Universitat Politècnica de Catalunya

Optical Communications Group

Coordination of Smart B5G Radio Access and Autonomous Optical Transport Networks

Shaoxuan Wang

Advisor:

Dr. Marc Ruiz Ramírez

A thesis presented in partial fulfilment of the requirements for the degree of

Philosophy Doctor

July 2024

Optical Communications Group (GCO)

Universitat Politècnica de Catalunya (UPC)

C/ Jordi Girona, 1-3

Campus Nord, D4-213

08034 Barcelona, Spain

Acknowledgements

First and foremost, I am extremely grateful to my supervisor, Prof. Marc Ruiz for his indispensable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I would like to thank my teacher Prof. Luis Velasco's help during my Ph.D. study as well. At the same time, I would also thank my teacher Ms King Wang and my best friend and younger sister Emma Wang for their support over the years. Without your help, this doctoral thesis would not have been successful.

I would like to thank all my colleagues, Fatemeh, Sima, Mariano, Morteza, Masab, Diogo, Prasunika, Hailey, Sadegh, and Pol for the cherished time spent together in the GCO lab, and all my Chinese friends in UPC whose contribution in my life is unforgettable.

Finally, I would like to express my gratitude to my parents, my brother, my sister, and my friends in the church of Bezalle. Without their tremendous understanding and encouragement over the past few years, it would be impossible for me to complete my studies.

Abstract

Future radio access network (RAN) will operate with massive and heterogeneous small-cell deployments and end-to-end (e2e) connectivity in support of diverse beyond fifth-generation (B5G) use cases. More and more connectivity services are requiring not only stringent but also more predictable Quality of Service (QoS) performance, measured in terms of key performance indicators (KPI) such as throughput and capacity. With the advent of Open RAN (O-RAN), the implementation of flexible function splits/placement for guaranteeing target latency requirements and improved reliability is enabled. This smart operation must also precisely match capacity requirements, which typically reduces energy consumption, by managing the number of active base stations (BS) that are required to support user traffic requirements.

In addition to RAN, access and metro optical networks play a fundamental role to meet e2e requirements, in terms of both capacity and latency. Thus, optical transport networks can operate autonomously, e.g., to adapt optical capacity to current traffic. Nevertheless, the foreseen B5G scenarios poses challenges to autonomous optical network operation, since smart RAN operation generates highly variable and unpredictable traffic. Indeed, smart operation of both RAN and fixed network makes difficult to achieve optimal e2e connectivity performance if they are done independently. Instead, both domains can share knowledge and coordinate with the objective of guaranteeing strict QoS requirements and efficient resource utilization of e2e connectivity services.

This Ph.D. thesis is dedicated to developing solutions that coordinate both smart and autonomous operation of RAN and fixed optical network segments under B5G foreseen scenarios. To this aim, three goals are defined. The first goal aims at providing a methodology for smart operation of RAN cells with dense deployment of BSs, which is one of the most challenging scenarios envisioned for B5G networks. Relying on Open-RAN capabilities regarding monitoring and control loops, an AI-based approach that integrates both supervised and unsupervised machine learning algorithms to achieve intelligent RAN operation is proposed. The objective is to

minimize energy consumption by switching on/off BSs while providing the desired coverage and required capacity needs.

From the previous contributions and lessons learnt, the second goal focuses on analyzing the impact in terms of traffic to be supported by the underlying access and metro optical networks assuming smart RAN operation. The main conclusion of this goal is that smart RAN operation can have a critical affectation on underlying optical transport, which requires coordination between RAN and optical networks for efficient e2e network management.

In light of the above, the third goal tackles two different use cases where coordination between smart RAN operation and autonomous optical network management provide benefits and allow e2e QoS assurance. On the one hand, a procedure for which RAN configuration changes to be performed are anticipated to the fixed network controller is proposed. By means of contextual data, fixed access and metro traffic prediction models are extended with RAN context in order to predict ongoing sharp traffic changes. On the other hand, a second use case focuses in the scenario of serving particular services where a maximum e2e delay need to be assured. In particular, a dynamic coordination mechanism is proposed, where actual RAN delay is informed in case that this exceeds a given level, so that the fixed network controller can adapt its budget and take decisions according to the new constraint.

The research leading to these results has received funding from the Smart Networks and Services Joint Undertaking under the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101096120 (SEASON), and the MICINN IBON (PID2020-114135RB-I00) projects and from the ICREA Institution.

Resumen

La futura red de acceso de radio (RAN) funcionará con despliegues masivos y heterogéneos de celdas pequeñas y conectividad extremo a extremo (e2e) para soportar casos de uso diversos más allá de la quinta generación (B5G) . Cada vez hay más servicios de conectividad que requieren no sólo un estricto rendimiento de calidad de servicio (QoS) sino también más previsible, medido en términos de indicadores de rendimiento clave (KPI) como la latencia y la capacidad. Con la llegada de Open RAN (O-RAN), se habilita la implementación de división de funciones flexibles para garantizar los requisitos de latencia objetivo y una fiabilidad mejorada. Esta operación inteligente también debe ajustarse con precisión a los requisitos de capacidad, que normalmente reducen el consumo de energía, mediante la gestión del número de estaciones base (BS) activas que se requieren para apoyar los requisitos de tráfico de datos de los usuarios.

Además de la RAN, las redes ópticas de acceso y metro juegan un papel fundamental a la hora de cumplir los requisitos e2e, tanto en términos de capacidad como de latencia. Así, las redes ópticas de transporte pueden funcionar de forma autónoma, por ejemplo, para adaptar la capacidad óptica a la demanda real. Sin embargo, los escenarios B5G previstos plantean retos para el funcionamiento autónomo de la red óptica, ya que el funcionamiento inteligente de la RAN genera tráfico muy variable e impredecible. De hecho, el funcionamiento inteligente tanto de la RAN como de la red fija hace difícil conseguir un rendimiento óptimo de conectividad e2e si se realizan de forma independiente. Por el contrario, ambos dominios pueden compartir conocimiento y coordinarse con el objetivo de garantizar requisitos estrictos de QoS y una utilización eficiente de los recursos de los servicios de conectividad e2e.

Esta tesis doctoral se centra en desarrollar soluciones que coordinan la operación inteligente y autónoma de segmentos de red óptica fija y RAN bajo escenarios previstos B5G. Sobre este objetivo general, se definen tres objetivos específicos. El primero tiene como objetivo proporcionar una metodología para el funcionamiento

inteligente de las redes RAN con un despliegue denso de BS, que es uno de los escenarios más difíciles previstos para las redes B5G. Basándose en las capacidades de Open-RAN en lo que se refiere a los bucles de monitorización y control, se propone un enfoque basado en IA que integra algoritmos de aprendizaje automático supervisados y no supervisados para conseguir un funcionamiento inteligente de la RAN. El objetivo es minimizar el consumo de energía mediante el encendido/desactivación de las BS a la vez que se proporciona la cobertura deseada y las necesidades de capacidad requeridas.

A partir de las contribuciones anteriores y de las lecciones aprendidas, el segundo objetivo se centra en analizar el impacto en términos de tráfico de datos que debe ser soportado por las redes ópticas de acceso y metro subyacentes asumiendo el funcionamiento inteligente de la RAN. La principal conclusión de este objetivo es que el funcionamiento inteligente de la RAN puede tener una afectación crítica en el transporte óptico subyacente, que requiere de la coordinación entre la RAN y las redes ópticas para una gestión eficiente de la infraestructura e2e.

En vistas de lo anteriormente citado, el tercer objetivo aborda dos casos de uso diferentes en los que la coordinación entre el funcionamiento inteligente de la RAN y la gestión autónoma de la red óptica ofrece ventajas y permite una garantía de QoS. Por un lado, se propone un procedimiento por el que se anticipan los cambios de configuración de la RAN a realizar al controlador de red fija. Mediante datos contextuales, se amplían los modelos de predicción de tráfico de datos de la red de acceso y metro con el contexto RAN para poder así predecir los cambios bruscos en curso. Por otro lado, un segundo caso de uso se centra en el escenario de proveer servicios específicos donde debe asegurarse un retardo máximo e2e. En particular, se propone un mecanismo de coordinación dinámica, donde se informa del retardo real de la RAN en caso de que éste supere un determinado nivel, de modo que el controlador de red fija pueda adaptar su umbral y tomar decisiones según la nueva restricción.

Table of Contents

List of Figures

Page

List of Tables

Page

Chapter 1

Introduction

1.1 Motivation

Future radio access networks (RAN) will operate with massive and heterogeneous small-cell deployments and end-to-end (e2e) connectivity in support of diverse beyond fifth-generation (B5G) use cases. The optical transmission will play a fundamental role in meeting B5G requirements, in terms of capacity and latency. Specifically, with the disaggregation of the 5G RAN and the definition of different functional splits [La18], the requirements for the front-haul (F-H) become stringent [Pe18]. The optical network is being extended toward the edges of operators' networks [Ve13], fostered not only by the increased amount of traffic coming from current and future access segments but also by the stringent requirements that they need to support, like low latency, high reliability, and high bandwidth. In fact, more and more connectivity services are requiring not only stringent but also more predictable Quality of Service (QoS) performance, measured in terms of key performance indicators (KPI) such as throughput and capacity.

Addressing the previous challenges demands more innovative and efficient solutions than those currently employed in the 5G existing industries. With the advent of Open RAN (O-RAN), which includes the RAN intelligent controller (RIC) as a key component for smart RAN management, the implementation of flexible function splits/placement for guaranteeing target latency requirements and improved reliability is enabled. This smart operation must also precisely match capacity requirements, eliminating resource overprovisioning in the context of dense B5G scenarios, which typically reduces energy consumption. This can be achieved by managing the number of active base stations (BS) that are required to support the current user traffic requirements [Zh21] while keeping the required QoS.

In addition to RAN, access and metro optical networks play a fundamental role to meet e2e B5G requirements, in terms of both capacity and latency. Thus, optical transport networks can operate autonomously, e.g., to adapt optical capacity to current traffic [Ef22]. Typically, these works assumed predictable network traffic to behave according to legacy 4G scenarios, i.e., back-haul (B-H) traffic injected by base stations (BSs). Nevertheless, the foreseen B5G scenarios dramatically change such assumption, since smart RAN operation generates highly variable and unpredictable traffic that mixes F-H, mid-haul (M-H), and B-H traffic.

In light of the above, it is clear that the smart operation of both RAN and fixed networks makes it difficult (and even unfeasible) to achieve optimal e2e connectivity performance if they are done independently. Recent advances in network control and orchestration have proposed enhanced solutions for enabling a scalable and decentralized architecture by leveraging intelligence ubiquitously and securely across different technologies, network layers, and segments of the B5G e2e network [Ve21.1]. Supported by that, different segments can share knowledge and coordinate with the objective of guaranteeing strict QoS requirements and efficient resource utilization of e2e connectivity services.

Aiming at covering open issues on smart management of e2e connectivity services in B5G networks, this Ph.D. thesis is dedicated to developing solutions that coordinate both smart and autonomous operation of RAN and fixed optical network segments under B5G foreseen scenarios. The different goals that will be explained in detail in the next sections aim at providing methods and procedures to achieve smart operation of energy-efficient RAN and cost-efficient autonomous optical network management while ensuring e2e connectivity requirements by means of different strategies of coordination and knowledge sharing between RAN and fixed network domains.

1.2 Goals of the thesis

This thesis goes further with state-of-the-art approaches for smart management of RAN and optical networks and targets a more comprehensive, coordinated operation of both network domains. The objective is to guarantee that smart RAN operation targeting energy-efficient RAN management can be supported with desired QoS by underlying optical transport networks that, in turn, aim at optimizing the use of capacity resources in an autonomous way. To achieve that, coordination strategies are required to communicate smart RAN and autonomous fixed network operations, to avoid conflicting decisions and poor QoS performance.

The path towards achieving the overall thesis objective has been divided into three main goals:

G.1 –Smart RAN operation in dense B5G scenarios

This goal targets providing a methodology for the smart operation of RAN cells with dense deployment of BSs, which is one of the most challenging scenarios envisioned for B5G. In this regard, we rely on Open-RAN capabilities regarding RAN monitoring and control loops (supported by the RIC) to propose an AI-based approach that integrates both supervised and unsupervised machine learning (ML) algorithms to achieve intelligent RAN operation with the objective of minimizing energy consumption (by switching on/off BSs) while providing the desired coverage and required capacity needs. The results obtained utilizing real collected datasets and simulated data show that energy-efficient RAN management is successfully achieved.

G.2 – Impact of smart RAN operation on fixed optical networks

This goal starts from the lessons learned from the previous goal and assumes a scenario where a highly dynamic RAN operates and injects sharply variable traffic into the fixed transport network. In particular, this goal evaluates different configurations that can be triggered by smart RAN operation such as activation/deactivation of BSs and flexible functional split, and analyzes the impact in terms of traffic to be supported by the underlying access and metro optical networks. The main conclusion of this goal is that smart RAN operation can have a critical affectation on underlying optical transport, mainly because it might affect autonomous operation typically based on local monitoring and closed control loops. Therefore, this goal allows the conclusion that coordination between RAN and optical networks is required for efficient e2e network management.

G.3 - Coordination of RAN and fixed optical networks

In view of the previous goals, this goal presents two different use cases (separated into sub-goals) where coordination between smart RAN operation and autonomous optical network management provides benefits and allows e2e QoS assurance:

- **G.3.1 - Context-based e2e autonomous operation in B5G networks:** this goal presents a procedure for which the RIC anticipates RAN configuration changes to be performed to the fixed network controller in the form of contextual data. With this contextual information, that is shared before actual RAN changes are implemented, the autonomous operation is enhanced; in particular, fixed access and metro traffic prediction models are extended with RAN context variables in order to predict ongoing sharp traffic changes. Note that that prediction is used for dynamic optical capacity allocation which has a critical impact on performance in case of capacity under-provisioning. Results show the great benefits of implementing such context sharing for e2e autonomous operation purposes.
- **G.3.2 - Coordination of RAN and optical transport for delay guarantee:** this goal focuses on the scenario of serving particular services where a maximum e2e delay needs to be assured. By allocating and managing

RAN and fixed optical network resources conveniently, the e2e QoS requirement can be achieved. However, in particular situations such as RAN congestion, the delay budget to be guaranteed in the fixed network segment can be too large if we rely on a static value configured at the provisioning time. Instead, we propose a dynamic coordination mechanism, where the RIC informs about the actual RAN delay in case this exceeds a given level so that the fixed network controller can adapt its budget and take decisions according to the new constraint. Results verify that adding this coordination allows efficient optical capacity resources while keeping the desired e2e QoS requirement.

A summary of the goals of the thesis is presented in Table 1-1.

Table 1-1: Thesis goals

Goals	RAN	Optical Transport Network
G1 - Smart RAN operation in dense B5G scenarios	✓	
G2 - Impact of smart RAN operation on fixed optical networks		✓
G3 - Coordination of RAN and fixed optical networks	✓	✓

1.3 Methodology

The methodology used in this thesis is a crucial aspect of the study, as it outlines the approach taken to answer the research questions, collect and generate data, and provide and disseminate results, conclusions, and lessons learned. In this thesis, a mixed-methods approach was employed, combining both qualitative and quantitative data collection and analysis methods. The study began with a comprehensive literature review, which provided a foundation for the research questions and informed the development of the research design.

The data collection process involved the usage of available real datasets used for G.1, as well as the deployment and usage of a simulation environment developed in Matlab and Python, used partially or totally for all the goals. In fact, the simulation environment was used as a platform for the deployment of models and methods, and the obtention of performance analysis results.

Lastly, the results were disseminated in international conferences and journals, as well as included in deliverables of related research projects (see Chapter 9).

1.4 Thesis outline

The remainder of this Ph.D. thesis is organized as follows.

Chapter 2 briefly describes the background to easier understand the researching throughout the whole Ph.D. thesis. Meanwhile, background on the 5G RAN architecture, O-RAN architecture, and some specific introduction to ML is also provided.

Chapter 3 reviews the state-of-the-art related to the objectives of this Ph.D. thesis and points out the research directions supported by the proposed goals.

Chapter 4 is related to the description of the reference B5G scenario, as well as illustrates the main simulation and performance evaluation environment.

Chapter 5 is related to G.1 and presents the methods and results regarding Smart RAN operation in dense B5G scenarios. This chapter is based on and extends the conference paper [ICTON2024].

Chapter 6 is related to G.2 and presents the methods and results of the impact of smart RAN operation on fixed optical networks. This chapter is based on and extends the conference paper [ICTON2023].

Chapter 7 is related to G.3.1 and aims to provide the different methods and results of context-based e2e autonomous operation in B5G networks. It is based on the contribution presented in the journal paper [SENSORS24].

Chapter 8 is related to G.3.2 and presents the proposed approach for coordinating RAN and optical transport networks for delay guarantee. This chapter is partially based on the conference paper [ONDM2023].

Finally, Chapter 9 concludes this Ph.D. thesis.

1.5 Contributions and references from the Literature

For the sake of clarity and readability, references contributing to this Ph.D. thesis are labelled using the following criteria: [<conference/journal> <Year(yy) [. autonum]>], e.g., [ECOC20] or [JSAC21]; in case of more than one contribution with the same label, a sequence number is added.

The rest of the references to papers or books, both auto references not included in this Ph.D. thesis and other references from literature are labeled with the initials of the first author's surname together with its publication year, e.g., [Ve17].

Chapter 2

Background

In this chapter, we present the needed background of the core topics to be covered in this Ph.D. thesis. Most precisely, in section 2.1, we provide a brief background on 5G RAN and its developments in the last years. Section 2.2 describes the innovation of O-RAN (Open RAN) for B5G, and makes a comparison between 5G RAN and O-RAN. Section 2.3 makes a short introduction to optical networks and their main characteristics. Then, section 2.4 presents a summary of the ML methods and algorithms used in this work. Finally, section 2.5 concludes this chapter.

2.1 Radio Access Networks (RAN)

2.1.1 5G RAN Architecture

In this subsection, we provide a background about 5G RAN and how many solutions have evolved in the last few years and the function combination of different solutions is also investigated, and their use cases on 5G communication system performance.

3GPP has proposed a variety of deployment methods for introducing 5G RAN and has made a high-level division of 5G base stations, which has a great impact on the 5G RAN networking architecture. The 5G RAN has redefined the architecture of base stations, dividing them into two functional entities based on the Packet Data Convergence Protocol (PDCP)/ Radio Link Control (RLC) layer: the CU) and the DU). The CU undertakes the Radio Resource Control (RRC)/ PDCP layer functions, while the DU handles RLC/ Medium Access Control (MAC)/ physical layer (PHY) functions. A CU can accommodate multiple DUs. This redefined functional separation will influence the design of various layer functionalities in the protocol.

With the segregation of functionalities, 5G RAN introduces an F1 interface between CU and DU, standardized by 3GPP, defining its structure and message interactions.

The separation of CU/DU functionalities implies that future base stations will offer multiple deployment options for operators to flexibly choose from. CU and DU can be deployed in different locations or integrated into a single device, depending on diverse factors such as latency demands, front-haul and mid-haul network conditions, facility installation conditions, equipment costs and power consumption, reliability requirements, and CU coverage range needs. Figure 2-1. illustrate the 5G RAN architecture and the internal structure, which split into two parts called CU and DU as shown below, and these two entities are connected by a new interface called F1.

In general, there are four deployment scenarios:

Scenario 1: Independent deployment of CU, DU, and RU devices. DU is placed at the comprehensive business access point, while CU is positioned at the aggregation node. This scenario suits applications with relaxed latency requirements and specific demands for both front-haul and mid-haul conditions. CU can be implemented using general-purpose hardware. By forming a centralized network through CU, connecting multiple DUs is advantageous in achieving unified RRC management across multiple cells and resource pool aggregation. It can be deployed in edge cloud data centers in combination with edge computing-related functionalities.

Scenario 2: Independent deployment of CU while DU and RU are deployed in the same location or integrated into a single device. This scenario is suitable for applications with relaxed latency requirements, unrestricted mid-haul, and limited front-haul conditions. It is applicable in scenarios where PoP facility conditions are restricted or in small-site setups, sharing similar advantages with Scenario 1.

Scenario 3: Deployment of CU and DU in the same location or integration into a single device positioned at the comprehensive business access point. This scenario caters to applications with high-latency demands, unrestricted front-haul, and restricted mid-haul conditions, allowing for small-scale pooling.

Scenario 4: Deployment or integration of CU/DU/Active Antenna Unit (AAU) in the same location or within a single device. This scenario targets applications with high-latency demands and limited front-haul and mid-haul conditions. Due to size constraints, it is suitable for integrated small stations used for hotspot coverage.

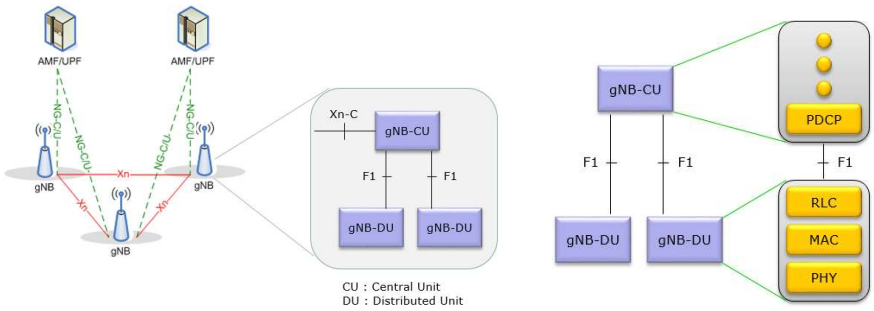

Figure 2-1: 5G RAN architecture [Te17]

2.1.2 O-RAN Architecture

In this subsection, we provide background about O-RAN and its architecture, the difference between 5G RAN and O-RAN is also investigated and their use cases on 5G communication system performance are also described.

The O-RAN alliance is committed to evolving 3GPP radio access networks based on principles of openness, cloudification, and intelligence. O-RAN networks can be built using multi-vendor, interoperable components and can be programmatically optimized through centralized abstraction layers and data-driven closed-loop control. For instance, the introduction of Non-Near-RT RIC and Near-RT RIC functionalities is aimed at advancing AI/ML capabilities from RAN management and operations to radio resource management [Ho23]. Research has been conducted on AI/ML workflows, deployment scenarios, and solutions for deploying AI/ML models, demonstrating their integration into O-RAN. In contrast to the work of International Telecommunication Union (ITU-T) and European Telecommunications Standards Institute (ETSI), O-RAN provides detailed procedures and protocols for Near-RT RIC, Non-Near-RT RIC, and related interfaces, serving as an open reference design to facilitate the identification of RAN intelligence use cases within 3GPP. It offers clear and detailed guidance for the industry's product development.

The O-RAN architecture, outlined in [Ra23], extends the 3GPP RAN standards to embrace openness and intelligence by integrating RAN splits, novel interfaces, RICs, and Service Management and Orchestration (SMO). The key components within O-RAN are denoted as the O-RAN Central Unit (O-CU), O-RAN Distributed Unit (O-DU), and O-RAN Radio Unit (O-RU).

For instance, Figure 2-2 illustrates the logical architecture. Within this setup, the O-RU represents a physical device. In contrast, the other components, such as gNB O-DU/gNB O-CU-CP/gNB O-CU-UP/O-eNB, can exist as physical appliances or as virtualized instances operating on an O-Cloud layer. The depicted green lines denote low-latency interfaces, while the purple lines represent management plane interfaces. Additionally, the black lines signify the interfaces defined by 3GPP.

The O-CU is divided into two segments: the control plane (O-CU-CP), responsible for RRC with Packet Data Convergence Protocol-Control Plane (PDCP-C) protocols, and the user plane (O-CU-UP), handling Service Data Adaptation Protocol (SDAP) with PDCP-User Plane (PDCP-U).

The O-DU encompasses RLC, MAC, and a high-physical layer, which includes the MAC scheduler. On the other hand, the O-RU covers low-physical layer functionalities such as Orthogonal Frequency Division Multiple Access (OFDMA) processing, beamforming, and Radio Frequency (RF) front end.

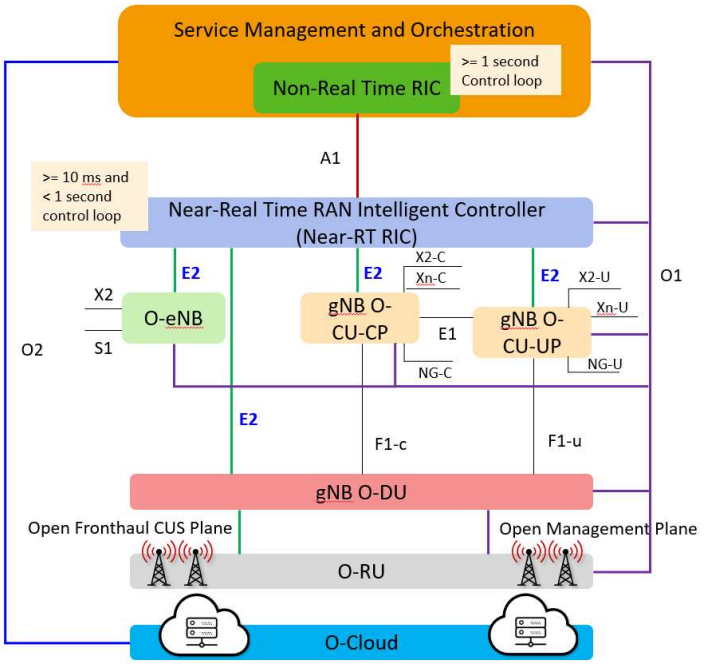

Figure 2-2: The O-RAN architecture [Ra23].

A fundamental addition within O-RAN is the RAN Intelligent Controller (RIC), a distinct entity separates from the processing units, providing access to Radio Resource Management (RRM) functions. The RIC is segregated into two components: The Non-Real-Time RIC (Non-RT RIC) and the Near-Real-Time RIC (Near-RT RIC). The Non-RT RIC operates on a timescale exceeding 1 second and is dedicated to non-real-time functions like radio resource management, optimizing higher layer procedures and policies in RAN. It facilitates the implementation of AI and ML workflows for RAN elements. Additionally, it furnishes policy-based guidance for Near-RT RIC applications and supplies Enrichment Information (EI) to support these applications. Conversely, the Near-RT RIC is an integral part of the RAN, specifically geared for controlling and optimizing algorithms pertinent to radio resource management. Functioning within a timescale longer than 10 milliseconds but shorter than 1 second, it operates in coordination with the control loop, employing use-case-specific applications termed xApps.

O-RAN outlines novel interfaces, including Open Fronthaul (OFH), which establishes the connection between O-DU and O-RU. Moreover, it introduces E2 and A1 interfaces serving as control loop connections, alongside O1, O2, and OFH M-plane interfaces dedicated to management functions. In the O-RAN architecture, O-CU-CP, O-CU-UP, and O-DU are termed "E2 Nodes" owing to their linkage through the E2 interface to the Near-RT RIC. This association enables their functionalities to be managed externally through applications such as the earlier-mentioned xApps.

Within these interfaces, the primarily significant components emphasized in O-RAN are E2 and A1:

- The E2 interface establishes a closed loop within the RAN sphere, facilitating the transmission of RIC control and policies to the E2 Nodes and receiving feedback from these nodes back to the Near-RT RIC

- The A1 interface delivers policies, EI, and ML models to the Near-RT RIC. Additionally, it retrieves policy feedback from the Near-RT RIC and routes it back to the Non-RT RIC.

The Near-RT RIC functions as a software platform, empowering xApps to manage the RAN. Supported by databases housing the network state for RAN and UE, it oversees xApp management, security measures, and conflict resolution. This platform facilitates nearly real-time control and optimization of E2 Nodes via action directives transmitted over the E2 interface, including CONTROL, INSERT, POLICY, and REPORT services [Ra23]. The comprehensive delineation of Near-Real-Time RIC is specified in [Ra22].

E2 Nodes, as previously mentioned, expose parameters and functionalities to RIC through the E2 interface, which xApps and rApps leverage to fine-tune the behavior of the radio network. Examples of xApps encompass mobility management, interference and beamforming control, traffic steering, load balancing, slice control, admission control, as well as signaling anomaly detection.

The RIC represents a software-defined element within the Open RAN structure. Its primary role involves overseeing and enhancing RAN functions. As a pivotal part of the Open RAN disaggregation strategy, the RIC introduces multi-vendor compatibility, intelligence, adaptability, and programmability to radio access networks. By facilitating the integration of third-party applications, it streamlines and enhances RAN operations on a larger scale, contributing to the reduced total cost of ownership (TCO) for mobile operators and improved quality of experience (QoE) for customers.

The RIC comprises two distinct segments: the non-real-time and near-real-time components. The non-RT RIC integrates into the centralized SMO Framework established by the O-RAN Alliance. Operating through specialized rApps, the non-RT RIC facilitates control over RAN elements and resources with a timeframe exceeding one second. It utilizes network insights, performance data, and subscriber information to furnish AI-driven suggestions for network enhancement and policy directives to xApps on the near-RT RIC. Additionally, it enables the running of external applications, known as rApps, aimed at delivering enhanced services for assisting RAN optimization and operations. These applications cover a range of functionalities such as policy guidance, enrichment information, configuration management, and data analytics. Meanwhile, the rApps are capable of providing control functionalities akin to xApps, such as traffic steering, scheduling control, and handover management, albeit over more extended periods. However, they have been standardized to create control policies that function at a higher level, exerting influence over a larger scope of users and network nodes. The near-RT RIC, situated within a telco edge or regional cloud infrastructure, executes network optimization

operations within a timespan ranging from 10 milliseconds to one second. For instance, near-RT RIC could adjust resource allocation based on network load, data traffic, and user demands, such as spectrum allocation and base station load management. Meanwhile, it also supports quick decision-making and operations to address network changes and sudden events, like dynamically adjusting transmission parameters to tackle network congestion. Tracks and optimizes user mobility, including seamless handovers, mobility management, and handover optimization, and improves service quality by instantly analyzing user data, boosting network performance, and enhancing user experience. Bellowing is the difference between 5G RAN and O-RAN:

	5G RAN	O-RAN
Openness	closed hardware and software solutions provided by specific vendors, lacking interoperability and flexibility	open interfaces and standards, hardware and software from different suppliers, enhancing scalability, and flexibility
Architecture	centralized architecture, consolidating functionalities within a single device	distributed architecture, decomposing functionalities into different functional entities (CU, DU, etc.), enabling more flexible network deployment and management.
Intelligence and Management	limited in intelligent control and management	O-RAN supports RIC, which leverages AI and ML technologies to optimize network performance and resource allocation
Cloudification and Virtualization	lack cloudification and virtualization capabilities, limiting dynamic resource allocation	O-RAN supports cloud-native architecture and virtualization technologies, allowing more flexible management and configuration of network resources

Overall, O-RAN, compared to conventional 5G RAN, is more open, and flexible, and supports advanced capabilities in intelligence, cloudification, and virtualization. This enables operators to construct and manage 5G networks more effectively.

2.1.3 RAN slicing

This section provides a short background on RAN slicing in B5G architecture. It is the basement of the 5G wireless communication systems technologies to which the G.3 is related. This section also describes the importance of RAN slicing in the 5G system.

Network slicing will be a fundamental feature of 5G networks. Network slicing is aimed at supporting several different logical networks on the same physical network infrastructure. Network slicing means cutting the physical network into multiple virtual end-to-end logical networks, and data of different requirements' accessing and transporting between a radio access network and core network are logically independent. Next Generation Mobile Networks Alliance (NGMN) 5G white paper has confirmed that the 5G slicing technology can be applied in the fields of car networking, medical treatment, and industrial manufacturing. The industry is exploring more potential use cases, including smart cities, drones, and another slicing network. Given the large variety of requirements on network functionalities (in terms of, e.g., security and mobility), performance (e.g., ultra-low-latency and ultra-reliability), and associated business models, it is envisaged that the support of network slicing will be one of the pillars for building the 5G ecosystem [Am16]. Network slicing aims to support several different logical networks on the same physical network infrastructure, and it will decrease the cost and energy consumption when compared to deploying separate physical for different use cases or business models [Si16], therefore, each network slice can be tailored to support specific applications and/or be operated by a communications provider other than the owner of the physical network infrastructure. Examples of communications providers, referred to herewith as tenants, are a mobile virtual network operator (MVNO) for the Mobile Broad-Band (MBB) consumer market or a vertical service provider [Sa16]. A network slice is composed of a collection of network functions and specific radio access technology settings that are combined for the specific use case or business model associated with a particular application and/or tenant [Ng15]. Such multi-tenant RAN are envisaged to be beneficial in localized areas with high user density (e.g., stadiums, malls, etc.) where dedicated small cell deployments per operator basis become impractical, therefore, the use of neutral host models, in which an infrastructure owner deploys several cells shared by multiple tenants, becomes an attractive solution.

In the RAN slicing part, efficient sharing of radio resources (i.e., spectrum) among multiple concurrent slices with diverging needs is a challenging problem. RAN slicing can rely on smart RRM functionality, such as spectrum planning, admission

control, etc., which can support the split of radio resources among different slices when considering the isolation requirements. There are so many algorithms and models that have been put forward to implement RAN slicing, such as service slicing mapping has been investigated in [Kh18], per group slicing, joint scheduling, and multi-objective problems considering latency or throughput have been studied in [Li12], and so on. According to this, 3GPP has settled some obstacles to network slicing, such as how to pair radio access network slicing and core network, and how the RAN slice chooses the core network slice have been solved. The main challenge facing the user side in the access network is that certain terminal devices (e.g., automobiles) need to access multiple slice networks at the same time. Authentication and user identification issues have also been researched by some enterprises and institutions.

Slicing a RAN and ensuring the necessary isolation between the different virtualized networks built on top of the same infrastructure becomes particularly challenging, due to the inherently shared nature of the radio channel and the potential influence that any transmitter may have on any receiver [Li15].In this context, RAN slicing can rely on smart RRM functionalities, such as packet scheduling, admission control, etc., which can support the split of radio resources among the different slices, taking into account the isolation requirements. In this respect, different works can be found in the literature, although mainly focusing on single-cell scenarios [Co13], while there are only a few works addressing slicing in multi-cell networks [Ma13]. Specifically, the work in [Sa17] analyses the RAN slicing problem in a multi-cell network with the RRM functionalities used to split the radio resources among the RAN slices and proposes four different RAN slicing approaches that are compared from different perspectives, such as the granularity in the assignment of radio resources and the degrees of isolation and customization. In turn, the work in [Pe17] proposes a multi-tenant admission control algorithm for performing the resource split among tenants with the target of ensuring efficient use of the radio resources by exploiting traffic multiplexing principles at both intra-cell and multi-cell levels to cope with heterogeneities in the spatial traffic distribution. The work in [Fe18] elaborates on a set of vendor-agnostic configuration descriptors that could be used to characterize the features, policies, and resources to be put in place across the radio protocol layers of a Next-Generation RAN node (called gNB) for the realization of multiple RAN slices over a shared cell using the New Radio (NR) interface.

The support of network slicing in 5G networks is a multi-faceted problem. A complete solution for network slicing combines multiple elements, ranging from virtualization techniques for the abstraction and sharing of radio resources (e.g., network virtualization substrate concept in [Pe13]) up to network slice lifecycle management solutions enabling the delivery of Network Slice as a Service (e.g., 5G network slice broker concept in [Sa16]). On the one hand, effective RAN slicing methods should be considered these days if we need to split radio resources from the RAN slice (i.e., four different slicing manners are introduced in [Sa17]), by doing this, different requirements will share the same slices, so they need an adaptation algorithm to

make it come true. On the other hand, dynamic management of slicing is an innovation in 5G when compared to 4G, it not only improves the spectrum utilization efficiency but also decreases the cost function, such as capital expenditure and operating fees. How to improve the profit of the slice is a vital question faced by operators. So, the considerations of integrating slice, finite radio resources' allocation, and the multi-objective optimization problem (MOOP) are emerging. The work in [Sh17] came up with a model to improve the flexibility and efficiency of 5G network slicing.

The RAN slicing of 5G could support more complex mechanisms for traffic differentiation when compared to conventional systems, and the RAN slicing is a mapping of slice-ID to a set of configuration rules. For instance, [Gu15] worked on the RAN configuration rules associated with each slice to accomplish the network services supported by the network slice, and the slice ID methods and applications are talked about in [Fe18]. At the same time, the mechanism for traffic differentiation should be able to treat different requirements differently. Therefore, to make efficient use of traffic flow in the RAN slicing, mechanisms to arrange different traffic flows through the radio resource and transmission interface need to be stood by, [Li17] has come up with two main methods to support those mechanisms in the current 3GPP system, the one is radio scheduler, and the other is different services. In their work, the radio scheduler will be focused.

The radio resource scheduler is the process through which eNB decides which UEs should be given resources to send or receive data. In LTE, scheduling is down each 1 ms Transmission Time Interval (TTI). A conventional resource fair scheduler gives the same fair share of resources to every user. In the RAN slicing part, the operators of the network will lay out the network slice model to adapt to different services. Compared to the conventional scheduler, the scheduler used in 5G RAN slicing is expected to share the radio resource dynamically with the different requirements of users. For instance, the packet scheduler is the radio resource scheduler with priority [Gu13]. The work is to set different weight parameters for each slice, and the scheduler arranges the radio resources according to the weight, the higher the weight, the higher the priority. And [Ko11] gives another slice scheduler induced by the NVS, which is an effective virtualized wireless resource in the RAN. The MAC scheduler is a media access control layer-slicing scheduler, which is established in the base station so that carriers can share resources according to a preset ratio. While operators and slice owners can set their resource ratios based on different service level agreements (SLAs) [Pa17]. So, scheduling the resources among different slices is a vital task in 5G RAN.

2.2 Optical communications

2.2.1 Optical transmission

We provide a background on optical communication systems in this section. It also illustrates the optical communication technologies on which G.3 is expected. This section describes the basic knowledge for setting optical communication scenarios for simulation goals used in this PhD thesis.

Since the 1980s, fiber optic communication technology has matured and been widely adopted, playing a crucial role in the expansion of information and the development of IP networks. Faced with increasing demands from users for higher communication network capacity, the optimal solution to meet high-capacity data transmission required to leverage the enormous potential bandwidth resource provided by fiber optics, approximately 30 THz. This approach allows for unimpeded transmission and exchange of information, driving the development of fiber optic communication at a pace that not only surpasses the growth rate of switches and routers constrained by Moore's Law but also exceeds the rate of growth in data services. Fiber optic communication has thus become the most important technology supporting the increase in communication service volume [Li02]. For example, industry applications primarily employ optical coherent detection technologies, polarization division multiplexing (PDM), digital signal processing (DSP) blocks, and optical amplification, particularly utilizing erbium-doped fiber amplifiers (EDFA) [EDFA02]. EDFA, along with additional optical network components such as optical splitters/couplers, wavelength blockers, and Wavelength Selective Switches (WSSs), enables the realization of transparent optical networks. In this context, there is no need for optical-electrical-optical (OEO) signal conversion along its path or light path, connecting the transmitter (Tx) and receiver (Rx).

There are three basic building blocks of the optical system: Transmitter, Fiber channel, and Receiver. The primary function of an optical transmitter is to convert electrical signals into optical form and launch the resulting optical signal into an optical fiber. It comprises an optical source, a modulator, and a channel coupler. Optical sources such as semiconductor lasers or light-emitting diodes are employed due to their compatibility with optical fiber communication channels. The optical signal is created by modulating the optical carrier wave. While an external modulator is occasionally utilized, in some cases, it can be omitted as the output of the semiconductor optical source can be directly modulated by varying the injection current. This approach streamlines transmitter design and is generally cost-effective. The coupler, typically a micro-lens, focuses the optical signal onto the entrance plane of an optical fiber with maximum efficiency.

The fiber channel also influences the carrier signal through different linear and non-linear impairments. The communication channel plays a crucial role in conveying

the optical signal from the transmitter to the receiver without introducing distortions. Optical fibers are widely adopted as the communication channel in most light-wave systems due to the minimal losses of about 0.2 dB/km in silica fibers, allowing for an optical power reduction of only 1% over a distance of 100 km. Fiber losses are a significant design consideration, in determining the spacing of repeaters or amplifiers in long-haul light-wave systems. Another critical design factor is fiber dispersion, which causes the broadening of individual optical pulses during propagation. If optical pulses extend beyond their designated bit slot, the transmitted signal experiences severe degradation, making it challenging to recover the original signal accurately. This issue is particularly pronounced in multimode fibers, where pulses spread rapidly (typically at a rate of approximately 10 ns/km) due to varying speeds associated with different fiber modes. Consequently, the majority of optical communication systems opt for single-mode fibers to mitigate these challenges.

Finally, the receiver transverses the optical signal into an electrical one by using coherent detection, at the same time, the signal is processed digitally to recover the data by relieving the linear and non-linear impairments. The optical receiver is responsible for converting the optical signal, received at the output end of the optical fiber, back into the original electrical signal. It comprises a coupler, a photodetector, and a demodulator. The coupler directs the received optical signal onto the photodetector, with semiconductor photodiodes commonly employed due to their compatibility with the overall system. The demodulator's design is contingent upon the modulation format utilized by the light-wave system.

In the majority of light-wave systems, the prevalent modulation scheme is known as "intensity modulation with direct detection" (IM/DD). In this setup, demodulation is accomplished by a decision circuit that discerns bits as either 1 or 0 based on the amplitude of the electric signal. The accuracy of the decision circuit is contingent upon the Signal-to-Noise Ratio (SNR) of the electrical signal generated at the photodetector.

Various modulation formats are employed in optical communication systems, with amplitude and phase modulation being the predominant techniques in current applications. Notably, there are five common modulation formats utilized in optical signal modulation: Binary Phase Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK), 8-level Quadrature Amplitude Modulation (8QAM), 16-level Quadrature Amplitude Modulation (16-QAM), and 4-level Pulse Amplitude Modulation (PAM4). Figure 2-3 illustrates the constellation diagram for these modulation formats.

In addition to modulation techniques aimed at enhancing the spectral efficiency of optical transmission systems, there is the option to multiplex two different polarization modes, known as Dual Polarization (DP) multiplexing. Furthermore, combining both modulation and polarization multiplexing proves effective in augmenting the spectral efficiency of optical networks.

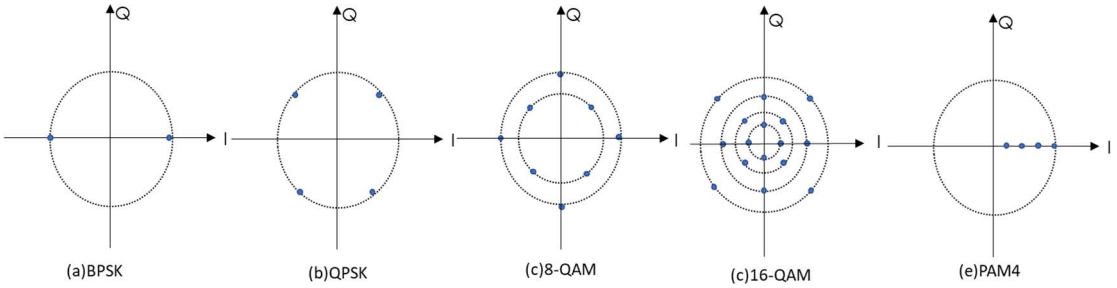

Figure 2-3: Constellation diagram of different modulation formats

From a geographical standpoint, optical networks can be categorized into three main segments (refer to Figure 2-4), primarily determined by their coverage areas: 1) access: This segment, located closest to end-users, spans various regions within cities and suburbs. It typically covers distances ranging from 20 to 100 km and serves numerous clients. 2) metro: Serving as a conduit between different cities, the metro segment is responsible for transmitting the aggregated data traffic originating from access networks. 3) backbone or core: Connecting metro networks within the same country or extending to a national level, this segment forms the backbone or core of the optical network infrastructure.

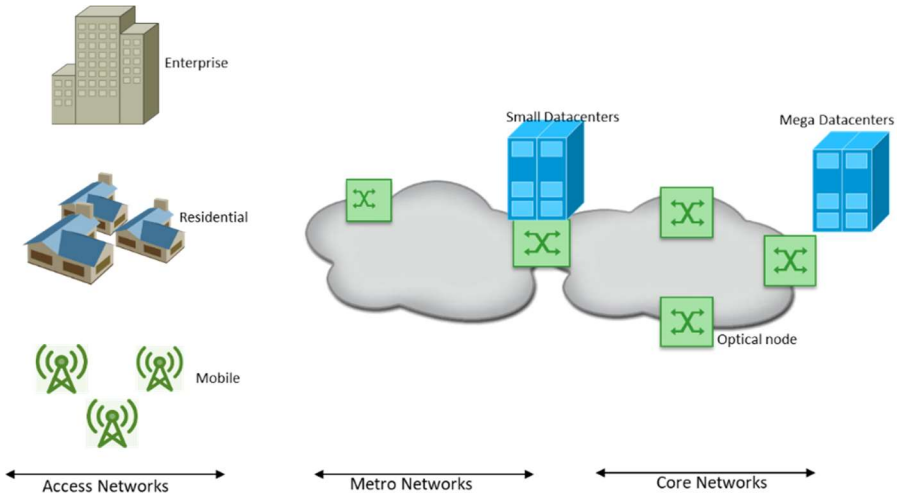

Figure 2-4: Heterogeneous Optical network architecture

2.2.2 Digital subcarrier multiplexing

Digital Subcarrier Multiplexing (DSCM) emerges as a pivotal technology in enhancing the overall capacity and adaptability of optical networks [Su20, We21, Ve21.2]. It serves as a crucial approach to augment the flexibility in the design and management of coherent optical networks, enabling independent propagation of multiple digital subcarriers through an optical channel. DSCM is implemented at the transmitter (Tx) end, where each Subcarrier (SC) is independently detected and

post-processed at the coherent receivers (Rx). SCs with varying modulation formats (MF), symbol rates (SR), and forward error correction (FEC) overheads can coexist.

DSCM plays a vital role in both P2P and P2MP transmission operations as well [Su20, We21]. For instance, P2P transmission technology utilizes an 8 DSCM signal, empowered by advanced parallelized Digital Signal Processing (DSP) blocks. On the other hand, P2MP facilitates finer granularity in optical network resource management, spanning from 25G to 400G, particularly influential in future metro or core optical networks leveraging multiple independent SC. The traffic data in such scenarios predominantly adheres to a hub-and-spoke pattern [Ho22]. In this Ph.D. thesis, P2MP is employed and modeled as a B5G scenario to address and eliminate the resource overprovisioning issue.

Subsequently, the mathematical formulation of DSCM systems is introduced, and the total DSCM signal is defined by [Ra17].

$$S_{DSCM}(t) = \sum_{i=1}^{N} S_i(t).e^{j2\pi f_i t} \tag{2-1}$$

Where N is the total number of SC and f_i is the frequency shift applied for a given general signal S_i to create a SC, and it is defined by:

$$f_{i=1\ldots N} = \left[(i-1) - \frac{N-1}{2}\right].\Delta f_{sc} \tag{2-2}$$

Where Δf_{sc} is the spectral width of a single SC, and it can be represented to:

$$\Delta f_{sc} = \frac{R_s}{N}(1+\beta) \tag{2-3}$$

Where R_s is the total symbol rate and β is the roll-off factor of the digital root-raise-cosine filter (RRCF) used to signal optical signal shaping.

Figure 2-5 shows a basic step for the step of creating a DSCM signal at an optical Tx. Firstly, a pseudo-random binary sequence is generated and mapped by QAM formats and shaped by the RRC filter. Then, Eq. (2-2) is used for the frequency shift computation, i.e., f_i. Finally, all the SCs at an optical multiplexer are combined and a DSCM signal is created, S_{DSCM} will be propagated through the optical fiber.

The utilization of DSCM in optical transport networks offers a significant advantage in maintaining elevated data rates, such as 400 Gb/s, while employing lower SR per Subcarrier (SC), typically around 8 or 11 GBaud. Concurrently, the flexibility inherent in DSCM systems can be leveraged to achieve substantial reductions in energy consumption.

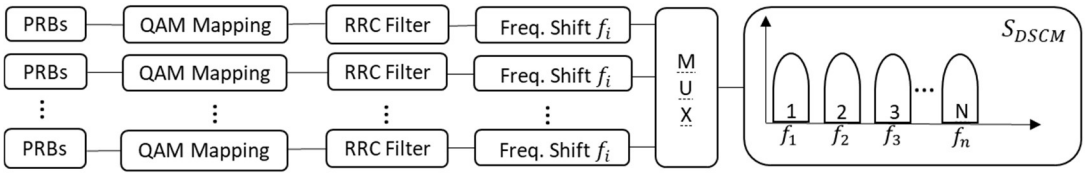

Figure 2-5: DSCM signal generation processing at the optical coherent Tx

2.3 Machine Learning (ML)

In this section, we focus on the background regarding the ML-based methods considered in this PhD thesis. Meanwhile, in subsection 2.3.1, we provide a short description of feature structure engineering based on the ML model. In subsection 2.3.2, the classification methods decision tree is explained.

2.3.1 Clustering

Recently, feature structure analysis for high-dimension big data has become necessary, it could extract useful feature information by using feature engineering to process the raw dataset. For instance, data clustering is a kind of feature engineering that organizes the raw data into clusters or groups with similar properties or characters.

Different from *k*-means, a two-step clustering model, SOM-K [Wa21] is a more precise clustering model, that describes each cluster in a dataset by evaluating the Euclidean distance between the inner cluster center and inter-cluster center. Since SOM simulation is complex and time-consuming, thus nearly accurate clustering results are possible and required in the initial assembly, after training is completed, the network makes each node of the output layer become a neuron, which is sensitive to a certain pattern class through the method of self-organization, and the corresponding internal weight vector of each node becomes the central vector of each input pattern class. This center vector can be used as a primary center vector in the *k*-method algorithm for performing accurate secondary aggregation. Finally, after applying the SOM-K model to a given dimension dataset, we can gain a better understanding of the feature structure analysis, and how we can perform this ML-based model. In this PhD thesis, we focus on the SOM-K to characterize and cluster cellular cells based on different cell parameters.

2.3.2 Classification

ML-based algorithms are considered an effective and strong tool, which is applied to optical communication as well as to turn them more intelligently [Ra18]. Decision trees are general-purpose machine learning algorithms that can perform complex

classification and regression tasks, even multi-output tasks, and are powerful enough to handle more complex data sets. It generally begins with a root node that divides into various potential decision nodes. Each of these decision nodes progresses to further end nodes, branching out into additional options, forming a structure reminiscent of a tree. There are three distinct kinds of nodes within this structure: nodes representing chance, nodes for making decisions, and end nodes. Figure. 2-6 illustrates an example of a single decision tree with three different represented nodes and an end node representing the classification results. In this PhD thesis, we focus on the application of a decision tree for cell classification and the BS switch on/off analysis. Additionally, we investigate a method of embedding a decision tree module in a non-real-time RIC. The idea of a decision tree for this use case will be described in Chapter 5.

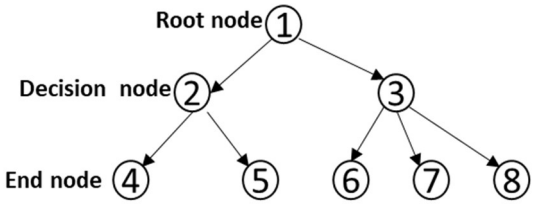

Figure 2-6: Single decision tree

2.3.3 Artificial Neural Network (ANN)

Machine learning-based models are considered powerful and interdisciplinary tools and have recently been applied to optical communications to make these systems more intelligent [Ra18]. Artificial Neural Networks (ANN) are information processing systems that mimic the biological behavior of neural networks, allowing them to learn the nonlinear interactions between network inputs and outputs. They are characterized by an input layer followed by a certain number of neurons, and based on their architecture, they can be classified into two main categories: feedforward and recurrent ANN. Feedforward ANN is defined by the absence of loops between layers or neurons, with inputs propagating forward through multiple hidden layers until reaching the output layer. Figure 2-7 illustrates an example of a feedforward ANN, with the respective layers and neurons represented by circles.

In this PhD thesis, we focus on applying feedforward ANN for traffic prediction in the access and metro fields of optical communication network modeling and analysis. Additionally, we explore the methodology of integrating context-aware mechanisms into the ANN prediction model to enhance prediction accuracy and reduce capacity overallocation in optical networks. A detailed illustration of the context-aware mechanism for the ANN prediction model will be provided in Chapter 7.

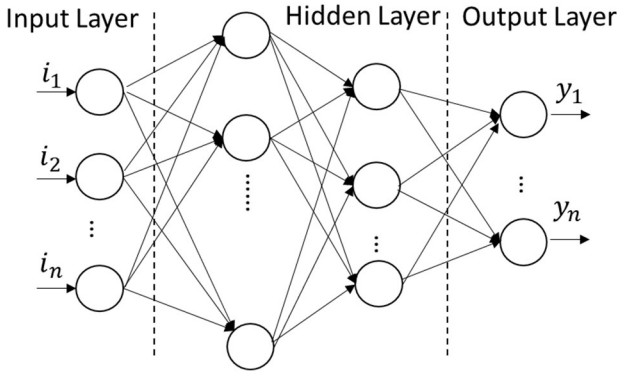

Figure 2-7 General structure of feedforward ANN

2.4 **Conclusions**

This section lays the foundation for the remainder of the Ph.D. thesis by introducing essential concepts and background about RAN, optical communications, and ML models. The next chapters start with this knowledge to present the different goals and achievements.

Chapter 3

State-of-the-Art

In this chapter, we present a review of the state-of-the-art different goals defined for this Ph.D. thesis with the twofold objective of ensuring that these goals have not yet been covered in the literature and serve as a starting point for this research work.

3.1 Smart RAN operation

Future RAN will extend current 5G technologies and will operate with massive and heterogeneous small-cell deployments in support of diverse beyond 5G (B5G) and 6th Generation (6G) use cases demanding high bandwidth and stringent latency requirements [Uu20]. To cope with such demand, RAN cells need to be planned with a high number of base stations (BS) per cell, which anticipates both large overprovisioning and energy consumption. To reduce both, the operational mode (active-sleep) of BSs can be dynamically managed as a function of current user equipment (UE) traffic requirements [Zh21]. For example, the authors of [Ch23] proposed a method that dynamically determines which small BSs to switch off based on the traffic load of small BSs, macro BSs, and high-altitude platform stations, and authors in [Zh15] proposed a distributed framework to improve energy efficiency and create a greener network through cooperation among base stations. Unlike authors of [Zh15] using a new framework, the authors in [Oh17], and [Oh13] proposed a dynamic BS sleeping strategy for shutting down the BS using network impact as a parameter to save network power for downlink/uplink transmission. In particular, authors in [Oh13] propose a distributed BS switching on/off scheme, which tries to sleep BSs one by one that minimally affects the network by introducing a network-impact parameter. Meanwhile, literature [Da14] proposed a hybrid traffic prediction model based on linear regression and ranked base stations according to coverage to

make the shutdown sequence more reasonable. Simulation results show that at least 14% of base stations can be shut down without affecting QoE. The authors in [Da22] present a traffic-driven cell zooming technique, where the coverage area of BSs can expand and contract as per the traffic volume. This is done by switching off BSs with low traffic and compensating for the coverage loss by expanding the neighboring BSs' coverage through increasing transmit power, but the bandwidth and the carrier frequency of MBS and μBSs in this work are the same. In addition, the works in [Gh20] and [Zh17] provide more details regarding Zoom technology and its integration with other elements in BS control and energy saving. Similarly, the authors in [Zh21] introduce a long-term short-term memory learning approach to predict the traffic distribution in the service area, by which they can determine when the BS sleeping operation is triggered. Finally, the authors in [Si23] propose analysis and classification of various handover techniques for providing recommendations for the selection of the most appropriate candidate BS for UEs. All the above work did not consider using both clustering and classification for BS switch on/off dynamically and precisely, the main objective, in line with the thesis goal G.1, is to find a suitable μBS to switch on/off based on the cells' features, while improving the radio resource efficiency and reducing the energy used in BS, consequently, reducing the cost.

3.2 Impact of RAN operation on fixed network

Several works in the literature have focused on implementing dynamic capacity management on fixed networks based on RAN operations.

Optical communication has experienced an innovative phase over the past decades, the optical networks should be able to adapt their capacity in response to the changes in current user traffic requirements across heterogeneous multi-application services (Multi-AS) networks is a challenging task. In this regard, the authors in [Ro20] proposed a specific split solution for an efficient F-H, which enables reducing the consumed bandwidth while being compliant with advanced cooperative radio technologies. Meanwhile, the authors in [Zh19] proposed a flexible functional split design to enable the dynamic functional configuration of each active remote radio head (RRH) of 5G RAN. The goal is to minimize the aggregate power consumption while considering limited F-H capacity, results showed that the F-H capacity constraint has a significant impact on aggregate power consumption. However, they did not consider dynamically optimizing the SCs allocation in the baseband unit pool according to using a flexible F-H function split. In their work, two traditional cooperative strategies in cloud RAN (C-RAN) that can maximize energy efficiency for downlink C-RAN are studied as a benchmark strategy in [Vu18]. However, thanks to RAN and 5G core virtualization, functional splits [La18] can be used to distribute the signal processing chain between a distributed unit (DU) and a centralized unit (CU) in, the RAN and the user plane function (UPF) in the core, which can be deployed at different sites of the network [Nd23]. The adoption of a

flexible function split is a promising solution that allows dynamic adaptation to different quality of service (QoS) requirements, which substantially improves RAN efficiency [Mo22].

In this PhD thesis, different from the literature, our target is to use different functional split combinations and CU/DU placement to match the requirements of every BS in a cell. A flow-based traffic model is presented aimed at formally quantifying the traffic contribution that each BS introduces to both access and metro segments according to the functional split. The study will focus on alternative solutions that could reduce optical capacity requirements and also improve the efficiency of capacity utilization.

3.3 Coordination for RAN and optical network

Recently, e2e autonomous operations for B5G networks have attracted much research interest among institutions and industries. For instance, the deployment of multilayer optical networks in access and metro segments plays a fundamental role in meeting the e2e requirements of slices [La23]. Several recent works have focused on 5G and B5G RAN provisioning supported by the resources (computation and connectivity) provided by underlying access and metro networks. For instance, authors in [Zo22] tackled the problem of DU/CU placement in access and metro networks for power consumption minimization subject to functional split, latency, and capacity requirements. Targeting more advanced B5G scenarios, authors in [Se23] considered functional split, traffic split, different placement options for virtual functions, and network slice-specific requirements in a joint provisioning problem. Moreover, the problem of combining DU/CU placement with connection provisioning in underlying optical networks was presented in [Wa22] and solved for different types of services. However, when autonomous fixed network operation deals with a mix of several slices from different tenants, as well as other fixed access flows, information sharing among the domains deserves dedicated consideration to avoid the revealing of internal domain details. Indeed, sharing aggregated data and/or models allows the preservation of privacy while keeping the value of transferred knowledge [Ru20]. Although those contributions present valid techno-economic and performance evaluation studies, they do not tackle the practical challenges of integrating dynamic and smart slice operation with fixed transport network operation [Mo23], since they focus from a planning perspective.

The main challenge for an autonomous transport network operation is to deal with highly variable traffic, which also becomes unpredictable as a consequence of smart slice operation. In a classical 4G scenario, the traffic injected by RAN cells to the fixed network typically fluctuates with smooth patterns highly correlated with UE demand [Be20]. However, depending on the functional split and DU/CU virtual function placement, B5G slices carry a mix of front-haul (F-H), mid-haul (M-H), and

back-haul (B-H) traffic that depends not only on UEs demand but also on slice operation. Thus, actions such as activating a new BS and changing the functional split or the placement of virtual functions, introduce large and sudden changes in the traffic of slices and consequently, in the underlying optical connections supporting them [Wa23].

Recently, some initiatives have explored the use of contextual information to improve the use of RAN resources in B5G scenarios by sharing data from UEs to the RAN control in an asynchronous and private way [Ko22]. Inspired by that concept, we extend the coordination between RAN and fixed networks in [Ba23] and the knowledge management in [Ra20] and present context-aware autonomous network operation that enables e2e smart operation by acompassing operation of slices and the fixed network in an effective and privacy-preserving way. The main objective of the proposed context-aware operation is to improve current autonomous fixed network operation approaches [Ve21.2] that fail to guarantee acceptable QoS assurance under extremely dynamic traffic originated by smart RAN operation. The key concept is the definition of context variables that are passed from the slice manager to the Software-Defined Networking (SDN) control performing dynamic capacity resource allocation in the fixed transport network. Context variables contain relevant information about the configuration of the slices in an aggregated way, so as to preserve the privacy of individual services and UEs. Moreover, context is updated asynchronously, e.g., before a significant slice reconfiguration is performed. Finally, the frequency and volume of data exchanged between domains is minimized.

Table 3-1: Literature review and contributions

Reference	**Smart RAN**	**Autonomous Optical Networks**	**5G/B5G RAN + Fixed Networks**	**Coord. RAN + Fixed Network**	**Context / Knowledge Sharing**
[Nd23, Mo23, Mo22]	X				
[Zo22, Se23, Wa22]			X		
[La23]	X		X		
[Ve21.2]		X			
[Ba23]		X		X	
[Ru20]		X			X
[Ko22]	X				X
Our work	**X**	**X**	**X**	**X**	**X**

3.4 Conclusions

In this chapter, we have reviewed the state-of-the-art of relevant works related to the goals of this thesis. Table 3-4 summarizes the literature review and contributions.

We can conclude that, although some previous works have worked on different aspects of energy saving and capacity estimation for B5G RAN and optical networks, a holistic approach for capacity coordination, energy saving, and e2e QoS assurance is needed.

Chapter 4

Preliminaries

This chapter focuses on the target B5G RAN architecture and presents the developed simulator as an accurate and efficient tool for reproducing network scenarios mixing RAN and fixed network technologies.

4.1 B5G Reference Architecture

In the B5G RAN scenario, we consider that a cell consists of a single macro BS (MBS) and several micro BSs (μBS). MBSs provide full coverage within their cells and provide the minimum capacity to absorb users' traffic, whereas μBSs complement the capacity of the MBS within a limited area of the cell. We assume that μBSs provide two operational modes: i) active, where the μBS is switched on and fully operational; and ii) sleep, where the μBS is switched off. Without loss of generality, we consider that radio units (RU) on both MBS and μBSs provide support for e2e traffic flows. RAN cells provide radio connectivity to UEs requiring one of the following main service classes [Su22]: i) enhanced Mobile Broad-Band (eMBB); ii) Ultra-Reliable Low Latency Communications (URLLC); and iii) massive Internet-of-Things (mIoT). It is worth mentioning that eMBB typically requires a large capacity (~150 Mb/s per UE and service) with relaxed e2e latency requirements (~4 ms from the UE to the core). On the opposite, the URLLC service has very stringent latency requirements (~ 1 ms) and reduced capacity. Finally, mIoT is typically highly distributed, which entails managing a large number of UEs injecting moderated bandwidth (in the order of tens of Mb/s) with intermediate target e2e latency assurance (~2 ms).

Figure 4-1(a) illustrates the 5G high-level reference architecture considered in this work, where the traffic generated by UEs in a cell sequentially traverses some functions, namely, RU, DU, and CU, until reaching the UPF serving as the breakout point of the 5G core [Ga21]. Thus, the resultant graph can be split into four different slice links, characterized by the RAN segment, i.e., radio (between UE and RU), F-H (between RU and DU), M-H (between DU and CU), and B-H (between CU and UPF). All these functions can be virtualized and run on the computing resources (servers, virtual machines, or containers) available at the different sites of the network. The B5G architecture is supported by resources in the fixed network infrastructure, for both connectivity, i.e., capacity and ensured latency. and computing. The e2e B5G reference topology assumed in this work is depicted in Figure 4-1(b), where the main network segments connecting sites and Central Offices (CO) are sketched. This topology is based on the reference high-level one from major European network operators presented in [Ru23]. Therefore, the traffic of a cell enters the fixed network. Specifically, an access optical network connects cell sites with their reference access CO (ACO). Typically, the distances between RAN cells and their ACO site are short, i.e., from a few to tens of km. Besides optical transport and switching capabilities, ACO sites are small data centers equipped with computing and storage resources that enable the deployment of virtualized DU/CU functions, as well as other UPF functions. Typically, ACOs aggregate traffic from various RAN cells in the proximity, as well as from other access technologies, such as residential gateways or customer edge premises. ACOs are interconnected among them and with regional COs (RCO) by metro-aggregation networks. RCOs are farther from UEs (around hundreds of km) and larger and more complex than ACOs and hence, they can host more virtualized functions and achieve higher efficiency. Finally, RCOs are interconnected with national COs (NCOs) by means of a meshed metro-core network, which provides large computational capabilities and serves as a gateway to other networks.

Figure 4-2 illustrates the overall architecture considered in this work, including the control and orchestration planes, which is an adapted version of the O-RAN architecture [Ra24]. The main entity responsible for RAN domain management is the RAN intelligent controller (RIC) which is in charge of a wide set of actions, such as QoS-based resource optimization, traffic steering, and RAN energy efficiency, just to mention a few. The RIC is divided into near-real-time RIC and non-real-time RIC. The near-real-time RIC controls RAN elements and their resources by means of local control loops that typically run in the range of 10 ms to 1 second; it receives policies from non-real-time RIC, running in the service management and orchestration system, that enables wide control loops requiring execution time above 1 second. For the sake of simplicity, hereafter we refer to simply RIC as the unified RAN control entity that combines near-real time and non-real-time operation. Specifically, we assume that the RIC deals with cell configuration, e.g., BS on/off switching, as well as manages DU/CU placement for each slice [Sa21.1]. The core network orchestrator is responsible for the core functions and specifically, we assume that it manages UPF

placement for each slice. A slice manager is in charge of making decisions about the configuration of each slice for service-level agreement assurance. Finally, in the transport network domain, the orchestrator coordinates actions with the SDN control plane. It is worth noting that the orchestrator layer provides O-Cloud functionality [Ra24], i.e., it manages the computing nodes running in each site, as well as the connectivity between sites [Ca18].

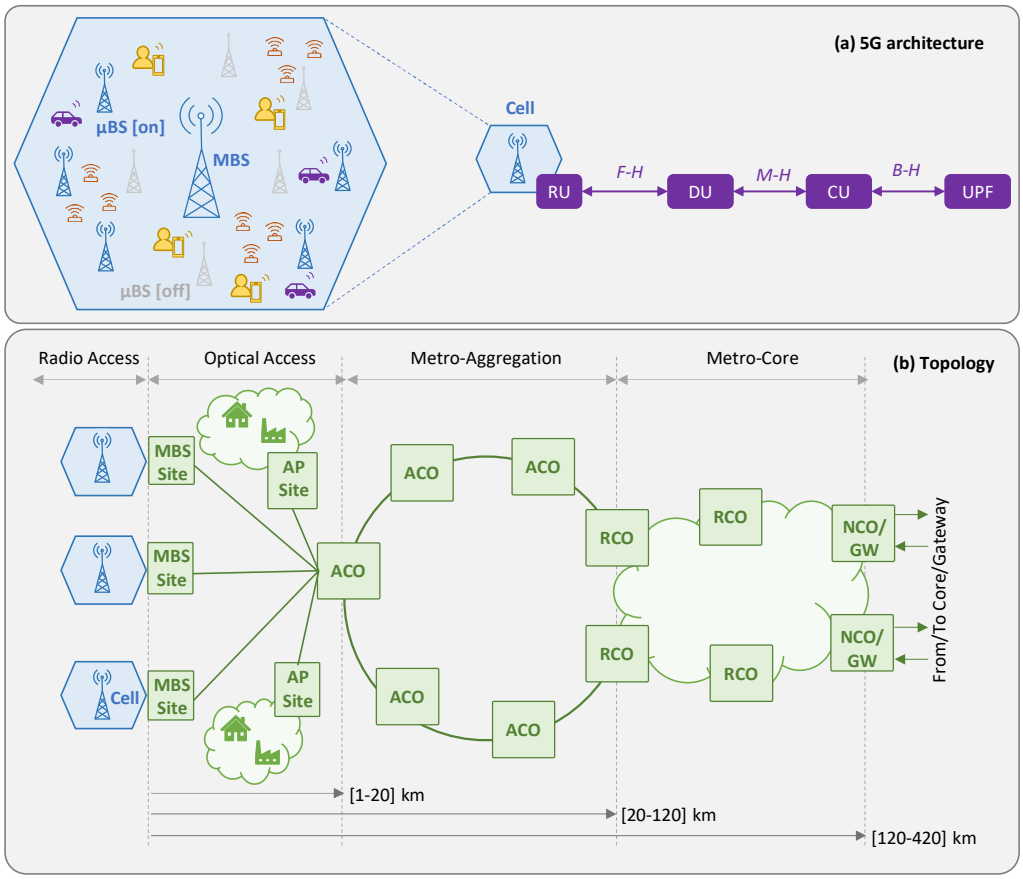

Figure 4-1: Reference 5G architecture (a) and topology (b)

Without loss of generality and in line with [Ru23], sites are equipped with optical transponders (TP) that allow connecting them to remote sites by establishing an optical connection. In addition, in this architecture, we assume DSCM TPs, which can allocate a variable number of sub-carriers to adapt the capacity to the traffic needs.

The mapping of slice links connecting functions onto optical connections depends on the slice configuration (capacity and placement of virtual functions) managed by the slice manager, which in turn, consumes resources (computing and connectivity). Note that the placement of the functions cannot be done in any potential location site due to constraints of each RAN segment, such as distance between sites and latency requirements [Et22]. Table 4-1 summarizes the mapping of virtual functions and site types, based on a typical network operator configuration [Ru23]. In the case

of DU and assuming split 7.2 for F-H, only MBS and ACOs are suitable for its deployment. However, M-H latency can be relaxed by means of split 2, which allows extending its placement to RCO if suitable, i.e., for eMBB services. Regarding UPF, without loss of generality, we assume that they consist of processes that require more intensive computation and centralization than those of DU/CU. Therefore, due to the very limited availability of resources at MBS, the placement of such functions is avoided at the very edge of the network. In addition, although function placement is allowed in ACOs, their computational resources are reserved for URLLC and mIoT services due to their limited capacity.

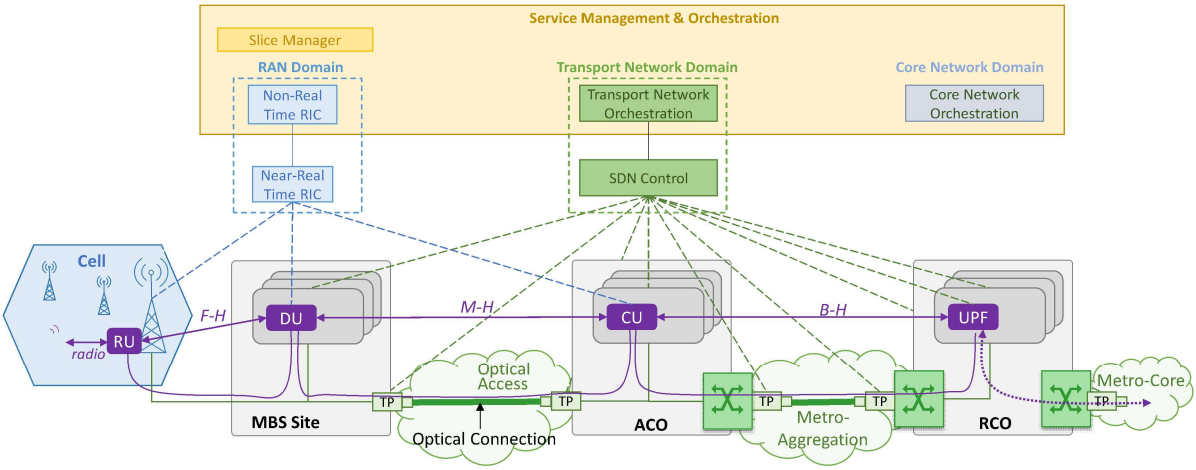

Figure 4-2: High-Level Architecture

Table 4-1: Virtualized function placement constraints

Function	MBS	ACO	RCO	NCO
DU	*Yes*	*Yes*	*Yes (eMBB)*	*No*
CU	*Yes*	*Yes*	*Yes*	*No*
UPF	*No*	*Yes (URLLC, mIoT)*	*Yes*	*Yes*

4.2 B5G Network Simulator

A Python-based simulator was implemented based on the one presented in [Be20], where a flow-based network simulator was proposed as an accurate and efficient tool for reproducing network scenarios mixing RAN and fixed network technologies. In a nutshell, the simulator generates flow traffic by means of statistical-based generators emulating RAN cells and fixed access points and propagates them through a system of fluid-flow continuous queues that model the different network

elements (packet and optical interfaces and connections) and segments (RAN cell, optical access, etc.).

Figure 4-3 shows the main blocks and components implemented in the simulator. It contains a data plane manager (Figure 4-3(a)) that simulates a topology consisting of several RAN cells, all of them individually connected to a reference ACO by means of access optical connections (i.e., each optical connection transports traffic from a single RAN cell). Then, the traffic received at the ACO (that aggregates several RAN cells and fixed access traffic) is propagated to the reference RCO by means of a metro optical connection. Without loss of generality, we assumed point-to-point optical connections, each supported by many sub-carriers of 25 Gb/s each [Ve21]. Collocated with this topology, a set of slices is served, each identified by the graph ρs that allows mapping each slice function with a computing node (we assume one single computing node per MBS and CO).

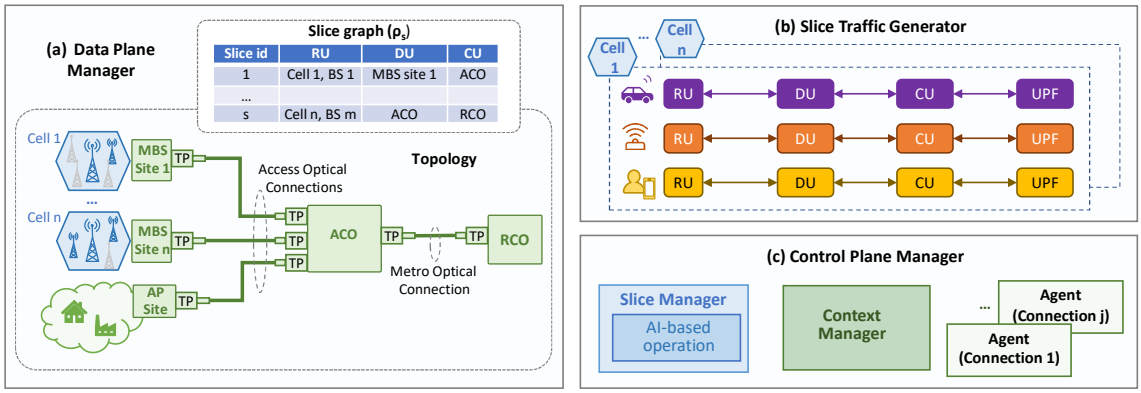

Figure 4-3: Simulator blocks and components

Available RAN traffic models in [Be20] are based on 4G technology and do not consider neither functional splitting nor slicing, i.e., traffic injected by RAN cells is always B-H. Since these models are not valid for the considered smart RAN operation, we developed a novel RAN *slice traffic generator* (Figure 4-3(b)) that, given a specific RAN cell configuration, generates synthetic slice traffic flows for each of the considered service classes (eMBB, URRLC, and mIoT) and each slice link (radio, F-H, M-H, and B-H). Note that this traffic can be easily mapped to each of the network segments according to the placement of DU and CU functions maintained in the data plane manager. The source code of the generator is openly available at [Bg24], where instructions to generate the main data used in this section are provided, as well as indications about how to reproduce other configurations and policies. Finally, the *control plane manager* (Figure 4-3(c)) implements the different modules that will be depicted in Figure 7-2 of Chapter 7 that relate to the context-aware operation, namely: *i*) the slice manager containing the AI-based operation module; *ii*) the context manager and related DBs at the fixed network orchestrator; and *iii*) a connection agent for each access and metro optical connection.

4.3 **Conclusions**

A complete introduction to B5G scenario architecture is illustrated in this chapter. This scenario includes two BS modes, one is sleep, and the other is active. RAN cells provide radio connectivity to UEs requiring one of the following main service classes: eMBB, URLLC, and mIoT services, respectively. Meanwhile, the B5G access part is divided into several modules (i.e., RU/DU/CU and UPF), with different services for UEs. The different combinations of modules are used to evaluate the required scenarios.

The B5G simulator is presented as an accurate and efficient tool for reproducing network scenarios mixing RAN and fixed network technologies. Moreover, the simulator creates network traffic flows using statistical generators that mimic RAN cells and stationary access points. These flows are then routed through a series of continuous fluid-flow queues, which represent various network components (such as packet and optical interfaces and connections) and segments (including RAN cells, optical access, etc.). This system effectively simulates the behavior and dynamics of network traffic through different elements of the network.

Chapter 5

Smart RAN operation in dense B5G scenarios

In this chapter, we discuss the methods defined in G1 for BSs management in B5G RAN dynamically in order to achieve remarkable energy consumption reduction, as well as maintaining committed QoS of end users.

5.1 Motivation

Future RAN will operate with massive and heterogeneous small-cell deployments and e2e connectivity in support of diverse B5G/6G use cases. At the same time, energy efficiency and consumption will be a major design criterion in 6G along with other metrics such as capacity, peak data rate, latency, and reliability. Thus, cost-effective networks require solutions providing high adaptivity that allow providing just the right capacity, thus eliminating overprovisioning and wasting. This requires near-real-time control that can be supported through closed control loops exploiting zero-touch and intent-based networking paradigms [Ve21.1]. In fact, a key operational objective in dense and heterogeneous RAN is to reduce energy consumption. This can be achieved by managing the number of active BSs that are required to support the current user traffic requirements [Zh21].

Smart-RAN is seen as a key enabler for realizing the full potential of B5G networks, offering the agility needed to support a wide range of applications and services while maintaining high levels of performance and efficiency. It is designed to meet the diverse and dynamic needs of modern telecommunications, enabling operators to

provide improved services while optimizing operational costs and energy consumption. Its implementation helps to address the growing complexity and demands placed on modern telecommunication networks. In addition, managing the operational mode (*active-sleep*) of BS as a function of current user traffic requirements reduces capacity overprovisioning and energy consumption [Zh21]. The key question is when and which BSs in the RAN should be activated/deactivated in order to achieve remarkable energy consumption reduction, as well as maintain the committed Quality of Service (QoS) of end users.

Among different next-generation RAN architectures, O-RAN promises great potential for flexible and dynamic RAN configuration based on monitoring data [Ra23]. Therefore, we also rely on O-RAN capabilities and propose an AI-based methodology that combines supervised and unsupervised ML procedures for smart RAN operation. In particular, different features that are continuously monitored in 6G RAN are processed in order to detect in which cell areas the QoS of UEs is worsening, thus triggering the activation of neighboring BSs. On the other hand, those BSs covering areas with high QoS are candidates to be deactivated for energy consumption minimization purposes.

5.2 Concept and Model

To carry out the studies needed to meet the goals of smart RAN, the working procedures and architecture of smart RAN illustrated in Figure 5-1 and Figure 5-2 will be followed. The model is based on the O-RAN architecture and it is mainly working on the Non-real time RIC module.

As a starting point, procedures in Figure 5-1 are conceived. Without loss of generality, we assume that a given RAN *area* contains a single macro BS (MBS) and several micro BSs (μBSs) providing coverage and connectivity to a number of *cells* (not depicted for the sake of clarity). MBSs provide full coverage within their cells and provide the minimum capacity to absorb users' traffic, whereas μBSs complement the capacity of the MBS within a limited area of the cell. We assume that μBSs provide two operational modes: (*i*) *active*, where the μBS is switched on and fully operational, and (*ii*) *sleep*, where the μBS is switched off. The RAN area serves a number of UEs that belong to a mix of services including enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive Internet of Things (mIoT) [Wa24]. a' also depicts an illustrative scenario happening at time *t*, where the current BSs status and UE demand are creating some saturation that is affecting the QoS of some services (represented by colored gauges). Note that few BSs are switched off, which enables the possibility to activate them and offload part of the traffic generating such congestion. However, it is necessary to know where the UE has worse performance and if it coincides, activate BSs

covering the cells where those affected UEs are. This will increase the overall energy consumption unless some active BSs that are covering cells with low UE demand and/or high QoS could be de-activated.

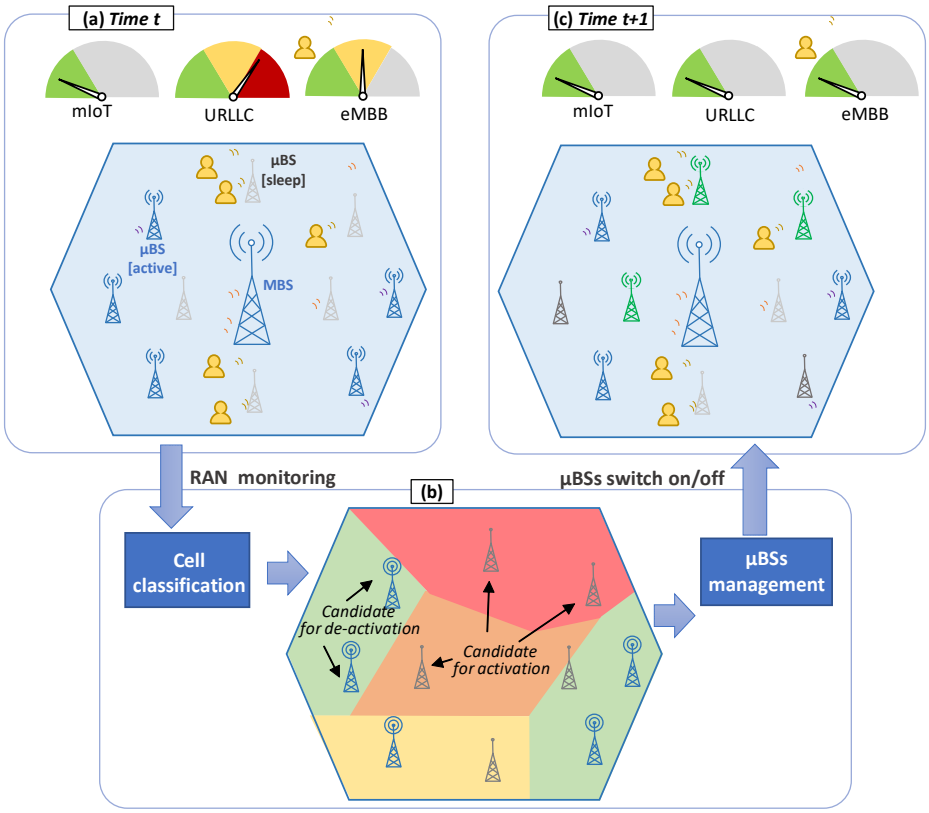

Figure 5-1: Smart RAN working procedure

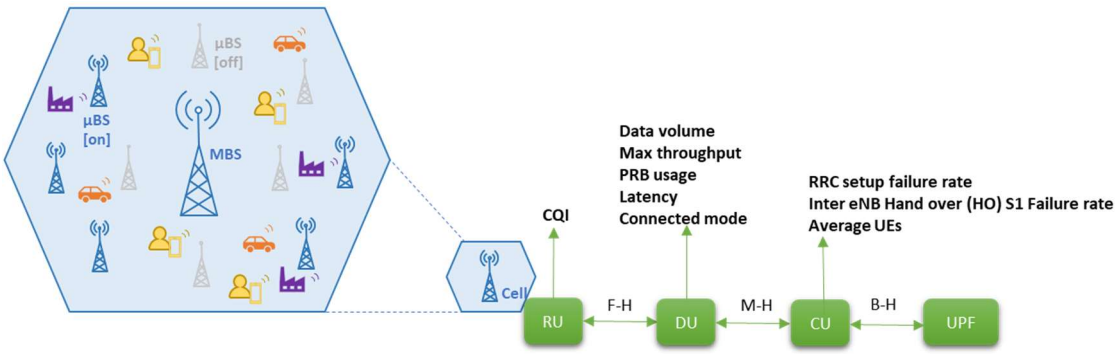

Figure 5-2: Feature structure distribution of the RAN

In view of the above, we propose the smart RAN operation sketched in b'. It consists of two main steps: *i*) *cell classification*, and *ii*) *µBS management*. The first procedure aims at processing RAN monitoring data obtained from different sources, including the radio units (RU), distributed units (DU), and centralized units (CU) of the different active BSs in the area [Wa24]. This monitoring data contains features such

as channel quality indicator (CQI), throughput, and latency, which are strongly correlated with UE QoS. Indeed, raw data is typically collected by UE and, since UE geo-location is also monitored, cell monitoring data can be easily obtained by aggregating UE monitoring data. This cell monitoring data is then used to classify cells into different categories. Step b' illustrates those categories by means of colors that highlight the need for activating or de-activating µBSs closer to such cells. In particular, the example shows that cells with bad QoS (orange/red) currently contain µBSs that are switched off, whereas cells with good QoS (green) contain an excess of active µBSs. Thus, this classification is processed by the second step, which applies a rule-based approach to make µBSs activation/de-activation decisions. Step c' illustrates the RAN status at time t+1 after applying the reconfiguration instructed by the smart RAN operation procedure. As a result of this reconfiguration, overall UEs QoS has improved with respect to previous time t.

Figure 5-2 is a design for inner of dynamic RAN based on the ML algorithm. First of all, the data is collected from the different BSs in the area and aggregated into an aggregator at Non-real time RIC. In order to simplify the cluster analysis results, 9 representative community data characteristics got from the SOM-K model as reference and we just consider the downlink situation, they are CQI, Data volume, Max throughput, Physical resource blocks (PRB) usage, latency, connected mode, Inter eNB Hand over (HO) S1 Failure rate, Average UEs and Radio resource control (RRC) setup failure rate. The RAN processing modules and topology diagrams that affect the above characteristics are presented in Figure 5-2, which means different features are all collected from the RAN part and the specific analysis of the features will be illustrated in section 5.3.1. Meanwhile, Table 5-1 summarizes the used notation.

Table 5-1: Notation

S	Set of µBSs
N	Number of neurons in SOM
$c_i(t)$	The cell belongs to class at time t
P_{opt}	Size of SOM topology
I	Number of cell samples
$W_j(t)$	Weight of j-th neuron at time t
A_t	Switch on/off actions for µBS
$x_i(t)$	The cell monitoring sample collected at time t in cell i
k^*	Minimizes the *Davies & Bouldin index* (DBI)
$z_s(t)$	The requirement of BS s at time t

5.3 Methodology and Use Cases

5.3.1 Cell monitoring

As mentioned in the previous section, the cell monitoring's function is to collect data from different modules among different BSs. For instance, RRC setup failure rate Inter eNB Hand over (HO) S1 Failure rate, and other statistics (i.e., Average UEs) related to network control are processed in the CU module, because CU is mainly responsible for high-level network control and management functions, such as session management, mobility management, including monitoring of user connection status and service quality, etc. The statistics and calculations of Data volume, Max throughput, PRB usage, latency calculation, and connected mode are usually performed in DU because these are directly related to the real-time processing of data transmission. Finally, RU is mainly responsible for measuring and reporting the quality of wireless channels, such as CQI information, but does not directly perform complex statistical analysis. All in all, all the feature information is collected in the cell monitoring, which is in the RAN function area.

5.3.2 Cell classification and μBS management

In this section, we describe the different models and algorithms involved in the smart RAN procedure sketched in Figure 5-1. The tackled scenario consists of a RAN area composed of a set I of cells that are covered by a set S of μBSs that can be switched on/off (recall that other BSs not in S are always active). The map Π contains the assignment of which cells can be covered from each μBS. Based on the reference O-RAN architecture [Ra23], we assume that the non-real-time RIC can collect and process cell monitoring samples. Thus, $x_i(t)$ denotes the cell monitoring sample collected at time t in cell i. Without loss of generality and according to [Wa21], each $x_i(t)$ sample contains an array of features that includes the following measurements:

- Number of average UEs
- Inter eNB Hand over (HO) S1 Failure rate
- RRC setup failure rate
- Data volume
- Max throughput
- PRB usage
- Latency
- Connected mode
- CQI

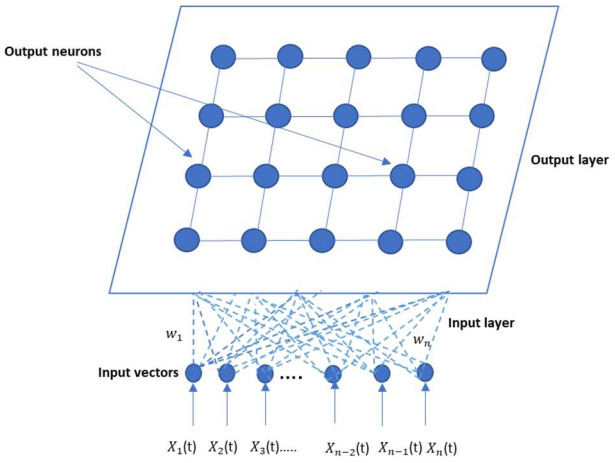

Figure 5-3: The topology of SOM

First, we implement the SOM algorithm, and the input data is clustered into the SOM for training. Since SOM simulation is complex and time-consuming, thus nearly accurate clustering results are possible and required in the initial assembly, and it is used for dimensionality reduction of the original input successfully [Wa21]. Second, after training is completed, the network makes each node of the output layer become a neuron, which is sensitive to a certain pattern class through the method of self-organization, and the corresponding internal weight vector of each node becomes the central vector of each input pattern class. This center vector can be used as a primary center vector in the *k* method algorithm for performing accurate secondary aggregation.

SOM is an unsupervised learning neural network model that can be used for clustering, high-dimensional reduction, and visualization [Li23]. The topology of the SOM neural network is illustrated in Figure 5-3. It consists of two layers, namely, the input layer and the competition layer. The working principle of SOM is projecting a collection or sequence of data input items from the input layer's neurons into the competition layer's neurons [Ul90]. Let us assume that a set of cell monitoring samples $X=\{x_i(t), \forall i \in I, t \in T\}$ collected during a large period T (e.g., a month) is available for training purposes. The first step in cell classification is to run the SOM procedure in order to identify clusters that will be used as classes in a supervised ML procedure. The competition layer of the SOM model is a rectangular or hexagonal grid of P neurons, each associated with a weight vector at time t is $W = \{W_j(t), \forall j \in N, t \in T\}$ where $j=1\ldots, N$. For the determination of the weight vectors, a training procedure is used, which typically implemented using the batch computation algorithm presented in [Ko13]. It is based on a competition approach in which input vectors are compared with weight vectors to select the winning neurons, referred to as the Best Matching Units (BMU) distance. During this process, when an input sample is fed to the SOM model, its Euclidean distance to all weight vectors is

computed and the weights of the BMU and the other neurons close to it in the SOM grid are adjusted to the input vector [La05]. The magnitude of the change decreases with many iterations and with the grid distance from the BMU [Ko07]. Let us denote P_{opt} to the size of the SOM model that delivers the best performance in terms of topographic error (TE)/ mean quantization error (QE), which are the main measurements to assess the quality of the SOM model [Ca19]. Formally, QE and TE are computed as follows:

$$\text{QE} = \frac{1}{R}\sum_{t=1}^{R}\left\|X(t) - W_{\text{q}}(t)\right\| \tag{5-1}$$

$$\text{TE} = \frac{1}{R}\sum_{t=1}^{R} d\big(X(t)\big) \tag{5-2}$$

Where $W_{\text{q}}(t)$ is the BMU's weight vector of the sample $X(t)$, $d\big(X(t)\big) = 1$, if the first BMU and the second BMU of $X(t)$ are not adjacent and $d(X(t)) = 0$ otherwise, and R is the number of iterations until network convergence. We train the SOM model using a different number of neurons ranging from 4 (i.e., 2x2 grid) up to 100 neurons (i.e.,10x10 grid). For each configuration, we compute the value of QE and TE, the one with the lowest QE and TE will be selected. Let us denote P_{opt} to the size of the SOM model that delivers the best performance in terms of TE/QE.

On the other hand, a k-means procedure uses the output of the SOM model as input to find the centroid of k clusters. In the next step, the k-means algorithm is used to cluster the P_{opt} weight vectors obtained by SOM into a smaller number of vectors $k \leq P_{\text{opt}}$ so that overall clustering accuracy is improved [Zh17]. The obtained k vectors are called centroids or cluster centers. To find the optimal number of clusters k*, we use the value that minimizes the Davies & Bouldin index (DBI) (eq. 5-3), which can be computed as the ratio of the sum of the centroid (x') intra-cluster distances (Δ) and inter-cluster distances (δ) (eq. 5-4). For specific evaluation criteria for DBI, please refer to our previous work [Wa21].

$$k^* = armgin_k DBI(k) \tag{5-3}$$

$$DBI(k) = \frac{1}{k}\sum_{i=1}^{k} max\, j \neq i \left\{\frac{\Delta\left(x_i'\right) + \Delta\left(x_j'\right)}{\delta\left(x_i', x_j'\right)}\right\} \tag{5-4}$$

Where x_i' and x_j' represent the i-th and j-th clusters within the set of k^* clusters. $\delta\left(x_i', x_j'\right)$ is the so-called inter-cluster distance between two generic clusters i and j.

As a result of the previous clustering, the new dataset Y= {<c, x'_c>$_i$, $\forall i \in I$ } is generated, containing for each cell i the id c and the centroid x'_c of the cluster the cell belongs to. This cluster id is the label to be predicted, taking as input the centroid

data. For the classifier, we use a decision tree (DT) that is trained to minimize the Gini impurity level, while keeping a moderated depth and a minimum number of elements in each leaf node [Po23]. The resultant DT f is then used to predict, given an input sample $x_i(t)$, the class $c_i(t)$ the cell belongs to at that specific time t, which will depend on the activity of UEs and the status of neighboring BSs.

Due to the good explain-ability properties of DT, an inspection of trained f is then performed in order to identify the set of cell requirements Q that each class holds. Specifically, we create the look-up table g that, for each class, assigns one of the following requirements, each identified with a numerical value q: *i*) the cell requires activating a µBS (q=+1), *ii*) the cell can admit switching off a µBS (q=-1), *iii*) no changes are required (q=0). Then, we define score $z_s(t)$ as the requirement of BS s at time t, computed as the average requirement of the cells that BS covers (eq 5-5). This score value needs to be eventually compared against a set of manually setup thresholds R to decide the main actions to perform in a BS, i.e., switch on, switch off, or keep the BS with its current status.

$$z_s(t) = \frac{1}{|\Pi(s)|} \sum_{i \in \Pi(s)} g(f(x_i(t))) \tag{5-5}$$

Finally, Algorithm 1 sketches the proposed smart RAN procedure that is run once the abovementioned models are trained and available for usage. It is called every time a new monitoring sample $X(t)=\{x_i(t), \forall i \in I\}$ is available and returns the set of actions $A(t)$ to perform, i.e., which µBS need to be switched on/off. After some initialization (line 1 of Algorithm 1), the numerical requirement of each cell is computed (lines 2-3). Then, for each µBS that can be dynamically managed, the score is computed according to the expression in eq. 5-5 (lines 4-8). The score value is then compared with thresholds R to find the potential action a (line 9). Finally, if and only if the action is different than keeping the current status, the action is stored and returned (lines 10-12).

5.3.3 Use cases

Due to B5G datasets being quite new and hard to get and collect, so we use a 4G/LTE high-dimension dataset, which is to test the cluster model. We try to obtain the cell patterns based on the long-term behavior (long-term behavior is the performance of cell patterns throughout the whole 15 days) of the cells as observed during the entire measurement collection period. According to average values of the entire measurement period were first computed for all the cells, and the input data onto 63-row vectors (one per cell) with 29-column vectors of each cell (one per feature) was selected. We represent input data as a 63x29 matrix. To determine the map size, the values of QE and TE are computed for configurations ranging from 2x2 neurons up

to 10x10 neurons. Obtained values are given in Table 5-2, as we can see, QE and TE get the minimum when the configuration reaches 6x6 (QE=0.296 and TE=0).

Algorithm 1 : Smart RAN procedure

INPUT: *X(t), f, g*

OUTPUT: *A(t)*

A(t)←∅
for $i \in I$ **do**
  $q_i \leftarrow g(f(X_i(t))$
for $s \in S$ **do**
  $z \leftarrow 0$
  for $i \in \Pi(s)$ **do**
    $z \leftarrow z + q_i$
  $z \leftarrow z / |\Pi(s)|$
  $a \leftarrow$ get_action (z, R)
  if $a \neq$ "keep" **and** $a \neq s$. status **then**
    $A(t) \leftarrow A(t) \cup <s.id, a>$
return *A(t)*

Table 5-2: Value of QE and TE

Map size	QE	TE
2*2	1.66	0
3*3	0.79	0
4*4	0.95	0
5*5	0.76	0
6*6	0.296	0
7*7	0.46	0
8*8	0.43	0
9*9	0.4	0
10*10	0.41	0.5

5.4 Results analysis

To evaluate the proposed approach, we used the real dataset in [Wa21] containing 1440 monitoring data samples of 63 cells collected during T=15 days in a major

European city. Every sample contains 29 different features with statistics of metrics related to the number of UEs, data volume, throughput, physical resource block (PRB) utilization, handover (HO) failure, radio resource control (RRC) failures, CQI, and latency. We start focusing on the evaluation of the two-phase clustering approach, whose performance is summarized in Figure 5-4. As we can find from Figure 5-4 (a), which shows the minimum value of DBI is 2.1 and its corresponding optimal cluster volume is 5, and we use different colors to represent these 5 clusters (Figure 5-4 (b)).

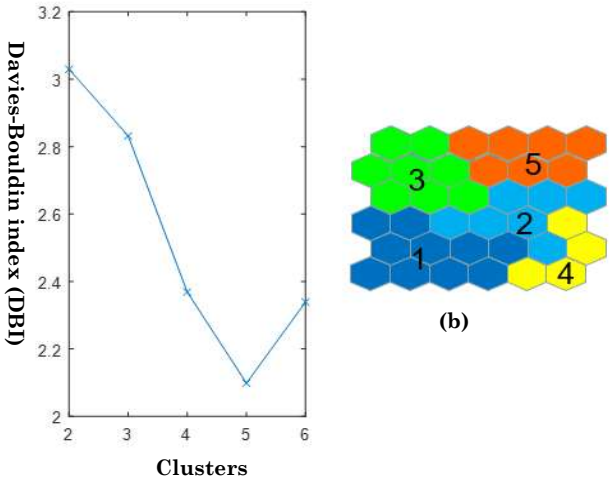

Figure 5-4: The minimum value of DBI and responding clusters

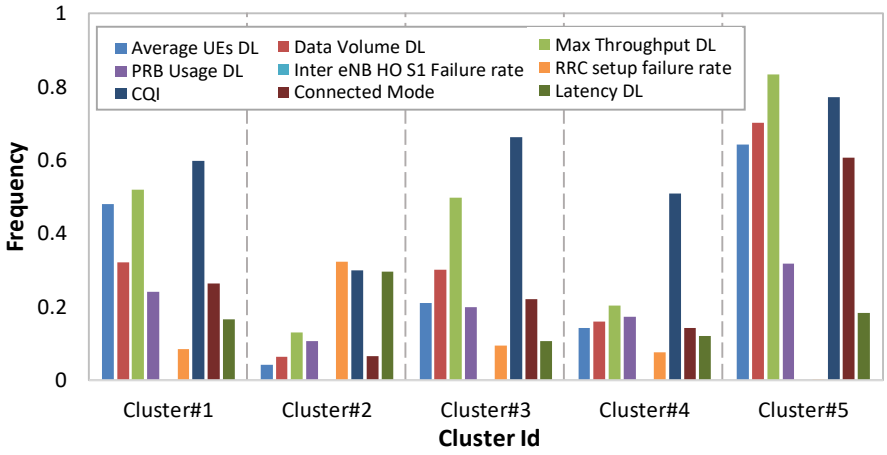

Figure 5-5: Features distribution of behavior cell patterns

We start focusing on the evaluation of the two-phase clustering approach, whose performance is summarized in Figure 5-5 and Table 5-3. First of all, it can be observed that SOM was allowed to reduce from 29 to just 9 relevant features (see legend for feature details). Using this reduced feature set, k-means clustering is executed with the DBI optimization, which results in an optimal number of clusters k*=5. The figure shows the distribution of features in each cluster, plotting the

relative frequency of each of them. In view of the graphs, it can be concluded that clusters identify cells with different feature patterns, which anticipates promising usefulness to characterize cells for RAN management purposes.

Table 5-3: Number of cells in different clusters by using SOM-K

Clusters	**Number of cells**	**Features summary**
1	16	Medium: Average UEs, Data volume, Throughput, RRC setup failure rate, PRB usage, CQI, Connected mode, and Latency Low: Inter eNB HO S1 failure rate
2	12	High: RRC setup failure rate and Latency Low: Average UEs, Data volume, Throughput, PRB usage, CQI, connected mode, and Inter eNB HO S1 failure rate
3	14	Medium: Average UEs, CQI, Data volume, Throughput, RRC setup failure rate, Connected mode, and PRB usage. Low: Latency and Inter eNB HO S1 failure rate
4	6	Medium: Connected mode and CQI Low: Average UEs, Data volume, Throughput, Inter eNB HO S1/RRC setup failure rate, PRB usage, Latency.
5	15	High: Average UEs, Data volume, Throughput, PRB usage, CQI, and Connected mode Medium: Latency Low: Inter eNB HO S1/RRC setup failure rate.

Taking the five clusters as true labels, a DT is trained for classification purposes. Figure 5-6 shows the obtained model f, indicating the Gini obtained at every node and the number of samples that belong to each class. Note that every leaf node has Gini=0, which means that perfect classification is done. In fact, the number of leaf nodes equals the number of classes, which allows remarkable significance of every decision rule. By inspecting the obtained tree, the lookup table g in Table 5-4 can be easily obtained. Note that class 1 comprises cells with low throughput and latency and high CQI, showing that sufficient QoS is provided, which opens the opportunity to switch off an active neighboring μBS (q=-1). Then, classes 2 and 3 represent situations where QoS is ensured but with no excess of resources, which leads to keeping the current status (q=0). Finally, classes 4 and 5 represent situations that show resource saturation and poor QoS, which lead to triggering BS activation (q=1).

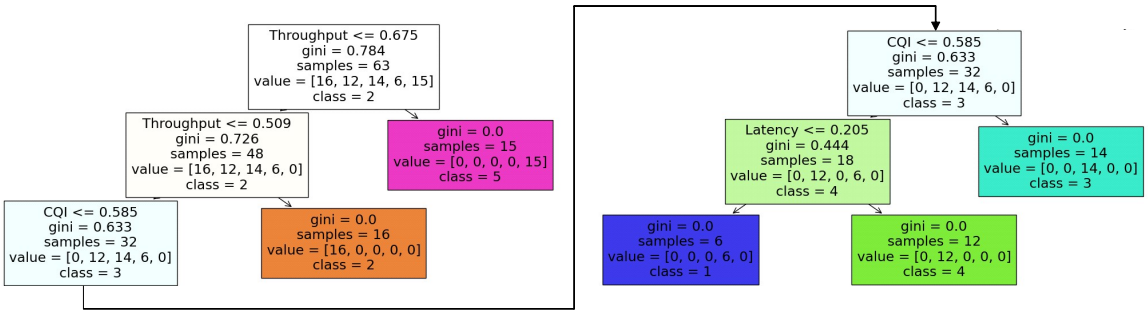

Figure 5-6 Classification results of decision tree

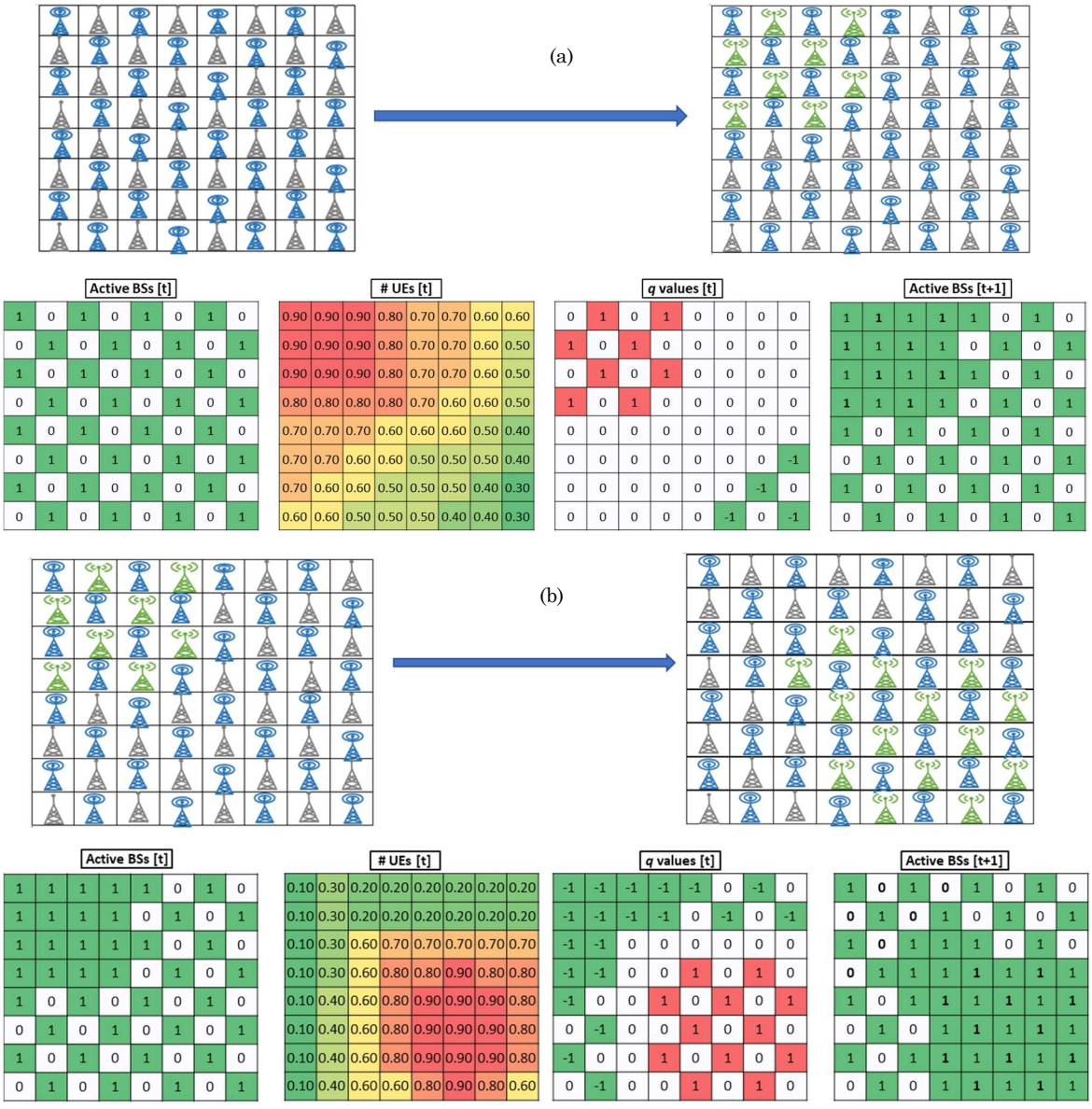

Figure 5-7: Illustrative results in a dense urban scenario

Once models *f* and *g* have been trained and validated with real data, we eventually show the application of Algorithm 1 in a simulated 6G scenario consisting of a dense urban area with 64 cells and one μBS per cell. Figure 5-7 shows two different cases and four plots for each case (from left to right): *i*) the active BSs at time *t*, *ii*) the normalized number of UEs per cell, *iii*) the *q* values of every cell after processing cell monitoring data, and *iv*) the active BSs at time *t*+1, after actions *A*(*t*) are implemented. In the first case (Figure 5-7 (a)), the initial BS map contains only those BSs that are always active (which are the blue BSs and represent 50% of μBS), and the other 50% of BS (the gray μBSs), which can be active or de-active based on the surroundings' requirements. Due to the concentration of UE demand in the upper left part of the map, some cells in that region require BS activation. As a result of this, 8 BSs that were in sleep mode at time *t* are switched on and become active at *t*+1 (i.e., the 8 green BSs in the upper left part).

In the second case (Figure 5-7 (b)), we start from the previous map and emulate the mobility of the UEs towards a different region in the map where they are now concentrated. This movement reduces the UE demand in the top left cells while increasing it in the bottom right ones. The *q* values catch that mobility, as well as the need to move BS capacity following the UE demand. Thus, 6 BSs in the top left region are switched off (i.e., from green BSs turn to gray BSs), while 12 BSs (i.e., the green BS) in the bottom right area are switched on. Note that the increasing RAN capacity will require coordination with fixed transport networks to properly adapt e2e connectivity capacity in support of new active BSs [Wa24].

Table 5-4: BS status evaluation by decision tree classifier model

Class	**Number of cells**	**Features summary**	**q**
1	6	Throughput<=0.509 CQI<=0.585, and Latency<=0.205	-1
2	16	0.509<Throughput<=0.675	0
3	14	Throughput<=0.509and CQI>0.585	0
4	12	Throughput<=0.509 CQI<=0.585, and Latency>0.205	+1
5	15	Throughput>=0.675	+1

5.5 Conclusion

A smart RAN operation procedure suitable for e2e autonomous network operation has been proposed. The overall approach is based on ML models and processes RAN

monitoring data to decide which BSs need to be switched on/off in order to reduce RAN energy consumption while guaranteeing UEs QoS. ML models have been trained and tested with real data and the complete procedure has been evaluated in a simulated B5G scenario.

Chapter 6

Impact of smart RAN operation on fixed optical networks

In this chapter, we focus on exploring and analyzing optical network traffic. The advent of 6G will revolutionize the way Radio Access Networks (RAN) will be operated. Expected massive small cell deployments and features, such as an adaptive functional splitting, are expected to change not only the volume but also the requirements of the traffic to be supported by the fixed transport network. This chapter presents an insight into 6G RAN operation, focusing on how such operation will impact the autonomous operation of the fixed network. As concluding remarks of such analysis, key requirements and challenges of fixed network operation for B5G/6G scenarios are identified.

6.1 Introduction

Future radio access networks (RAN) will operate with massive and heterogeneous small-cell deployments and end-to-end (e2e) connectivity in support of diverse B5G/6G use cases. The optical transmission will play a fundamental role to meet 6G requirements, in terms of capacity and latency. At the same time, energy efficiency and consumption will be a major design criterion in 6G, along with other metrics such as capacity, peak data rate, latency, and reliability in addition, cost-effective networks require solutions providing high adaptivity that allow providing just the right capacity, thus eliminating overprovisioning and wasting. This requires near-real-time control that can be supported through closed control loops exploiting zero-touch and intent-based networking paradigms [Ve21.1]. In fact, a key operational objective in dense and heterogeneous RAN is to reduce energy consumption. This

can be achieved by managing the number of active base stations (BS) that are required to support the current user traffic requirements [Zh21]. Note that such operations change the capacity requirements of the fronthaul, and therefore, the optical layer should be able to adapt its capacity in response.

At the same time, the use of functional splits [La18] allows distributing the processing of the 5G chain between a distributed unit (DU) and a centralized unit (CU), which can be deployed at different sites of the RAN and fixed network. With the disaggregation of the 5G RAN and the definition of different functional splits [La18], the requirements for the front-haul (F-H) become stringent [Pe18]. Recently, the adoption of adaptive function split is a promising solution that allows adapting dynamically to different quality of service (QoS) requirements, which substantially improves efficiency [Mo22]. In addition, managing the operational mode (active-sleep) of BS as a function of current user traffic requirements reduces capacity overprovisioning and energy consumption [Zh21]. Hence, by combining both dynamic RAN capacity management and adaptive functional split operation, the impact of smart 6G RAN operation on fixed networks can be investigated.

Access and metro optical networks play a fundamental role in meeting e2e 6G requirements, in terms of both capacity and latency. Similarly, to the smart RAN operation above, optical networks can operate autonomously, e.g., to adapt optical capacity to current traffic [Ef22]. Typically, these works assume fixed network traffic to behave according to legacy 4G scenarios, i.e., back-haul (B-H) traffic injected by BSs. Nevertheless, the foreseen B5G scenarios dramatically change this assumption, since smart 6G RAN operation generates highly variable and unpredictable traffic that mixes front-haul (F-H), mid-haul (M-H), and B-H traffic, which is also called an X-haul mechanism. This scenario fits very well with digital subcarrier multiplexing (DSCM) optical systems, due to: 1) point-to-multipoint (P2MP) connectivity can be easily implemented to connect several BSs to the access/metro network [Ve21.2], which results in cost-saving by reducing transponders (TP) count; and 2) their ability to activate independently each subcarrier (SC) in near real-time [Be20], which also reduces energy consumption. The problem here is that sharp traffic changes coming from the activation and deactivation of BSs in the RAN could lead to temporal bottlenecks that would produce high delay and even packet loss until the optical capacity is adapted.

In view of this, we analyze the impact of smart RAN operation on the traffic injected into the optical transport in 6G scenarios. To achieve this aim, a flow-based traffic model is presented aimed at formally quantifying the traffic contribution that each BS introduces to both access and metro segments according to the functional split and BS operational mode. X-haul traffic highly depends on both μBS management and the adopted B5G RAN functional split. RAN capacity is dynamically adjusted by switching on/off μBSs to serve user traffic.

6.2 Reference scenario under smart RAN operation

The e2e B5G/6G scenario assumed in this work is depicted in Figure 6-1a, where three different network segments, i.e., macro BS (MBS), access-metro, and metro-core, are sketched. First, a RAN cell consists of a single MBS, working at the sub-6 GHz band, and a number of micro BSs (μBS) configured in the mmWave band. MBSs provide full coverage within their RAN cells and provide a minimum capacity to absorb users' traffic, whereas μBSs complement the capacity of the MBS within a limited area. Without loss of generality, we consider that the radio units (RU) of both MBS and μBSs are endpoints of e2e traffic flows. Regarding μBSs, we assume that they provide two operational modes: active, where the μBS is switched on and fully operative, and sleep, where it is switched off. The traffic generated by the MBS (always active) and each active μBSs in a RAN cell is injected into the fixed network through an access optical network that connects cell sites with the reference access-metro site. Typically, the distance between both RAN cell and access-metro sites is short, i.e., from a few to tens of km. Besides optical transport and switching capabilities, access-metro sites are equipped with computing resources that enable the deployment of virtualized DU/CU functions. Finally, traffic injected by access-metro sites traverses the optical metro network segment and reaches the reference metro-core site, where the 5G core endpoint is deployed.

As introduced in Section 6-1, smart RAN operation is built upon two main pillars: i) dynamic capacity management by means of switching on/off the μBSs in a cell with the objective of reducing energy consumption while ensuring the minimum RAN capacity needed to support users' traffic; and ii) adaptive functional split operation, where both functional split and DU/CU placement are adapted to match the requirements of every BS in a cell. We consider the five functional split options (I to V) represented in Figure 6-1 (b), where the e2e latency is reduced, which is opposite to that where DU/CU computing cost (including energy consumption) is increased. Therefore, when user services demand strict latency requirements, higher options will be configured and DU/CU functions are placed close to users, which increases the cost by replicating DU/CU functions. On the contrary, when latency requirements can be relaxed, lower splits will reduce cost by concentrating functions at fewer sites.

Let us now illustrate the impact of using different functional splits on the traffic to be transported by access and metro segments. Specifically, two functional splits are considered, namely 7.2 and 2/4. According to [La18], under split 7.2, F-H bitrate is highly correlated with user traffic, as it mainly depends on the actual capacity used in the cell. However, the bitrate of split 2/4 depends on the actual configuration of MBSs and μBSs, so it sharply increases (decreases) when μBSs are switched on (off).

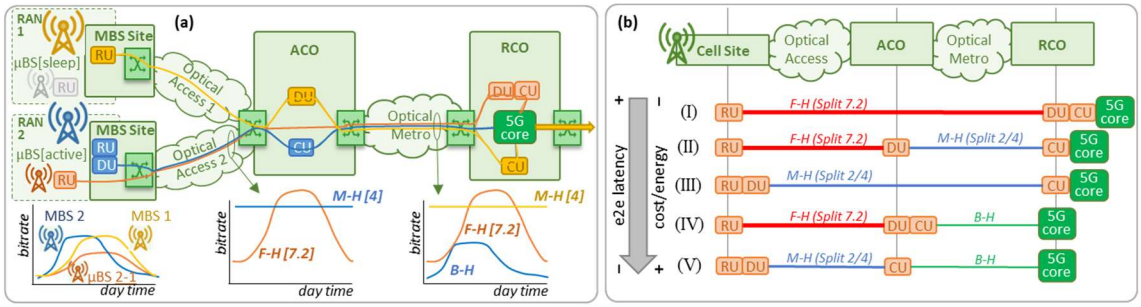

Figure 6-1: Reference B5G scenario (a) and considered options for functional split and DU/CU placement for flexible split (b)

In the example of Figure 6-1 (a), three active BSs are configured with different functional splits emulating different QoS needs: option II for MBS 1 (moderately low latency), option V for MBS 2 (ultra-low latency), and option I for µBS 2-1 (no strict latency requirements). Assuming a simple static RAN configuration (no changes in time), the three inset graphs (from left to right) show the input traffic in one day served by each BS, the traffic injected in the fixed access network, and the traffic injected in the metro network. It can be observed that different BSs generate completely different traffic patterns in access and metro. For instance, µBS 2-1 injects input-dependent F-H traffic in both access and metro domains (no intermediate function at the access-metro site), whereas MBS 2 injects constant M-H traffic into the access network and, after passing the CU function at the access-metro site, generates B-H traffic into the metro network.

On top of the above, the smart operation of B5G/6G RAN will manage both µBS operational mode and functional split as a function of the users' traffic. Therefore, a more complex traffic flow model than that for traditional 5G is needed to characterize traffic flows injected into the transport network. The next section is devoted to presenting such a model.

6.3 Fixed network traffic flow model for B5G scenario

The proposed model aims at characterizing, for every time t, the access and metro traffic flow components (variables y_{cat} and z_{amt}, respectively), as a function of input traffic at every BS (variables x_{bt}). We assume that a given RAN cell $c \in C$ connects to one single access site $a \in A$ and metro site $m \in M$ and, consequently, all BS $b \in B$ in cell c have the same reference access and metro sites. Independently of where DU and CU functions are placed, the traffic generated from BS b will traverse the fixed access segment until reaching reference access site a and then, will traverse the fixed metro segment from a to reference metro site m. Table 6-1 summarizes the used notation.

The traffic flow model is defined through the following equations. Eq. (6-1) models the traffic that a given BS b injects into the access network (y_{bt}). The value is zero if the BS is not active; otherwise, it can be F-H, M-H, or B-H depending on the placement of DU/CU functions. Similarly, Eq. (6-2) characterizes the traffic injected into the metro network (z_{bt}). Note that these two variables do not depend on the actual network configuration, e.g., where a given function is placed. The output is the expected F-H or M-H capacity for each cell i for the next period, z_i(t+1), which depends on the functional split and y_i(t+1) be the traffic monitored at time t. The generic model for split s, based on models in [Pe18], is defined in Eq. (6-3), and where η_{ij}^{s}(t+1) ∈ [0,1] is a factor that scales the component K_j^s that accounts for the F-H traffic that cell j injects at maximum load when split s is used. The scaling factors η for the considered splits in this work are in Eq. (6-4) and Eq. (6-5). At the same time, C_j is the maximum RAN capacity of cell j. From Eq. (6-4) and (6-5), we observe how F-H traffic in split 7.2 scales proportionally to user traffic clearly, whereas split 2/4 produces a constant F-H traffic per BS. In addition, although component K depends on multiple BS parameters, such as the number of antennas, layers, and chosen modulation format (see [Pe18] for further details), $K_{2/4}^{j} > K_{7.2}^{j}$ for any BS j. Eq. (6-6) and (6-7) compute the target access and metro traffic flow components, respectively. For a given pair cell-access site <c, a> and pair access-metro site <a, m>, y_{cat} and z_{amt} aggregate the components of every BS that is in cell c and have assigned the access site a and metro site m as reference ones.

$$y_{bt} = \rho_b \cdot \begin{cases} x_{bt}, & if\ cu_b == "cell" \\ mh_b, & if\ cu_b \neq "cell"\ \&\ du_b == "cell" \\ x_{bt} \cdot \dfrac{fh_b}{k_b}, & otherwise \end{cases} \tag{6-1}$$

$$z_{bt} = \rho_b \cdot \begin{cases} x_{bt} \cdot \dfrac{fh_b}{k_b}, & if\ du_b == "metro" \\ mh_b, & if\ du_b \neq "metro"\ \&\ cu_b == "metro" \\ x_{bt}, & otherwise \end{cases} \tag{6-2}$$

$$z_i(t+1) = \sum_{j=0..N} y_{ij}(t+1) \cdot \eta_{ij}^{s}(\mathrm{t}+1) \cdot K_j^s \tag{6-3}$$

$$fh_{b_{7.2}}(t+1) = \frac{x_{ij}(t+1)}{C_j} \tag{6-4}$$

$$\eta_{ij}^{2/4}(t+1) = 1 \tag{6-5}$$

$$y_{cat} = \sum_{b \in B} \delta_{bc} \cdot \delta_{ba} \cdot y_{bt} \tag{6-6}$$

$$z_{amt} = \sum_{b \in B} \delta_{bc} \cdot \delta_{bm} \cdot z_{bt} \tag{6-7}$$

Table 6-1: Parameters and variables

ρ_b	1, if BS b is active
k_b	Capacity of BS b [Gb/s]
du_b	Position of DU of BS b [site type]
cu_b	Position of CU of BS b [site type]
δ_{bc}	1, if BS b is in cell c
δ_{ba}	1, if BS b sends to access site a
δ_{bm}	1, if BS b sends to metro site m
fh_b	Max F-H traffic of BS b [Gb/s]
mh_b	Max M-H traffic of BS b [Gb/s]
x_{bt}	User traffic in BS b at time t [Gb/s]
y_{bt}	Access traffic by BS b at time t [Gb/s]
z_{bt}	Metro traffic by BS b at time t [Gb/s]
y_{cat}	Traffic in pair <c, a> at time t [Gb/s]
$\underline{z_{amt}}$	Traffic in pair <a, m> at time t [Gb/s]

6.4 **Illustrative results**

For numerical evaluation purposes, we have built a Python-based flow-based simulator that reproduces the e2e B5G scenario presented in Fig. 6-1a. To simplify the analysis of access (y_{cat}) and metro (z_{amt}) traffic components, we configured a scenario consisting of one dense RAN cell with 1 MBS and 64 µBS, one access site, and one metro site. We consider typical configurations for MBS (2x2 MIMO, 20 MHz bandwidth) and µBSs (8x8 MIMO, 100 MHz bandwidth). The maximum F-H and M-H for every BS (fh_b and mh_b) for all splits in Fig. 1b was computed from the models in [La21]. User traffic was generated following realistic daily patterns and scaled according to [Er22] to emulate a medium-term scenario with traffic peaks of 60 Gb/s for the whole cell.

With the configuration above, two different RAN operation policies were evaluated: i) static, where all µBS are always active and all BSs implement the same functional split option; ii) dynamic, where the split is still fixed but capacity is dynamically adapted by switching on/off µBSs according to actual traffic needs.

Figure 6-2 shows the performance under static RAN operation policy. Figure 6-2 (a) shows an example of one-day total user traffic and the total capacity, i.e., aggregating MBS and active µBSs. Note that the total capacity remains constant at 96 Gb/s and the traffic usage of UEs fluctuates with the time of day. Under this configuration,

Figure 6-2 (b) and Figure 6-2 (c) show the traffic at access and metro network, respectively, generated with every functional split option. We observe that this policy results in either predictable time-variant traffic (closely correlated with input traffic) or constant traffic, depending on the different functional split options. Additionally, it is worth noting that the traffic volume is dramatically affected by the chosen split, e.g., metro traffic of I and V follows a similar daily pattern but with a largely different magnitude. Moreover, access and metro traffic remain constant in options I and III, whereas they drastically vary for the rest of the options. In consequence, under key functional splits foreseen for B5G scenarios, metro traffic does not correspond to the aggregated access traffic, which is against the assumptions of typical traffic models and motivates the proposed ones clearly.

Regarding the dynamic policy (Figure 6-3), we have implemented µBSs switching on/off based on simple threshold-based criteria. Specifically, two rules were implemented at every µBS b: i) when the load (input traffic over capacity) of b exceeds 0.6, then the closest to b inactive µBS is switched on; ii) when the load of b drops below 0.3 and the closest neighboring µBS is active, b is switched off. These rules provided committed users QoS (no loss) during the whole simulation. Figure. 6-3 (a) shows that capacity savings up to 80% can be achieved when user traffic is minimal (i.e., only 11 out of 64 µBSs are active). In addition, we observe that those options that provided constant traffic in static operation are sensitive to RAN capacity changes, i.e., options III and V in access (Figure 6-3 (b)) and II and III in metro (Figure 6-3 (c)). In fact, traffic reduction in both segments is equivalent to RAN capacity reduction, which is an outstanding feature of those split options, e.g., to minimize optical capacity requirements and also reduce energy consumption. This comes at the cost of added unpredictability to access and metro traffic since constant periods are combined with varying periods, which hinders those widely used traffic forecast models based on short-term past window predictors.

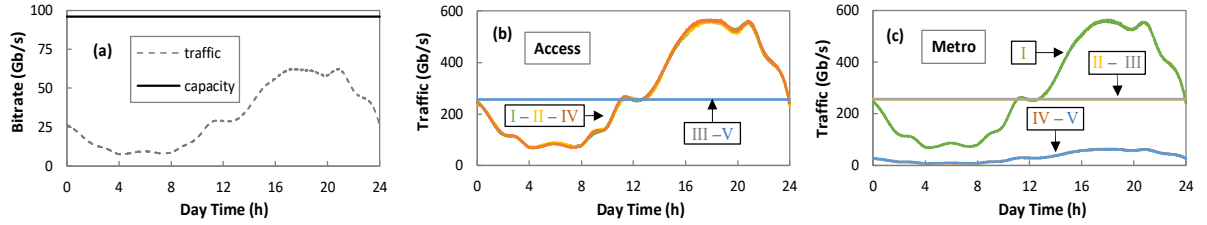

Figure 6-2: Static operation

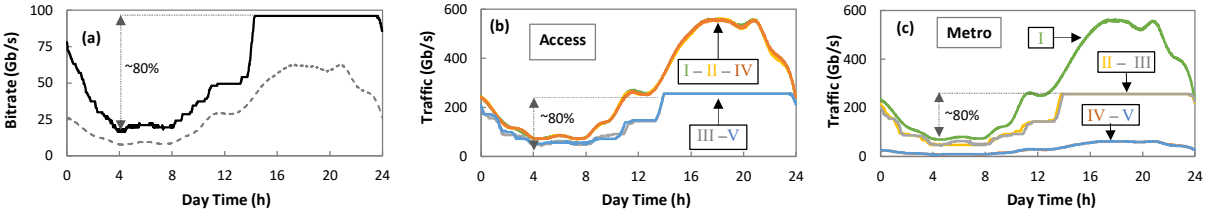

Figure 6-3: Dynamic operation

6.5 Conclusion

Smart RAN operation in B5G/6G scenarios will induce access and metro network smart operation to implement novel solutions to manage unprecedented variables and sharply changing traffic flows. Autonomous fixed network operation in tight coordination with RAN control is foreseen as a key challenge to achieve target e2e requirements. In this regard, we identify the need for RAN controllers to periodically collect user traffic monitoring data gathered from different cells and perform traffic prediction to estimate the expected traffic to be required for the next time interval. Note that this prediction is necessary for deciding which μBSs need to be powered on/off. Then, the RAN controller needs to be extended with additional modules to perform estimation of the traffic injected in the fixed network, which will depend on both users' demand and the functional split implemented, as well as on the RAN operation approach. That estimation is necessary to allow optical capacity setup, i.e., dynamic allocation of optical SCs based on traffic monitoring and capacity forecasting to be performed autonomously in the optical node agent. This optical capacity update needs to ensure both RAN traffic requirements and local capacity prediction forecasts.

Chapter 7

Context-based e2e Autonomous Operation in B5G Networks

The research and innovation of 5G networks carried out during the last years has settled the fundamentals of a smart slice in radio access networks (RAN), as well as autonomous fixed network operation. One of the most challenging objectives of beyond 5G (B5G) and 6th Generation (6G) networks is to deploy mechanisms for enabling smart end-to-end (e2e) network operation, which is required for achieving stringent service requirements of the envisioned use cases to be supported in the short-term. Therefore, smart actions, such as dynamic capacity allocation, flexible functional split, and dynamic slice management need to be performed in tight coordination with the autonomous capacity management of the fixed transport network infrastructure. Otherwise, the benefits of smart slice operation (i.e., cost and energy savings while ensuring per-slice service requirements) might be cancelled due to uncoordinated autonomous fixed network operation. Note that the transport network in charge of supporting slices from the user equipment (UE) to the core, expands across access and metro fixed networks. The required coordination needs to be done while keeping the privacy of the radio and fixed network domains, which is important in multi-tenant scenarios where both network segments are managed by different operators. In this chapter, we propose a novel approach that explores the concept of context-aware network operation, where the slice control anticipates aggregated and anonymized information of the expected slice operation to the fixed network orchestrator in an asynchronous way. Context is then used as input for artificial intelligence (AI)-based models used by the fixed network control for predictive capacity management of optical connections in support of RAN slices. Exhaustive numerical results show that slice context availability remarkably improves benchmarking fixed network predictive methods (90% reduction of maximum error prediction) under foreseen B5G scenarios, for both access and metro

segments, and under heterogeneous service demand scenarios. Moreover, context-aware network operation enables robust and efficient operation of optical networks in support of dense RAN cells (>32 base stations), where benchmarking methods fail for different operational objectives.

7.1 Introduction

Recently, open radio initiatives, such as O-RAN [Ra24], have allowed the deployment and management of slices to exploit the aforementioned dynamicity and adaptability capabilities, as well as the achievement of a virtualized, interoperable RAN among multiple vendors [Sa21.1]. Hence, smart slice operation can be achieved by combining dynamic RAN resource allocation and slice management with flexible functional split management [Oj23], which can be extended with advanced artificial intelligence (AI) capabilities for QoS assurance of stringent 6G services [Sa21.2]. Similarly, to smart slice operation, autonomous network operation is required to allow fixed (optical) networks to operate efficiently. Such operation paradigm is typically based on autonomous control loops, where monitoring data is continuously gathered and analyzed by means of AI models and algorithms that trigger actions to be performed in the network, e.g., to adapt optical capacity to current and expected traffic demand. This is particularly interesting when the digital sub-carrier multiplexing (DSCM) technology is used, as sub-carriers can be activated and deactivated in near real-time, which leads to operational benefits including energy savings [Ve21.2]. Note that such AI-based operation can be built on top of a distributed architecture that enables local control loops and efficient resource management without saturating centralized systems in charge of e.g., e2e service provisioning [Ve23].

In view of the above, it is clear that smart slice operation and autonomous fixed network operation must coordinate among themselves to achieve the required e2e performance. Note that per-slice coordination has been recently proposed, e.g., to guarantee e2e QoS target performance by exchanging service-related parameters [Ba23]. This approach is valid only for restricted scenarios where RAN operation is fixed and both RAN and fixed network domains belong to one single operator, so detailed information from UE activity can be shared from one domain to the other. However, when autonomous fixed network operation deals with a mix of several slices from different tenants, as well as other fixed access flows, information sharing among domains deserves dedicated consideration to avoid revealing internal domain details. Indeed, sharing aggregated data and/or models allows preserving privacy while keeping the value of transferred knowledge [Ru20]. In this chapter, we propose a novel approach to explore the concept of context-aware network operations, where slicing control predicts aggregated and anonymized information about expected slicing operations, which is sent asynchronously to a fixed network coordinator.

The rest of the chapter is as follows. Section 7.2 introduces the main concepts and actions of smart slice operation, highlighting the impact of slice management actions on the fixed transport operation and setting up the main challenges and requirements of contextual information sharing. Section 7.3 illustrates the work using contextual information sharing between slice management and fixed transport domains for autonomous e2e network operation coordinating and encompassing both RAN and fixed network autonomous operation. The section presents the main algorithms and models involved, which include asynchronous context updates and context-aware AI-based capacity reconfiguration modules. Section 7.4 presents numerical results to illustrate the benefits of the proposed context-aware autonomous network operation compared to benchmarking approaches, where slice management and the fixed network operate independently. Finally, Section 7.5 concludes the chapter.

7.2 B5G RAN and Slice Operation

As introduced in the Introduction, smart slice operation is built upon three main pillars: i) dynamic μBSs management, by switching on/off μBSs with the objective of reducing energy consumption in the RAN, while ensuring the minimum capacity needed to support UE traffic; ii) dynamic RAN capacity slicing, with the aim of managing physical radio blocks (PRBs) to assign resources to each of the different slices in order to provide the required QoS; and iii) flexible functional split operation, where the placement of virtual functions (DU/CU) is adapted to match the requirements of the UEs served by each BS in a cell. In this section, we aim to illustrate how that smart slice operation dramatically affects the traffic supported by the underlying transport network in each of the segments of the reference topology.

Figure 7-1 (a) shows the RAN state at a given time t_a of an example consisting of one cell with one active MBS that provides connectivity to a mix of UEs from different services. For the sake of simplicity, we assume that one slice per service type is deployed. Let us assume that the core network orchestrator decides, at slice provisioning time, the placement of UPF according to slice type and QoS requirements. This UPF placement remains fixed during the slice lifetime. Moreover, each slice has its placement of DU/CU along the different CO sites (see Figure 7-1(b)). In this case, DU/CU placement can be dynamically reconfigured by the slice manager according to current and expected UE traffic conditions in order to guarantee that e2e latency (i.e., from UE to UPF) meets the requirements of the slice service type. The figure also shows a simplified view of the PRBs used by each of the slices. Table 7-1 shows the RAN segment per service class that is transported in each of the network segments. Note that the traffic in each segment is a heterogeneous mix of F-H, M-H, and B-H traffic, depending on the slice configuration. The selected time instant ta illustrates a scenario where radio

resources are reaching a point of saturation that is negatively affecting services e2e QoS (represented by colored gauges). In particular, URLLC service is strongly affected by such saturation, even when DU/CU functions are currently placed as close as possible to UEs to reduce e2e latency. In view of this, let us imagine that such QoS degradation is detected by the slice manager and, after analysis, some slice reconfiguration has been identified, which entails several actions to be performed. First, the slice manager triggers the activation of an available µBS (in light grey in Figure 7-1(a)). Due to the physical location of the antenna and its proximity to the majority of URLLC UEs, activating such a new antenna will relieve the MBS from serving most of URLLC traffic. Figure 7-1(b) shows the RAN state after activating such µBS at time instant tb. Since the activated µBS (now in green) captures most of the URLLC traffic, the overall RAN load is reduced and, consequently, the delay introduced by the RAN segment [Mo23], which in turn makes the e2e QoS of all services reach the desired target performance.

Nonetheless, smart slice operation goes beyond µBS activation. For instance, in order to reduce the cost associated with virtual function placement, URLLC and mIoT functions can be now located far from the edge (where available resources are typically cheaper) without a major impact on the QoS of those services. This action might require re-allocating some functions of other slices, for the sake of global optimality (as illustrated with the re-allocation of eMBB functions). Therefore, as a consequence of this smart slice reconfiguration (both µBS activation and virtual function re-allocation), the traffic supported in fixed network segments sharply changes. Table 7-2 updates Table 7-1 after slice reconfiguration, where we observe segments that are greatly affected, e.g., traffic in the optical access sharply increases due to the addition of large F-H traffic volumes. It is worth noting that the change in the transport network traffic between time t_a and t_b cannot be predicted by typical monitoring and data analytics control loops in the fixed network [Se23], since the reason for that change is uncorrelated with past observed traffic. Hence, some *contextual information* about slice operation needs to be provided to the transport network orchestrator before it actually happens, so that the latter can prepare the fixed transport network accordingly.

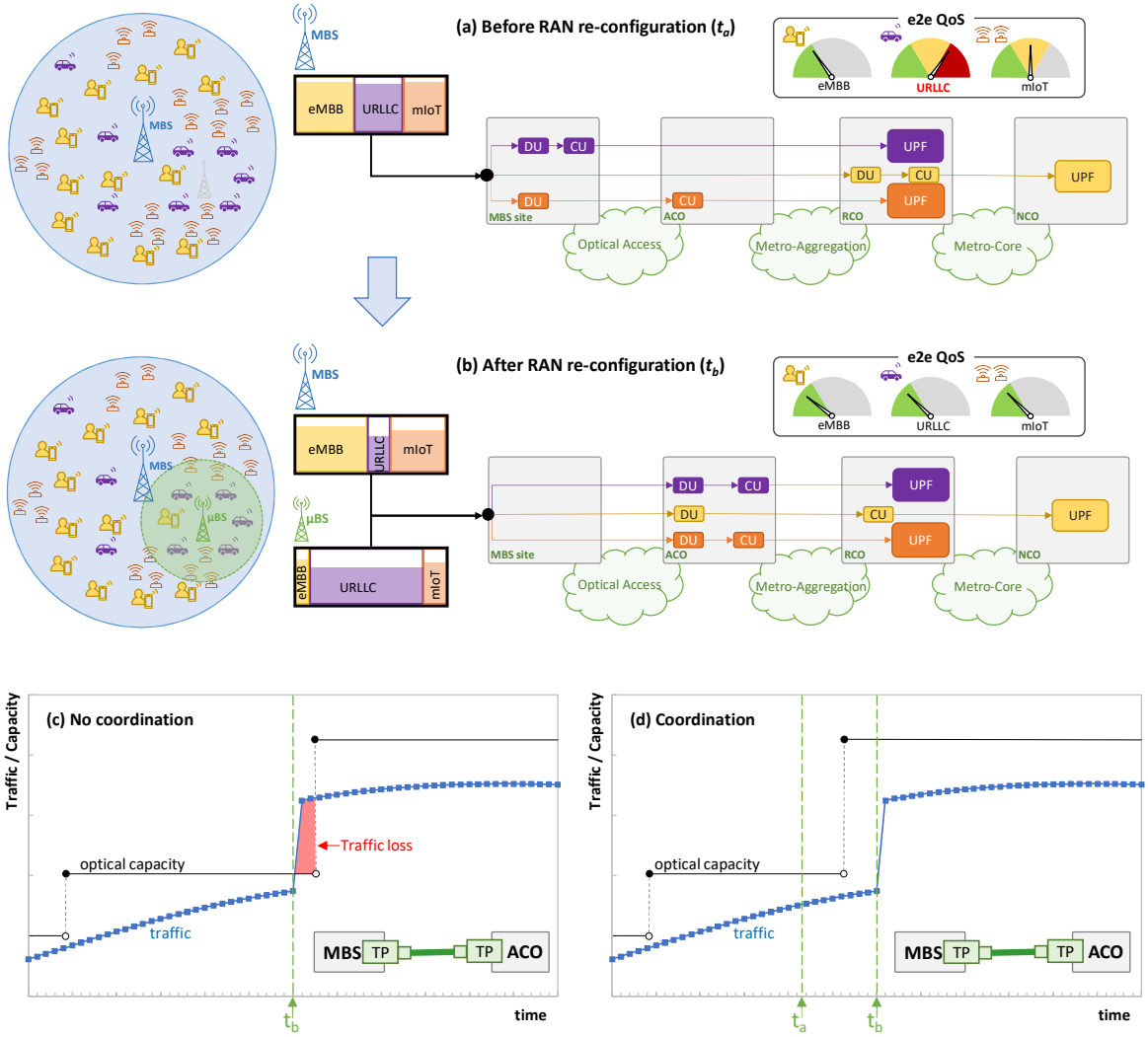

Figure 7-1: Example of RAN reconfiguration: before (a) and after (b) BS activation and function placement reconfiguration. Capacity allocation in optical access without (c) and with (d) RAN-fixed network coordination.

Table 7-1: Network Traffic Before Reconfiguration (time t_a)

Service Class	**MBS <-> ACO [Optical Access]**	**ACO <-> RCO [Metro-Aggregation]**	**RCO <-> NCO [Metro-Core]**
URLLC	B-H	B-H	-
eMBB	F-H (7.2)	F-H (7.2)	B-H
mIoT	M-H (2/4)	B-H	-

It is worth noting that the change in the transport network traffic between time t_a and t_b cannot be predicted by typical monitoring and data analytics control loops in the fixed network [Ve21.2], since the reason for that change is uncorrelated with past observed traffic.

Table 7-2: Network Traffic After Reconfiguration (time t_b)

Service Class	MBS <-> ACO [Optical Access]	ACO <-> RCO [Metro-Aggregation]	RCO <-> NCO [Metro-Core]
URLLC	**F-H (7.2)**	B-H	-
eMBB	**F-H (7.2)**	**M-H (2/4)**	B-H
mIoT	**F-H (7.2)**	B-H	-

To better illustrate this problem, Figure 7-1c sketches the autonomous operation of an access optical connection between the MBS site and ACO (following a typical control loop as the one proposed in [Ve21]) in the event of the smart RAN operation exemplified in Figure 7-1(a) and. Based on traffic prediction, capacity (in the granularity of optical subcarrier units) can be smoothly adapted to traffic changes, so that service is guaranteed and capacity (and consequently energy) is minimized. However, due to the lack of coordination between network domains, once RAN reconfiguration is performed (at time t_b), traffic suddenly increases, which can lead to traffic loss until optical capacity is reactively extended to mitigate the loss (Figure 7-1(c)). Hence, in order to overcome this problem, some *contextual information* about slice operations need to be provided to the transport network orchestrator before it actually happens (at time t_a), so as the latter can prepare the fixed transport network accordingly. This is illustrated in Figure 7-1(d), where the anticipation of the RAN reconfiguration can be added to the process of predictive optical capacity allocation. Note that, under this assumption, optical capacity can be increased before traffic changes. Although this action might produce some over-provisioning before t_b, the benefits in terms of traffic loss mitigation deserve the need for coordination.

In view of the above, the next section proposes a *context-aware autonomous network operation* solution based on sharing contextual information between the slice manager and the fixed transport network orchestrator in order to allow AI-based autonomous operation to efficiently control the optical capacity allocated to the optical connections supporting e2e slice connectivity.

7.3 Operation Context-aware autonomous network operation

Figure 7-2 sketches the architecture and workflows of the proposed context-aware autonomous network operation. By means of this procedure, the RIC shares with the

fixed network orchestrator the relevant information about slice reconfigurations to come before slice changes are actually performed. With this relevant input, the transport network orchestrator can generate a context for the different agents running autonomous network control loops. In particular, those agents are in charge of dynamically allocating capacity to the optical according to current and predicted traffic [Se23]. Although the procedure is focused on the reconfiguration of existing slices, note that it can be easily extended to the provisioning of new e2e slices, i.e., the first reconfiguration notification can be the provisioning of the slice. Table 7-3 introduces the notation consistently used in this section.

Table 7-3: Notation

N	Set of MBS sites and COs
E	Set of optical connections
V	Set of computing nodes
S	Set of e2e slices
$F(s)$	Set of virtual functions i.e., DU, CU, UPF, of slice s
$L(s)$	Set of slice links of slice s
C	Set of service classes i.e., eMBB, URLLC, mIoT
M	Set of RAN segments, i.e., radio, F-H, M-H, B-H
δ_{ne}	1, if the MBS site/CO n is adjacent to the optical connection e
δ_{vn}	1, if computing node v is located in MBS site/CO n
δ_{sc}	1, if slice s belongs to service class c
δ_{lm}	1, if slice link l belongs to segment m
δ_{lf}	1, if slice link l is adjacent to function f
δ_{le}	1, if slice link l is supported by an optical connection e
κ_s	Capacity (in normalized PRBs) assigned to slice s
ρ_s=[<f,v>]	Node-based graph of slice s, consisting of an ordered vector of tuples. Each tuple contains the function f and the computing node v where the function is placed.
π_s=[<l,e>]	Link-based graph of slice s, consisting of an ordered vector of tuples. Each tuple contains the slice link l and connection e that supports the link.
$y_e(t)$	Traffic monitored in optical connection e at time t

$x_{ecm}(t+1)$	Capacity (in normalized PRBs) supported by optical connection e and belonging to segment m of class c expected at time $t+1$
$z_e(t+1)$	Capacity (in optical capacity units) to be allocated to connection e at time $t+1$

The proposed procedure can be divided into two main processes. Figure 7-2 (a) illustrates the asynchronous context update workflow that is executed when the slice manager decides to perform a reconfiguration, which is assumed to happen at time ta (labeled as 1 in Figure 7-2 (a)). Let us assume that that reconfiguration decision is computed by an AI-based operation module (e.g., similar to that of [Zh21]), that analyzes slice monitoring data and predicts a new (better) configuration for one or several e2e slices. Without loss of generality, we consider that such reconfiguration will be performed at a given time $t_b > t_a$. For instance, the AI-based module can predict an increase in UE activity and decide, at time ta, that a new μBS in a given cell will be needed in 5 minutes from now. Consequently, the capacity of some slices will increase accordingly once the μBS is active, i.e., at time $t_b = t_a + 5$ min. Note that, in case of a reactive RAN reconfiguration due to some unexpected and unpredictable event, it might happen that $t_b = t_a$, which will limit the performance of the context-aware operation in the short period between the reconfiguration is performed and the context is conveniently updated.

As soon as the decision is taken, a slice reconfiguration notification is sent to the context manager at the fixed network orchestrator for each of the slices that will suffer some modification (2). This notification for slice s contains the time where the reconfiguration is expected to happen (t_b), and the new characteristics of the slice, namely, the capacity k_s (in capacity units like the number of PRBs) allocated at the radio segment and the graph ρ_s with the placement of each virtual function $f \in$F(s) in the set of computing nodes V. Although both characteristics are related to RAN configuration, they are also indicative of the actual UE activity and, consequently, correlated with the actual slice traffic. Hence, this notification procedure enables passing relevant information from RAN to the fixed network in an asynchronous and efficient way, without the need for continuous monitoring data exchange between domains.

Once the context manager receives a slice reconfiguration notification, it executes the context computation process (3). This process aims at computing the changes in the context due to the new reconfiguration. Those changes are timely updated in each of the agents managing optical connections (4), in order to guarantee that they can use up-to-date contextual information when running autonomous operation actions. Recall that the slice manager will trigger the needed reconfiguration workflows with sufficient anticipation to ensure that they will be active at time t_b.

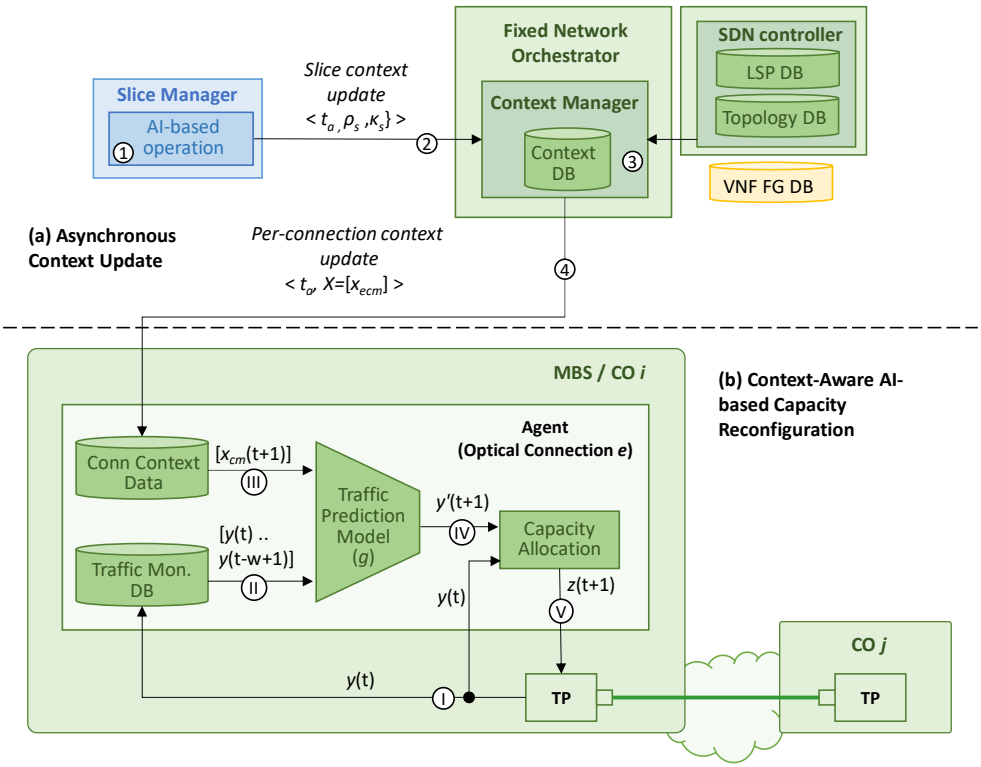

Figure 7-2: Context-aware autonomous network operation scheme

More formally, we denote X as the context variable set at a given reconfiguration time t_b. This variable contains, for each optical connection $e \in E$, the expected capacity (in normalized capacity units) that such connection will support for each service class $c \in C$ and segment $m \in M$. More formally, each $x_{ecm} \in X$ can be expressed as follows:

$$x_{ecm} = \sum_{s \in S} \delta_{sc} \sum_{l \in L(s)} \delta_{lm} \cdot \delta_{le} \cdot \kappa_s \qquad (7\text{-}1)$$

Note that this context omits particular details of individual slices since it aggregates all the slice capacity per class and segment. Thus, the privacy of individual slices (clients) is guaranteed when operating optical connections. Moreover, in order to compute context variables, static information about the mapping of slices to classes (δ_{sc}) and links to segments (δ_{lm}) is needed from the slice definition, which can be easily obtained at slice provisioning time. However, the mapping of slice links to optical connections (δ_{le}) is dynamic and depends not only on the slice graph ρ_s but also on the placement of computing nodes in COs (δ_{vn}), which can change in time depending on fixed network reconfiguration actions.

Algorithm 1 details the context computation process, which is called for every single slice reconfiguration notification. Thus, the algorithm receives all the data of one slice reconfiguration notification at a given time t_b, computes the new context for that time, and updates the context for those planned reconfigurations to happen at

time t_b (and after it). Moreover, the process has access to the internal DB that stores static slice characteristics, context status, and transport network configuration, which is conveniently kept up to date from the fixed network controller.

Algorithm 1 Context computation (Context Manager)

INPUT: s, t_b, ρ_s, κ_s

OUTPUT: -

1: **if** $DB[t_b] == \emptyset$ **then**
2: $t' \leftarrow$ getPreviousTime(DB, t_b)
3: $DB[t_b] \leftarrow DB[t']$
4: $t'' \leftarrow$ getHighestTime(DB)
5: $T \leftarrow$ getTimeRange(t_b, t'')
6: $c \leftarrow DB[s]$["c"]
7: **for each** $t_i \in T$ **do**
8: $X=\{x_{ecm}\} \leftarrow DB[t_i]$["$X$"]
9: $\kappa^0 \leftarrow DB[t_i]$["$\kappa_s$"]
10: $\Pi_s^0 \leftarrow DB[t_i]$["$\Pi_s$"]
11: $\Pi_s \leftarrow$ computeMapping(ρ_s) (Alg. 2)
12: **for each** $<l, e> \in \Pi_s^0$ **do**
13: $m \leftarrow DB[s][l]$["m"]
14: $x_{ecm} \leftarrow x_{ecm} - \kappa^0$
15: **for each** $<l, e> \in \Pi_s$ **do**
16: $m \leftarrow DB[s][l]$["m"]
17: $x_{ecm} \leftarrow x_{ecm} + \kappa_s$
18: $DB[t_i]$["Π_s"] $\leftarrow \Pi_s$
19: $DB[t_i]$["κ_s"] $\leftarrow \kappa_s$
20: $DB[t_i]$["X"] $\leftarrow X$
21: **return -**

Note that the DB can store multiple context variable sets and fixed network configurations if several reconfigurations are planned to happen at different times. Then, for convenience, slice reconfiguration and context data are stored and indexed by time, so they can be easily retrieved and modified every time a new

reconfiguration notification is processed. For those slice static parameters, we assume that they follow non-temporal indexing by slice id.

The algorithm starts by generating a new entry in the DB for time t_b (if it does not exist yet) copying the same data in the entry indexed in the immediately previous time to t_b (lines 1-3 in Algorithm 1). Then, the vector of times T is obtained; this vector contains the list of available times in the DB between t_b and the highest available time (lines 4-5). In other words, this identifies all the different contexts that need to be updated due to the new reconfiguration. Next, after retrieving the class of the slice, a loop is executed to update all contexts (lines 6-7). Thus, for each $t_i \in T$, the stored context X, capacity κ^0, and mapping Π_s^0 of slice links to optical connections are retrieved from DB (lines 8-10). Then, the new mapping Π_s is computed from graph ρ_s. (line 11). After this computation, the stored slice capacity is subtracted from all the x_{ecm} variables belonging to the stored mapping (lines 12-14), whereas the new capacity is added to the x_{ecm} variables of the new mapping (lines 15-17). Finally, slice capacity, mapping, and context are updated (lines 18-20).

Algorithm 2 details the process of computing a mapping Π_s from a new slice graph ρ_s, i.e., line 11 of Algorithm 1. After some initialization, the process iterates in the graph in order to obtain the pair of tuples <f,v> and <f',v'> that identify each graph adjacency (lines 1-4 of Algorithm 2). Then, queries to the DB are done to retrieve the slice link l that supports that graph adjacency (line 5), as well as the optical connection e that connects the sites where the computing nodes are running (lines 6-8). After that, the tuple <l,e> is added to Π_s, which is eventually returned (lines 9-10).

Algorithm 2 Compute Mapping

INPUT: ρ_s

OUTPUT: Π

1: Π_s= []
2: **for** i=1.. $|\rho_s|$-1 **do**
3: <f, v>← $\rho_s[i]$
4: <f', v'>← $\rho_s[i+1]$
5: $l \leftarrow DB[s]$.query("l" | $\delta_{lf} == 1$ & $\delta_{lf'} == 1$)
6: $n \leftarrow DB[s]$. query("n" | $\delta_{vn} == 1$)
7: $n' \leftarrow DB[s]$. query("n" | $\delta_{v'n} == 1$)
8: $e \leftarrow DB$. query("e" | $\delta_{ne} == 1$ & $\delta_{n'e} == 1$)
9: Π_s.add(<l,e>)
10: **return** Π

As already mentioned, besides asynchronous context update workflow, context-aware AI-based capacity reconfiguration workflow (Figure7-2 (b)) is executed

periodically every time t, i.e., when a new traffic monitoring sample y(t) is collected (labeled as I in Figure 7-2 (b)). To support this workflow, the following elements are deployed in the agent: a) the traffic monitoring DB storing the last w traffic values; b) the context DB with the current and future contexts affecting the optical connection; c) the traffic prediction model that estimates expected traffic y(t+1); and d) the capacity allocation module that decides the amount of optical capacity units z(t+1) (e.g., digital sub-carriers) to be configured in the optical transponder according to traffic prediction y(t+1) to stay below a desired (target) maximum load.

Once the new measurement is collected, the traffic monitoring DB is updated (II) and the expected context at prediction time, i.e., t+1, is retrieved from context DB (III). For convenience, since connection context DB contains only context variables for the specific connection, we removed the sub-index e from X variables. Then, the traffic prediction model g can be formally expressed as follows:

$$y'(t+1) = g([y(t-i), \forall i = 0..w-1], [x_{cm}(t+1), \forall c \in C, m \in M\,]) \qquad (7\text{-}2)$$

which results into $w + |C| \cdot |M|$ inputs and provide one single output (IV). Training model g can be autonomously done by the agent as soon as it receives context variables and can learn from them. Without loss of generality, we assume that the agent, at its provisioning time, receives an initial model that has been offline trained with generic traffic data and null context inputs. Then, the model is online re-trained combining contextual variables and monitored traffic, improving the accuracy of the initial one.

The traffic prediction $y'(t+1)$ is then processed by the capacity allocation module (V), jointly with the current traffic y(t). Thus, being λ_{max} the target maximum load, u the traffic supported by each optical capacity unit, and z_{max} the maximum number of optical capacity units supported by a transponder, $z(t$+1) can be formally defined as follows:

$$z(t+1) = \max(z_{max}, \left\lceil \frac{\max\big(y(t), y'(t+1)\big)}{\lambda_{max} \cdot u} \right\rceil) \qquad (7\text{-}3)$$

Without loss of generality, we assume that the transponder internally implements an agent that allocates the target number of capacity units detailed in $z(t+1)$ [Ve21]. Note that, while model g is being retrained online and the $y'(t+1)$ predictions have not reached a target accuracy performance, a more conservative policy can be temporarily used for capacity allocation, e.g., $z(t+1) = z_{max}$.

As a final remark, it is worth mentioning that the proposed context-aware approach does not increase execution time with respect to the existing optical capacity management approaches based on optical connection monitoring and local decision-making [Ve21]. Note that the connection context is stored and available at the local agent and is simply added as another input to the prediction model. Moreover, this connection context is updated asynchronously, e.g., every time a slice reconfiguration update is processed by the context manager. Thus, context management and AI-

based optical capacity reconfiguration are completely independent processes; therefore, the former does not increase the computing time of the latter.

7.4 Illustrative results

In this section, we first describe the details of the smart slice operation implemented in the simulator. After that, the context-aware autonomous network operation in Section 7-3 is evaluated by means of two numerical studies. On the one hand, the performance of connection traffic prediction using contextual information is analyzed and compared against a benchmarking approach. On the other hand, the proposed capacity reconfiguration method is validated for different operational objectives.

7.4.1 Simulation setup

As mentioned in the previous section 4.2, the theory and frame of the simulator are illustrated, this section just makes a deep description of the workflow of the simulator. Figure 7-3 shows the simulator workflow that is periodically executed every minute of simulated time; it shows the blocks and components in Figure 4-3 as well as the simulator manager that controls the execution of the workflow for a number of pre-configured time steps. At a given time step t_a, the simulator manager firstly requests the slice manager to perform a programmed RAN slice reconfiguration decided in a previous time step (labelled as 1 in Figure 7-3). In this case, no changes need to be applied and the reply is sent back to the simulator manager. Then, the simulator manager triggers slice traffic generation for the current time step t_a (2), which is propagated through the topology assuming the current graph ρ_s (3). The result of this propagation generates relevant data for performance evaluation purposes (not depicted), as well as monitoring data that is collected by the slice manager in charge of AI-based operation actions (4). These actions, which will be later explained in Section 7.4.2 below, can generate slice context updates that need to be timely announced to the context manager (5). Upon the reception of a context update, the context manager computes per-connection context updates that are distributed to the agents controlling the optical connections in the data plane (6). After receiving confirmation from all agents, a context update reply is sent back to the simulator manager (7).

The last phase to reproduce at every time step is to reconfigure the optical capacity of the connections with the prediction expected for the next time step t+1, which now can be done with updated context (8). Thus, the agents of the connections perform context-aware AI-based capacity reconfiguration as explained in Section 7.3, and notify the manager when done (9). This concludes the execution of a time step. Note that, at time step t_b, the request of BS and slice reconfiguration will produce changes

in the network, that need to be executed (10), before continuing with traffic generation and propagation.

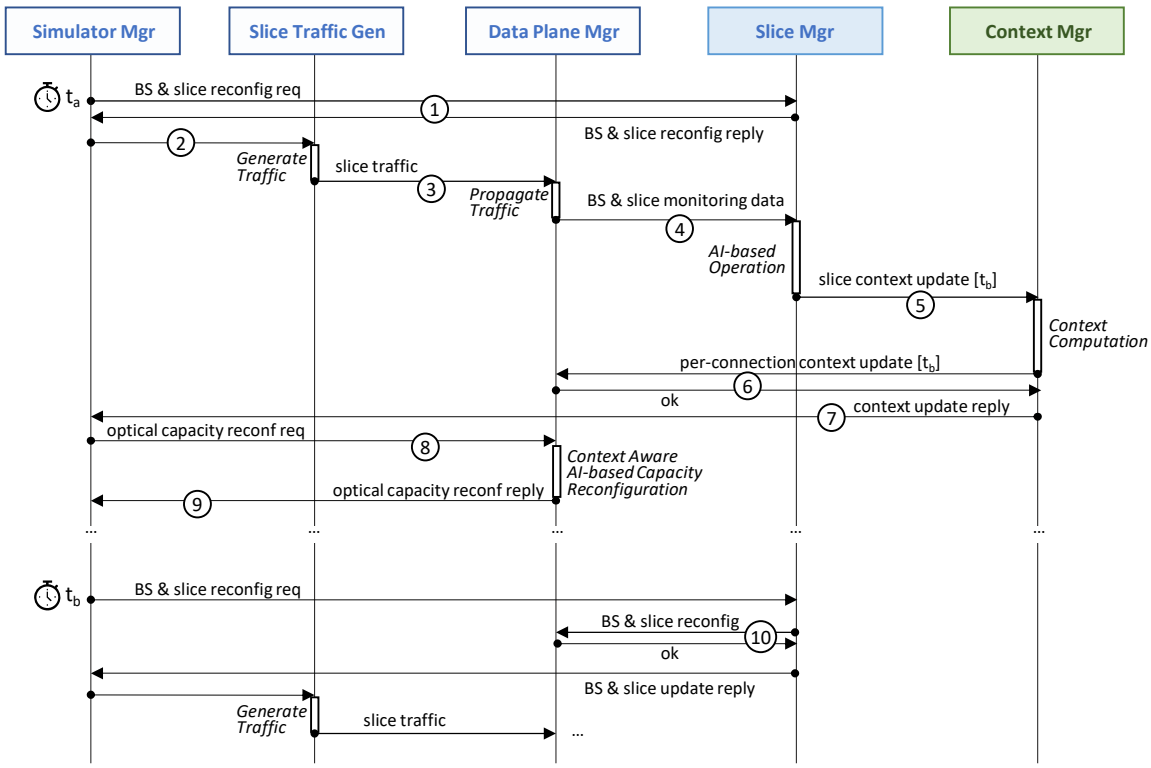

Figure 7-3: Simulator workflow

7.4.2 RAN configuration and AI-based operation

Regarding RAN configuration, we considered RAN cells consisting of 1 MBS working at the sub-6 GHz band (2 × 2 MIMO, 20 MHz bandwidth, 30 KHz subcarrier spacing, and 55 PRBs), and a variable number of µBSs configured in the mmWave band (8 × 8 MIMO, 100 MHz bandwidth, 30 KHz subcarrier spacing, and 275 PRBs). Without loss of generality, we assumed that every active BS has one slice per service class. Moreover, the F-H and B-H traffic were computed according to the formulation and references in [Wa23]. Figure 7-4 illustrates the UE traffic demand per service class (normalized to maximum) in a typical day as a function of time; this was obtained according to the assumptions in [Ru23] and [Er22] for a medium-term scenario. As can be seen, eMBB presents a clear peak in the afternoon/evening time, whereas URLLC inversely increases in the morning/noon period. Regarding mIoT, it fluctuates with peaks and valleys distributed throughout the whole day. Overall, the percentages of the traffic of each type are 45%, 15%, and 40% for eMBB, URLLC, and mIoT, respectively.

Two different RAN cell scenarios were configured according to the distribution of UEs within cell areas. For the sake of a fair comparative analysis, the total demand and demand per service class are the same in both scenarios. Then, in the *clustered* scenario, those UEs of the same class are grouped in a given region within the area and separated from the other classes. It means that a given µBSs will typically provide service to (mainly) UEs of one single class. Therefore, the clustered scenario tends towards µBSs with a predominant slice and more variable load (depending on the predominant class) along the day. On the other hand, the *distributed* scenario considers that UEs are randomly distributed and therefore, there are no clusters of UEs. As a consequence of this, all µBSs have a similar load (fluctuating with the overall traffic) with a mix of slices of all service classes.

Recall from Section 7.2 that the main objective of smart slice operation is to minimize capacity utilization by adjusting the number of active µBSs and slice capacity (which leads to energy savings) while ensuring the minimum capacity to support UE traffic and guaranteeing committed QoS of the different slices. Thus, we implemented a policy running in the slice manager that consists of rules oriented to reconfigure RAN capacity resources and slices by means of computing traffic loads (defined as UE traffic over RAN capacity) and comparing them with thresholds in order to determine the actions to perform. Whenever an action is detected at time ta, slice reconfiguration notification (if needed) is created with time $t_b = t_a + 1$, thus giving time for context update and proper use of contextual information for AI-based optical connection operation. Note that slice reconfiguration actions detected and notified at time t_a are actually performed at time t_b.

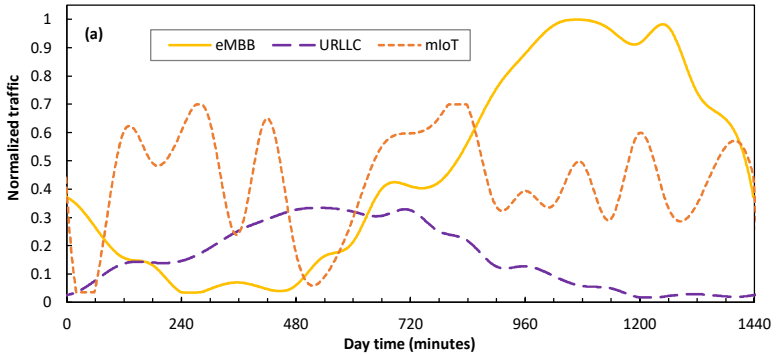

Figure 7-4: UE traffic per service class

We conducted an exhaustive set of simulations in order to find the configuration of rules and thresholds that produce a remarkable energy consumption reduction (measured in terms of active µBS) that guarantees robust QoS (no traffic loss). Specifically, the set of rules and actions that have been configured for each of the three pillars identified in Section 7.2 are the following:

Table 7-4: Flexible functional split configuration

		eMBB	**URLLC**	**mIoT**
Load threshold (low->high)		60%	30%	50%
Low Load Regime	*Split F-H/M-H*	7.2	7.2/2	7.2
	Placement DU/CU	RCO/RCO	ACO/RCO	RCO/RCO
High Load Regime	*Split F-H/M-H*	7.2/2	7.2/2	7.2/2
	Placement DU/CU	ACO/RCO	MBS/ACO	ACO/RCO

- *Dynamic μBSs management*: when one μBS exceeds a total load of 75%, the closest inactive μBS (if any) is switched on. Otherwise, if the load drops below 15% and neighbor/s μBS are active, then the μBS is switched off and its current demand is now served by active neighbor μBS.
- *Dynamic slice capacity:* PRBs assigned to each slice are assigned and released in order to keep (or approach) a slice load between 60% and 70%.
- *Flexible functional split operation*: depending on the number and utilization of μBS, we define both *low load* and *high load* regimes, in which different splits and virtual function placement are considered. Table 7-4 details the different regimes for each service class. For instance, URLLC slices follow a semi-centralized approach (DU at ACO and CU at RCO) when the RAN cell load is below 30%. When the load exceeds that threshold, functions are approached to UEs to reduce the risk of a potential delay degradation due to high load (DU at MBS and CU at ACO).

Figure 7-5 shows the total allocated RAN capacity (as a result of the number of active μBS and the PRBs allocated to each slice) of a cell with 1 MBS and 16 μBSs serving the mix of UE traffic in Figure 7-4. Note that no traffic loss is experienced by UEs (total traffic is always below capacity), which validates the proposed RAN operation. Moreover, RAN capacity resources are smoothly adapted to demand, which results in energy savings [Ve21.2]. As a direct consequence of such a smart slice operation, Figure 7-6 shows an example of traffic injected in an access connection for both clustered and distributed scenarios. As can be seen, sudden changes are observed, much sharper than that of the UE traffic in Figure 7-4. Hereafter, we consider that RAN operates according to this policy and proceed to evaluate the context-aware autonomous network operation in Section 7.3.

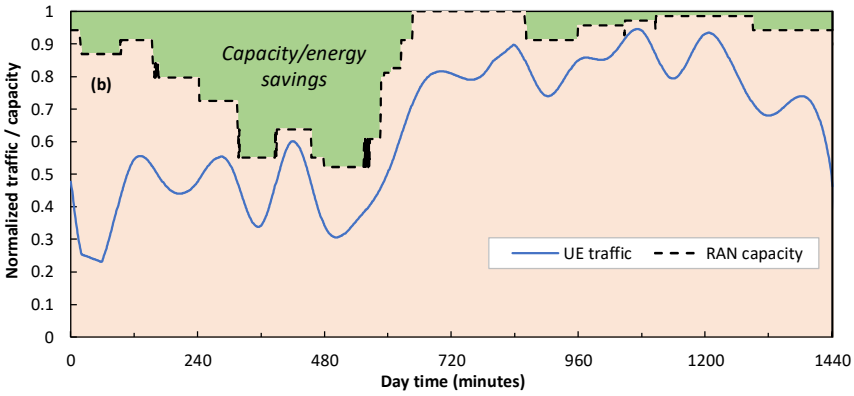

Figure 7-5: UE Smart RAN capacity allocation

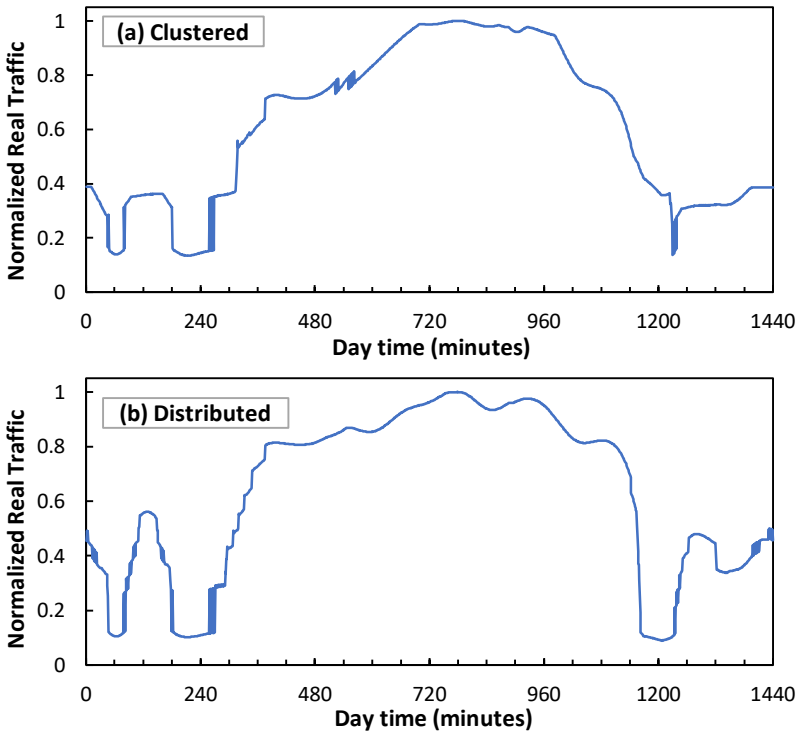

Figure 7-6: Access optical connection traffic for clustered (a) and distributed (b) scenarios

7.4.3 Optical connection traffic prediction

As previously introduced, this section is devoted to showing the accuracy of the optical connection traffic prediction model that uses RAN context data, compared against a benchmarking approach where optical network autonomous operation is performed in the absence of RAN-fixed network coordination. This study will be

conducted for both access and metro optical connections and for both distributed and clustered scenarios.

Let us start focusing on the traffic prediction carried out in the agent of an access optical connection connecting the MBS of a RAN cell with 1 MBS and 16 μBSs with its associated ACO site. In particular, we aim to evaluate the accuracy of predicting $y(t+1)$ using the model in eq. (7-2) (named as *context*) and compare it with the model presented in [Et22] (named as the *benchmark*) that showed high accuracy in predicting traffic in fixed networks supporting 4G and residential services. Note that the benchmark model predicts $y(t+1)$ only as a function of the last w traffic values stored in the monitoring DB. Both models have been implemented as deep neural networks with 2 hidden layers with 24 and 12 neurons, respectively, and a hyperbolic tangent activation function. Thus, the comparison between both models enables a fair way to evaluate the use of contextual data in B5G scenarios.

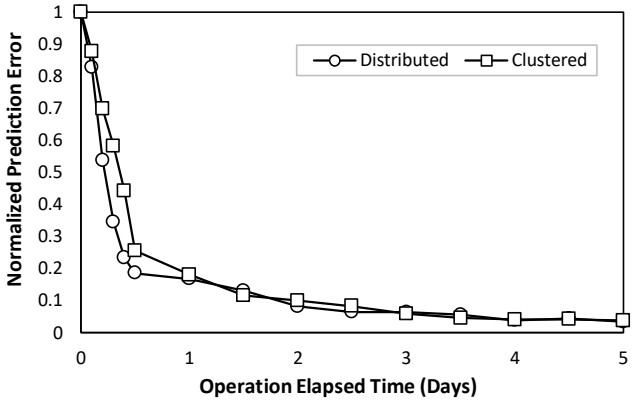

Figure7-7: Online training performance of context traffic prediction model

Our first numerical evaluation concentrates on the training of the context-based model at the provisioning time of a given optical connection, i.e. when the connectivity between a RAN cell and the ACO is established for the first time. At this provisioning time, as already mentioned in Section 7.3, an initial model trained with generic traffic data and null contextual variables is loaded to the connection agent. Then, as soon as real traffic measurements and connection context variables are received, the prediction model is retrained online until the prediction error reaches a stable value. Figure 7-7 shows the normalized prediction error (normalized to zero-day error) as a function of the elapsed operation time (in days) from the provisioning time instant. As can be seen, for both distributed and clustered scenarios, convergence to a stable error (one order of magnitude lower than that of the initial model) is reached after a few days of operation (3 to 5). Since connectivity between RAN cells and ACOs typically lasts for very long times (weeks to months), the duration of this initial tuning phase is negligible with respect to optical connectivity lifetime. Thus, contextual-based prediction models can be practically

deployed since they are easily trained and remarkably improved with the data available in the optical connection agent.

Next, assuming the error convergence performance obtained in Figure 7-7, where stable prediction error for both benchmark and context models is reached, Figure 7-8 and Figure 7-9 show a detail of the performance of traffic prediction for one of the sudden traffic changes introduced by smart slice operation in clustered and distributed scenarios, respectively. The figures show real traffic and predicted traffic using both benchmark and context methods, as well as the relative error. Note that benchmark prediction presents large inaccuracies (errors around 50%) in the presence of those sudden changes. On the contrary, the context model provides accurate and smooth prediction since it anticipates sudden changes thanks to context information.

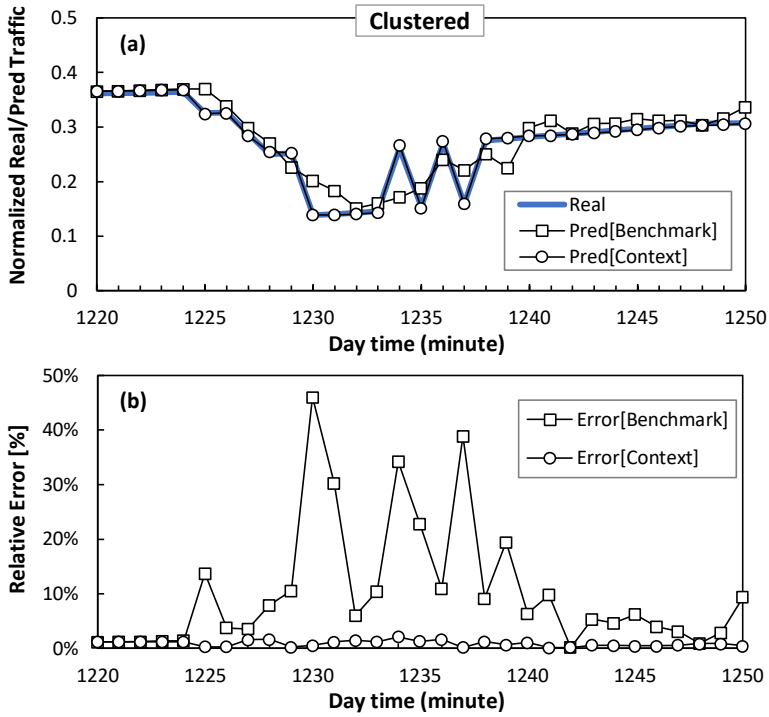

Figure7-8: Traffic prediction detail in access optical connection for clustered scenario

Table 7-5 summarizes the average and maximum error along the whole day time, for both models, using mean square error (MSE) and relative error as accuracy metrics. Moreover, the reduction of error (in %) that the context model provides with respect to benchmark one is detailed. In view of these results, we can conclude that the context model clearly improves benchmarking one, by remarkably reducing both average and maximum error. Especially interesting is the case of the maximum relative error: the context model reduced 93% and 89% of the error of the benchmark model for clustered and distributed scenarios, respectively.

Let us now focus on evaluating the accuracy of traffic prediction for a metro optical connection connecting an ACO with its reference RCO. Recall that this type of connection supports traffic from a number of RAN cells, as well as traffic from fixed AP sites (which typically show smoother daily patterns mixing residential and business traffic [Et22]). Figure 7-10 shows the average and maximum error of a metro optical connection that supports 10 RAN cells (mixing clustered and distributed scenarios) and a variable number of fixed APs. In particular, we defined several configurations with different ratios of fixed / RAN (1 means that fixed and RAN traffic volumes are equal, 2 means that fixed traffic doubles RAN one). It is worth noting that improved prediction by context model persists even for large ratios (i.e., when overall traffic becomes more predictable). In other words, sharp fluctuations induced by RAN operation have a large impact on aggregated flows, which validates extending context updates to metro optical connections (and not only access ones). Figure 7-11 depicts the evolution of maximum relative error as a function of the number of RAN cells for ratios 0.5 and 2. In line with previous results, the error reduction achieved by the context model is outstanding. Indeed, the benchmark model produces relative maximum errors of 30% in optical connections supporting large traffic volumes, which can have a large impact in terms of absolute errors (dozens to hundreds of Gb/s).

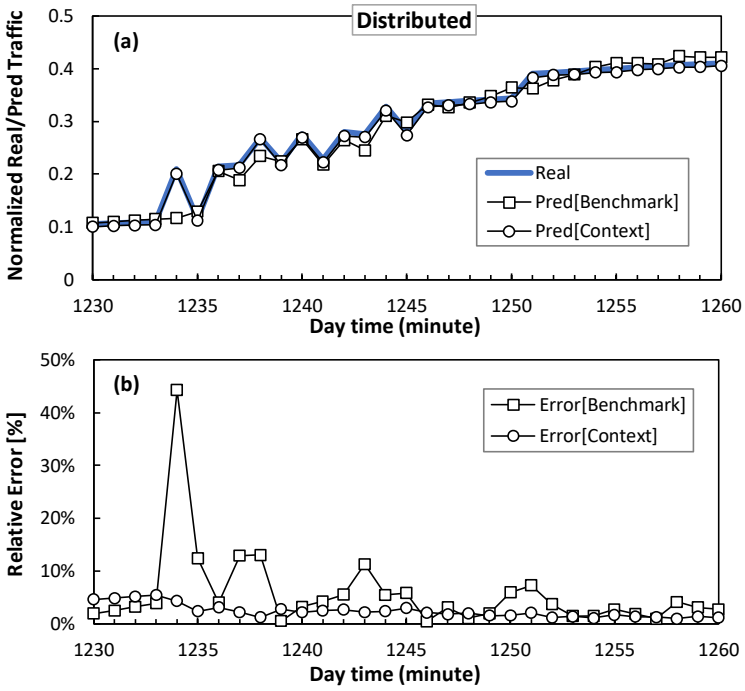

Figure 7-9: Traffic prediction detail in access optical connection for distributed scenario

Table 7-5: Summary of access optical connection traffic prediction

	Benchmark		**Context**		**Reduction [%]**	
MSE	avg	max	avg	max	avg	max
Clustered	0.19	5.03	0.14	0.38	26%	92%
Distributed	0.26	5.43	0.17	0.84	35%	85%
Rel. Error	avg	max	avg	max	avg	max
Clustered	0.018	0.913	0.008	0.063	55%	93%
Distributed	0.019	0.551	0.012	0.059	37%	89%

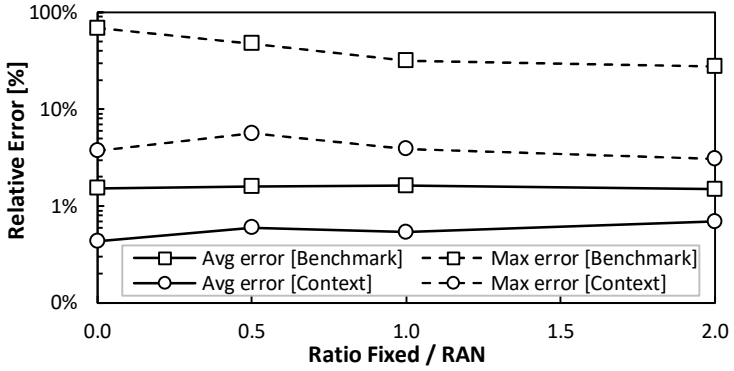

Figure 7-10: Traffic prediction in Metro optical connection

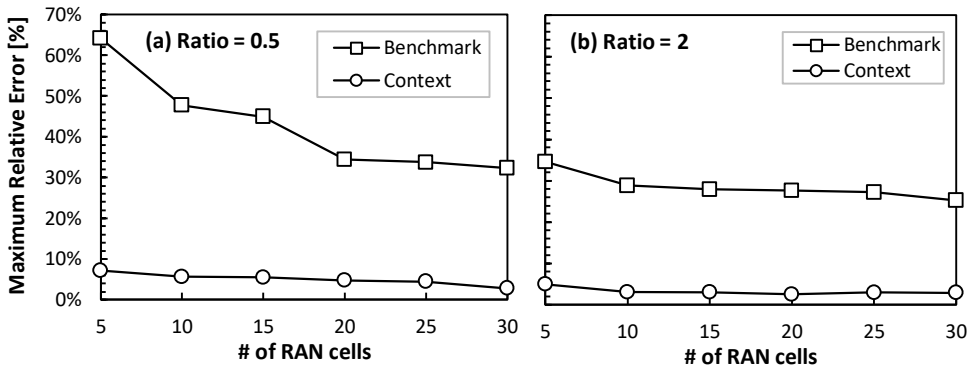

Figure 7-11: Maximum error for ratio 0.5 (a) and 2 (b)

In view of these results, we can conclude that using RAN context significantly increases the accuracy of models predicting the traffic supported by optical connections. In terms of reduction of maximum error, the context model reduces between 85% and 93% of the error of the benchmarking model when predicting traffic from access connections. Regarding metro connections, where RAN and fixed access traffic are mixed, the benchmarking model provides a much larger prediction error (~30%) than context one (~2%), even when fixed access traffic doubles RAN traffic.

7.4.4 Optical connection capacity reconfiguration

Once the accuracy of the context model has been validated and its improvement with respect to benchmark one has been shown in terms of prediction error, let us finally focus on evaluating the impact of such predictions to dynamically allocate optical connection capacity. For the sake of simplicity, we focus on access connections and evaluate the size of RAN cells for the distributed scenario, i.e., the number of µBS ranges between 16 and 128 µBS to consider different cell densities [Bg22]. Moreover, we consider two different operational objectives to be configured in the capacity allocation module in the connection agent. On the one hand, the capacity minimization objective aims at adjusting the allocated optical capacity as close as possible to actual traffic. Thus, the load of optical connections can grow up to 95% in order to exploit optical capacity resources. On the other hand, the delay reduction objective allocates an excess of capacity to avoid congestion in the fixed segment that could introduce an additional delay to RAN slices. Therefore, the maximum load is kept below 80%.

Figure 7-12(a-c) and Figure 7-12(d-f) show the obtained results for capacity minimization and delay reduction operational objectives, respectively. The results include the average and maximum values of capacity allocated, load, and traffic loss as a function of RAN cell density. One initial observation is that both models allocate similar overall optical capacity (small differences in average that are hard to observe). Moreover, the (large) relative errors in model prediction shown in the previous section for low-density RAN cells (16 µBS) do not produce an effect on the decisions made by the connection capacity allocation module. This is because absolute fluctuations stay between the size of the optical capacity units (recall that we consider 25 Gb/s sub-carriers) and consequently, the benchmark model stays as a valid traffic predictor.

However, as soon as the density of cells increases (which is the expected trend [Bg22]), the real magnitude of prediction inaccuracies exceeds the sub-carrier capacity. Thus, the benchmark model (sometimes) underestimates the required capacity to support the target load, which leads to unacceptable performance (traffic loss) for RAN cells as dense as 32 µBS. On the contrary, optical connection capacity allocation based on context model prediction provides accurate operation, always below the desired target load and eliminating traffic loss, for a wide range of RAN cell densities. In light of these results, we can eventually validate the proposed context-aware autonomous network operation procedure to guarantee smooth and smart fixed transport network operation in the presence and under the activity of smart B5G slice operation.

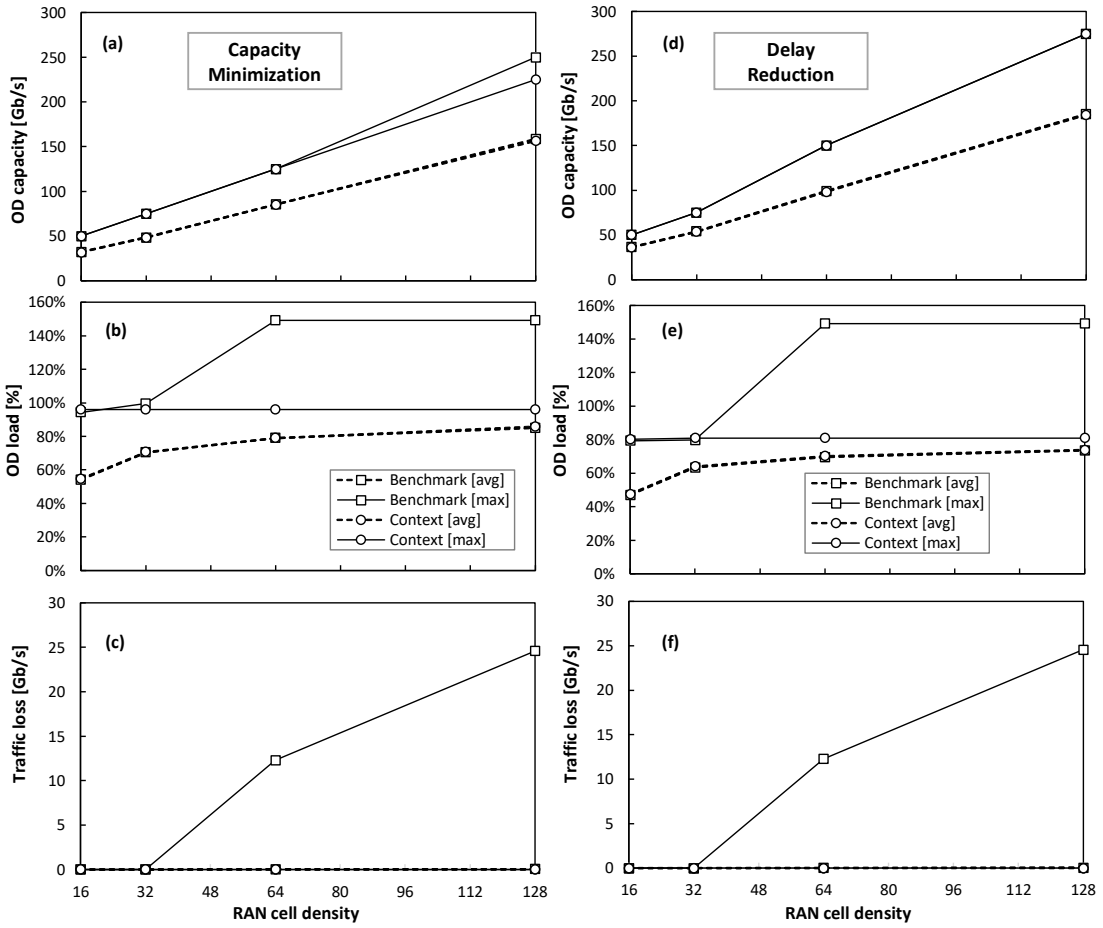

Figure 7-12: Performance of optical connection capacity reconfiguration for capacity minimization (a-c) and delay reduction (d-f) objectives

7.5 Conclusions

In this chapter, we focused on the impact of B5G slice operation on the underlying transport network infrastructure in charge of providing computing and connectivity resources. In particular, we analyzed how smart slice operation based on pillars such as dynamic capacity reconfiguration, adaptive μBSs management, and flexible functional split configuration injects highly variable and hard-to-predict traffic in access and metro fixed networks. Thus, effective autonomous fixed network operation cannot be achieved without coordination between the slice manager and fixed networks. In order to maximize autonomous fixed network performance while hiding sensitive information of individual slices, we explored the concept of context sharing, which allows the slice manager to asynchronously inform about slice reconfigurations to be done in an aggregated and private way with enough anticipation to synchronize with autonomous fixed network operation. This slice

context is then incorporated into AI-based prediction models used for dynamic capacity allocation in optical access and metro networks.

The proposed architecture and algorithms for context update and usage for prediction purposes in optical connections were validated by means of exhaustive simulations considering a realistic network operator infrastructure. The results showed that adding slice context to traffic prediction greatly enhances prediction accuracy. In particular, maximum prediction error is remarkably reduced as compared with benchmarking approaches for traffic prediction. This prediction accuracy gain is observed under different RAN demand scenarios with heterogeneous services, as well as for optical access networks carrying only RAN traffic and for metro networks transporting a mix of RAN and fixed (residential and business) traffic.

The value of this prediction accuracy improvement was eventually evaluated for dynamically allocating capacity to the optical connections supporting slice links, under two different network operational objectives: capacity minimization and delay reduction. In both cases, the benefits of the proposed context-based operation increased with the size and density of RAN cells. Therefore, we can conclude that, with the increase of RAN traffic and density of RAN cells foreseen for B5G scenarios, the need for context-based coordination to achieve actual e2e smart operation will be required.

Chapter 8

Coordination of Radio Access and Optical Transport for Delay Guaranteeing

New 5G and beyond applications demand strict delay requirements. In this chapter, we propose coordination between radio access and optical transport to guarantee such delay while optimizing optical capacity allocation. Illustrative results show near real-time autonomous capacity adaptation benefits based on radio access delay requirements.

8.1 Introduction

The support of 5G and beyond use cases requires bringing the optical network to the very edge of the network not only to increase capacity but also to guarantee end-to-end (e2e) quality of service (QoS), e.g., delay. Such e2e QoS requires that both 5G radio access network (RAN) and optical transport operate under strict QoS constraints [Be20]. However, establishing fixed capacity optical connections to connect RAN to 5G core entails large capacity overprovisioning, increasing thus the total cost of ownership to network operators. An option is to implement digital subcarrier multiplexing (DSCM) optical systems, which allow to activation/deactivation of each subcarrier (SC) independently in near real-time to provide just the needed capacity and meet the maximum delay requirement [Ve21.2]. The near real-time operation needs to be implemented as close as possible to the data plane, to liberate the software-defined networking (SDN) controller from those tasks.

Even though the use of DSCM systems can reduce costs, there is still a large amount of overprovisioning in the optical network just because of the lack of coordination between radio and optical segments. coordination between both network segments is considered, so the e2e delay is ensured and optical capacity overprovisioning is decreased. With such a solution, the maximum delay allowed for the optical network can be changed based on the requirements from the RAN and adapt the optical capacity accordingly; this will bring capacity overprovisioning to a minimum.

8.2 **Automatic Operation**

Figure 8-1(a) shows the analyzed e2e scenario, where the user equipment (UE) requests virtualized 5G services placed in a remote location in the fixed network, e.g., a metro/core site. Without loss of generality, we assume that UEs and the 5G core are the endpoints of e2e traffic and that some maximum e2e delay needs to be ensured. Hence, the e2e traffic flow consists of two components for the RAN and the optical network.

The component of such e2e traffic flow that traverses the RAN is represented by a blue thick arrow. We consider that a RAN cell consists of a single macro base station (i.e., a next- Generation Node B - gNB) that covers the whole cell area. The configuration of the gNB, e.g., numerology, bandwidth, power, etc., can be configured to support capacity and latency requirements. This configuration has a direct impact on the actual QoS. As an example, the inset graphs in Figure 8-1(a) show the behavior of the RAN delay component as a function of gNB load, assuming a typical 5G configuration. We observe that, in order to achieve low delay, the RAN controller needs to operate the gNB up to some percentage of its capacity (e.g., 60%); otherwise, the RAN delay component would increase up to the point that the committed e2e delay requirement (e.g., 2 ms) cannot be achieved. Even when the RAN works in that low to moderate load regime, delay fluctuations can be observed since traffic typically varies throughout the day, making load also variable in time. In this example, the RAN delay component oscillates between 1 and 1.5 ms in a day, which entails a stringent delay budget for the optical network delay component that such a network has to guarantee.

Let us assume that a cell site gateway (CSG) is the boundary between the RAN and the fixed optical transport network. For the sake of simplicity, we assume that traffic flow at the fixed network (green thick arrow) transparently traverses single or multiple optical domains inside a single e2e light path. The capacity of such a light path can be properly dimensioned by dynamically activating/deactivating SCs to provide the required QoS. In line with [Ve21.2], autonomous optical capacity management with QoS guarantees can be realized in the fixed transport network segment by means of the control architecture sketched in Figure 8-1(a), where different entities are considered (from bottom to top): i) the transponder (TP) agent

that is in charge of collating telemetry data, e.g., traffic and measured delay from the TPs, as well as to manage SCs to ensure the committed QoS; ii) the capacity manager that uses telemetry to run policies, models, and rules to find the required capacity that better satisfies the target QoS; iii) the SDN controller that is in charge of the initial light-path setup and of communicating the capacity manager key parameters, such as the required QoS. It is worth noting that both RAN and optical network domains operate without any coordination among them, which entails overprovisioning capacity in the light path to meet a fixed target optical network delay component that absorbs delay variations introduced by the RAN. This is illustrated in Figure 8-1(b) (left), where the optical capacity is dynamically adjusted to keep the optical network delay component under control. In our example, if the RAN can introduce up to 1.5 ms of delay, the target optical network delay needs to be setup at 0.5 ms to guarantee that the maximum 2 ms of e2e delay is met. Although this uncoordinated strategy can guarantee e2e QoS and provide some dynamic capacity adaptation, it results in large overprovisioning if the RAN delay is far from the maximum.

In view of the above, we propose coordinating RAN and optical network operation to dynamically adjust the target optical network delay component to the current traffic conditions. We claim that overprovisioning can be greatly reduced while guaranteeing e2e delay (Figure 8-1b right).

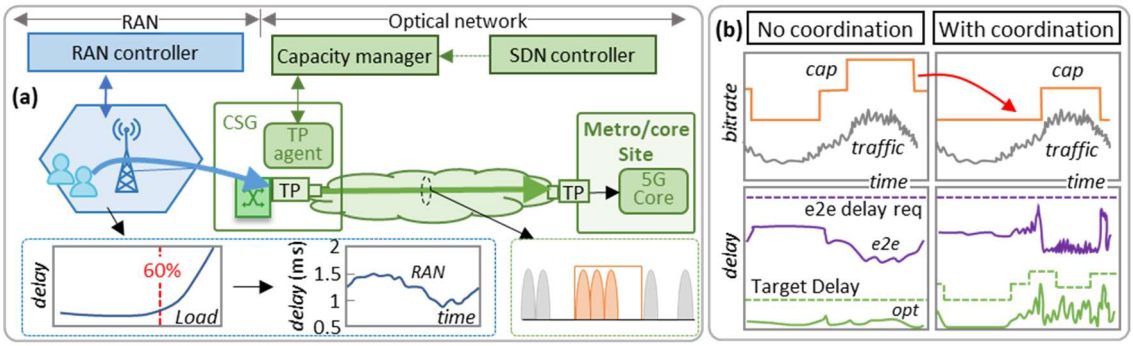

Figure 8-1: Reference e2e scenario, b) autonomous capacity management performance

8.3 Autonomous Capacity Management Architecture

Figure 8-2 details the proposed architecture for capacity management with QoS guarantees. We adapted the architecture for packet flow management in single network domains in [Ve21.2] to deal with i) coordination between RAN and optical network domains; ii) QoS assurance; and iii) DSCM-based optical capacity management. At the core of the autonomous operation system, the capacity manager

module implements an external interaction dynamically. The module finds, at every time interval t, the minimum optical capacity $z(t)$ to ensure that the optical network delay component $d_{opt}(t)$ does not exceed a given target $dmax_{opt}$. Thus, at every time interval t, telemetry traffic $x(t)$ and delay $d_{opt}(t)$ are retrieved from the TP agent and fed into the capacity manager module (1). Then, this module makes a learning, computing, and decision-making action to translate into the required capacity for the next time interval $z(t+1)$ (2), which is processed by the TP agent to activate/deactivate SCs (3). Finally, coordination between domains is implemented to satisfy the e2e delay requirement. Let us assume that, upon provisioning of the e2e service, the optical network controller receives the required e2e delay $dmax_{e2e}$ (I). Once the operation starts, the RAN controller is able to asynchronously notify its maximum delay $dmax_{RAN}$ to the optical network controller (II). The optical network controller computes the requirement for the optical segment as $dmax_{opt} = dmax_{e2e} - dmax_{RAN}$, and pushes this value to the capacity manager (III). At this point, an AI engine embedded in the capacity manager will work to guarantee such updated $dmax_{opt}$ requirements.

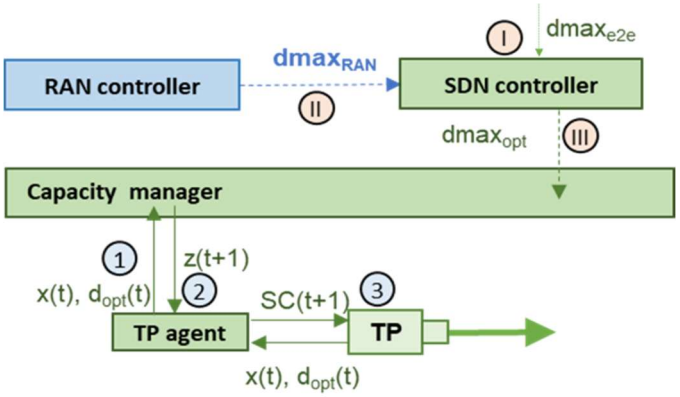

Figure 8-2: Coordinated network operation scheme

8.4 Results and Concluding Remarks

For numerical evaluation, we developed a simulation environment combining different tools for RAN and optical network emulation. Regarding RAN simulation, we adopted the open-source discrete-event ns-3 network simulator with the 5G-LENA module extension [Pa19] to simulate the 5G New Radio (NR) technology. In particular, we simulated a scenario with a gNB (one omnidirectional antenna, single bandwidth part at a central frequency of 28GHz and 400MHz of bandwidth) and several UEs sending uplink UDP traffic. Traffic from video on demand (VoD) and online gaming services was generated according to the characterization in [Ru18]; we assume a typical daily profile to create a variable gNB load in a time not exceeding 60%. For the sake of simplicity, we considered interference free radio links

with line-of-sight between the gNB and the UEs. The proportional fair scheduler was considered. UDP traffic was scaled up to emulate a dense area with 40 gNBs with the same input traffic and delay behavior.

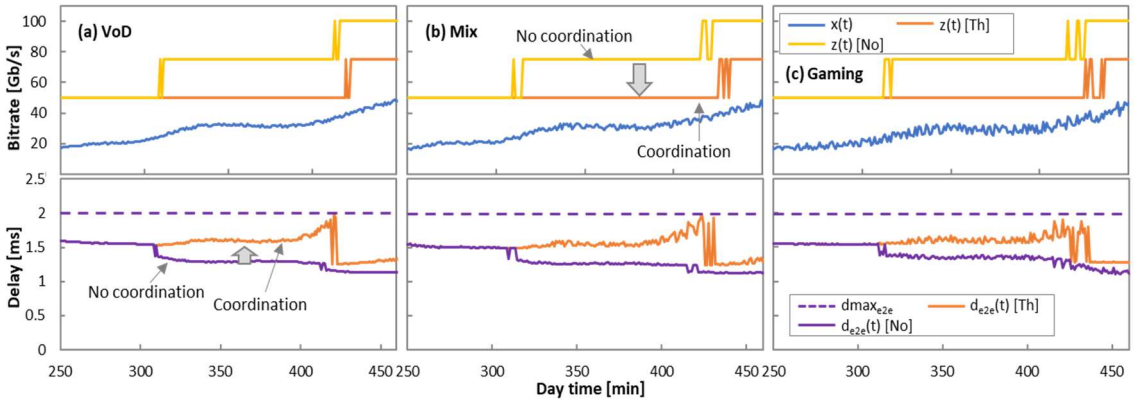

Figure 8-3: Capacity management and e2e delay assurance for scenarios: VoD only (1), VoD + Gaming (2), and Gaming only (3).

The UDP traffic and maximum delay obtained with the RAN simulator were then aggregated with a granularity of 1 minute and injected as x(t) and $dmax_{RAN}$ in a Python-based optical simulation environment. This tool, built upon those ones in [Ve21.2] and [Ba21], implements all the blocks and procedures shown in Figure 8-2. The capacity management module was implemented using Twin Delayed Deep Deterministic Policy Gradients (TD3), an off-policy ML algorithm that uses a pair of critic networks and an actor-network that is updated with some periodicity [Fu18]. A set of TD3-based models, trained with traffic with similar characteristics to that of VoD and gaming for different operational ranges for $dmax_{opt}$, were loaded in the sandbox domain before starting simulations. Finally, a DSCM-based light-path was emulated, assuming a 100 Gb/s optical TP equipped with 4×25 Gb/s SCs.

The proposed coordination method described in Section 8-3 has been numerically evaluated and compared against two benchmarking approaches. Aiming at evaluating the benefits of coordination, the first method (labeled *No*) assumes no coordination and hence, $dmax_{opt}$ is fixed to a restrictive value to ensure $dmax_{e2e}$. Then, aiming at evaluating the performance of the proposed coordination scheme, the method (labeled *Th*) implements coordination but it implements a threshold-based method to adjust capacity with *perfect knowledge* of the actual future delay, which is unfeasible to implement in a real network.

Figure 8-3 shows the allocated optical capacity (top row) and e2e delay (bottom row) for part of a day with increasing traffic; results for the evaluated approaches and different traffic scenarios (only VoD, mix of VoD and gaming, and only gaming) and $dmax_{e2e}$ = 2ms are plot. Table 8-1 complements Figure 8-3 with the optical capacity allocated in a day and the total number of SC activation/deactivations per day. We

observe that coordination allows a remarkable reduction of overprovisioned capacity in all evaluated scenarios without violating $dmax_{e2e}$. Interestingly, low loads produced slightly higher $dmax_{RAN}$ (and consequently, a stringent $dmax_{opt}$ requirement), which is due to signaling overhead [Pa18]. We observe that the proposed method is able to improve the unrealistic threshold-based method with *a priori* perfect knowledge of $d_{opt}(t)$. Such improvement is small in terms of the number of SC changes but significant in terms of capacity allocation. Note that our approach requires less changes, which indicates that its operation is able to anticipate increments or decrements in optical capacity, thus reducing unnecessary capacity fluctuations as well as overall management complexity.

Table 8-1: Allocated capacity and SC changes per day

Approach	**Capacity [Tb/s]**			**SC changes**		
	VoD	**Mix**	**Gaming**	**VoD**	**Mix**	**Gaming**
No	109.9	109.7	10.9.1	19	17	25
Th	87.9	87.7	87.2	12	18	30

8.5 Conclusions

The benefits of coordination between RAN access and optical transport for e2e QoS assurance have been demonstrated through simulation. Demonstrated findings reveal the advantages of adapting network capacity in almost real-time, guided by the demands for radio access latency, and the proposed operation showed optimal and smooth optical capacity allocation.

Chapter 9

Closing Discussion

9.1 Main Contributions

This Ph.D. thesis focuses on applying multiple techniques for coordinating wireless and optical networks. The main contributions are summarized as follows:

- In Chapter 5, a ML-based cell clustering and classification compensator has been presented that might significantly improve the efficiency of radio resource allocation and enable the switching on/off of BS dynamically. The proposed system is based on three main components: i) a SOM-K clustering model able to precisely cluster the current cells while minimizing overhead; ii) a decision tree-based classifier that is able to accurately classify the cluster results with fine granularity; iii) a dense urban scenario construction, which decides when and which BS need to be turned on/off based on dynamic cell management.
- In Chapter 6, part of the proposed adaptive functional splitting is expected to change not only the volume but also the requirements of the traffic to be supported by the fixed transport network. A RAN controller is proposed and extended with additional modules to perform estimation of the traffic injected into the fixed network, which will depend on both users' demand and the functional split implemented, as well as on the RAN operation approach. The experiments caused us to come up with both users' demands and functional split, as well as on RAN operation to allow optical capacity setup.
- In Chapter 7, a novel approach that explores the concept of context-aware network operation, where the slice control anticipates the aggregated and anonymized information of the expected slice operation that is sent to the

fixed network orchestrator in an asynchronous way is proposed. This is the first work using contextual information sharing between slice management and fixed transport domains for an autonomous e2e network operation that coordinates and encompasses both the RAN and fixed network operation. The proposed context-aware method aims at two main objectives: i) Enhance the performance of the self-operating fixed network while concealing the confidential details of each individual slice. ii) Asynchronously deliver details on the required slice reconfigurations in a consolidated and confidential manner, ensuring sufficient lead time for synchronization with the operations of an autonomous fixed network.

- In Chapter 8, we propose coordination between radio access and optical transport to guarantee such delay while optimizing optical capacity allocation. Simple coordination between both network segments is explored, so the e2e delay is ensured and optical capacity overprovisioning is decreased. The maximum delay allowed for the optical network can be changed based on the requirements of the RAN, which will allow adapting the optical capacity accordingly; this will bring capacity overprovisioning to a minimum.

9.2 List of Publications

9.2.1 Publications in Journals

[SENSORS24]	**S. Wang**, M. Ruiz, and L. Velasco, "Context-Based e2e Autonomous Operation in B5G Networks," MDPI Sensors, vol. 5, pp. 1-23, 2024. DOI: 10.3390/s24051625

9.2.2 Publications in Conferences

[ICTON24]	**S. Wang**, M. Ruiz, and L. Velasco, "Smart Operation of 6G Radio Access Networks," in Proc. International Conference on Transparent Optical Networks (ICTON), 2024
[ICTON23]	**S. Wang**, M. Ruiz, and L. Velasco, "Optical network traffic analysis under B5G/6G RAN operation," in Proc. International Conference on Transparent Optical Networks (ICTON), 2023
[ONDM23]	Barzegar, S.; Richart, M.; **Wang, S.**; Castro, A.; Ruiz, M.; Velasco, L. Coordination of Radio Access and Optical Transport. In Proceedings of the International Conference on Optical Network Design and Modeling (ONDM), 2023

9.2.3 Other publications

[IEEE Access21]	**S. Wang** et al.," Extracting Cell Patterns from High-dimensional Radio Network Performance Datasets using Self-Organizing Maps and K-means Clustering", IEEE Access, vol.9, pp.2169-3536, 2021
[ICCBDA20]	**S. Wang** et al., "Analysis of Self-Organizing Maps (SOM) Methods for Cell Clustering with High-Dimensional OAM Collected Data," in Proc. IEEE International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2020
[ISCC19]	**S. Wang** et al., "On the use of prioritization and network slicing features for mission critical and commercial traffic multiplexing in 5G Radio Access Networks," in Proc. IEEE Symposium on Computers and Communications (ISCC),2019

9.3 List of Research Projects

9.3.1 EU-US Funded Projects

- **SEASON**: Smart Networks and Services Joint Undertaking under the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101096120 (SEASON).

9.3.2 National Funded Projects

- **IBON:** *AI-Powered Intent-Based Packet and Optical Transport Networks and Edge and Cloud Computing for Beyond 5G*, Ref: PID2020-114135RB-I00, 2021-2024.
- **CSC:** China Scholarships Council (No. 201808390034)

9.4 Future Work

In order to verify the applicability and feasibility of the proposed methods and algorithms, it is planned to perform experimental assessment of the RAN smart operation by collecting and processing real monitoring data in an O-RAN testbed. By means of x-apps, RAN telemetry can be collected and pushed to a multi-agent system in charge of holding and communicating the different processes that allow the coordination of both RAN and fixed network domains [Ve23.2].

List of Acronyms

AAU	Active antenna unit
AC	Admission control
ACO	Access central office
AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
B5G	Beyond 5G
B-H	Back-haul
BPSK	Binary Phase Shift Keying
BS	Base station
CO	Central offices
CQI	Channel quality indicator
C-RAN	Cloud RAN
CSG	Cell site gateway
CU	Centralization unit
DBI	Davies & Bouldin index
DL	Deep learning
DNN	Deep neural network
DP	Dual polarization
DSP	Digital signal processing
DU	Distribute unit
DSCM	Digital subcarrier multiplexing

EDFA	Erbium-doped fiber amplifiers
EI	Enrichment information
ETSI	European telecommunications standards institute
E2E	End to End
FEC	Forward error correction
F-H	Front haul
GNN	Graph neural network
HO	Hand over
IBN	Intent-based networking
ICI	Inter-cell interference
ITU	International telecommunication union
LSTM	Long-short term memory
KNN	K-nearest neighbors
KPI	Key performance indicators
MAC	Media access control
MBS	Macro base stations
μBS	Micro base stations
MBB	Mobile broadband
MF	Modulation formats
mIoT	Massive internet-of-things
M-H	Mid-haul
ML	Machine learning
MNO	Mobile network operators
MOOP	Multi-objective optimization problem
Multi-AS	Multi-application and services
MVNO	Mobile virtual network operator
NB-IoT	Narrow band-internet of things
NCO	National COs
NGMN	Next generation mobile networks alliance
NR	New radio

NT-RIC	Near real time RIC
NVS	Network virtualization substrate
OFDMA	Orthogonal frequency division multiple access
O-RAN	Open RAN
O-CU	O-RAN CU
O-CU-CP	O-RAN CU control plane
O-CU-UP	O-RAN CU user plane
OEO	Optical-electrical-optical
PAM4	4-level Pulse amplitude modulation
P2MP	Point to multi-point
P2P	Point to point
PDM	Polarization division multiplexing
PF	Proportional fairness
PDCP	Packet data convergence protocol
PDCP-C	Packet data convergence protocol-control
PDCP-U	PDCP-user plane
PHY	Physical layer
PRB	Physical resource block
8QAM	8-level Quadrature amplitude modulation
QoE	Quality of experience
QoS	Quality of service
QPSK	Quadrature phase shift keying
QE	Quantization error
RAN	Radio access network
RCO	Regional COs
RIC	RAN intelligent controller
RF	Radio frequency
RLC	Radio link control
RRC	Radio resource control
RRCF	Root-raise-cosine filter

RRH	Remote radio head
RRM	Radio resource management
RU	Radio unit
SANIT	Secure and latency-aware DT-assisted resource scheduling algorithm
SC	Subcarrier
SDN	Software-defined networking
SLAs	Service level agreements
SDAP	Service data adaptation protocol
SDN	Software-defined networking
SMO	Service management and orchestration
SNR	Signal-to-noise ratio
SR	Symbol rates
SVM	Support vector machine
TCO	Total cost of ownership
TE	Topographic error
TP	Transponder
TTI	Transmission time interval
TX	Transmitter
μBS	Micro base stations
UE	User equipment
URLLC	Ultra-Reliable Low Latency Communication
UPF	User plane function
VoD	Video on demand
WSSs	Wavelength selective switches

References

[Ah23] M. Ahsan, A. Ahmed, A. Al-Dweik, A. Ahmad," Functional Split-Aware Optimal BBU Placement for 5G Cloud-RAN Over WDM Access/Aggregation Network", IEEE Systems Journal., vol. 17, no. 1, pp. 123–133, Mar. 2023.

[Al22] G. M. Almeida, L. D. L. Pinto, C. B. Both, and K. V. Cardoso, "Optimal joint functional split and network function placement in virtualized RAN with splittable flows," IEEE Wireless Commun. Lett., vol. 11, no. 8, pp. 1684–1688, Aug. 2022.

[Am16] 5G America "Network slicing for 5G Networks & Services" November 2016

[An23] D. Anand, M. A. Togou, and G.-M. Muntean," A Multi-Classification Machine Learning-based Solution to Mitigate Co-Tier Interference", in Proc ICC 2023 - IEEE International Conference on Communications, Rome, Italy, pp.4840-485,28 May - 01 Jun 2023

[Ba21] S. Barzegar et al., "Packet Flow Capacity Autonomous Operation based on Reinforcement Learning," MDPI Sensors, vol. 21(8306), pp.1-24, 2021.

[Ba23] S. Barzegar, M. Richart, S. Wang, A. Castro, M. Ruiz, L. Velasco, "Coordination of Radio Access and Optical Transport", In Proc. International Conference on Optical Network Design and Modelling (ONDM), Coimbra, Portugal, pp.1-3, 8–11 May 2023.

[Be20] A. Bernal, M. Richart, M. Ruiz, A. Castro, and L. Velasco, "Near real-time estimation of end-to-end performance in converged fixed-mobile networks," Elsevier Computer Communications, vol. 150, pp. 393-404, 2020.

[Bg22] B5G-OPEN D2.1: Definition of Use Cases, Requirements, and Reference Network Architecture. Version 1.0, 2022. Available online: https://www.b5g-open.eu/deliverables/

[Bg24] B5G RAN slice traffic generator, https://gitlab.com/gco-upc/b5g-ran-slice-traffic-generator, Access on 2024

[Ca18] R. Casellas, R. Martínez, R. Vilalta, R. Muñoz, "Control, Management, and Orchestration of Optical Networks: Evolution, Trends, and Challenges". IEEE J. Light. Technol., vol. 36, pp.1390–1402. 2018

[Ca19] R.R Cai, H.W. Zhang, H.L. Bu, and Y. Zhang, "Research on variation of runoff and sediment load based on the combination patterns in the middle and lower yellow river (Chinese)," Shuili Xuebao, vol. 5, no. 6, pp.732-742, Jun. 2019.

[Ch23] G. Chopra," An Efficient Base Station Sleeping Configuration for Ultra-dense Networks", in Proc. International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp.1-5, 01-03 Mar. 2023

[Co13] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, "Radio Access Network Virtualization for Future Mobile Carrier Networks", IEEE Communications Magazine, vol.51, no.7, pp.27-35, July, 2013.

[Da14] S. Dawoud et al., "Optimizing the power consumption of mobile networks based on traffic prediction," in Proc. IEEE COMPSAC'14, Vasteras, Sweden, Jul. 2014.

[Da22] M. S. Dahal, "Energy saving in 5G mobile communication through traffic driven cell zooming strategy," Energy Nexus, vol. 5, pp. 1-13, 2022

[Ef22] F. Effenberger et al., "Fixed 5th Generation Advanced and Beyond," ETSI White Paper, no. 50, 2022.

[Er21] C. C. Erazo-Agredo, M. Garza-Fabre, R. A. Calvo, L. Diez, J. Serrat, and J. Rubio-Loyola, "Joint route selection and split-level management for 5G C-RAN," IEEE Trans. Netw. Service Manag., vol. 18, no. 4, pp. 4616–4638, Dec. 2021.

[Er22] Ericsson Mobility Report 2022. Available online: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2022

[Et22] "Fixed 5th generation advanced and beyond," ETSI White Paper, version 1.0, Sep. 2022.

[Fe18] R. Ferrus, O. Sallent, J. Perez-Romero and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," in IEEE Communications Magazine, vol. PP, no. 99, pp. 2-10, January 2018.

[Fu18] Fujimoto et al., "Addressing Function Approximation Error in Actor-Critic Methods," in Proc. ICML, pp.1-10, 2018.

[Fu20] Y. Fu *et al.*, "End-to-End Energy Efficiency Evaluation for B5G Ultra Dense Networks," in Proc. IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium pp.1-6, 25-28 May 2020.

[Ga21] A. Gavras, Ö. Bulakci, M. Gramaglia, M. Iordache, M. Ghoraishi, A. Garcia, T. Cogalan, J. Gutiérrez, A. Tzanakaki, D. Warren, X. Li, G. Landi, J. Mangues, K. Tsagkaris, V. Frascolla, and H. Lee, "5G PPP Architecture Working Group - View on 5G Architecture," Version 4.0, 2021.

[Ga22] Z.G. Gao, S.Y. Yan, J.W. Zhang, B.T. Han, Y.C. Wang, Y.M. Xiao," Deep Reinforcement Learning-Based Policy for Baseband Function Placement and Routing of RAN in 5G and Beyond", Journal of Lightwave Technology., vol. 40, no. 2, pp. 470–480, Jan. 2022.

[Gh20] S. Ghosh, D. De, P. Deb, A. Mukherjee, "5G-ZOOM-Game: small cell zooming using weighted majority cooperative game for energy-efficient 5 G mobile network," Wireless Networks, vol. 26, pp. 349–372, 2020.

[Gu15] A. Gupta, R. KuMmar Jha, "A survey of 5G Network: Architecture and Emerging Technologies", IEEE Access, vol.3, pp.1206-1232, July 2015,

[Gu22] H. Gupta, A. A. Franklin, M. Kumar, B. R. Tamma," Traffic-Aware Dynamic Functional Split for 5G Cloud Radio Access Networks", In Proceedings of the 2022 International Conference on Network Softwarization (NetSoft), Milan, Italy, 27 Jun. - 1 Jul. 2022, pp. 297–301.

[Ho22] M. Hosseini, J.Pedro, A.Napoli, N.Costa, J.Prilepsky and S.Turitsyn,"Optimization of survivable filterless optical networks exploiting digital subcarrier multiplexing," IEEE/OSA Journal Optical Communications and Networking,vol.14,pp.586-594,2022

[Ho23] M. Hoffmann, S. Janji, A. Samorzewski, Ł. Kułacz, C. Adamczyk, M. Dryjański, P. Kryszkiewicz, A. Kliks, H. Bogucka," Open RAN xApps Design and Evaluation: Lessons Learnt and Identified Challenges," IEEE Journal on Selected Areas in Communications, vol.42(2), pp.1-14,2023

[Hs22] Y. H. Hsu, W. J. Liao, "eMBB and URLLC Service Multiplexing Based on Deep Reinforcement Learning in 5G and Beyond," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10-13 Apr. 2022, pp1467-1472.

[It15] ITU. Framework and overall objectives of the future development of IMT for 2020 and beyond：ITU-RM.2083 IMT Vision. Geneva：ITU-R,2015.

[Kh18] B. Khodapanah, A. Awada, I. Viering, D. Ohmann, M. Simsek, Gerhard P. Fettweis. "Fulfillment of Service Level Agreements via Slice-Aware Radio Resource Management in 5G Networks". in Proc. IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, Portugal, 03-06 June 2018, pp.1-6

[Ko07] T. Kohonen and T. Honkela, "Kohonen network," Scholarpedia, vol.2, no.1, pp.1568, Oct. 2007.

[Ko11] R. Kokku, R. Mahindra, H.H Zhang, and S. Rangarajan," NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks", IEEE/ACM Transactions on Networking, vol. 20(5), pp.1333-1346, Oct.2012

[Ko22] N. Koursioumpas, S. Barmpounakis, I. Stavrakakis, N. Alonistioti, "AI-driven, Context-Aware Profiling for 5G and Beyond Networks". *IEEE Trans. Netw. Serv. Manag.* vol.19(2), pp.1036–1048, Jun. 2022.

[La18] Larsen, L., Checko, A., Christiansen, H.L. A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks. IEEE Commun. Surv. Tutor. 2018, 21, 146–172.

[La21] S. Lagén et al., "Modulation Compression in Next Generation RAN: Air Interface and Fronthaul Trade-offs," IEEE Comm. Mag., vol. 59, pp. 89-95, Jan.2021.

[La23] Larrabeiti, D.; Contreras, L.M.; Otero, G.; Hernández, J.A.; Palacios, J.P.F. "Toward end-to-end latency management of 5G network slicing and fronthaul traffic". Elsevier Opt. Fiber Technol. vol.76, pp.1-9, Mar.2023,

[Li02] S.H. Liu," Latest Development of Optical Communication Technology", Optoelectronic Technology & Information, pp.1-8, 2002

[Li12] R.R Lian, H. Tian, We. C. Fei, J. Miao and C. Wang, "QoS-Aware Load Balancing Algorithm for Joint Group Call Admission Control in Heterogeneous Networks", in Proc. IEEE 75th Vehicular Technology Conference (VTC Spring), Yokohama, Japan, 06-09 May,2012, pp1-5.

[Li15] C. Liang, F. R. Yu, "Wireless Network Virtualization: A survey, some research issues and challenges", IEEE Communications Surveys and Tutorials, vol. 17, no. 1, pp. 358 – 380,2015.

[Li19] X. Liu, G. Chuai, W.D. Gao, K.S. Zhang, and X.Y. Chen, "KQIs-driven QoE anomaly detection and root cause analysis in cellular networks," in *Proc. 2019 IEEE Global Commun. Workshops*, Waikoloa, Hi, USA, 9-13 Dec. 2019, pp.1-6

[Li23] Y. Liming and L. Tianyao," A K-means Optimized Self-Organizing Map Neural Network Video Recommendation Framework," in Proc. IEEE Int. Conf. on Electronic Technology, Communication and Information (ICETCI), Changchun, China,26-28 May. 2023, pp.176-181

[Ma13] R. Mahindra, M. Khojastepour, H. Zhang, S. Rangarajan, "Radio Access Networks Sharing in Cellular Networks", in Proc. 21st IEEE International Conference on Network Protocols (ICNP), Goettingen, 07-10 Oct.2013, pp.1-10

[Mo22] F. Z. Morais, G. M. F. De Almeida, L. L. Pinto, K. Cardoso, L. M. Contreras, R. D. R. Righi, and C. B. Both, "Place RAN: Optimal placement of virtualized

network functions in beyond 5G radio access networks," IEEE Trans. Mobile Comput., vol.22(9), Apr. 29, 2022, pp. 5434 – 5448.

[Mo23] E. Moro, G. Gemmi, M. Polese, L. Maccari, A. Capone, T. Melodia, "Toward Open Integrated Access and Backhaul with O-RAN". in Proc.21st Mediterranean Communication and Computer Networking Conference (MedComNet), Island of Ponza, Italy, 13–15 Jun. 2023, pp.61-69

[Mu21] F. W. Murti, J. A. Ayala-Romero, A. Garcia-Saavedra, X. Costa-Pérez, G. Iosifidis," An Optimal Deployment Framework for Multi-Cloud Virtualized Radio Access Networks", IEEE Trans. Wireless Commun., vol. 20, no. 4, pp. 2251–2265, Apr. 2021.

[Mu22] F. W. Murti, S. Ali, and M. Latva-Aho, "Constrained deep reinforcement based functional split optimization in virtualized RANs," IEEE Trans. Wireless Commun., vol. 21, no. 11, pp. 9850–9864, Nov. 2022

[Nd23] A. Ndao, X. Lagrange, N. Huin, G. Texier, L. Nuaymi, "Optimal placement of virtualized DUs in O-RAN architecture". in Proc. of the Vehicular Technology Conference (VTC), Hong Kong, China, 10–13 Oct. 2023, pp.1-6.

[Ng15] NGMN Alliance, "5G White Paper", February, 2015.

[Oh13] E. Oh, K. Son and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," IEEE Trans. on Wireless Commun., vol. 12, no. 5, pp. 2126-2136, May 2013.

[Oh17] E. Oh and K. Son, "A Unified Base Station Switching Framework Considering Both Uplink and Downlink Traffic," IEEE Wireless Commun. Lett., vol. 6, no. 1, pp. 30-33, Feb. 2017.

[Oj19] B. Ojaghi, F. Adelantado, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "Sliced-RAN: Joint slicing and functional split in future 5G radio access networks," in Proc. IEEE Int. Conf. Commun. (ICC), Shanghai, China,20-24 May 2019, pp. 1–6.

[Oj22] B. Ojaghi, F. Adelantado, A. Antonopoulos, and C. Verikoukis, "Sliced RAN: Service-aware network slicing framework for 5G radio access networks," IEEE Syst. J., vol. 16, no. 2, pp. 2556–2567, Jun. 2022.

[Oj23] B. Ojaghi, F. Adelantado, C. Verikoukis, "On the Benefits of vDU Standardization in Softwarized NG-RAN: Enabling Technologies, Challenges, and Opportunities." *IEEE Commun. Mag.* vol.61, pp.92–98,2023.

[Or18] ORAN Alliance White paper：O-RAN：towards an open and smart RAN,2018

[Pa18] N. Patriciello et al., "5G new radio numerologies and their impact on the end-to-end latency," in Proc. IEEE CAMAD, Barcelona, Spain, 17-19, Sep.2018, pp.1-6

[Pa19] N. Patriciello et al., “An E2E Simulator for 5G NR Networks,” Elsevier Simulation Modelling Practice and Theory, vol.96, pp.1-3 2019.

[Pe13] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," in IEEE Communications Magazine, vol. 51, no. 7, pp. 27-35, July 2013

[Pe17] J. Pérez-Romero, O. Sallent, R. Ferrus, R. Agustí, “Admission Control for Multi-tenant Radio Access Networks”, in Proc. IEEE International Conference on Communications (ICC) - workshop on Smart Communication Protocols and Algorithms, Paris, France ,21-25 May, 2017, pp.1073-1078.

[Pe18] G. Perez, J. Hernandez, and D. Larrabeiti, “Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G”, IEEE/OPTICA J. of Optical Communications and Networking, vol. 10, pp. 573-581,2018.

[Po23] F. Posch; A. Fakhreddine; E. Caballero; C. Bettstetter,” A Classifier for Aerial Users in 5G Networks”, in Proc 2023 IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, Malaysia, 04-08 Dec. 2023, pp.775-780

[Qu21] S.H. Qu, Z.J. Guo, L.F. Xu. “Practice and verification of 5G technology in smart grid.” Distribution & Utilization, vol.38(5). pp.2-9, 2021

[Ra17] T. Rahman, Flexible and high data-rate coherent optical transceivers. Diss. Ph.D. Thesis. Technische Universiteit Eindhoven,2017

[Ra18] D. Rafique and L. Velasco,” Machine Learning for Optical Network Automation: Overview, Architecture and Applications,” IEEE/OSA Journal of Optical Communications and Networking, vol.10, pp. D126-D143,2018

[Ra22] O-RAN ALLIANCE WG3, “Near-real-time ran intelligent controller near-rt ric architecture, v.2.1,” Tech. Rep., 03 2022.

[Ra23] O-RAN ALLIANCE WG1, “O-ran architecture description, v.8.0,” Tech. Rep., 03 2023.

[Ra24] O-RAN Alliance. Available online: https://www.o-ran.org/, Access on 2024

[Ro20] V. Q. Rodriguez, F. Guillemin, A. Ferrieux and L. Thomas, “Cloud-RAN functional split for an efficient fronthaul network,” in Proc. International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus ,15-19Jun.2020, pp.245-250.

[Ru18] M. Ruiz et al., “CURSA-SQ: A Methodology for Service-Centric Traffic Flow Analysis,” IEEE/OSA J. of Optical Communications and Networking, vol. 10, pp. 773-784, 2018.

[Ru20] M.Ruiz, F. Tabatabaeimehr, L. Velasco, “Knowledge Management in Optical Networks: Architecture, Methods and Use Cases [Invited]. “IEEE/OSA J. Opt. Commun. Netw., vol.12, pp. A70–A81, 2020.

[Ru23] M. Ruiz, J. A. Hernandez, M. Quagliotti, E. Hugues-Salas, E. Riccardi, A. Rafel, L. Velasco, and O. Gonzalez De Dios, “Network Traffic Analysis under Emerging Beyond-5G Scenarios for Multi-Band Optical Technology Adoption,” IEEE/OPTICA Journal of Optical Communications and Networking (JOCN), vol. 15(11), pp. F36-F47,2023.

[Sa16] K. Samdanis, Xavier Costa-Pérez, Vincenzo Sciancalepore “From Network Sharing to Multi-tenancy: The 5G Network Slice Broker”. IEEE Communications Magazine, vol.54(7), pp. 32-39, July 2016.

[Sa21.1] E. Sarikaya and E. Onur, “Placement of 5G RAN slices in multi-tier O-RAN 5G networks with flexible functional splits,” in Proc. 17th Int. Conf. Netw. Service Manag. (CNSM), Izmir, Turkey, 25-29 Oct.2021, pp. 274–282.

[Sa21.2] A. Salh, L. Audah, N.S.M. Shah, A. Alhammadi, Q. Abdullah, Y.H. Kim, S.A. Al-Gailani, S.A. Hamzah, B.A.F. Esmail, A.A. Almohammedi, “A Survey on Deep Learning for Ultra-Reliable and Low-Latency Communications Challenges on 6G Wireless Systems”. IEEE Access, vol. 9, pp. 55098–55131,2021.

[Se23] N. Sen, A. Franklin A,” Slice Aware Baseband Function Placement in 5G RAN Using Functional and Traffic Split”, IEEE Access., vol.11, pp. 35556–35566, Apr. 2023

[Si16] I. Silva, G. Mildh, A. Kaloxylos, P. Spapis, E. B. Alessandro, T. G. Zimmermann, N. Bayer,” Impact of network slicing on 5G Radio Access Networks”, in Proc. European Conference on Networks and Communications (EuCNC), Athens, Greece, 27-30 Jun.2016, pp.153-157.

[Si23] H. Singh, C.S. Rai, B.V. Ramana Reddy,” Analysis and Classification of Vertical Handover Decisions in Small Cell Scenarios Deploying Heterogeneous Wireless Networks”, in Proc 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 01-03 Nov. 2023,pp.589-595.

[Su20] H. Sun et al.,”800G DSP ASIC Design Using Probabilistic Shaping and Digital Sub-Carrier Multiplexing,” IEEE/OSA Journal of Lightwave Technology, vol. 308, pp.4744-4756,2020

[Su22] M. Sulaiman, A. Moayyedi, M. A. Salahuddin, R. Boutaba, and A. Saleh, “Multi-agent deep reinforcement learning for slicing and admission control in 5G C-RAN,” in Proc. IEEE/IFIP Network Operations and Management Symposium, 2022.

[Te17] Technical Specification Group Radio Access Network: Study on new radio access technology: Radio access architecture and interfaces (Release 14), White paper,3GPP TR 38.801 V14.0.0, Mar.2017.

[Ul90] A. Ultsch and HP. Siemon, "Kohonen's self-organizing feature maps for exploratory data analysis," in Proc. Int. Conf. Neur. Netw., Dordrecht, Netherlands, Kluwer, 9-13 Jul.1990, pp. 305–308.

[Uu21] M. Uusitalo, P. Rugeland, M. Boldi, E. Strinati, P. Demestichas, M. Ericson, G. Fettweis, M. Filippou, A.Gati, M. Hamon, et al. "6G Vision, Value, Use Cases and Technologies from European 6G Flagship Project Hexa-X". IEEE Access, vol 9, pp.160004–160020. 2021

[Ve13] L. Velasco, P. Wright, A. Lord, and G. Junyent, "Saving CAPEX by Extending Flexgrid-based Core Optical Networks towards the Edges," IEEE/OSA Journal of Optical Communications and Networking, vol. 5, pp. A171-A183, 2013.

[Ve23] L. Velasco, P. González, M. Ruiz, "Distributed Intelligence for Pervasive Optical Network Telemetry." IEEE/OPTICA J. Opt. Commun. Netw. (JOCN), vol.15, pp.676–686,2023.

[Ve23.2] L. Velasco, M. Ruiz, P. Gonzalez, F. Paolucci, A. Sgambelluri, L. Valcarenghi, C. Papagianni, "Pervasive Monitoring and Distributed Intelligence for 6G Near Real-Time Operation," in Proc. European Conference on Networks and Communications (EuCNC), 2023.

[Vu18] T. T. Vu, D. T. Ngo, M. N. Dao, S. Durrani, D. H. Nguyen, and R. H.Middleton, "Energy efficiency maximization for downlink cloud radio access networks with data sharing and data compression," IEEE Transaction Wireless Communication, vol. 17, pp. 4955–4970, 2018.

[Ve21.1] L. Velasco *et al.*, "End-to-End Intent-Based Networking," IEEE Communications Magazine, vol. 59, pp. 106-112, 2021.

[Ve21.2] L. Velasco *et al.*, "Autonomous and Energy Efficient Lightpath Operation Based on Digital Subcarrier Multiplexing," IEEE Journal on Selected Areas in Communications, vol. 39, pp. 2864-2877, 2021.

[Wa21] S. Wang, R. Ferrus," Extracting Cell Patterns from High-dimensional Radio Network Performance Datasets using Self-Organizing Maps and K-means Clustering", IEEE Access, vol.9, pp.2169-3536, 2021.

[Wa22] R.K. Wang, J.W. Zhang, Z.Q. Gu, S.Y. Yan, Y.M. Xiao, Y.F. Ji," Edge-enhanced graph neural network for DU-CU placement and lightpath provision in X-Haul networks", Journal of Optical Communications and Networking., vol. 14, no. 10, pp. 828–839, Oct. 2022.

[Wa23] S. Wang, M. Ruiz, L. Velasco, "Optical network traffic analysis under B5G/6G RAN operation". In Proc. IEEE International Conference on Transparent Optical Networks (ICTON), Bucharest, Romania, 2–6 Jul. 2023, pp.1-4.

[Wa24] S. Wang, M. Ruiz, and L. Velasco, "Context-based e2e Autonomous Operation in B5G Networks," MDPI Sensors, vol. 24, pp. 1-25, 2024.

[We21] D. Welch, A. Napoli, and J. Bäck, "Point-to-Multipoint Optical Networks Using Coherent Digital Subcarriers," IEEE Journal of Lightwave Technology, vol. 39, 2021.

[Wu23] Y.F. Wu, L. Liang, Y.J. Jia, Z.C. Chen, and W.L. Wen," Slicing Enabled Flexible Functional Split and Resource Provisioning in 5G-and-Beyond RAN", in Proc. IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, United Kingdom, 26 – 29 Mar. 2023; pp. 1–6.

[Xi20] X. Xia, X. Yuan, Y. Liang, et al. "Research on 5G network slicing enabling the smart grid" Application of Electronic Technique, 46(1). pp.17-21, 2020.

[Zh15] J. Zheng, Y. Cai, X. Chen, R. Li, and H. Zhang, "Optimal Base Station Sleeping in Green Cellular Networks: A Distributed Cooperative Framework Based on Game Theory," IEEE Trans. on Wireless Commun., vol. 14, no. 8, pp. 4391-4406, Aug. 2015.

[Zh17] F. Zhang, J. Wang, X.P. Wang, "Recognition of spatial framework for water quality and its relation with land use/cover types from a new perspective: A case study of Jinghe Oasis in Xinjiang, China," Natural Hazards and Earth Syst. Sci., vol. 358, pp. 1-18, Oct. 2017.

[Zh17] Z. Zhang, F. Liu, Z. Zeng, "The cell zooming algorithm for energy efficiency optimization in heterogeneous cellular network," in Proc. 9th International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China ,11-13 Oct. 2017, pp.1-5.

[Zh19] Y. Zhou, J. Li, Y. M. Shi, and V. W. S. Wong, "Flexible Functional Split Design for Downlink C-RAN with Capacity-Constrained Fronthaul," IEEE Transactions on Vehicular Technology, vol. 68, pp.6050-6063, 2019.

[Zh21] Y. Zhu and S. Wang, "Joint Traffic Prediction and Base Station Sleeping for Energy Saving in Cellular Networks," in Proc. IEEE Int. Conf. on Communications (ICC), 2021.

[Zo22] L.M.M. Zorello, M. Sodano, S. Troia, G. Maier, "Power-Efficient Baseband-Function Placement in Latency-Constrained 5G Metro Access." IEEE Trans. Green Commun. Netw. vol.6, pp.1683–1696,2022.