UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Dynamic resource optimization and orchestration techniques for 5G new radio and beyond

Massimiliano Maule

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

PhD program in Signal Theory and Communications

Dynamic Resource Optimization and Orchestration Techniques for 5G New Radio and Beyond

Doctoral thesis by:

Massimiliano Maule

Thesis advisor:

PhD Supervisor: Dr. Christos Verikoukis

PhD Supervisor: Dr. John Vardakas

PhD Tutor: Dr. Ramon Antonio Ferrús Ferré

A thesis submitted in fulfillment for the degree of Doctor of Philosophy

at the

Department of Signal Theory and Communications

Universitat Politècnica de Catalunya

Barcelona, March 2023

Abstract

The mobile industry has started to deploy 5G in multiple diversified areas, creating an ecosystem that powers society's digital transformation. New vertical markets are growing, driven from the latest technological features in different fields such as Artificial Intelligence (AI), Internet of Things (IoT), cloud computing, big data, and edge computing. This new technology is expected to be faster than 4G, reaching 3.5 billion 5G subscriptions globally by 2026, accounting for around 40 percent of all mobile subscriptions at that time[1]. As a consequence, it is foreseen that new demands, such as more traffic volume, multiple devices with diverse service requirements, better Quality of user Experience (QoE) and better affordability by further reducing costs, will require an increasing number of innovative solutions.

Through the joint collaboration of standardization bodies, three main use-case families with distinct connectivity requirements have been defined in 5G: i) enhanced Mobile Broadband (eMBB) enables high data rates and human-centric scenarios like mobile data and media content, ii) massive Machine Type Communications (mMTC) provides connectivity to an high number of narrow-bandwidth and low cost devices such as sensors, trackers, and wearables and, iii) ultra-Reliable Low Latency Communications (uRLLC) focuses on reliable and strict latency solutions as real-time applications, autonomous vehicles, and advanced robotic. All together, these new 5G use cases require unprecedented network performance, pushing telecom operators and standardization bodies to evolve the current network architecture towards a novel enriched architecture, adding new sites, spectrum and system complexity. As promising solution, an evolution of the concept of Heterogeneous Network (HetNet) has been introduced with new paradigms for the interconnection of multi-access technologies, e.g., 5G New Radio (5G NR), 4G, WiFi, and fixed broadband networks have been defined. The existing Radio Access Network (RAN) solutions have been extended with the Cloud Radio Access Network (C-RAN) architecture, enforcing resource pooling, capacity scalability, layer interworking, and spectral efficiency. In the next-generation architecture, the C-RAN concept is extended to ultra-dense networks that incorporates macro, micro and small cells in diverse spectrum bands. From a functional point of view, the most logical approach is to build a set of dedicated networks each adapted to serve one type of business customer. However, this trend would introduce the design of tailor-made functionalities customized for each business customer, which might be challenging to realize at large scale due to physical infrastructure complexity.

Earmarked as the revolutionary 5G key enabler, Network Slicing (NS) is the embodiment of the concept of running multiple logical networks as virtually independent business operations on a common physical infrastructure in an efficient and economical way[2]. NS enables the sharing of 5G

[1] Ericsson, "Ericsson Mobility Report", June 2021.

[2] GSMA Association, "An Introduction to Network Slicing", White Paper, 2017.

RAN functional elements and the Transport Network (TN), securing the effective allocation and prioritization of the limited resources available, and offering to the customer improvement in service flexibility and scalability. By combining 5G and NS solutions, Service Providers (SPs) will pave the ways to cutting-edge industry fields such as Massive IoT, high resolution video on-demand services, smart manufacturing, and advanced solutions like Augmented Reality (AR), Virtual Reality (VR) and cloud gaming, gaining access to potential new revenue sources and smarter solutions to support end customers[3].

This dissertation provides a contribution to the design, analysis, and evaluation of novel multi-tenant RAN NS techniques for the dynamic orchestration of the radio resources in 5G NR. Our research aims to define in real-time the optimal resources' allocation per slice, able to overcome the resources' under-over provisioning issue, by exploiting the latest 5G NR features and flexibility principles of the 5G infrastructure. All the proposed mathematical models have been extensively validated using the 5G simulation platforms and\or real 4G-5G testbed configurations. The main thesis contributions are divided into two parts. In the first part, starting from a common baseline scenario, four different resource orchestration frameworks are presented, where the optimization problem is gradually refined through the exploitation of distinct mathematical models, 5G features, and metrics. In the second part, the design of a novel Open-RAN (O-RAN) Functional Split (FS) architecture for Software Defined Radio (SDR) has been proposed; this innovative concept represents a step forward to the creation of a shared RAN environment for the fair coexistence of multi-vendor solutions, as massively promoted in 5G.

During the elaboration of this thesis, two general key conclusions have been extracted. First, higher performance, system stability, and advanced service customization can be achieved using real-time NS in the RAN, outperforming the current state of the art limitations. This is possible through the new set of capabilities, sharing principle, and infrastructure scalability introduced in 5G NR. Second, we highlight the role of virtualization\softwarization of the network functionalities in 5G, illustrating the advantages of a shared infrastructure compared with a single vendor approach.

[3]Ericsson, "5G RAN Slicing: Capture new business revenues in the 5G era", Ericsson Business Review, October 2021.

Resumen

La industria móvil ha comenzado a implementar 5G en múltiples áreas diversificadas, creando un ecosistema que impulsa la transformación digital de la sociedad. Están creciendo nuevos mercados verticales, impulsados por las últimas novedades tecnológicas en diferentes campos como la Inteligencia Artificial (IA), Internet of Things (IoT), cloud computing, big data y edge computing. Se espera que esta nueva tecnología sea más rápida que 4G, alcanzando 3500 millones de suscripciones 5G en todo el mundo para 2026, lo que representa alrededor del 40% de todas las suscripciones móviles en ese momento[4]. Como consecuencia, se prevé que las nuevas demandas, como un mayor volumen de tráfico, múltiples dispositivos con diversos requisitos de servicio, una mejor calidad de la Quality of Experience (QoE) y una mayor asequibilidad mediante la reducción adicional de costos, requerirán un número cada vez mayor de soluciones innovadoras.

A través de la colaboración conjunta de los organismos de estandarización, se han definido tres familias principales de casos de uso con distintos requisitos de conectividad en 5G: i) la enhanced Mobile Broadband (eMBB) permite altas velocidades de datos y escenarios centrados en el ser humano, como datos móviles y contenido multimedia, ii) Las comunicaciones massive Machine Type Communications (mMTC) brindan conectividad a una gran cantidad de dispositivos de ancho de banda estrecho y de bajo costo, como sensores, rastreadores y dispositivos portátiles y, iii) las comunicaciones Ultra Reliable Low Latency Communications (uRLLC) se enfocan en soluciones de latencia estrictas y confiables como aplicaciones en tiempo real, vehículos autónomos y robótica avanzada. En conjunto, estos nuevos casos de uso de 5G requieren un rendimiento de red sin precedentes, lo que empuja a los operadores de telecomunicaciones y a los organismos de estandarización a evolucionar la arquitectura de red actual hacia una nueva arquitectura enriquecida, agregando nuevos sitios, espectro y complejidad del sistema. Como solución prometedora, se ha introducido una evolución del concepto de Heterogeneous Network (HetNet) con nuevos paradigmas para la interconexión de tecnologías de acceso múltiple, por ejemplo, 5G New Radio (5G NR), 4G, WiFi y redes de banda ancha fija. definido. Las soluciones de red de Radio Access Network (RAN) existentes se han ampliado con la arquitectura de la Cloud Radio Access Network (C-RAN), lo que hace cumplir la agrupación de recursos, la escalabilidad de la capacidad, el interfuncionamiento de capas y la eficiencia espectral. En la arquitectura de próxima generación, el concepto C-RAN se extiende a redes ultradensas que incorporan celdas macro, micro y pequeñas en diversas bandas de espectro. Desde un punto de vista funcional, el enfoque más lógico es construir un conjunto de redes dedicadas, cada una adaptada para servir a un tipo de cliente empresarial. Sin embargo, esta tendencia introduciría el diseño de funcionalidades personalizadas para cada cliente comercial, lo que podría ser difícil de realizar a gran escala debido a la complejidad de la infraestructura física.

[4]Ericsson, "Ericsson Mobility Report", junio de 2021.

Designado como el habilitador clave revolucionario de 5G, Network Slicing (NS) es la encarnación del concepto de ejecutar múltiples redes lógicas como operaciones comerciales virtualmente independientes en una infraestructura física común de una manera eficiente y económica[5]. NS permite compartir elementos funcionales de 5G RAN y la Transport Network (TN), asegurando la asignación y priorización efectivas de los recursos limitados disponibles y ofreciendo al cliente una mejora en la flexibilidad y escalabilidad del servicio. Al combinar las soluciones 5G y NS, los Service Provider (SP) allanarán el camino hacia campos industriales de vanguardia como Massive IoT, servicios de video a pedido de alta resolución, fabricación inteligente y soluciones avanzadas como Augmented Reality (AR), Virtual Reality (VR) y juegos en la nube, obteniendo acceso a posibles nuevas fuentes de ingresos y soluciones más inteligentes para ayudar a los clientes finales[6].

Esta disertación proporciona una contribución al diseño, análisis y evaluación de nuevas técnicas RAN NS multiinquilino para la orquestación dinámica de los recursos de radio en 5G NR. Nuestra investigación tiene como objetivo definir en tiempo real la asignación óptima de recursos por porción, capaz de superar el problema de aprovisionamiento insuficiente de los recursos, mediante la explotación de las últimas funciones 5G NR y los principios de flexibilidad de la infraestructura 5G. Todos los modelos matemáticos propuestos han sido ampliamente validados utilizando las plataformas de simulación 5G y\o configuraciones reales de banco de pruebas 4G-5G. Las principales contribuciones de tesis se dividen en dos partes. En la primera parte, a partir de un escenario de referencia común, se presentan cuatro marcos de orquestación de recursos diferentes, donde el problema de optimización se refina gradualmente mediante la explotación de distintos modelos matemáticos, características 5G y métricas. En la segunda parte, se ha propuesto el diseño de una novedosa arquitectura Open-RAN (O-RAN) Functional Split (FS) para Software Defined Radio (SDR); este concepto innovador representa un paso adelante en la creación de un entorno RAN compartido para la coexistencia justa de soluciones de múltiples proveedores, como se promueve masivamente en 5G.

Durante la elaboración de esta tesis se han extraído dos conclusiones generales clave. En primer lugar, se puede lograr un mayor rendimiento, estabilidad del sistema y personalización avanzada del servicio utilizando NS en tiempo real en la RAN, superando las limitaciones actuales del estado de la técnica. Esto es posible a través del nuevo conjunto de capacidades, el principio de uso compartido y la escalabilidad de la infraestructura introducidos en 5G NR. En segundo lugar, destacamos el papel de la virtualización/software de barra invertida de texto de las funcionalidades de red en 5G, lo que ilustra las ventajas de una infraestructura compartida en comparación con el enfoque de un solo proveedor.

[5]GSMA Association, "An Introduction to Segmentación de red", Informe técnico, 2017.

[6]Ericsson, "5G RAN Slicing: Capture nuevos ingresos comerciales en la era 5G", Ericsson Business Review, octubre de 2021.

Keywords: 5G NR, Radio Resource Orchestration, Functional Split, SDR, Network Slicing, Game Theory, O-RAN, Testbed, RAN, Network Optimization.

Acknowledgement

Four European countries, four universities, three job positions in three different companies. This is the summary of my path so far, this is what you can read in my CV or Linkedin profile. But those who have had the opportunity to know me more closely know that this is only a small part of what I personally consider essential in my life and career.

The passing of my mother nine years ago was the spark of a profound personal change, from the way I see things, to what values are important in life. I was forced to leave my comfort zone, and react to this unexpected situation. Facing new academic and work challenges was the answer to my insecurity in feeling lost, and only now that I am at the end of my PhD I understand how this experience has contributed to my inner peace, to believe in myself, and to have a different perspective on the problems and challenges of the future, personal and work.

I would be a fool to think that it is all my doing, because the values I have mentioned are not taught in any manual, lecture, or book. The people and the relationships are the foundation of who we are and of our actions. For this reason, I would like to infinitely thank my supervisors Dr. Christos Verikoukis and Dr. John Vardakas, for guiding me throughout this journey, for believing in me, and for the respect you have always shown me. Your teachings will always be part of my personal background, and I will always value them. I would like to thank Dr. Walter Nitzold and Clemens Felber, not only for being my supervisors during my secondment period, but also for being outstanding mentors. You made me love even more the world of research, thanks to your ability to appreciate my skills and ideas. I would also like to thank all of Iquadrat RD, especially Dr. Kostas Ramantas and Melani Gurdiel, for the trust you have always placed in me. I thank Dr. Ferrus who reviewed the thesis before the final deposition.

Undoubtedly, I would like to express my gratitude to the people who have also supported me in all the other aspects of my life. I am immensely grateful to my father Carlo, my sister Chiara and my grandparents Giuseppe and Mariarosa. I can't imagine how difficult it was for you to indulge my dreams away from home. I am grateful that you have been there for me every moment, and I know you will always be there. Thank you for your advices, wisdom, and trust in me along this path. I keep learning from your teachings and unconditional love.

Finally, I would like to dedicate this thesis to the person I want by my side in the remaining steps of my life. To you Maria Alejandra and your family, who have always supported me and made me see the light even in the most difficult moments. You are always by my side, with the correct word, making me proud to be with you. You taught me to live life fully, to always find a solution, and to never lose the faith.

Massimiliano Maule

Ringraziamenti

Quattro paesi Europei, quattro università, tre posizioni lavorative in tre aziende diverse. Questo è il riassunto del mio percorso fino ad ora, questo è quello che si può leggere nel mio CV o profilo Linkedin. Ma chi ha avuto la possibilità di conoscermi più da vicino, sa che tutto questo è solo una minima parte di quello che personalmente reputo fondamentale nella mia vita e carriera.

La scomparsa di mia madre nove anni fa è stata la scintilla di un profondo cambiamento personale, dal modo in cui vedo le cose, fino a capire quali sono valori importanti nella vita. Sono stato obbligato a lasciare la mia confort zone, e reagire a questa inattesa situazione. Affrontare nuove sfide accademiche e lavorative è stata la risposta alla mia insicurezza nel realizzarmi e sensazione di sentirmi perso, e solo ora che mi trovo alla fine del mio dottorato capisco come questa esperienza abbia contribuito alla mia pace interiore, credere in me stesso, ed avere una prospettiva differente ai problemi e le sfide del futuro, personali e lavorativi.

Sarei uno sciocco nel pensare che è tutto merito mio, perché i valori che ho menzionato non vengono insegnati in nessun manuale, conferenza, o libro. Le persone ed i legami che si creano sono le fondamenta di chi siamo e delle nostre azioni.

Per questo voglio ringraziare infinitamente i miei supervisori, il Dr. Christos Verikoukis e il Dr. John Vardakas, per avermi guidato in tutto questo percorso, per aver creduto in me, e per il rispetto che mi hanno sempre dimostrato. I vostri insegnamenti faranno sempre parte del mio bagaglio personale, e sempre gli valorizzerò. Vorrei ringraziare il Dr. Walter Nitzold e Clemens Felber, per essere stati non solo i miei supervisori durante il mio periodo di secondment, ma anche degli accezionali mentors. Mi avete fatto amare ancora di più il mondo della ricerca, grazie alla vostra capacità di esaltare le mie competenze e idee. Vorrei anche ringraziare tutta Iquadrat RD, in particolare Dr. Kostas Ramantas e Melani Gurdiel, per la fiducia che sempre mi avete dato. Ringrazio Dr. Ferrus che ha esaminato la tesi prima della finale deposizione.

Indubbiamente, vorrei esprimere la mia gratitudine alle persone che mi hanno sempre sostenuto anche in tutti gli altri aspetti della mia vita. Sono immensamente grato a mio padre Carlo, a mia sorella Chiara e ai miei nonni Giuseppe e Mariarosa. Non riesco ad immaginare quanto sia stato difficile per voi assecondarmi per seguire i miei sogni lontano da casa. Vi sono grato che ci siete stati per me in ogni momento, e sò che sempre ci sarete. Grazie per le vostre parole, la vostra saggezza, e fiducia in me durante tutto questo percorso. Non smetterò mai di imparare dalle vostra esperienza ed amore incondizionato.

Infine, vorrei dedicare questa tesi alla persona che voglio al mio lato nei rimanenti passi della mia vita. A te Maria Alejandra e alla tua famiglia, che sempre mi avete sostenuto e fatto vedere la luce anche nei momenti più difficili. Sempre sei al mio fianco, con la parola corretta, rendendomi orgoglioso di stare con te. Mi hai insegnato a vivere pienamente la vita, a trovare sempre una

soluzione, e che non bisogna mai perdere la fede.

Massimiliano Maule

List of Publications

- Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., Verikoukis, C., "Real-time dynamic network slicing for the 5G radio access network", GLOBECOM IEEE Global Communications Conference, 2019.
- Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., Verikoukis, C., "Dynamic partitioning of radio resources based on 5G RAN Slicing", GLOBECOM IEEE Global Communications Conference, 2020.
- Maule, M., Vardakas, J., Verikoukis, C., "Multi-service Single Tenant 5G Fronthaul Resource Orchestration Framework based on Network Slicing", GLOBECOM IEEE Global Communications Conference, 2022.
- Maule, M., Vardakas, J., Verikoukis, C., "5G RAN Slicing: Dynamic Single Tenant Radio Resource Orchestration for eMBB Traffic within a Multi-Slice Scenario", IEEE Communications Magazine, 2021.
- Maule, M., Vardakas, J. S., Verikoukis, C., "A Novel 5G-NR Resources Partitioning Framework Through Real-Time User-Provider Traffic Demand Analysis", IEEE Systems Journal, 2021.
- Maule, M., Kohjogh, O., Rezazadeh, F., "Advanced Cloud-Based Network Management for 5G C-RAN", in Enabling 6G Mobile Networks (pp. 371-397). Springer, Cham, 2022.
- Maule, M., Vardakas, J. S., Verikoukis, C., "Multi-service network slicing 5G NR orchestration via tailored HARQ scheme design and hierarchical resource scheduling", IEEE Transactions on Vehicular Technology, 2022.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols and abbreviation that will be later used within the body of the document

$3GPP$	3rd Generation Partnership Project
$5GPPP$	5G Public-Private Partnership
$5G-EIR$	5G-Equipment Identity Register
$5G\,NSA$	5G Non-Standalone
$5G\,SA$	5G Standalone
$5G\,NR$	5G New Radio
ACK	Acknowledge
AD	Analog Device
AF	Application Function
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AN	Access Network
API	Application Programming Interfaces
APN	Access Point Name
AR	Augmented Reality
AS	Access Stratum
$AUSF$	Authentication Server Function
$B2B$	Business to Business

BBU	Baseband Processing Unit
BER	Bit Error Rate
BWP	Bandwidth Parts
$C-RAN$	Cloud RAN
$CAGR$	Compound Annual Growth Rate
$CAPEX$	CAPital EXpenditure
CES	Constant Elasticity of Substitution
$CHDR$	Condensed Hierarchical Datagram for RFNoC
$CoMP$	Coordinated Multi-Point
CP	Control Plane
CQI	Channel Quality Indicator
CRF	Central Repository Function
CSA	Communication Service Availability
CSF	Channel State Feedback
CSI	Channel-State Information
$CSIF$	Communication Service Interface
CSP	Communications Service Provider
CSR	Communication Service Reliability
CTI	Cooperative Transport Interface
CU	Centralized Unit
$D-DU$	O-RAN Distributed Unit
$D-RAN$	Decentralized RAN
DB	Database
DCI	Downlink Control Information
DL	Downlink

DMA	Direct Memory Access
DN	Data Network
DNF	Data NF
DP	Data Plane
DU	Distributed Unit
$E-UTRAN$	Evolved-Universal Terrestrial Radio Access Network
$eCPRI$	enhanced Common Public Radio Interface
$eCPRI$	enhanced Common Public Radio Interface
$eMBB$	Enhanced Mobile Broadband
EPC	Evolved Packet Core
$eURLLC$	Enhanced Ultra-Reliable Low-Latency Communication
$FBMC$	Filter Bank Multicarrier
$FCFS$	First Come First Served
FDD	Frequency-Division Duplexing
FP	Functional Split
$FPGA$	Field-Programmable Gate Array
$FRAND$	Fair, Reasonable and Non-Discriminatory
$FTTH$	Fiber to the Home
$FTTP$	Fiber to the Premises
FWA	Fixed Wireless Access
GBR	Guaranteed Bit Rate
$GERAN$	GSM EDGE Radio Access Network
GSM	Global System for Mobile Communication
$HARQ$	Hybrid Automatic Repeated Request
$HPMN$	Home Public Mobile Network

HR	Home-Routed
HW	Hardware
IAB	Integrated Access and Backhaul
$ICMP$	Internet Control Message Protocol
$IMEI$	International Mobile station Equipment Identity
IMS	IP Multimedia Subsystem
$IMSI$	International Mobile Subscriber Identity
IoT	Internet of Things
$ISAC$	Integrated Sensing and Communication
ISI	Inter-Symbol Interference
$ITU-R$	International Telecommunication Union – Radio Sector
$L2$	Layer 2
$L3$	Layer 3
LB	Load Balancer
LBO	Local-Break Out
LLS	Lower-Layer Split
LoS	Line-of-Sight
$LPWA$	Low Power Wide Area
LTE	Long Term Evolution
$M2M$	Machine-to-Machine
MAC	Medium Access Control
MCC	Mission-Critical Communication
$MCDM$	Multi Criteria Decision Making
MCS	Modulation Coding System
MEC	Mobile Edge Computing

$MILP$	Mixed Integer Linear Programming
$MIMO$	Massive Multiple-Input and Multiple-Output
$mMTC$	Massive Machine-Type Communication
MN	Master Node
MP	Management Plane
$MR-DC$	Multi-Radio Dual Connectivity
$multi-TRP$	Multiple Transmission Points
$NaaS$	NS as a Service
$NaaS$	Network as a Service
NAS	Non-Access Stratum
NEF	Network Exposure Function
$NESP$	Nash Equilibrium Strategy Profile
$ng-gNB$	New Generation - gNB
$NGMN$	Next Generation Mobile Networks
NIM	Network Infrastructure Manufacturers
NLP	Natural Language Processing
$NOMA$	Non-orthogonal Multiple Access
$non-GBR$	non-Guaranteed Bit Rate
$non-RT$	non-Real-Time
NRA	National Regulatory Authority
NRF	NF Repository Function
$nrUE$	New Radio UE
NS	Network Slicing
$NSC-VNF$	Non-SDN-Enabled Virtual NF
$NSSAI$	Network Slice Selection Assistance Information

$NSSF$	Network Slice Selection Function
$NSSMF$	Network Slice Subnet Management Function
$O-CU$	O-RAN Centralized Unit
$O-RAN$	Open-RAN
$O-RU$	O-RAN Radio Unit
$OFDM$	Orthogonal Frequency Division Multiplexing
OP	Operator Key
$OP-ID$	Operator-ID
$OPEX$	OPerating EXpense
OS	Operating System
OSA	OAI Software Alliance
PCC	Policy and Charging Control
PCF	Policy Control Function
$PDCCH$	Physical Downlink Control Channel
$PDSCH$	Physical Downlink Shared Channel
PER	Packet Error Rate
PHY	Physical Layer
$PLMN$	Public Land Mobile Network
$PLMNID$	Public Land Mobile Network-ID
PNE	Physical Network Elements
PNF	Physical Network Functions
PRB	Physical Resource Block
PTP	Precision Time Protocol
$PUCCH$	Physical Uplink Control Channel
$PUSCH$	Physical Uplink Shared Channel

QoS	Quality of Service
RAC	Radio Admission Control
RB	Resource Block
RF	Radio Frequency
$RFIC$	Radio-Frequency Integrated Circuit
$RFNoC$	RF Network-on-Chip
RIC	Radio Intelligent Controller
RL	Reinforcement Learning
RoE	Radio over Ethernet
ROI	Return on Investment
RRC	Radio Resource Control
RRM	Radio Resource Management
RRS	Reconfigurable Radio System
RRU	Remote Radio Unit
RSI	RAN Slice Instance
RSO	RAN Slicing Orchestrator
RTC	Real-time Controller
RTT	Round-Trip Time
RU	Remote Unit
$S-NSSAI$	Single Network Slice Selection Assistance Information
$S-Plane$	Synchronization Plane
SBA	Service-Based Architecture
$SC-FDMA$	Carrier Frequency Division Multiple Access
SCS	Subcarrier Spacing
SD	Slice Differentiator

SDK	Software Development Kit
SDN	Software Defined Networking
SDR	Software-Defined Radio
$SEPP$	Security Edge Protection Proxy
$SFP+$	Small Form-factor Pluggable +
SM	Slice Manager
SME	Small and Medium-sized Enterprise
SMF	Session Management Function
SMO	Service Manager and Orchestrator
SN	Slave Node
SOA	Socially Optimal Allocations
SoC	System-on-a-Chip
SP	Service Providers
SP	Synchronization Plane
SW	Software
$SyncE$	Synchronous Ethernet
TA	Transport Adapter
TCO	Total Cost of Ownership
TDD	Time-Division Duplexing
THz	Terahertz
TN	Transport Network
TTI	Transmission Time Interval
UDM	Unified Data Management
UDP	User Datagram Protocol
UDR	Unified Data Repository

$UDSF$	Unstructured Data Storage Function
UE	User Equipment
UHD	USRP Hardware Driver
UL	Uplink
UPF	User Plane Function
$uRLLC$	Ultra Reliable Low Latency Communications
$USRP$	Universal Peripheral Radio Software
$USRP$	Universal Software Radio Peripheral
$UTRAN$	Universal Terrestrial Radio Access Network
$V2X$	Vehicle-to-Everything
VNF	Virtual Network Function
$VPMN$	Visited Public Mobile Network
VR	Virtual Reality
WG	Working Group
$WTTx$	Fixed Broadband Fibre-like Connectivity
WUS	Wakeup Signal
$XaaS$	Anything as a Service

Chapter 1

Introduction

1.1 Motivation

The Fifth-Generation (5G) mobile network targets novel use-cases and business models expected to globally convert the role of telecommunications technology in the society. The definition of a service-oriented architecture combined with enhanced computing power dislocated in the network defines an ecosystem involving vertical markets such as automotive, energy, food and agriculture, city management, government, healthcare, manufacturing, public transportation, and many more. This solution will serve a larger portfolio of applications with a corresponding multiplicity of requirements ranging from high reliability to ultra-low latency going through high bandwidth and mobility (1). As consequence of this service-oriented vision, the SPs share infrastructures to deliver mobile services to end users, following two modalities: passive sharing consists of sharing network infrastructure such as masts, sites, cabinet, power, cooling, and active sharing for the sharing of RAN elements such as antennas and controllers. According with the National Regulatory Authority (NRA), the infrastructure sharing is translated in cost saving for the operators (2), paving the way to the development of new services and applications in the history of mobile and wireless communications. As early 5G investigation phase, different organizations have been formed to establish research requirements and set the path for the definition of the next generation of mobile solutions. In particular, 3rd Generation Partnership Project (3GPP), during the fourth quarter of 2015, approved a 5G study titled "Study on Architecture for Next Generation Systems", introducing the concept of NS as a promising future proof framework adhering to the technological and business needs of different industries (3). NS enables business customers to deploy connectivity and data processing following ad-hoc requirements defined through the operator Service Level Agreement (SLA) (4). Each operator orchestrates its slices as an end-to-end logical network on top of a shared physical infrastructure, spanning over different technology domains (e.g., core, transport and access networks) and administrative domains (e.g., different mobile network operators), while this

technology is transparent to the business customer.

In the field of NS, multiple approaches have been investigated to define the optimal configuration and management of a slice. In (5), authors developed an adaptable grouping edge cloud architecture to authorize 5G optical mobile FH NS. To support several Quality of Service (QoS) flow and resource orchestration schemes per slice, a united network resource management technique was developed. Final outcomes shown that the proposed solution was capable of cooperatively allocate the radio resources to network slices and perceived cloud-computing offloading to reach several QoS necessities and consequently, inline with the FH bandwidth constraint. In (6), authors proposed a heuristic 5G core NS allocation algorithm, based on a Multi Criteria Decision Making (MCDM) technique. In the initial phase, the algorithm categorizes the node significance with the MCDM method through the evaluation of the topological features and node capacity. Gradually, the slice node with the ranking value is stored and, in the case of the slice link provisioning step, the shortest path algorithm is executed to accomplish the candidate physical tracks for the slice link. Different tests revealed that this algorithm maximize slice acceptance ratio and the highest provisioning revenue to cost factor, matching the security constraints of 5G Core Network (CN) slice request.

Recent works on NS primarily focus on the analysis of the CN capabilities in order to provide dedicated virtual networks (7)-(8). However, from the RAN perspective, NS presents extremely diverse services and use cases requirements, with different service configurations in terms of slice resources. These disparities could be in either baseband (e.g., frame structure, Subcarrier Spacing (SCS), etc.) and/or in Radio Frequency (RF) front-end (e.g., processing bandwidth or sampling rate, etc.). In RAN slicing, specific Physical (PHY) Layer and Medium Access Control (MAC) Layer configurations can be tailored for slice policies, since 5G NR PHY comes with new features such as the support of Bandwidth Parts (BWPs) that could facilitate the implementation of RAN Slice Instance (RSI) with diverse PHY requirements. The BWP feature in 5G NR allows for UEs to operate only in a portion of the channel, reducing UE processing complexity and power consumption (9). In (10), authors present a functional framework for the NS management for a New Generation-Radio Access Network (NG-RAN) infrastructure, identifying the necessary information models and interfaces to support the dynamic provisioning of RAN slices. This work highlights how the type of user traffic affects the resource provisioning in the radio part of the network, while only minor changes are applied in the CN slice structure. In (11), the 5GDRIVE research project focuses on the integration of virtualized network services running at the edge, able to facilitate vehicular end-users and their respective systems. For the RAN section, the project proposes a slice orchestrator able to minimize the Round-Trip Time (RTT) through the manipulation of the Time Division Duplex (TDD) slot to Downlink-Uplink (DL-UL) pattern, and reduction of the scheduling request period. Although the proposed method improves the performance of the RAN domain, 5G

frame format features are not fully used, limiting the final performance.

Especially for low latency traffic, timing is a crucial feature required to guarantee a seamless and secure service. To better support service with short latency requirement, the 3GPP proposes in the time domain scalable Transmission Time Intervals (TTIs) where the 5G NR flexible frame structure scales up and down depending on specific service requirements (12). This flexibility allows the air interface to reduce latency using shorter TTI (e.g., hundreds of microseconds) or trade-off for higher spectral efficiency in case of delay-tolerant use cases with longer TTI. In (13), the authors first analyze the uRLLC service requirements for 5G C-RAN architecture, and then a novel resource allocation scheme using flexible frame structures and SCS is illustrated. This work illustrates how in the frequency domain,a higher SCS shortens the symbol and TTI duration, improving the performance of low latency services. The combination of numerology/SCS and TTI determines the amount of bits and manner they are transmitted over the air interface. Authors in (14) show how different numerologies of5G can be used to maximize the energy efficiency for eMBB and uRLLC services, exploiting both frequency and time dimensions. Similar work is illustrated in (15) where authors explore the Physical Resource Block (PRB) allocation in a Non-orthogonal Multiple Access (NOMA)-based mixed numerology system using PRB location reutilization, resource fragmentation, and service QoS.

Multiple types of data of different nature are required to be examined, and processed to perform life cycle orchestration of a network slice in order to fulfill the QoS necessities of the service deliberated to be distributed through it, in spite of the network conditions and time varying workloads. In this complex scenario, Machine Learning (ML) and AI have been introduced to predict future trends and improve the analytics behind NS resource management (16). As novel approach, the use of Reinforcement Learning (RL) methods for NS control and management has gained interest due to their promising performance. In (17), a RL RAN slicing admission control system is presented, where the system learns the services with the potential to bring high profit (i.e., high revenue with low degradation penalty), and hence to be accepted. In (18), the authors designed a slice admission block based on traffic prediction, where the forecasted load is adjusted based on measured deviations. In (19)-(20), authors present a slice admission strategy based on RL in the presence of services with different priorities. The use case considered is a 5G flexible RAN, where slices of different mobile SPs are virtualized over the same RAN infrastructure.

1.2 Scope

In this dissertation, novel dynamic RAN NS solutions are presented, where access network resources are dynamically customized according to specific slice policies, service characteristics, infrastructure/architecture sharing principles, and real-time evolution of the network environment. A series of NS resource orchestration mathematical models are illustrated, designed following distinct optimization techniques, and the same objective: the real-time optimal coexistence of a multi-tenant slice-centric RAN environment, where the slice resources are tailored following service specific properties, features, distinct granularity degrees, and the exploitation of 5G NR enablers (i.e., flexible frame structure, resource disaggregation, etc.). Apart from their inherent benefits, our methodology overcomes the main RAN resource management challenges of the current NS Fronthaul (FH) architectures, listed as follow:

- Distributed management of the RAN resources: one-size-fits-all orchestrators aim at orchestrating the entire End-to-End (E2E) networks, providing a central point of management for the entire E2E infrastructure. However, these solutions are difficult to realize, since: i) they tend to oversimplify each network domain with a limited set of technologies and standards, and ii) they focus principally on the process of instantiating and deploying the network slice, while ignoring how they are enforced in the mobile network.

- On-demand real-time service design: due to the stochastic behavior of networks, static allocation of network resources does not represent an efficient approach. In particular, when applied to wireless resources, due to burst of traffic, user mobility, and time-varying channel, only through resource overprovisioning techniques we are able to deal with multiple services, without exploiting the flexibility of short-medium term fluctuations in terms of resource requirements. For this reason, dynamic NS represent the leading approach for the future networks. With this technique, network resources are dynamically assigned to meet tailored performance requirements (e.g. capacity, latency, priority, security) through a seamless and virtually continuous network propagated across the 5G networks architecture.

- Disaggregated infrastructure: while the previous standards were characterized by monolithic network infrastructures of inelastic elements, software, and functionalities, the 5G architecture and NS introduce a novel resource abstraction technique able to overcome the previous legacy systems restrictions. The decoupling between the virtualised and the physical infrastructure enables scaling and flexibility of the slices, defining an environment where the resources are adapted on demand, introducing new NS challenges in terms of data isolation among multi-tenancy solutions, management of different E2E QoS within a slice, NFs optimisation for automatic selection of network resources and functions, monitoring the NS behaviour

in a multi-domain scenario, and capability exposure for NS API for slice configuration and interaction.

1.3 Structure of the Dissertation and main Contributions

The remainder of the thesis consists of six chapters, where the contents and the contributions of each chapter are described in detail as follows:

- **Chapter 2**. The second chapter provides the technical background for the main concepts treated in this work. First, the principal motivations driving the investigation and deployment of this new "G" technology are highlighted, followed by the vision and contribution of the main standardization bodies, industry organizations, and agencies towards the composition of the 5G technical aspects and requirements. Without the cross combination of different types of technologies, 5G would not be able to realize seamless connectivity on top of a heterogeneous infrastructure. For this reason, the main features and enablers are illustrated, with special attention to the innovative features in 5G NR. Finally, the new 5G architecture is illustrated, together with the standardized deployment options.

- **Chapter 3**. This chapter contains the details of our experimental testbeds, and an illustration their components. Two platforms have been developed: a legacy 4G E2E system equipped with 5G features, and the first open source 5G Standalone (5G SA) system. Even though this chapter does not represent the main part of this thesis, the design, configuration and testing of the aforementioned platforms represented an extremely complex task, which contributed to strengthen the performance analysis of the proposed RAN NS solutions. The contributions of this chapter have been published/submitted in part of the following publications and workshops:

  - Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., Verikoukis, C., "Real-time dynamic network slicing for the 5G radio access network", GLOBECOM IEEE Global Communications Conference, 2019.
  - Maule. M., "NI-OpenAirInterface 5G testbed and O-RAN USRP Functional Split design proposal", NI secondment, Dresden, (Germany), 2021.
  - Maule, M., Ramantas, K., "5G ERA testbed", first year review meeting of the 5G-ERA project, January 2022.

- **Chapter 4**. The fourth chapter illustrates our proposed innovative solutions based on 5G NS. The main achievements are subdivided into four subsections, with increasing complexity in terms of new features, algorithmic structure, and optimization technique. Each method

represents a feasible implementation of a real-time NS orchestrator equipped with multi purpose capabilities in terms of platform scalability and flexibility, such that specific 5G services can be precisely customized and seamless performed. Most of the solutions are evaluated through a two level process, first using a simulated environment (written in Matlab), and secondly through our experimental testbeds. Moreover, supplementary material is reported with two appendixes, as it will be recalled in the corresponding subsections inside this chapter. The contributions of this chapter have been published/submitted in part of the following publications and conferences:

  - Maule. M., "Joint XHaul Network Resources Allocation", 5GSTEP-FWD Session I, IEEE CAMAD, Barcelona, (Spain), 2018.
  - Maule, M., "Real-time Dynamic Network Slicing for the 5G Radio Access Network", Photonic Technologies in 5G and Beyond workshop, IEEE EuCNC, Valencia, (Spain), 2019.
  - Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., Verikoukis, C., "Dynamic partitioning of radio resources based on 5G RAN Slicing", GLOBECOM IEEE Global Communications Conference, 2020.
  - Maule, M., Vardakas, J., Verikoukis, C., "5G RAN Slicing: Dynamic Single Tenant Radio Resource Orchestration for eMBB Traffic within a Multi-Slice Scenario", IEEE Communications Magazine, 2021.
  - Maule, M., Vardakas, J. S., Verikoukis, C., "A Novel 5G-NR Resources Partitioning Framework Through Real-Time User-Provider Traffic Demand Analysis", IEEE Systems Journal, 2021.
  - Maule, M., Vardakas, J. S., Verikoukis, C., "Multi-service network slicing 5G NR orchestration via tailored HARQ scheme design and hierarchical resource scheduling", IEEE Transactions on Vehicular Technology, 2022.

- **Chapter 5**. 5G has triggered a new wave of interest in network sharing, introducing new set of challenges that telecom operators need to resolve in their overall network-sharing strategies. For this reason, in this chapter an Open-RAN (O-RAN) FH interface for NI Universal Software Radio Peripheral (USRP) Software Defined Radio (SDR) is designed. After an initial background subsection, two different O-RAN integration approaches are presented, with a deep analysis of benefits and weak points of each architecture. The contributions of this chapter have been published/submitted in part of the following workshops and proposal:

  - Maule. M., "NID Thinktank: Functional Split", NI, Dresden, (Germany), 2021.

- **Chapter 6**. To conclude, this chapter collects the main results of this thesis, and provides some hints to future extension of this work towards the incoming 6G technology.

Chapter 2

5G Contextualization and Background

2.1 Ecosystem and Drivers

5G represents a game changer technology with the potential to create significant social and economic impact. Unlike the previous "G" eras, it transforms the role of mobile technology in the society, since the demand for continuous connectivity grows, together with the opportunity to create a dynamic service-tailored network to the different needs of citizens and the economy. Between 2021 to 2028, the global telecom services market is expected to growth of 5.4% on the Compound Annual Growth Rate (CAGR), primarily driven by the increasing demand of mobile data services[1] (Fig. 2.1).

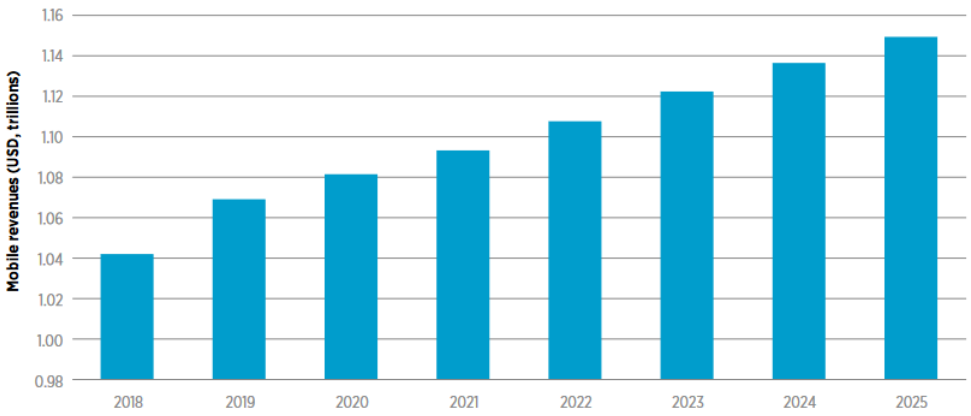


Figure 2.1: Mobile revenue forecast

According to Ericsson[2], the 5G consumer market could be worth USD 31 trillion by 2030 in the ICT industry, with up to USD 3.7 trillion earned by Communications Service Providers (CSPs) in cumulative 5G-enabled consumer revenues. To reach this goal, IDC research house[3] estimated that telecom companies will invest around USD 57 billion on the rollout of 5G by 2022, divided among

[1]Grand View Research Market Analysis Report April 2021.
[2]Ericsson Press Release November 17, 2020.
[3]IDC, "Market Analysis Perspective: Worldwide Carrier Network Infrastructure" Sept. 2018.

eight pillar industries, as illustrated in Fig. 2.2.

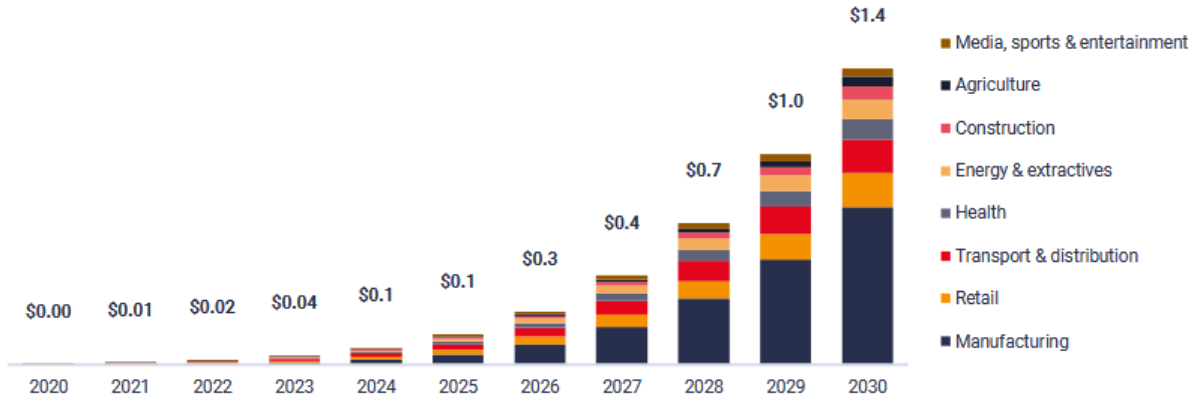


Figure 2.2: Extra yearly Gross Domestic Product (GDP) in USD trillions introduced with 5G

Unlike the launch of 4G where operators'revenues have been primarily consumer-driven, the wide 5G solutions'portfolio introduces new challenges for operators in terms of Return on Investment (ROI) strategies. For this reason, new monetization models are defined for 5G networks (21):

1. *Wireless fiber deployment*: it provides fiber like speed by 4G/5G Fixed Wireless Access (FWA) technology. Introduced as alternative of Fiber to the Premises (FTTP) and Fiber to the Home (FTTH), it reuses legacy mobile network infrastructure, does not need access permission and civil construction, and can easily provide inclusive fixed broadband with a low Total Cost of Ownership (TCO) (22). From a mobile operator perspective, this technology activates new business opportunities against fixed broadband, providing fast services with reduced costs.

2. *Enhanced consumer and vertical services*: three new macro classes of 5G use cases allow CSPs to deal with application-specific and partner-specific requirements, enabling multiple pricing models, and foster complex partner ecosystems:

    - *eMBB*: it delivers ultra-high wireless bandwidth capabilities, and can be used to provide fast, high-bandwidth consumer and business broadband service. It will help to develop scenarios such as emerging AR/VR media and applications, UltraHD, or 8K video streaming.

    - *uRLLC*: it applications in this category are characterized by high reliability and ultra-low latency, which enable consistent responses in real or near real time. Since uRLLC requires a 5G SA core, it will have a greater impact on operators' ROI than eMBB, which might operate over a 5G Non-Standalone (5G NSA) environment, meaning it can fall back onto a 4G core.

    - *mMTC*: it enables industry players to connect massive numbers of devices with specific

connectivity requirements, in sectors such as manufacturing, IoT, utilities, and logistics, where high scalability and low power consumption are required.

3. *Private wireless solutions*: better connectivity is not the only feature required for new businesses. Through the cross implementation of the aforementioned use cases, competitive 4.0 industry solutions are customized for specific scenarios. A well-known cases are harbours, warehouses and factories where there is either a lot of machinery causing radio interference, or a lot of indoor/outdoor devices connected over a wide area. This paves the way to specialized revenue plans, bringing the CSPs closer to their enterprise customer needs.

4. *E2E NS*: 5G is expected to open lucrative new business opportunities for mobile operators and other newer entrants. This key technology enabler permits to multiple vertical industries the execution of their solutions on top of a shared infrastructure, where the Service Providers (SP) customize the network capabilities (security, connection, processing power, storage, etc.) (23). With the definition of a network slice, service requirements and prices can be better matched with customer demand, without affecting other third-party configurations.

While many 5G concepts are still under investigation, it is clear that a novel service infrastructure design embracing all the network domains is essential to cope with the new network parameters, as latency, throughput, reliability, scaling, etc. In an initial phase, a soft technological transition will guide operators to include the new 5G capabilities at the speed that matches their own business models, while simultaneously preserving and enhancing existing 4G capabilities by reducing risk and making best use of current infrastructure.

2.2 Standardization Bodies and Vision

2.2.1 3GPP 5G specifications

The 3GPP has finalized the initial set of 5G specifications in 3Q 2019 with Release 15[4]. It represents the first full set of 5G standards, after the completion of the 5G NSA specification in late 2017, with a clear path towards a well-defined 5G SA system. In addition to the enhancement of 4G radio system and the Evolved Packet Core (EPC) capabilities, this release extends the 5G support to new scenarios, upgraded performance and management models, enabling vendors to progress rapidly with chip design and initial network implementation (24). The extension of the 5G NSA specifications with the 5G SA requirements turns on the first independent installation of 5G NR solutions, brand new E2E network architecture, accelerating the smart information and communications technology improvement process of enterprise customers and vertical industries.

[4]https://www.3gpp.org/release-15

To expand 5G with new services, spectrum, and deployments, 5G NR Release 16[5] is considered the major milestone for the entire mobile and broader vertical ecosystem, as this new set of 5G specifications unlocks many new 5G opportunities beyond the classic broadband services. It introduces few important benefits across a broad range of 5G use cases (25):

- *Massive Multiple-Input and Multiple-Output (MIMO)* enhancements: in order to provide higher throughput, reduced overhead, and additional robustness, enhanced beam handling and Channel-State Information (CSI) feedback have been introduced, as well as support for transmission to a single User Equipment (UE) from Multiple Transmission Points (multi-TRP) and full-power transmission from multiple UE antennas in the UL.

- *Enhanced Ultra-Reliable Low-Latency Communication (eURLLC)*: one key technology to overcome this system challenge is Coordinated Multi-Point (CoMP), which provides a valid solution to enhanced throughput and coverage performance by reducing the interference, especially for cell-edge users. To cope with high-reliability and extremely low latency simultaneously, Hybrid Automatic Repeated Request (HARQ) retransmission mechanism has been restructured. Unlike Long Term Evolution (LTE), timing between data transmission and HARQ response is flexibly set in NR, with novel feedback mechanism and enhanced Channel State Feedback (CSF).

- *Integrated Access and Backhaul (IAB)*[6]: introduced to overcome the inherent limitations of 5G NR operating in the mid-band and above, IAB leverages the spectral efficiencies of new radio and the increased capacity afforded by the higher bands available in 5G to deliver an alternative to optical cell site Backhaul (BH). IAB can open doors to a more flexible densification strategy, supporting more efficient mmWave densification costs, allowing operators to quickly add new base stations dynamically, before having to install additional fibers to increase BH capacity.

- *New power-saving features*: several power reduction features have been introduced: i) a new Wakeup Signal (WUS) alerts UE devices when transmissions are incoming, while keeping them in low-power mode when inactive, ii) a cross-slot scheduling to aware the UE with the minimal time interval between DL transmissions, such that unnecessary RF operations are avoided, iii) a dynamic UE MIMO layer reduction for advanced multiple-input/multiple-output configurations for data transmission and reception, and iv) a reduced radio resource measurements for less power consumption when managing low level radio parameters.

[5]https://www.3gpp.org/release-16.
[6]3GPP TS 38.401 "5G NG-RAN Architecture description".

During December 2020, 3GPP Release 17[7] timeline was defined. As main objective, new mobile communication use cases have been investigated such as 5G Low Power Wide Area (LPWA) IoT communication based on the LTE massive MTC (mMTC) technologies, IAB enhancements, MIMO improvements from real life deployments learning, 5G NR operation on high frequencies, and improved Broadcast/Multicast solutions.

2.2.2 International Telecommunication Union – Radio Sector (ITU-R)

The International Telecommunication Union – Radio Sector (ITU-R) TU allocates the global radio spectrum and satellite orbit resources, develops the technical standards that ensure networks and technologies seamlessly interconnection, and strives to improve access to ICTs to underserved communities worldwide (26). Multiple radio interfaces (i.e., 3G, 4G) are developed under the name of International Mobile Telecommunications (IMT), and in early 2012, ITU-R embarked on a programme to develop "IMT for 2020 and beyond", setting the stage for 5G research activities emerging around the world. The following trends were identified, leading in a later phase to concrete requirements for the new 5G system defined by 3GPP (27):

- Support for very low latency and high reliability communications.
- Support of high user density (in one area, one cell, etc.).
- Support of high accurate positioning methods.
- Support of the IoT.
- Support of high quality communications at high speeds.
- Support of enhanced multimedia services and converged applications.

To transform the aforementioned requirements into 5G use case scenarios, IMT-2020 considers in Recommendation ITU-R M.2083[8] as three main 5G services eMBB, uRLLC, and mMTC. As illustrated in Fig. 2.3, each service is composed by a great diversity of requirements in order to meet the flexibility and diversity to serve many different use cases and scenarios.
As timeline, IMT-2020 as well as future enhancement of the existing IMT started to be implemented from early 2020 in some countries, according to different factors (spectrum availability, technology development, regulatory considerations, etc.). Meanwhile, further studies focused on traffic analysis, spectrum management, access network, and novel radio interfaces are under investigation for future IMT releases.

[7] https://www.3gpp.org/release-17.
[8] IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond

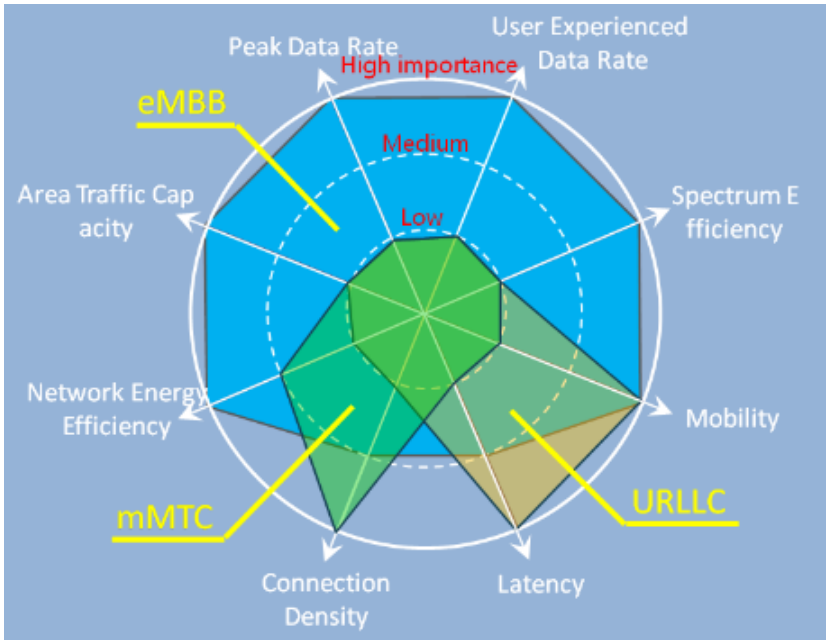


Figure 2.3: Usage scenarios of IMT for 2020 and beyond

2.2.3 Next Generation Mobile Networks (NGMN)

The NGMN Alliance is an open forum founded by world-leading mobile network operators, with the goal of design the next generation network infrastructure, service platforms and devices, satisfying end user demand and expectations. Starting from 2015, a series of white papers have been published describing the 5G vision requirements, technology trend, and system architecture proposals. In particular, their last work (28) highlights some of the vertical industries that 5G will progressively support, showing how mobile network operators can provide both network and services to meet the varied requirements of the different markets. Four main areas have been identified, particularly in the context of new and emerging paradigms in 5G business enablers, service delivery, and social responsibility:

- *Business context*: 5G will bring a massive change to cloud computing. The evolution of operator cloud may involve mixed private/public cloud models, for example involving more than one hyper-scale cloud providers. This innovation introduces several investment opportunities for cloud businesses, like streaming data and analytic, industrial IoT, edge computing, AI, Natural Language Processing (NLP), and AR.

- *Service activation*: the cloud native 5G system with Service Based Architecture (SBA) will be software-based and programmable, with separation of planes and RAN functions, network disaggregation and openness, distributed intelligence, composable core and hybrid clouds. From an infrastructure perspective, Network as a Service (NaaS) models have been introduced such that SPs easily operate the network without owning, building, or maintaining the

physical infrastructure. Users benefit by scaling up and down as demand changes, rapidly deploy services, and eliminate HW costs.

- *Business models*: 5G is a game changer in the business ecosystem, where MNOs have the unique possibility to differentiate and enrich their service portfolio and to perform multiple business roles, achieving new market opportunities. New 5G era monetizable attributes include NS and edge computing, which represent two of the main profitable sources for MNOs. While there is no direct mass market consumer play for NS, it is highly relevant for many consumers use cases, such as mobile gaming, and the many applications of AR and VR. In addition, there's a myriad of direct Business to Business (B2B) opportunities to organizations in industries from manufacturing to automotive, in mines and ports and across energy and utility networks (29).

- *Social and environmental responsibility*: multiple number of use cases, with diverse requirements, traffic volume, and connection density, are enabled by 5G and its evolution, through significant technological, operational, and business transformation. New 5G capabilities and use cases will be key components and enablers for future innovation in sectors such as the medicine, manufacturing, transportation, agriculture and several others. All these improvements require the simultaneously cooperation of different features: continue industry standardization to ensure interoperability and innovation, open access platforms to extend cross-industry ecosystem, innovation initiatives in national regulation, and harmonization efforts across countries and sectors for better services and cost levels (30).

In the NGMN's vision, 5G is expected to increase the service capabilities, flexibility, resilience, and efficiency. NGMN suggest to the worldwide research community to focus on the societal and environmental requirements, challenges and opportunities that future wireless systems can help address.

2.2.4 Global System for Mobile Communications Association (GSMA)

The GSMA is a global organization unifying the mobile ecosystem to discover, develop and deliver innovation foundational to positive business environments and societal change (31). Thanks to technology advances in many different fields, GSMA identified five goals for the 5G era aligned with the industries and society purposes (see Fig. 2.4):

1. *Worldwide boundless connectivity*: 5G will naturally evolve from existing 4G networks, but will mark an inflection point in the future of communications, bringing instantaneous high-powered connectivity to billions of devices. Moreover, it must coexist with the current G

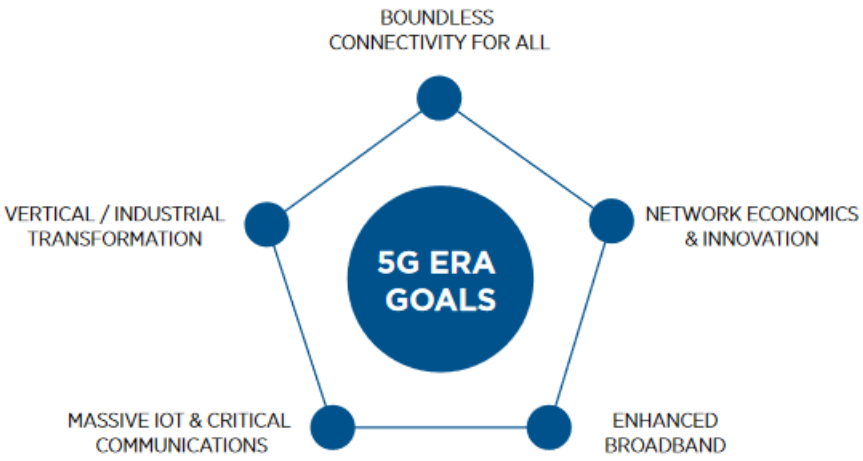


Figure 2.4: GSMA 5G mobile industry goals

technologies to deliver a boundless, high-speed, reliable and secure broadband experience, and support a plethora of use cases for society.

2. *Sustainable network economics and innovation*: 5G has triggered a new wave of interest in network sharing. The introduction of new technologies such as 5G NR architectures, active antennas, virtualization, edge computing, NS, and open networking reshape sharing principles that minimize CAPEX and TCO for operators, and improve overall network quality.

3. *Mobile broadband transformation*: 5G networks will provide between 10-times and 100-times faster data rates, at latencies of up to 10 times smaller when compared to current 4G networks. With carrier aggregation, for example, operators have not only harnessed the potential of their spectrum holdings to augment capacity and performance, but the technology is also the foundation for entirely new capabilities, such as operating LTE in unlicensed bands.

4. *Expand IoT use cases*: cellular IoT technologies will drive the digitalization of the society by delivering Machine-to-Machine (M2M) and machine-to-person communications on a massive scale. For all applications, solutions need to be integrated on platforms that can scale and handle millions of devices efficiently. SPs are in an excellent position to capture a share of the added value generated by the emerging cellular IoT market, as they are largely responsible for wireless connectivity on a global scale (32).

5. *Fast deployment in vertical markets*: while 5G technical activities are scaling up globally, requirements analysis of key vertical sectors is rapidly progressing. This requires new value propositions, new partnerships, business models, and improved cost structures for the benefit of the whole society and economy. While services and business models are in an early stage of development, industrial 5G activity is expected to pick up over the next two years, and the first sectors to benefit are likely to be automotive, manufacturing and media.

The GSMA expects commercial 5G networks to be widely deployed in the post-2020 period as a platform that enhances existing services, and enables new business models and use cases.

2.2.5 5G Public-Private Partnership (5G PPP)

The 5GPPP is joint initiative between the European ICT industry and the European Commission to rethink the infrastructure and to define key performance targets for the 5G service classes: eMBB, mMTC, and uRLLC (33). The 5GPPP projects are building pre-standards consensus among their partners and provide relevant contributions to standardization, focusing on innovation areas such as:

- Flexible air interface design that serves the three main services classes for 5G.
- Integration of different radio access technologies within the 5G TN.
- Flexible architecture that allows quick setup of slices on multi-tenant networks.
- Integration of fog computing and mobile edge computing near the end user and within vertical premises.

In 5GPPP vision, 5G technology acts as a catalyst for the development of new business relationships providing opportunities for SPs, Network Infrastructure Manufacturers (NIM), IT-SP, and business customers including Small and Medium-sized Enterprises (SMEs). New business opportunities emerge for telecom/network operators, manufacturers, solution providers, Software (SW) houses, brokers, startups, and SMEs that use 5G for creating innovative products and services for existing and new customers and markets, leveraging on the Anything as a Service (XaaS) model. These opportunities are conditioned by the ability of 5G technologies to provide the right performance that convince vertical stakeholders and allow the creation of this new dynamic 5G ecosystem.
To avoid fragmentation of 5G solutions deployment, 5GPPP collaborates with standardization bodies like 3GPP, ETSI and ITU for the definition of common interfaces among all the network domains. Standards must be flexible to support and sustain the diversity of business models and deployments of 5G networks, covering all use cases classes (eMBB, mMTC, uRLLC). The contributions of results include (33):

- 3GPP-RAN specifications of the PHY of the radio Interface for UE as well as radio interface architecture and protocols, radio resource control and management and the services provided to the upper layers.
- 3GPP-SA specifications of services and features, definition and evolution of the overall architecture, and addressing security and privacy by design.

- ETSI contributions to Mobile Edge Computing (MEC), Reconfigurable Radio System (RRS) and cyber security.

The 5GPPP future trend encourages the inclusion of the European use cases in the standards such that the technology required to realize them is developed. Finally, it is important that demand for new services requiring new technology is stimulated so that the equipment is developed and brought to market, fostering a healthy ecosystem.

2.3 IMT-2020 5G Mobile Services, Use Cases and Network Requirements

Research on 5G services and their technical requirements has been performed by the joint collaboration of standardization bodies, industries, and academics. From this global effort, in February 2021, ITU-R has published the definitive 5G design goals under the Recommendation ITU-R M.2150 titled '*Detailed specifications of the radio interfaces of IMT-2020*[9]. This Recommendation identifies and provides the detailed specifications of the radio interfaces for the terrestrial component of IMT-2020 and provides the radio interface specifications (34), as summarized in Table 2.1.
Considering as baseline the aforementioned three services defined by 3GPP, other standards and industry organizations have provided significant input to novel use cases, according to their needs. In general, 5G use cases can be grouped into five main categories (35), as illustrated in Fig. 2.5 [10]:

1. eMBB
    - e.g., Mobile Broadband, UHD / Hologram, High-mobility, Virtual Presence
2. Critical Communications
    - e.g., Interactive Game / Sports, Industrial Control, Drone / Robot / Vehicle, Emergency
3. mMTC
    - e.g., Subway / Stadium Service, eHealth, Wearables, Inventory Control
4. Network Operation
    - e.g., NS, Routing, Migration and Interworking, Energy Saving
5. Enhancement of Vehicle-to-Everything
    - e.g., Autonomous Driving, safety and non-safety aspects associated with vehicle

Table 2.1: IMT-2020 Minimum Technical Performance Requirement

<table>
<tr><th>Technical requirement</th><th colspan="4">eMBB</th><th>mMTC</th><th>uRLLC</th></tr>
<tr><td>Peak data rate (Gbps)</td><td colspan="4">DL: 20 Gbps
UL: 10 Gbps</td><td></td><td></td></tr>
<tr><td>Peak spectral efficiency (bps/Hz)</td><td colspan="4">DL: 30 bps/Hz
UL: 15 bps/Hz</td><td></td><td></td></tr>
<tr><td>User experienced data rate (Mbps)</td><td colspan="4">DL 100 Mbps
UL 50 Mbps</td><td></td><td></td></tr>
<tr><td rowspan="4">5th percentile user spectral efficiency (bit/s/Hz)</td><td>Environment</td><td colspan="2">DL</td><td>UL</td><td rowspan="4"></td><td rowspan="4"></td></tr>
<tr><td>Indoor Hotspot</td><td colspan="2">0.3</td><td>0.21</td></tr>
<tr><td>Dense Urban</td><td colspan="2">0.225</td><td>0.15</td></tr>
<tr><td>Rural</td><td colspan="2">0.12</td><td>0.045</td></tr>
<tr><td rowspan="4">Average spectral efficiency (bit/s/Hz)</td><td>Environment</td><td>DL</td><td colspan="2">UL</td><td rowspan="4"></td><td rowspan="4"></td></tr>
<tr><td>Indoor Hotspot</td><td>9</td><td colspan="2">6.75</td></tr>
<tr><td>Dense Urban</td><td>7.8</td><td colspan="2">5.4</td></tr>
<tr><td>Rural</td><td>3.3</td><td colspan="2">1.6</td></tr>
<tr><td>Area traffic capacity (Mbit/s/m2)</td><td colspan="4">10</td><td></td><td></td></tr>
<tr><td>User plane latency (ms)</td><td colspan="4">4</td><td></td><td>1</td></tr>
<tr><td>Control plane latency (ms)</td><td colspan="4">10 - 20</td><td></td><td>10 - 20</td></tr>
<tr><td>Connection density (devices/km2)</td><td colspan="4"></td><td>1,000,000</td><td></td></tr>
<tr><td>Energy efficiency</td><td colspan="4">Qualitative measure</td><td></td><td></td></tr>
<tr><td>Reliability</td><td colspan="4"></td><td></td><td>1-10-5 Success Probability for TX 32B in 1ms</td></tr>
<tr><td rowspan="4">Mobility</td><td>Environment</td><td colspan="2">Normalize traffic channel link data rate (Bit/s/Hz)</td><td>Mobility (km/h)</td><td rowspan="4"></td><td rowspan="4"></td></tr>
<tr><td>Indoor Hotspot</td><td colspan="2">1.5</td><td>10</td></tr>
<tr><td>Dense Urban</td><td colspan="2">1.12</td><td>30</td></tr>
<tr><td>Rural</td><td colspan="2">0.8 – 0.45</td><td>120 - 500</td></tr>
<tr><td>Mobility interruption time (ms)</td><td colspan="4">0</td><td></td><td>0</td></tr>
<tr><td>Bandwidth</td><td colspan="6">At least 100MHz, up to 1GHz for higher frequency bands (above 6 GHz).</td></tr>
</table>

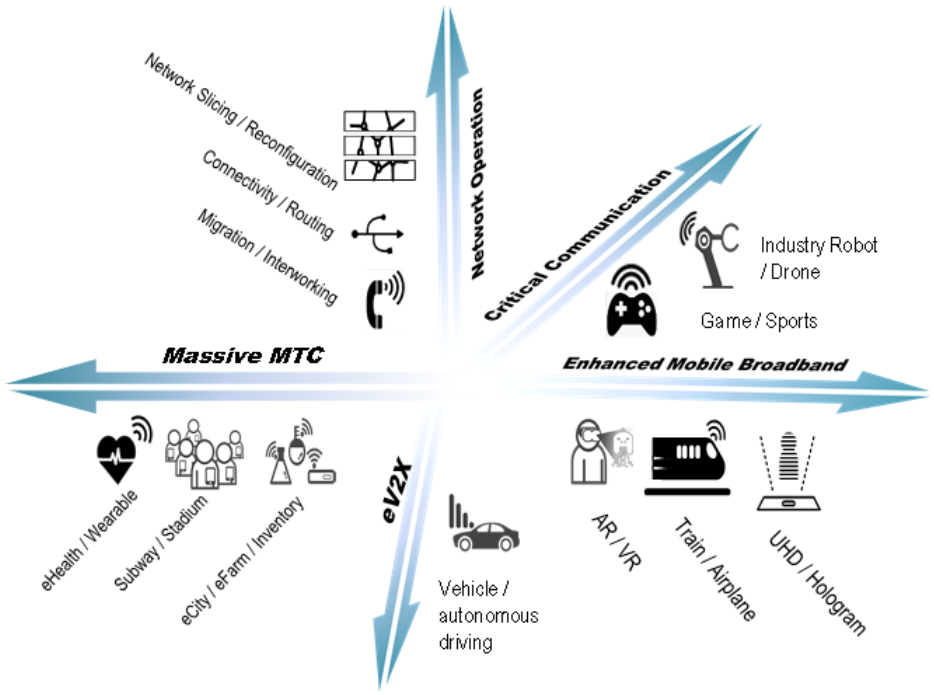


Figure 2.5: 5G general use cases

In light of the above use cases, the following technical requirements have been identified. They have been assessed from a technical point of view (36):

- *Reliability and availability*: Communication Service Reliability (CSR) has been defined in 3GPP TS 22.104 (37) and IEC 61907 (38) the ability of the communication service to perform as required for a given time interval under given conditions. As direct consequence of reliability, Communication Service Availability (CSA) has been defined by 3GPP as the amount of time the E2E communication service is provided in line with the agreed QoS, expressed as a percentage (39)-(40).

- *Security*: 5G has been designed to address many of the threats faced in today's 2G/3G/4G networks. These controls include new mutual authentication capabilities, enhanced subscriber identity protection, and additional security mechanisms. There are already various security architectures, infrastructures for credentials, and security rules implemented in industrial communication networks. Furthermore, 5G provides preventative measures to limit the impact to known threats, but the adoption of new network technologies introduces potential new threats for the industry to manage.

- *Transmission time (latency)*: transmission time or latency is the time taken to transfer a given piece of information from a source to a destination, measured at the Communication

[9]https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2150-0-202102-I\!\!PDF-E.pdf
[10]3GPP TR 22.891 V2.0.0 (2016-02)

Table 2.2: Potential latency requirements of different 5G scenarios

Scenario	Application	Latency Requirements
Enhanced Mobile Broadband	UHD video (4K, 8K), 3D video (including broadcast services)	Low latency (real-time video)
	Virtual Reality	ultra-low latency
	Augmented Reality	low latency
	Tactile Internet	Ultra-low latency
	Cloud Gaming	low latency
	Broadband kiosks	low latency
	Vehicular (cars, buses, trains, aerial stations, etc.)	low latency
Ultra-Reliable Communications	Industrial Automation	Low to ultra-low latency
	Mission-critical applications e.g. e-health, hazardous environments, rescue missions, etc.	Low to ultra-low latency
	Self-driving vehicles	Low to ultra-low latency
Massive Machine-Type Communications	Smart home	medium to high latency
	Smart office	medium to high latency
	Smart city	medium to high latency
	Sensor networks (industrial, commercial, etc.)	medium to high latency

Service Interfaces (CSIFs), from the moment it is transmitted by the source to the moment it is successfully received at the destination. According to the type of application, different 5G latency requirements have been defined (see Table 2.2), as described in the following two sub items (41):

– *System analysis and automation*: In the 5G era, operators clearly understand the need for automation: 65% of operators surveyed by GSMA Intelligence believe that the automation of business and Network Functions (NFs) is "extremely" or "very" important (42). Monitoring traffic characteristics and performance (e.g., data rate, packet drop, and latency), the end user's geographical distribution, per session-user-slice instance-based monitoring, and more will be key to the success of 5G technology. Automating operations such as quality monitoring, fault isolation, and service management can deliver immediate benefits for existing HW-based 4G networks and help prepare for the operation of new virtualized and cloud-native 5G networks. Following six phases, Table 2.3 illustrates how operators'management vision has evolved with the growth of 5G networks.

– *Network isolation*: Isolation is one of major challenges of slice networking with 5G networks. It refers to the degree of resource sharing that could be tolerated by the industry partner. The 5G TN and CNs can implement different isolation solutions,

Table 2.3: Phase vision of intelligent autonomous 5G networks (42)

-2*Phase		-2*Key Feature	Evaluation dimension					
			Execution	Perception	Analysis	Decision-making	Intent-driven	Scenario
L0	Network with manual operation	All manual operations	Manual	Manual	Manual	Manual	Manual	Nope
L1	Network with computer-assisted operation	Computer-assisted data collection, manual analysis and decision-making	Mainly System	Mainly Manual	Manual	Manual	Manual	Few scenarios
L2	Preliminary intelligent autonomous network	Automatic analysis and manual decision-making based on static policies in some scenarios	System	Mainly System	Mainly Manual	Manual	Manual	Some scenarios
L3	Intermediate intelligent autonomous network	Automatic analysis of dynamic policies in specific scenarios and system-assisted manual decision-making in pre-designed scenarios	System	System	Mainly System	Mainly Manual	Manual	Most scenarios
L4	Advanced intelligent autonomous network	The system implements complete closed-loop operation of dynamic policies, and automatically performs intent perception and implementation in pre-designed scenario	System	System	System	Mainly System	Mainly Manual	Overwhelmingly most scenarios
L5	Fully intelligent autonomous network	The system implements closed-loop operation of all scenarios, and automatically performs intent perception and implementation	System	System	System	System	System	All scenarios

through either HW (HW) or SW. On the 5G NR side, QoS scheduling mechanisms are mainly used to achieve SW-based isolation on WANs. The level of isolation is dependent upon the baseline security level and also upon customer needs and willingness to pay a higher isolation level per slice. As a consequence, the higher the isolation, the more resource-heavy and costly for the customer due to having more dedicated resources allocated to that slice regardless of workloads.

2.4 5G Features and Technical Enablers

5G mobile system is expected to build on the success of the current 4G technology, offering support for dedicated use cases and specific types of services to satisfy simultaneously various customer demands and diverse performance requirements. In order to address these requirements, a paradigm shift is taking place in the technologies that drive the networks, introducing new concepts and techniques are being developed to power the next generation mobile networks. In this section, the principal 5G features and technical enablers are illustrated, which are responsible of the major change in the way network services are deployed and operated.

2.4.1 5G Spectrum

Significant increase of bandwidth is necessary to support the new 5G services. The 3GPP 5G NR specification includes traditional mobile bands as well as newer, wider bands designed for 5G. Unlike conventional Global System for Mobile Communication (GSM) and LTE network uses frequency range below 4 GHz range, the new standard supports channel bandwidths ranging from 5 MHz to 100 MHz for bands between 3.4-3.6 GHz (5G NR sub-6GHz), and channel sizes from 50 MHz to 400 MHz in bands between 24-100 GHz (5G NR mmWave). Fig. 2.6 illustrates how from 2019 the main regulatory bodies have started to define 5G spectrum bands around the world (43); Europe has prioritized the 700 MHz band for wide area 5G and a growing number of countries globally are supporting the 600 MHz band (including the US which already uses it for 5G).

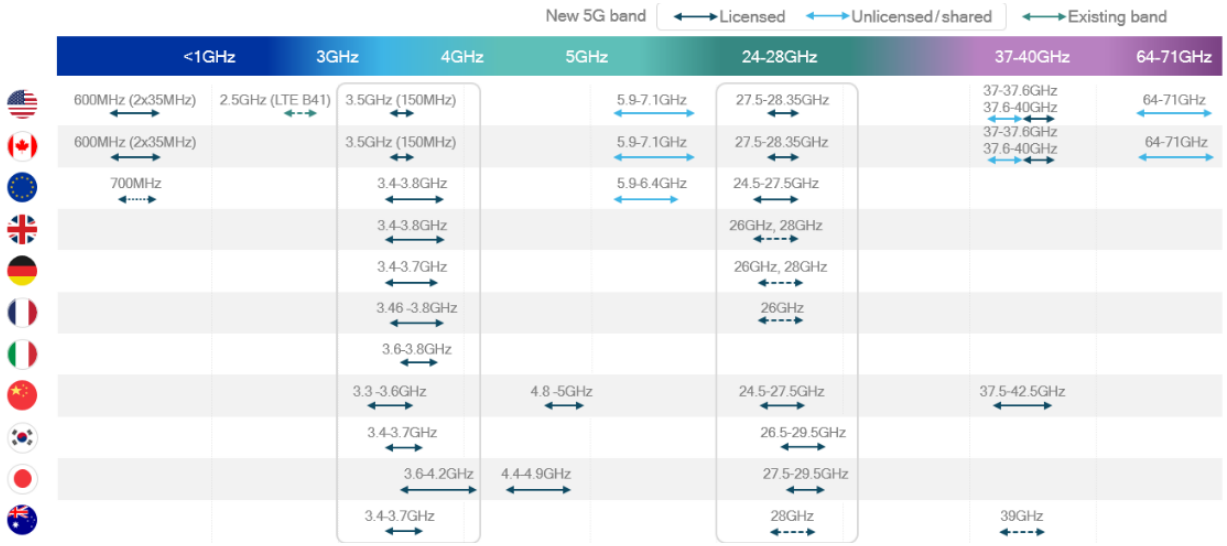


Figure 2.6: Global 5G spectrum. Allocated and targeted bands for 5G NR

Regulators have assigned 5G spectrum in three broad ranges:

- *High bands* (above 24 GHz, mmWave): support the fastest 5G speeds (extreme bandwidths). 5G NR represents the first mobile technology generation to use mmWave. Introduced as a complementing technology to the existing wireless deployments, it supplies new and enhanced experiences with multi-Gbps data rates, low latency, and virtually unlimited capacity while supporting a wide range of devices. At least 800 MHz of contiguous spectrum per 5G network should be available to meet the 5G requirement of very high capacity, especially in hotspot areas as well as for Fixed Broadband Fibre-like Connectivity ("WTTx") (44).

- *Mid bands* (1-6 GHz): offer a good mixture of coverage and capacity (eMBB and mission-critical). While the millimeter-wave high-band aspect of 5G gets all the attention, mid-band frequencies are what's going to unlock much of 5G's promise. Companies will use it to improve supply chains, autonomous driving, and VR, providing a better online experience than 4G

afforded and create a sizable and solid foundation for the M2M communications in IoT.

- *Low bands* (below 1 GHz): help provide strong wide area and in-building coverage (e.g., mobile broadband and massive IoT). In particular, the C-band (3300-4200 MHz and 4400-5000 MHz) has emerged as the primary frequency band for the introduction of 5G by 2020, providing an optimal balance between coverage and capacity for cost efficient implementations (44).

2.4.2 5G Air Interface: System Design Improvements

2.4.2.1 Flexible 5G Frame Structure

One of the main 5G challenges is the design of a system able to handle all the type of services (eMBB, mMTC, uRLLC) under the same ecosystem. Especially in the wireless domain, a performance trade-off between guaranteeing high spectral density, high reliability, scalability, and low latency must be designed in order to meet the future mobile broadband traffic requirements. For fulfilling such mixed conditions, a highly flexible and configurable air interface is needed. In this vision, a novel flexible frame structure is designed, able to fulfill the challenging 5G requirements for efficient support of a mixture of diverse services. The new structure includes Frequency-Division Duplexing (FDD) mainly for below 3 GHz frequencies, while Time-Division Duplexing (TDD) is preferred above 3 GHz, with some exceptions. Due to this fragmentation, efficient utilization of spectrum below 6 GHz calls for a flexible air interface design, introducing contiguous carrier bandwidths up to 100 MHz in the spectrum range 3–6 GHz, while narrower carrier bandwidths (up to 40–100 MHz) for sub 3 GHz FDD deployments are only possible with carrier aggregation mechanisms.
In the air interface, the type of radio waveform plays an important role in the design of 5G frame structure: so far, Orthogonal Frequency Division Multiplexing (OFDM), Single Carrier Frequency Division Multiple Access (SC-FDMA), and Filter Bank Multicarrier (FBMC) have been most widely considered (45), where their selection depends from the network deployment scenarios (46). Even though new waveforms bring advantages in terms of increase of bandwidth efficiency, relaxed synchronization requirements, and reduced inter-user interference, novel challenges in increased transceiver complexity and MIMO integration are introduced. For this reason, OFDM is still selected from 5G NR as waveform baseline due to its several merits such as low complexity, easy integration with MIMO, and plain channel estimation (47).
In order to allow efficient adaptation for each user in coherence with its service and radio requirements, time-frequency users multiplexing have been redesigned in 5G. In the time domain, the subframe length of 1 ms is fragmented into an integer number of slots, each one composed by 14 OFDM symbols. Each slot can carry control signals/channels at the beginning and/or ending

Table 2.4: Supported transmission numerologies and additional info

Numerology μ	**SCS** ($\Delta f = 2^{\mu} * 15KHz$)	N_{symb}^{slot}	$N_{slot}^{frame,\mu}$	$N_{slot}^{subframe,\mu}$	**Slot length (ms)**	**CP**
0	15KHz	14	10	1	1 ms	normal
1	30 KHz	14	20	2	0.5 ms	normal
2	60 KHz	14	40	4	0.25 ms	normal, extended
3	120 KHz	14	80	8	0.125 ms	normal
4	240 KHz	14	160	16	0.06 ms	normal

OFDM, with a variable DL/UL slot pattern (all DL, all UL, or at least one DL part and at least one UL part). Unlike LTE, scalable TTI sizes is standardized in 5G. At Layer 2 (L2), users are multiplexed flexibly over the resource grid with different TTI durations. The value Δt (integer number of OFDM symbols) determines the minimum TTI size for scheduling a user, as well as the resolution for other TTI scheduling options (48). As consequence, the research community established that a minimum TTI size of no more than 0.2–0.25 ms is needed most stringent latency requirement of 1 ms for Mission-Critical Communication (MCC). The reduced TTI size, combined with stricter network and device processing requirements, allows a sufficient delay budget for sending the payload, receiving and processing it, followed by sending the corresponding Acknowledge (ACK) message (49).

In the frequency domain, 5G employs scalable Subcarrier Spacings (SCS) as a subset or superset of 15 kHz. Feasible SCS can be 15 kHz x 2^m, where m can be a positive integer or zero. For each SCS value, multiple Cycling Prefix (CP) lengths can be inserted to adapt with various levels of Inter-Symbol Interference (ISI) at different carrier frequencies and mobility. As in LTE, the basic scheduling unit in 5G NR is a PRB, which is composed of 12 subcarriers, with the exception of supporting multiple SCS. For this reason, when PRBs of different bandwidth ranges are multiplexed in the time domain, boundaries of PRBs should be aligned, forming a PRB grid, as illustrated for each UE in Fig. 2.7.

Compared to LTE, 5G numerology (combination of SCS and symbol length, indicated with μ) represents the most outstanding innovation in the radio domain: as illustrated in Table 2.4, 3GPP Release 15 defines different numerologies for mmWave and FR1 frequencies, with larger sub-carrier spacing for higher frequencies[11].

2.4.2.2 5G Hybrid Automatic Repeat Request (HARQ) Enhancements

Basic concept of HARQ in NR is similar to LTE HARQ, but there are some minor differences in terms of the details. In LTE, HARQ procedure uses asynchronous mechanism for DL, and synchronous mechanism for UL. Even though synchronous implementation reduces signaling overhead, the LTE HARQ approach increases the system complexity since transmitter and receiver in

[11] 3GPP TS 38.211 version 15.2.0 Release 15.

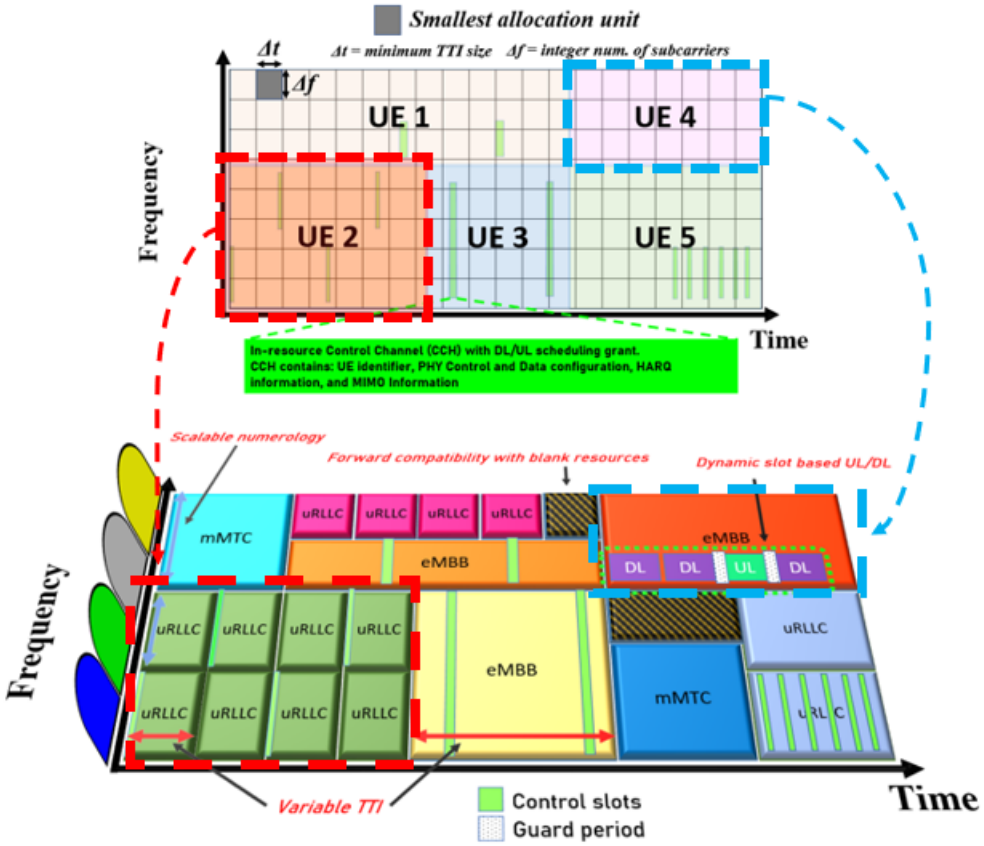


Figure 2.7: 5G dynamic user multiplexing and scheduling grant

the HARQ process should know the exact HARQ process number for each transmission/reception of the HARQ data to keep track of each HARQ process even when they are not running in order. As first improvement, to cope with the flexibility of using variable symbol duration, HARQ feedback is set asynchronous in both DL and UL directions. This methodology uses DL and UL Downlink Control Information (DCI) to carry the HARQ process number, ensuring a lower over the air latency. Moreover, it is a fundamental feature for solutions where PHY-MAC HW are disaggregated, in order to support centralized radio network implementations with different FH latencies. As second improvement, dynamic scheduling has been introduced to control the delay between different paired of control and data transmissions. While in LTE the timing between data transmission and HARQ response is fixed (i.e., 4 ms in FDD), in 5G NR a table listing all the possible timing between data, HARQ, and DCI is specified in the Radio Resource Control (RRC) message[12]. As third improvement, similar to the K parameters of LTE[13], 5G NR introduced novel scheduling/HARQ-ACK feedback timing parameters named K0, K1, and K2 to control the timings that govern the communications between gNB and UE (50) - (51).

As illustrated in Fig. 2.8, the timing parameters have a different roles:

[12]3GPP TS 38.214 version 15.3.0 Release 15
[13]3GPP TS 36.211 version 14.2.0 Release 14

- *K0*: it represents the interval between the DL scheduling Slot DCI (in Physical Downlink Control Channel (PDCCH)), and the corresponding scheduled data (in Physical Downlink Shared Channel (PDSCH)).
- *K1*: it represents the PDSCH HARQ-ACK feedback slot spacing, meaning the delay between DL data (PDSCH) reception and corresponding HARQ-ACK feedback transmission on UL (in Physical Uplink Control Channel (PUCCH)).
- *K2*: it corresponds to the number of slots between UL scheduling interval DCI (PDCCH), and its scheduled UL data (in Physical Uplink Shared Channel (PUSCH)).

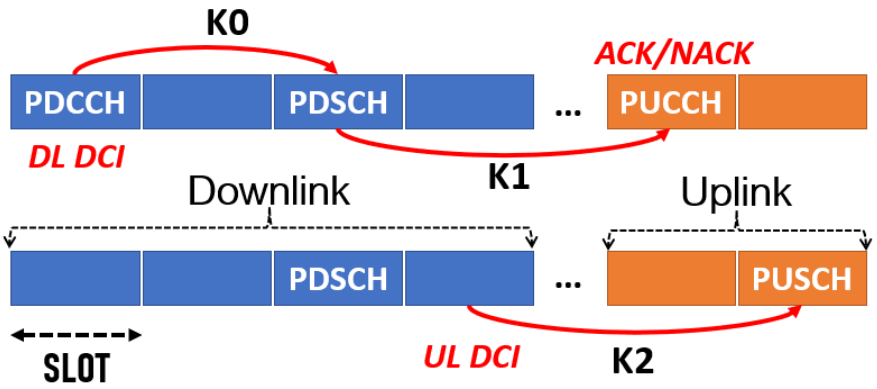


Figure 2.8: K0, K1, K2 timing parameters roles

As final mayor enhancement, 5G NR does not explicitly implement HARQ ACK-NACK for PUSCH. The UE understands the successfully PUSCH delivery if it receives or not a retransmission request from the gNB.

2.4.3 SDN and NFV roles in 5G

Nowadays a MNO faces increasing difficulties to launch a new service due to a variety of proprietary HW appliances, power costs, capital investment challenges, complex HW-based interoperability, etc. Moreover, network equipment life cycle becomes shorter as technology and services innovation accelerate, pushing a growing group of companies and standardization bodies to investigate the NFV paradigm to improve cost efficiency, flexibility, and performance guarantees of cellular networks in general. NFV allows network operators to manage and expand their network capabilities on demand using virtual, SW-based applications where physical boxes once stood in the network architecture. This makes it easier to load-balance, scale up and down, and move functions across distributed HW resources. With continual updates, operators can keep things running on the latest SW without interruption to their customers[14].

[14]https://www.ericsson.com/en/nfv.

Through NFV, vendors deploy SW-based network components called Virtual Network Functions (VNFs), which have the advantage to be easily deployed on cloud infrastructures instead of specific HW. As an example, for disaggregated RAN solutions, signal processing resources are virtualized in a cloud environment instead of using multiple Baseband Processing Units (BBUs) (52).
Among the multiple benefits introduced with NFV, the more relevant are (53):

- Reduced CAPEX/OPEX costs and exploitation of scalable IT industry solutions.
- Reduced network operator cycle for the launch of a new service or product. Since NFV is HW independent, massive investments in IT equipment is no longer required, making feasible other modes of feature evolution. NFs virtualization should enable network operators to significantly reduce the maturation cycle.
- Platform sharing allows the use of a single platform for different applications, users and tenants, contributing to a multi-version and multi-tenancy environment. This allows network operators to share resources across services and across different customers.
- Services can be rapidly scaled up/down on demand, breaking the geographic boundaries faced with previous service architectures.
- Openness and scalability drive the market to novel SW entities, small players and academia, encouraging more innovation to bring new services and new revenue streams quickly at much lower risk.

NFV is highly complementary to Software Defined Networking (SDN). It adopts two main ideas: logically centralized control of the Data Plane (DP), and network state management across distributed controllers. This separation facilitates increasing traffic volumes and improves network reliability, predictability, and performance. SDN makes it possible to manage the entire network through intelligent orchestration and provisioning systems. Thus, it allows on-demand resource allocation, self-service provisioning, truly virtualized networking, and secures cloud services. The value of SDN in 5G wireless networks lies specifically in its ability to provide new capabilities like network virtualization, automating and creating new services on top of the virtualized resources, in secure and trusted networks. Also, SDN enables the separation of the control logic from vendor-specific HW to open and vendor-neutral SW controllers (54).
NFV and SDN were developed by different standardization bodies; however, they are complementary technologies and as a result are often collectively referred as NFV/SDN. Although they have substantial value when exploited separately, in combined they offer significant additional value.
Fig. 2.9 illustrates the integration architecture between SDN and NFV proposed by 5GPPP (55). NF are classified according to the compatibility degree with SDN/NFV implementation. The set of

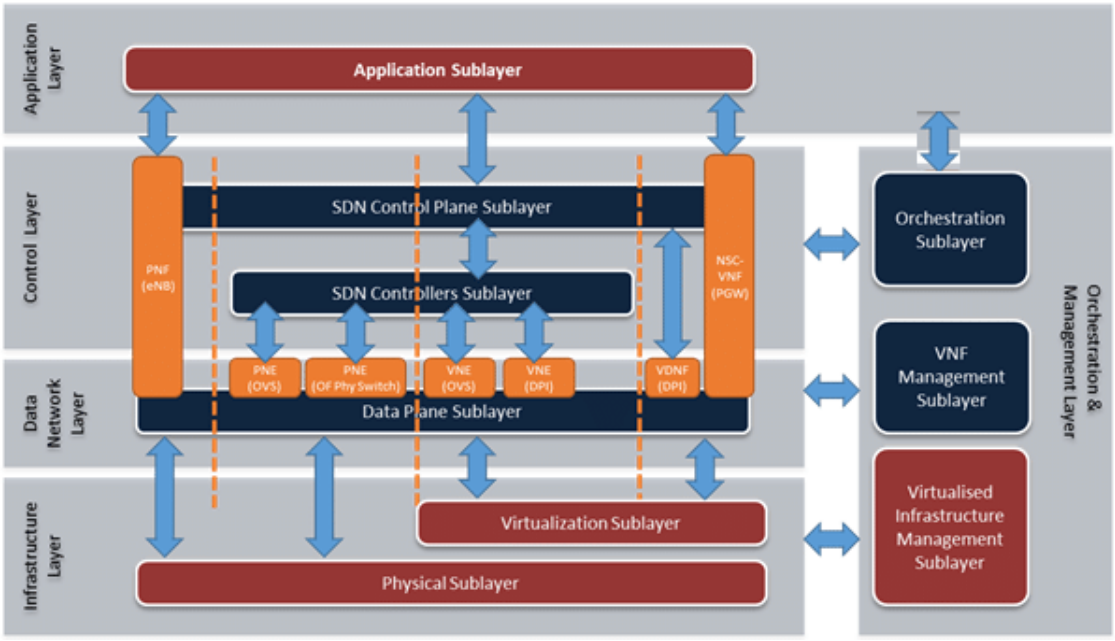


Figure 2.9: Overall architecture of SDN/NFV integration and management

physical NFs such as Physical Network Elements (PNEs) and Physical Network Functions (PNFs) are generally controllable using SDN, with some exceptions (i.e., Non-SDN-Enabled Virtual NF (NSC-VNFs)). VNFs with DP-only functionalities (i.e., Data NF (DNF)) exploit SDN to separate the two planes'implementations. Architecturally, from the bottom up, the infrastructure layer provisions both physical and virtualized resources and the operation environment for the various VNFs. The control layer is broken down to a SDN Controller sublayer and a SDN CP sublayer on top of it (56).

2.4.4 5G Functional Split

Recent advances in NFV enabled MNOs to transit from the fully Decentralized RAN (D–RAN) architecture, where baseband processing and radio elements are co–located, to the fully centralized C–RAN architecture (57) (see Fig. 2.10).

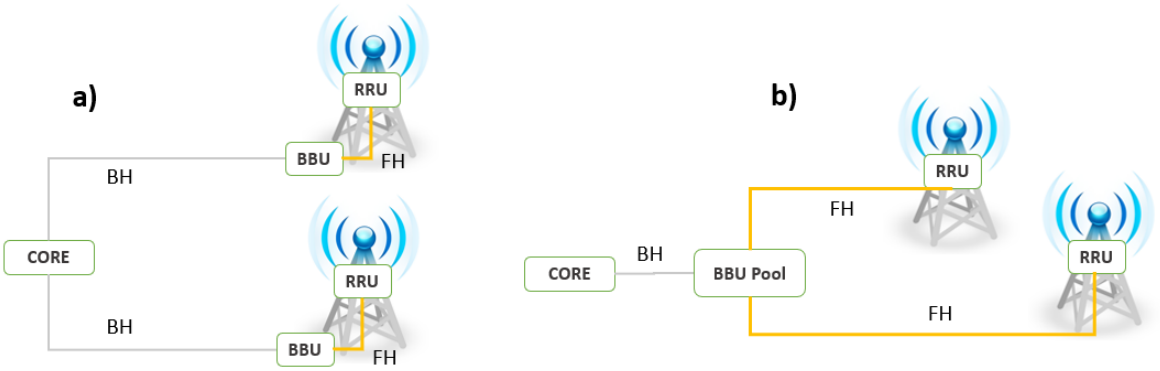


Figure 2.10: a) Decentralized RAN b) Cloud RAN

Both architectures are two distinct concepts, both with advantages and disadvantages: D–RAN requires relatively low BH capacity, while C–RAN enables joint signal processing techniques, such as CoMP, introducing higher BH requirements. In order to tackle the aforementioned challenges, a number of intermediate functional splits, each characterized by a different demarcation point between Distributed Units (DUs) and Centralized Units (CUs), have been proposed (58). FP determines the number of functions left locally at the antenna site, and the number of functions centralized at a high processing powered datacenter. It aims to leverage the benefits of virtualization (e.g., reducing costs and dynamic scalability) and centralization (e.g., statistical multiplexing gains). In 5G NR, this method consists into splitting the processing and baseband functionalities of the 3GPP protocol stack between CU and DU. To be ready for the future mobile networks, standardization bodies and joint alliances are proposing different FP deployments, according to precise use cases:

- 3GPP FP provides eight different suggestions numbered from 1 to 8 (3GPP TR 38.801), where gNB may consist of a gNB-CU and one or more gNB-DU(s) (59). Since the majority of these options present a set of issues and challenges that will difficult their short-term implementation, industries are more inclined to option 2 for CU-DU splitting, which could be implemented on the basis of Dual Connectivity (DC) standard (60).

- The ITU-T Technical Report (TR) on TN adds an extra refinement to the 3GPP 5G FP architecture, referring to one or two tier FP, where gNB is split between Remote Unit (RU), DU and CU (61). This disaggregation supports multiple final application scenarios due to higher flexibility in terms of wireless specifications, applications (i.e., eMBB vs URLLC), transport technology, and operators'deployment requirements.

- enhanced Common Public Radio Interface (eCPRI): the Common Public Radio Interface (CPRI) protocol which only considers split option 8 has been extended to eCPRI which covers more 3GPP compliant options (1, 2, 4, 6, 7, and 8), labelled with letters (A to I).

Although 3GPP FP architecture is widely adopted, its single vendor-oriented structure and lack of specification regarding the interoperability in the RAN represent a barrier for the virtualization and sharing principles introduced with 5G. For this reason, new FP options have been investigated where innovative solutions from different vendors can be integrated. Founded in February 2018 by AT&T, China Mobile, Deutsche Telekom, NTT DOCOMO and Orange, O-RAN aims to create easier interoperability on existing 3GPP RAN interfaces (62). The O-RAN alliance has specified an eCPRI-based 7.2x open interface between the O-RAN Radio Unit (O-RU) and O-RAN Distributed Unit (D-DU), and it is the only standardized FH interface that enables multi-vendor

interoperability. Fig. 2.11 offers an overview of the different FP options of the main alliances and standardization bodies.

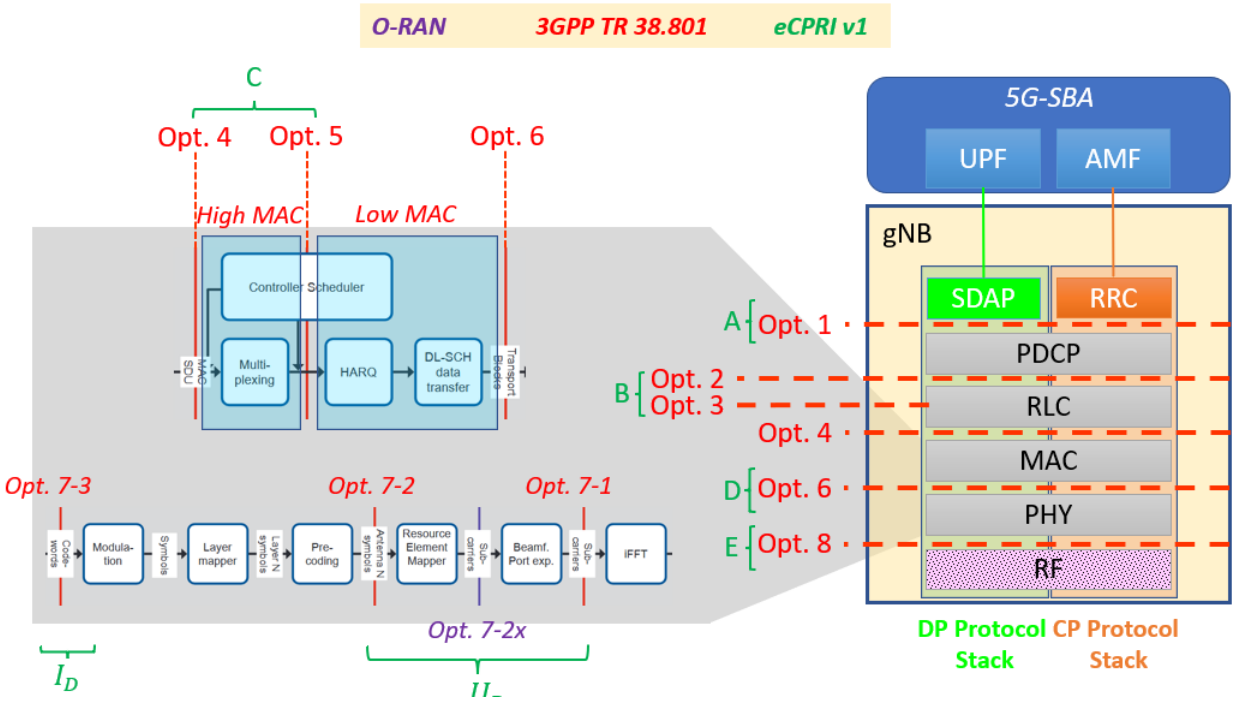


Figure 2.11: 5G NR FP nomenclature from different standardization bodies

2.4.5 Network Slicing: the key 5G Enabler

Earmarked as a prominent feature of 5G for enabling the aforementioned technological capabilities, the concept of NS has been introduced. This key technology enabler permits to multiple vertical industries the execution of their solutions on top of a shared infrastructure (see Fig. 2.12), where the SP customizes its own network capabilities (security, connection, processing power, storage, etc.) (63). From a SP point of view, this solution represents a new type of business model, known as NS as a Service (NSaaS), which will drive the future network trend of the next years. Multiple customers can be allocated within a single slice entity, as well as different services belonging to multiple network slices may be gathered together and supplied as a single slice to a customer with diverse requirements. To realize this new service model, a SP configures a network slice instance following some baseline principles (64):

- The network section (RAN, TN, and CN): each section presents different requirements and performance which should be carefully analyzed before instantiating a service.
- The SLAs: latency, Guaranteed Bit Rate (GBR), non-Guaranteed Bit Rate (non-GBR), availability, and packet loss are some parameters to evaluate when the slice should be defined.

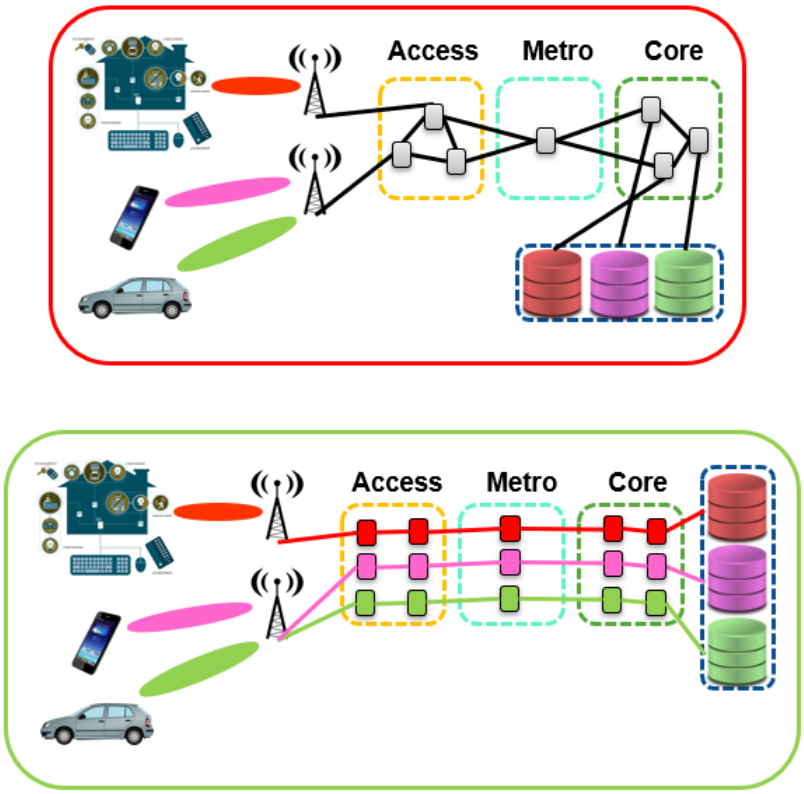


Figure 2.12: a) "one size fits all" approach b) custom tailored solutions

- The type of vertical market application: industry 4.0, Vehicle-to-Everything (V2X), smart cities, IoT, etc.
- The deployed access technology: cellular and fixed networks (i.e., LTE, WiFi, optical access solutions) present different technological performance and limitations. Since a slice could potentially exploit all the architecture, the SP should consider how each access technology affects the global service performance.
- The type of service provided (eMBB, uRLLC, mMTC): according to the service requirements, the traffic flows belong to different types of services, facilitating the management and service supervision.
- Cross domain services: services across multi-provider domains such as L2, Layer 3 (L3), and VPN services.

The customization of different slice granularities introduces multiple challenges, especially in terms of slices orchestration and resource sharing mechanisms. An early solution proposed by 3GPP is called *static* NS (65), where the slices are already instantiated, and the device has to proceed only with the slice selection. To avoid SLA violation, this technique uses static resource over-provisioning, introducing possible waste of resources which might be allocated to other services. To

overcome this limitation and considered the spatially inconsistent time varying traffic trend within a slice, *dynamic* resource allocation schemes permit the tenants to apply customized strategies by monitoring the KPIs, so as to maximize its own utility. With dynamic NS, multiple tenants adjust their network capacities during different time periods, according to service and system variations. From a 3GPP perspective, the System Aspects 2 (SA2), System Aspects 3 (SA3), and System Aspects 5 (SA5) Working Groups (WGs) focused on the standardization of the management, system architecture, and orchestration of 5G NS[15]. The process to request, instantiate, and manage a slice is described in Fig. 2.13, where the E2E slice life cycle process is illustrated as a series of interactions (preparation, commisioning, operation, and decommissioning) using different tenants northbound Application Programming Interfaces (APIs), or the underlying management systems (i.e., NFV Orchestrator).

5G NR protocol stack introduces two new functional blocks for the E2E slice management: the Network Slice Management Function (NSMF) and the underlying Network Slice Subnet Management Function (NSSMF) for RAN, TN, or CN (66). In particular, with a focus on the RAN domain, Fig. 2.14 provides an overview of the radio protocol stack workflow performed by the RAN Orchestrator when the slice template is forwarded from the NSMF to the NSSMF.

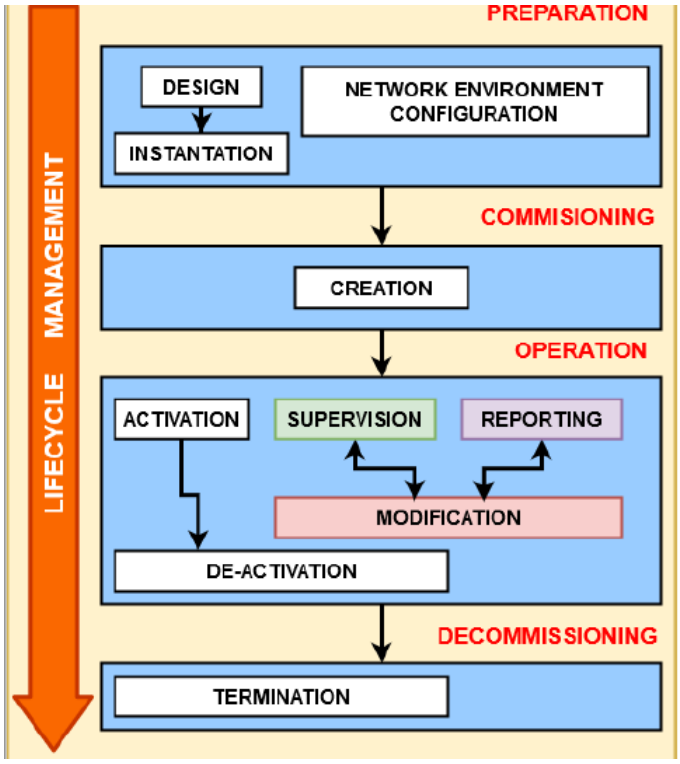


Figure 2.13: Slice life cycle management

At the top layers, each slice is identified and characterized using a RAN slice descriptor, which contains the Operator-ID (OP-ID), the Public Land Mobile Network-ID (PLMNID), and a slice ID

[15]3GPP TS 23.501, TS 28.533.

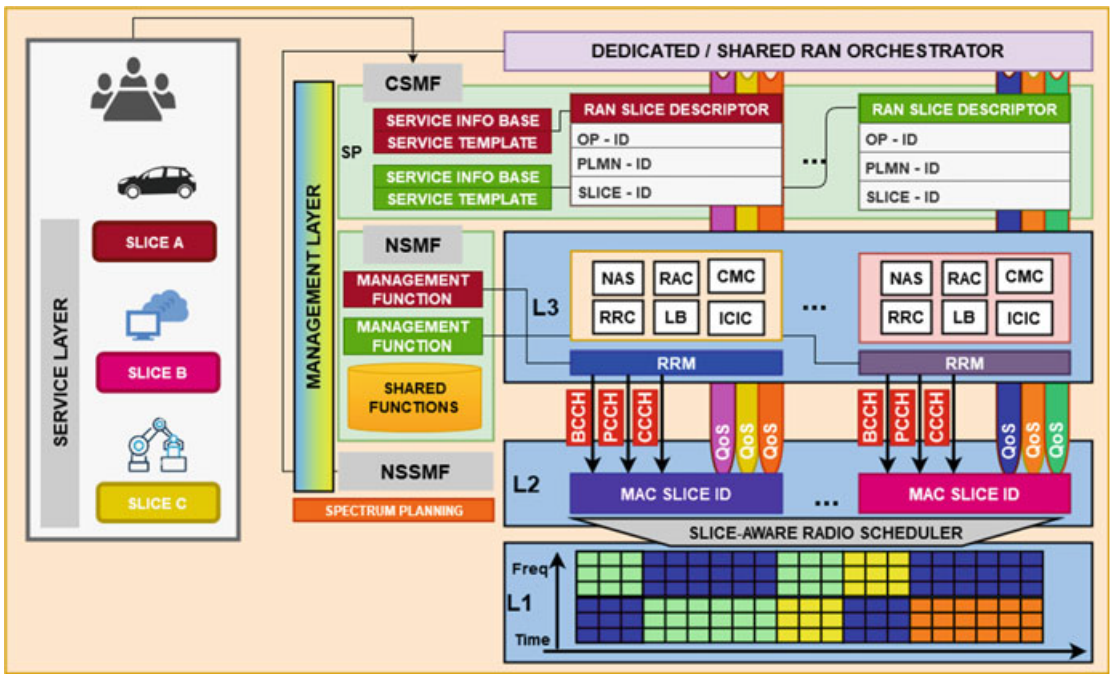


Figure 2.14: E2E RAN protocol stack

(SLICE-ID) required to identify a specific slice within the SP domain. At L3, different functionalities cooperate for radio resource management aspects: the Non-Access Stratum (NAS) handles the CP between the UE and the CN, the Radio Admission Control (RAC) analyses the QoS requirements and resource status before accepting or rejecting a new radio bearer, the Connection Mobility Control (CMC) facilitates the radio resource management according to the UE mobility, the RRC reconfigures the UE according to the radio bearer, measurements, and cell group, the Load Balancer (LB) manages the real-time traffic under different load conditions, and the Inter-cell Interference Coordination (ICIC) keeps the inter-cell interferences under control by Radio Resource Management (RRM) methods. For each slice, the bearers information are forwarded to the L2, where a first scheduling procedure is executed among the traffic flows belonging to the same slice. Following the sharing infrastructure principles of 5G, the radio section may also be shared among SP. For this reason, before reaching the L1, another slice-based radio scheduler is defined, which may introduce a further prioritization policies among the SPs. A flexible PRBs control is an important property for RAN slicing at L1. The allocation of PRBs to each slice based on each slice's minimum and maximum sharing portions is done according to the operator's SLA and the use case scenario. Finally, to guarantee the correct resource orchestration, the concept of slice isolation has been introduced. Isolation can be defined as the ability of a NSP to ensure that congestion, attacks and life cycle-related events (e.g., scaling in/out) on one NSI does not negatively impact other existing NSIs (67). Isolation in terms of performance represents the ability to ensure that service KPIs are always met on each NSI, regardless of the workloads or faults of other existing NSIs. Isolation in terms of management represents the ability to ensure that individual NSIs can

be managed as separate networks, with the possibility of the customer to retain control of the slice. Following the main five isolation flavors applied to NS are illustrated:

- *Function isolation*: each slice can have a different set of functions, with different services and requirements.
- *Configuration isolation*: different types of configurations can be used for the same function or service.
- *Resource isolation*: the administration of the resources among the slices is defined through policies.
- *Security isolation*: definition of security policies to prevent unexpected intrusion among the slices.
- *Slice life cycle isolation*: guarantees the creation, support and removal of each slice in the network.

2.5 5G Architecture, Sub-domains and Deployment Options

To achieve ultra-high data transmission rate and ultra-low response time (latency), 5G redesigns the current architecture using SDN and NFV techniques. This virtualized approach has a minimal impact on the physical infrastructure, which is expensive and complex to renovate. The highlights of this architecture are summarized as follows (68):

- A clear separation of the CP and UP functions.
- Scalability, flexibility and rapid establishment of NFs.
- Intelligent resource management that reduces the ability to reuse a network service for different tenants.
- A vision of the network as a single system, reducing the costs and management procedures for the interconnection of different NFs.
- A unique authentication system that allows a simple interaction between the network devices.
- Minimization of the private infrastructure concept. AN and CN belonging to different domains must be equipped with control and management systems capable of communicating with each other.
- Migration of NFs between AN and CN for service support and low latency and reduction of congested situations between them.

Compared to LTE, 5G must combine the previous listed capabilities over the entire E2E architecture, seamless, and transparent to the end customer. As illustrated in Fig. 2.15, given the type of service, the NFs are ad-hoc placed on top of the three domains characterizing the 5G architecture (69): i) the FH domain, which is the network between the Remote Radio Unit (RRU) and the DU, ii) the MH domain in the middle between the DU and the CU, and at the end by iii) the BH domain, which is the bridge between the CU and the 5G CN.

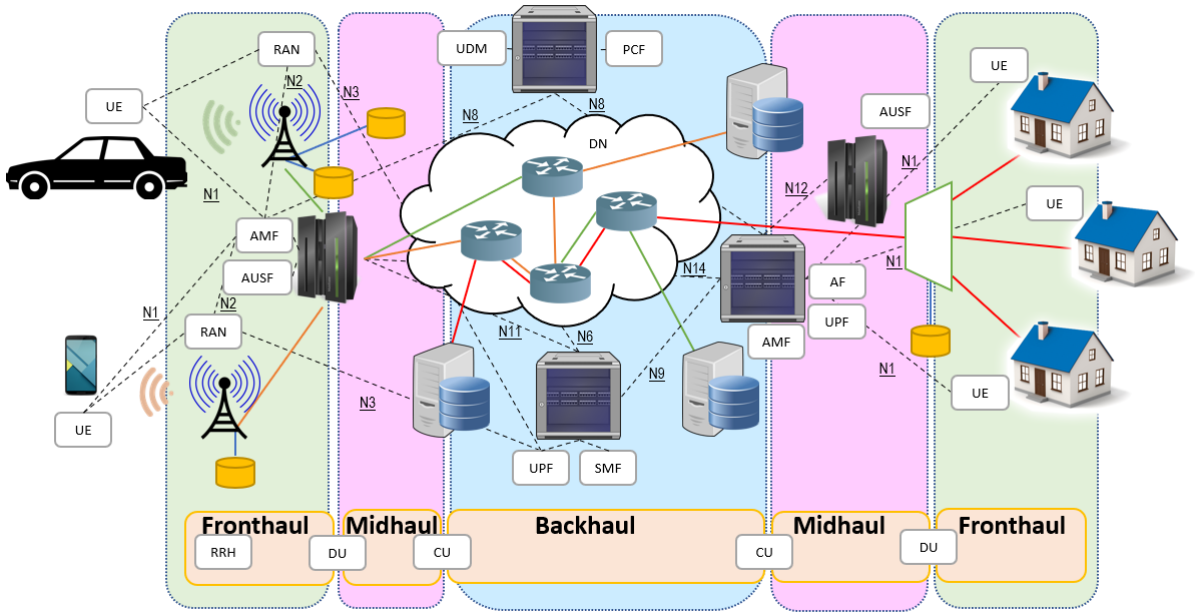


Figure 2.15: 3GPP 5G E2E TN Architecture

Each node is interconnected with one or more network entities (server, data-center, storage, router) equipped with different network functionalities, such as firewall, authentication service, proxy, virtual switch, and virtual machines. The combination of NS with an SDN controller allows defining an optimal combination between the functions of each node and the service requirements. Moreover, at the FH, additional nodes with computational resources are directly connected to the eNBs acting as MEC nodes. This technology allows content caching close to the network edge, which can support the deployment of NS in the RAN part.

2.5.1 5G System Architecture: the first "G" Service-Based Architecture

The existing mobile network architecture designed for 4G is able to meet voice and conventional MBB service requirements. Its "box-driven" architecture is executed on dedicated HW structures, introducing barriers in terms of network differentiation. The main limitations of the 4G architecture are:

- Point-to-point interconnection between specific functionalities. The extension of new functions to the architecture requires the definition of new policies and parameterization settings.

- High dependency among the EPC functionalities. MME, SPGW, and PCRF shares part of the functionalities, increasing the complexity deployment.
- 4G DP masks the session UE information of each architecture component. There is no easy method for NFs to expose and share their session data.

To meet the needs of new 5G services with different requirements across multiple industries, virtualization and service-based mechanisms are a significant industry trend especially relevant for the 5G ecosystem. 3GPP defines a SBA for 5G able to provide a modular framework from which common applications can be deployed using components of varying sources and suppliers. The architecture elements are defined as NFs[16] able to supply services via interfaces of a common framework to other NFs. Under this model, a service is defined as an atomized capability with the characteristics of high-cohesion, loose-coupling, and independent management from other services.
Unlike the classical legacy networks, SBA shall bring the following benefits to 5G (70):

- *Updating Production Network*: new degree of scalability than legacy network allows services to be modified without impacting the other services. This introduces multiple advantages in service management, reducing processing time and costs.
- *Extensibility*: light-weighted protocol facilitates the autonomous communication among services without human interaction. This facilitates the introduction of new features without interrupting the workflow.
- *Modularity and Reusability*: network capabilities are decomposed into modules, while their concatenation enables 5G features such as NS. A service can be easily invoked by other services, enabling each service to be reused as much as possible.
- *Openness*: a novel suit of third-party interfaces is designed in 5G for higher management and control capabilities without complex protocol conversions.

Fig. 2.16 and Fig. 2.17 illustrate the Local-Break Out (LBO) and Home-Routed (HR) roaming reference 5G SBA respectively, primarily the 5G CN NFs, with a single interconnect to the rest of the system as defined in 3GPP TS 23.501 (71). These two roaming schemes have been introduced to correctly use IP Multimedia Subsystem (IMS) services even when the UE is not geographically located in the service area of the Home Public Mobile Network (HPMN); as an example, in Fig. 2.16, the roaming UE interfaces the DN in the Visited Public Mobile Network (VPMN), and the HPMN enables it with subscription information, subscriber authentication, and UE specific policies.

[16]3GPP has introduced NF Services in 3GPP for release 15 (3GPP TS23.501 "System Architecture for the 5G System"). A NF represents a set of services called NF Service through a service-based interface that is consumed by other authorized NFs.

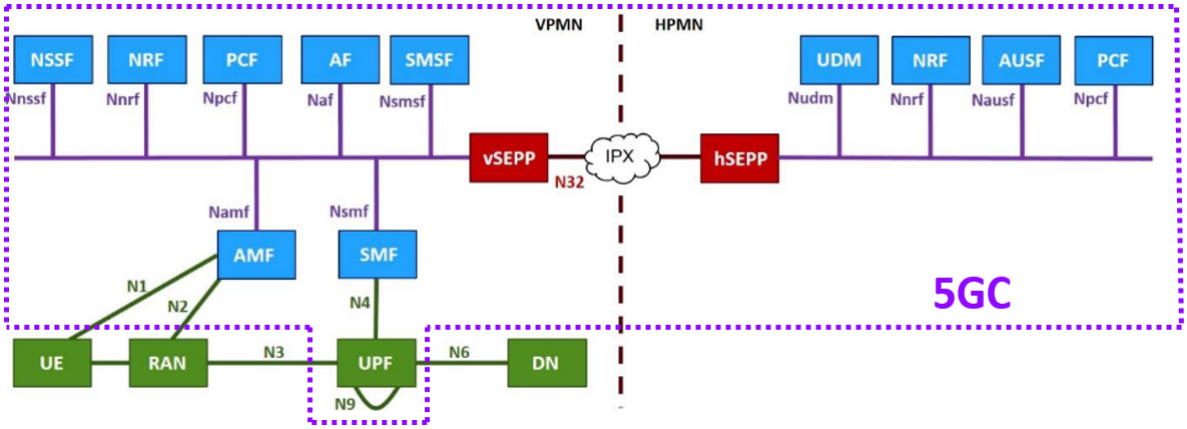


Figure 2.16: 5G System Architecture – LBO

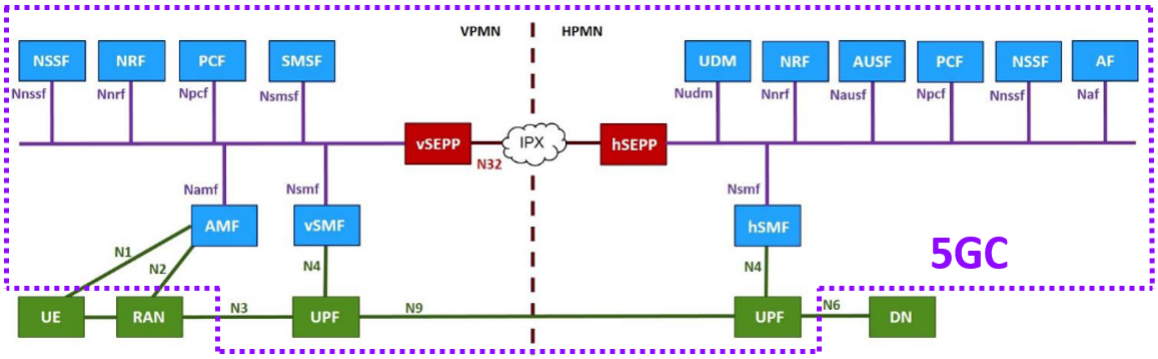


Figure 2.17: 5G System Architecture – HR

Unlike the monolithic point-to-point 4G architecture, 5G adopts a common protocol framework as basic paradigm to provide NF services via APIs. A Central Repository Function (CRF) collects all the NFs and their provided services such that other NFs retrieve a functionality by querying the central repository on the instances providing the services. This paradigm enables each service to have its own life cycle, to easily add new services without impact to any other service and to decouple the SP and service user in a sense that a new version of a service, or a completely new service, doesn't impact the user until the user is ready to use this new version or the new service. The 5G SBA adopts principles like modularity, reusability, and self-containment of NFs to enable deployments considering the latest virtualization and SW technologies. The aforementioned SBA figures depict those service-based principles by showing the NFs, primarily CN functions, with a single interconnection to the rest of the system. The 5GS architecture, as fully described in (72)-(73)-(74), comprises of the following NFs (75):

- *Authentication Server Function (AUSF)*: this functionality, closely correlated to the Unified Data Management (UDM), is always places in the Home PLMN of the subscriber. It is responsible of the early authentication and key agreement procedures carried out to enable mutual authentication between UE and network.

- *Access and Mobility Management Function (AMF)*: it terminates the Access Network (AN) to CN CP interface (N2) as well as the UE to CN NAS signalling interface (N1). This functionality manages the UE access control of the UE registration in the 5G System, UE access security management, and UE mobility management.

- *Data Network (DN)*: e.g., operator services, Internet access or 3rd party services

- *Unstructured Data Storage Function (UDSF)*: it is introduced to store dynamic state data. Instead of a NF holding its own storage resources, it stores UE context data in the UDSF.

- *Network Exposure Function (NEF)*: this functionality controls the access of external domains to the data resources within the 5G CN CP. Other activities covered from NEF are secure data exposure and mapping of identifiers and concepts.

- *NF Repository Function (NRF)*: the NRF provides a service discovery function for other NF to use. It also includes NF profiling, NF instance management, and NF service discovery. The profile includes address of the instance, the services supported, and key attributes defining the domain in which these services are authorized to be used, such as network slice related identifiers and the PLMN ID. As complementary part of the discovery mode, NRF is notified every time a set of instances is modified.

- *Network Slice Selection Function (NSSF)*: the NSSF selects the network slice(s) to be used for a UE. This operation consists of mapping Single Network Slice Selection Assistance Information (S-NSSAI) to the actual Network Slice Instance Identifier (NSI). Moreover, the NSSF determines the set of AMF instances that can be used to serve the UE. Each PLMN runs its own NSSF, which has an overview of all slices of the network.

- *Policy Control Function (PCF)*: it is responsible for the definition of any type of policies in the network and for delivering related policy rules to the other CP NF. The PCF supports retrieving subscription information or specific application requirements, retrieving network conditions, policy rule determination, and policy rule delivery to a NF.

- *Session Management Function (SMF)*: this functionality supports Protocol Data Unit (PDU) session control, termination of the session management part of the NAS interface, UE IP address allocation, User Plane Function (UPF) selection and control, control of policy enforcement, and charging. In 5G, the PDU replaces the concept of Packet Data Network (PDN) connection in EPC. A PDU session provides the application connectivity between the UE and a Data Network (DN), as i.e., private company network and internet.

- *Unified Data Management (UDM)*: the role of this function is to support, inside the Home PLMN, the access to data storage. With the UDM, it is possible to control the subscription data management, access, and service authorization, user identification storage and management, user authentication, and support of Short Message Service (SMS) service.

- *Unified Data Repository (UDR)*: it is a converged repository of subscriber information and can be used to service a number of NFs. It can for example be used by the 5G UDM to store and retrieve subscription data. Alternatively, the PCF can use the UDR to store and retrieve policy related data. From a IoT domain perspective, the NEF may use the UDR to store subscriber related data that is permitted to be exposed to third-party applications.

- *User plane Function (UPF)*: it provides capabilities on top of the N4 interface through the SMF, which allows flexibility for the deployment of 5G UP features. UPFs can be geographically distributed or centralized. Moreover, the UPF supports traffic detection, traffic forwarding, data buffering, handover support, QoS enforcement, resource consumption toward the SMF via the N4 interface, and event reporting.

- *Application Function (AF)*: it is a logical element of the 3GPP Policy and Charging Control (PCC) framework which provides session related information to the PCRF in support of PCC rule generation.

- *Access Network*: it identifies a RAN that connects to the 5G CN. Examples include the 5G NR and radio systems with NR extensions.

- *5G-Equipment Identity Register (5G-EIR)*: it stores International Mobile station Equipment Identity (IMEI) numbers in order to filter unwanted handsets of the network. Handsets are "white listed", "grey listed" or "black listed" as appropriate, and may have their service revoked if they are allocated to the grey or black list. The GSM EDGE Radio Access Network (GERAN), Universal Terrestrial Radio Access Network (UTRAN) and Evolved - Universal Terrestrial Radio Access Network (E-UTRAN) systems can all use the EIR.

- *Security Edge Protection Proxy (SEPP)*: it is used to protect CP traffic that is exchanged between different 5G PLMNs. As such, the SEPP performs message filtering, policing, and topology hiding for all API messages.

To conclude this subchapter, Fig. 2.18 illustrates the 4G and 5G architecture models, and highlights how the principal functionalities are migrated inside the new architecture. Some NFs present the same role (i.e., HSS, MME, SPGW), while others are specific for the new type of services and advanced mobility techniques introduced with 5G. Even though 5G allows LTE RAT integration,

this capability requires E-UTRAN eNBs upgrade to support the interfaces defined for the 5G CN. To enable an independent deployment and evolution of the different access technologies, a CN based interworking solution between legacy core (EPC) and 5G CN is needed, as it will be presented in the next subsection.

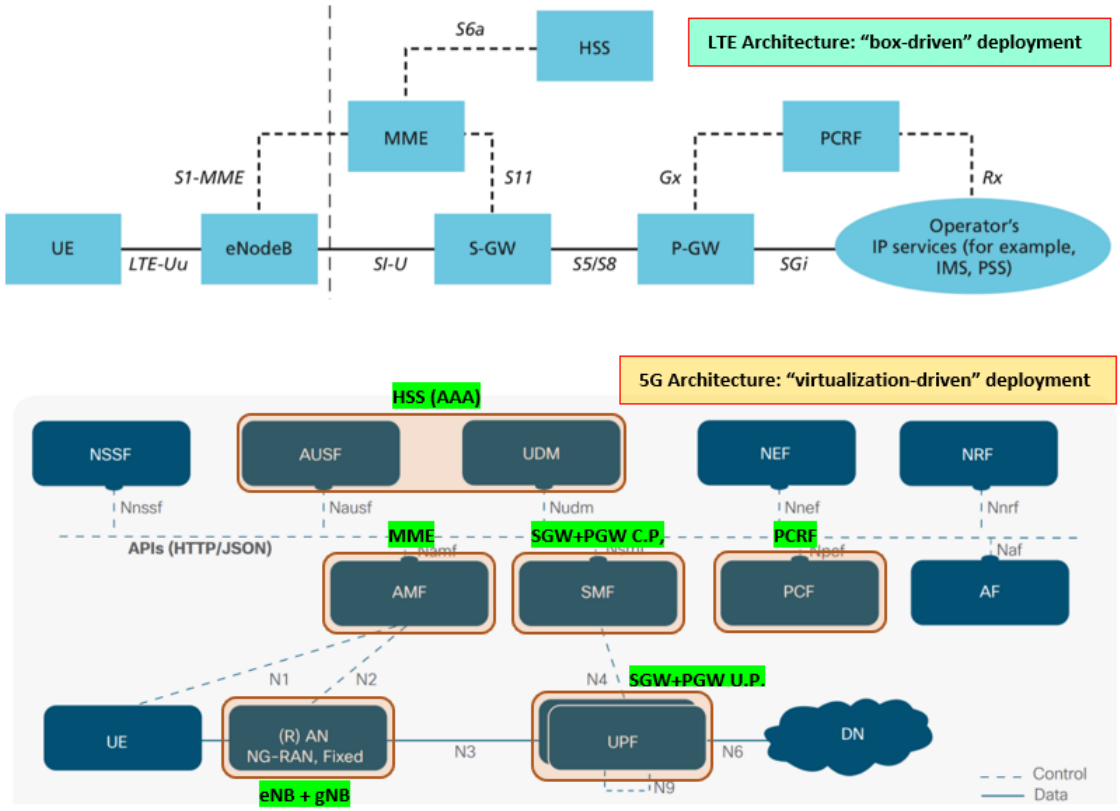


Figure 2.18: 4G-5G architectures comparison

2.5.2 5G Deployment Options

Unlike previous standards were access and CN from the same generation are paired, 5G supports the integration of elements from previous "G" generations in different configurations. The 3GPP introduces six architecture options for 5G NR deployment[17], as illustrated in Fig. 2.19, grouped into two deployment scenarios: the 5G SA options (number 1, 2, and 5) consist of only one generation of radio access technology, while the 5G NSA options (number 3, 4, and 7) consist of two generations of radio access technologies (4G LTE and 5G). However, since option 1 is a legacy LTE system, it is not considered when dealing with 5G NR deployment scenarios. This ample solutions portfolio appeases the operators'migration to the new standard, without completely dismantle the current network infrastructure.

In every NSA deployment, a node labelled Master Node (MN) is always directly connected with the 4G/5G CN, while the other (Slave Node (SN)) can optionally be connected to a CN, depending

[17]3GPP TR 38.912 version 16.0.0 Release 16.

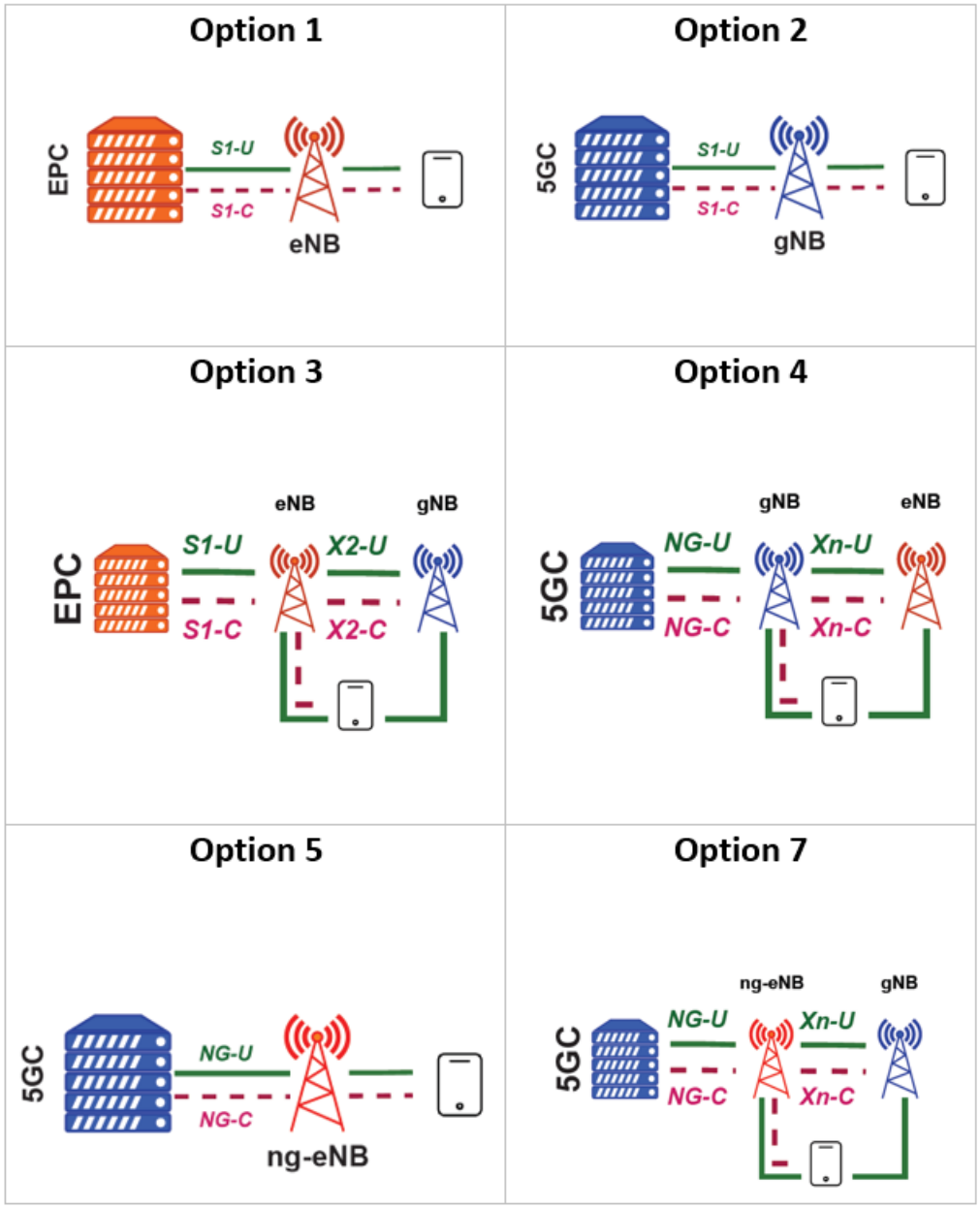


Figure 2.19: 5G NR deployment architecture options

from the option type. In most of the deployments, two distinct RAN access points provide Multi-Radio Dual Connectivity (MR-DC) to the UE, where the SN CP information are routed through the MN, while the UP information are obtained directly from each node. To follow, for each 5G option, a technical, economical, and strategic analysis is illustrated (76) - (77):

- *Option 2*: this option is suitable for greenfield 5G operators where inter-RAT migration mechanisms are used to move UEs from 4G LTE under EPC coverage and 5G NR under 5G CN coverage. This is the only option which introduces an entire 5G legacy network able to fully support new 5G services such as eMBB, mMTC, and uRLLC, and features like E2E NS and MEC. This deployment may be considered as the choice for operators who need to fulfil

the market requirements, especially of vertical industries, without operating with previous 2G/3G domains.

- *Option 3*: this configuration only requires the development of NSA 5G NR access as part of E-UTRAN connected to the EPC, in which the en-gNB is deployed in the LTE network, and thus does not need a 5G CN. This option may be preferred by operators that already have an LTE infrastructure, allowing a flexible deployment where capacity is needed using the same or different vendors for LTE and 5G NR. However, 5G services are restricted to RAN capability due to its dependency on the legacy EPC, which may represent a possible bottleneck (e.g., latency) that limits the performance that could otherwise be extracted from NR. This option is usually selected from operators which final task is the migration to option 2.

- *Option 4*: unlike option 3, the 5G CN is used to replace the EPC in serving 5G use cases, the gNB is connected to the 5G CN, and both gNB and New Generation - gNB (ng-eNB) are connected with each other. Users can take full advantage of 5G E2E network capabilities, since this option can restore the data rate competitiveness of the 5G SA customer to match a 5G NSA customer, while keeping the advantages of NR SA (78). In current scenarios, many operators have most of their sub-3 GHz spectrum dedicated to LTE with a high number of subscribers. In consequence, option 4 is the only way for them to transition to NR SA deployment in a competitive way compared with LTE or option 3.

- *Option 5*: in this option, the network has transitioned toward the 5G CN, but still continues to use LTE access, in which the ng-eNB is connected without dual connectivity with NR systems. The eNB is upgraded in order to interwork with the 5G CN, enabling 5G functionalities such as NS. Even though it represents an important step towards 5G functionalities, this does not exploit the benefits of 5G NR air interface such as mmWave, multiple numerologies, and flexible frame structure.

- *Option 7*: this option combines the advantages of Option 5 with the added benefit of MR-DC allowing data aggregation with any co-existing 5G NR carriers to improve throughput. Both the eNB and gNB are connected with each other, and it represents the natural evolution of option 3, where 5G network performance and coverage are obtained connecting to a 5G CN.

Table 2.5 summarizes the main advantages and drawbacks of each deployment options, while extra sub-deployment configuration is documented in (79).

Table 2.5: Deployment options details

Option	**Configuration**	**Advantages**	**Disadvantages**
Option 1	CN: EPC RAN: SA	RAN: - legacy deployment CN: - legacy deployment - leverage existing - EPC deployment	RAN: -hardware-driven -service limited CN: -hardware-driven -service limited
Option 2	CN: 5GC RAN: SA	RAN: -simple to manage inter-generation handover between 4G-5G CN: -cloud native - easier to support multiple access	RAN: - not able to leverage existing LTE deployments if NR is used in SA CN: - new deployment required
Option 3	CN: EPC RAN: NSA	RAN: -leverage existing LTE deployments CN: -leverage existing EPC deployment	RAN: -tight interworking, between LTE and NR required -may impact end user experience CN: - cloud support in optional

Option 4	CN: 5GC RAN: NSA	RAN: -leverage existing LTE deployments CN: -cloud native easier to support multiple access	RAN: - tight interworking, between LTE and NR required - may impact end user experience CN: - cloud support in optional
Option 5	CN: 5GC RAN: SA	RAN: -simple to manage inter-generation handover between 4G-5G CN: -cloud native -easier to support multiple access	RAN: -not able to leverage existing LTE deployments if NR is used in SA CN: - new deployment required
Option 7	CN: 5GC RAN:NSA	RAN: -leverage existing LTE deployments CN: -cloud native -easier to support multiple access	RAN: - tight interworking, between LTE and NR required - may impact end user experience CN: - new deployment required

Chapter 3

Experimental Platforms: 4G and 5G SA Testbeds

In this chapter, the configuration and components of the two testbeds used for our experiments are presented. At the initial stage, due to HW and SW limitation, our dynamic slicing algorithms have been tested on top of a LTE architecture, equipped with 5G features, following the virtualization principles and specifications defined for 5G from 3GPP. Through this fully virtualized environment, the migration towards a fully 5G SA\NSA platform has been straightforward, simplifying the integration of novel enablers and third-party functionalities.

3.1 OpenAirInterface

OpenAirInterface (OAI) (80) is an open experimentation and prototyping platform created by the Mobile Communications Department at EURECOM[1] to enable innovation in the area of mobile/wireless networking and communications. OAI distinguishes itself from other similar projects through its unique open-source license, the OAI public license v1.1, which was created by the OAI Software Alliance (OSA) in 2017. This license is a modified version of Apache v2.0 License, with an additional clause that allows contributing parties to make patent licenses available to third parties under Fair, Reasonable and Non-Discriminatory (FRAND) terms similar to 3GPP for commercial exploitation. The usage of OAI code is free for non-commercial/academic research purposes. Such a license allows contributions from 3GPP member companies while at the same allowing commercial exploitation of the code, which is not at all possible with other open-source projects (81).
With OAI, the transceiver functionality (of a base station, access point, mobile terminal, CN, etc.) is realized via a SW radio front end connected to a host computer for processing; this approach is similar to other widely used Software-Defined Radio (SDR) prototyping platforms in the wireless

[1] https://www.eurecom.fr/en/home

networking research community such as SORA (82) -(83). Two unique features of the OAI platform are:

- open-source SW implementation of the 4G mobile cellular system that is fully compliant with the 3GPP LTE standards and can be used for real-time indoor/outdoor experimentation and demonstration.
- Built-in emulation capability that can be used within the same real execution environment to seamlessly transition between real experimentation and repeatable, scalable emulation. Specifically, two PHY emulation modes are supported which differ in the level of detail at which PHY is realized.

Before diving into the next subsection, Fig. 3.1[2] illustrates the OAI protocol stack architecture running over general-purpose platform such as Intel/ARM.

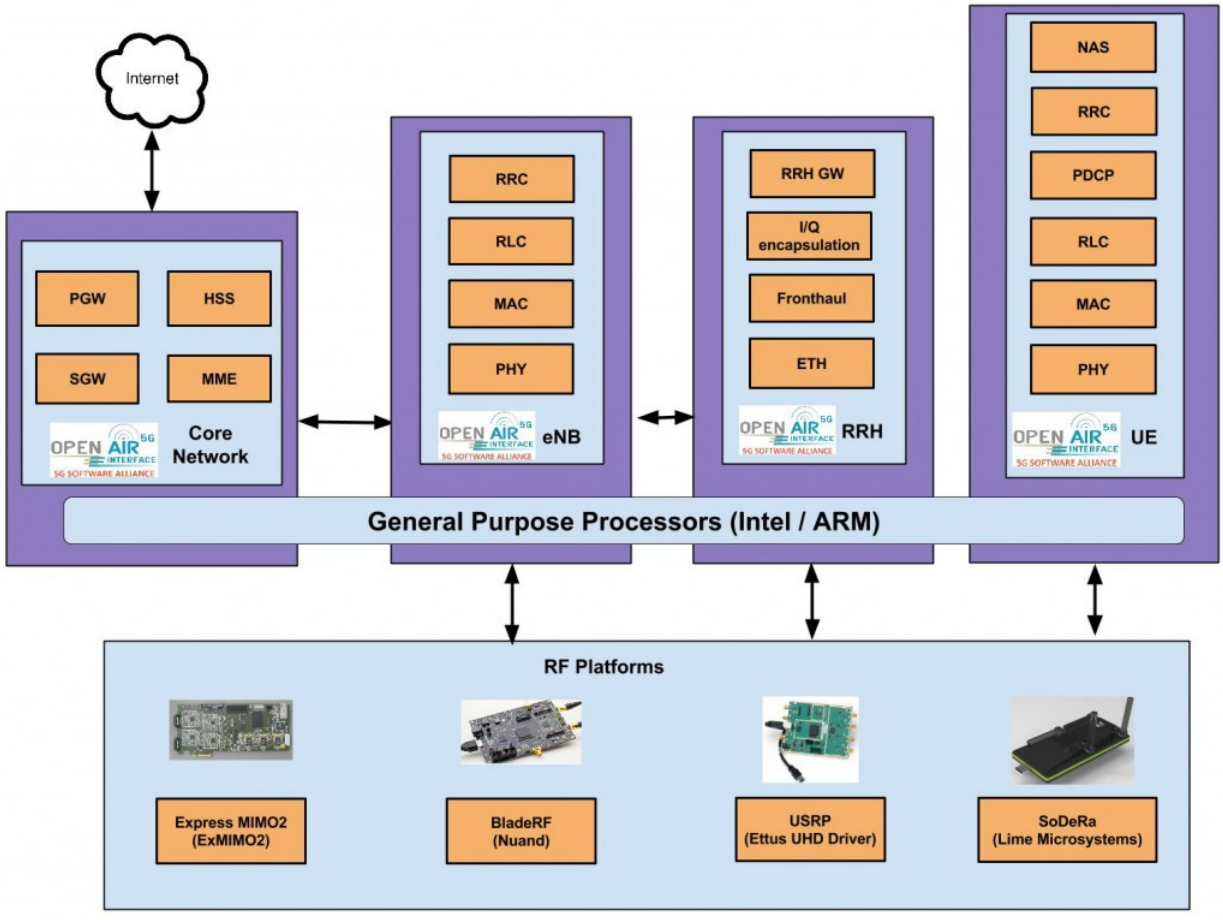


Figure 3.1: OpenAirInterface protocol stack architecture

[2]https://openairinterface.org/community/whitepapers/towards-building-a-5g-ue-demonstrator-using-openairinterface/

3.1.1 OpenAirInterface and RAN Vision

The scope of the OAI 5G RAN project is to build the 5G protocol stack for both gNB and UE allowing for E2E deployment of a 5G network. Its objective consists to develop and provide the 5G NSA RAN SW and enable connection and traffic flow with an NSA-capable 5G commercial UE. In the 5G NSA setting the gNB is supplemented by the LTE eNB that carries the CP of the 5G signaling while the data bearer is set up on the gNB. An NSA capable 3GPP Rel-15 4G EPC is connected through the S1 interface to eNB and X2-C interface enables connection between the eNB and gNB for routing and managing the flow of IP traffic (84).
Starting from 2020, three phases have been defined for the OAI RAN deployment, as illustrated in Fig. 3.3:

1. Phase I: also defined as "noS1", this phase has a similar implementation to the 4G eNB-UE OAI testbed, where gNB and UE work without the existence of a CN.

2. Phase II: using as CN the EPC from the 4G OAI testbed, this solution implements a fully 5G NSA architecture, as illustrated in the 5G deployment option 3.

3. Phase III: a fully 5G SA deployment is expected in this phase. The gNB will connect to a legacy 5G Core implementation.

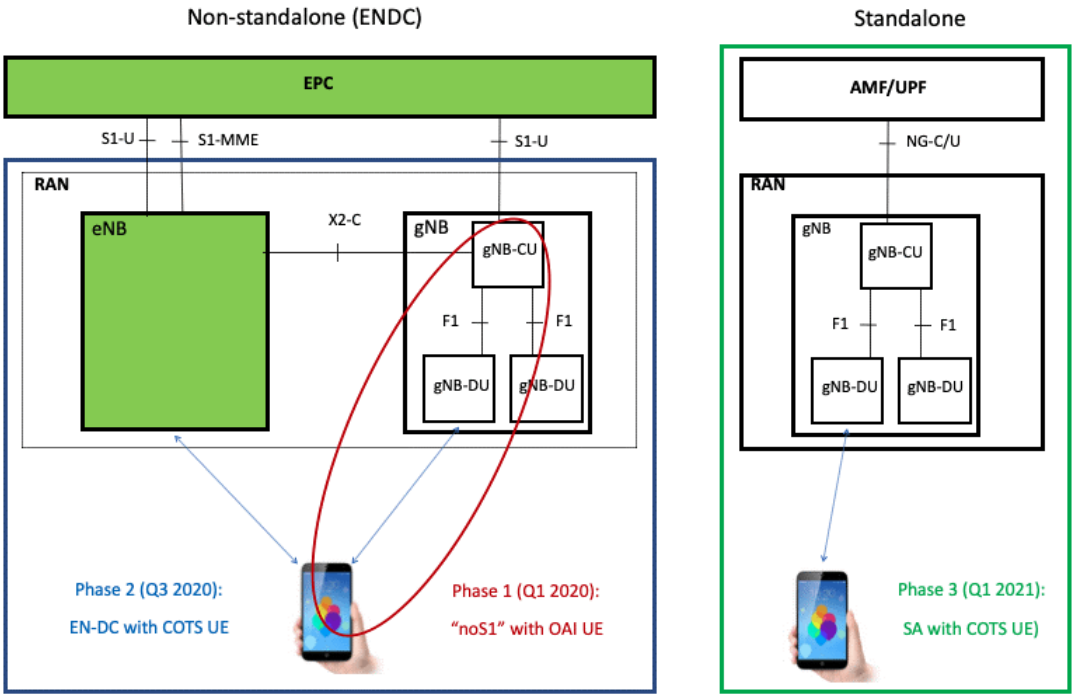


Figure 3.2: OAI RAN deployment phases

3.2 4G Testbed Design and Deployment

In this section, we describe the main features of our testbed configuration and the motivation behind our system choice implementation. Since a substantial contribution for the management of resources through slicing techniques in the BH and MH networks already exists in the literature, we focused our attention on the FH part, because dynamic RAN NS has attracted a strong attention especially in the last years. Fig. 3.3 shows the structure of our testbed used to perform dynamic NS in the FH network, based on OAI, where the division of NFs into independent SW components further emphasized as NFV represents a dominant technology in future networks.

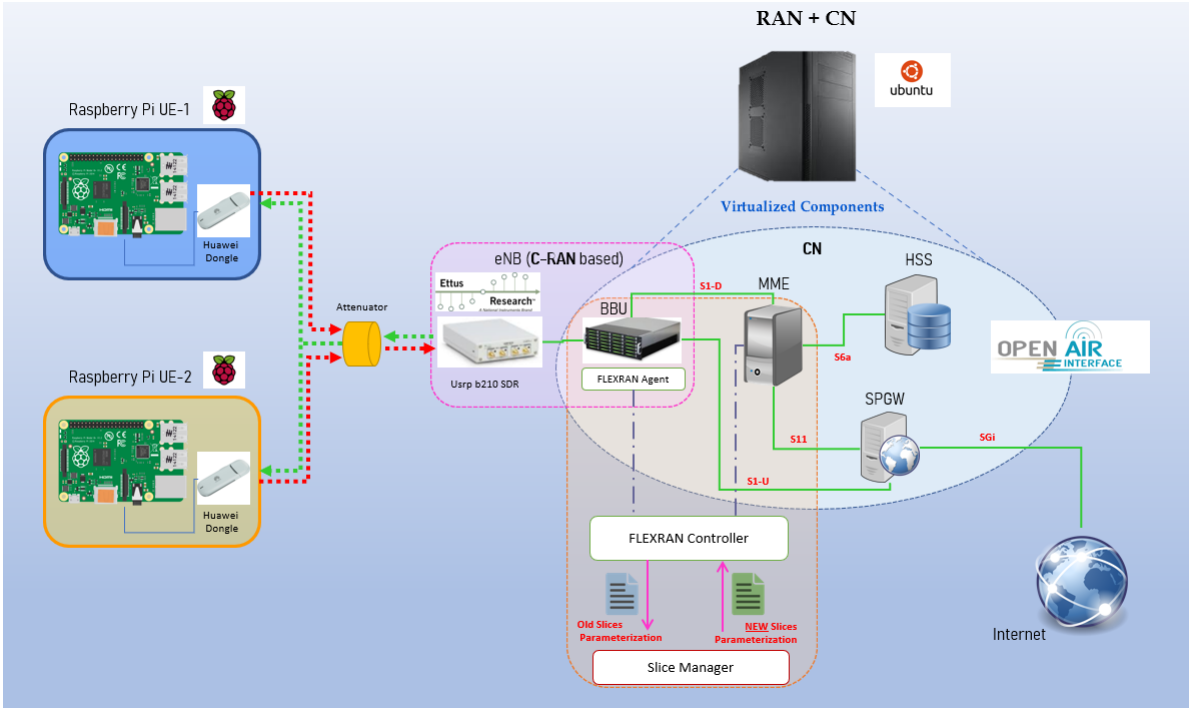


Figure 3.3: 4G Testbed architecture

In a real system, this allows the placement of each entity in a different node of the network, significantly reducing the installation and maintenance costs present in today's systems. To support multi-tenant capabilities, the synchronization of each entity is provided by standardized interface based on IPv4 and IPv6 protocols. The radio system used is Universal Peripheral Radio Software (USRP) B210, from Ettus Research (85), able to provide up to 56 MHz of real-time bandwidth. In the default scenario, the testbed handles UEs based on Raspberry Pi 4 Model B[3], equipped with a 4G/LTE HAT board[4], running over RaspBian Operating System (OS)[5], and a Huawei Dongle[6] antenna for the interface toward the eNB. To enable the interaction of OAI RAN with a third-

[3]https://www.raspberrypi.com/products/raspberry-pi-4-model-b/
[4]https://sixfab.com/product/raspberry-pi-base-hat-3g-4g-lte-minipcie-cards/
[5]https://www.raspberrypi.com/software/
[6]https://consumer.huawei.com/en/routers/e3372/

party SW, OAI FlexRAN[7] SDN-based platform has been designed (86). As illustrated in Fig. 3.4, the FlexRAN platform is made up of two main components: the FlexRAN Service and Control Plane, and FlexRAN Application plane. The first follows a hierarchical design and is composed of a Real-time Controller (RTC) connected to a number of underlying RAN runtime, one for each RAN module (e.g., one for monolithic 4G eNB, or multiple for a disaggregated 4G and 5G). The latter can be developed both on the top of the RAN runtime and RTC Software Development Kit (SDK) allowing to monitor, control, and coordinate the state of RAN infrastructure.

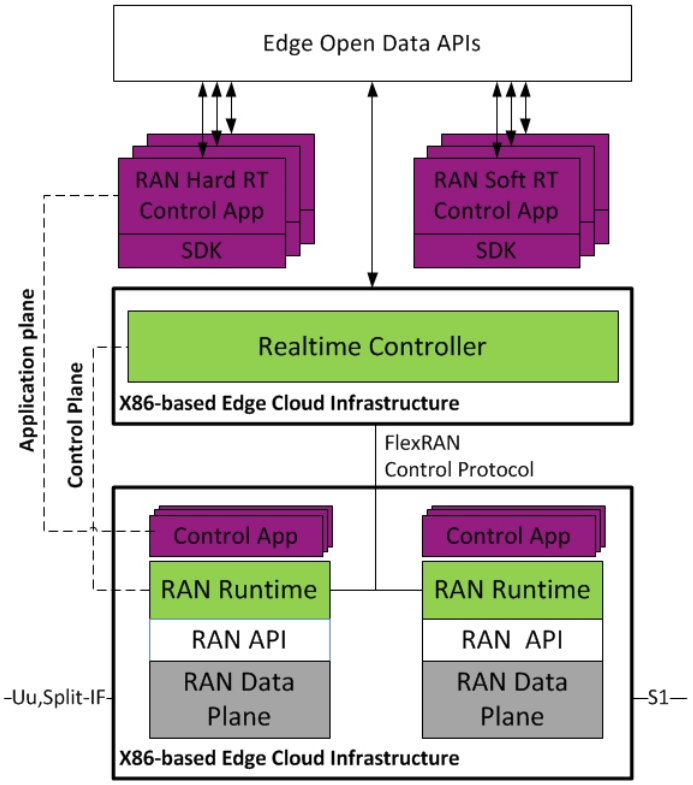


Figure 3.4: High level schematic of the FlexRAN platform

Using FlexRAN APIs as input/output interface with the OAI RAN platform, we developed a new NF component called *Slice Manager (SM)*. Through the FlexRAN RESTful APIs, the SM acquires real-time network information of each RAN slice, elaborates a new parameterization, and post a new slices configuration in the system, following the 5G 3GPP compliant specifications.

Fig. 3.5 illustrates different 4G testbed architectural options supported by OAI, while Table 3.1 collects the main system parameters. With the current configuration, two main OAI implementations (simulation or emulation), disaggregated and centralized are supported, according to the HW capabilities and tested scenario.

It is important to remark that our proposed solutions are backward compatible with existing 3GPP 5G protocol stack, since they utilize standardized interfaces and functionalities to communicate with

[7]https://mosaic5g.io/flexran/

Table 3.1: 4G testbed components

Type	Value
OS	Ubuntu 16.04.6 LTS (Xenial Xerus) 64 bits
RAM	8 Gb
UE antenna	Huawei E3372
UE HW	Raspberry Pi model B
UE OS	Raspbian
eNB SDR	USRP B210 SDR - Dual Channel Transceiver (70MHz-6GHz) Ettus Research
Radio splitter	LTE band7 Duplexer

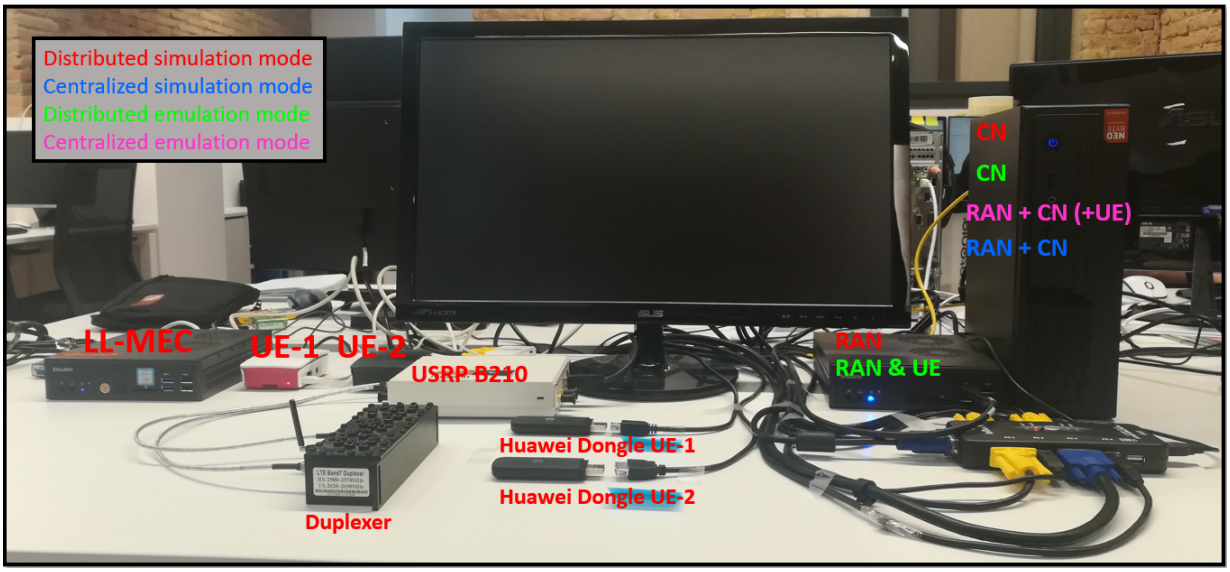


Figure 3.5: 4G testbed HW and deployment options

the main network architecture elements. Moreover, even though our testing platform is mainly focused on dynamic RAN NS, using a suitable set of input data and output configuration files settings, it is possible to extend our SM orchestrator to other network sectors (MH and BH), without downgrade the optimization capabilities of our models.

Table 3.2: 5G testbed parameters

Type	Value
CPU	Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz
RAM	16 Gb DDR4 up to 3200MHz
ROM	240 GB SSD
NIC	Intel X722 series. 2x 10 Gbps SFP+ ethernet interface
OS	Ubuntu 18.04 LTS low latency kernel
UHD driver versions	3.15/4.0
Optional	Data Plane Development Kit (DPDK)

3.3 5G Testbed Design and Deployment

Fig. 3.6 illustrates the OAI 5G "noS1" configuration using OAI gNB and OAI new radio UE (nrUE), following a full-cabling implementation[8]. At the time of testing, 5G NR OAI was only a prototype with limited capabilities (no NS support). For this reason, this subsection provides an early overview of the SW and HW components required for an initial 5G setup without COTS UE, which will be easily extended in future with the latest OAI 5G NR SW releases.

Unlike the 4G case, NI USRP N310[9] are used on both sides (gNB and UE): this SDR is one of the highest channel density devices in the SDR market, offering four RX and four TX channels in a half-wide RU form factor. The RF front end uses two AD9371 transceivers, the latest RFIC technology from Analog Devices[10]. Each channel provides up to 100 MHz of instantaneous bandwidth and covers an extended frequency range from 10 MHz to 6 GHz. The processing requirements for 5G NR are much higher than for 4G, so a high-end PC or server is needed.

In our implementation, Intel Supermicro servers[11] are configured on both gNB and UE sides to overcome the strict latency and high processing requirements of the 5G OAI protocol stack. On each SDR, dual 10Gb streaming Small Form-factor Pluggable + (SFP+) port 0/1 is enabled, together with the "XG" FPGA image. Finally, for enhanced synchronization performance, NI OctoClock[12] guarantees high-accuracy time and frequency reference distribution. Table 3.2 summarizes the main technical HW and SW requirements of our configuration.

[8]This work has been co-supervised with National Instruments GmbH (Dresden, Germany), as part of a secondment plan during the project implementation.

[9]https://www.ettus.com/all-products/usrp-n310/

[10]https://www.analog.com/en/products/ad9371.html.

[11]https://www.supermicro.com/en/products/xeon-e

[12]https://www.ettus.com/all-products/octoclock/

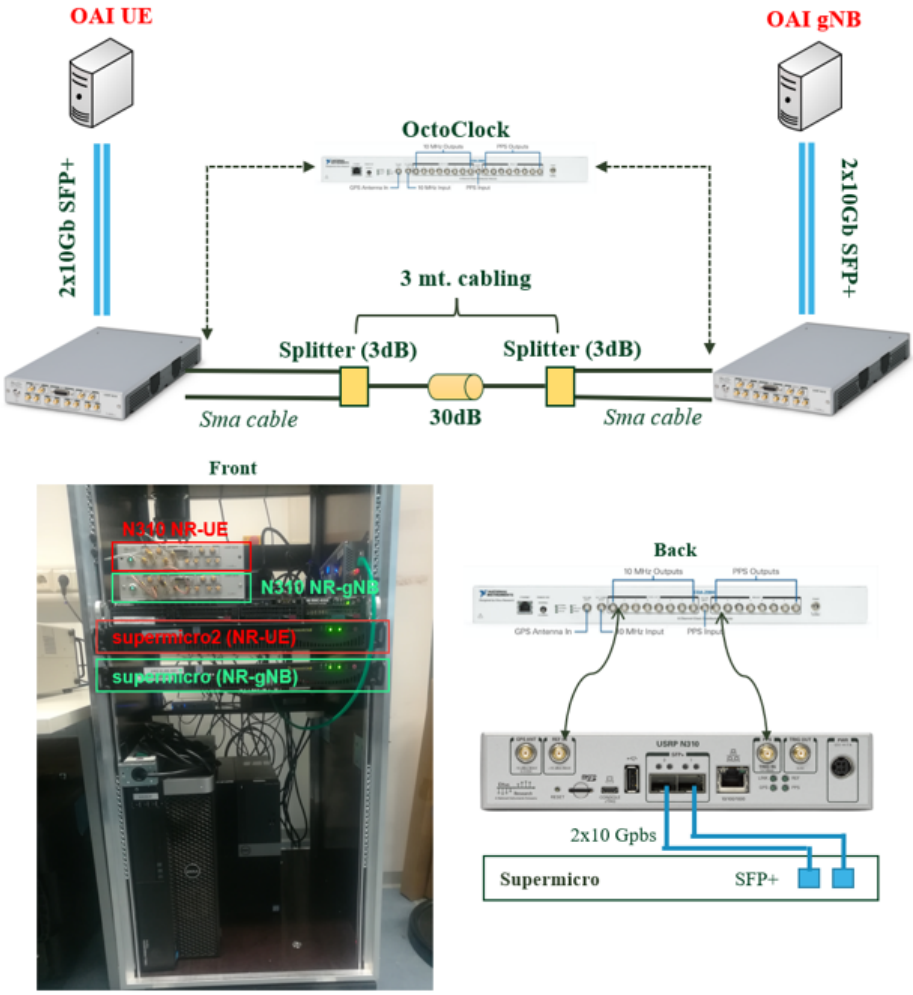


Figure 3.6: OAI 5G "noS1" testbed architecture

Chapter 4

Dynamic Network Slicing SDN-based Orchestrators: Proposed Solutions

In this section, four main achievements of this research work on novel dynamic NS solutions are proposed and detailed illustrated. Using as baseline our SM SDN-based controller mentioned in Subsection 3.2, each proposed solution represents an advanced version of the previous one, since new features and more advanced optimization techniques are gradually applied to the SM. Different system scenarios and flexible services are tested to validate our models and features, while the performance are benchmarked against current state of the art solutions.
Each proposed solutions aims to highlight the advanced resource sharing and orchestration capabilities introduced with 5G, in a multi-service and multi-tenant NS environment.

4.1 Real-time Slice Manager Framework for 4G-5G RAN: a Selfish MNO-driven Approach

4.1.1 Motivation

With the fusion of SDN and NFV technologies, the new 5G architecture is expected to support a wide range of services and requirements through NS. With SDN, a network is capable to manage the traffic requirements dictated by the new forms of distributed processing, while NFV introduces the concept of virtualization of NFs, translating HW-based appliances into high volume servers, switches and storage. In the current state of the art, NS is mainly focused on a static NS approach

This contribution of this chapter has been published in the following publication: Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., & Verikoukis, C. (2019, December). Real-time dynamic network slicing for the 5G radio access network. In 2019 IEEE Global Communications Conference (GLOBECOM) (pp. 1-6). IEEE.

since the current network architecture model does not fully support NFV and SDN techniques necessary for the correct establishment of the slices. In (87), a prototype of an E2E NS testbed is proposed, where the optimal slice parameterization is discussed, together with the limitations of static slicing. More sophisticated techniques based on ML and AI have been proposed for the derivation of the optimal dynamic NS approach. In (88), the authors formulate the dynamic slicing problem as a Mixed Integer Linear Programming (MILP) problem. Also, in (89), authors focus on the Deep Q-Learning technique for slicing resource management while in (90) the authors propose a new dynamic NS scheme based on BBU capacity allocation and PRBs management.
According to our analysis, even though these approaches provide the optimal solutions for dynamic slice selection and configuration, they are far from a physical application in a real scenario, where the communication channel is subjected to sudden variations that can significantly compromise learning models. In this first solution, we present a dynamic NS approach whose objective is to minimize the over-provisioning of resources by guaranteeing the optimal requirements of each service requested by the users. The innovative part of our solution consists in the real-time slice parameterization allocation according to the traffic requirements and the management of unpredictable variation due to the aleatory behavior of the wireless channel. The obtained results illustrate how the correct resource allocation can maintain high QoS while increasing the number of users served. Hence, our contribution is as follows:

- First, we provide an innovative dynamic model for slice management able to filter and analyze in real time the traffic parameters (packet received, packet transmitted, delay, buffer queue size, error rates) of each data flow, and setup the slice considering the minimum number of Resource Blocks (RBs) necessary to guarantee the correct service supply.
- Second, we define a real scenario within our testbed, where two flows are appropriately parameterized to simulate a slice with multiple video traffic (steams with different video quality), and another slice with mobile broadband traffic (messaging, email, updates).
- Third, we test our dynamic slicing model alongside a static slicing baseline model on top of our testbed over the same scenario, in order to compare their performance and extract useful insights.

For more details about the proposed solution, Appendix A illustrates a eHealth use case scenario where the orchestration of the FH resources is done following the approach illustrated in this section.

4.1.2 Real-time Dynamic NS Algorithm

In this section, we present our dynamic NS algorithm in detail. The algorithm description is divided into three subsections: i) the user acceptance and service requirement analysis, ii) the statistical

analysis and slice configuration, and iii) the background real-time dynamic system optimization for the optimal resource balancing among the slices.

A *System initialization and user admission*

This part describes the 5G network architecture initialization and the acceptance procedure of the users, as it is illustrated in Fig. 4.1. In the initialization part, the operator registers the UE parameters (i.e., IMSI, PLMN, APN, Operator Key (OP)) in the authentication database, which is then communicated to the 5G CN block responsible for the management of the tenant subscriber. Each block of the 5G CN owns a specific functionality of the 5G

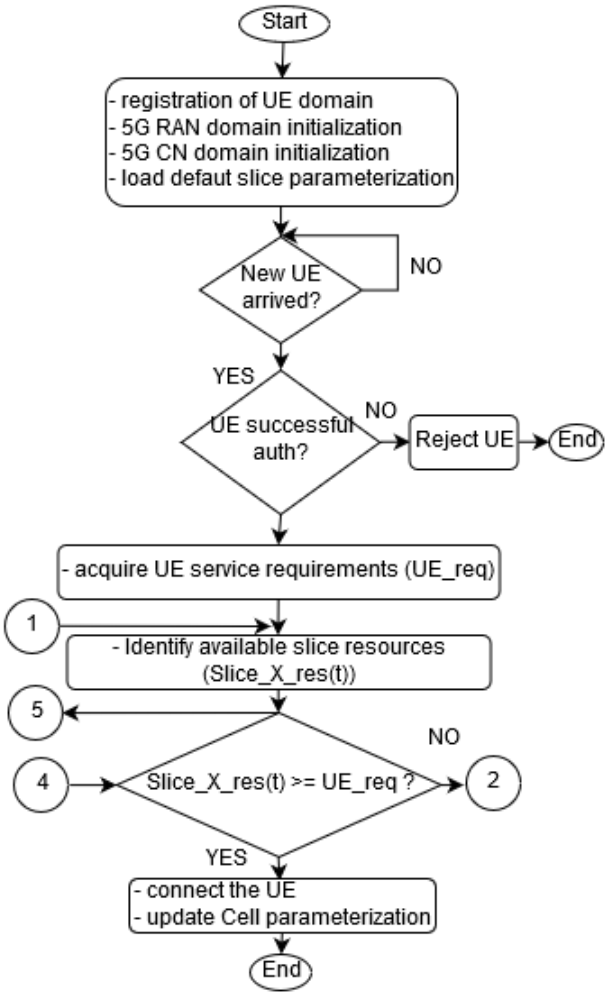


Figure 4.1: Dynamic slice algorithm: UE acquisition

CP and DP, which communicates among each other through specific network interfaces and protocols. Once they are active and synchronized, the 5G RAN is initialized and connected to the 5G CN. In our approach, at this point, the system communicates an initialization slice parameterization to the 5G CN and 5G RAN where the RBs are equally divided. The SDN controller, whose role is to monitor the status of the underlying network, is monitoring the RAN and CN parts, collects network statistics and posts new slice configurations when

required. The gNBs are equipped with an SDN agent for the signaling between the SDN controller and the gNB CP. On top of the CP, the SDN controller and the SDN agent share the radio parameters such as frequency, number of RBs, transmission gain, etc., that are necessary for the configuration of the SDR. When a UE sends a connection request to the gNB, the authentication procedure is activated. If its authentication parameters match the configuration of the 5G subscriber authentication function, the UE is accepted, otherwise, it is refused. As a first step, each accepted UE activates an initialization messaging phase with the gNB, where the main traffic requirements are communicated such as the traffic priority, average packet size, maximum packet delay, isolation restrictions and type of service (UE_req). This information is utilized by the 5G RAN and 5G CN to recognize the slice for the UE, and quantifies the resources needed ($Slice\_X\_res$) for the correct service supply. If the slice resources guarantee a proper service, the UE requirements are satisfied, otherwise, the system activates a reallocation resource procedure, as it will be explained in the next part.

B *Statistical analysis and slice configuration*

In order to reallocate the RBs among the slices, we designed a SM orchestrator, which operating principle is illustrated Fig. 4.2. This procedure is activated by our SM when a slice does not fulfill the service requirements for the incoming UE. Moreover, the SM system communicates with the RAN and CN domains through the 5G CP specifications, compliant with 3GPP standard. For each slice belonging to the gNB, the SM determines the amount of RBs ($Slice\_X$). If at least one slice has free available RBs ($Slice\_X\_free\_res$), the SM monitors if the inter-slice functionality option is enabled, otherwise the UE is redirected to another gNB (point 3). When the inter-slice functionality is active, the SM extracts a certain amount of RBs from each slice according to a specific slice weight ($Slice\_X\_w$), until the sum of unused RBs from each slice is greater or equal to the minimum UE SLAs. Another advanced functionality is the intra-slice sharing. If active, the SM reconfigures the RBs and the scheduling procedure among the UEs served within the same slice. The new parameterization obtained from the reallocation procedure is translated in a 5G CP compliant file format, which is sent from the SM to the SDN controller. The SDN controller communicates the new settings to the corresponding gNB SDN agent that applies the new system changes.

C *Background real-time dynamic system optimization*

The proposed innovative method for the optimal parameterization of the slices is illustrated in Fig. 4.3. According to a specific granularity (i.e., one frame length), the output of this method is the optimal slice parameterization by taking into account:

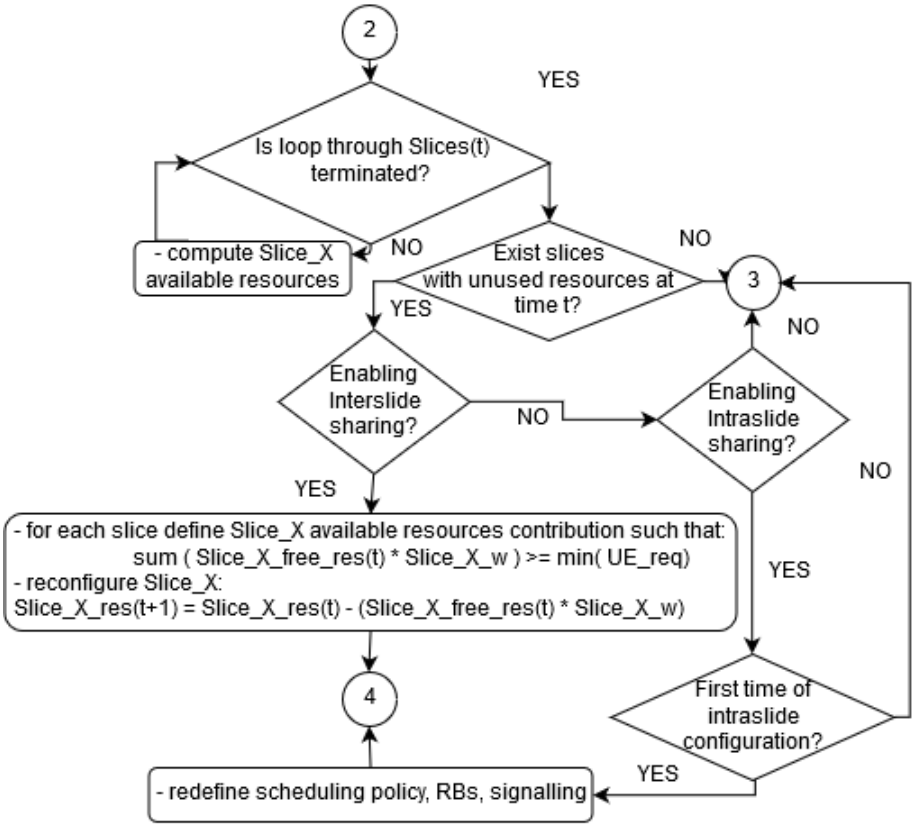


Figure 4.2: Dynamic slice algorithm: slice configuration

- the ongoing services,
- the unpredictable traffic variations,
- the release of RBs from UEs that completed the service session,
- the changes of the service type (i.e. from eMBB to URLLC) during an ongoing transmission for the same UE.

Moreover, this algorithm, which is implemented inside the SM, is executed in background mode in order to reduce the CPUs load.

When the procedure is enabled (point 5 in Fig. 4.3), the SM adjusts the number of RBs until the estimated slice throughput is as close as possible to the measured real slice datarate. This technique guarantees the allocation, according to the system granularity, of the optimal amount of RBs to each slice (eMBB, uRLLC, mMTC), without the isolation of unused RBs. This method allows our system to be always equipped with the optimal configuration in line with the services evolution. As a consequence, the delay due to the RB reconfiguration when a new connection request arrives is reduced, and a homogeneous resource distribution is applied among the slices.

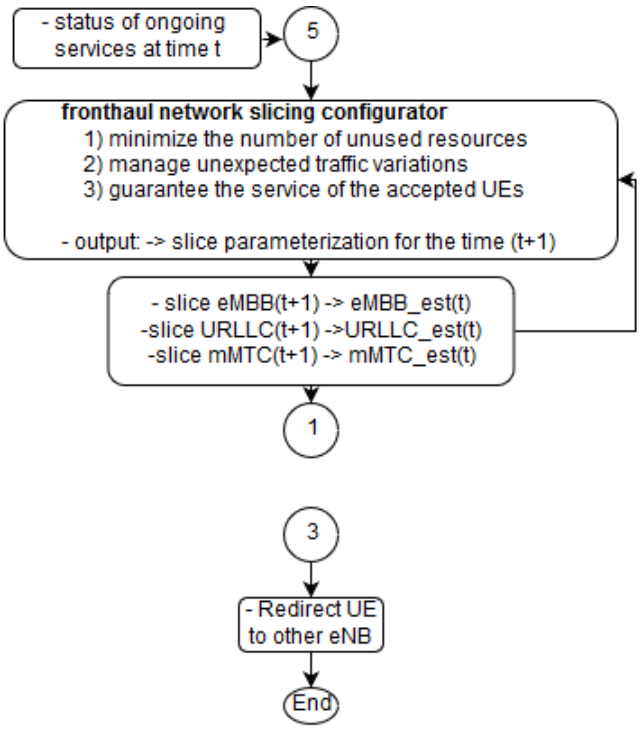


Figure 4.3: Dynamic slice algorithm: system runtime optimization

4.1.3 Experimental Testbed Configuration

Fig. 4.4 shows the structure of our testbed used to perform dynamic NS in the FH network, based on OAI, following the setup illustrated in Subsection 3.2.

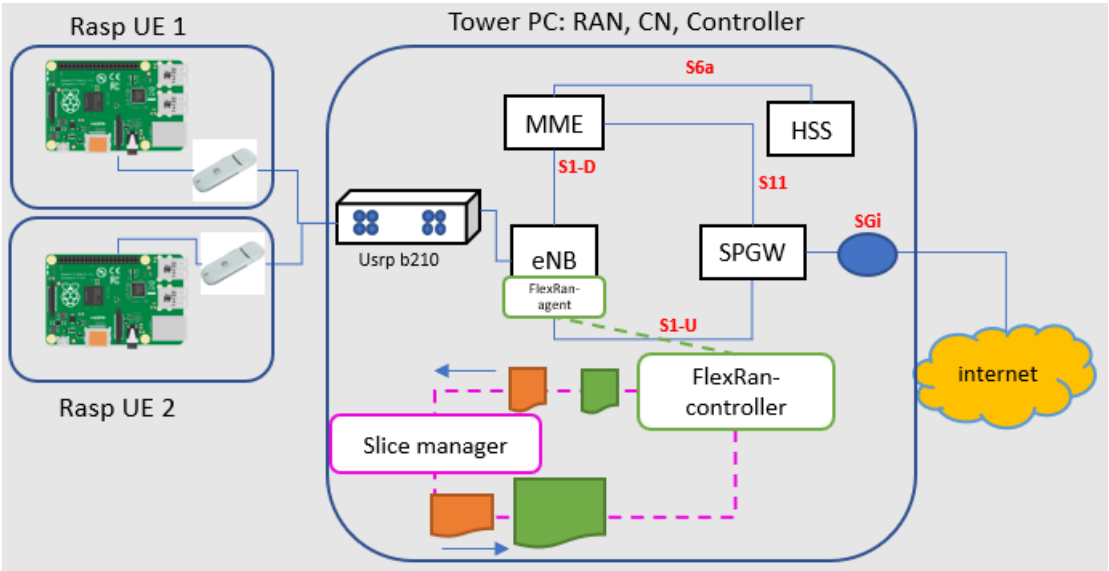


Figure 4.4: FH testbed architecture

4.1.4 System Parameterization and Performance Evaluation

In our experiments, two UEs subjected to different traffic requirements are tested on top of the previous discussed testbed. To evaluate our system under constant optimal conditions, the Channel

Quality Indicator (CQI) is fixed to 15, while the Modulation Coding System (MCS) to 28. The total bandwidth is 20 MHz, in FDD mode, allowing a total of 50 RBs for each direction. To provide a realistic scenario, we defined a variable packet load (from 500 to 1200 bytes) and the variable type of traffic (UDP, ICMP), taking into consideration different types of inter-arrival packet distributions. The fundamental parameters of our implementation are illustrated in Table 4.1.

Table 4.1: Experimental scenario parameters

Variable	**Value**
Frame type	TDD
Downlink frequency	2650 MHz
Uplink frequency	2530 MHz
Downlink bandwidth	10 MHz
Uplink bandwidth	10 MHz
Downlink RB	50
Uplink RB	50
Nr. UE	2
Packet size	500-1200 bytes
CQI	15
MCS	28
Inter-departure packet distribution	Constant, Poisson
Packet protocols	UDP, ICMP

Using the aforementioned implementation, we performed different types of experiments: i) first, we emphasize the importance of a correct configuration of the slices, and ii) we compare our algorithm against static NS. The first experimental scenario performance is illustrated in Fig. 4.5-a, where the maximum estimated throughput (blue line) is compared against the real measured throughput (red line), assuming a constant traffic trend during the entire transmission. In parallel with this traffic, Fig. 4.5-b shows the number of RBs assigned to the slice along with the packet delay metric. With a configuration equal to 100 percent of the available resources (50 RBs), the service is abundantly guaranteed (around 94 Kbps) and much greater than the minimum throughput required (8-9 Kbps). With this configuration, a significant number of RBs are assigned, but they are not fully used. To optimize the system, it is necessary to refine the slice RBs percentage so that the maximum estimated throughput presents a trend close to the real measured throughput. The SM defines a variable *security threshold* in order to keep the estimated throughput always slightly higher than the real throughput. This functionality is introduced to handle channel fluctuations and queues saturation during the service transmission. The importance of this parameter is highlighted in the red section of Fig. 4.5-a: with a small *security threshold*, the estimated throughput is almost equal to the measured throughput, and the packet delay increases as there are not enough resources available to handle the queues. The configuration of resources between slices is particularly important in the case of scenarios where multiple slices coexist with time variant services.

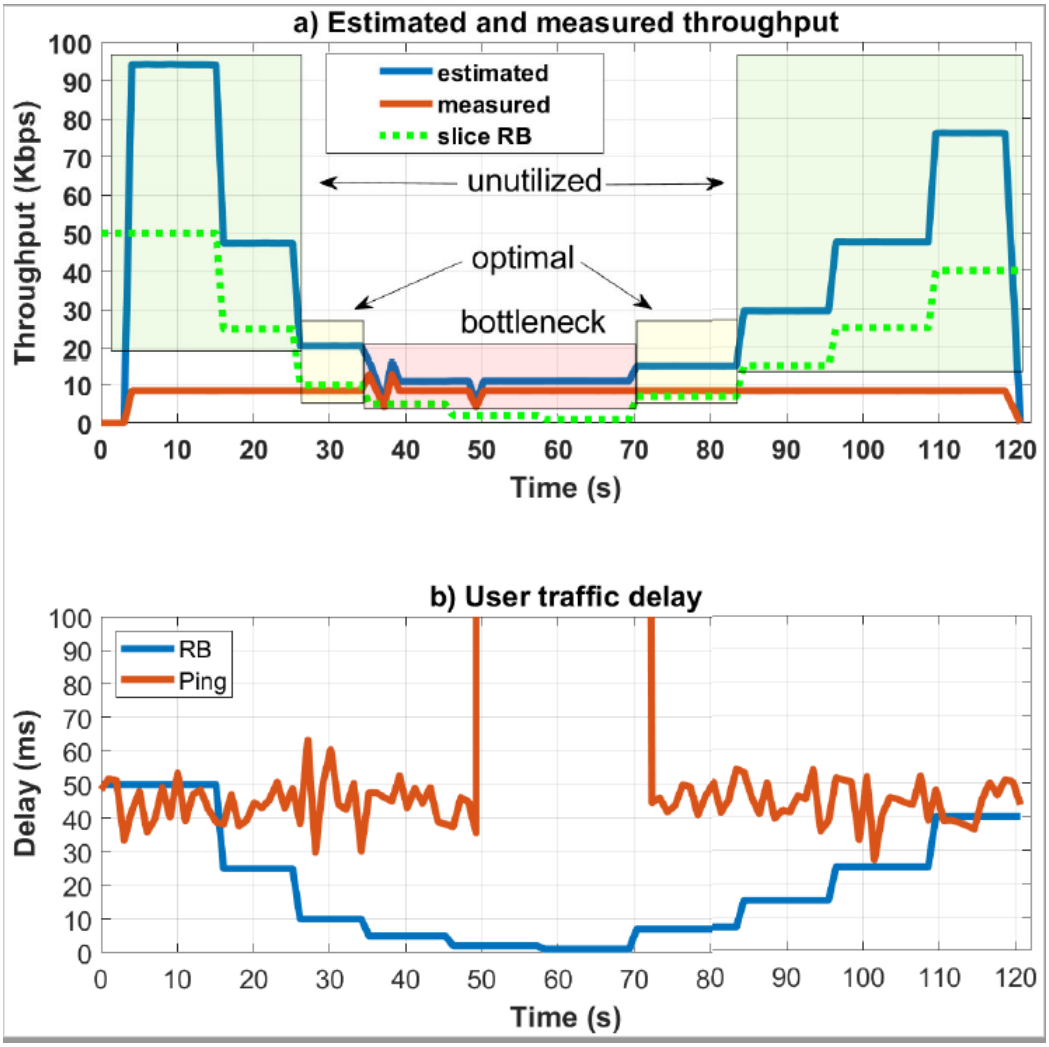


Figure 4.5: Performance comparison: a) estimated throughput vs. real throughput and b) packet delay vs. number of resource blocks assigned

Fig. 4.6 illustrates the performance of a two slices scenario, one with constant low-bitrate traffic (blue line), and the other with a time-varying burst traffic (orange line). For each of these traffic patterns, and Fig. 4.7 shows the corresponding RBs assignation, comparing the static and dynamic NS approaches. In order to reflect a realistic scenario, the low bitrate slice emulates 5G broadband mobile traffic, while the burst slice emulates video streaming, where packet buffer is filled multiple times during the transmission. In Fig. 4.7, the dashed lines represent the ideal configurations to guarantee the services in the same scenario under the static slicing method. The limited flexibility of the static approach forces to configure each slice according to the maximum throughput requested by a service. This involves a high allocation of resources especially in the case of variable traffic in the network. In Fig. 4.7, the application of the dynamic approach to the variable burst traffic slice guarantees the SLAs of the user's served, while reducing the use of RBs up to 23 percent. The granularity between consecutive new configurations of a slice parameters depends from the bit rate of the served traffic. In this experiment, it was possible to assign a new parameterization every

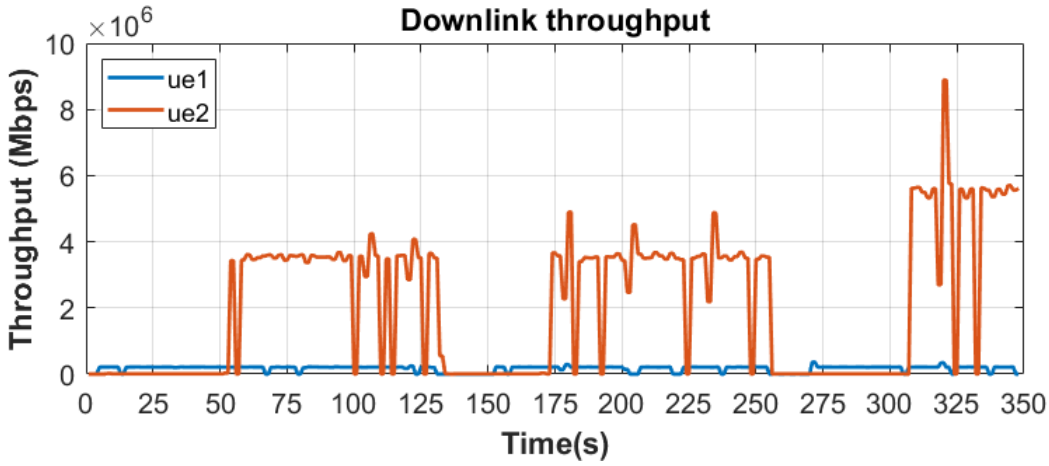


Figure 4.6: RB traffic-based assignation

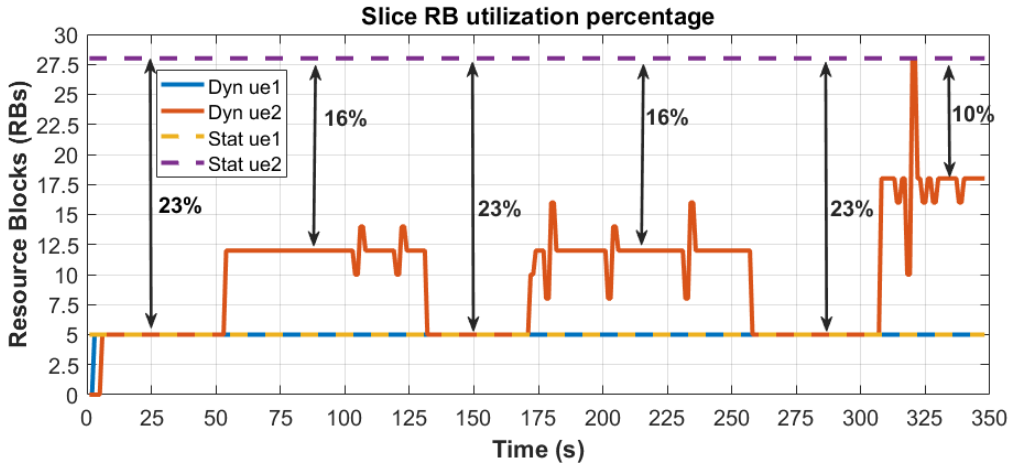


Figure 4.7: Comparison static slicing versus dynamic slicing

10 milliseconds, like the LTE frame size. This level of granularity permits to optimally handle the traffic variation, avoiding congestion in the queues and reducing packet delay. In the case of constant low bitrate traffic (blue line), a dynamic approach does introduce further performance improvements than the static method. This result shows a case where it is preferable to manage the traffic with the static approach, and to keep the dynamic configuration for the management of variable and complex traffic. Regarding the dynamic slices, Fig. 4.6 illustrates the throughput performance achieved with our solution. The correct parameterization of the slices, explained in Fig. 4.7, is highlighted in this figure, where the measure throughput follows a trend proportional to the number of RBs assigned to each slice. Burst traffic is subject to multiple peaks (up to 9 Mbps) due to queue adjustments in the scheduler or retransmission of a large number of packets due to a channel nature issues. These unexpected variations do not damage the traffic of other users, as the resources are appropriately divided and independent among the slices.

4.2 Cooperative Multi-slice Intra-Tenant 5G NR Slicing Framework

4.2.1 Contribution and Innovation

This subsection presents a novel NS resource distribution methodology where the sharing degree of each slice is evaluated according to the type of services, maximum slice capacity, QoS requirements, SLA, and ad-hoc isolation policies. This solution aims to optimize the resource management in the 5G NR part of the network, create a new level of flexibility where the slice settings are dynamically configured according to the real-time traffic, and an intelligent control of the critical traffic peaks through a small-scale flexible resource over-provisioning functionality.

4.2.2 Framework Integration following 3GPP Compliant Architecture

This section describes the integration of our solution with the aforementioned slice SBA, as illustrated in Fig. 4.8. The system consists in two slices (i.e.,eMBB and mMTC) managed by a single AMF entity, in a shared APN. Nevertheless, the solution can be extended to a higher number of slices, and multiple AMFs can perform handover operations among different AN.

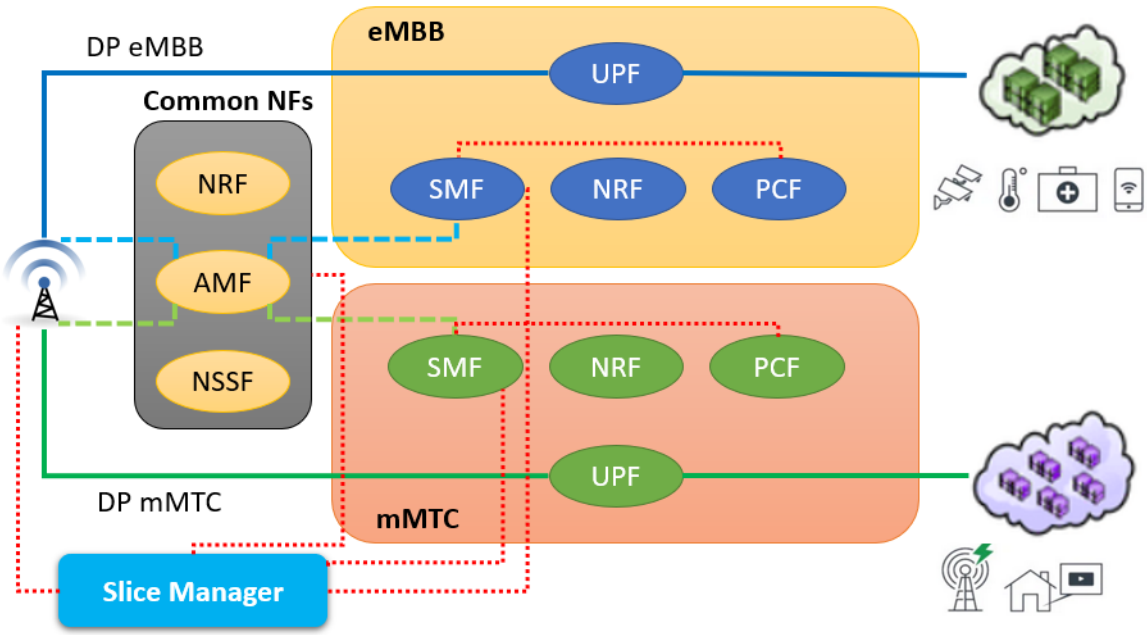


Figure 4.8: 5G compliant system architecture

For enhanced isolation, in addition to the AMF, only the NRF and NSSF are common NFs among

This contribution of this chapter has been published in the following publication: Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., & Verikoukis, C. (2020, December). Dynamic partitioning of radio resources based on 5G RAN Slicing. In GLOBECOM 2020-2020 IEEE Global Communications Conference (pp. 1-6). IEEE.

the slices, while the SMF, NRF, PCF, and UPF NFs are implemented within each slice. This implementation choice improves decoupling capabilities between the CP and DP, optimize the slice PDU session control and management, and guarantees a dynamic NFs deployment and scaled-up on demand in a completely automated manner. Our innovative NF for this architecture is represented by SM. With the support of RESTful APIs, the SM acquires real-time network information of each slice and the RAN, elaborates a new parameterization, and post a new slices configuration in the system, compliant with the 5G 3GPP architecture standard. The acquired system statistics are utilized as input for our novel dynamic RAN slicing solution, which performs the resources migration from one slice to another according to each slice SLA. In this work, the scenario is tested within a simulation environment, equipped with 5G functionalities. The target slice KPIs are the average slice data rate and the transmission error probability. Nevertheless, the degree of flexibility of the SM allows the SP to evaluates other types of metrics, as for example latency constraints, slice priority, packet error probability, advanced resource isolation paradigms, etc. Following, we will use the term *Master* slice for the slice acquiring PRBs, and *Slave* slice the one lending/releasing RBs.

4.2.3 Proposed Dynamic 5G NR Network Slicing Optimization Algorithm

The proposed algorithm for real-time RBs sharing among the slices is illustrated in Fig. 4.10. As initial input parameters, the algorithm receives the Transmission Mode (TM), a set of thresholds (th1, th2) for each slice modes (*Support*, *Conservative*, and *Critical*), the initial bandwidth partition among the slices, and the slice granularity value. Fig. 4.9 illustrates the initial modes configuration for each slice of the tested scenario. Once the initialization phase is completed, the system is ready to accept incoming users. When a new UE connection request arrives and belongs to the SP (row 4), the SM acquires from the RAN and the common NFs set, the corresponding SLA and KPI. If the existing number of available resources satisfies the users' SLA (rows 9-11), the UE is automatically served and the system parameters updated. Otherwise, the slice is in *Critical* mode, labelled *Master*, and the SM activates the sharing procedure (row 13). Until the UE SLA is not guaranteed, for each time interval t (defined through the granularity value), the SM compares the *Master* slice SLA with the current quality system indicators (SINR, CQI, packet error rate, etc.), and estimates the amount of RBs needed to fully accomplish the *Master* slice traffic load request. If exists another slice whose amount of utilized resources is less or equal to its *Support* mode, a *Slave* slice is identified, and RBs sharing with the *Master* slice can be performed. Otherwise, the slice is in *Conservative* mode, meaning that its amount of RBs must be preserved to guarantee the QoS of its users served.

Labelled all the slices, the SM performs the migration of RBs from the *Slave* to the *Master*, until

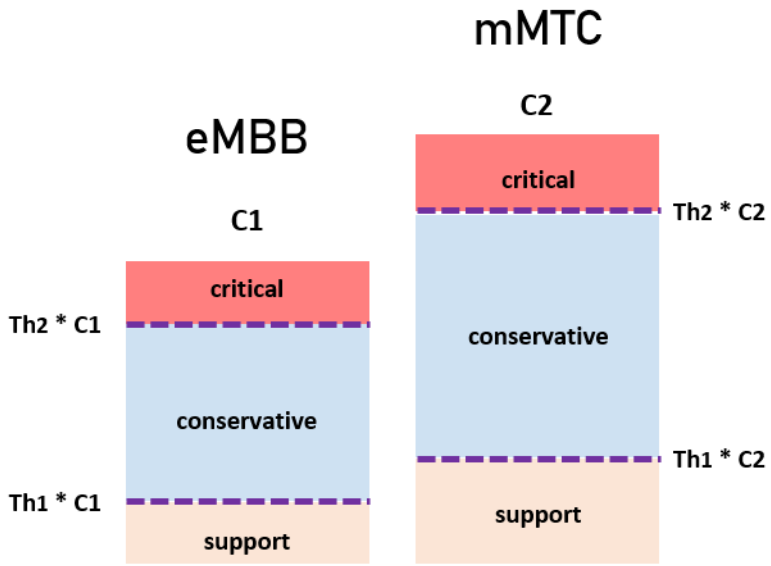


Figure 4.9: Initial slice configuration

the *Master* SLA are guaranteed, or the *Slave* slices do not have more available RBs in *Support* mode (rows 18-21). Completed the migration, the SM posts the new slice parameterization in the RAN (row 22), update the system statistics (row 23), and waits for a new incoming UE (back to row 4). Every time a UE completes the service, its amount of RBs is carefully released and distributed among the slices (row 31), taking into account the traffic load, priority, and operating mode of each slice.

4.2.4 Scenario Description

The tested scenario includes 2 slices, sharing the same RAN infrastructure, focusing on the downlink channel. An eMBB slice has the role of *Master* slice, while a mMTC slice acts as the *Slave*. Most of the real cases scenarios barely meet the 5G flexibility in terms of maximum transmission bandwidth configuration, due to the limited number of operators with segments of continuous bandwidth of 100 MHz. For our system, the optimal balance between real implementation and 5G performance is achieved using 40 MHz bandwidth at 30 KHz SCS, achieving 106 RBs in the PHY layer. As secondary simulator parameters, the cycling prefix is set Normal, the code rate for the transport block size is 490/1024, the number of PDSCH transmission antennas is 8, the number of UE receive antennas is 2, and the PDSCH modulation varies between 16 and 64 QAM, according to the CQI value. To highlight the RBs sharing principle of our RAN NS solution, at the initial phase, the total amount of available RBs is partitioned as 25% for the eMBB slice, and 75% for the mMTC, without any users connected. The *Support* mode, for both the slices, is set between 0 to 40% of the slice size, the *Conservative* mode from 40 up to 70%, and the remaining 30% represents the *Critical* mode. Progressively, multiple high demand data rate users attach to the *Master* slice,

Algorithm 1 real-time resource block slice sharing

Input: Transmission mode, slice configuration, initial slice setting, decision granularity

Output: amount of resource blocks for each slice

Initialisation :

compute the support, conservative and critical modes for *s_embb*
compute the support, conservative and critical modes for *s_mmtc*
compute the support, conservative and critical modes for *s_urllc*
if (new UE 'ue_i' arrived) **then**
  if (*ue_i* ∈ s_embb ‖ s_mmtc ‖ s_urllc) **then**
    slice_par = GET slice SLA and KPI for ue_i
    slice_sta = GET real-time slice statistics
    mode_master = Mode(slice_par,slice_sta, ue_i)
    if (mode_master ≠ critical) **then**
      connect the UE to the corresponding slice
      update system variables
    else
      while *ue_i* is not completely served **do**
        acquire channel status
        organize slices according to the priority
        for $i = 0$ to $Slices.length - 1$ **do**
          if (mode slice[i] is *support* and ≠ mode_master) **then**
            while slice[i] is *support* ‖ mode_master is critical **do**
              migrate slice[i] RBs towards slice master
              update slices RBs availability
            end while
            POST new slices parameterization
            update system resources variables
          end if
          refresh slices status (modes, av. RBs)
        end for
      end while
    end if
  else
    ue_i completed the service
    redistributes its resources
    disconnect ue_i
  end if
end if

Figure 4.10: real-time resource block slice sharing

while the *slave* slice supports a constant GBR users' traffic. With the current injected traffic, the SLA of slice eMBB are not guaranteed, and the slice shifts from *Support* to *Critical* mode. On the other side, the *Slave* slice, equipped with a generous amount of RBs, easily guarantees the SLA of its users, remaining in *Support* mode. The SM activates the slice sharing optimization procedure, and gradually moves RBs among the slices, keeping track of the real-time service traffic demand and the KPI of the ongoing served users for each slice.

4.2.5 Performance Analysis: Multi-phases Evaluation

4.2.5.1 Experiment 1: 2 Milliseconds Granularity Sharing Activation

The first test illustrates the migration of RBs from the *Slave* to the *Master* slice, under a granularity of 2 frames, when the sharing procedure is activated. After the initialization phase, two clusters of users, respectively requiring an average data rate of 45 and 22 Mbps, send a connection request to the slice eMBB. In parallel, a connection request from another cluster of users requiring 15 Mbps average data rate arrives at the mMTC slice. While mMTC remains in *Support* mode even after the user connection, the eMBB slice is located in *Critical* mode, and cannot accomplish the slice SLA. Under these conditions, the SM activates the sharing procedure, and the capacity of each slice is updated, as shown in Fig. 4.11. At each iteration, the available RBs from the *Slave* slice (mMTC) are migrated to the *Master* slice (eMBB). The graph shows how this migration is done gradually, following the granularity of 2 frames. With a SINR of 20 dB, the CQI is optimal, and the system needs only 2 iterations of 2 frames each to balance the slice load (green circle). After this stabilization phase, the SLA is guaranteed and the KPI of the ongoing services are matched. The SM continues to share RBs among the slices until the *Critical* section of the *Master* slice is restored, and the eMBB slice is safely placed in *Conservative* mode. In this test, the optimal balance of the radio resources is achieved in just 24 frames (240 ms). After this time interval, each slice presents a stable configuration, where the *Support*, *Conservative*, and *Critical* modalities are proportionally re-established.

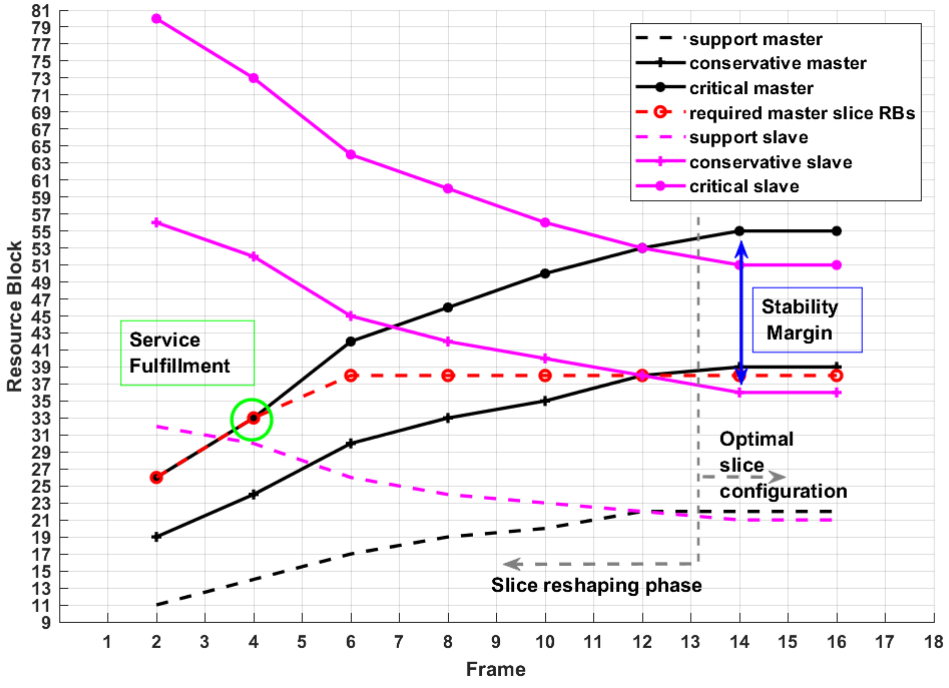


Figure 4.11: Slice RBs sharing. Optimal channel condition

4.2.5.2 Experiment 2: Sharing Performance under different SINR Values

In this subsection, Experiment 1 is repeated using different channel SINR intervals: 20-12 dB, and 8-4 dB. In Fig. 4.12, the blue lines illustrate the behavior of the eMBB slice (*Master* slice), when the *Support* mode is 40% of the whole slice capacity. Under optimal channel conditions (blue dashed line), the system rapidly achieves the slice stability point, while the *Slave* slice is still in *Support* mode after the sharing procedure. For this scenario, a more generous parameterization of the *Slave Support* mode would not increase the whole system capacity and sharing potentials, but it might redirect to a resources over-provisioning situation. A completely different case appears when the SINR is between 8-4 dB, and the system parameterization is equal to the Experiment 1 (blue solid line). The stability is not completely achieved due to the little modulation order (16 QAM) and the preserving behavior of the *Slave* slice towards its served users.

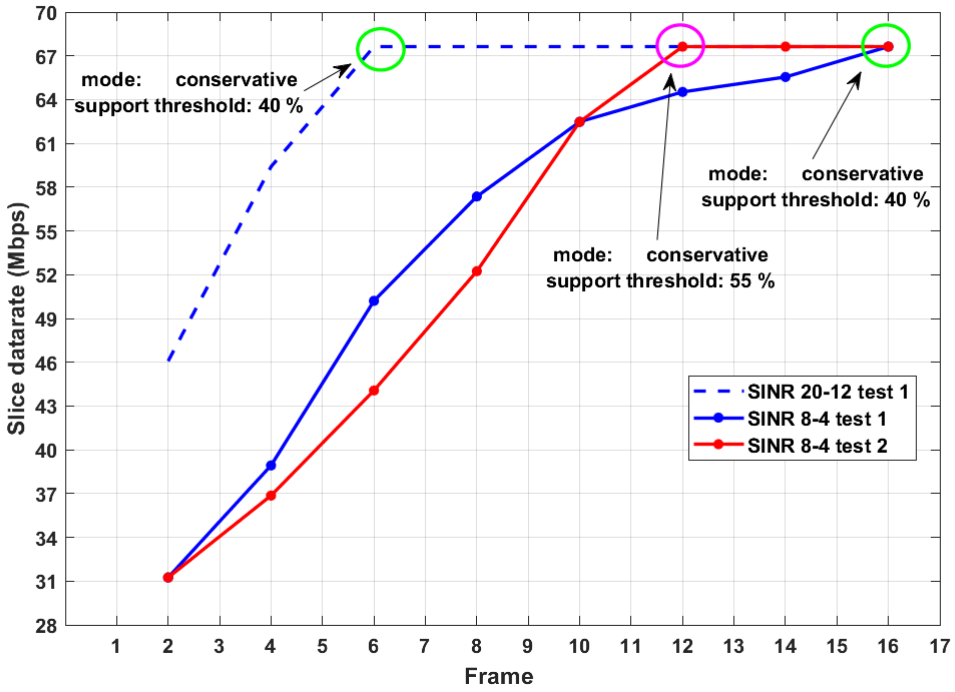


Figure 4.12: Performance comparison for different slice parameterization

To overcome this low-quality parameterization, our solution permits different degrees of configuration of the slices parameters to handle multiple types of scenarios, defining ad-hoc flavors of slice isolation, delimit the amount of resources to share, and advanced scheduling capabilities.

4.2.5.3 Experiment 3: Optimal Slice Configuration at Low SINR

Experiment 3 is represented with the red line in Fig. 4.12, showing how our solution responds at low channel quality condition when a more relaxed sharing policy is applied to the *Slave* slice. In this test, the *Slave Support* mode is increased from 40 to 55% of the whole capacity, raising the

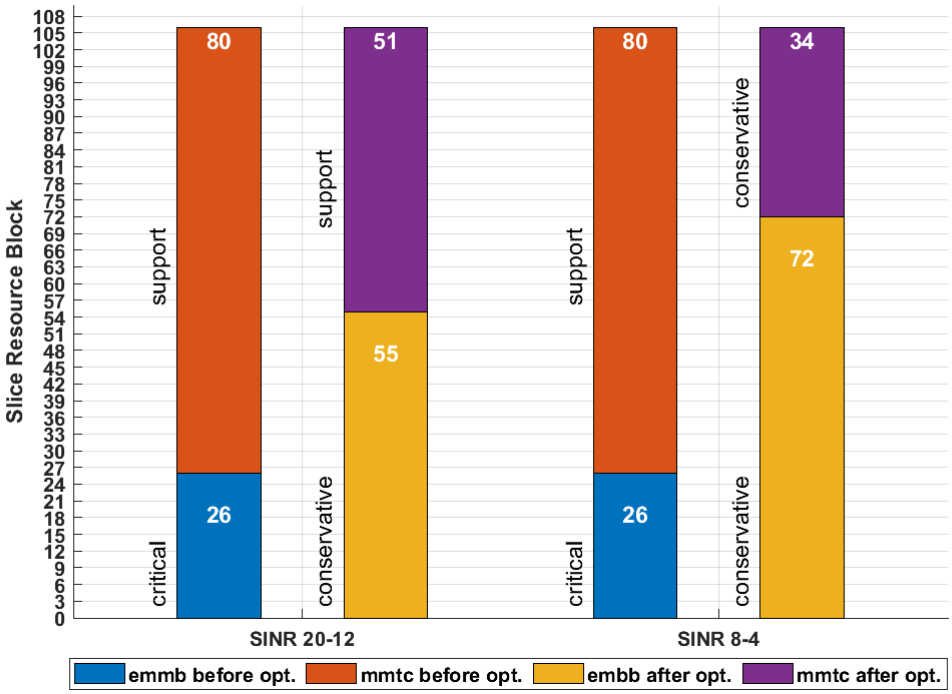


Figure 4.13: Slice resources partitioning

amount of RBs that can be shared. With the new parameterization, under bad channel condition, the eMBB slice reaches the stability condition 40 ms before the Experiment 1, and the slice SLA are completely satisfied, without impacting the KPI of the served *Slave* services. A bar representation of the slice's resources provisioning before and after the execution of the resource optimization algorithm is illustrated in Fig. 4.13, for different SINR intervals.

As previously described, at low SINR the SM must share a higher number of RBs to guarantee the SLA, while under optimal channel conditions, the mMTC slice is still in *Support* mode, meaning that not all its available resources are fully employed.

4.2.5.4 Conclusive Experimental Considerations

To conclude the result section, Fig. 4.14 illustrates another tuning option for the slice sharing algorithm: the frame granularity. This parameter defines the window size used by the SM to collect the input information (CQI, SINR values, and services statistics) necessary for the estimation of the appropriate amount of RBs to be shared among the slices. Even though employing a conspicuous number of RBs can result problematic for time sensitive applications, it increases the sharing accuracy capabilities of the algorithm. Fig. 4.14 top part compares the data rate of slice eMBB for a granularity of 1 frame (blue line) with 8 frames (red line), while bottom part illustrates the amount of RBs released by slice mMTC under the same granularities. For 1 frame granularity, the "rough" resource sharing estimation brings the slice mMTC to over-release RBs during the first

50 ms, and then converging to the system stability around 60 ms. This behavior, even though presents higher data rate during the initial phase, can compromise the SLA of the *Slave* slice if the granularity time is too small for correctly evaluates the system resources, channel quality, etc. On the other side, a granularity of 8 frames represents the optimal approach in case there are few available RBs, and the SM must be meticulous during the selection process.

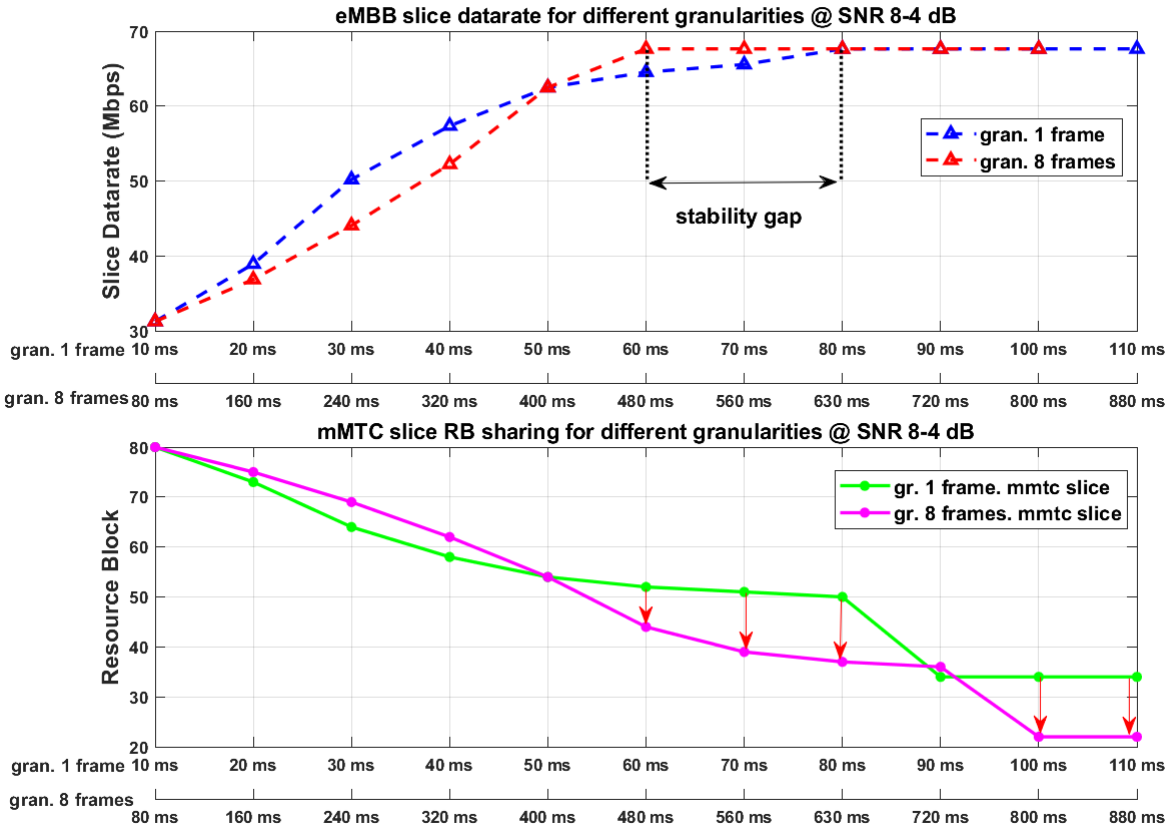


Figure 4.14: Performance under different frame granularity

As final observation, in Fig. 4.14 bottom part, for the first 50 ms of the 8 frames granularity case, the SM releases slowly the RBs for the slice mMTC, at the price of a lower data rate for the slice eMBB (Fig. 4.14, top part). After this initialization phase, the system acquires sufficient understanding of the surrounding environment, and applies a more relaxed sharing policy, as shown by the red arrows in Fig. 4.14, increasing the system accuracy and rapid alignment with the required slices SLAs.

4.2.6 Real Use Case Implementation: eMBB Traffic Orchestration in a Multi-slice Scenario

4.2.6.1 Scope and novel Features

In this subsection, the aforementioned dynamic NS framework is implemented and tested on top of a real testbed. To the best of our knowledge, our solution represents the first application of dynamic NS solution through a joint evaluation of the slice tenant SLA and the real-time service performance from the user perspective. The main contributions of our solutions includes:

- Creation of a network traffic generator tool for testing different services inline with the testbed capabilities. The traffic requirements are proportionally scaled according to the performance of the testbed.
- Highly customizable configuration of the SP traffic requirements, slices isolation policies, processing resources for new slice instances, and logging.
- Real-time analysis of the slice performance and resources configuration through a joint evaluation of the user traffic trend and slice SLA.
- Implementation of the dynamic NS solution using platform independent interfaces for advanced scalability with third party SW and HW.

4.2.6.2 System Architecture

The proposed dynamic RAN slicing framework architecture is illustrated in Fig. 4.15. Our solution consists of three macro layers (fetch, management, execution), nestled together following a bottom-up workflow. It is important to remark that our proposed solution is backward compatible with existing 3GPP 5G stack, since it utilizes standardized protocols and functionalities to communicate with the main network architecture elements. Moreover, even though this work is focused on RAN slicing, using a suitable set of input data and output configuration files settings, it is possible to extend our approach to other network sectors (MH, BH), without downgrade the optimization capabilities of our model.

This contribution of this chapter has been published in the following publication: Maule, M., Vardakas, J., Verikoukis, C. (2021). 5G RAN Slicing: Dynamic Single Tenant Radio Resource Orchestration for eMBB Traffic within a Multi-Slice Scenario. IEEE Communications Magazine, 59(3), 110-116.

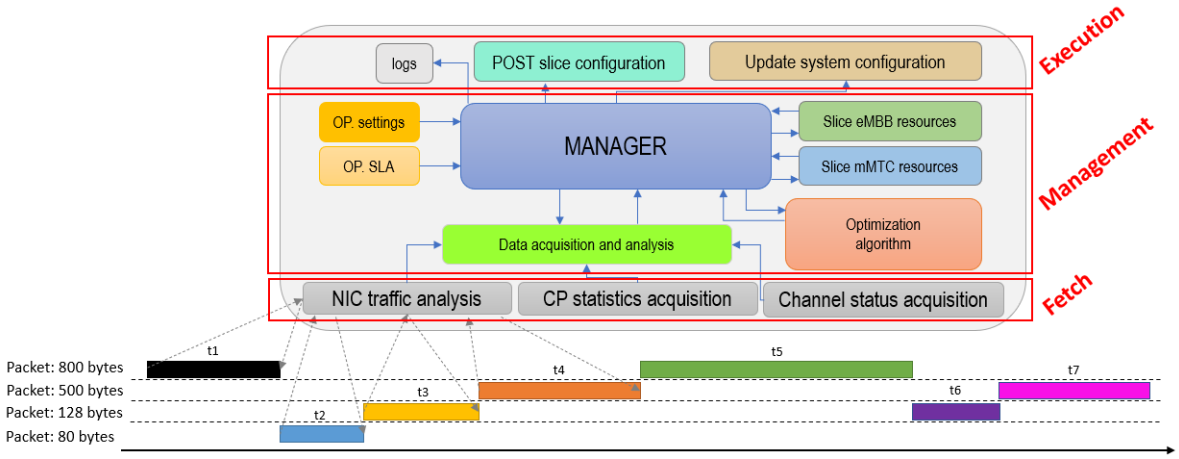


Figure 4.15: Framework structure of our solution for dynamic RAN slicing

4.2.6.3 Fetch Phase

In this subsection, all the information from the data acquisition blocks are collected, sorted, and skimmed to define a network data model utilized as input of our model. The *NIC traffic* block filters the incoming traffic according to the MNOs slices requirements, using variable window granularities. Different type of filters can be customized based on the operator's policies: packet size, colored traffic, IP subnet, etc.

The *CP acquisition* block monitors the behavior of the served users to evaluate if the SLAs and custom user traffic requirements are guaranteed. As the current served users have the priority, this block can dynamically modify the incoming services acquisition rules. Through this methodology, new incoming users are rejected in the case the slice resources are saturated, or no resource sharing policies are available. Moreover, if the slice tenant decides to modify the SLA requirements after the service is instantiated, this block is responsible to reconfigure the acquisition criteria for the new and served users, while maintaining a seamless service.

As last low layer functionality, the *Channel Acquisition block* monitors the access and FH parts to identify physical variations of the medium, which would affect not only the service acquisition rate, but also the configuration of the slices in the upper layers. For example, if a channel degradation is identified, the amount of resources for users must be redefined, together with the slices acceptance rates.

4.2.6.4 Management Phase

This part illustrates the leading block of the presented approach. As central block, the *Manager* synchronizes and coordinates the tasks of each other blocks. Its implementation can be centralized or distributed, as most of the system components are virtualized. The correct placement of the *Manager* improves the system tolerance against failures, and minimize traffic overloading in sensi-

tive network nodes. For this reason, in this work we assume the proposed framework installed in the edge part of the network, which represents a strategic point for the management of multiple RAN aspects. As minor task of the *Manager*, it is responsible to instantiate or delete the slices, according to the slice owner decision policies and/or the system performance.

As first step, the *Manager* receives in input the real-time data scenario model from the fetch layer, the SP specific settings, and per slice operator SLA. These information are encapsulated following a specific pattern, and forwarded to the *Optimization Algorithm* block, which return back the optimal slices parameterization for the next system processing window. For each slice of the MNO is defined a container with the amount of available resources, and predefined scheduling policies. In Fig. 4.15, only two slice containers are represented, one eMBB and one mMTC, in order to keep our explanation aligned with the subsequent testing part.

For the identification of RAN resources changes, the *Manager* estimates a slice parameterization every time new data are received from the fetch phase. If the novel slice configuration differs from the current one in terms of required resources per slice, the *Manager* reconfigures the amount of resources assigned to each slice according to the service's need. From a practical perspective, this procedure corresponds to shift portion of RBs among the slices, while preserving the SLA and QoS of each user. Following a tunable granularity, this operation can be executed dynamically, in accordance with the TTI of our system. The new slices parameterization is structured following the JSON syntax, and forwarded to the top tier of the architecture.

4.2.6.5 Execution Phase

The top system layer implements a set of RESTful-based APIs for the exchange of control information between the system and third-party radio SW. Once a new slice configuration is posted to the RAN, the *POST* block sends back to the *Manager* an ACK with the outcome of the operation. As final operation, the *Manager* calls the *Update System Statistics* block, which is responsible to update the system variables required for the optimal processing of the forthcoming slices configurations. Optionally, a log file can be provided to trace potential issues during the entire workflow.

4.2.6.6 Testbed Scenario Deployment

In order to assess the performance of our solution, a HW/SW experimental platform is deployed, as illustrated in Fig. 4.16. We focus our experiments in downlink traffic and for the proof of concept we use two types of slices, eMBB and mMTC, as part of a single tenant scenario. Since we want to evaluate our solution under heavy traffic condition, significant eMBB traffic is injected (92), while the primary role of the mMTC slice is to assist the supply of eMBB radio resources along the simulation. Typical examples of eMBB services are VR, gaming, and entertainment, where a

remarkable number of users are simultaneously connected. It is important to highlight that our system applies inter-slice resource prioritization only if negotiated during the SLA definition process. Otherwise, *First Come First Served (FCFS)* policy is applied. Due to the current limitation in open source 5G SA platforms, in this work, the performance of the dynamic NS solution are evaluated on top of a LTE-based testbed, equipped with 5G virtualization functionalities.

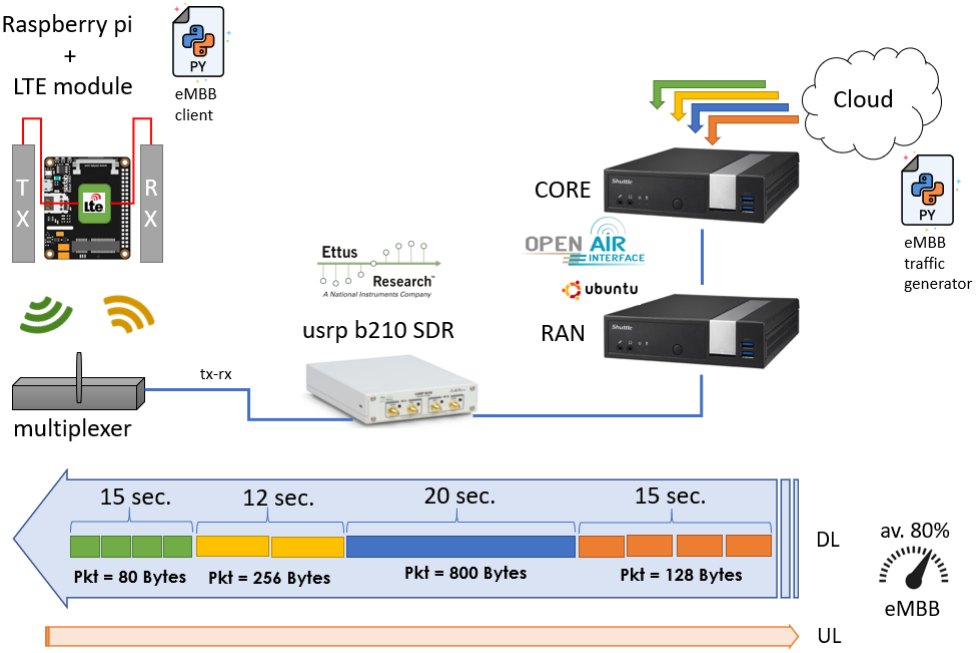


Figure 4.16: Testbed scenario and system architecture

As shown in Fig.4.16, we separated in two different machines the RAN and CORE parts. This implementation choice balances the processing workload of the system, allows flexibility in terms of centralized or distributed management, and the deployment of multiple split options as standardized in 5G (93). Table 4.2 summarizes the main parameters of the tested scenario. The initial slice division represents the percentage of RBs assigned to each slice given the total system capacity, while the thresholds refer to the maximum capacity of each slice mode.

4.2.7 Testbed Scenario Performance Analysis

Using the aforementioned scenario as baseline of our experiments, we evaluate the efficiency of the proposed dynamic NS solution when a high load eMBB traffic is injected, and as consequence the SM selects radio resources from the mMTC slice to balance the system until the SLAs are not fulfilled.

This condition is confirmed through the evaluation of the final slice eMBB mode occupancy percentage (8% *Support*, 24% *Conservative*, and 68% *Critical*), where the high presence of critical calls force the SM to perform resource migration among the slices during the entire simulation time.

Fig. 4.17 illustrates the behavior of the eMBB slice during the entire simulation time. To correctly

Table 4.2: System and simulation parameters

Total duration			150 sec		
Nr. slices			2		
Bandwidth			10 MHz (50 RBs)		
MCS DL			28		
MCS UL			8-20		
Granularity			3 sec.		
Protocol			UDP		
Direction			Downlink		
Nr. input flows	**Duration (sec)**	**Packet size**	**Packet size 128**	**Packet size 500**	**Packet size 800**
10	5	1	2	1	1
12	4	1	1	0	2
15	5	2	0	0	3
20	6	2	1	2	1
Support th. eMBB			20%		
Conservative th. eMBB			50%		
Support th. mMTC			20%		
Conservative th. mMTC			50%		
Initial slice division DL			50%		
Initial slice division UL			50%		

interpret the achieved results, the reader should take into account that the maximum downlink capacity using OAI-based testbed is around 30 Mbps.

The blue line indicates the injected eMBB high load traffic (average 20.62 Mbps, standard dev. 3.83, variance 14.74), while the red line represents the slice eMBB capacity growth trend until the SLA are not reached. For every variation of the traffic flow, the SM evaluates if a new slice configuration must be applied. This procedure is displayed with the bar plot, where each bar represents the amount of RBs required by the incoming service to reach the optimal slice eMBB SLA. When the bar sequence has a growing trend, the SM increases the amount of RBs for the slice eMBB in the next slice configuration. Conversely, a decreasing bar sequence indicates that the current slice configuration correctly match the served users traffic, and the SM might decide to redistribute part of the free RBs among other slices.

As stress downlink test for the slice eMBB, a series of flows are generated from the cloud network towards the users, with an average traffic load equals to 80% the total system capacity. As expected, the consecutive resource optimization calls of the SM brings a continuous increment of the eMBB slice capacity (from 25 RBs of the initial setup, until a final slice capacity of 44 RBs), with an occupation of 88% of the total system capacity before the end of the simulation. With our solution, using a SM decision granularity of 3 seconds, after approximately 60 seconds the system reaches a steady resource equilibrium, optimally configured for the type of injected traffic. Moreover, as the sharing policies take into account the performance of the complete set of slices within the

system, the SLAs are guaranteed for both the slices during the entire process, until a final stable slice configuration. With an average amount of received packets of 20.50 Mbps, this experiment presents a Packet Error Ratio (PER) equals to 0.005. The dynamic approach of our solution, as expected, slighly affect the PER, which is nevertheless acceptable if compared with the standardized eMBB requirements (27).

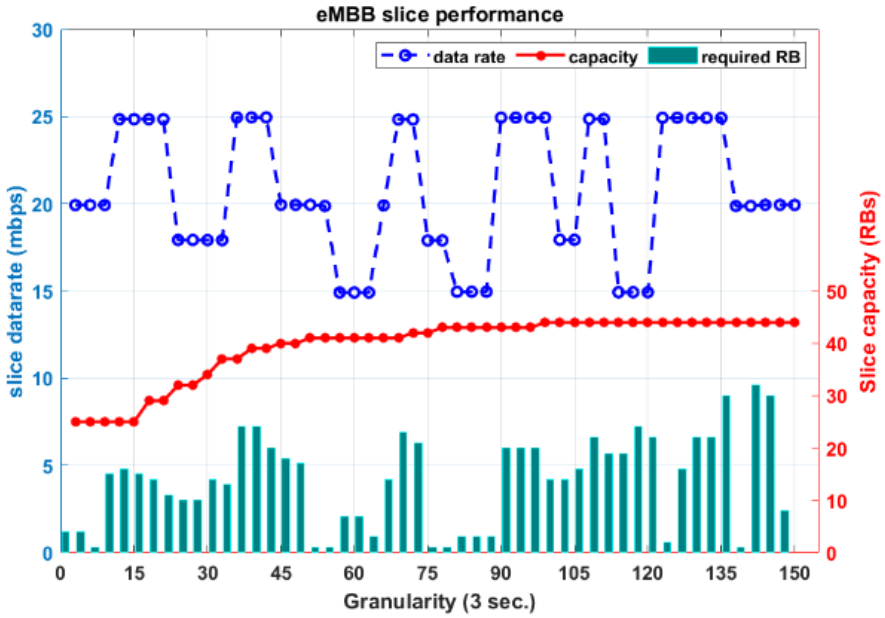


Figure 4.17: eMBB data rate and slice capacity variation

For every iteration, the impact on the slice eMBB capacity proportionally affects all the modes, as illustrated in Fig. 4.18. This proportional scaling of the RBs of each mode reduces the ping-pong effect among the slices, which appears when a highly variable traffic is injected, and a continue scale up and down of the resources impacts the PER and management complexity. Without the division of each slice in different modes, the system would continue moving radio resources among the slices with the objective to balance the maximum capacity of each slice, defining a new slice configuration even for irrelevant traffic variations.

The results of Fig. 4.18 brings to light an intrinsic principle regarding the initial parameterization of the *Support* and *Conservative* eMBB slice modes. Allocating a tiny amount of RBs in *Support* (20% of the initial capacity), pushes the system in *Conservative* mode even when a small amount of traffic is injected in the system. Moreover, even a reduced *Conservative* threshold benefits the eMBB slice, since the slice is more inclines to enter the *Critical* mode, requesting other slices to share part of their RBs. This observation highlights how the initial slice modes setup should be carefully investigated taking into account the type of slice traffic, the slice isolation policies, and specific tenant's requirements.

To conclude, an average jitter of 0.206 ms confirms the optimal configuration of the tested scenario,

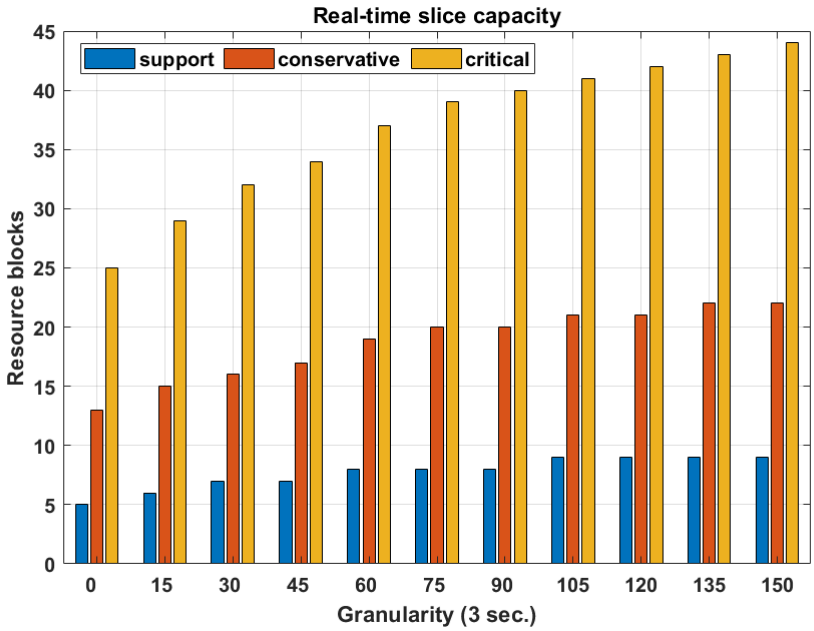


Figure 4.18: eMBB slice capacity mutation

ensuring high order modulation scheme during the entire simulation. The high customization degree of the SM parameters permits to tackle multiple scenarios, and simultaneously optimize different type of slices. A trade-off between slice resource assignation accuracy and a new slice configuration processing time should be carefully inspected. Depending on the type of traffic, this feature may have an impact on the slice throughput, latency and responsiveness to real-time traffic pattern changes.

4.3 User-Slice Stochastic Provisioning of Radio Resources in a Multi-slice 5G NR Scenario

4.3.1 Novel Framework Objective

In this work a novel dynamic RAN NS orchestrator is presented, capable to provide customized network services able to guarantee the SP's requirements while effectively utilize network resources. The novel mathematical model is applied to our proposed SM, which represents a 3GPP backward compatible framework able to dynamically perform the optimal radio resources partitioning among the served slices through the simultaneous analysis of the tenant SLAs and real-time service requirements of the served users connected to each slice. The heterogeneous evaluation of the tenant's slice requirements with the user-based traffic perspective outperforms the ordinary dynamic radio slicing approaches, since the proposed system is constantly aware of the real-time user's traffic performance, reducing the impact of unexpected wireless channel variations, whose intrinsic stochastic behavior is complex to predict.

4.3.2 Optimization Framework and Slice Life Cycle Consolidation

In this subsection, the integration of the proposed RAN slicing optimization framework with the standardized 3GPP NS life cycle management is illustrated. Fig. 4.19 shows the integration of the proposed SM solution (red part) with the 3GPP standardized handshake procedure between the UE and the 5G network. The proposed SM NF is detached from the 5G architecture, and communicates with the 5G service-based NFs through a set of RESTful APIs. This implementation facilitates the interaction with 5G SBA of distinct ecosystems, and raises the degree of flexibility and scalability of our solution since the SM can be placed in different part of the system. The set of slices used by the UE and its PDU sessions are controlled by the SP considering the user's subscription. Multiple AMF are placed inside the system, each one with a different vision of available slices. Since a UE may require various slices, the optimal AMF selection is done by the 5G AN after receiving the NSSAI by the UE, during the initialization phase. The NSSAI is composed by multiple Single-NSSAI (SNSSAI) consisting of a Slice Service Type (SST) field indicating the slice characteristics, and a Slice Differentiator (SD) field with extra SST information. Once the 5G AN assigned the correct AMF to the user, a copy of the NSSAI and NSIs is sent to the SM, which receives a complete overview of the involved slices and user specific traffic requirements. At PHY layer, the

This contribution of this chapter has been published in the following publication: Maule, M., Vardakas, J. S., & Verikoukis, C. (2021). A Novel 5G-NR Resources Partitioning Framework Through Real-Time User-Provider Traffic Demand Analysis. IEEE Systems Journal.

SM constantly monitors the channel conditions, resource utilization, and the RAN KPIs. This information, together with the SP SLAs, the slices isolation features, and the priority policies are utilized by the SM to process the optimal slices RAN resources configuration in line with the real-time service traffic needs. While S-NSSAI indicated the expected characteristics from a slice, further parameters could be required to select the optimal slice. For this reason, when the 5G AN requires a PDU session, the 5G CN sends the S-NSSAI corresponding to the PDU session to the 5G AN, refining the selection criteria. To complete the current subsection, the proposed slice selection

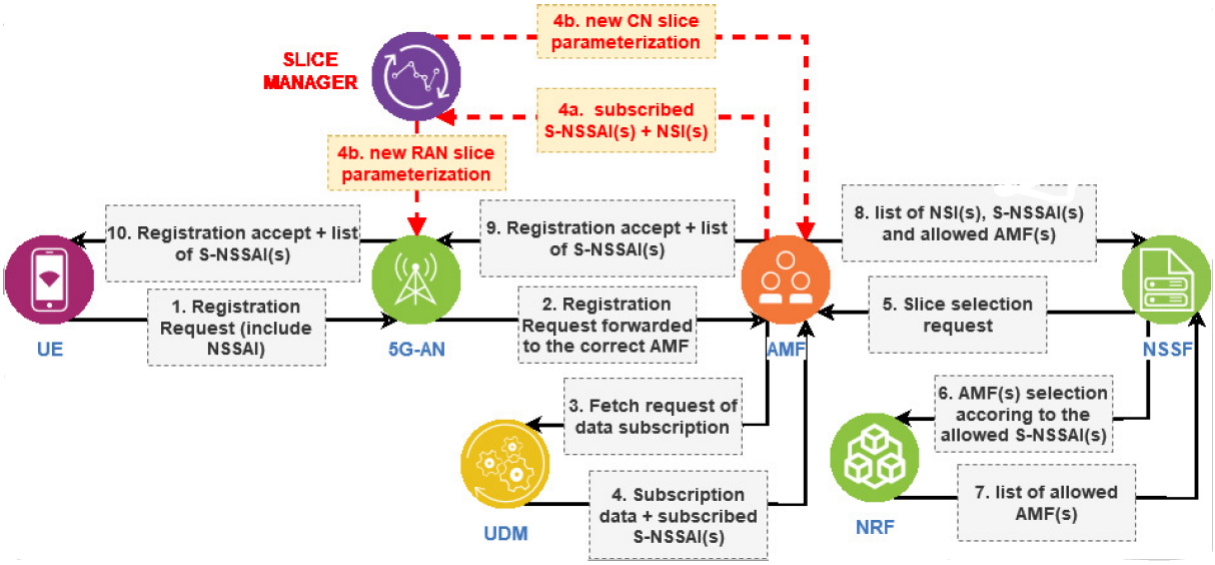


Figure 4.19: Service allocation and RAN resources provisioning

and optimization message flow is presented, starting from the initial handshake procedure between a new UE request until the complete service supply (27):

- The UE sends a Registration Request and the NSSAI. The requested NSSAI is configured according to the PLMN (i.e., provisioned to the UE) when accessing the PLMN for the first time.
- The primary role of the AN is the AMF first selection, based on the requested NSSAI. The RAN may use requested NSSAI in access stratum signaling to handle the UE CP connection before the 5G-CN informs the RAN of the available NSSAI.
- The AMF verifies the Requested NSSAI and authenticates the subscriber information with the support of the AUSF. Next, the AMF fetches subscription data from UDM, providing also the requested NSSAI.
- The UDM elaborates the received information and return the S-NSSAI(s) for those the UE has a subscription:

  - The list of S-NSSAI(s) and subscription information are forwarded to the SM, which are used to initialize the network optimization procedure for the current UE.
  - The SM adds the network status statistics parameters to the slicing optimization model. At this point, the SM initiates the computation of a novel radio resource parameterization.

- The AMF requests the NSIs compatible with the subscribed S-NSSAIs to the NSSF. Moreover, information regarding other AMFs able to support the allowed NSIs are forwarded.
- NSSF discovers the suitable AMFs through the forwarding of S-NSSAs to NRF.
- The allowed AMFs are forwarded back to the NSSF.
- At this point, the selected AMF is aware of service requirements of the UE: NSI IDs, suitable S-NSSAIs, and AMF candidates.
- The registration status and the S-NSSAIs are received by the 5G-AN. According to the 3GPP 5G standard, another AMF can be selected to meet the slice requirement. Since in our work we are considering a system of two slices of a single tenant, we consider one single AMF implementation.
- To conclude, the service acceptance request is received by the UE, and its services can start.

4.3.3 Dynamic Network Slicing Mathematical Model

This subsection illustrates the proposed radio sharing resources decision algorithm implemented inside the SM, where multiple tenants own a set of slices, each one customized according to the type of service and customer specifications. The optimal system configuration is based on the determination of the optimal set of parameters for each slice able to minimize the blocking probability of an incoming user connection request. Table 4.3 summarizes the mathematical model parameters. Let $S_n^{T_j}$ denote the slice $n, (n = 1, ..., N_j)$ at time t in the system belonging to the tenant T_j, where $j, (j = 1, ..., J)$, is the number of tenants sharing the FH infrastructure, N_j is the number of slices of the tenant T_j, and J is the total number of tenants. Each slice $S_n^{T_j}$ is characterized by an initial capacity $C_n^{T_j}$ such that:

$$\sum_{j=1}^{J} \sum_{n=1}^{N_j} C_n^{T_j} = C_{tot} \tag{4.1}$$

where C_{tot} represents the total physical system capacity of the radio section. The system parameters

Table 4.3: Mathematical model parameters

$S_n^{T_j}$	Slice n of tenant T_j
T_j	Tenant j
$C_n^{T_j}$	Initial capacity slice n of tenant T_j
C_{tot}	Total system capacity
λ_n^j	Arrival rate slice n of tenant j
μ_n^j	Service rate slice n of tenant j
$T^j{}_{n,sup}$	*Support mode* slice n of tenant j
$T^j{}_{n,cons}$	*Conservative mode* slice n of tenant j
$P^j{}_{n,lim}$	Slice n *blocking probability* of tenant j
$P^j{}_{n,bloc}$	Total *blocking probability* for slice n of tenant j
P_b^{tot}	*Total blocking probability*

of each slice $S_n^{T_j}$ are defined by the quintuple:

$$[\lambda_n^j;\ \mu_n^j;\ T^j{}_{n,sup};\ T^j{}_{n,cons};\ P^j{}_{n,lim}], \tag{4.2}$$

where λ_n^j is the connection request arrival rate for n-th slice of the j-th tenant, μ_n^j is the service rate for n-th slice of the j-th tenant, $T^j{}_{n,sup}$ is the percentage of radio resources assigned to the n-th slice of the j-th tenant in *support* mode, $T^j{}_{n,cons}$ is the percentage of radio resources assigned to the n-th slice of the j-th tenant in *conservative* mode, and $P^j{}_{n,lim}$ is the *slice blocking probability* upper bound for n-th slice of the j-th tenant.

Given the aforementioned set of parameters, the scope of our solution is:

$$\begin{aligned} &min\{[T^1{}_{1,sup}, T^1{}_{1,cons}], \ldots, [T^J{}_{N_j,sup}, T^J{}_{N_j,cons}]\} \\ &s.t.\ P^1{}_{1,bloc} \leq P^1{}_{1,lim}, \ldots, P^J{}_{N_j,bloc} \leq P^J{}_{N_j,lim} \\ &P^j{}_{n,bloc}, P^j{}_{n,lim} \in (0,1), (n = 1, ..., N_j), (j = 1, ..., J), \end{aligned} \tag{4.3}$$

where $P^j{}_{n,bloc}$ is the *total blocking probability* for n-th slice of the j-th tenant.

Considering all the tenants T_j, the combination of slices $S_n^{T_j}$ and corresponding capacities $C_n^{T_j}$ allows to define the state space of the system as:

$$\begin{gathered} \vec{n} = (i_1^{T_0}, \ldots, i_{N_j}^{T_j}), \quad 0 < i_1^{T_0}, \ldots, i_{N_j}^{T_j} < C_1^{T_0}, \ldots, C_{N_j}^{T_j}, \\ \Omega = \{\vec{n}\}, \end{gathered} \tag{4.4}$$

where $i_n^{T_j}$ is the amount of occupied resources by the n-th slice of the j-th tenant, and Ω is the *state space* of the system.

For presentation purposes, the presented analysis refers to a single tenant system of 2 slices; however, in a multi-slice case, the presented analysis can be applied to groups of two slices, since the proposed solution is based on the exchange of resources between pair of slices. In case of an odd number of slices, a prioritization of the radio resources is applied beforehand, based on the availability and slice modality. Fig. 4.20 illustrates the system space Ω of two slices S_1, S_2 belonging to the same tenant, where $\vec{n} = (i, j)$ for $0 < i < C_1$, $0 < j < C_2$.

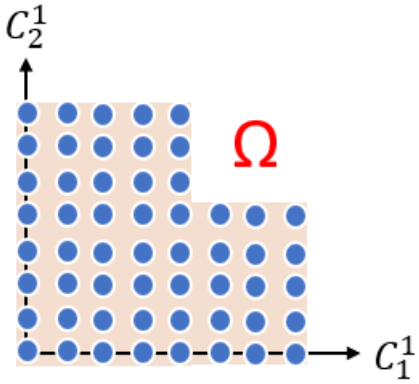


Figure 4.20: Complete two slices state space

We assume that for slice S_n, the establishment requests follow a Poisson distribution with a mean rate λ_n^j, which is a widely accepted process for the arrival of slicing requests (94), while the slice service times are generally distributed with a mean value of μ_n^j (95). Also, please note that the presented analysis considers that the number of occupied resources is equal to the number of users served from the slice. Under the latter assumption, we can describe the system as a discrete time stochastic process. Stochastic processes are meant to model the evolution over time of real phenomena for which randomness is inherent. We assume that the state space $\vec{n} = \{[0,0],[0,1],\dots,[1,0],\dots\}$, is a proper subset of Z, with $\vec{n} \geq 0$. Moreover, since the system is memory-less, the stochastic process can be represented as a two-dimensional Markov process. Each transition from one state (i, j) to another involves one radio unit per time defines the *one-step transition probability* $P_n(i, j)$.The collection of all the possible $P_n(i, j)$ is a matrix of dimension $C_1 \times C_2$ called *one-step transition matrix*:

$$\begin{aligned} \vec{P} &= P_n(i,j) \quad i,j \in \Omega \\ P_n(i,j) &\rightarrow (i \pm 1, j) \vee (i, j \pm 1). \end{aligned} \tag{4.5}$$

Assuming the transition probabilities do not depend on the time t, using $P_n(0,0) = 1$ yields $P_n(i,j) = P(i,j|0,0)$.

Since the system is a Markov process, starting from a random state it is possible to reach any other in one or more steps. This property allows to describe our problem as a multidimensional birth-death process, where the representation of a generic state as shown in Fig. 4.21 permits to derive the *balance equation* of the system:

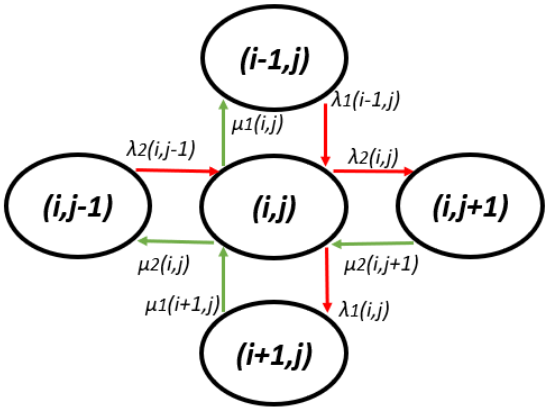


Figure 4.21: Generic Markov state representation

$$\begin{aligned}&\lambda_1^1(i-1,j)\cdot P_n(i-1,j)+\lambda_2^1(i,j-1)\cdot P_n(i,j-1)+(i+1)\cdot\mu_1^1\cdot P_n(i+1,j)+\\&(j+1)\cdot\mu_2^1\cdot P_n(i,j+1)=\lambda_1^1(i,j)\cdot P_n(i,j)+\lambda_2^1(i,j)\cdot P_n(i,j)+i\cdot\mu_1^1\cdot P_n(i,j)+\\&j\cdot\mu_2^1\cdot P_n(i,j)\end{aligned}\tag{4.6}$$

To determine the *one-step transition probabilities* $P_n(i,j)$, an opportune manipulation of Eq. 4.6 permits to derive a matrix formulation of our system in the form:

$$\begin{cases}\underset{C_1\cdot C_2\times C_1\cdot C_2}{\mathrm{A}}\times\underset{1\times C_1\cdot C_2}{P^T}=0\\\sum_{i=1}^{C_1\cdot C_2}Pi=1\end{cases}\tag{4.7}$$

A is the matrix of coefficients, and the final *transition probability P matrix* will have dimension $C_1\times C_2$. From Eq. 4.6, we obtain:

$$\begin{aligned}&\lambda_1^1(i-1,j)\cdot P_n(i-1,j)+\lambda_2^1(i,j-1)\cdot P_n(i,j-1)+(i+1)\cdot\mu_1^1\cdot P_n(i+1,j)+\\&(j+1)\cdot\mu_2^1\cdot P_n(i,j+1)-\lambda_1^1(i,j)\cdot P_n(i,j)-\lambda_2^1(i,j)\cdot P_n(i,j)-i\cdot\mu_1^1\cdot P_n(i,j)\\&-j\cdot\mu_2^1\cdot P_n(i,j)=0\end{aligned}\tag{4.8}$$

$$\begin{aligned}&\lambda_1^1(i-1,j)\cdot P_n(i-1,j)+\lambda_2^1(i,j-1)\cdot P_n(i,j-1)-(\lambda_1^1(i,j)+\lambda_2^1(i,j)+\\&i\cdot\mu_1^1+j\cdot\mu_2^1)\cdot P_n(i,j)+(j+1)\cdot\mu_2^1\cdot P_n(i,j+1)+(i+1)\cdot\mu_1^1\cdot P_n(i+1,j)=0\end{aligned}\tag{4.9}$$

For an easier interpretation, we rename the coefficients of expression Eq. 4.9 as:

- $\alpha(i,j) \rightarrow \lambda_1^1(i-1,j)$
- $\beta(i,j) \rightarrow \lambda_2^1(i,j-1)$
- $\gamma(i,j) \rightarrow -(\lambda_1^1(i,j) + \lambda_2^1(i,j) + i \cdot \mu_1^1 + j \cdot \mu_2^1)$
- $\delta(i,j) \rightarrow (j+1) \cdot \mu_2^1$
- $\epsilon(i,j) \rightarrow (i+1) \cdot \mu_1^1$

Obtaining a final approximated form of the matrix of coefficients A:

$$\begin{pmatrix} 0 & .. & 0 & .. & .. & \gamma(0,0) & .. & \delta(0,0) & .. & \epsilon(0,0) & .. & .. & 0 \\ 0 & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & 0 \\ 0 & .. & \alpha(i,j) & .. & \beta(i,j) & .. & \gamma(i.j) & .. & \delta(i,j) & .. & \epsilon(i,j) & .. & 0 \\ 0 & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & 0 \\ 0 & .. & .. & .. & .. & \alpha(C1,C2) & .. & .. & \beta(C1,C2) & .. & .. & .. & \gamma(C1,C2) \end{pmatrix} \tag{4.10}$$

As mentioned in subsection 4.3.2, a slice accepts an incoming connection user in *Critical* mode if and only if exists at least another slice in *Support* mode. Otherwise, the user request is rejected. In our Markovian process, this behavior is modelled through a set of constraints and boundaries conditions on the arrival rate of each slice, as illustrated with Eq. 4.11:

$$\lambda_l^1(i,j) = \begin{cases} \lambda_l^1 & \text{if } i \leq \lfloor {T^1}_{l,sup} C_l \rfloor \\ \lambda_l^1 & \text{if } \lfloor {T^1}_{l,sup} C_l \rfloor < i \leq \lfloor {T^1}_{l,cons} C_l \rfloor \cap j \leq \lfloor {T^1}_{l,cons} C_m \rfloor \\ \lambda_l^1 & \text{if } i > \lfloor {T^1}_{l,cons} C_l \rfloor \cap j \leq \lfloor {T^1}_{l,sup} C_m \rfloor \\ 0, & otherwise \end{cases} \tag{4.11}$$

for $l = (1,2)$ and $l \neq m$.

Given the pair of slice, the user-connection blocking state consists of one slice fully occupied while the other in *Conservative* mode. The case *Critical-Critical* does not exist, since it would require a capacity per slice exceeding the maximum system capacity. A graphical representation of the blocking states of our two slices system is illustrated in Fig. 4.22, where the yellow area represents the blocking area for the slice 2, while the green area is for the slice 1.

From the previous analysis, we can derive the formulation of *static blocking probability* for the user-connection request in each slice:

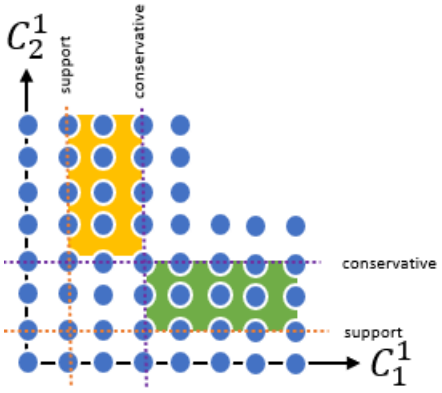


Figure 4.22: Two slices blocking states

$$P^1{}_{l,b} = \sum_{n=\lfloor T^1{}_{l,sup} \cdot C_l \rfloor}^{\lfloor T^1{}_{l,cons} \cdot C_l \rfloor} P_n(C_1, n) \tag{4.12}$$

$$P^1{}_{m,b} = \sum_{n=\lfloor T^1{}_{m,sup} \cdot C_m \rfloor}^{\lfloor T^1{}_{l,cons} \cdot C_m \rfloor} P_n(n, C_2) \tag{4.13}$$

For $l = (1, 2)$ and $l \neq m$.

Finally, we calculate the *total blocking probability* P_b^{tot} in the system via the following weighted summation of Eq. 4.12 and 4.13:

$$\begin{aligned} P_b^{tot} = \frac{\lambda_1^1}{\lambda_1^1 + \lambda_2^1} \cdot \sum_{n=\lfloor T^1{}_{l,sup} \cdot C_l \rfloor}^{\lfloor T^1{}_{l,cons} \cdot C_l \rfloor} P_n(C_1, n) + \\ \frac{\lambda_2^1}{\lambda_1^1 + \lambda_2^1} \cdot \sum_{n=\lfloor T^1{}_{m,sup} \cdot C_m \rfloor}^{\lfloor T^1{}_{m,cons} \cdot C_m \rfloor} P_n(n, C_2) \end{aligned} \tag{4.14}$$

4.3.4 Proposed Framework Performance: Structural Analysis

In the next subsections, the validation and practicality of our proposed dynamic RAN slicing algorithm is illustrated, following a three step methodology specifically designed to exhaustively prove both the analytical and experimental capabilities of our solution. As a first step, the mathematical model is carefully calibrated such that its setup is compliant with the network simulation and emulation environments. The accuracy of this process are explained in Subsection 4.3.4.1, where results from the proposed analytical model are compared with corresponding results from simulation, together with an illustration of the employed system tools and scenario settings. While local minimum points/solutions are a sufficient condition to test the calibration of our proposed

algorithm, the investigation of the global minimum solution enhances the stability and performance of our system. For this reason, in Subsection 4.3.4.2 we examine the impact of the optimal NS configuration using our proposed algorithm, with some specific considerations regarding the intrinsic resource sharing cooperation among the slices to reach the correct system balancing. Finally, in Subsection 4.3.4.3, the knowledge of the previous experiment are applied to our testbed scenario, which is equipped with 5G functionalities and a real user device. This final experiment further consolidates the capabilities of our algorithm through a real implementation and evaluation of the global minimum slice configuration setup using real devices. Furthermore, the experimental tools and scenario settings are detailed explained before presenting the results.

4.3.4.1 Analytical and Simulation Results Comparison

This subsection validates the accuracy of the proposed model, by comparing analytical results with corresponding results from our network simulator. Since the complexity of our analytical solution increases proportionally with the system radio resources, the mathematical model presented in Subsection 4.3.3 has been defined using the Python programming language. Beside other tools, Python presents powerful libraries (i.e. NumPy) for matrix elaboration and high order system of equations processing (i.e. SciPy), facilitating the mathematical implementation and reducing the computational resources. For the simulation part, the Matlab language has been chosen, using the same metrics of the analytical part for the exact comparison of the two methods. This programming language includes native modules for modelling the users and their services according to our system requirements, and different toolbox for communication systems. As development tools, the analytical part is written in Python, while the simulator in Matlab.

The simulation scenario consists of a single tenant with two slices, each one characterized by different input parameters such as the arrival and service rates, where two Poisson arrival processes, one for each slice, inject traffic in the system. The total system capacity of 40 PRBs is divided between the two slices. For both the models (analytical and simulation), once defined the input parameters, the analytical model returns as output the *total blocking probability* (Eq. 4.14), while the same result is obtained from the simulation model by averaging the number of users connection requests rejected with the total number of users requests received by the system. Since our priority is to define a seamless and unique scenario configuration among all the experiments, each slice has an initial capacity of 20 PRBs, which provides the optimal trade-off between performance and accuracy. The percentage value of each slice mode does not impact the accuracy evaluation, since the same mode parameterization is applied to both the analytical and simulation systems. For this reason, we defined the *Support mode* as 30% of the occupied resources, while the *Conservative mode* as 80% of the occupied resources, for both the slices. This slice modes configuration tends

to maintain each slice in *Conservative mode*, reducing the risk of metric mismatch between the analytical and simulation models caused by an excessive number of *Support* PRBs requests in *Critical* mode.

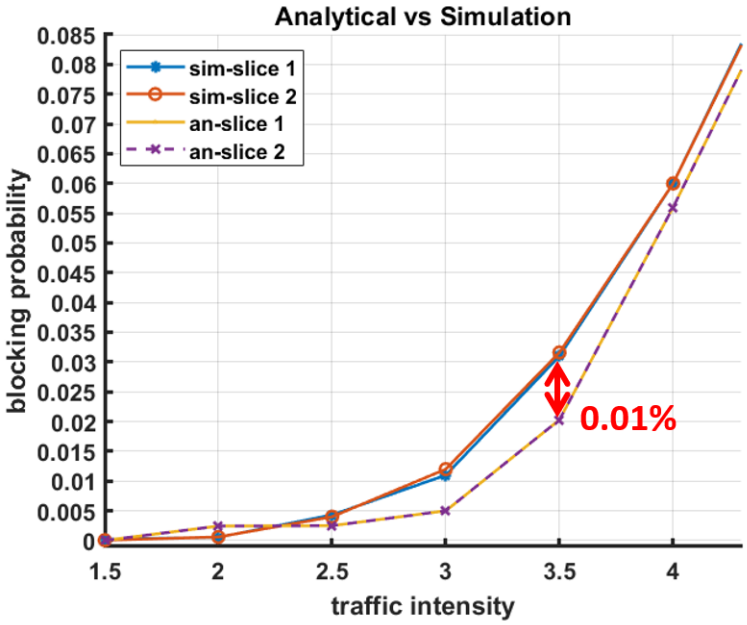


Figure 4.23: Analytical vs simulation results

Given the service time generally distributed with a mean value $\mu^{-1} = 1$, the blocking probability of an incoming service request increases proportionally with the traffic intensity (ratio between arrival and service rates) of each slice. As the comparison of analytical (yellow and purple lines) and simulation (blue and orange lines) results of Fig. 4.23 reveals, the accuracy of the proposed mathematical model is highly satisfactory. The same parameterization was tested multiple times for both the models in order to increase the confidence level of our experiment. The maximum discrepancy value among the two models is encountered when the traffic intensity value is 3.5 erl, and it is equals to 0.01%, as highlighted with the red arrow in Fig. 4.23. After an initial phase where both the models assume almost a constant trend, the blocking probability of all the slices follows an exponential behavior, as expected with the increase of the injected traffic volume. After an initial phase where both the models assume almost a constant trend, the blocking probability of all the slices follows an exponential behavior, as expected with the increase of the injected traffic volume. The performance analysis for a traffic intensity higher than four times the service time is out of the scope of this work, since it involves the collapse of the entire system and other techniques should be involved to overcome the critical situation. To conclude the subsection, we demonstrated that the accuracy of the proposed analytical model is satisfactory, even for high traffic intensity values; therefore, the proposed approach can be used for the determination of the optimal slice configuration under any traffic conditions. Moreover, this first experiment confirms the correct

calibration of the proposed analytical model, inline with the system specification of a multipurpose 5G scenario.

4.3.4.2 Optimal Slice Configuration: Performance Analysis

In the previous subsection, the percentage of *Support* and *Conservative* modes was randomly chosen and pairwise assigned to the analytical and simulation set of slices. Since the target of our model is the minimization of the blocking probability for each slice, an optimal slice modes parameterization must be estimated.

Given a scenario composed by a pair of slices, each one with an initial capacity of 10 PRBs, the objective of the second experiment is to identify the set of modes for each slice which guarantees the minimum blocking probability. Both slices have a Poisson arrival distribution, with corresponding rates λ equal to 4 and 2 user-request, respectively. The service rate is the same for both the slices, with mean value $\mu = 1\ s^{-1}$. Assigned to the *Support* mode an amount of PRBs between 10-30% of the whole slice capacity, and between 40-70% for the *Conservative* mode, Fig. 4.24 depicts the corresponding *total blocking probability* P_b^{tot} versus all the possible combination of slice modes of the two slices. Considering only integer values of PRBs for each mode, the x axis of the graph

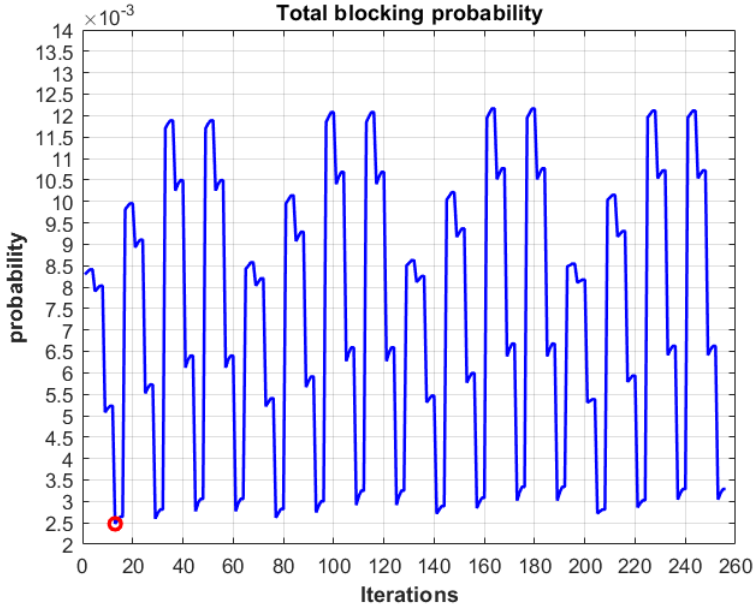


Figure 4.24: Optimal blocking probability case

illustrates 256 possible mode combinations. Among all the trials (Table 4.4 collects some examples of thresholds parameterization for different values of *total blocking probability* bound), the case with the lowest total blocking probability (equals to 0.247%), represented with a red cycle, corresponds to:

- *Support mode* slice 1: 10%

- *Conservative mode* slice 1: 20%
- *Support mode* slice 2: 40%
- *Conservative mode* slice 2: 50%

Table 4.4: Optimal thresholds configuration for different blocking probabilities

Blocking probability bound	**Support slice 1**	**Conservative slice 1**	**Support slice 2**	**Conservative slice 2**	**Average Blocking Probability slice 1**	**Average Blocking Probability slice 2**
0.0026	1.20	5.20	4	6.00	0.000328	0.00718
0.0027	2.40	5.40	4	5.40	0.000449	0.00728
0.0028	2.36	5.63	4	6.27	0.000490	0.00740
0.0029	2.66	6.00	4	5.83	0.000558	0.00756
0.003	2.91	6.50	4	6.00	0.000552	0.00790
0.0031	1.75	6.75	4	6.75	0.000513	0.00821
0.0032	3.50	7.00	4	6.00	0.000626	0.00835
0.0033	3.00	7.00	4	7.00	0.000725	0.00843

Fig. 4.25 illustrates for each slice, the amount of resources allocated in *Support* and *Conservative* modes (y right axis), and the trend of the slice blocking probability (y left axis), given an upper bound for the *total blocking probability* (x axis). As expected, due to the plenty of traffic pattern combinations, the slice characterized by a higher traffic intensity (slice 1) presents a fluctuating increasing slice blocking probability trend, while for lower traffic intensity values, the trend is more stable, assuming almost a constant behavior (slice 2).

Another direct correlation with the traffic intensity is the dimensioning of each slice modes. Fig. 4.26 illustrates the PRBs difference between the *Support* and *Conservative* modes for each slice, given an upper bound for the user *total blocking probability*. As illustrated, to maintain small slice blocking probability values, the proposed algorithm tends to increase the PRBs interval between the modes when the traffic intensity is high (top part Fig. 4.26), while increasing the *Support* mode of the low traffic intensity slice (bottom part Fig. 4.26). Following this principle, slice 1 lies most of the time in *Conservative* mode, increasing the preservation of its own resources, since sharing policy is limited. On the other side, slice 2 presents greater chances to share its resources, since a sizable amount of PRBs are allocated in *Support* mode. It is important to highlight that these results are illustrated with a confidence level equals to 97%, since the same experimental scenario was tested multiple times with a random number generator seed for each iteration.

The aforementioned analysis regarding the effects of an optimal slice configuration foresees the intrinsic intelligence behind the proposed algorithm. Our algorithm prefers to define small *Support* mode and greater *Conservative* mode for the slices with high traffic intensity, while increasing the *Support* mode for the slices with a low traffic intensity.

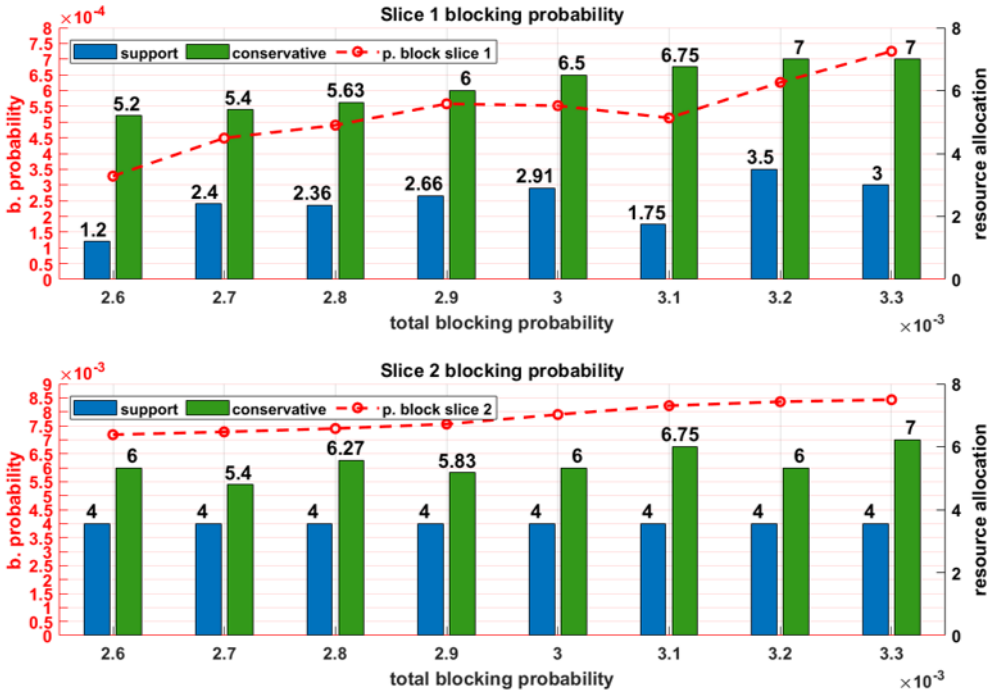


Figure 4.25: Slice resource mode allocation

4.3.4.3 Testbed Description and Results Evaluation

This subsection explains the testbed settings and performance for an eMBB user traffic scenario, under the application of the optimal slice modes configuration derived from subsection 4.3.4.2 experiment. Fig. 4.27 illustrates the testbed model for a single tenant system equipped with two slices, one eMBB and one mMTC. From a technological perspective, at the time of testing our scenario, the OAI platform offered a bandwidth of 10 MHz at 2.5 GHz, reaching up to 30-32 Mbps in downlink. For this reason, starting from an initial equal bandwidth division among the two system slices, we have decided to consider a high load downlink eMBB traffic (corresponding to 80% of the total system capacity) able to fully trigger the radio resources sharing capabilities of our proposed model. Moreover, in order to further highlight the radio resource flow migration from a *Support* towards a *Critical* slice, following a heuristic approach we identified the optimal thresholds (*Support* and *Conservative*) values for each slice such that the correct amount of radio resources are moved from the *Support* slice until the *Critical* slice returns to a stable status (meaning the slice is not in *Critical* mode anymore). Despite the fact that the testbed architecture is based on 4G, the aforementioned system setup represents an optimal solution for testing 5G-based features and functionalities. At the time of testing, only primitive open-source 5G testbeds are available, with limited capabilities and functionalities.

To simulate the 5G traffic requirements, we developed a client-server traffic generator and analyzer platform, written in Python, which scales the 5G traffic features according to the testbed capabil-

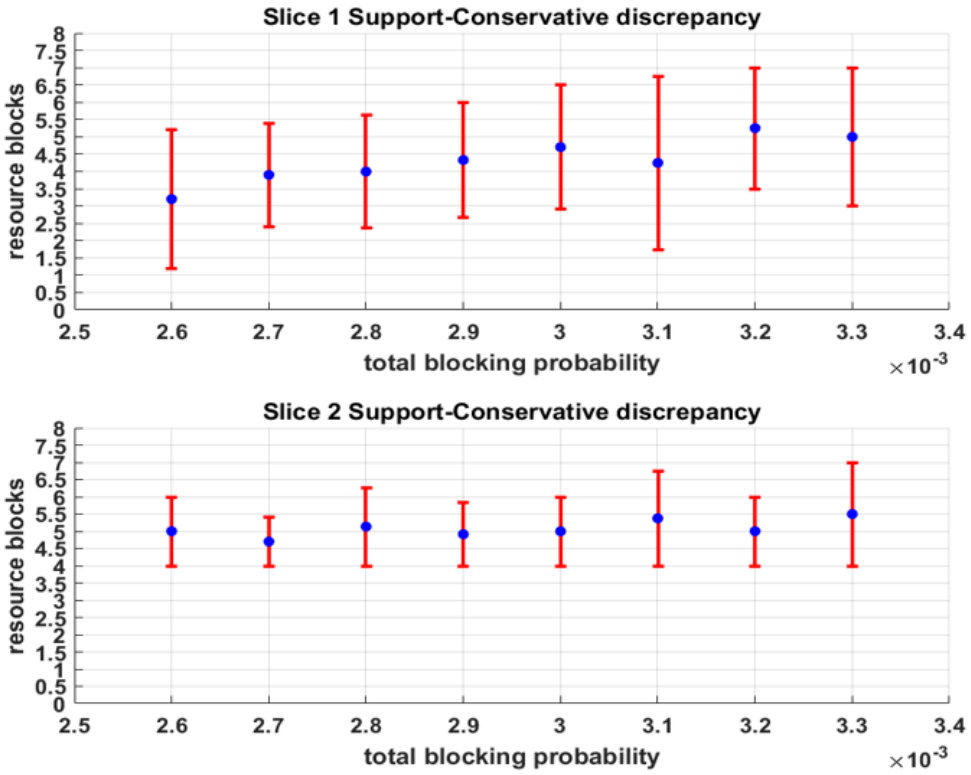


Figure 4.26: Slice modes discrepancy

ities. Multiple users can be served since the client application is based on SW containers. This feature permits to simultaneously manage multiple user traffic requests for the same slice service type. In addition, the deployment of the RAN and CN on two separated machines guarantee better workload balancing inside the system. Moreover, it enhances the solution flexibility in terms of centralized or distributed management, and the deployment of multiple split options as standardized in 5G (96).

Table 4.5: System and testbed parameters

Total duration	70 sec	**Direction**	Downlink
Nr. slices	2	**Support eMBB**	20%
Bandwidth	10 MHz (50 RBs)	**Conservative eMBB**	40%
MCS DL	28	**Support mMTC**	35%
MCS UL	8-20	**Conservative mMTC**	45%
Granularity	2 sec.	**Initial slice PRB division DL**	25-25
Protocol	UDP	**Initial slice PRB division UL**	25-25

Table 4.5 summarizes the parameters necessary to configure the SM NF. The scenario is firstly emulated offline using the analytical model (subsection A) to determine the set of *Support* and *Conservative* modes which guarantee the lowest *total blocking probability* for each slice. Then, these parameters are forwarded to the SM, and the dynamic NS performance of the current scenario is compared against the same scenario for the static slicing case, used as benchmark.

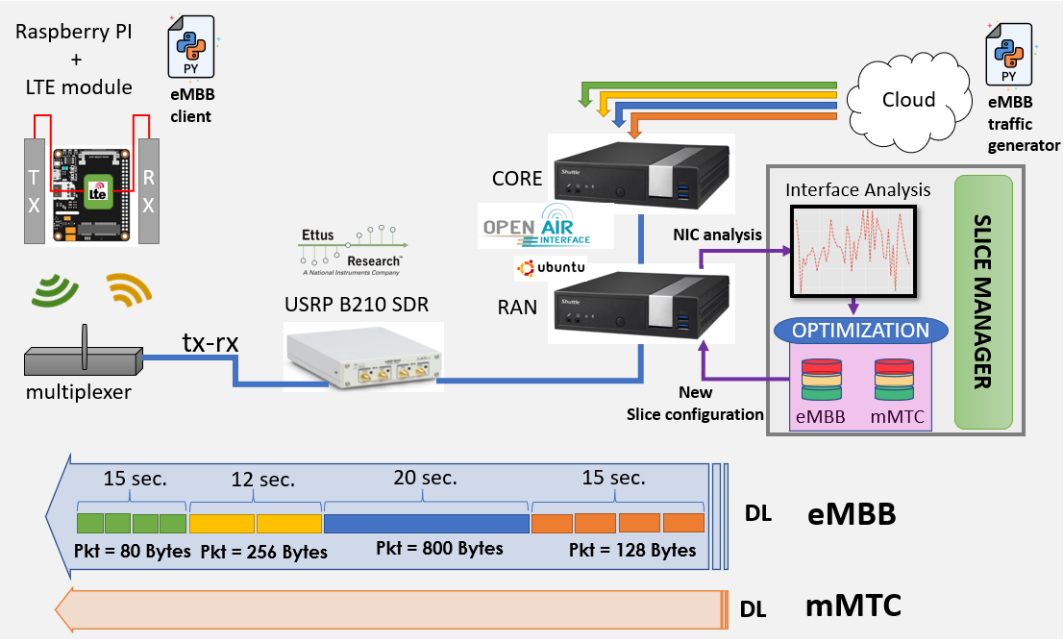


Figure 4.27: Two slices testbed design

Fig. 4.28 illustrates the performance of the testbed scenario, for three different cases. Given a Poisson input traffic able to occupy approximately 80% of eMBB slice capacity during the entire simulation, the red line in Fig. 4.28 shows the response trend for the static slicing case. The eMBB slice capacity is fixed at 25 PRBs, which corresponds to a maximum average data rate of 15 Mbps in our testbed. While the input traffic has an average data rate of 17 Mbps, under static approach, the receiver only reaches an average data rate of 9-10 Mbps, which corresponds to a poor QoS. This low performance is the result of an inappropriate initial static allocation of the radio resources, further affected by the traffic congestion at the transmitter side and reduced MCS.

The second experiment (green line, *dynamic_default*) illustrates the performance of our proposed dynamic slicing solution, when a equal PRBs partition is applied among each mode of the two slices: *Support* mode equals to 30%, and *Conservative* mode equals to 80%. As illustrated, even though this second case does not constitute the global optimum solution of our model, it performs better than the static approach (2-3 Mbps more on the receiver side), proving that even low-accuracy dynamic solutions are able to outperform static methods.

The desirable performance is achieved with the third case (blue line, *dynamic_opt*), where the computation of the slice's modes through the analytical model guarantees low blocking probability and high QoS for the eMBB services. As it is shown, the input data traffic is completely served to the end users. Moreover, given the stochastic behavior of the wireless channel, unexpected traffic peaks are also correctly handled due to the PRBs safety gap between the *Conservative* and total capacity of the eMBB slice. This phenomena is evident after 30 seconds experimental time, where two consecutive traffic peaks are received and promptly handled, reaching the maximum testbed

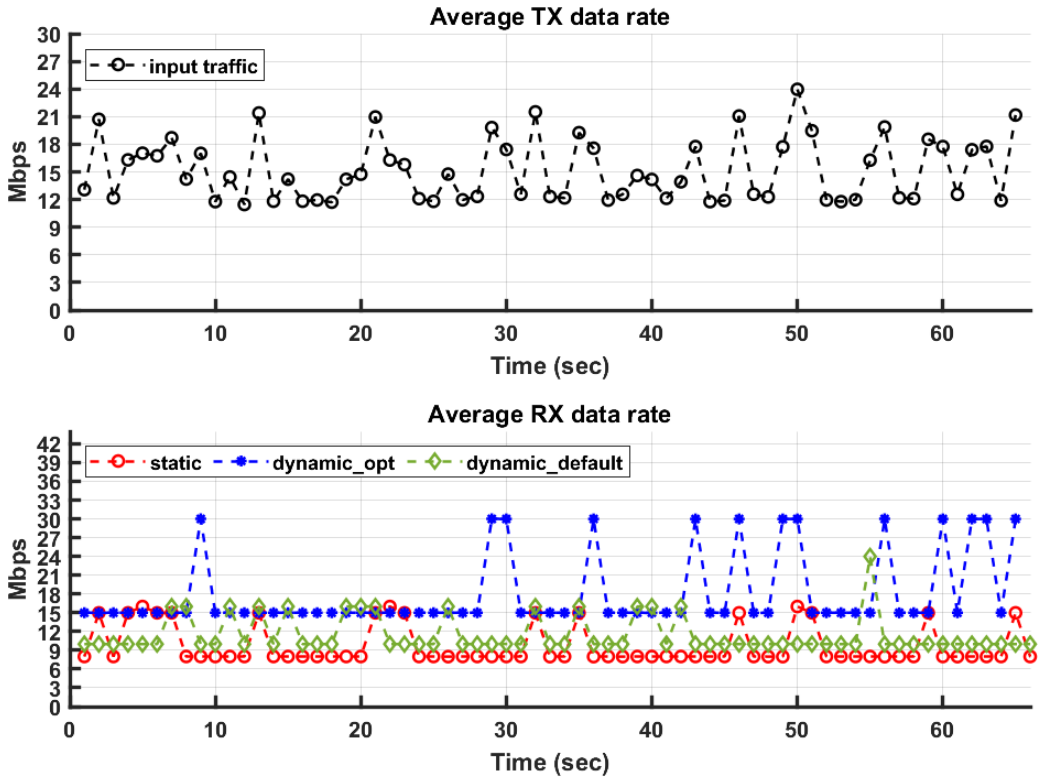


Figure 4.28: Slice configuration performance: a) Av. TX data rate b) Av. RX data rate

data rate capacity of 30-32 Mbps, without affecting the integrity of our platform and the next service traffic.

The benefits of the proposed method are further confirmed in Fig. 4.29, where the Bit Error Rate (BER) of the optimal dynamic solution is compared with the static approach. As expected, the percentage of error is higher for the static approach, with peaks up to 0.57%. The low BER is the consequence of limited capabilities in terms of customization and management of the RAN resources, which is a well-known drawback of static slicing approaches, since the SM does not adjust the slice resources according to the real-time traffic behavior.

On the other side, the dynamic approach presents just three peaks up to 0.24%, in line with some of the most critical traffic peaks from the transmitter side. Despite the time consuming and increasing computational processing time due to the real-time investigation of the optimal slice configuration, dynamic NS presents a low BER, since our algorithm performs the slice reconfiguration only after the real-time evaluation of all the slices performance. Moreover, taking into account the complexity behind the elaboration of a suitable RAN slices resources division in a very tight time interval, the average BER of the dynamic approach is around 0.10%, which is acceptable for this type of real-time environment. This final experiment proved that the proposed SM NF constantly analyzes the input traffic and regulates accordingly the capacity of each slice, ensuring a low BER even under heavy traffic scenarios.

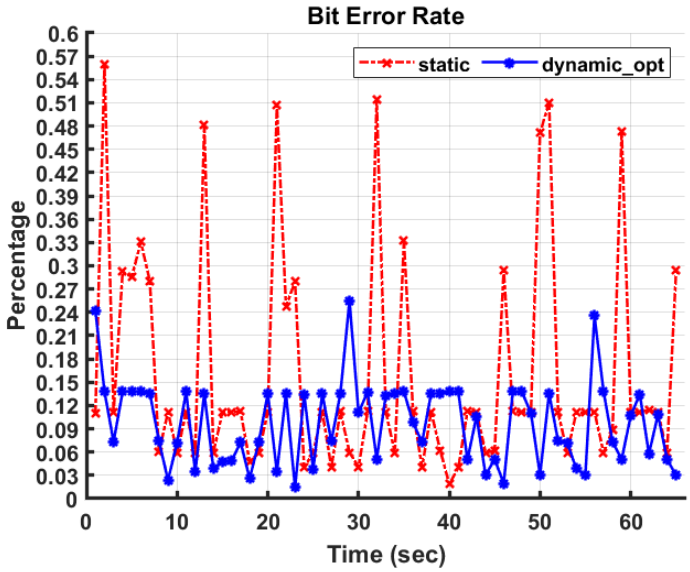


Figure 4.29: Static vs optimal dynamic slice parameterization

4.4 5G Fronthaul Multi-Service Customization: a Dynamic Functional Split and Hierarchical Double-tier Scheduler 5G NR Framework

4.4.1 Innovation and Contribution

Through NS, techniques such as cloud computing, edge computing, NFV, and SDN can be orchestrated in an effective manner, persuading SPs to discover E2E network slice orchestration and operations solution able to create and provision network slices spanning over the different network domains. Different one-size-fits-all orchestrators have been already faced in literature (97) (98) (99) (100), which aim at orchestrating the entire E2E networks, providing a central point of management for the entire E2E infrastructure. However, a real 5G implementation of the proposed solutions is difficult to realize, since: i) they tend to oversimplify each network domain with a limited sets of technologies and standards (101), making extremely complex extending new functionalities and features, and ii) they focus principally on the process of instantiating and deploying the NS, while ignoring how they are enforced in the mobile network (102).

From a management and orchestration perspective, traditional cloud-centric NS solutions are limited, since the high number of virtual and physical resources are spread across different technological domains (AN, CN, and TN, and administrative domains), which observe proprietary regulations

This contribution of this chapter has been published in the following publication: Maule, M., Vardakas, J. S., Verikoukis, C., "Multi-service network slicing 5G NR orchestration via tailored HARQ scheme design and hierarchical resource scheduling", IEEE Transactions on Vehicular Technology, 2022.

and tools (103). Moreover, these solutions are mainly business driven and only address the needs of use-cases, which do not reflect the characteristics of the RANs in the NS instance (104). For this reason, enforcing slices in the RAN still remains a difficult task, since it is challenging to provide different levels of isolation to a RAN slice owner, to cleverly customize the services across multiple planes, and effectively allocate the available radio resources to enable the multiplexing gains.

In this light, this work presents a real-time NS RAN orchestration framework, which models the allocation of network resources and placement of NFs in a single optimization model. Given the FH infrastructure and SLA delay constraints, the FS technique is employed to arrange the network service functionalities between the CU, the DUs, and RRUs, introducing massive cost reduction for the SPs, harmonization of the infrastructure design, and maximization of the traffic KPIs. Among the RAN slices, the radio resources are shared and dynamically distributed, following a cross-correlated optimization between the real-time traffic load and data rate SLA. For each slice, the users' flow is prioritized following a weighted distribution, proportional to the maximum slice capacity.

In terms of performance, while LTE presents a HARQ scheme with subframe granularity, our method exploits the slot-based TTI flexibility of 5G, executing the HARQ RTT response 37%-88% times faster than LTE. At RAN inter-slice level, the classical static NS approach are outdated, since fast service responses entail quasi-real-time radio resources' parameterization. Our dynamic inter-slice radio resources allocation approach improves the resource's utilization up to 50% compared to static NS method, enhancing the service continuity. In the last optimization phase, within each slice, fair RBs allocation (i.e., Round Robin scheme) does not further distinguish among the users, which are however characterized by distinct traffic characteristics even if associated to the same slice. For this reason, with our proposed method, at intra-slice level the under- and over-resource provisioning has been decreased by 21.35% and 60.08%, respectively, allowing a higher number of users to be served simultaneously, without affecting the slice' SLA.

Using as reference Fig. 4.30, the contributions and novelties of this work are summarized as follows:

1. The phase 1 is subdivided into two sub-parts. Given the physical infrastructure delay constraints, we propose an adaptive FS algorithm aimed at satisfying the latency SLAs of all the users served in a slice. The resulting split option defines the NFs placement on top of the physical architecture, and the RTT delay upper bound of the slice's service, which is integrated in the maximization problem to determine the DL-UL scheduling timing parameters (K0, K1, K2) per user's traffic (105).

2. In phase 2 we propose an autonomous inter-slice scheduling algorithm able to dynamically orchestrate the radio resources of the RAN slices. This work is an extension of (106), where distinct resource sharing policies are applied according to the real-time slice load. This algo-

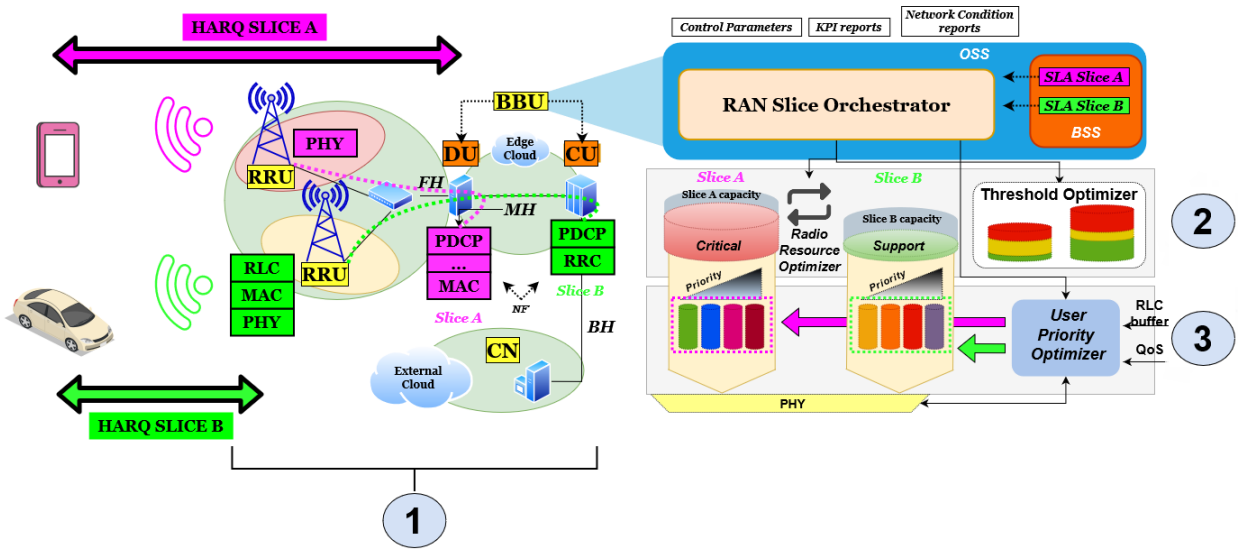


Figure 4.30: RAN slicing orchestrator design over 5G E2E architecture

rithmic design minimizes the blocking probability of incoming user's request, while preserving the SLA.

3. In phase 3, given the number of users served per slice, we develop a priority-based intra-slice algorithm to perform the users' queuing order proportionally to their traffic load. Unlike the previous phase, the resource granularity is raised at user's level. The weight distribution is designed using a game-theoretic approach based on the SOA, which aims to maximize an overall network utility function. Additionally, the user's weight regulates the radio resources' distribution on top of the physical radio resource grid, maximizing the resources' utilization.

With the proposed framework, through the division of the proposed RAN orchestrator framework into three phases, we provide a structured, up-to-date, and hierarchical management of the whole RAN slicing life cycle, by taking advantage of the latest features introduced with 5G such as scalable numerologies, variable TTI, distributed processing, and flexible timing. To the best of our knowledge, this is the first work that correlates all the aforementioned to maximize the RAN resources' utilization,, since most of the current RAN NS management methods prefer to control separately these paradigms, considering high level resource granularity abstraction (107), independent slice-based flow optimization (108), and long-term slice life cycle implementation (109).

4.4.2 5G NR Flexible Scheduling Timing

Inline with the theoretical background of subsection 2.4.2.2, the DL HARQ RTT is measured from the time instant the user receives the first DL-PDCCH, until the retransmitted data are correctly decoded in the corresponding PDSCH subframe. By exploiting the flexibility of the 5G HARQ RTT, the algorithm proposed in subsection 4.4.4.2 aims to maximize the number of slots between

K0 and K1 values, introducing multiple advantages in the RAN:

- The gap between K0 and K1 regulates the delay between the HARQ feedback and the retransmission. Fixed the slot duration, greater is the gap duration, more relaxed are the latency constraints. In this case, the processing capability is migrated towards the CU unit, while preserving the latency constraints.
- Multiple contiguous PDSCH are acknowledged within the same PUCCH, thus reducing the signaling and processing time of the frame structures. This method allows a higher amount of data can be transmitted within the same TTI, reducing the number of the HARQ feedback responses.
- Using a centric management approach, the performance of the network entities converge to a single node of the FH infrastructure. This accelerates the processing time in response to a new slice configuration, since the synchronization with the remaining nodes is not needed, and a cooperative resource allocation is performed.
- Reduced CAPEX and OPEX for the infrastructure and SP is achieved. Although MEC solutions are beneficial for low latency applications, they are costly and expensive to maintain.

4.4.3 Dynamic NS and Game Theory Applications

A key problem underlying NS refers to the efficient sharing of radio network resources, by avoiding unnecessary overheads and complexity. An early solution proposed by 3GPP is called *static* NS (110), where the slices are already instantiated, and the UE has to proceed with the slice selection. However, since the slice traffic load is spatially inconsistent and time varying, *dynamic* resource allocation schemes permit the tenants to apply customized strategies by monitoring the KPIs, so as to maximize their own utility. In this work, the proposed model combines a dynamic allocation solution with a bidirectional communication channel between the tenant and the RAN infrastructure orchestrator in order to exchange service preferences, and improve the management of the slice's resources to better customize the allocation of the customers. This type of resource optimization problem can be modelled as a Fisher market model (111), which is one of the most fundamental resource allocation schemes in economics, with multiple applications in computer science and algorithmic game theory approaches.

This model guarantees the existence of a market equilibrium under mild conditions (112), which can be efficiently computed when the utility function belongs to the CES class. In particular, in

this work, we refer to the following CES utility function:

$$u_i(x_i) = \sum_{j=1}^{m} (a_{ij} \times x_i^{p_i})^{1/p_i}, -\infty < p \leq 1, p \neq 0 \tag{4.15}$$

where $u_i(x_i) : [0;1]^m \rightarrow R$ represents the buyer's utility when receiving x_i amount of products, $p \in (0;1]$ parameterizes the family, and a_{ij} is a weight quantifies the impact of j on the buyer i's utility.

Even though the Fisher market model have been heavily investigated, most of the studies consider linear utility functions ($p = 1$), which is characterized by an excessive unfair resource allocation when customization policies are applied to each buyers (113). As it will be illustrated in Subsection 4.4.4.4, phase 3 takes advantage of the Fisher market model with the linear utility function ($p = 1$) to define a intra-slice scheduling algorithm where UEs are prioritized according to a proportional evaluation of their specific real-time traffic flow characteristics.

4.4.4 Three-phases Optimization Framework: Model Design and Mathematical Formulation

4.4.4.1 System Design and Characterization

Let S_n^j denote the slice $n \in (n = 1, ..., N^j)$ in the system belonging to the tenant $j \in (1, ..., J)$, where N^j is the number of slices of the tenant j and J is the total number of tenants in the system. Each slice S_n^j is characterized by an initial capacity C_n^j such that:

$$\sum_{j=1}^{J} \sum_{n=1}^{N^j} C_n^j = C_{rrh}^{cell,j} \tag{4.16}$$

where $C_{rrh}^{cell,j}$ represents the total cell capacity of the Remote Radio Head (RRH) element rrh, where $rrh \in (1, ..., R_{rrh})$, and R_{rrh} is the total number of RRHs in the system. Each slice S_n^j is characterized by the following parameters:

$$\begin{aligned} &[\ \boldsymbol{D_n^{j,harq}};\ \boldsymbol{R_n^j};\ th_{n,1}^j;\ th_{n,2}^j;\ \boldsymbol{W_n^j}] \\ &0 < th_{n,1}^j < th_{n,2}^j < 1 \end{aligned} \tag{4.17}$$

where, for each tenant j:

- $\boldsymbol{D_n^{j,harq}}$ is the vector of the RTT HARQ delays of the users served from the slice n,
- $\boldsymbol{R_n^j}$ is the vector of required guaranteed bitrate of the users served by the slice n,
- $th_{n,1}^j$ is the percentage of *Support* mode radio resources assigned to the slice n,

Table 4.6: General purpose scenario parameters

Parameter	Description	Parameter	Description
S_n^j	Slice n of tenant j	$\boldsymbol{U_n^j}$	Users of slice S_n^j
C_n^j	Initial capacity slice n of tenant j	$C_{rrh}^{cell,j}$	Maximum cell capacity
$th_{n,1}^j$	*Support mode* slice n of tenant j	$th_{n,2}^j$	*Conservative mode* slice n of tenant j
N^j	Number of slices of tenant j	J	Number of tenants in the system
R_{rrh}	Number of RRH in the system	$\boldsymbol{D_n^{j,harq}}$	RTT HARQ delay vector of each UE served by the slice n
$\boldsymbol{R_n^j}$	average expected throughput vector of each UE served by the slice n	$\boldsymbol{W_n^j}$	vector of UEs'weights of slice n

- $th_{n,2}^j$ is the percentage of *Conservative* mode radio resources assigned to the slice n,
- $\boldsymbol{W_n^j}$ is the vector of users' weights belonging to the slice n.

The meaning of the *Support* and *Conservative* modes will be illustrated in Subsection 4.2.3, while Table 4.6 summarizes the main variables used in our model. To model each 5G service, we refer to the system parameterization illustrated in (114), which analyses fundamental multi-tenant 5G use-cases and their corresponding scenarios and traffic models.

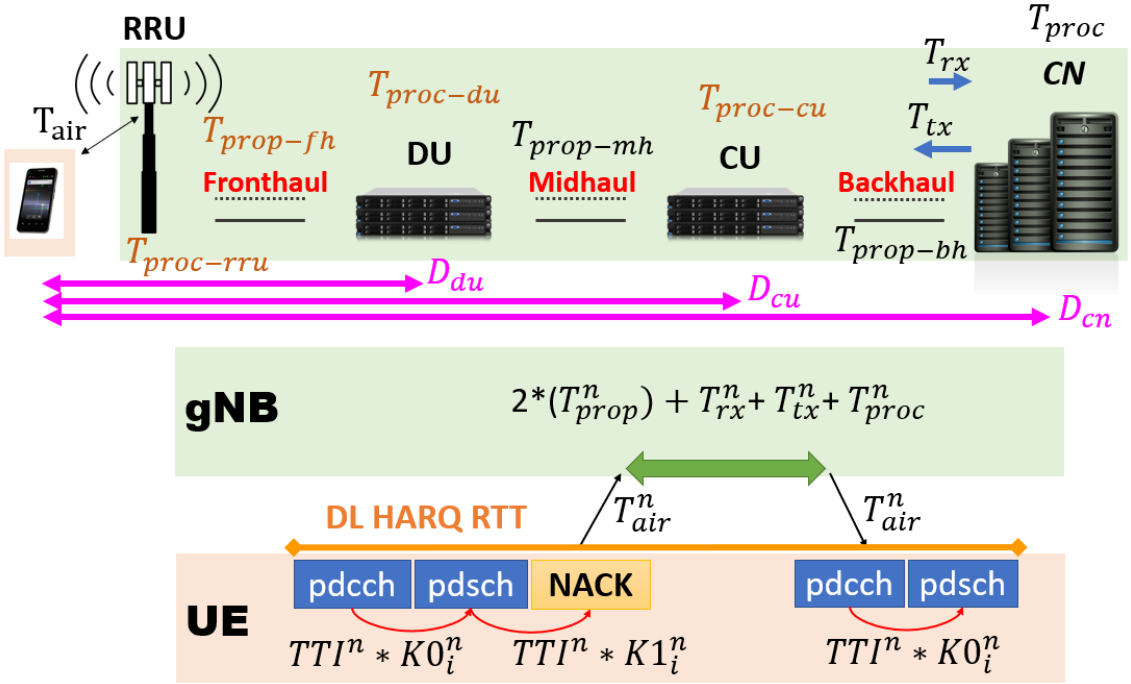


Figure 4.31: HARQ RTT parameters for a generic purpose scenario

Table 4.7: Infrastructure timing parameters

Parameter	Description	Parameter	Description
T_{air}	Air time	$T_{proc-rru}$	RRU processing time
$T_{prop-fh}$	FH propagation time	$T_{proc-du}$	DU processing time
$T_{prop-mh}$	MH propagation time	$T_{proc-cu}$	CU processing time
$T_{prop-bh}$	BH propagation time	T_{tx}	Transmission time
T_{rx}	Receiver time	$T_{proc-cn}$	CN processing time
D_{du}	RTT delay user-DU	D_{cu}	RTT delay user-CU
D_{cn}	RTT delay user-CN	TTI_n^j	TTI value of slice n in tenant j

4.4.4.2 Phase 1: DL HARQ Timing Parameters and 5G NR FP Selection

As illustrated in Fig. 4.31, the network infrastructure is characterized by multiple delays, introduced by the physical links and processing units. This affects the selection of the optimal slice FS option, since the placement of the 5G MAC NF is crucial for the management of the HARQ procedures. In our approach, inline with the 5G standard, users with similar service requirements belong to the same slice type. For this reason, given the HARQ RTT ($D_{u,n}^{j,harq}$) for each user u in slice n of tenant j, the latency constraint of the FS option FS_n^j is formulated as:

$$FS_n^j = \frac{1}{|U_n^j|} \times \sum_{u}^{U_n^j} D_{u,n}^j \quad \forall n \in j$$

$$D_{u,n}^j = \begin{cases} (T_{air}^n + T_{prop-mh}^n + T_{prop-bh}^n) \times 2 + T_{rx}^n + \\ T_{tx}^n + T_{proc}^n \text{ if } D_{cn} \leq D_{u,n}^{j,harq}, \\ (T_{air}^n + T_{prop-mh}^n) \times 2 + T_{rx}^n + T_{tx}^n + T_{proc}^n \\ \text{ if } D_{cu} \leq D_{u,n}^{j,harq} < D_{cn}, \\ (T_{air}^n) \times 2 + T_{rx}^n + T_{tx}^n + T_{proc}^n \\ \text{if } D_{du} \leq D_{u,n}^{j,harq} < D_{cu}. \end{cases} \tag{4.18}$$

where the $|U_n^j|$ is the number of users served by slice u, and $D_{u,n}^j$ is the infrastructure delay constraint required by the RTT HARQ of user u in slice n. The Eq. 4.18 represents the average infrastructure RTT HARQ delay needed to satisfy all the users served by slice n, such that the users' SLA are guaranteed. A detailed explanation of the $D_{u,n}^j$ network infrastructure delay is summarized in Table 4.7.

From a theoretical perspective, the FS concept is represented using four macro classes, which gather together the 3GPP FS options where the MAC NF is placed in the same network domain (115):

- *RRU option* ($FS_n^j < D_{du}$): it includes all the FS options where MAC layer is placed with the

RRU unit.

- *Fully distributed* ($D_{du} \leq FS_n^j < D_{cu}$): it includes all the FS options where MAC layer is placed in the DU unit.
- *Hybrid implementation* ($D_{cu} \leq FS_n^j < D_{cn}$): it includes all the FS options where MAC layer is placed in the CU unit.
- *Fully centralized* ($D_{cn} \leq FS_n^j$): it includes all the FS options where MAC layer is placed close to the CN unit.

In the definition of $D_{u,n}^j$, we decided to neglect the propagation and processing delays of Fig. 4.31, since they do not produce any remarkable variation to the final FS option selection.
Through the aforementioned procedure, the optimal FS option FS_n^j defines a latency bound to all the users in n. However, in our work, the service reliability is further enhanced through an ad-hoc customization of the DL HARQ timing parameters ($K0_{u,n}^j$, $K1_{u,n}^j$) of each user u served by slice n of tenant j. By recalling subsection 4.4.2, maximizing the number of slots between $K0_{u,n}^j$ and $K1_{u,n}^j$ has multiple advantages from a management perspective. This operation considers the RTT infrastructure latency constraint introduced with FS_n^j, and the different numerologies standardized in (116). For this reason, in order to have a realistic configuration, three types of numerologies and corresponding TTI values are selected, suitable for different use-cases:

- numerology 0 ($\mu = 0$, SCS = 15 KHz) and 1 ms TTI.
- numerology 1 ($\mu = 1$, SCS = 30 KHz) and 0.5 ms TTI.
- numerology 2 ($\mu = 2$, SCS = 60 KHz) and 0.25 ms TTI.

According to our investigation, this selection represents an optimal trade-off between the current state-of-the-art (117), and real 5G implementation at Frequency Range 1 (FR1) (118). For each slice n of tenant j, the timing parameters $K0_{u,n}^j$ and $K1_{u,n}^j$ of each user u are defined as:

$$\max_{K0^j_{u,n};K1^j_{u,n}} [K0, K1]^j_{u,n} \quad \forall n \in j$$

$$s.t.$$

$$[K0, K1]^j_{u,n} = \sum_{u=1}^{U^j_n} TTI^j_n \times (K1^j_{u,n} - \beta - 2K0^j_{u,n}) + D^j_n,$$

$$\beta = 2 \times (\frac{\mu^{pdsch}}{\mu^{pdcch}}), \tag{4.19}$$

$$K0^j_{u,n} \in (0, 32),$$

$$K1^j_{u,n} \in (0, 15),$$

$$0 < D_{du} < D_{cu} < D_{cn},$$

$$0 \leq K0^j_{u,n} < K1^j_{u,n},$$

$$TTI^j_n \times (K1^j_{u,n} - \beta - 2K0^j_{u,n}) + D^j_n \leq D^{j,harq}_{u,n} \leq FS^j_n,$$

where TTI^j_n is the TTI of slice n of tenant j, U^j_n is the number of users served by slice n, and β is a correction factor due to different numerologies between control (μ^{pdcch}) and data (μ^{pdsch}) channels. If not specified, the default value for K0 is zero, as defined in (119).

In our formulation, all the infrastructure and service delay parameters are embedded in Eq. 4.19, enhancing the cooperative capabilities of the FH infrastructure, since each $[K0, K1]^j_{u,n}$ couple is estimated considering cause-effect relationship over the entire RAN orchestration. To maintain the RAN parameterization up to date with the network and service performance, the optimal frame pattern can be dynamically reconfigured according to the RAN traffic load. Similar approach is considered to perform dynamic FS, where a new NF placement is evaluated for the entire system every time a slice is instantiated or removed from the service infrastructure.

4.4.4.3 Phase 2: inter-slice scheduler and thresholds optimization

Once the FS per slice is applied using Phase 1, our proposed orchestrator initiates the inter-slice scheduler for dynamically controlling the radio resources of each slice through the joint real-time evaluation of the slices' SLA per tenant with the users' KPIs. This method outperforms the ordinary dynamic radio slicing approaches where the radio resources are managed only from a tenant perspective. Moreover, since the proposed system is constantly aware of the real-time user's traffic KPIs, unexpected wireless channel fluctuations are promptly handled without affecting the service performance. For each tenant j and slice S^j_n, the amount of PRBs needed to serve the

corresponding users is defined as:

$$
\begin{aligned}
& PRB_n^j = \sum_{j=1}^{J} \sum_{n=1}^{N^j} \sum_{u=1}^{\boldsymbol{U_n^j}} \frac{r_{u,n}^j \times 10^3}{C_f \times \log_2(QAM_n^j) \times 2^\mu} \\
& s.t. \\
& C_f = N_{n,sc}^j \times N_v^{cell} \times f^{cell} \times R_{n,code}^j \times 168 \times (1 - OH_n^j), \\
& \sum_{j=1}^{J} \sum_{n=1}^{N^j} PRB_n^j \leq C_{rrh}^{cell,j}
\end{aligned}
\tag{4.20}
$$

where $r_{u,n}^j$ is the data rate required by the user u, QAM_n^j is the modulation order, μ is the slice numerology, $N_{n,sc}^j$ is the number of carriers, N_v^{cell} is the number of layers, f^{cell} is a scaling factor, $R_{n,code}^j$ is the coding rate, and OH_n^j is the overhead factor.

Since the PRBs are a finite quantities, resource sharing is a key enabler for the 5G NR. In Algorithm 1, we propose our inter-slice resource sharing technique, which defines the optimal amount of radio resources per slice, without decreasing the performance of the others. The values of $th_{n,1}^j$ and $th_{n,2}^j$ are set initially randomly, and adjusted gradually. This solution is divided into two sections:

1. *Slice Mode Selector* (line 10 to 26): this functionality is responsible for assigning the corresponding mode to each slice according to the real-time PRB load. Each mode has distinct radio resource sharing policies:

   - *Support* mode ($0 < PRB_n^j \leq th_{n,1}^j \times C_n^j$): the slice accepts incoming service requests, and it can share radio resources (when necessary) with the other slices.
   - *Conservative* mode ($th_{n,1}^j \times C_n^j < PRB_n^j \leq th_{n,2}^j \times C_n^j$): the slice prioritizes its own traffic by disabling sharing of radio resources with others.
   - *Critical* mode ($th_{n,2}^j \times C_n^j < PRB_n^j \leq C_n^j$): the amount of allocated resources does not fully satisfy the SLA. If exists at least another slices in *Support* mode, radio resource sharing from the *Support* slice to the *Critical* slice can be activated.

   When a slice is labelled *Critical*, a capacity adjustment procedure is activated if exists at least another slice in *Support* mode, otherwise a new capacity configuration able to satisfy the system load, the *Threshold Optimizer* method is called.

2. *Threshold Optimizer* (line 28 to 34): for each slice, it aims to maximize $th_{n,1}$ in order to increase the probability of one or more slices to be categorized as *Support* mode, and share radio resources. The constraint on the *slice blocking probability* $P_n(i, C_n)$ guarantees that no slice violates its SLA by exceeding the maximum *system blocking probability* P_n^j. Through $P_n(i, C_n)$, the slice radio capacity is dynamically adjusted, preventing under-over provisioning

of radio resources. An exhaustive mathematical analysis of the role of the *slice blocking probability* is provided in our previous work (106).

The recursive structure of the proposed algorithm decreases the system time complexity. The PRB_n^j is also recomputed recursively (line 4) such that the resources per slice are up-to-date with the system evolution. Advanced details of the sharing capabilities and bandwidth utilization are illustrated in (91).

The presented slice resource optimization algorithm guarantees a stable FH system when distinct services run simultaneously on the same radio infrastructure.

Algorithm 1: Slice_Capacity_Customization()

input:
$th^j_{n,1} \leftarrow$ Support threshold slice n
$th^j_{n,2} \leftarrow$ Conservative threshold slice n
$PRB^j_n \leftarrow$ Slice n PRB load
$C^j_n \leftarrow$ Slice n initial capacity
output:
$[C^j_n, th^j_{n,1}, th^j_{n,2}]\ \forall S^j_n \in k$

start:
for each $k \in \mathcal{I}^j_{rrh}$ **do**
  for each $S^j_n \in k$ **do**
    if $0 \leq PRB^j_n < th^j_{n,1} \times C^j_n$ **then**
      set *Support* mode.
    end if
    if $th^j_{n,1} \times C^j_n \leq PRB^j_n < th^j_{n,2} \times C^j_n$ **then**
      set *Conservative* mode.
    end if
    if $th^j_{n,2} \times C^j_n \leq PRB^j_n < C^j_n$ **then**
      set *Critical* mode.
      if $\exists$ slice $n_{su} \in$ *Support* **then**
        while $n \in$ *Critical* $\vee\ n_{su} \in$ *Support* **do**
          $\rightarrow$ increase C^j_n
          $\rightarrow$ decrease $C^I_{n_{su}}$
        end while
      end if
      Slice_Capacity_Customization()
    else
      ${}_{th^j_{n,1}}(th_{n,1})\ , \forall S^j_n \in k$
      s.t.
      $\frac{\lambda^j_n}{\sum_{N^j} \lambda^j_n} \sum_{u=\lfloor th^j_{n,1} \cdot C^j_n \rfloor}^{\lfloor th^j_{n,2} \cdot C^j_n \rfloor} P_n(u, C^j_n) \leq P^j_n,$

      $0 < th^j_{n,1} < th^j_{n,2} < 1$
      Slice_Capacity_Customization()
    end if
  end for
end for
stop:

4.4.4.4 Phase 3: Intra-Slice Scheduler: User Weight Definition

Once the radio resources are arranged among the slices, most of the dynamic RAN NS solutions in the literature apply classic scheduling algorithms (i.e., Round Robin) at MAC layer to serve the users, without considering supplementary capabilities introduced with 5G traffic characterization. In this subsection, we believe it exists an extra margin of improvement at intra-slice layer (label 3 of Fig. 4.30), where each user can be scheduled considering also its own traffic characteristics, and not only through the slice SLA from the SP. For each user u served from the slice S_n^j of tenant j, a weight $w_{u,n}^j \in \boldsymbol{W_n^j}$ proportional to the user's real-time traffic load is defined. The weight $w_{u,n}^j$ corresponds to the percentage of physical resources assigned to the user u, given the total amount of physical resources per slice C_n^j. The weight distribution is calculated with the following game-theoretic approach:

$$
\begin{aligned}
& max(U(\boldsymbol{W_n^j})) = \sum_{n=1}^{N^j} \sum_{u=1}^{\boldsymbol{U_n^j}} \varphi_{u,n}^j \times f_q(w_{u,n}^j) \quad \forall j \in J \\
& \text{s.t.} \\
& f_q(w_{u,n}^j) = \begin{cases} \frac{1}{1-\alpha_q} \left(\frac{w_{u,n}^j \times PRB_{u,n}^j}{\sum_{\boldsymbol{U_n^j}} w_{u,n}^j}\right)^{1-\alpha_q}, & \text{if } \alpha_q \neq 1 \\ \log\left(\frac{w_{u,n}^j \times PRB_{u,n}^j}{\sum_{\boldsymbol{U_n^j}} w_{u,n}^j}\right), & \text{if } \alpha_q = 1 \end{cases} \\
& \sum_{u=1}^{\boldsymbol{U_n^j}} w_{u,n}^j = C_n^j, \qquad (4.21) \\
& w_{u,n}^j > 0, \\
& C_n^j > 0, \\
& \varphi_{u,n}^j > 0, \\
& PRB_{u,n}^j = \frac{r_{u,n}^j \times 10^3}{C_f \times \log_2(QAM_n^j) \times 2^\mu}, \\
& C_f = N_{n,sc}^j \times N_v^{cell} \times f^{cell} \times R_{n,code}^j \times 168 \times (1 - OH_n^j),
\end{aligned}
$$

where $PRB_{u,n}^j$ is the radio capacity of the user u of slice n, $\boldsymbol{U_n^j}$ are the users of slice n, $\varphi_{u,n}^j$ is the priority factor of the user u of slice n, f_q is the $\alpha_q fair$ utility function, $r_{u,n}^j$ is the data rate required by the user u, QAM_n^j is the modulation order, μ is the slice numerology index, $N_{n,sc}^j$ the number of carriers, N_v^{cell} the number of layers, f^{cell} is a scaling factor, $R_{n,code}^j$ is the coding rate, and OH_n^j is the overhead factor (121). As a payoff to the decision maker we take the generalized $\alpha_q fair$ utility function, where the proportional fairness resource allocation is reached for $\alpha_q = 1$. Other well-known utility functions are obtained for $\alpha_q = 0$ (linear utility), $\alpha_q = 2$ (delay fairness), and $\alpha_q \to \infty$ (max-min fairness). Multiple resource allocation schemes can satisfy the users' SLA.

However, the global optimal resource allocation policy is achieved when no users has incentive to reallocate. Given the $\alpha_1 fair$ utility function, we characterize the best slice strategy with the following lemma:

Lemma 1 *Assuming a set of slices $n \in (1, .., N^j)$, and corresponding users $u \in \boldsymbol{U_n^j}$, where $|N^j|$ and $|\boldsymbol{U_n^j}|$ are ≥ 2. Given a slice S_n^j with $\alpha_1 fair$ utility function, and a strategy $\hat{\boldsymbol{W_n^j}} \geq 0$ for each slice $S_{n'}^j \neq S_n^j$, the slice S_n^j best response $\boldsymbol{W_n^j}$ is the solution of the equation:*

$$w_{u,n}^j = \frac{\varphi_{u,n}^j}{\varphi_{u',n}^j} \times w_{u,n'}^j \times \frac{L(\hat{\boldsymbol{W_n^j}}) + \sum_{u' \setminus u \in U_n^j} w_{u',n}^j}{L(\hat{\boldsymbol{W_n^j}}) + \sum_{u \setminus u' \in U_n^j} w_{u,n}^j}, \quad \forall u \in \boldsymbol{U_n^j}$$
$$L(\hat{\boldsymbol{W_n^j}}) = \sum_{n' \setminus n}^{N^j} \sum_{u}^{U_{n'}^j} w_{u,n'}^j \tag{4.22}$$

where $L(\hat{\boldsymbol{W_n^j}})$ is the weight strategy of all the slice $S_{n'}^j \neq S_n^j$.

The proof of this result is technical and been relegated to the Appendix B. If Eq. (4.22) is applicable for each slice, the result system strategy reaches an important type of market equilibrium, called *Nash Equilibrium Strategy Profile* (NESP) (120), where no users can unilaterally deviate and get a better resource allocation. Even though it is not obvious how to compute one NESP efficiently, focusing on the $\alpha_1 fair$ utility function, the existence of at least one NESP is proven through the Rosen's theorem (122), while a technical demonstration of a NESP for generic $\alpha_q fair$ utility function is illustrated in (113), where the case of a log utility function is subjected to extra assumptions to be satisfied. For a more intuitive explanation of the NESP proof with $\alpha_1 fair$, this work provides in Subsection 4.4.5.4 a graphical representation and explanation of the best strategy among different slices, in a real case scenario. We believe this type of proof provides a more incisive understanding of the NESP for our system model.

4.4.5 Experimental Environment and Performance Analysis

4.4.5.1 Scenario Configuration

In this section, we validate the performance of our proposed algorithms, by using a 5G event-driven simulator based on Matlab R2019a (123), where the DL and UL transport channels have been modelled using the Matlab 5G-NR Toolbox. The simulated E2E multi-tenant scenario includes three tenants ($J = 3$), each one serving a different number of slices and services:

- Tenant A ($j = 1$): $S_1^1 = eMBB, S_2^1 = mMTC,$ and $S_3^1 = uRLLC$.
- Tenant B ($j = 2$): $S_1^2 = eMBB, S_2^2 = mMTC,$ and $S_3^2 = eMBB$.
- Tenant C ($j = 3$): $S_1^3 = eMBB, S_2^3 = mMTC, S_3^3 = uRLLC,$ and $S_4^3 = uRLLC$.

Table 4.8: Infrastructure delay and radio parameters

Parameter	**Value**	**Parameter**	**Value**
T_{air} (ms)	$\sim$N(180,1)$\times 3 \times 10^{-5}$	$T_{prop-fh}$ (ms)	$\sim$P(4)
$T_{prop-mh}$ (ms)	$\sim$P(10)	$T_{prop-bh}$ (ms)	$\sim$P(20)
T_{tx} (ms)	$\frac{1518\times\lvert U_n^j\rvert}{\sum r_n^j} \times 10^3$	T_{rx} (ms)	$\frac{1518\times\lvert U_n^j\rvert}{\sum r_n^j} \times 10^3$
T_{proc} (ms)	$\sim$P(5)	$C_1^{cell,1}$	100 MHz
$C_2^{cell,2}$	90 MHz	$C_3^{cell,3}$	100 MHz

As illustrated in Table 4.8, each network domain is interconnected using fiber-based connections subjected to multiple infrastructure delays, while in the RAN section, three rrh elements ($R_{rrh} = 3$) are installed, all working at band n78 TDD in FR1, with different bandwidth sizes. This radio configuration has been carefully selected such that there are the network conditions necessary to trigger all our optimization algorithms, and the reader has a clear vision of the sharing and optimization capabilities of the entire RAN orchestrator. Each slice is modelled as an $M\backslash M\backslash 1$ queue, where the average number of users in the slice depends from the type of service: 3 for eMBB, 8 for mMTC, and 6 for uRLLC. Since each slice is characterized by different service requirements, the following parameterization is applied to each user:

- eMBB users (127): data rate $r_{u,n}^j \in \mathcal{P}(80)$ Mbps, latency constraint $D_{u,n}^{j,harq} \in \mathcal{P}(20)$ ms,
- mMTC users (128): data rate $r_{u,n}^j \in \mathcal{P}(15)$ Mbps, latency constraint $D_{u,n}^{j,harq} \in \mathcal{P}(13)$ ms.
- uRLLC users (129): data rate $r_{u,n}^j \in \mathcal{P}(25)$ Mbps, latency constraint $D_{u,n}^{j,harq} \in \mathcal{P}(8)$ ms.

In aforementioned scenario, all the FH domain features have been included to emphasize the scalability of the proposed approach in an heterogeneous environment. Moreover, through the accurate calibration of the physical and radio resources, all the optimization phases of the proposed framework are triggered, maximizing the efficiency of the entire FH system.

4.4.5.2 Multi-slice FP Design and HARQ Timing Scheme

Given the users' traffic requirements and the infrastructure delays (see Table 4.8), we calculate the optimal FS option per slice, using the Eq. 4.18. In Fig. 4.32, each bar represents the average RTT delay per slice required to satisfy the corresponding users' HARQ RTT requirement. The average RTT delays introduced by the main processing units (DU, CU, CN) are displayed using horizontal dashed line. By recalling Subsection 4.4.4.2, each horizontal dashed lines can be seen as the threshold between the FS options. With an average RTT delay above D_{cn}, the eMBB slices of Tenant 1 and Tenant 2 are suitable for a *fully centralized* split, taking advantage of a centralized management of the NFs. A different split configuration should be adopted for the eMBB slice of

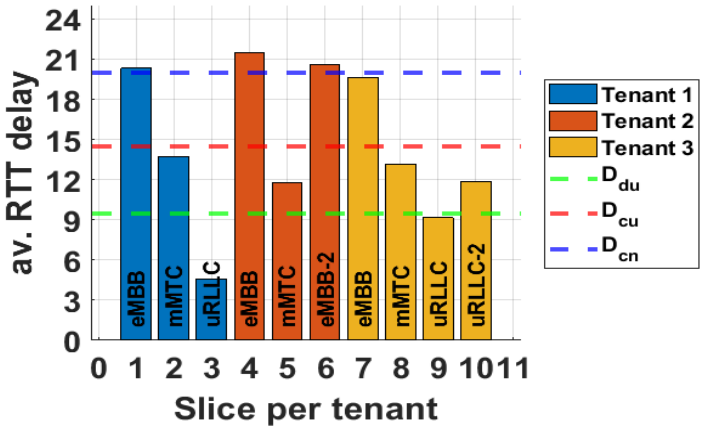


Figure 4.32: Functional Split candidate per slice

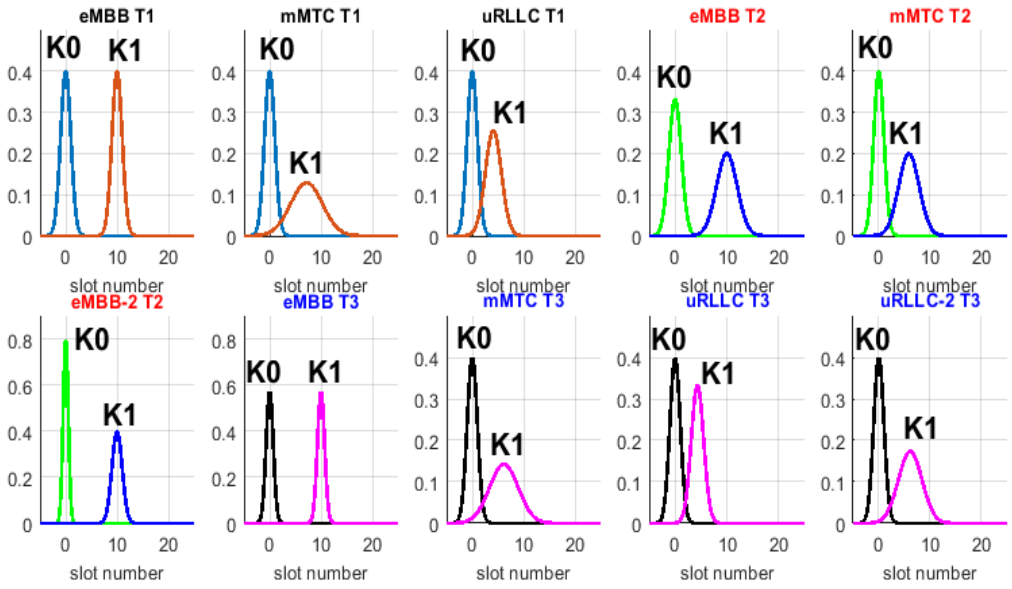


Figure 4.33: Slice-based HARQ parameters definition

Tenant 3, which is the only slice of the entire system requiring a *hybrid* FS installation, where distributed RRHs are controlled from a single BBU, as expected for C-RAN approach. All the mMTC slices and the one uRLLC slice from Tenant 3 require a *fully distributed* installation, since service latency requirements are more stringent, and the distributed NFs deployment is the only suitable approach to meet the infrastructure's delay requirements. Finally, the uRLLC slices of Tenant 1 and 2 require extremely low RTT latency, which is achieved with a fully implementation on the RRH unit.

This FS option requires an extra management effort since the resources should be controlled individually, without any centralization paradigm. To the best of our knowledge, the proposed solution represents the first FS deployment where the optimal NFs placement is established by considering the user's latency constraint at MAC layer. The majority of the FS methods in literature consider only the global latency SLA defined by the SP as main constraints, without refining the decision algorithm at user's flow (124)-(125).

In the second part of Phase 1, the users' timing parameters are calculated, considering 3 numerologies ($M_n^j = [0, 1, 2], \forall n, \forall j$) for each slice. The resulting frame patterns are compared with the current LTE TDD frame pattern in UL-DL configuration 3 (Table 4.2-2 of (126)). Fig. 4.33 illustrates for each slice of each tenant the average $K0$ and $K1$ values obtained from Eq. (4.19). Through the joint analysis of the tenant and user's service requirements, Phase 1 is able to estimate the optimal set of HARQ timing parameters per slice, which satisfies the slice FS RTT latency requirement estimated in 4.18. As expected, the eMBB slices present the larger interval in slots units between $K0$ and $K1$, since their services are characterized by relaxed latency constraints. With an average interval of $\approx$ 9.9 slots, multiple *ACKs* are embedded in the last slot before the TTI expires, promoting services with high DL traffic. Therefore, the resulting HARQ RTT (9.9 ms) is similar to LTE TDD (10 ms), since they have the same frame pattern and numerology ($\mu = 0$). On the other hand, extreme latency constraints require short TTIs, which corresponds to group the *ACKs* within a single subframe, every few slots. This is performed using a small interval in slot units between $K0$ and $K1$, as illustrated in Fig. 4.33, for each uRLLC slice. The average timing interval is $\approx$ 4.8 slot units, which is translated in an HARQ RTT from 52% to 88% times faster than LTE TDD. This is achieved using higher numerologies (μ = 1,2), where the slot duration variates between 0,5 and 0,25 ms. Between the two aforementioned services, mMTC slices accept higher flexibility in terms of frame timing parameterization, since the average interval is $\approx$ 6.34 slot units. The resulting sets of $K0$ and $K1$ parameters represent a trade-off between eMBB and uRLLC latency performance. For numerologies 0 and 2, the HARQ RTT is 37% and 85% times faster than LTE TDD, respectively.

4.4.5.3 Hierarchical Scheduler: Slice Optimization and User Priority Performance

In the previous subsection, the optimal FS option and timing parameters have been defined for each slice to accomplish the latency requirements. However, traffic congestion might still compromise the system equilibrium if the resources at PHY and MAC layers are intelligently orchestrated. Phase 2 of our proposed framework applies a dynamic radio resource allocation method able to satisfy the data rate constraint of each slice. During the initialization step of the *Algorithm 1*, given the maximum cell capacity for each tenant, a set of thresholds (*Support* and *Conservative*) is assigned randomly to each slice as follows:

- Tenant 1: $th^1_{1,1} = 0.3$, $th^1_{1,2} = 0.6$, $th^1_{2,1} = 0.4$, $th^1_{2,2} = 0.7$, $th^1_{3,1} = 0.4$, $th^1_{3,2} = 0.8$.
- Tenant 2: $th^2_{1,1} = 0.3$, $th^2_{1,2} = 0.6$, $th^2_{2,1} = 0.4$, $th^2_{2,2} = 0.7$, $th^2_{3,1} = 0.4$, $th^2_{3,2} = 0.8$.
- Tenant 3: $th^3_{1,1} = 0.3$, $th^3_{1,2} = 0.6$, $th^3_{2,1} = 0.4$, $th^3_{2,2} = 0.7$, $th^3_{3,1} = 0.4$, $th^3_{3,2} = 0.8$, $th^3_{4,1} = 0.4$, $th^3_{4,2} = 0.5$.

For each tenant, the initial configuration is illustrated in Fig. 4.34, Fig. 4.35, and Fig. 4.36 respectively. Each mode is differentiated with different colors, while an horizontal dashed line represents the real-time traffic load of the slice. For Tenant 1, the slice mMTC lacks of radio resources, since *rt-load* belongs to the *Critical* mode. Similar behavior appears for the slice eMBB of Tenant 2, and the slices eMBB and uRLLC-2 of Tenant 3. Under this unstable resource distribution, the RAN orchestrator activates the *Slice Mode Selector* function to identify if exists at least a slice in *Support* mode able to share part of its radio resources. For Tenant 1, since the uRLLC slice is in *Support* mode (*rt-load* = 170 RBs), part of its RBs are shared with the *Critical* mMTC slice (*rt-load* = 160 RBs), until an equilibrium is reached, where the *rt-load*s off all the slices are in *Conservative* modes, as illustrated in the Fig. 4.37. Similar procedure is applied for Tenant 2, where the optimal resource parameterization is obtained by migrating a portion of the RBs from the mMTC (*rt-load* = 60 RBs) and the eMBB-2 (*rt-load* = 130 RBs) slices to the *Critical* eMBB slice (*rt-load* = 330 RBs), as shown in Fig. 4.38. To this end, Fig. 4.39 highlights the final optimal resource partitioning of Tenant 3, obtained from a starting configuration of 2 slices in *Critical* mode (the eMBB slice with *rt-load* = 190 RBs, and the uRLLC-2 slice with *rt-load* = 80 RBs). In particularly overloaded scenarios, radio resource sharing is insufficient to reach the network stability. In this case, the *Threshold Optimizer* is activated, and the thresholds of the *Conservative* slices are increased. The *rt-load* has higher probability to enter the *Conservative* mode, which preserves the resources assigned to the slice by disabling the sharing policy.
For all the initial slice tenant configurations, the scenario represents the case of a static RAN NS slicing approach, where the radio resources are fixed for the entire service time. Our dynamic

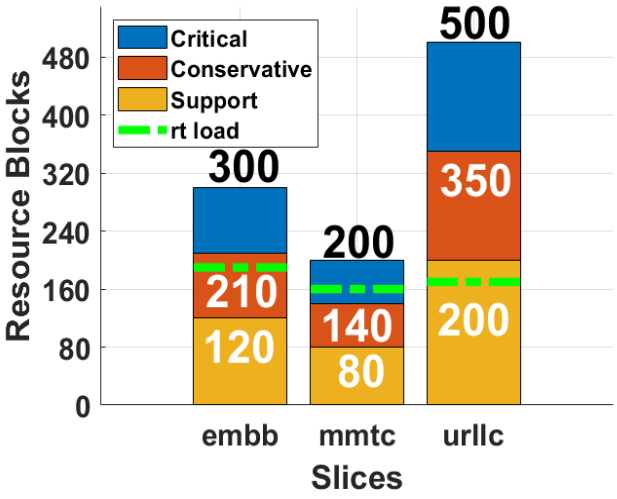


Figure 4.34: Initial Tenant 1 slice set

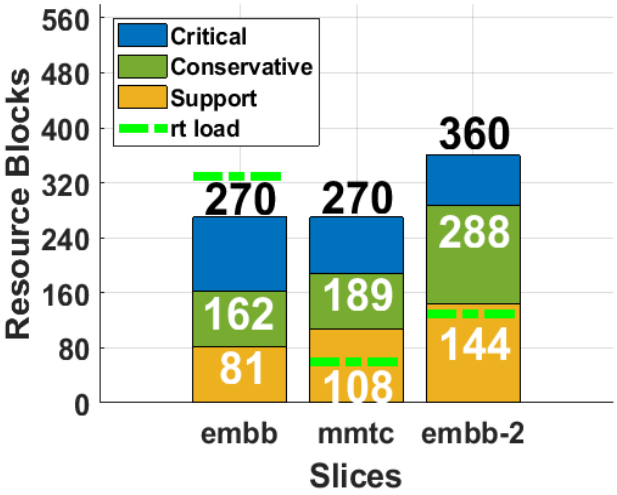


Figure 4.35: Initial Tenant 2 slice set

approach outperforms the static approach, increasing the available radio resources of the most congested RAN slices. This is highlighted in Fig. 4.38 and Fig. 4.39, where the capacity of the *Critical* slices is increased by 20.3% and 50% respectively, without downgrading the SLA of the others. Besides stabilizing the system, this operation increases the number of users served simultaneously.

In the final stage of our optimization method, the *intra-slice* scheduling algorithm described in Subsection 4.4.4.4 is performed to prioritize the users' service, proportionally to the maximum slice capacity. For each tenant $j \in (1, 2, 3)$, a priority factor $\varphi^j_{u,n}$ is assigned to each user u such that the resources of slice n are maximized, according to the system of Eq. (4.21). The weight distribution for each tenant is illustrated in Fig. 4.40, Fig. 4.41, and Fig. 4.42 respectively, where each color identifies a slice, and each bar is a user's weight. Compared to the classical fair weight distribution ($\alpha_q >> 1$), our proportional allocation overcomes under-over resource provisioning up to 28.5% and

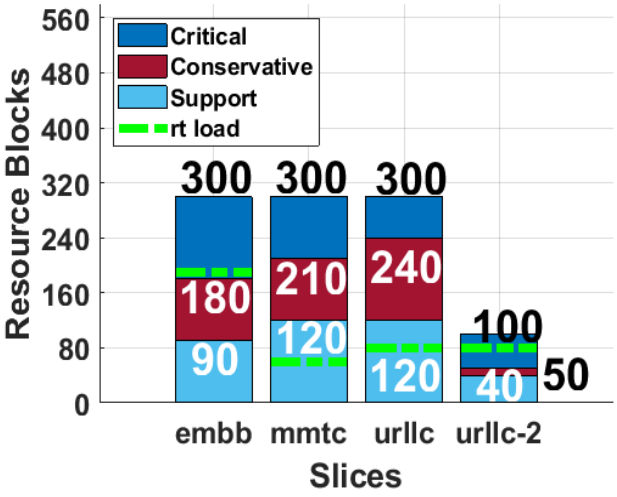


Figure 4.36: Initial Tenant 3 slice set

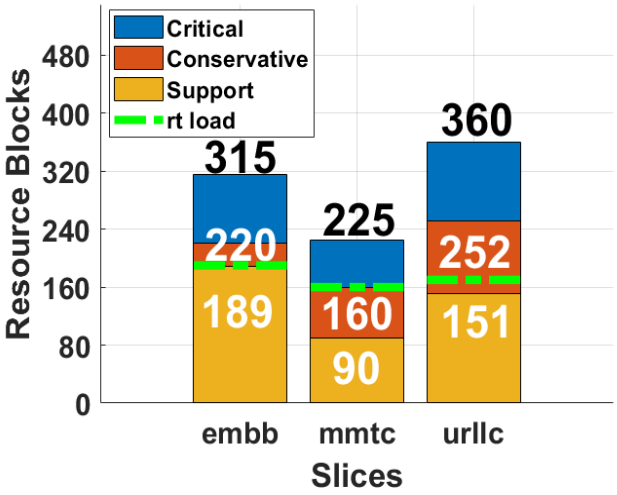


Figure 4.37: Final Tenant 1 slice set

-120% (blue and green rectangles) respectively, introducing remarkable benefits in resource saving policies. Indeed, for Tenant 1, the correct provisioning in the eMBB slice brings to a 94.6% saving in radio resources, which could be used to serve more users. The role of α_q is further illustrated in Fig. 4.43, where for each slice, the Jain index determines resource sharing fairness among the users. As expected, for our case ($\alpha_q = 1$) the Jain index per slice is less than 1, emphasizing the proportional resources' allocation of our approach, while proportional distribution is reached for $\alpha_q >> 1$. By showing the Jain index, we aim at highlighting the correct calibration of our system, and the effectiveness of our game theoretic approach.

4.4.5.4 Graphical Abstraction of Nash Equilibrium Strategy Profile

Fig. 4.44 illustrates the existence of a unique equilibrium point (global maximum) for the utility function f_q (as shown in Eq. (4.21)), under an increasing number of users' served per slice. In total,

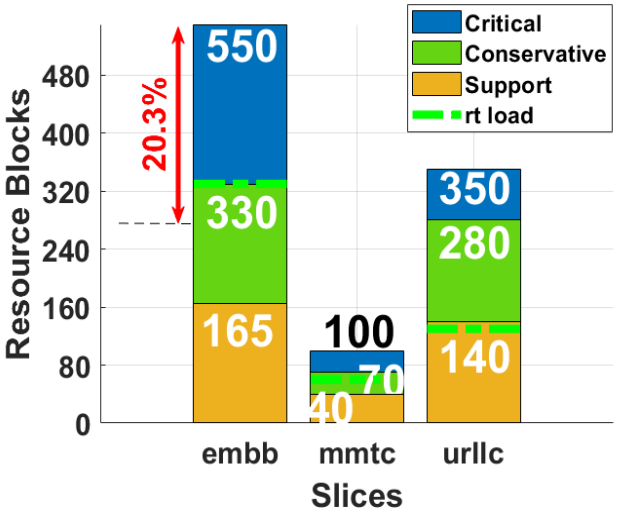


Figure 4.38: Final Tenant 2 slice set

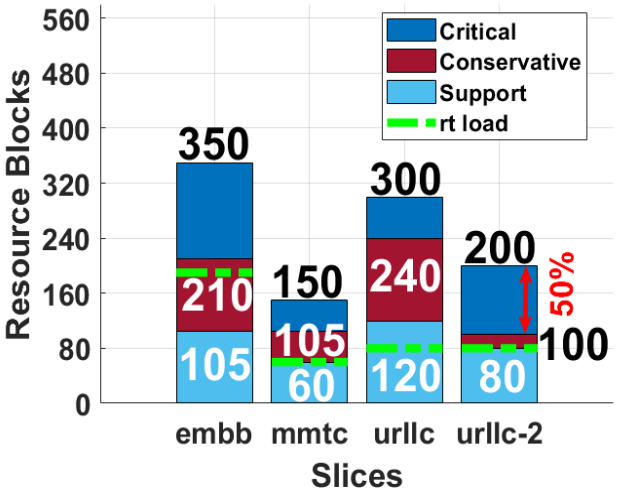


Figure 4.39: Final Tenant 3 slice set

6 experiments are shown, starting from 3 users served per slice (Exp. 1, $f_q \approx 3.5$), until 20 users served per slice (Exp. 6, $f_q \approx 2.3$), where with the same color is indicated the slices of the same tenant. For each experiment, the outcome is obtained by repeating the simulated scenario multiple times ($\approx$ 1000) under different initial random weights, in order to increase the accuracy of our solution. When a small amount of users are simultaneously served by a slice, the optimal weights' distribution is reached in few iterations, as illustrated in Exp. 1. The proportional fair weights distribution becomes complex under the increasing number of users' served per slice. Indeed, the intra-slice scheduling algorithm needs to select and apply a constrained nonlinear optimization algorithm (i.e., Matlab *fmincol*, *fminsearch*, *fminunc*) able to solve the system of Eq. (4.21). This is the case of Exp. 6, where the convergence of the utility function is obtained after 37-39 iterations. For each set of slices under different loads, the proposed algorithm reaches the equilibrium point of f_q under a finite set of iterations. This solution presents an intuitive and alternative methodology

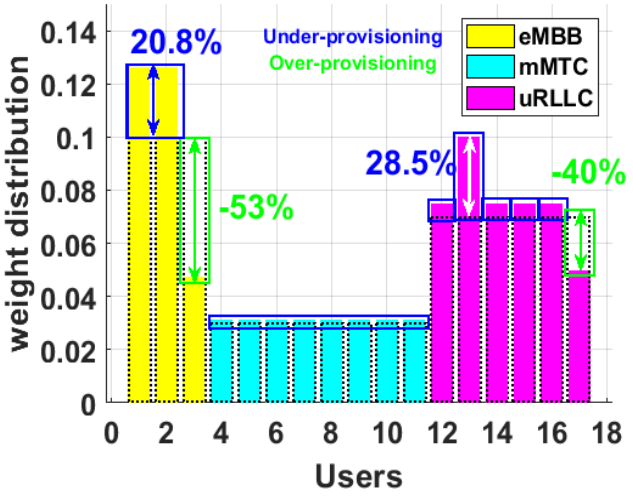


Figure 4.40: Users' weight Tenant 1

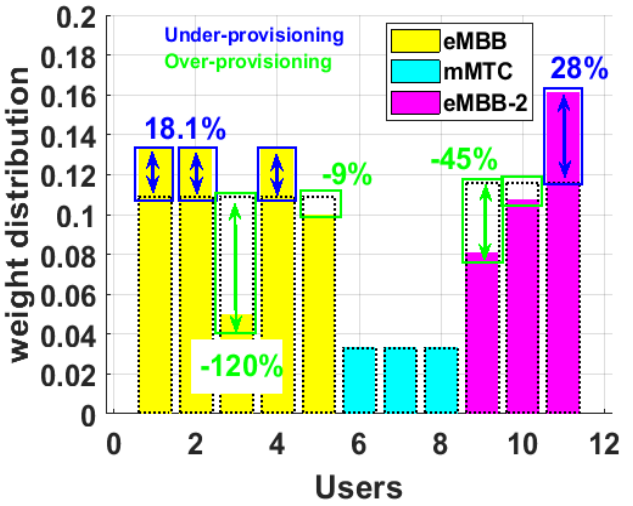


Figure 4.41: Users' weight Tenant 2

to proof the existence of a Nash Equilibrium in our model, where the desired outcome is achieved by not deviating from the initial problem constraints.

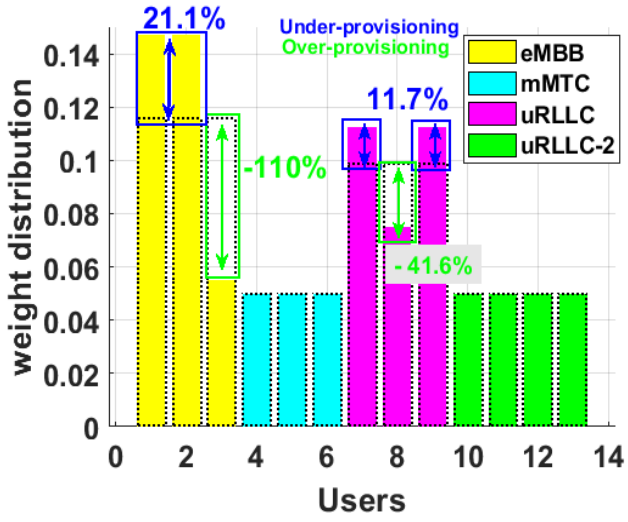


Figure 4.42: Users' weight Tenant 3

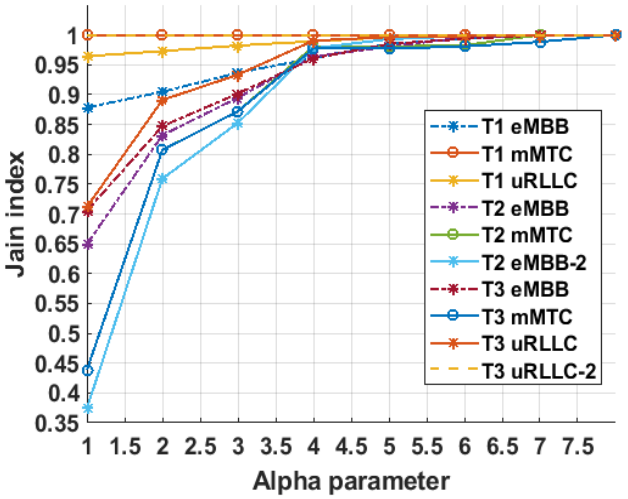


Figure 4.43: Jain index under increasing alpha-q

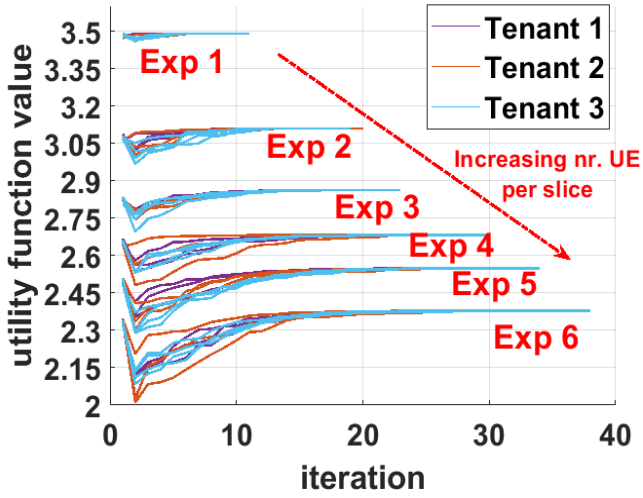


Figure 4.44: NESP trend per tenant under increasing alpha-q

Chapter 5

O-RAN Fronthaul Interface Prototype for NI USRP N310 SDR

5.1 Theoretical Background

5.1.1 O-RAN Alliance: High Level Architecture Overview

The O-RAN ALLIANCE is committed to evolving RAN, making them more open, smarter, interoperable, and scalable than contemporary deployments. It consists in the standardization of RAN elements including a unified interconnection standard for white-box HW and open source SW elements from different vendors. The O-RAN vision focuses on two main aspects:

- *Openness*: with the introduction of open interfaces, O-RAN allows small size vendors, MNO, and operators to rapidly deploy their own services, or integrate their solutions on top of other networks.This approach transforms the current ecosystem with the introduction of novel capabilities such as multi-vendor deployments, open source SW and HW, and easier innovation.
- *Intelligence*: with the exponential traffic growth and new services demand, new automated solution must be deployed to manage the 5G network. With the O-RAN architecture, emerging deep learning are integrated in every RAN layer, enabling dynamic RRM techniques and optimized network efficiency. In combination with O-RAN's open interfaces, AI-optimized closed-loop automation is achievable and will enable a new era for network operations.

Fig. 5.1 illustrates the reference architecture designed by O-RAN. The most important functional components introduced by O-RAN are the non-Real-Time (non-RT) Radio Intelligent Controller (RIC), placed in the Service Manager and Orchestrator (SMO), and the near-RT RIC, installed

This work has been co-supervised with National Instruments GmbH (Dresden, Germany), as part of a secondment plan during the project implementation.

inside the 3GPP compliant O-CU and/or O-DU. At the bottom of the figure, the O-Cloud represents an O-RAN compliant cloud platform that uses HW acceleration addons when needed (e.g., to speed up fast Fourier transform operations or forward error correction tasks) and a SW stack that is decoupled from the HW to deploy eNBs/gNBs as VNFs in vRAN scenarios (130). Each component of the aforementioned architecture is individually studied inside the O-RAN project and, as it will be illustrated in the next sections, we will deeply analyse the O-RU node since it hosts Low-PHY layer and RF processing based on a lower layer functional split.

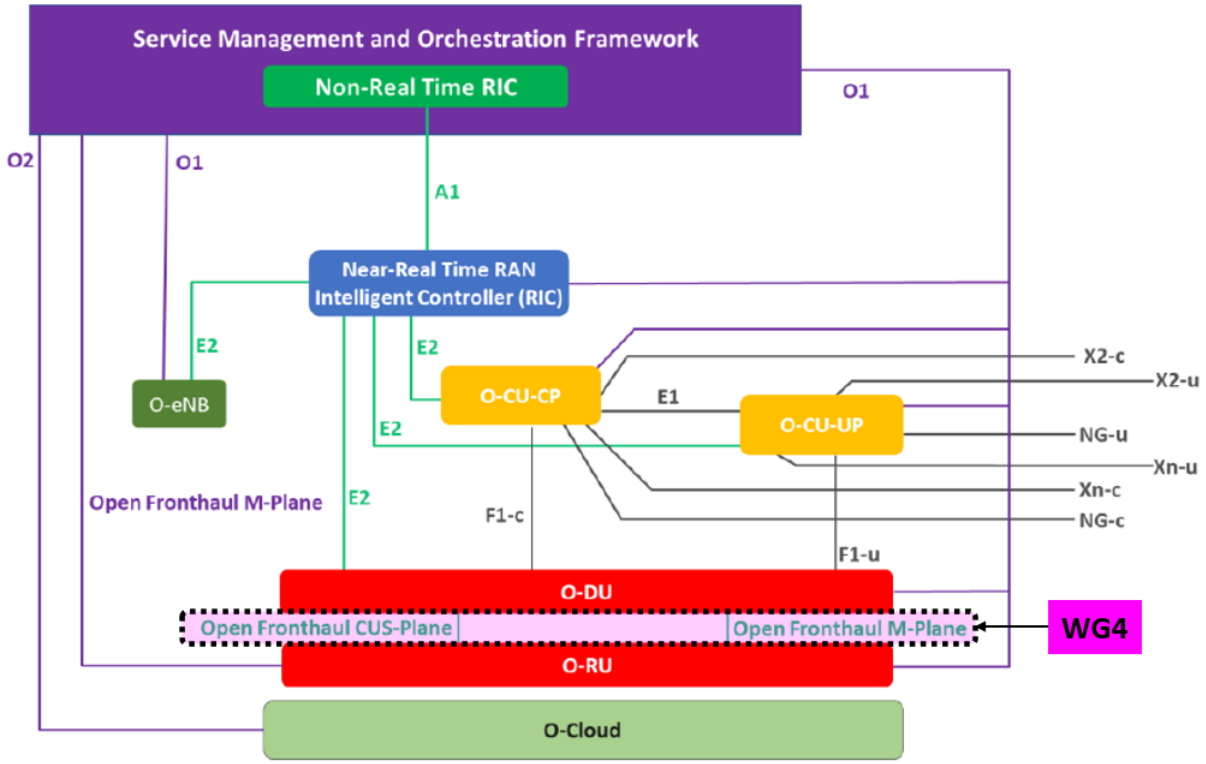


Figure 5.1: O-RAN Alliance Reference Architecture

5.1.2 NI USRP N310 SDR: Design and Core Features

The USRP N310 is a networked SW defined radio that provides reliability and fault-tolerance for deployment in large-scale and distributed wireless systems[2]. It is composed by four RX and four TX channels in a half-wide RU form factor, where the RF front end uses two AD9371[3] transceivers, and the latest Radio-Frequency Integrated Circuit (RFIC) technology from Analog Devices (ADs). Each channel provides up to 100 MHz of instantaneous bandwidth and covers an extended frequency range from 10 MHz to 6 GHz. The open-source USRP Hardware Driver (UHD) API and RF Network-on-Chip (RFNoC) Field-Programmable Gate Array (FPGA) development framework reduce SW development effort and integrate with a variety of industry-standard tools such as GNU

[2]https://www.ettus.com/all-products/usrp-n310/
[3]https://www.analog.com/en/products/ad9371.html

Radio. The UHD is a user-space library that runs on a General Purpose Processor (GPP) and communicates with and controls all of the USRP device family. This concept is illustrated in Fig. 5.2[4], where UHD is written in C/C++ while the code developed for the FPGA is written in Verilog. There is a C/C++ API that can interface to other SW frameworks, as in the case of GNU Radio[5], or a user can simply build custom signal processing applications directly on top of the UHD C/C++ API. For the baseband processing, a Xilinx Zynq-7100[6] System-on-a-Chip (SoC) allows

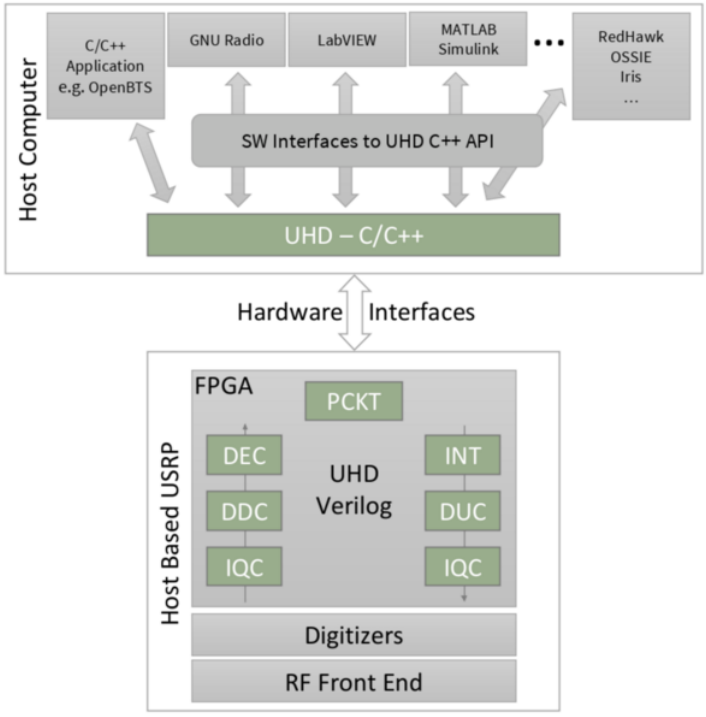


Figure 5.2: UHD components

for real-time and low-latency processing on FPGA, and a dual-core ARM CPU for stand-alone operation. Morever, using the preinstalled Linux OS, custom applications can be embedded, or stream samples to a host computer using the two high-speed 1 or 10 Gigabit Ethernet SFP+ ports. In addition, the USRP N310 has a flexible synchronization architecture with support for traditional SDR synchronization methods such as clock reference, PPS time reference, and GPSDO, and the open-source White Rabbit timing protocol[7].

Fig. 5.3 illustrates the block diagram of the USRP N310 motherboard[8].

[4]https://kb.ettus.com/UHD
[5]https://www.gnuradio.org/
[6]https://www.xilinx.com/products/silicon-devices/soc/zynq-7000.html
[7]https://white-rabbit.web.cern.ch/
[8]https://www.ettus.com/all-products/usrp-n310/

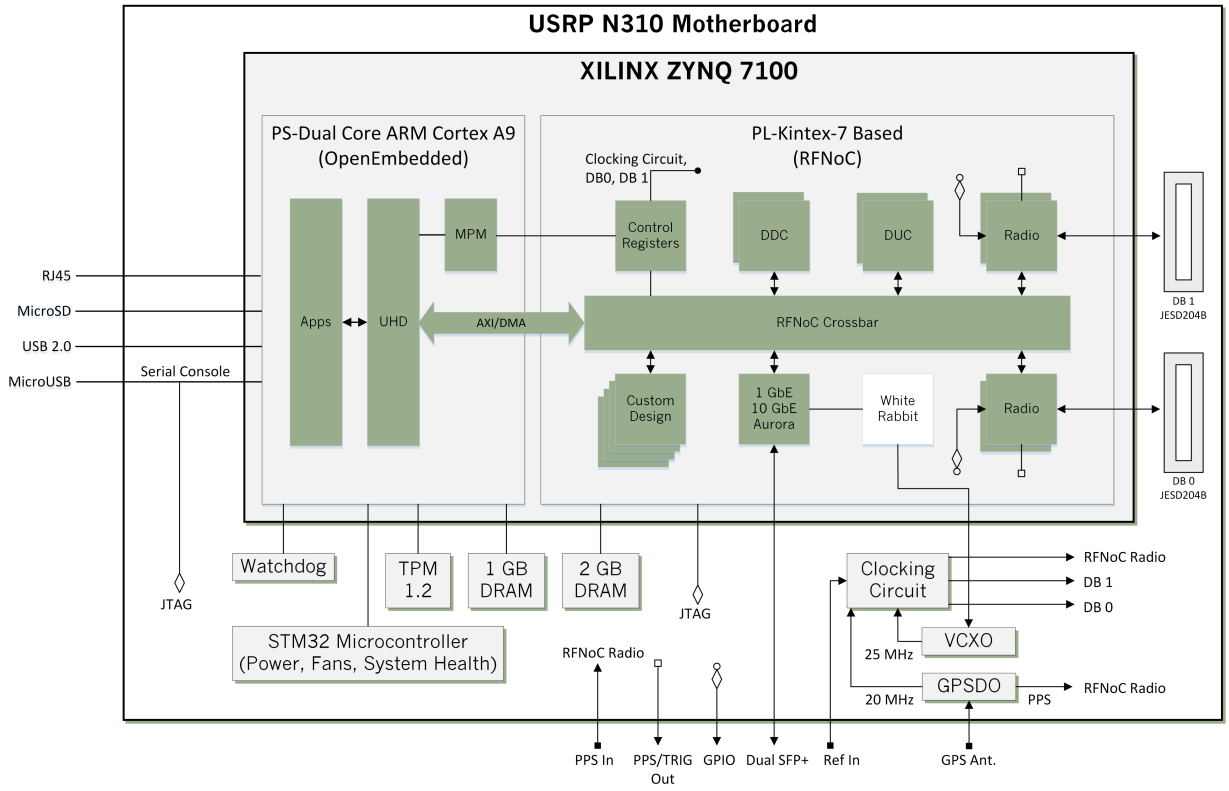


Figure 5.3: USRP N310 motherboard block diagram

5.2 O-RAN Open Fronthaul SDR Integration: Proposed Contribution

O-RAN's open FH is a logical interface consisting of Lower-Layer Split (LLS) Control Plane (LLS-CP), LLS User Plane (LLS-UP), Synchronization Plane (SP), and Management Plane (MP), in addition to specifying a new Cooperative Transport Interface (CTI). The WG4 is responsible to deliver truly open FH interfaces, in which multi-vendor DU-RRU interoperability can be realized. The O-RAN split architecture utilizes this novel interface to interconnect the O-DU with the O-RU in a vendor-agnostic manner, focusing on intra-PHY split, which reduces FH bandwidth requirements compared to conventional split RAN architectures. In this section, we propose an initial draft of O-RAN open FH interface design for NI USRP N310: subsection 5.2.1 displays the role of our SDR device inside the O-RU unit, including the subset of the eNB/gNB functions as defined in split option 7-2x, and subsection 5.2.2 provides and high level overview of the O-RAN transport protocol stack integration with UHD transport drivers. Finally, two O-RAN USRP N310 architecture proposals are illustrated in subsection 5.2.3, which are also compliant with the latest NI SDR product, the USRP X410[9].

[9]https://www.ettus.com/all-products/usrp-x410/. At the time of writing, the USRP X410 was still a prototype, limiting our design know-how for this device. Nevertheless, since its architecture is an enhanced version of the USRP N310, we believe the proposed solutions are straightforward compliant with USRP X410.

Even though our illustration is a pure theoretical concept, the proposed solution represents a first step towards the interoperability of NI SDR devices with third-party O-DU vendors/platforms using O-RAN architecture.

5.2.1 7.2x Function Split and NI SDR

Fig. 5.4[10] illustrates the "7.2x" split defined in O-RAN between the O-DU and O-RU, where the dotted block are not mandatory according to different implementation categories (131). Compared to other splits, it is the best balance between bringing 5G to market quickly and deployment cost, reducing confusion about split specifics while making traffic reduction gains and improvements. When the precoding phase is not performed (therefore of lower complexity) in the RUs, the O-RUs belong to "Category A", while O-RUs within which the precoding is done are called "Category B" O-RUs. Two main benefits come from this approach: i) simplicity, since the size, weight, and power draw represent a limitation of the current RU, and ii) interface complexity, since higher level interfaces tend to reduce the interface throughput relative to a lower-level interface.

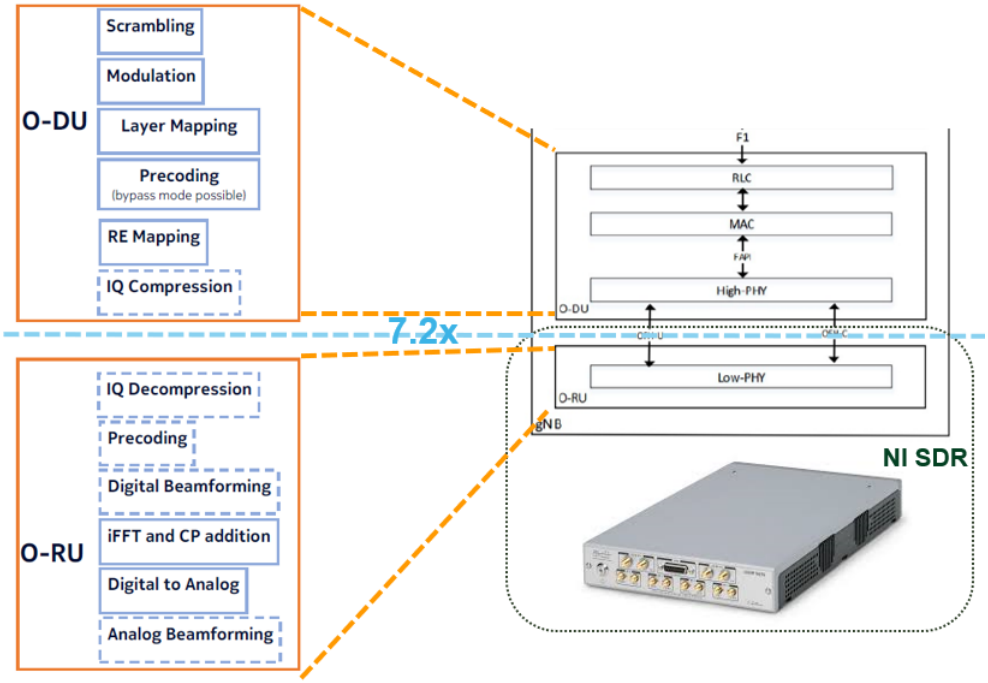


Figure 5.4: DL O-RAN Split point between O-DU and O-RU

In the DL, OFDM phase compensation, iFFT, CP addition, and digital beamforming functions reside in the O-RU as well as precoding for Category B O-RUs. The rest of the PHY functions including resource element mapping, precoding, layer mapping, modulation, scrambling, rate matching and coding reside in the O-DU. In the UL, OFDM phase compensation, FFT, CP removal

[10]O-RAN ALLIANCE, "O-RAN Fronthaul Cooperative Transport Interface Transport Control Plane Specification 2.0", O-RAN Working Group 4, O-RAN.WG4.CTI-TCP.0-v02.00. Available: https://www.o-ran.org/specification-access

and digital beamforming functions reside in the O-RU. The rest of the PHY functions including resource element de-mapping, equalization, de-modulation, de-scrambling, rate de-matching and de-coding reside in the O-DU (dotted boxes in Fig. 5.4 are optional).

5.2.2 Fronthaul O-RAN Transport Protocol Integration over USRP N310

The protocol stack of the transport planes in O-RAN FH specifications are shown in Fig. 5.5. In the CP and UP, the FH specifications support eCPRI or Radio over Ethernet (RoE), while the SP FH uses Precision Time Protocol (PTP) and SyncE over Ethernet (132).

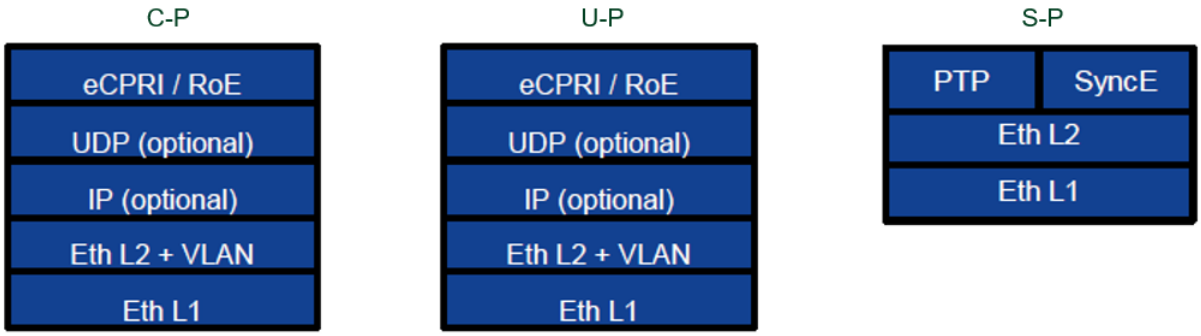
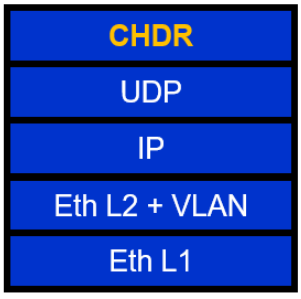


Figure 5.5: Protocol stack of each transport plane

The integration of the eCPRI interface represents the first challenge in our solution, since the USRP with UHD\RFNoC utilizes Condensed Hierarchical Datagram for RFNoC (CHDR) packet format which is used for data streaming to and from the USRP and is a mandatory requirement for packet processing within RFNoC. CHDR is based on UDP but does not support eCPRI header, and it is structurally different from the O-RAN protocol stack, as illustrated in Fig.5.6. Adding a CHDR header on top of the eCPRI header of the O-RAN packet does not represent an O-RAN compliant solution, since third-party O-DUs might not support CHDR packet format. Similarly, replacing the CHDR field of UHD with the eCPRI is also not correct, since CHDR field is used inside the USRP for source routing and deterministic routing. To overcome this issue, in the next subsections two distinct solutions are designed and illustrated, with different level of complexity.



#	Memory Layout <------ CHDR_W = 64 bits ------>								Required?
0	VC (6)	EOB (1)	EOV (1)	PktType (3)	NumMData (5) *Value=M*	SeqNum (16)	Length (16) *Value=L*	DstEPID (16)	Y
1	Timestamp (64)								N
2	Metadata[0] (CHDR_W)								N
.	...								.
M+1	Metadata[M-1] (CHDR_W)								N
M+2	Payload[0] (CHDR_W)								Y
.	...								.
M+N+1	Payload[N-1] (CHDR_W)								N

Figure 5.6: CHDR packet protocol stack

5.2.3 USRP N310 Transport Adapter Upgrade

In the current architecture, the Transport Adapter (TA) of a USRP N310 is external to the RFNoC, and customized to only accepts CHDR-based packet format, allowing only UHD-based SW platforms to interact with it. In order to integrate O-RAN interface, Fig.5.7 presents a solution where non-CHDR packets are straightforward redirected from the TA to the ARM processor (equipped with Linux RT OS), where a new SW function adds a specific CHDR header to every eCPRI\O-RAN packet. This operation classifies O-RAN traffic inside the device, and allows the correct packet routing and processing within the RFNoC/FPGA, compliant with the UHD drivers. Even though this solution might be relatively complex to implement since it requires a modification of the current TA, it represents an initial step to the integration of third-party platform within NI SDR, paving the way towards the interoperability of new transport protocols, interfaces, and 5G HW independent principle.

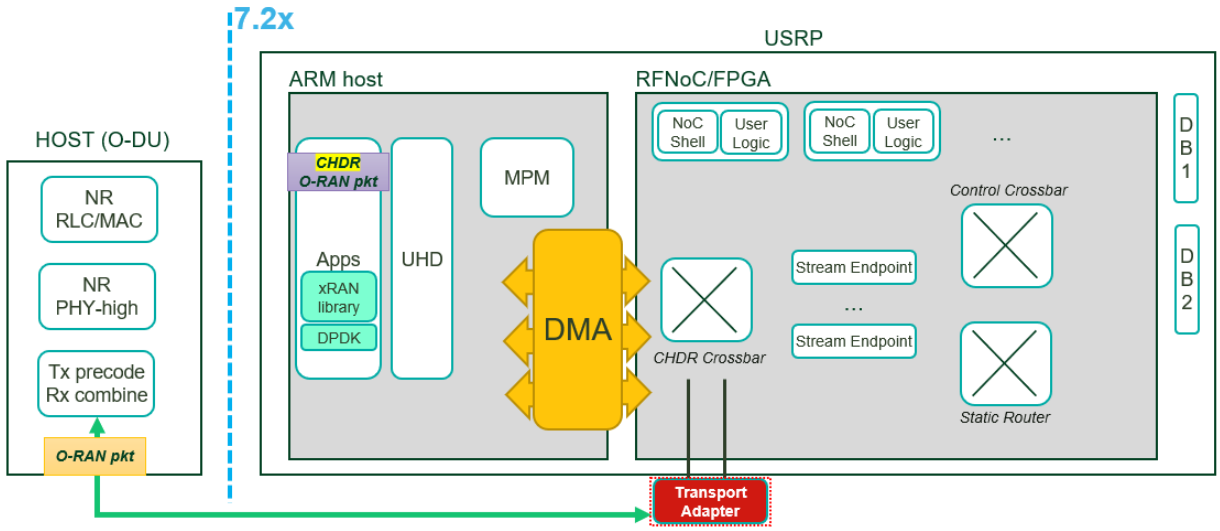


Figure 5.7: External TA design with CHDR header aggregation

Once the CHDR-O-RAN header is attached, the packet is redirected to an O-RAN module for the data extrapolation. In our solution, we provide two different methods, as illustrated in the following subsections.

5.2.3.1 ARM-based O-RAN Processing

As part of O-RAN open FH project, xRAN are C\C++ libraries to perform UP and CP functionality according to the ORAN FH specification[11]. They are responsible most of the packet processing such as Transport header, Application header, Data section header and interactions with the rest of the PHY processing pipeline. Moreover, xRAN are built on top of DPDK to perform UP and CP functionality according to the ORAN FH specification.

[11] https://docs.o-ran-sc.org/projects/o-ran-sc-o-du-phy/en/latest/overview1.html

Assuming that the xRAN library and the DPDK are installed inside the *Apps* section (see Fig.5.7), the O-RAN packet with CHDR header is now processed following O-RAN specs, and then forwarded to the RFNoC for the final DSP processing before reaching the antenna elements. This solution allows the coexistence of CHDR and non-CHDR traffic on top of the same SDR. However, the SW implementation of O-RAN using xRAN libraries requires a deep analysis of the processing requirements, since the ARM processor of an USRP N310 does not guarantee extremely high performance. Forwarding all the O-RAN traffic to the ARM first and then to the RFNoC double the traffic handled by the Direct Memory Access (DMA). This could bring latency and synchronization issues on our device, which should require a careful implementation of the interactions among the different components. We conclude that this solution represents a potential alternative for O-RAN traffic use cases where a trade-off between performance and datarate can be achieved.

5.2.3.2 Custom Eth-CHDR and Third-party O-RAN IP NoC Blocks

As mentioned in 5.2.3.1, the ARM processing capabilities might affect the performance of our O-RAN integration. As valid alternative, Fig. 5.8 illustrates a fully O-RAN FPGA implementation of a USRP N310. This figure is the same as Fig.5.7, with more detailed analysis of the RFNoC/FPGA section. In our solution, we decided to dedicate one SFP+ port exclusively for non-CHDR traffic (in our case, O-RAN traffic), while the other operates following the default CHDR compliant workflow. At the non-CHDR port, the Ethernet processing is shifted to a NoC block, transforming the TA as a transparent bridge between input port and the DSP unit inside the RFNoC. This approach

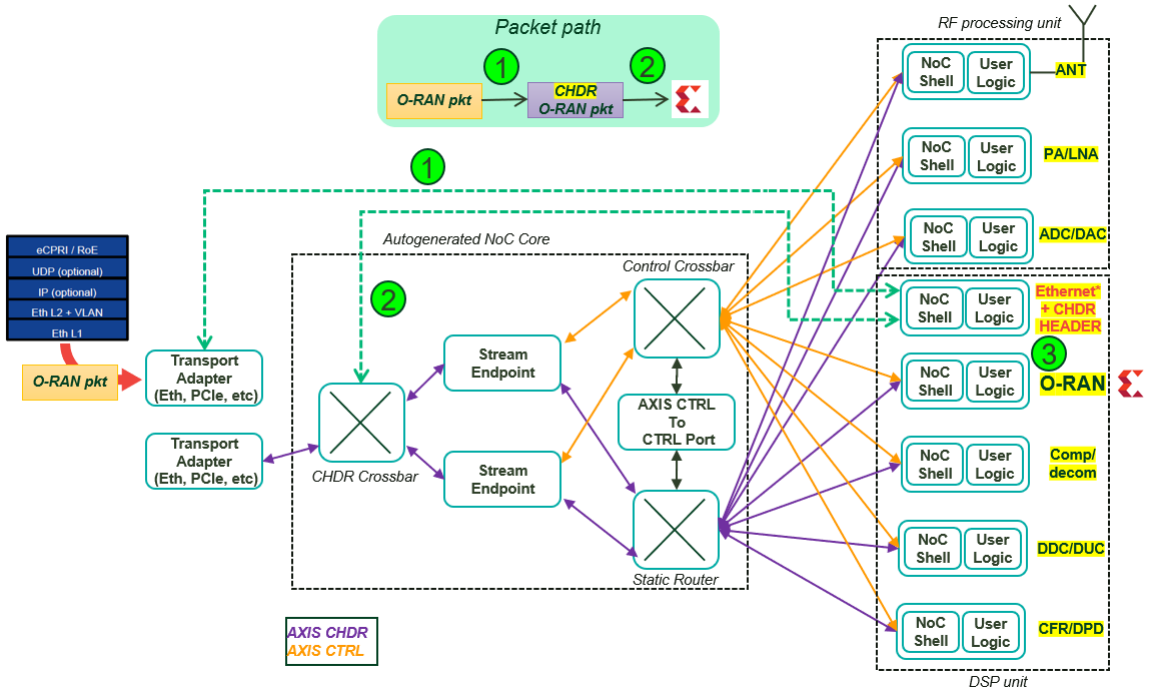


Figure 5.8: Full O-RAN implementation over FPGA

presents several advantages: first, the TA does not introduce extra delay due to transport header packet processing, and second, the modification of the Ethernet interface such that a CHDR header is added to the eCPRI\O-RAN packets is simpler to implement as a NoC Block, since it is NI USRP proprietary design, and it does not require modification of third-party HW.

Applied the CHDR-O-RAN header (point 1 of Fig. 5.8), the packet is redirected to the CHDR *Crossbar* (point 2 of Fig. 5.8), which is responsible for the routing among the different NoC blocks. An implementation from scratch of a O-RAN FPGA module might require massive manpower and time; for this reason, we opted for already existing O-RAN SoC module from Xilinx[12], as illustrated in Fig. 5.9. This module is integrated as a NoC block inside our USRP architecture (point 3 of Fig. 5.8), and it is responsible for the O-RAN packet processing, similar to the xRAN libraries of our previous solution.

At this point, the packet can follow the ordinary DSP chain (IQ decompression, precoding, digital beamforming, iFFT and CP addtion, etc.), and be forwarded through the antenna element inside the RF processing unit.

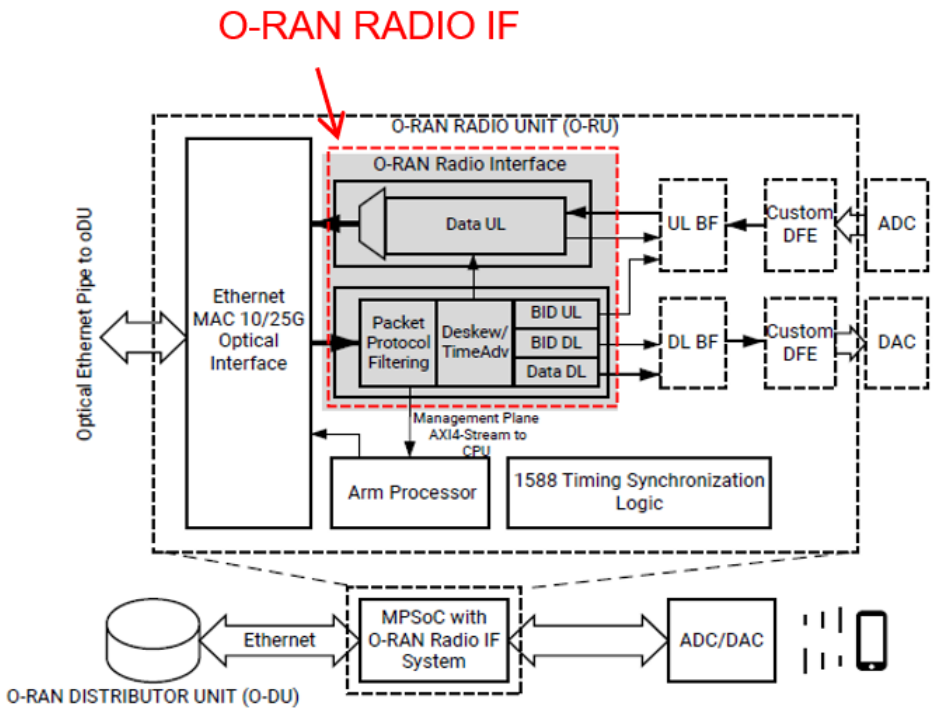


Figure 5.9: Xilinx O-RAN radio interface subsystem

Through the RFNoC, multiple architecture flavours could be designed: in our design, all the NoC blocks are interconnected with the NoC Core such that multiple routing options and NoC chain structures are feasible. Moreover, CHDR and non-CHDR traffic can coexist on the same platform, increasing the flexibility and scalability of the device. For extreme low latency applications, packet routing between NoC Core and DSP unit might increase the system latency and processing time.

[12]https://www.xilinx.com/products/intellectual-property/ef-di-oran-radio-if.html

For this reason, another design could include the direct connection of the NoC block inside the DSP and RF units, without routing back to the NoC Core. Anchoring in a chain the NoC block simplifies the HW and SW implementation on our device, but it limits the use cases of the USRP to specific protocols, interfaces, and applications.

5.2.4 S-Plane: O-RAN Synchronization over USRP N310

Synchronization Plane (S-Plane) has been identified as a key aspect of O-RAN specifications, with requirements defined in the working groups for Fronthaul, Open Interfaces and Transport. O-RAN standard involves time and frequency synchronization distributed to the O-DU and O-RU in different manners (131):

- Frequency synchronization where clocks are aligned in frequency,
- Phase synchronization where clocks are aligned in phase,
- Time synchronization where clocks are aligned to a common base time.

Together the above parameters define a *profile* for the network, requiring a set of features and options selections for bridges and end stations operation. Further, the profile also states the conformance requirements for supporting equipment and user applications.
Different O-RAN synchronization topologies have been designed to address different deployment market needs. In our first draft, configuration LLS-C1 has been selected, as illustrated in Fig. 5.10, where the O-DU is part of the synchronization chain towards the O-RU, and network timing is distributed from O-DU to O-RU via direct connection between O-DU site and O-RU site.

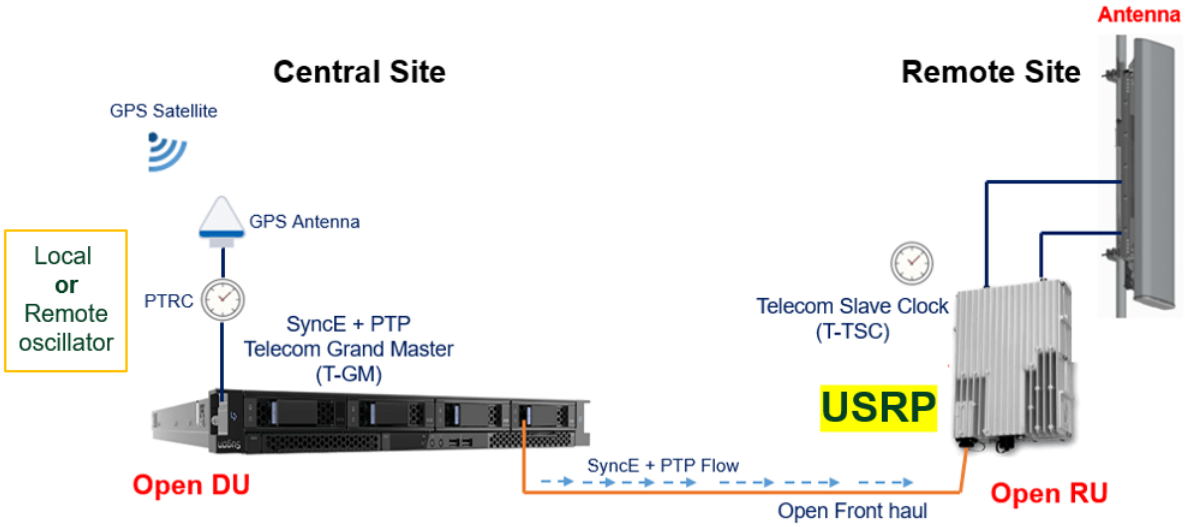


Figure 5.10: LLS-C1 synchronization design for USRP N310

Configuration LLS-C1 is based on point-to-point connection between O-DU and O-RU using network timing option. As shown in figure, it is basically the simplest topology for network timing

option, where O-DU directly synchronizes O-RU. At the current status, the USRP N310 is predisposed for SyncE and PTP support from HW perspective, while Sync SW part needs to be developed. Regarding White rabbit protocol, the implementation ready for USRP devices, but it is not suitable for O-RAN, since requires third-party infrastructure equipped with White rabbit SW and HW support.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

This dissertation has dealt with the design an testing of novel 5G NR resource orchestration solutions, given the modern challenges of heterogeneity, interoperability, scalability, and flexibility of the future 5G and beyond infrastructures and architectures.

As wireless domain grows and becomes increasingly complex, a higher degree of automation and less manual intervention is needed. For this reason, we focused on the definition and implementation of innovative real-time resource management schemes, platform-independent, which are able to produce an efficient RAN parameterization based on the dynamic analysis of the E2E 5G infrastructure, traffic's evolution, and SP's SLA. The new features and requirements introduced with 5G standard have been precisely integrated in our mathematical models, using different mathematical models based on constrained non-linear programming (i.e., Interior-point method, trust region reflective algorithm), stochastic dynamic programming (i.e., Markov decision processes), and strategic stochastic methods (i.e., Game theory). In our attempt of testing the proposed solutions on top of realistic scenarios, two testbeds have been assembled, based on open-source software and proprietary hardware components. Moreover, inline with the transforming the RAN industry towards open, intelligent, virtualized and fully interoperable RAN, a design of a O-RAN compliant interface for SDR has been proposed.

The major contributions of the thesis have been divided into two main parts: i) the first part is described in Chapter 4, and ii) the second part includes Chapters 5. A detailed summary of each chapter is presented in the following:

- Chapter 4: In this chapter, we presented our proposed RAN orchestration frameworks based on different optimization algorithms, under increasing modelling complexity, for the optimal management of radio resources, dynamic FH parameterization, and ad-hoc placement of the NFs according to the service's constraints. Combining various management techniques based

on the concept of NS, the scalability and flexibility advantages for the 5G architecture enable multiple scenarios characterized by diversified services.

In Subsection 4.1, our first dynamic NS management solution have been presented. By taking advantage of our early testbed deployment, we evaluated the proposed dynamic resource allocation algorithm on top of a real 5G scenario. We illustrated how a simple dynamic NS resources management scheme is able to outperform the current state-of-the-art, which is primarily based on a static NS principle. Moreover, an enhancement of QoS performance per user is revealed, increasing the service capability.

Balancing the RAN resources based on greedy algorithm might imply a ping-pong resource sharing effect among the slices, causing instability and congestion in the centralized unit. To overcome this issue, in Subsection 4.2, a novel framework for real-time RAN NS resources sharing has been proposed, where a specific label is assigned to each slice according to its service traffic load and slice tenant's SLA. Three different modes are defined, each one characterizing a different resource sharing policy. Through each slice mode, the RAN controller (called Slice Manager) is aware of the traffic load and resource saturation threshold per slice, allowing the activation of resource sharing policies among slices in order to keep the stability of the system. Since the proposed solution is technology independent, its implementation is suitable for 5G Option 3 deployment, which reflected the initial launch strategy being adopted by multiple operators.

Forecasting the resource utilization of a slice in a real-time RAN system is crucial to maintain the stability of a 5G network. For this reason, in Subsection 4.3, using as baseline the previous slice-mode model, a novel slice resource sharing algorithm is proposed, where a *blocking probability* is defined to predict the resource utilization per slice in the system. With this technique, the system is modelled as a Markov process, and the impact of the slice traffic flow is forecast such that resource sharing capabilities can be performed on time, avoiding traffic bottleneck situations. The results show that the proposed method increases the average receiver data rate per slice by 20% compared to a classical static slicing approach. Moreover, under unexpected traffic peaks variations, the average BER of the dynamic solution is 52% less than the static case, while preserving the QoS of the served users.

Lastly, in Subsection 4.4, a complete FH NS orchestrator is proposed, exploiting the latest paradigms of the 5G standard: FS, 5G numerology, flexible timing parameters, and NS. This novel RAN orchestration framework is able to compute the optimal RAN network functionalities placement, flexible HARQ design, and dynamic radio resource management. The obtained results show that specific FS option and flexible HARQ timing parameterization

per service improves the performance in terms of latency, infrastructure design, and system scalability. Simultaneously, a real-time spectrum sharing technique based on a joint UE-slice evaluation of the data rate requirements overcomes saturation and overprovisioning of radio resources. Compared to our previous solutions, to further improve the serving time and resource assignation for each UE, at intra-slice layer, a novel scheduling scheme based on game theory is designed, capable of optimize the UEs' service priority according to the single UE traffic characterization.

- Chapter 5: among the multiple features introduced in 5G, the HW\SW disaggregation has a key role in the 5G NR evolution, where the HW components are replaced with flexible and reusable software-defined functions. As drivers for this new trend, two solutions are involved: i) Virtualized RAN assists vendors in the abstraction \virtualization of their RAN components such as gNB and baseband units, and ii) the SDR migrates the entire radio function runs on a General-Propose Processor (GPP), interacting with the physical resources either directly or through a full or partial hardware emulation layer.

  In this chapter, the design of a O-RAN compliant interface on top of NI USRP SDR has been proposed. Since NI SDRs use UHD protocol for the communication among the protocol stack functions, the interoperability\integration with third-party O-RAN solutions is not straightforward. For this reason, in this work, two strategies have been proposed, where the architectures of the latest NI USRP N310 and X410 SDR models have been re-thinked to meet the O-RAN requirements. The first fully-SW solution is a lightweight implementation of the O-RAN protocol, which embeds the O-RAN SW libraries directly in the radio, exploiting the ARM-based processor functionalities of the NI SDRs. As advantage of this approach, from an HW perspective, only the NI SDR ethernet transport adapter have been modified, such that non-UHD O-RAN traffic is directly redirected to the ARM Host. On the other side, the limited processing capabilities of the ARM host might introduce different limitations: i) the traffic on the DMA is doubled due to the CHDR header encapsulation and following routing of the O-RAN packets inside the RFNoC/FPGA, and ii) no DPDK has been currently designed for USRP, limiting the throughput and real-time capabilities of the radio device.

  To overcome the aforementioned limitations, the second proposed strategy offers a fully HW integration of the O-RAN protocol for NI SDRs. Since the O-RAN protocol is represented by a third-party DSP component embedded in the RFNoC, this solution takes advantage of the FPGA capabilities to route the O-RAN traffic until the RF unit. Moreover, the scalable RFNoC design of the NI SDRs facilitate the instantiation of multiple configurations, enabling different protocol coexistence policies.

The two proposed solutions aim to illustrate the potential of NI SDRs in terms of customization, scalability, and flexibility towards shared RAN protocols. Different use-cases and architectures can be configured, tested, and easily integrated thanks to the open-source NI SDR capabilities.

6.2 Future Works

While the 5G deployment phase recently started, the evolving smart infrastructure, efficient technologies, and diversified wireless applications make the launch of Sixth Generation (6G) networks inevitable. The research topics and contributions of this dissertation will continue to be under discussion also in 6G, paving the way to new challenges and investigation:

- Energy harvesting technologies and the use of new materials will greatly improve the RAN energy efficiency and realize sustainable green networks. NS will continue to evolve as a key enabler in 6G Massive RANs, enabling advanced RAN flexibility, massive inter-connectivity and energy-efficient communications. By introducing AI-based optimization techniques in the proposed SM, the sharing of radio resources among RAN slices might present higher performance, advanced resource forecasting, and implicit power saving capabilities.

- Mature 5G\6G testbeds will be available in the next years, integrating new waveforms like filter bank multi-carrier, universal filtered multicarrier, or generalized frequency division multiplexing. Non-orthogonality constraint will lead to a more efficient and flexible modulation schemes, such as Non-Orthogonal Multiple Access (NOMA) or Rate-Splitting Multiple Access (RSMA). Testing the proposed optimization algorithms under these complex modulation schemes will provide higher throughput, spectral efficiency, new resource granularities, and advanced self-interference while minimizing the amount of active resources.

- Wireless transmission at THz frequencies represents a particularly attractive and flexible solution in 6G. The THz spectrum will certainly open up more bandwidth, providing an additional degree of scalability in assigning the most suitable frequency resources for given scenarios. Future C-RAN THz solutions will enable ultra-dense deployment of small cells in the FH, mitigating the high path loss of THz signals and guarantee connectivity via Line-of-Sight (LoS) links. Under this novel frequency region, the proposed FH orchestration framework might be suitable for the future 6G V2X scenarios, which require extreme low latency and accurate management of the radio infrastructure.

- Finally, Integrated Sensing and Communication (ISAC) has been recently proposed for numerous emerging applications, including but are not limited to, vehicular network, environ-

mental monitoring, remote sensing, IoT, smart city as well as indoor services such as human activity and gesture recognition. Novel RAN sharing SDR design will be needed to enable heterogeneity among distinct technologies, such that seamless ecosystem will play a key role in 6G.

Appendix A: eHSO: a Real-time eHealth Service Orchestrator for Medical Vehicles in 5G Networks[1]

[1]This work has been honored among the top fourteen projects participating to the IEEE ComSoc student competition 2020.

eHSO: a real-time eHealth Service Orchestrator for medical vehicles in 5G networks

M. Maule, *Student Member, IEEE*, A. Dalgkitsis, *Student Member, IEEE*, P.-V. Mekikis", *Member, IEEE*, A. Antonopoulos*, *Senior Member, IEEE,* John S.Vardakas", *Senior Member, IEEE,* and C. Verikoukis*, *Senior Member, IEEE*

Universitat Politècnica de Catalunya, Barcelona Spain
"Iquadrat Informatica S.L., Barcelona Spain
**Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Castelldefels Spain*

Abstract:
The next years will be revolutionary in the healthcare sector. 5G technology introduces novel use cases, enabling the cooperation of vertical markets for the definition of innovative cutting edge solutions. Autonomous healthcare vehicles like drones and smart ambulances will be everywhere in our cities, and advanced medical solutions will guide medical staff across different cases of emergency. As key factor of this futuristic vision, an excellent orchestration framework of the healthcare services should be investigated in order to guarantee system stability and reliability on top of the shared network infrastructure. Inspired from this challenge, this work presents a novel real-time orchestrator for the definition, provisioning, and supervision of 5G-based healthcare services, where the the network resources are dynamically tailored according to the healthcare vehicular requirements in terms of autonomous driving and quality of the services on board. The architecture and functional blocks of the proposed framework are detailed illustrated in the first part, while at the end the performance of an experimental healthcare scenario is examined, using a network simulator and a real testbed.

Index Terms: eHealth, eHealth drone, Connected ambulances, 5G orchestration, Advanced 5G services.

1. Introduction

The current world population of 7.6 billion is expected to reach 8.6 billion in 2030, introducing new challenges in terms of healthcare costs, pushing the healthcare industry to investigate novel solutions to improve the current healthcare system, while also boosting the access to greater healthcare services.

In our vision, through the exploitation of the novel 5G telecommunication standards, innovative healthcare solutions such as connected ambulances, healthcare drones, and remote real-time healthcare assistance represent key factors for the future healthcare system. Given the sharing and coverage capabilities expected with 5G, advanced healthcare use cases will be possible: ambulances will be able to communicate with multiple traffic entities (i.e., traffic lights, traffic APIs, traffic management services) to define the optimal route towards and emergency point, doctors may real-time remotely assist the medical staff on board the ambulance until reaching an hospital, and drones may be use to supply first aid and transport to injured persons inside impassable areas.

This novel trend not only improves the quality of the healthcare services: according to hundreds of executives from the healthcare and telecom sectors, the cooperation of 5G technology and healthcare solution could bring cost savings of approximately $94 billion USD to the healthcare industry in 2030 [2]. This alliance brings to the launch of a novel vertical market and service called *eHealth*, which the World Health Organisation (WHO) defined as the transfer of health resources

and healthcare by electronic means [3].

Despite the economical benefits of this technology applied in the eHealth market, the real value comes from the smart management and reuse of the resources to treat more patients and extend the healthcare coverage to every corner of the globe, leading to almost a billion extra patients treated globally each year by 2030.

In this work, we present a real-time eHealth slice-based service orchestration framework for the provisioning, control, and planning of the network edge and radio resources required to enable on-site medical aid, in response to an emergency request. The proposed platform could efficiently drive the eHealth vehicles towards the shortest path from the emergency site until a suitable destination healthcare infrastructure, while simultaneously planning the network resources along the predicted path so that advanced aid functionalities like high-definition (HD) video streams between the paramedics and the healthcare infrastructure are supplied with high Quality of Service (QoS).

Concerning our proposed framework, three fundamental enablers are needed to support the heterogeneous Key Performance Indicators (KPIs) for the novel eHealth use cases in a cost-efficient way: Network Slicing (NS), Software Defined Network (SDN), and Network Function Virtualization (NFV) [5].

As social impact on humanity or local community, the proposed solution drastically reduces the medical reaction time in response to an emergency call, increasing the probability to save lives. Medical staff with different specialization may remotely execute multiple operations without physically change site, the ambulances will move autonomously with a reduced human interaction, drones could monitor and forward relevant emergency information to an hospital, etc. The proposed resource orchestration framework will optimize the way we are employing the healthcare resources: real-time provisioning of healthcare equipment, monitoring of ambulance and drones availability, service extension to rural areas, and the acquisition of advanced data patterns will contribute to reuse in an efficient way all the healthcare resources which are nowadays available. As a consequence, we will have multiple benefits such as reduced human error, well defined emergency information, fast setup of specialized equipment, etc.

From a technological perspective, these benefits clearly show the relevance of 5G as the baseline technology to provide advanced quality care through intelligent data driven insights, increases the awareness/knowledge for the first responders, provides seamless critical eHealth services, and presents augmented flexibility and scalability adaptation capabilities according to dynamic variations of the network architecture and services priorities.

The rest of this document is structured as follows: Section II illustrates the framework architecture and functional blocks of the proposed solution. Section III describes the experimental scenario for an eHealth service, and the obtained performance. To conclude, Section IV presents the conclusions the future vision of our innovative idea.

2. eHealth System and Architecture Framework

2.1. 5G Ecosystem integration and Key Technology Enablers

An application of the proposed method is illustrated in Fig. 1, where the eHealth service architecture of our proposed solution is applied on top of a 5G network. Since each 5G network segment has specific capabilities and requirements, the architecture can be broken up into three layers [6]:

- The *Backhaul* is the connection between the midhaul and the internet\core network. In this layer we decided to setup a core functionality of our framework, the eHealth Service Orchestrator (eHSO). This application is responsible to estimate, monitoring, and orchestrating the network resources and thr QoS for a eHealth service, deployed among all the layers.
- The *Midhaul* network carries traffic between the Distributed Unit (DU) and the Centralized Unit (CU) and has tighter latency requirements such as $\sim$1-5 milliseconds (ms). This layer hosts Multi-Access Edge Computing (MEC) technology, facilitating the collection and processing of

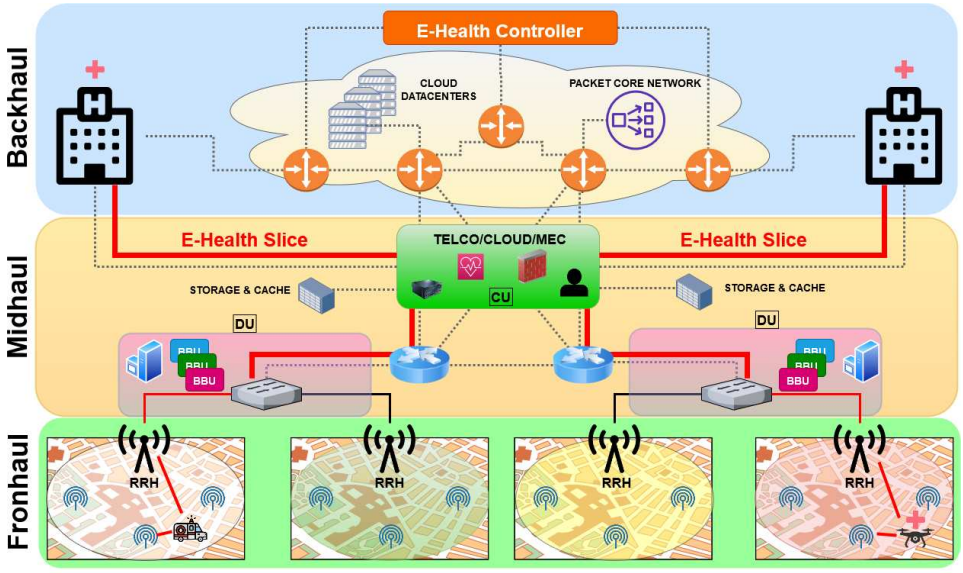


Fig. 1. 5G ecosystem architecture for eHealth

the data closer to the eHealth vehicle, reducing the latency and bringing real-time performance to high performance E-Heath applications (i.e., real-time healthcare video streaming, patient monitoring, eHealth vehicle autonomous driving, etc.).

- The *Fronthaul* is defined as the fiber-based connection in RAN infrastructure between the Baseband Unit (BBU) and Remote Radio Head (RRH). This structure guarantees a seamless eHealth services in the RAN part of the network, which is characterized by multiple small radio cells anchored to a macro cell.

While this section illustrated the technological aspects involved in our solution, the next part illustrate the eHealth service life-cycle management procedure performed by the eHSO.

2.2. eHealth Framework Features

Following a eHealth use case application, this section explains the system model and functional workflow of our solution, as illustrated in Fig. 2.

When an emergency call is executed, the local or national emergency call center automatically interacts with our proposed eHSO platform, which initializes the emergency protocol in response to the specific critical situation.

The definition and orchestration of a new service process by the eHSO is subdivided into four phases:

- **Preparation**: When an emergency request arrives to the eHSO, this block acquires the location and multiple information related to the emergency degree. Then, the eHSO selects the suitable healthcare infrastructure (i.e. hospital) able to supply the right assistance (i.e remote assistance, medical staff, etc) to the eHealth vehicle (ambulance, drone) taking into account the type of remote assistance, required equipment, and distance to reach the emergency point. This block should also estimate which type of physical (i.e. cache, specialized hardware, storage, radio resources) and virtualized (i.e., virtual machines, software applications, policies) resources must be considered before instantiating the eHealth slice service.
- **Creation**: Once the E2E network resources are estimated, the eHSO provisions and allocates them on top of the network infrastructure, according to the estimated physical route that will be followed by the eHealth vehicle. The premature allocation of the network resources along

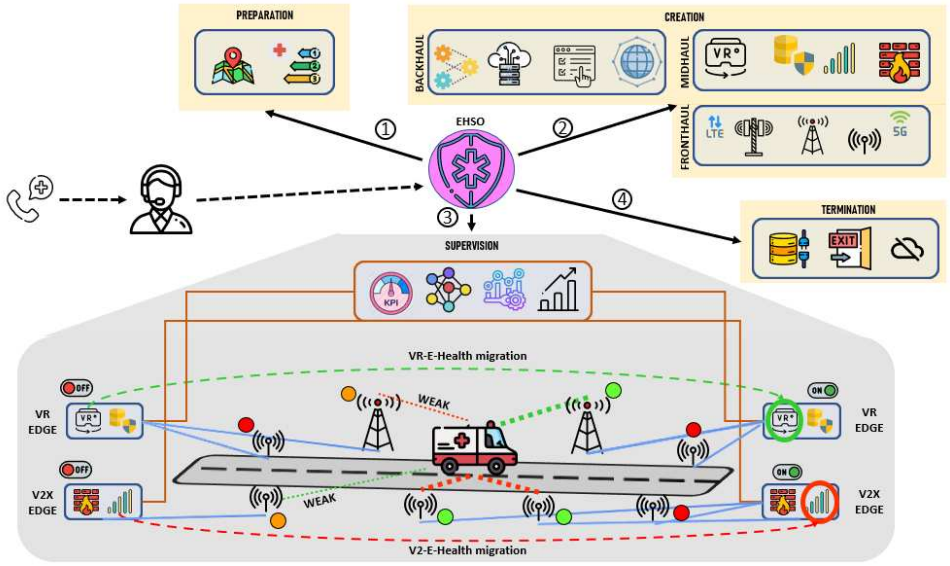


Fig. 2. Proposed framework features and workflow

the optimal vehicle path represent another advantage of our solution, since the timing to change the status from idle to active is much faster than dynamically allocate it according to the vehicle mobility. This principle increases the service stability and seamless properties of the eHealth service.

- **Supervision**: The eHealth vehicle and potential medical staff onboard must receive remote assistance from specialized doctors, until the vehicle reaches its final destination. For this reason, the eHSO progressively activates the network service functionalities according to the mobility of the eHealth vehicle towards the healthcare destination infrastructure. During this phase, the eHSO predicts the closest next position of the eHealth vehicle and monitors in real-time the Key Performance Indicators (KPIs) of the eHealth service and the network infrastructure to guarantee seamless service and minimal traffic congestion.
- **Termination**: During the service execution, the deprecated network resources are progressively decommissioned to mitigate the network load. This final block verifies that all network resource are correctly released and all slice policies are deactivated once the eHealth vehicle reaches the healthcare infrastructure. This phase concludes the E2E management procedure performed by the eHSO, and the system is ready to handle the next emergency call.

The eHSO has the capability to handle multiple services simultaneously, and the aforementioned procedure is applied to each one of them.

3. eHealth use case: Scenario and Results

This section illustrates the experimental tools and system performance of a typical eHealth scenario, considering the management of the eHealth entities through our proposed framework.

3.1. Scenario setting and experimental tools

As testing scenario, we selected an area of San Francisco Bay, as illustrated in Fig. 3. Given the expected coverage requirements of 5G inside an urban environment, the experimental area is divided into tree sectors, each one covered by a MEC entity. The MEC infrastructure is responsible for the management of a 5G macro cells cluster, illustrated with different color. Moreover, each macro cell is characterized by a master antenna and multiple small antennas, which cooperates together to supply a seamless service.

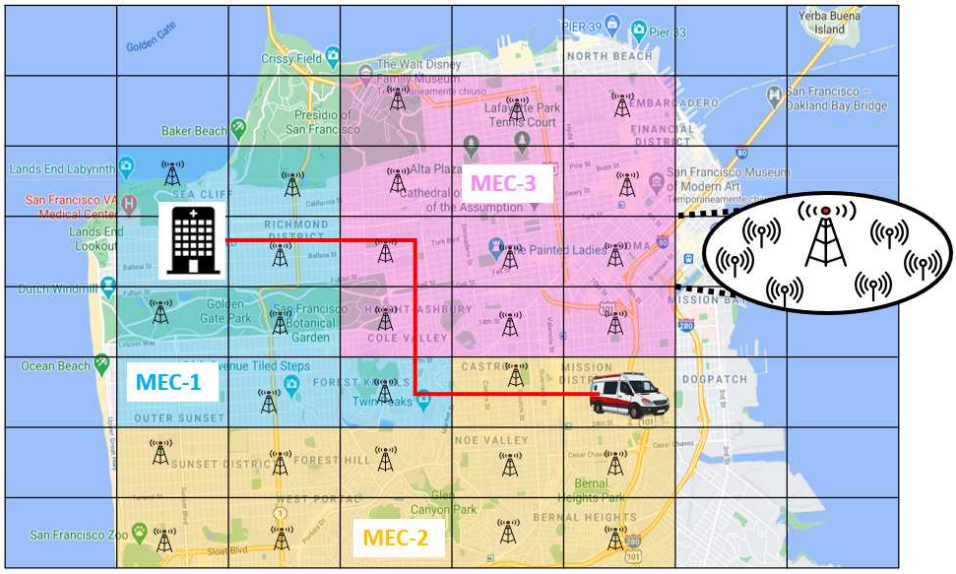


Fig. 3. Experimental scenario

Since autonomous driving is one of the main 5G features, for the analysis of the traffic mobility, we consider real data from vehicles roaming the San Francisco Bay area [7], where each vehicle is equipped with a 5G service. Two types of service priority are evaluated: i) *critical*, such as eHealth, police, firefighters, and ii) *normal*, like usual autonomous driving traffic, best effort traffic, etc.

Following our network resources training model, the eHSO is able to forecast when the eHealth slice settings should be migrated among the MEC infrastructures before the eHealth vehicle and/or the medical staff experience low QoS or service interruption.

For the mobility prediction of the eHealth vehicles, *TensorFlow* [8] and *Keras API* [9] are used to define a Convolutional Neural Network (CNN). Inconsistencies in the data set were filtered using *MATLAB*, and the dynamic service orchestration algorithm was coded using *Python*.

Based on the amount of *critical* services, the eHSO may decide to redirect *normal* priority traffic towards secondary nodes, while keeping eHealth service instances inside the nodes close to the predefined path of the eHealth vehicle. This functionality introduces a further degree of strength to our proposed model, since short range distance among the eHealth entity has a positive impact to the QoS in terms of latency and maximum data rate, while *normal* services performance are not affected since their requirements are less strict.

The prioritization of the eHealth vehicle and other services in the MEC infrastructures is modelled as a Multi-dimensional Knapsack Problem, and its solution is estimated with the *Pyeasyga* module, using genetic algorithms [10].

Regarding the management of the network and radio resources in the fronthaul network layer, other tools and consideration are deployed. On our experimental scenario, the management of the radio resource among the small cells is perform using OpenAirInterface (OAI) [11], while the radio system is based on Universal Peripheral Radio Software (USRP) B210 [12]. Raspberry Pi [13] equipped with an LTE module is utilized as end device to simulate the mobility and streaming traffic performance of an eHealth vehicle.

In the next section, the performance of the aforementioned scenario are illustrated, emphasizing the time response and throughput performance of the migration procedure among adjacent cells or MEC areas.

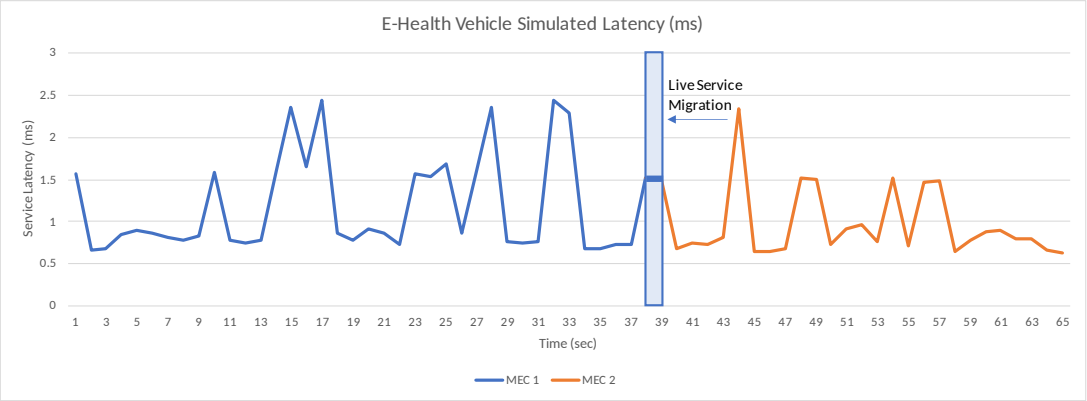


Fig. 4. E2E latency between an eHealth vehicle and the dynamic MEC service for 65 seconds of operation.

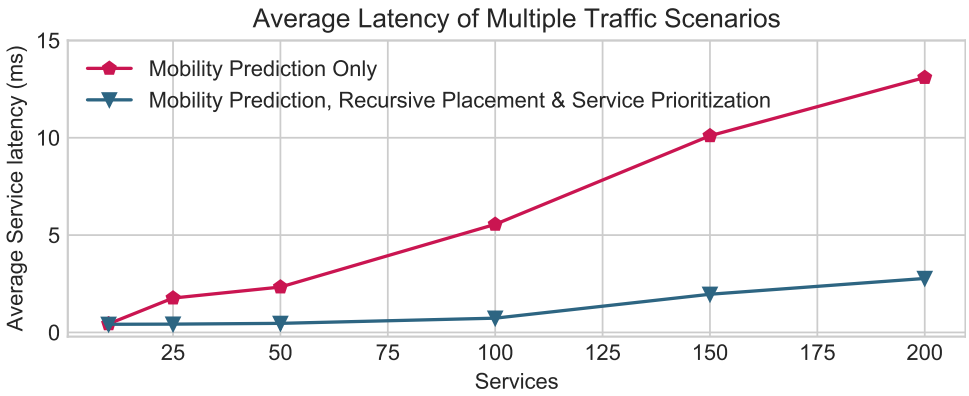


Fig. 5. Average service latency of multiple scenarios.

3.2. Results

The performance evaluation of the proposed eHSO framework is divided into two parts: the MEC status live migration among macro cells belonging to different MECs, and the slice service handover among small cells within a single macro cell.

Fig. 4 highlights the latency response of the proposed orchestrator when the eHealth vehicle across two macro cells of different MEC instances and a live migration of the eHealth midhaul Virtual Machine(VM) is performed to maintain a seamless service. Thanks to the forecasting method performed by the eHSO, a copy of the VM is sent on time towards the next MEC instance, and progressively activated according to the eHealth service KPI and vehicle mobility. As highlighted in the figure, this technique maintains a seamless service between the two MEC entities, without unexpected latency peaks during the switching operation.

Service prioritization and dynamic placement bring further performance enhancement among the MEC instances. As illustrated in Fig. 5, for an increasing number of vehicles (services), considering just the mobility factor (red line) as input variable for the MEC migration, a performance degradation is experienced, since the latency increases almost linearly with the number of vehicles. Adding to the model a service prioritization property and a reallocation of low priority flows have a concrete impact on the eHealth service latency (blue line), since leading MEC resources can be assigned to *critical* services, while *normal* traffic can be rerouted to other nodes, without affecting the required QoS.

Focusing our analysis to a single macro cell, Fig. 6 illustrates the activation/deactivation process among small cells, according to the mobility of the eHealth vehicle and its services. Through

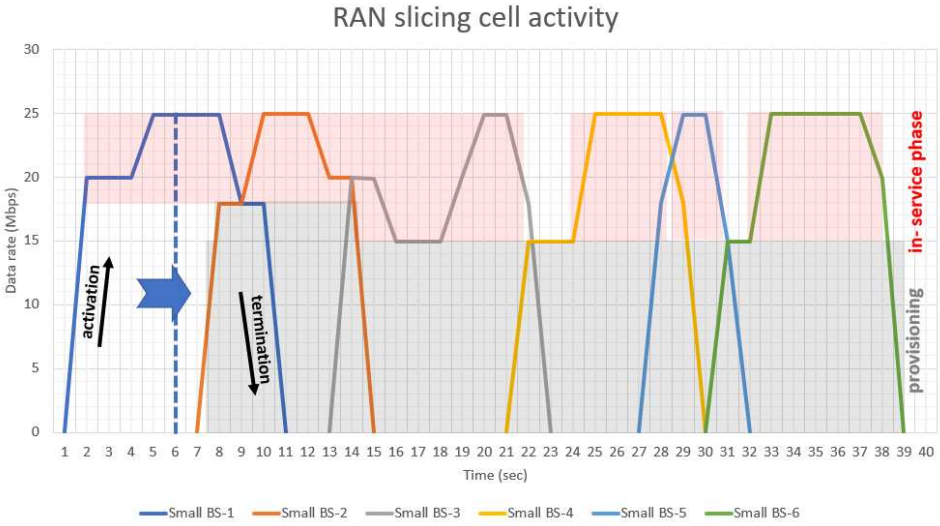


Fig. 6. Average service latency of multiple scenarios.

the evaluation of the channel quality, it is possible to estimate which small cell can supply the optimal service to the eHealth vehicle. Since the eHeath vehicle is moving, novel small cell with better coverage capabilities are constantly activated, while the old one are deactivated and decommissioned.

The eHSO performs the MEC and RAN slicing procedures as two parallels threads. This implementation choice reduces the process scheduling timing with a relevant impact on the real-time performance of our system.

4. Conclusions & Future Work

This work presented a real-time slice-based service solution for the optimal management of healthcare resources under an emergency situations. By following our approach, the healthcare system can benefit from novel 5G telecommunication technologies to upgrade the current medical capabilities and, thus, bring remarkable benefits to the society. In our vision, a dedicated eHealth dynamic network carefully orchestrated by the EHMS represents the innovative factor for a seamless multi-service future architecture, free from network domains obstacles, where ultra reliable low latency and guarantee throughput services are the reality, reaching high quality standards in the healthcare sector never seen before.

The experimental results confirm that the proposed novel eHealth architecture is able: i) to reduce the healthcare management costs, ii) to enhance the healthcare resources overview, iii) to supply advanced medical services in areas hard to reach currently, and iv) to increase the possibility to save lives exploiting new technological solutions.

In the future, we plan to include advanced antenna patterns to reach higher data rates, while increasing the system scalability through the distribution of functionalities at multiple technological domains, e.g., edge, cloud, RAN. As a result, the effectiveness of our proposal will be improved even further allowing the service of more patients.

References

[1] United Nations, ”World Population Prospects”, New York, June 2017
[2] Huawei, ”5G’s healthcare impact: 1 billion patients with improved access in 2030”, October 2019
[3] 5GPPP, ”5G and e-Health”, September 2015
[4] 5GPPP Architecture Working Group, ”View on 5G Architecture”, Version 3.0, 2019
[5] Alliance, N. G. M. N. ”Description of network slicing concept.” NGMN 5G P 1.1 (2016).
[6] 5G Americas Whitepaper, ”5G at the Edge”, October 2019
[7] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, “CRAW-DAD dataset epfl/mobility (v. 2009-02-24),” Downloaded from https://crawdad.org/epfl/mobility/20090224/cab, Feb. 2009, traceset: cab.
[8] M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/
[9] F. Chollet et al., “Keras,” https://keras.io, 2015.
[10] S. Shah, “Genetic algorithm for the 0/1 multidimensional knapsack problem,” 2019.
[11] Nikaein, N. et al. ”OpenAirInterface: A flexible platform for 5G research.” ACM SIGCOMM Computer Commun. Review 44.5 (2014): 33-38.
[12] https://www.ettus.com/all-products/UB200-KIT/
[13] https://www.raspberrypi.org/products/raspberry-pi-4-model-b/

Appendix B: Proof of Lemma 1

Consider a scenario with the following parameters:

- $w_{u,n}^{j} \in \boldsymbol{W_n^j}, \forall i \in \boldsymbol{U_n^j}$
- $\boldsymbol{W_n^j}$ is the weight vector
- $\boldsymbol{U_n^j}$ = users served by slice S_n^j
- $\varphi_{u,n}^{j}$ = priority of user u of slice S_n^j

Determine

$$\boldsymbol{W_n^j} = \max_{w_{u,n}^{j}} \sum_{u \in \boldsymbol{U_n^j}} \varphi_{u,n}^{j} \times \log(\frac{w_{u,n}^{j} \times PRB_{u,n}^{j}}{\sum_{u \in \boldsymbol{U_n^j}} w_{u,n}^{j}})$$

s.t.

$$\sum_{u \in \boldsymbol{U_n^j}} w_{u,n}^{j} = C_n^j,$$

$$PRB_{u,n}^{j} = \frac{r_{u,n}^{j} \times 10^3}{N_{n,sc}^{j} \times N_v^{cell} \times f^{cell} \times R_{n,code}^{j} \times 168 \times (1 - OH_n^j) \times \log_2(QAM_n^j) \times 2^{\mu}}. \tag{6.1}$$

proof:

Given slice S_n^j, the overall load at cell rrh can be decomposed as:

$$L_b(w_{u,n}^{j}) = \sum_{u \in \boldsymbol{U_n^j}} w_{u,n}^{j} = a_{rrh}(\overline{w}_{u,n}^{j}) + d_{rrh}(w_{u,n}^{j})$$

where

$$a_{rrh}(\overline{w}_{u,n}^{j}) = \sum_{n' \backslash n}^{N^j} \sum_{u}^{U_{n'}^j} w_{u,n}^{j}, \text{(sum of weights of all users served by cell } rrh\text{, except slice } S_n^j)$$

$d_{rrh}(w^j_{u,n}) = \sum_{u}^{U^j_n} w^j_{u,n}$, (sum of all weights of users of all users served by slice S^j_n in cell rrh)

Since the weights of other slices are given, we can assume $a_{rrh}(w^j_{u,n})$ constant.

Using the Lagrange multipliers method, it is possible to determine a strategy for finding the local maxima of a function subject to equality constraints (i.e., subject to the condition that one or more equations have to be satisfied exactly by the chosen values of the variables).

$$L(w,\lambda) = \sum_{u\in \boldsymbol{U^j_n}} \varphi^j_{u,n} \times log(\frac{w^j_{u,n} \times PRB^j_{u,n}}{\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n}}) - \lambda \times (\sum_{u}^{U^j_n} w^j_{u,n} - C^j_n)$$

and solve $\dfrac{\partial L(w,\lambda)}{\partial w, \lambda} = 0$.

Solving the derivative respect $w^j_{u,n}$:

$$\frac{\partial L(\boldsymbol{W^j_n},\lambda)}{\partial w^j_{u,n}} =$$

$$= \sum_{u\in \boldsymbol{U^j_n}} (\varphi^j_{u,n} \times (\frac{\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n}}{w^j_{u,n} \times PRB^j_{u,n}} \times \frac{\partial}{\partial w^j_{u,n}}(\frac{w^j_{u,n} \times PRB^j_{u,n}}{\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n}})) - \lambda \times \frac{\partial}{\partial w^j_{u,n}}(\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n}) - \frac{\partial}{\partial w^j_{u,n}}(\lambda \times C^j_n))$$

$$= \sum_{u\in \boldsymbol{U^j_n}} \varphi^j_{u,n} \times (\frac{L_b(w^j_{u,n})}{w^j_{u,n} \times PRB^j_{u,n}} \times \frac{PRB^j_{u,n} \times L_b(w^j_{u,n}) - (w^j_{u,n} \times PRB^j_{u,n}) \times \frac{\partial}{\partial w^j_{u,n}} \sum_{u'\in \boldsymbol{U^j_n}} w^n_{u'}}{L_b(w^j_{u,n})^2}) - \lambda$$

$$= \sum_{u\in \boldsymbol{U^j_n}} (\frac{\varphi^j_{u,n} \times L_b(w^j_{u,n})^2 \times PRB^j_{u,n}}{w^j_{u,n} \times PRB^j_{u,n} \times L_b(w^j_{u,n})^2} - \frac{(w^j_{u,n} \times PRB^j_{u,n} \times \varphi^j_{u,n} \times L_b(w^j_{u,n}) \times \frac{\partial}{\partial w^j_{u,n}} \sum_{u'\in \boldsymbol{U^j_n}} w^n_{u'})}{w^j_{u,n} \times PRB^j_{u,n} \times L_b(w^j_{u,n})^2}) - \lambda$$

$$= \sum_{u\in \boldsymbol{U^j_n}} (\frac{\varphi^j_{u,n}}{w^j_{u,n}} - \frac{(\varphi^j_{u,n} \times \frac{\partial}{\partial w^j_{u,n}} \sum_{u'\in \boldsymbol{U^j_n}} w^n_{u'})}{L_b(w^j_{u,n})}) - \lambda$$

focusing only on $w^j_{u,n}$:

$$= \frac{\varphi^j_{u,n}}{w^j_{u,n}} - \frac{(\varphi^j_{u,n})}{L_b(w^j_{u,n})} - \frac{\frac{\partial}{\partial w^j_{u,n}} \sum_{u'\setminus i\in \boldsymbol{U^j_n}} \varphi^{S^j_n}_{u'}}{L_b(w^j_{u,n})} - \lambda = \frac{\varphi^j_{u,n}}{w^j_{u,n}} - \frac{(\varphi^j_{u,n})}{L_b(w^j_{u,n})} - \lambda.$$

$$\frac{\partial L(\boldsymbol{W^j_n},\lambda)}{\partial \lambda_i} = \frac{\varphi^j_{u,n}}{w^j_{u,n}} - \frac{(\varphi^j_{u,n})}{L_b(w^j_{u,n})} - \lambda$$

$$= \frac{\partial}{\partial \lambda_i}(\sum_{u\in \boldsymbol{U^j_n}} \varphi^j_{u,n} \times log(\frac{w^j_{u,n} \times PRB^j_{u,n}}{\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n}}) - \lambda_i \times (\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n} - C^j_n))$$

$$= \frac{\partial}{\partial \lambda_i}(-\lambda_i \times (\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n} - C^j_n)) = -\sum_{u\in \boldsymbol{U^j_n}} w^j_{u,n} + C^j_n \tag{6.2}$$

$$\text{The final system of equations is: } \begin{cases} \frac{\varphi^j_{u,n}}{w^j_{u,n}} - \frac{\varphi^j_{u,n}}{L_b(w^j_{u,n})} - \lambda_i = 0 \\ -\sum_{u \in \boldsymbol{U^j_n}} w^j_{u,n} + C^j_n = 0 \end{cases} \tag{6.3}$$

To proof that system of equations admits unique solution for $w^j_{u,n}$, we need to equalize the previous equation for two different users of the same slice:

$$\frac{w^j_{u,n}}{w^j_{u',n}} = \frac{\varphi^j_{u,n} \times (L_b(w^j_{u,n}) - w^j_{u,n})}{\varphi^{S^j_n}_{u'} \times (L_b(w^j_{u,n}) - w^j_{u',n})} = H \times \frac{(L_b(w^j_{u,n}) - w^j_{u,n})}{(L_b(w^j_{u,n}) - w^j_{u',n})} = H \times \frac{a_{rrh}(\overline{w}^j_{u,n}) + \sum_{u' \setminus i \in \boldsymbol{U^j_n}} w^j_{u',n}}{a_{rrh}(\overline{w}^j_{u,n}) + \sum_{u \setminus u' \in \boldsymbol{U^j_n}} w^j_{u,n}}$$

$$w^j_{u,n} = H \times w^j_{u',n} \times \frac{a_{rrh}(\overline{w}^j_{u,n}) + \sum_{u' \setminus i \in \boldsymbol{U^j_n}} w^j_{u',n}}{a_{rrh}(\overline{w}^j_{u,n}) + \sum_{u \setminus u' \in \boldsymbol{U^j_n}} w^j_{u,n}}$$

$$w^j_{u,n} \times (a_{rrh}(\overline{w}^j_{u,n}) + \sum_{u \setminus u' \in \boldsymbol{U^j_n}} w^j_{u,n}) = H \times w^j_{u',n} \times (a_{rrh}(\overline{w}^j_{u,n}) + \sum_{u' \setminus u \in \boldsymbol{U^j_n}} w^j_{u',n}) \tag{6.4}$$

where H is a constant greater than 0.

If we fix $w^j_{u,n}$ to some positive value, there exists a unique positive value of $w^j_{u',n}$ that satisfies the above equation. Indeed, the LHS of the equation is fixed to a value greater than 0, while the RHS can grow following $w^j_{u',n}$. From above, we can compute $w^j_{u',n}$ values as a function of single $w^j_{u,n}$. Once we have all the $w^j_{u',n}$, we can uniquely compute $w^j_{u,n}$. Inserting the weights into the second equation of the Lagrange system we have $\sum_{u \in \boldsymbol{U^j_n}} w^j_{u,n} = C^j_n$, which is an equation of a single unknown. This equation admits unique solution, as the LHS is a linear increasing function, while RHS is constant. Since all relationships are bijective, this is the only solution of the system.

Bibliography

[1] 5G-PPP, "5G: Serving Vertical Industries", the 2nd 5G Verticals Workshop, Brussels, 9 November 2015.

[2] BEREC, "Report on infrastructure sharing", 14 July 2018.

[3] Redana, S., Bulakci, Ö., Zafeiropoulos, A., Gavras, A., Tzanakaki, A., Albanese, A., ... Zhang, Y. (2019). 5G PPP architecture working group: View on 5G architecture.

[4] GSMA, "Network Slicing: use case requirements", April 2018.

[5] C. Song, M. Zhang, Y. Zhan, D. Wang, L. Guan, W. Liu, L. Zhang, S. Xu, Hierarchical Edge Cloud Enabling Network Slicing for 5G Optical Fronthaul, J. Opt. Commun. Netw. 11 (2019) B60–B70.

[6] X. Li, C. Guo, L. Gupta, R. Jain, Efficient and Secure 5G Core Network Slice Provisioning Based on VIKOR Approach, IEEE Access 7 (2019) 150517–150529.

[7] Sama, Malla Reddy, et al. "Reshaping the mobile core network via function decomposition and network slicing for the 5G era." 2016 IEEE Wireless Communications and Networking Conference. IEEE, 2016.

[8] Arteaga, Carlos Hernan Tobar, Armando Ordo˜nez, and Oscar Mauricio Caicedo Rendon. "Scalability and performance analysis in 5G core network slicing." IEEE Access 8 (2020): 142086-142100.

[9] Navarro-Ortiz, Jorge, et al. "Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond." IEEE access 8 (2020): 79604-79618.

[10] Ferrus, Ramon, et al. "On the automation of RAN slicing provisioning and cell planning in NG-RAN", European Conference on Networks and Communications, 2018.

[11] Kourtis, Michail-Alexandros, et al. "Network slicing for 5G edge services." Internet Technology Letters: e289.

[12] Bhushan, Naga, et al. "5G air interface system design principles." (2017): 6-8.

[13] K.Jihas, and L. Jacob. "Resource Allocation for CoMP Enabled URLLC in 5G C-RAN Architecture." IEEE SJ (2020).

[14] Sui, Wenshu, et al. "Energy-Efficient Resource Allocation With Flexible Frame Structure for Hybrid eMBB and URLLC Services." IEEE Transactions on Green Communications and Networking 5.1 (2020): 72-83.

[15] Wang, Ru-Jun, et al. "Resource Allocation in 5G with NOMA-Based Mixed Numerology Systems." GLOBECOM 2020- 2020 IEEE Global Communications Conference. IEEE, 2020.

[16] Lloret, Jaime, et al. "An architecture and protocol for smart continuous eHealth monitoring using 5G." Computer Networks 129 (2017): 340-351.

[17] Raza, M. R., Natalino, C., O¨ hlen, P., Wosinska, L., & Monti, P. (2019).Reinforcement learning for slicing in a 5G flexible RAN. Journal of Lightwave Technology, 37(20), 5161-5169.

[18] Sciancalepore, V., Samdanis, K., Costa-Perez, X., Bega, D., Gramaglia, M., & Banchs, A. (2017, May). Mobile traffic forecasting for maximizing 5G network slicing resource utilization. In IEEE INFOCOM 2017-IEEE Conference on Computer Communications (pp. 1-9). IEEE.

[19] Dandachi, Ghina, et al. "An artificial intelligence framework for slice deployment and orchestration in 5G networks." IEEE Transactions on Cognitive Communications and Networking 6.2 (2019): 858-871.

[20] Zhao, et al. "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks", IEEE Transactions on Wireless Communications, 18(11), 5141-5152.

[21] Nokia, "Mapping demand: the 5G opportunity in enterprise for communications service providers", Nokia 5G enterprise report,January 2020.

[22] Huawei, "Wireless Fiber: 4G/5G FWA Broadband", Huawei Technologies Co., Ltd, 2019.

[23] GSMA, "5G Network Slicing for Vertical Industries", White Paper, September 2017.

[24] ATIS, "5G Specifications in 3GPP: North American Needs for the 5G Future", ATIS-I-0000078 Report, July 2020.

[25] Qualcomm, "Propelling 5G forward. A closer look at 3GPP Release 16", Qualcomm Report, July 2020.

[26] Series, M. (2015). IMT Vision–Framework and overall objectives of the future development of IMT for 2020 and beyond. Recommendation ITU, 2083, 0.

[27] Chandramouli, D., Liebhart, R., and Pirskanen, J. (Eds.). (2019). 5G for the Connected World. John Wiley & Sons.

[28] Sampson N., Erfanian J., and Hu N., "5G White Paper 2", NGMN Alliance, 27th July 2020.

[29] Deasy M., "Value is at the heart of network slicing monetization", Ericsson blog, July 2021.

[30] Telenor Group, "Capturing the societal benefits of 5G", white paper, November 2019.

[31] GSMA, "The 5G guide: a reference for operators", white paper, April 2019.

[32] Ericsson, "Cellular networks for Massive IoT", Ericsson white paper, January 2020.

[33] 5GPPP, "5G innovations for new business opportunities", white paper, March 2017.

[34] Eiman M., "Minimum Technical Performance for Requirements for IMT-2020 radio interface(s)", ITU-R Workshop on IMT-2020 terrestrial radio interfaces, Geneva 2021.

[35] 3GPP TR 22.891, "Feasibility Study on New Services and Markets Enablers: Stage 1", Release 14, V2.0.0 (2016-02).

[36] 5G-ACIA, "Key 5G Use Cases and Requirements", white paper, May 2020.

[37] 3GPP TS 22.104. "Service requirements for cyber-physical control applications in vertical domains", Release 16, September 2020.

[38] IEC 61907. "Communication network dependability engineering".

[39] 3GPP TS 22.104, "5G; Service requirements for cyber-physical control applications in vertical domains", September 2020.

[40] 3GPP TS 22.186, "5G; service requirements for enhanced V2X scenarios", v. 15.3.0 Release 15, July 2018.

[41] 5G Americas, "5G Spectrum Recommendations", white paper, April 2017.

[42] GSMA Intelligence, "Network automation revisited: the 5G priority", technical report, April 2021.

[43] Qualcomm, Nokia, "Making 5G a reality: Addressing the strong mobile broadband demand in 2019 and beyond", technical white paper, September 2017.

[44] Huawei, "5G Spectrum Public Policy Position", White Paper, 2017.

[45] Vihriala, J., Ermolova, N., Lahetkangas, E., Tirkkonen, O., & Pajukoski, K. (2015, May). On the waveforms for 5G mobile broadband communications. In 2015 IEEE 81st Vehicular Technology Conference (VTC Spring) (pp. 1-5). IEEE.

[46] Doré, J. B., Gerzaguet, R., Cassiau, N., & Ktenas, D. (2017). Waveform contenders for 5G: Description, analysis and comparison. Physical communication, 24, 46-61.

[47] Lien, S. Y., Shieh, S. L., Huang, Y., Su, B., Hsu, Y. L.,& Wei, H. Y. (2017). 5G new radio: Waveform, frame structure, multiple access, and initial access. IEEE communications magazine, 55(6), 64-71.

[48] Pedersen, K. I., Berardinelli, G., Frederiksen, F., Mogensen, P., & Szufarska, A. (2016). A flexible 5G frame structure design for frequency-division duplex cases. IEEE Communications Magazine, 54(3), 53-59.

[49] Mogensen, P., Pajukoski, K., Tiirola, E., Lähetkangas, E., Vihriälä, J., Vesterinen, S., ... & Cattoni, A. F. (2013, December). 5G small cell optimized radio design. In 2013 IEEE Globecom Workshops (GC Wkshps) (pp. 111-116). IEEE.

[50] 3GPP TS 38.331, TSG RAN; NR; Radio Resource Control (RRC) protocol specification, Release 15, v15.3.0, Sept. 2018.

[51] 3GPP TS 38.214, TSG RAN; NR; Physical layer procedures for data, Release 15, v15.3.0, Sept. 2018.

[52] Abdelwahab, S., Hamdaoui, B., Guizani, M., & Znati, T. (2016). Network function virtualization in 5G. IEEE Communications Magazine, 54(4), 84-91.

[53] ETSI, "Network Functions Virtualisation:An Introduction, Benefits, Enablers, Challenges & Call for Action", White Paper, Darmstadt-Germany, October 2012.

[54] Hakiri, A., & Berthou, P. (2015). Leveraging SDN for the 5G networks: Trends, prospects and challenges. arXiv preprint arXiv:1506.02876.

[55] 5GPPP Architecture Working Group, "View on 5G Architecture", Version 2.0, 2017.

[56] 5GPPP, "Vision on Software Networks and 5G", SN WG white paper version 2.0, January 2017.

[57] Chih-Lin, I., Huang, J., Duan, R., Cui, C., Jiang, J., & Li, L. (2014). Recent progress on C-RAN centralization and cloudification. IEEE Access, 2, 1030-1039.

[58] Harutyunyan, D., & Riggio, R. (2018). Flex5G: Flexible functional split in 5G networks. IEEE Transactions on Network and Service Management, 15(3), 961-975.

[59] 3GPP TS RAN, "Study on new radio access".

[60] L. M. P. Larsen, A. Checko and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," IEEE Commun. Surveys Tuts., vol. 21, no. 1, pp. 146-172, Firstquarter 2019.

[61] ITU-T, "Transport network support of IMT-2020/5G", Geneva, Switzerland, Rep. GSTR-TN5G, 2018.

[62] Niknam, Solmaz, et al. "Intelligent O-RAN for beyond 5G and 6G wireless networks."

[63] GSA, "5G Network Slicing for Vertical Industries", White Paper, September 2019.

[64] Makhijani, K., et al. "Network slicing use cases: Network customization and differentiated services." draft-netslices-usecases-02 (2017).

[65] 3GPP TR 22.852,"Study on Radio Access Network (RAN) sharing enhancements", v13.1.0, Sep. 2014.

[66] Flinck, Hannu, et al. "Network Slicing Management and Orchestration." Internet Engineering Task Force.

[67] GSMA, "E2E Network Slicing Architecture", Version 1.0, June 2021.

[68] Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., & Verikoukis, C. (2019, December). Real-time dynamic network slicing for the 5G radio access network. In 2019 IEEE Global Communications Conference (GLOBECOM) (pp. 1-6). IEEE.

[69] GSTR-TN5G, ITUT. "Technical Report "Transport network support of IMT-2020/5G", February 2018.

[70] D. Wang ,T. Sun, "Service-Based Architecture in 5G", NGMN Alliance, Deliverable V1.0, January 2018.

[71] GSMA, "5G Roaming Guidelines", Version 2.0, May 2020.

[72] "3GPP TS 23.501, "5G; System architecture for the 5G System (5GS)", version 16.6.0 Release 16, October 2020.".

[73] "3GPP TS 23.502, "5G; Procedures for the 5G System (5GS) ", version 16.10.0 Release 16, September 2021.".

[74] "3GPP TS 23.503, "5G; Policy and charging control framework for the 5G System (5GS); Stage 2", version 15.5.0 Release 15, April 2019.".

[75] "Lei, W., Soong, A. C., Jianghua, L., Yong, W., Classon, B., Xiao, W., ... & Saboorian, T. (2021). 5G system architecture. In 5G System Design (pp. 297-339). Springer, Cham.".

[76] Samsung, "5G Standalone Architecture", Technical White Paper, January 2021.

[77] GSMA, "Road to 5G: Introduction and Migration", Technical White Paper, April 2018.

[78] NGMN, "Option 4 as a 5G SA Complement - Option 4 for smooth 5G NSA-SA Migration", Technical White Paper v1.01, February 2021.

[79] Zakeri, A., Gholipoor, N., Tajallifar, M., Ebrahimi, S., Javan, M. R., Mokari, N., & Sharafat, A. R. (2020, December). Digital transformation via 5G: Deployment plans. In 2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K) (pp. 1-8). IEE

[80] The OpenAirInterface Initiative, http://www.openairinterface.org/.

[81] Kaltenberger, F., Silva, A. P., Gosain, A., Wang, L., & Nguyen, T. T. (2020). OpenAirInterface: Democratizing innovation in the 5G Era. Computer Networks, 176, 107284.

[82] Nikaein, N., Marina, M. K., Manickam, S., Dawson, A., Knopp, R., & Bonnet, C. (2014). OpenAirInterface: A flexible platform for 5G research. ACM SIGCOMM Computer Communication Review, 44(5), 33-38.

[83] K. Tan et al. Sora: High-Performance Software Radio Using General-Purpose Multi-Core Processors. Communications of the ACM, 54(1):99–107, Jan 2011.

[84] Kaltenberger, F., Knopp, R., Roux, C., Matzakos, P., Mani, F., & Velumani, S, "OpenAir-Interface 5G NSA system with COTS phone", International Workshop on Wireless Network Testbeds, Experimental evaluation Characterization, September 2020.

[85] https://www.ettus.com/all-products/UB200-KIT/

[86] Foukas, X., Nikaein, N., Kassem, M. M., Marina, M. K., Kontovasilis, K. (2017, October). Flexran: A Software-Defined RAN Platform. In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (pp. 465-467).

[87] Gebremariam, A. A., Usman, M., Du, P., Nakao, A., & Granelli, F. (2017, December). Towards E2E slicing in 5G: a spectrum slicing testbed and its extension to the packet core. In 2017 IEEE Globecom Workshops (GC Wkshps) (pp. 1-6). IEEE.

[88] Raza, M. R., Fiorani, M., Rostami, A., Öhlen, P., Wosinska, L., & Monti, P. (2018). Dynamic slicing approach for multi-tenant 5G transport networks. Journal of Optical Communications and Networking, 10(1), A77-A90.

[89] Li, R., Zhao, Z., Sun, Q., Chih-Lin, I., Yang, C., Chen, X., ... & Zhang, H. (2018). Deep reinforcement learning for resource management in network slicing. IEEE Access, 6, 74429-74441.

[90] Lee, Y. L., Loo, J., Chuah, T. C., & Wang, L. C. (2018). Dynamic network slicing for multi-tenant heterogeneous cloud radio access networks. IEEE Transactions on Wireless Communications, 17(4), 2146-2161.

[91] Maule, M., Mekikis, P. V., Ramantas, K., Vardakas, J., & Verikoukis, C. (2020, December). Dynamic partitioning of radio resources based on 5G RAN Slicing. In GLOBECOM 2020-2020 IEEE Global Communications Conference (pp. 1-6). IEEE.

[92] Qualcomm and Nokia, white paper, "Making 5G a reality: Addressing the strong mobile broadband demand in 2019 beyond", Sept. 2017.

[93] Nomor research GmbH, "3GPP 5G Adhoc:Any Decisions on RAN Internal Functional Split?", Munich, Germany, January 26, 2017.

[94] N. Van Huynh, "Optimal and fast real-time resource slicing with deep dueling neural networks" IEEE Journal on Selected Areas in Communications, 2019.

[95] J. Kaufman, "Blocking in a Shared Resource Environment," IEEE Trans. Commun., vol. 29, no. 10, Oct. 1981.

[96] Nomor research GmbH, "3GPP 5G Adhoc:Any Decisions on RAN Internal Functional Split?", Munich, Germany, January, 2017

[97] L. Peterson et al., "Central Office Re-Architected as a Data Center," IEEE Commun. Mag., vol. 54, no. 10, Oct. 2016.

[98] R. Riggio et al., "Programming Abstractions for Software-Defined Wireless Networks," Trans. Network and Service Management, vol. 12, no. 2, 2015, pp. 146–62.

[99] ECOMP (Enhanced Control, Orchestration, Management & Policy)Architecture White Paper, AT&T Inc., 2016.

[100] OPEN-O Project at a Glance, open-o.org, 2017.

[101] Katsalis et al., "Multi-Domain Orchestration for NFV: Challenges and Research Directions," Int'l. Conf. Ubiquitous Computing and Commun. and Int'l. Symp. Cyberspace and Security, 2016, pp. 189–95.

[102] Ksentini, Adlen, and Navid Nikaein. "Toward enforcing network slicing on RAN: Flexibility and resources abstraction." IEEE Communications Magazine 55.6 (2017): 102-108.

[103] Xiang, Hongyu, et al., "Network slicing in fog radio access networks: Issues and challenges. IEEE Communications Magazine", 2017, vol. 55, no 12, p. 110-116.

[104] Chang, Chia-Yu, NIKAEIN, Navid, SPYROPOULOS, Thrasyvoulos, "Radio access network resource slicing for flexible service execution", IEEE Infocom, 2018. p. 668-673.

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

[106] Maule, Massimiliano, John S. Vardakas, and Christos Verikoukis. "A Novel 5G-NR Resources Partitioning Framework Through Real-Time User-Provider Traffic Demand Analysis." IEEE Systems Journal (2021).

[107] Oladejo, Sunday Oladayo, and Olabisi Emmanuel Falowo. "Latency-aware dynamic resource allocation scheme for multi-tier 5G network: A network slicing-multitenancy scenario." IEEE Access 8 (2020): 74834-74852.

[108] Abiko, Y., Saito, T., Ikeda, D., Ohta, K., Mizuno, T., & Mineno, H. (2020). Flexible resource block allocation to multiple slices for radio access network slicing using deep reinforcement learning. IEEE Access, 8, 68183-68198.

[109] Mei, Jie, Xianbin Wang, and Kan Zheng. "An intelligent self-sustained RAN slicing framework for diverse service provisioning in 5G-beyond and 6G networks." Intelligent and Converged Networks 1.3 (2020): 281-294.

[110] "Study on Radio Access Network (RAN) sharing enhancements," TR 22.852, v13.1.0, Sep. 2014.

[111] Zhang, Li. "Proportional response dynamics in the Fisher market." Theoretical Computer Science (2011).

[112] Arrow, K J., et al. "Capital-labor substitution and economic efficiency." The review of Economics and Statistics (1961).

[113] Caballero, Pablo, et al. "Network slicing games: Enabling customization in multi-tenant networks." IEEE INFOCOM 2017-IEEE Conference on Computer Communications. IEEE, 2017.

[114] Navarro-Ortiz, Jorge, et al. "A survey on 5G usage scenarios and traffic models." IEEE Communications Surveys & Tutorials 22.2 (2020): 905-929.

[115] LARSEN, Line MP; CHECKO, Aleksandra; CHRISTIANSEN, Henrik L. A survey of the functional splits proposed for 5G mobile crosshaul networks. IEEE Communications Surveys Tutorials, 2018, vol. 21, no 1, p. 146-172.

[116] 3GPP TS 38.306, TSG RAN; NR; User Equipment (UE) radio access capabilities, Release 16, V16.1.0, July 2020.

[117] Kihero, Abuu B., Muhammad Sohaib J. Solaija, and Huseyin Arslan. "Multi-numerology multiplexing and inter-numerology interference analysis for 5G." arXiv preprint arXiv:1905.12748 (2019).

[118] Marijanović, Ljiljana, Stefan Schwarz, and Markus Rupp. "Multiplexing Services in 5G and Beyond: Optimal Resource Allocation Based on Mixed Numerology and Mini-Slots." IEEE Access 8 (2020): 209537-209555.

[119] 3GPP TS 38.214, 5G; NR; Physical layer procedures for data, version 16.2.0 Release 16, July 2020.

[120] Jalota, Devansh, et al. "Fisher Markets with Linear Constraints: Equilibrium Properties and Efficient Distributed Algorithms." arXiv preprint arXiv:2106.10412 (2021).

[121] 3GPP TS 38.306, 5G; NR, User Equipment (UE) radio access capabilities, version 17.0.0, Release 17, May 2022.

[122] J.B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. Econometrica: Journal of the Econometric Society, pages 520–534, 1965.

[123] https://es.mathworks.com/products/new_products/release2019a.html.

[124] Almeida, Gabriel M., et al., "Optimal joint functional split and network function placement in virtualized RAN with splittable flows", IEEE Wireless Communications Letters, 2022.

[125] Marotta, Andrea, et al., "Exploiting flexible functional split in converged software defined access networks", Journal of Optical Communications and Networking, 2019, vol. 11, no 11, p. 536-546.

[126] 3GPP TS 36.211, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, version 15.2.0, Release 15, October 2018.

[127] 3GPP TS 22.261, "5G; Service requirements for next generation new services and markets", Rel. 15, 03-2019.

[128] 3GPP TR 37.868, "Study on RAN Improvements for Machine Type Communications", Rel. 11, 2011.

[129] 3GPP TR 23.725, "Technical specification group services and system aspects: study on enhancement of ultra-reliable low-latency communication (urllc) support in the 5G core network (5GC)," Rel. 16, 2019.

[130] Garcia-Saavedra, A., Costa-Perez, X. (2021). O-ran: Disrupting the virtualized ran ecosystem. IEEE Communications Standards Magazine.

[131] O-RAN.WG4.CUS.0-v07.00, "O-RAN Fronthaul Working Group Control, User and Synchronization Plane Specification", Technical Report, July 2021.

[132] Yajima, A. U. T., Uchino, T., & Okuyama, S. (2019). Overview of O-RAN Fronthaul Specifications.