Flexible 4G/5G Testbed Setup for Mobile Edge Computing Using OpenAirInterface and Open Source MANO. In: Barolli, L., Amato, F., Moscato, F., Enokido, T., Takizawa, M. (eds) Web, Artificial Intelligence and Network Applications. WAINA 2020. Advances in Intelligent Systems and Computing, vol 1150. Springer, Cham

[31] N. Makris, C. Zarafetas, A. Valantasis and T. Korakis, "Service Orchestration Over Wireless Network Slices: Testbed Setup and Integration," in *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 482-497, March 2021

[32] R. Inam, A. Karapantelakis, K. Vandikas, L. Mokrushin, A. Vulgarakis Feljan and E. Fersman, "Towards automated service-oriented lifecycle management for 5G networks," 2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA), Luxembourg.

[33] NGMN 5G Initiative Team, “A Deliverable by the NGMN Alliance: NGMN 5G White Paper”. [Online]. Available: https://www.ngmn.org/. wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf.".

[34] Study on management and orchestration of network slicing for next generation network (Release 15), Technical Specification Group Services and System Aspects, Telecommunication management, 3GPP TR 28.801 V15.1.0, January 2018.

[35] E. Chirivella-Perez, P. Salva-Garcia, I. Sanchez-Navarro, J. M. Alcaraz-Calero and Q. Wang, "E2E network slice management framework for 5G multi-tenant networks," in Journal of Communications and Networks, vol. 25, no. 3, pp. 392-404, June 2023, doi: 10.2.

[36] P. Vanichchanunt, O. Ritruechai, N. Wuttiananchai, P. Thossaporn, L. Wuttisittikulkij and S. Paripurana, "Implementation of 5G Network Slicing Using Open Source Software," *2024 12th International Electrical Engineering Congress (iEECON)*, Pattaya, Thailand, 2024, pp. 1-6, doi: 10.1109/iEECON60677.2024.10537902.

[37] X. Li, R. Ni, J. Chen, Y. Lyu, Z. Rong and R. Du, "End-to-End Network Slicing in Radio Access Network, Transport Network and Core Network Domains," in IEEE Access, vol. 8, pp. 29525-29537, 2020, doi: 10.1109/ACCESS.2020.2972105.

[38] X. Shen et al., "AI-Assisted Network-Slicing Based Next-Generation Wireless Networks," in IEEE Open Journal of Vehicular Technology, vol. 1, pp. 45-66, 2020, doi: 10.1109/OJVT.2020.2965100.

[39] B. Zhao, H. Lu, S. Chen, J. Liu and D. Wu, "Convolutional neural networks for time series classification," in *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162-169, Feb. 2017, doi: 10.21629/JSEE.2017.01.18.

[40] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

[41] S. Zhao, S. Lin and J. Xu, "Time Series Traffic Prediction via Hybrid Neural Networks," *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand, 2019, pp. 1671-1676, doi: 10.1109/ITSC.2019.8917383.

[42] K. Samdanis, A. N. Abbou, J. Song and T. Taleb, "AI/ML Service Enablers and Model Maintenance for Beyond 5G Networks," in *IEEE Network*, vol. 37, no. 5, pp. 162-172, Sept. 2023, doi: 10.1109/MNET.129.2200417.

[43] K. Govindarajan et al., "Closed loop optimization of 5G network slices," 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2023, pp. 186-188, doi: 10.1109/COMSNETS56262.2023.10041293.

[44] N. Slamnik-Kriještorac et al., "AI-Empowered Management and Orchestration of Vehicular Systems in the Beyond 5G Era," in IEEE Network, vol. 37, no. 4, pp. 305-313, July/August 2023, doi: 10.1109/MNET.008.2300024.

[45] L. Blanco *et al.*, "AI-Driven Framework for Scalable Management of Network Slices," in *IEEE Communications Magazine*, vol. 61, no. 11, pp. 216-222, November 2023, doi: 10.1109/MCOM.005.2300147.

[46] M. Ramachandran, T. Archana, V. Deepika, A. A. Kumar and K. M. Sivalingam, "5G Network Management System With Machine Learning Based Analytics," in IEEE Access, vol. 10, pp. 73610-73622, 2022, doi: 10.1109/ACCESS.2022.3190372.

[47] A. Sheoran, S. Fahmy, L. Cao and P. Sharma, "AI-Driven Provisioning in the 5G Core," in *IEEE Internet Computing*, vol. 25, no. 2, pp. 18-25, 1 March-April 2021, doi: 10.1109/MIC.2021.3056230.

[48] J. Jeong *et al.*, "Mobility Prediction for 5G Core Networks," in *IEEE Communications Standards Magazine*, vol. 5, no. 1, pp. 56-61, March 2021, doi: 10.1109/MCOMSTD.001.2000046.

[49] S. D'Oro, L. Bonati, M. Polese and T. Melodia, "OrchestRAN: Orchestrating Network Intelligence in the Open RAN," in *IEEE Transactions on Mobile Computing*, vol. 23, no. 7, pp. 7952-7968, July 2024, doi: 10.1109/TMC.2023.3342711.

[50] Service requirements for the Evolved Packet System (EPS), 3GPP TS 22.278 version 13.4.0 Release 13. [Online]".

[51] D. Giannopoulos, P. Papaioannou, C. Tranoris and S. Denazis, "Monitoring as a Service over a 5G Network Slice," 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Porto, Portugal, 2021, pp. 329-334, doi: 10.1109/E.

[52] Juju, [Online] https://juju.is

[53] T. Italia, ""Telecommunications - SMS, Call, Internet - MI", Harvard Dataverse, V1," 2015. [Online]. Available: https://doi.org/10.7910/DVN/EGZHFV.

Chapter 7 CONCLUSIONS AND FUTURE WORK

This thesis focused on the system-level design of mobility management solutions for advanced mobile networks that are deployed in cloud infrastructures. The relevant work considered various architectural models and technologies relevant in 5G/B5G networks. The different architectural options under consideration were implemented and evaluated adopting suitable mathematical models, protocols and algorithms. In order to experimentally validate the proposed solutions, a 5G cloud testbed was developed. This test-bed comprised interconnected compute, networking and storage elements and was equipped with enhanced featured supported by a monitoring platform, an orchestration platform (MANO) as well as AI/ML tools, all based on open-source software.

Optimal design of the RAN, CN and DN segments was performed taking into account user mobility. To solve this problem the concept of resource disaggregation was adopted, that allows separation of the functions of each domain from the underlying hardware components. Thus, for the RAN domain the BBU functionalities can be split into three local components: the CU, DU and RU. For the CN domain we investigated the CUPS paradigm which separates the CP functionalities from the UP. The resulting network segments can be placed across different geographical locations depending on their requirements in terms of latency, bandwidth etc. Furthermore, by leveraging the feature of live VM migration, each of these components can be migrated across different geographical locations, without disrupting the provided services. We then formulated a mathematical model that aims to optimize the deployment of 5G systems in environments with user mobility. The objective was to identify the optimal hosting locations for the virtualized functions of the 5G system and the VMs that provide the services, with the aim to minimize the overall cost of the resulting network configuration. Additionally, an analysis was carried out that used realistic measurements extracted from an actual 5G standalone system that has been deployed in a private cloud environment, increasing the validity of the model. By performing extensive profiling, critical parameters could be determined that affect the performance of the system such as live VM migration-induced traffic and overheads, as well as the impact of wireless network traffic on RAN/CN components and the correlation between wireless access network traffic and mobility. The performance of the system was finally evaluated under different 5G deployment strategies including the RAN, CN and AF. Each strategy succeeds a specific optimization objective leading to the minimization of compute and network costs, minimization of overheads during migration of the AS and balancing of compute resources utilization. Based on our results, the strategy which considers mobility in the placement of network functions demonstrates the best performance, while network and cost savings increase with UE mobility.

Next a multi-technology network comprising 3GPP and non-3PP access technologies was considered where end-to-end slices interconnecting UEs with the DN were established. In this system, if a user was connected to a 3GPP RAN, the corresponding slice was formed through the gNB, UPF and DN nodes, whereas, if it was connected through a non-3GPP AN, the corresponding slice comprised the AN, N3IWF,UPF and DN nodes. This network was modeled as mixed network of queues where mobility

induced handovers were managed either by maintaining the same access technology or by switching to a different AN. Seamless service provisioning wass ensured by reserving parts of the available physical resources for prospective UE arrivals. Depending on their mobility profiles which were formed based on speed and arrival rates, UEs were directed to the appropriate access technology. Then, an overall network selection scheme was developed through a simple two-dimensional MC model, where the relative parameters are defined by experimental data extracted by the 5G cloud testbed. The model identified an optimal threshold related to the amount of reserved resources.

The final work reported in this thesis concentrated on the development of a B5G framework targeting to automate the provisioning, deployment, management and orchestration of network slices and services. In this framework, 5G functions can be easily deployed, managed, modified and deleted and, at the same time re-configuration actions can be performed autonomously without the need of human intervention. To perform this, the framework relied on four building blocks: a 5G cloud platform, a monitoring platform, a MANO platform and an AI/ML block. The validity and efficiency of the deployed system was verified by a two-stage evaluation process. The first stage focused on an orchestrator triggered multi-slice network deployment where each slice was dedicated to specific service use cases. The second stage included a ZSM operation designed to reconfigure the number of UPF nodes by predicting future resource consumption, based on real user mobility statistics. The framework was evaluated in terms of packet loss, where it showed significant improvement compared to other strategies.

Several challenges remain open for future work regarding this research area. Having deployed a platform that is capable of generating useful data, we plan to develop algorithms based on AI/ML tools that will facilitate increased optimization capabilities addressing mobility management and resource allocation challenges. Furthermore, we plan on exploring the O-RAN paradigm in more depth, since it is placed at the center of the design of future mobile systems. Finally, security presents a key challenge. Due to the increase in complexity, 5G/B5G systems are vulnerable to several cyber threats and attacks. Hence, the identification of those threats and development of countermeasures will be one of our primary goals in the future. This work is aiming to also explore and expand our research activities reported in this thesis focusing on the untrusted access of non-3GPP networks.