UNIVERSIDADE DO VALE DO RIO DOS SINOS — UNISINOS
ACADEMIC UNIT OF RESEARCH AND POSTGRADUATE STUDIES
POSTGRADUATE PROGRAM IN APPLIED COMPUTING

GUSTAVO ZANATTA BRUNO

ADAPTIVE NETWORK MANAGEMENT IN 6G O-RAN: A FRAMEWORK FOR DYNAMIC USER DEMANDS

São Leopoldo
2024

Gustavo Zanatta Bruno

ADAPTIVE NETWORK MANAGEMENT IN 6G O-RAN:
A Framework for Dynamic User Demands

A thesis presented as a partial requirement to obtain the Doctor's degree from the Postgraduate Program in Applied Computation of the University of Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Dr. Cristiano Bonato Both

Co-advisor:
Prof. Dr. Alexandre Huff

São Leopoldo
2024

For everyone who supports me on this journey.

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor, Dr. Cristiano Bonato Both, for his guidance, patience, and continuous support throughout the research and writing process. His expertise and advice were fundamental to the completion of this work. I am especially grateful for the great opportunities he provided me, which significantly enriched my academic and professional experience.

I am deeply grateful to Kleber Vieira Cardoso, a research partner of my advisor, for his significant help and contributions to my work.

I also sincerely thank my co-advisor, Dr. Alexandre Huff, for his invaluable insights and support.

I am grateful to my research colleagues, especially Gabriel Matheus F. de Almeida, whose collaboration and support were paramount. I also thank Karlla Chaves Rodrigues, Fernando Zanferrari Morais, Thiago Becker, Julio Renner, and Vikas Krishnan Radhakrishnan for their assistance and camaraderie.

I am grateful to UNISINOS for providing me with the opportunity to do this PhD. I also thank the funding agencies RNP and FAPESP, whose financial support was crucial for conducting this research.

I want to thank professors Luiz DaSilva, João Santos, and Aloizio da Silva from CCI/VA/USA, who contributed to my education and intellectual growth throughout my academic journey.

Finally, I thank the research participants for their willingness and cooperation, essential for this study's success.

Thank you all very much.

"*Who you least expect can be your best teacher.*"

ABSTRACT

The advent of Sixth Generation (6G) mobile networks heralds a transformative era in wireless communication, demanding unprecedented adaptability and energy efficiency to meet burgeoning dynamic user demands. This thesis introduces an innovative framework for adaptive network management within the Open Radio Access Network (O-RAN) paradigm, focusing on the integration of its dynamic architecture. The framework emphasizes the exploration of integration components to manage and optimize the Radio Access Network (RAN). Central to this framework are the architectural components of O-RAN: the Service Management and Orchestration (SMO), the Near-Real-Time RAN Intelligent Controller (Near-RT RIC), and the Non-Real-Time RAN Intelligent Controller (Non-RT RIC), each playing a crucial role in enhancing network adaptability and operational efficiency.

The proposed framework leverages the open and intelligent architecture of O-RAN, deploying various applications such as the rApp Energy Savings to optimize energy consumption and data flow management. It aims to establish a new benchmark for network management in the 6G era by integrating real-time data analytics, intelligent policy implementation, and adaptive energy management strategies. While energy savings is a primary use case for validating our dynamic architecture, it is important to note that the framework's flexibility allows for the integration of other applications as well. The dynamic clustering mechanism for radio nodes, coupled with the RIC, facilitates efficient resource management by adjusting network configurations based on current and predicted traffic loads, significantly improving resource utilization and energy efficiency.

A prototype implementation validates the framework under various network conditions and user demands, demonstrating substantial improvements in resource utilization and energy efficiency compared to traditional static network management approaches. The adaptive framework significantly reduces energy consumption during low-demand periods while maintaining high performance during peak times.

The effectiveness of the proposed framework is demonstrated through its dynamic management of network resources, facilitated by specialized rApps and xApps. These applications interface with the SMO, Near-RT RIC, and Non-RT RIC, contributing to substantial improvements in network performance. Key components such as the modified VespaMgr and A1 Mediator within the Near-RT RIC, along with custom SMO elements, ensure seamless communication and integration across the network. This dynamic architecture has shown significant enhancements in resource utilization and energy efficiency. The results indicate efficient data handling and communication through the O1 interface for VES, validating the framework's capability to adapt to varying network conditions and demands, thereby establishing new benchmarks in 6G network management.

Keywords: Adaptive Network Management. Energy Efficiency. O-RAN. RIC. RAN. SMO. 6G.

RESUMO

O advento das redes móveis de Sexta Geração (6G) anuncia uma era transformadora na comunicação sem fio, exigindo uma adaptabilidade e eficiência energética sem precedentes para atender às crescentes demandas dinâmicas dos usuários. Esta tese apresenta um quadro inovador para a gestão adaptativa de redes dentro do paradigma da Open Radio Access Network (O-RAN), focando na integração de sua arquitetura dinâmica. O quadro enfatiza a exploração de componentes de integração para gerenciar e otimizar a Radio Access Network (RAN). Centrais a este quadro são os componentes arquitetônicos do O-RAN: a Gestão de Serviços e Orquestração (SMO), o Controlador Inteligente de RAN em Tempo Real Próximo (Near-RT RIC) e o Controlador Inteligente de RAN em Tempo Não Real (Non-RT RIC), cada um desempenhando um papel crucial no aprimoramento da adaptabilidade e eficiência operacional da rede.

A solução proposta aproveita a arquitetura aberta e inteligente do O-RAN, implantando várias aplicações como o rApp Energy Savings para otimizar o consumo de energia e a gestão do fluxo de dados. Ele visa estabelecer um novo padrão para a gestão de redes na era 6G, integrando análises de dados em tempo real, implementação inteligente de políticas e estratégias adaptativas de gestão de energia. Embora a economia de energia seja um caso de uso principal para validar nossa arquitetura dinâmica, é importante notar que a flexibilidade do quadro permite a integração de outras aplicações também. O mecanismo de agrupamento dinâmico para nós de rádio, juntamente com o RIC, facilita a gestão eficiente de recursos ajustando as configurações da rede com base nas cargas de tráfego atuais e previstas, melhorando significativamente a utilização de recursos e a eficiência energética.

Uma implementação protótipo valida o quadro sob várias condições de rede e demandas dos usuários, demonstrando melhorias substanciais na utilização de recursos e na eficiência energética em comparação com abordagens tradicionais de gestão de redes estáticas. O quadro adaptativo reduz significativamente o consumo de energia durante períodos de baixa demanda, mantendo alto desempenho durante os picos.

A eficácia do quadro proposto é demonstrada através da gestão dinâmica dos recursos da rede, facilitada por rApps e xApps especializados. Essas aplicações interagem com o SMO, Near-RT RIC e Non-RT RIC, contribuindo para melhorias substanciais no desempenho da rede. Componentes chave como o VespaMgr modificado e o Mediador A1 dentro do Near-RT RIC, juntamente com elementos SMO personalizados, garantem comunicação e integração contínuas em toda a rede. Esta arquitetura dinâmica mostrou melhorias significativas na utilização de recursos e na eficiência energética. Os resultados indicam um manuseio eficiente de dados e comunicação através da interface O1 para VES, validando a capacidade do quadro de se adaptar a condições e demandas variáveis da rede, estabelecendo novos padrões na gestão de redes 6G.

Palavras-chave: Gerenciamento Adaptativo de Rede. Eficiência Energética. O-RAN. RIC. RAN. SMO. 6G.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3GPP 3rd Generation Partnership Project

5G Fifth Generation

6G Sixth Generation

A1 A1 Interface

AI Artificial Intelligence

ARQ Automatic Repeat Request

API Application Programming Interface

B5G Beyond 5G

BBU Baseband Unit

BH Backhaul

BLER Block Error Rate

BS Base Stations

CNN Convolutional Neural Networks

COTS Commercial Off-The-Shelf

CQI Channel Quality Index

CP Control Plane

CPU Central Processing Unit

CU Central Units

CWDM Coarse Wavelength Division Multiplexing

C-RAN Cloud Radio Access Network

CN Core Network

DBaaS Database as a Service

DCDPandUA Dynamic CU and DU Placement and User Association

DRL Deep Reinforced Learning

Dis Disaggregated

Dis-RIC Disaggregated near-RT RAN Intelligent Controller

DL Downlink

DP Data Plane

DRAandNFP Dynamic Resource Allocation and Network Function Placement

D-RAN Distributed Radio Access Network

DU Distributed Unit

DU Distributed Units

E2 E2 Interface

E2AP E2 Application Protocol

E2N E2 Node

E2Sim E2 Simulator

E2T E2 Termination

EE Energy Efficiency

EFK Elasticsearch, Fluentd and, Kibana

eNB Evolved Node B

EPC Evolved Packet Core

ETSI European Telecommunications Standards Institute

FCAPS Fault, Configuration, Accounting, Performance, and Security

FH Fronthaul

FL Federated Learning

FlexRIC Flexible RAN Intelligent Controller

gNB Next Generation Node B

gNB-CU Next-Generation NodeB - Central Unit

gNB-DU Next-Generation NodeB - Distributed Unit

GoB Grid of Beams

GPP General-Purpose Processor

HARQ Hybrid Automatic Repeat Request

HTTP Hypertext Transfer Protocol

IIoT Industrial Internet of Things

IL Internal layer

INFOCOM IEEE International Conference on Computer Communications

IOEE Indirect Optimization of Energy Efficiency

ITU-T International Telecommunication Union - Telecommunication Standardization

JISA Journal of Internet Services and Applications

JSON JavaScript Object Notation

JSAC IEEE Journal on Selected Areas in Communications

K8s Kubernetes

KPI Key Performance Indicator

LTE Long Term Evolution

MAC Media Access Control

MC Monolithic Co-located

MC-RIC Monolithic Co-located near-RT RAN Intelligent Controller

MD Monolithic Distributed

MD-RIC Monolithic Distributed near-RT RAN Intelligent Controller

MEC Multi-access Edge Computing

MILP Mixed-Integer Linear Programming

MH Midhaul

MIMO Multiple-Input, Multiple-Output

ML Machine Learning

ML Metropolitan layer

mMIMO Massive MIMO

MRO Mobility Robustness Optimization

Near-RT Near Real-Time

Near-RT RIC Near Real-Time RAN Intelligent Controller

NR New Radio

NFV Network Function Virtualization

NFVI Network Function Virtualization Infrastructure

NG-Core Next Generation Core

ng-eNB Next-Generation - Evolved NodeB

NIC Network Interface Card

NIB Network Information Base

Non-RT Non Real-Time

Non-RT RIC Non Real-Time RAN Intelligent Controller

NSI Network Slice Instance

NSSI Network Slice Subnet Instance

O1 O1 Interface

O2 O2 Interface

O-CU Open Central Unit

O-DU Open Distributed Unit

O-Cloud Open Cloud

OPlaceRAN Orchestrator Placement for Radio Access Network

O-RAN Open Radio Access Network

O-RAN SMO Open RAN Service Management and Orchestration

OSC Open Radio Access Network Software Community

O-RU Open Radio Unit

P2P Point-To-Point

PDCP Packet Data Control Protocol

PHY Physical Layer

PON Passive Optical Network

QAM Quadrature Amplitude Modulation

QoE Quality of Experience

QoS Quality of Service

RAM Random Access Memory

rApp Application on Non-Real-Time RAN Intelligent Controller

RAN Radio Access Network

REST Representational State Transfer

RF Radio Frequency

RFT RAN Intelligent Control Fault Tolerance

RIC RAN Intelligent Controller

RIC-O RAN Intelligent Controller Orchestrator

RLC Radio Link Control

RMR RAN Intelligent Controller Message Router

RRC Radio Resource Control

RRH Remote Radio Head

RRM Radio Resource Management

RSRP Reference Signal Received Power

RSRQ Reference Signal Received Quality

RSSI Received Signal Strength Indicator

RU Radio Units

SCA Successive Convex Approximation

SCTP Stream Control Transmission Protocol

SDL Shared Data Layer

SDN Software-Defined Networking

SDR Software Defined Radio

SDAP Service Data Adaptation Protocol

SD-RAN Software-Defined Radio Access Network

SBRC Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos

SBRT Simpósio Brasileiro de Telecomunicações

SINR Signal-to-Interference-plus-Noise Ratio

SLA Service Level Agreement

SM Service Model

SMO Service Management and Orchestration

SNR Signal-to-Noise Ratio

srsRAN Software Radio Systems Radio Access Network

STSL Shared Time-Series Layer

TDR Time to Disaster Recovery

TNSM IEEE Transactions on Network and Service Management

UAV Unmanned Aerial Vehicle

UE User Equipment

UL Uplink

UP User Plane

URL Uniform Resource Locator

USRP Universal Software Radio Peripheral

VES VNF Event Streaming

VESPA VES Prometheus Adapter

vCPU Virtual Central Processing Unit

vCU Virtualized Central Unit

vDU Virtualized Distributed Unit

vNG-RAN virtualized Next Generation Radio Access Network

vRAN Virtualized Radio Access Network

vRU Virtualized Radio Unit

VM Virtual Machine

VNF Virtual Network Function

xApp Application on Near-Real-Time RAN Intelligent Controller

Y1 Y1 Interface

YAML YAML Ain't Markup Language

CONTENTS

1 INTRODUCTION

The advent of the Sixth Generation (6G) of mobile networks marks a pivotal moment in technological evolution, addressing modern society's dynamic demands and lifestyle changes. The architecture of these networks, encompassing the Core Network (CN), the Radio Access Network (RAN), and transport networks, is undergoing substantial enhancements. These improvements are essential to support increased data rates, reduced latencies, and more reliable connections, which are crucial for the expected surge in traffic and service demands. The core will manage connectivity and policies, the RAN will ensure the wireless connection of user equipment over a wide area through Base Stations (BS), and the transport networks will connect the CN and RAN, accommodating the advanced requirements of 6G (TATARIA et al., 2021; TOMKOS et al., 2020; KATZ; MATINMIKKO-BLUE; LATVA-AHO, 2018; SIDDIQI; YU; JOUNG, 2019).

A significant evolution within the RAN in 6G is the disaggregation of BS into Central Units (CU), Distributed Units (DU), and Radio Units (RU), leading to more flexible network configurations. This flexibility is key to optimizing energy use, enabling dynamic resource allocation based on demand. The transition towards virtualization of network functions (vRAN) is another critical development, replacing traditional hardware-dependent setups with Virtual Network Functions (VNFs) on a Network Function Virtualization Infrastructure (NFVI). This approach facilitates efficient resource sharing and scalable operations (JAAFARI; CHUBERRE, 2023; 3GPP, 2017a; ITU-T, 2017; 3GPP, 2019; TSUKAMOTO et al., 2019; YOUSAF et al., 2019; SABELLA et al., 2018).

The O-RAN RAN Intelligent Controller (RIC) represents a leap forward in network optimization, pivotal for adapting to the dynamic requirements of 6G. This controller consists of two critical components: the Near-RT RIC and the Non Real-Time RAN Intelligent Controller (Non-RT RIC). The Near-RT RIC, operating close to the network edge, swiftly adjusts resources in response to changing demands. At the same time, the Non-RT RIC focuses on longer-term network planning and policy management, including strategies for sustainable operations (ALAVIRAD et al., 2023; O-RAN Alliance, 2023a).

Integrating these two RIC components is crucial for achieving the dynamic performance envisioned for 6G networks. This dual-controller setup enables a network that is both responsive and adaptive, meeting immediate demands and evolving to accommodate long-term trends. It is designed to ensure high data rates, low latencies, reliable connections, and energy efficiency. One notable use case for this architecture is the optimization of energy usage, effectively managing resources to respond to varying user and application demands (O-RAN Alliance, 2023b,c).

1.1 Motivation

The transition towards 6G necessitates reimagining network management to meet the twin challenges of dynamic user demands and environmental sustainability. This calls for introducing a virtualized Next Generation Radio Access Network (vNG-RAN) and the advanced integration of RAN RICs, incorporating containerization technologies for enhanced flexibility and efficiency.

The vNG-RAN emerges as a critical element in this transformative journey, offering unparalleled resource allocation and functionality distribution agility. It is poised to address the challenges posed by varied network topologies, bandwidth fluctuations, and the highly dynamic nature of user traffic. Incorporating the Near-RT RIC within the vNG-RAN architecture is particularly notable. This integration supports adaptive network management, essential for effectively implementing 6G networks. The design of this system is focused on multiple use cases, including optimizing energy consumption and ensuring strategic placement of RAN components in real-time scenarios, thereby enabling efficient and sustainable network service delivery (DRYJAńSKI; KLIKS, 2022; D'ORO et al., 2022; BESHLEY et al., 2022).

Energy efficiency takes center stage in the vNG-RAN architecture. Our research aims to integrate hardware optimizations, software mechanisms, energy-efficient protocols, and system-wide enhancements to create sustainable and cost-effective network solutions. This approach aligns with the dynamic-aware characteristics of 6G technology, allowing for proactive adjustments in network operations to minimize energy consumption. We seek to harness the open architecture of O-RAN to improve the energy efficiency of RAN nodes, contributing to a telecommunications infrastructure that is both adaptive and environmentally sustainable (BESHLEY et al., 2022).

Lastly, the O-RAN specifications, particularly the E2 interface protocols (E2 Application Protocol (E2AP) and E2AP), enable communication between the Near-RT RIC and RAN nodes. This communication is key to controlling and optimizing network performance. Our research delves into the intricacies of these protocols to ensure that network orchestration aligns with the performance requirements of 6G networks. We aim to enhance the capabilities of the SMO framework and the Non-RT RIC, enabling them to work in tandem with the Near-RT RIC. This collaboration will ensure the network's adaptability to dynamic demands, contributing to the broader goal of a sustainable, efficient, and flexible 6G network architecture (O-RAN Alliance, 2023d).

In exploring the orchestration of RAN to efficiently address dynamic user demands, a key research question emerges: How can Open Radio Access Network (O-RAN) components be integrated to meet these fluctuating demands efficiently? This inquiry is pivotal in the evolution towards 6G networks. Innovations in resource allocation, data traffic management, and adaptive network orchestration, as highlighted in studies by (MUNGARI, 2021; VILA et al., 2022; KASULURU et al., 2023), provide a foundational understanding of dynamic resource

management. These approaches underline the importance of real-time distribution of network resources, such as bandwidth and power, to maintain optimal network performance amidst varying user demands. Furthermore, (ORHAN et al., 2021; YOO et al., 2022; KOUCHAKI et al., 2022) offer insights into effective data traffic management, which includes strategies such as traffic steering and intelligent handover, crucial for maintaining service quality across different network segments. The orchestration frameworks discussed by (D'ORO et al., 2022; BONATI et al., 2023) further emphasize the role of automated decision-making in resource allocation and traffic management, highlighting the dynamic adaptability essential in modern telecommunications networks.

Another research question of significance revolves around the most accurate use case that represents fluctuating user demand in 6G networks. This inquiry delves deep into the potential of O-RAN in addressing the challenges of future telecommunications networks. The advancements in Massive MIMO, network slicing, and energy-saving strategies, as delineated in the works of the O-RAN Working Group (O-RAN Alliance, 2023e,f,g), reflects the necessity of adapting to the dynamic nature of user demands expected in 6G. These documents detail key areas such as Grid of Beams optimization, adaptive beam shaping, and intelligent automation in RIC, crucial for energy conservation and efficient network operations. The focus on managing Network Slice Instances (Network Slice Instance (NSI)) and Network Slice Subnet Instances (Network Slice Subnet Instance (NSSI)) further underscores the evolving role of O-RAN in future networks. The research efforts summarized in these works provide a roadmap for understanding and addressing the dynamic requirements of next-generation networks, emphasizing the need for agility and sustainability in the face of evolving user demands.

Reflecting on the comprehensive insights from existing literature in the field of RAN, it becomes apparent that while there have been significant advancements, substantial areas still beckon further exploration, especially in the context of the emerging 6G networks. This realization foregrounds the importance of our research, which is positioned to probe into the uncharted territories of RAN orchestration, particularly emphasizing strategies that can dynamically adapt to the complexities inherent in 6G networks. Our study aims to delve deeper into these emerging challenges, addressing pivotal questions around the optimal integration and orchestration of O-RAN components, accurately representing fluctuating user demands in the 6G landscape, and effectively managing Near-RT RIC functions. This inquiry, outlined in the forthcoming section, is strategically positioned to contribute significantly to the evolving 6G network technology domain and its operational effectiveness.

1.2 Research Questions

The emergence of 6G in wireless communications marks a pivotal era defined by dual imperatives: achieving unprecedented performance and upholding sustainability. This evolution in network technology necessitates a critical examination and refinement of radio unit orches-

tration. Such orchestration must efficiently balance user and application demands with strict energy efficiency standards.

Research Question (RQ): How can RAN be orchestrated optimally to accommodate fluctuating user demands?

This central question directs our investigation toward the intersection of software-centric network design and the novel approaches introduced by the O-RAN alliance. Our focus is on the SMO framework and the collaborative operation of the Non-RT RIC and the Near-RT RIC. We aim to enhance the 6G network management, ensuring it is both robust, as directed by the strategic vision of the Non-RT RIC, and agile, as required by the operational flexibility of the Near-RT RIC.

Our exploration delves into the response times of these controllers, spanning milliseconds to seconds, examining their impact on network energy consumption and service quality. One significant use case we investigate is the optimization of network energy consumption and service quality through these rapid response times. To further focus our inquiry, we address the following sub-questions:

Our exploration delves into the response times of these controllers, spanning milliseconds to seconds, examining their impact on the use case that we investigate is the optimization of network energy consumption. To further focus our inquiry, we address the following sub-questions:

- **SRQ1**: How can components of an open radio access network be integrated to address dynamic user demands efficiently?
  - *Objective*: To develop integration strategies for O-RAN components that ensure seamless and efficient handling of dynamic user demands.
  - *Impact*: Improved network responsiveness and resource utilization.
  - *Amswer*: The efficient integration of O-RAN components to address dynamic user demands can be achieved through a multifaceted approach involving adaptive resource allocation, effective data traffic management, and intelligent orchestration. This involves the synergy of Radio Resource Management (RRM), user-cell association, load balancing, and the orchestration of Near-RT RIC, Non-RT RIC, and the SMO framework.
- **SRQ2**: What specific use case most accurately represents fluctuating user demand in 6G networks?
  - *Objective*: To identify and validate a representative use case that encapsulates the typical fluctuations in 6G network demand.
  - *Impact*: Enhanced relevance and applicability of research findings to real-world scenarios.

  - *Answer*: Energy-saving strategies within the O-RAN framework, such as Carrier and Cell Switch Off/On, RF Channel Reconfiguration, and Sleep Modes, most accurately represent fluctuating user demand in 6G networks. These strategies enable dynamic adaptation to changing user demands, optimizing energy efficiency and operational performance, which are crucial for managing the variability in user activity typical of 6G networks.

- **SRQ3**: What strategies can effectively disaggregate Near-RT RIC functions to manage fluctuating user demands?

  - *Objective*: To develop and evaluate strategies for disaggregating Near-RT RIC functions to improve dynamic demand management.
  - *Impact*: Increased flexibility and efficiency in real-time network management.
  - *Answer*: Effective strategies for disaggregating Near-RT RIC functions include implementing centralized and distributed models. These approaches ensure scalability, rapid adaptability, and efficient resource management in response to dynamic user demands in 6G networks (ALMEIDA et al., 2023).

- **SRQ4**: What criteria and methods effectively evaluate the proposed orchestration strategy in response to fluctuating user demands?

  - *Objective*: To establish robust evaluation criteria and methods for assessing the effectiveness of orchestration strategies.
  - *Impact*: Validated and optimized orchestration strategies for practical deployment.
  - *Answer*: Effective evaluation of the proposed orchestration strategy involves criteria such as scalability, flexibility, rapid adaptability to changing network conditions, and adherence to O-RAN specifications like throughput and resource utilization. Methods include using advanced simulation tools and real-time monitoring systems to track key performance indicators such as latency, throughput, and resource utilization. Integrating AI and ML algorithms enhances predictive capabilities and proactive adjustments, ensuring the orchestration strategy's responsiveness to user demands and network dynamics (ALMEIDA et al., 2023).

Our research aims to contribute significantly to developing efficient and sustainable 6G network management practices by addressing these questions.

1.3 Research Targets

Our research is centered on developing *"Adaptative Network Management in 6G O-RAN: A Framework for Dynamic User Demands"*, aiming to address the complex challenges in the

evolving landscape of wireless access networks. The essence of our work is to design a system that is flexible and responsive to varying user demands using energy efficiency and a use case. Integral to this system are various interconnected components and methodologies, each contributing to the overall efficacy of the framework.

A key feature of our proposed solution is its dynamic adaptability, which is crucial for maintaining uninterrupted service quality and continuity. It aims to provide robust, dynamic, real-time network management solutions. This management involves adjusting to high-demand situations, as depicted in Figure 1, and scaling down resources during low-demand periods, as shown in Figure 2. Such adaptability is pivotal in optimizing energy consumption while ensuring service reliability. One of the primary use cases for this adaptability is energy optimization, aligning with our commitment to energy efficiency and sustainability in mobile networks. The primary components and objectives of our solution include:

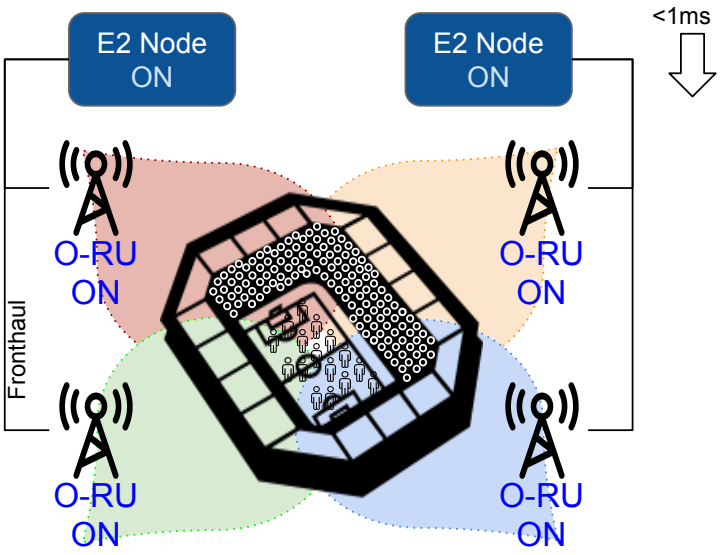


Figure 1 – High demand scenario.

- **Adaptative Network Management:** Development of an adaptive management framework aligned with O-RAN Alliance standards, involving a versatile orchestrator for managing Near-RT RIC functions. This framework dynamically allocates resources based on real-time demands.
- **Theoretical Framework:** Formulation of a comprehensive mathematical model for real-time optimization of energy consumption, ensuring the system's efficiency and sustainability. This model serves as the backbone for the adaptive management system.
- **Event-Driven Optimization Framework:** Creation of a specialized framework tailored for environments with fluctuating demands, such as large-scale events. This framework efficiently manages resource scaling and deploying Near-RT RIC instances to maintain service quality.

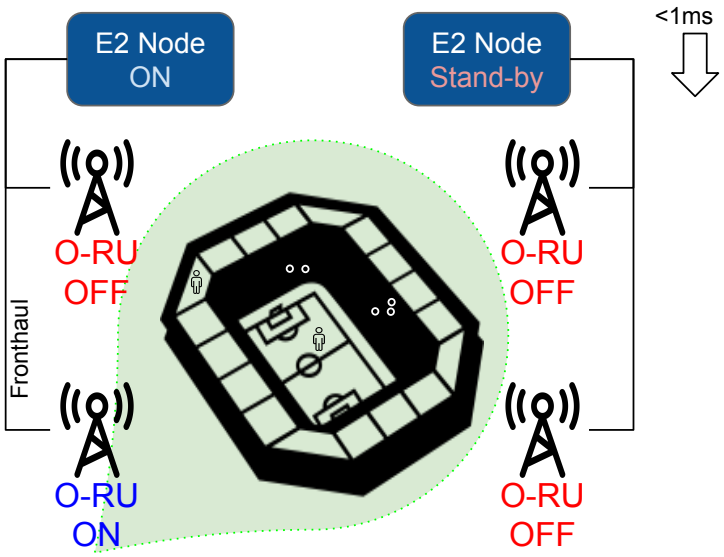


Figure 2 – Low demand scenario.

- **Architectural Components:** In-depth exploration and integration of key architectural components, including the SMO, Near-RT RIC, Non-RT RIC, and specialized applications like smoApps, rApps Energy Saver, and Application on Near-Real-Time RAN Intelligent Controllers (xApps). Each component is designed to interact seamlessly, enhancing the overall network performance.

- **Open-Source Contributions:** Sharing of developed software tools and datasets as open-source resources to encourage further research and innovation in 6G network technology. This initiative aims to foster a collaborative environment and accelerate advancements in the field.

Each component and objective within our research is intricately designed to address the multifaceted requirements of modern 6G wireless networks, emphasizing adaptive management and sustainable network operations. By focusing on these targets, we aim to develop a robust framework capable of meeting the dynamic demands of future network environments.

1.4 Text Organization

The remaining work is organized into five chapters. Initially, Chapter 2 introduces fundamental concepts of the evolution and virtualization of the RAN architecture to provide a comprehensive understanding of the rest of the work. Subsequently, Chapter 3 reviews related works relevant to the themes of this research, with a focus on their relationship to the research questions. This chapter aims to present the current state of the art and identify gaps and opportunities in the literature.

Chapter 4 presents the developed ***"Adaptive Network Management in 6G O-RAN: A***

Framework for Dynamic User Demands" solution and explains the design decisions made to address the gaps identified in the previous chapter and achieve the study's objectives. The evaluation methodology for the conceptual analysis of the mathematical formulation and the emulation of the experimental prototype is presented in Chapter 5.

The results obtained from the simulation and emulation are presented and discussed in Chapter 6. Finally, Chapter 7 provides the conclusion of the work and suggests possibilities for future research.

In summary, the structure of the thesis is as follows:

- **Chapter 2**: Introduction to fundamental concepts and the evolution of RAN architecture.
- **Chapter 3**: Review of related work and identification of literature gaps.
- **Chapter 4**: Presentation of the proposed framework and design decisions.
- **Chapter 5**: Methodology for evaluation and emulation.
- **Chapter 6**: Presentation and discussion of results.
- **Chapter 7**: Conclusion and future research directions.

2 BACKGROUND

This chapter introduces the critical aspects of vNG-RAN, their architecture, and the role of disaggregation and virtualization in enhancing network capabilities for 5G and beyond. In Section 2.1, we explore the evolution, architecture, and conceptual underpinnings of RAN disaggregation, highlighting its significance in meeting the demands of low latency, high data rates, and energy efficiency. The evolution from Distributed Radio Access Network (D-RAN) to Cloud Radio Access Network (C-RAN) and finally to vNG-RAN is discussed, illustrating the architectural progress and challenges at each stage.

Further, Section 2.2 discusses the O-RAN architecture, a pivotal development aimed at fostering openness in RANs. The discussion covers the structure's components, including the Non-RT RIC and the Near-RT RIC, elucidating their roles in enhancing network operations and efficiency. Section 2.3 addresses the crucial aspect of power management within the RAN, particularly in the context of O-RAN. We examine the primary energy consumption components in radio and transport networks, outlining strategies to optimize energy efficiency. This study aims to understand and define the use case for energy optimization within the RAN context.

2.1 Next Generation Radio Access Networks

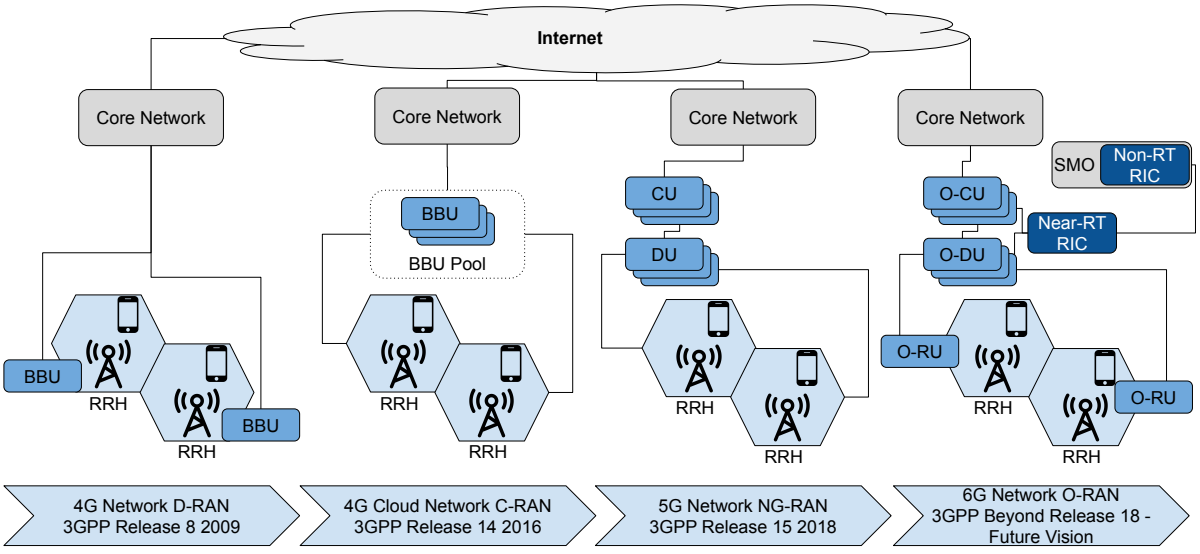
This section explores the evolution, architecture, and disaggregation of RAN in the context of Fifth Generation (5G) and beyond mobile networks. The discussion begins by tracing the evolution of RAN through the 3rd Generation Partnership Project (3GPP) standardization process, highlighting the improvements brought by Long Term Evolution (LTE) networks and the emergence of Evolved Node Bs (eNBs). The section emphasizes the need for re-evaluating RAN architecture to meet the demands of low latency, high data rates, adaptability, scalability, and energy efficiency. We introduce the architectural evolution from D-RAN to C-RAN and, ultimately, to the vNG-RAN architecture, emphasizing the advancements and challenges associated with each stage. The section also addresses the concept of RAN disaggregation, discussing its significance in flexible hardware and protocol composition and its influence on vNG-RAN development. Moreover, we present the various options for disaggregation and their implications on bandwidth and latency requirements within the Crosshaul transport network. The role of virtualization in realizing the disaggregated RAN functions is highlighted, along with its alignment with European Telecommunications Standards Institute (ETSI) standards and industry initiatives. The subsections cover topics such as protocol stack disaggregation, crosshaul transport networks, and the virtualization of RAN functions, collectively providing a comprehensive understanding of the evolution and prospects of radio access networks.

2.1.1 Evolution of Radio Access Networks

The evolution of RAN, through the 3GPP standardizing specifications for the LTE network, established from Release 8, brought significant improvements for the immersion in statistical packet networks and increased data rates. From that, 3GPP named the eNB. An eNB comprises two network elements, the Baseband Unit (BBU) and the Remote Radio Head (RRH). BBU is responsible for baseband data processing functions and is usually distributed and co-located with RRH. At the same time, RRH consists of radio units and antennas. However, the Release 8 specifications were developed based on the D-RAN architecture, on particular communication standards, and processing capacity for the services proposed for the LTE network (broadband and voice), providing support for some fixed loads.

The current demand for new requirements needed re-evaluating the RAN architecture to support low latency, high data rates, adaptability in non-uniform traffic, scalability, ultra-dense coverage, and low power consumption. Figure 3 details the architectural evolution of RAN. This illustration represents the evolution of the RAN architecture, starting with the D-RAN (LTE technology), explaining the evolution of the C-RAN architecture (in cloud computing environments) and the centralization of BBU, and ending in the current development, the vNG-RAN architecture, which provides high distribution, flexibility, and improvement of radio functions (AGRAWAL et al., 2017; CHEN et al., 2015).

Figure 3 – Architectural evolution of RAN.



Although not specified by 3GPP, the concept of RAN centralized in cloud computing environments, the C-RAN architecture emerged as a candidate to address the deficiencies presented by RAN from Release 14. Therefore, the C-RAN architecture proposes a significant modification in BSs, given that BBUs are decoupled from RRHs and centralized in one or more sets of BBUs in a cloud computing infrastructure, as expressed in Figure 3. Furthermore, this decoupling allows obtaining a high density of geographically distributed RRHs closer to UEs.

In addition, since the propagation of the communication signal is concentrated close to UEs, it offers greater system capacity and lower energy consumption. In this context, the advantages related to the cost reduction of physical locations (building infrastructure), operation, and maintenance stand out with the C-RAN architecture (3GPP, 2017b; PLIATSIOS et al., 2018).

C-RAN enables sharing and elasticity (allocating more or fewer resources according to demand). In this way, the network started to operate with non-uniform user standards. For example, allowing adaptation to everyday urban social movements, i.e., moving from residential to commercial areas during the day and returning at night. Another relevant point of the C-RAN architecture is introducing the new transport layer called Fronthaul, responsible for the communication between RRHs and BBUs (HABIBI et al., 2019; AGRAWAL et al., 2017).

The C-RAN architecture proposes significant improvements in the performance of radio attributes (processing efficiency, mobility, interference, and others). Moreover, the division between BBU and RRH requires high data rates from the fronthaul network. Therefore, it is impossible to have a high centralization of the BBU pool and, consequently, the advantages of this architecture for radio functions. The Academy conducted several types of research intending to seek alternatives to reduce the number of bits per second in the Fronthaul connection. However, this workaround includes the need to include more local functions in BSs and high processing before transmission (AL-DULAIMI; WANG; I, 2018).

To address these challenges, the 3GPP specified Release 15 for vNG-RAN, which drives the development of C-RAN mainly regarding disaggregation flexibility (up to two different points and different protocol groups) and data processing efficiency. In addition, it provides opportunities for grouping the lower-order radio units into a higher-order node (a group of RUs in a DU and a group of DUs in a CU). Furthermore, Release 15 specified the interoperability between the fourth and fifth-generation systems for vNG-RAN, introducing four variations for the BSs independent of CN in operation. In the case of LTE, the BS was kept as eNB when connected to the CN and Next-Generation - Evolved NodeB (ng-eNB) for connection with the 5G CN. While for fifth-generation mobile networks, the Next Generation Node B (gNB) is equivalent to BSs for connection with CN of this same generation and en-gNB (without defined acronym) for CN of fourth-generation (3GPP, 2019; MARSCH et al., 2018; BERTENYI et al., 2018).

A significant transformation refers to the architecture, where disaggregation is possible up to two different points and is a possible use up to three distinct units: CU, DU, and RU, as per Figure 3. The three-unit partitioned architecture conceptually divides the LTE BBU element into the gNB-CU and gNB-DU elements while maintaining the RRH only with the naming change to RU. Such transformation aims to enable a wide distribution of RUs and DUs nodes in the network for efficiently handling computational resources and to guarantee the requirements based on services (such as the latency requirement) (AL-DULAIMI; WANG; I, 2018). In industry, O-RAN initiatives also adopt the three-unit architecture, but with the insertion of the word Open RAN on the Open Central Unit (O-CU), Open Distributed Unit (O-DU), and Open Radio

Unit (O-RU) nodes. It is important to mention three key concepts in this architecture. SMO is a part of O-RAN that helps manage and orchestrate the network resources. It provides necessary services, configurations, and fault management for the network. The Non-RT RIC is a software function that uses non-real-time intelligence to create and manage policies for optimal radio resource management. In contrast, the Near-RT RIC uses near-real-time intelligence for radio resource management, such as load balancing, interference management, etc. These functionalities provide intelligence and enable efficient use of the network and resources, thereby offering improved performance (BERTENYI et al., 2018; 3GPP, 2017b; O-RAN Alliance, 2018, 2023a).

In this new architecture, the transport network is redesigned and developed considering segments of Fronthaul (FH), Midhaul (MH), and Backhaul (BH). FH is responsible for the communication between RU and DU (gNB-DU), differently from C-RAN, which is responsible for the communication between RRHs and BBUs. The MH is where communication between DU and CU (gNB-CU) occurs. Finally, BH is responsible for communication between CU and the core of the 5G network. The integration between these transport segments is called Crosshaul. In the same way, the division proposed in the architecture provides an opportunity to provide flexibility on the composition of the functions that make up the radio nodes and their respective protocols, as presented in the following subsection.

2.1.2 Disaggregation of Next Generation Access Radio Networks

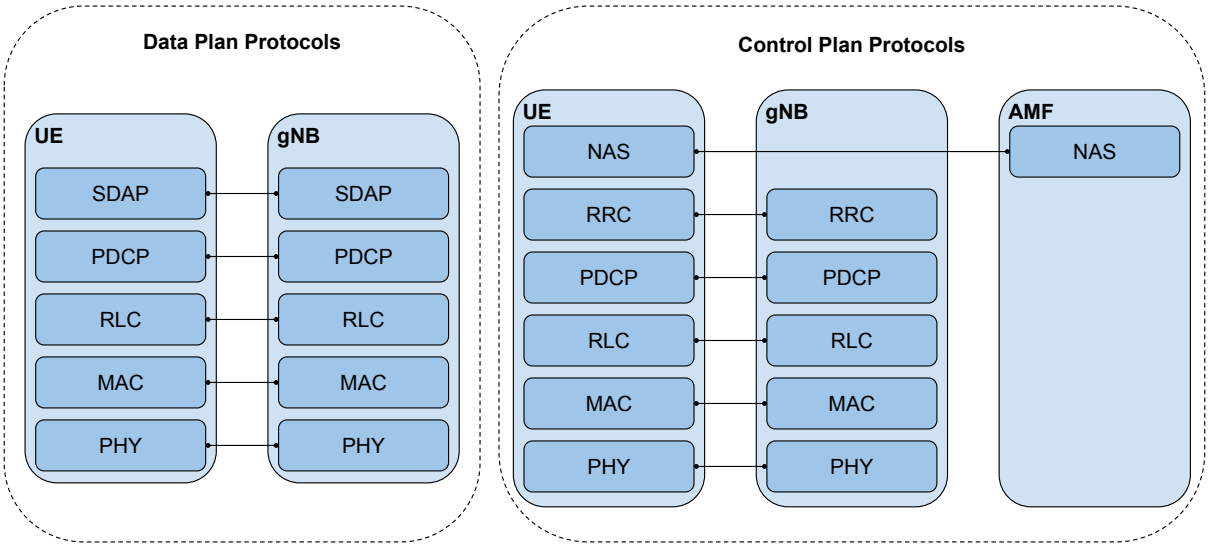
The flexibilization of radio functions in the literature is called RAN disaggregation, which allows radio functions not to be rigid in the composition of their hardware protocols and resources. This independence happens if they comply with six normative protocols stack composition and sequence. Unfortunately, this does not happen in the LTE RAN (D-RAN), oriented to a protocol stack and uses an architecture based on monolithic hardware where few interactions between logical nodes are specified (resulting in a limited and straightforward RAN architecture).

The development of vNG-RAN foresees additional benefits, mainly in the RAN functions that make up the CU and DU units, as (i) flexible, shared implementation of hardware, enabling cost-effective and scalable solutions; (ii) a split architecture provides resource performance coordination, workload management, and real-time performance optimization and allows for implementations based on software, and; (iii) the disaggregation of radio functions allows flexibility based on applications from different use cases (MARSCH et al., 2018; BERTENYI et al., 2018).

The choice of separating the functions of the vNG-RAN protocol stack depends on the scenario, limitations, and desired service, e.g., low latency, high bandwidth, and the density of specific users within a geographic area. The division of the vNG-RAN protocol stack has two groups: (i) the protocols that make up the data plane (also called the user plane) and (ii) those that make up the control plane. Figure 4 demonstrates the protocols that compose the

data plane and control plane (3GPP, 2017b; CARDOSO et al., 2020; BERTENYI et al., 2018; 3GPP, 2019).

Figure 4 – vNG-RAN data and control plane protocol stacks.



Based on (BERTENYI et al., 2018).

The division of the gNB protocol stack into network functions plays a vital role in reducing traffic over the FH and facilitating synchronization and latency requirements. The more processing is done in RU, the lower the requirements on the transport network. 3GPP proposed eight options for separating functions between centralized and distributed units, considering transport requirements, particularly throughput and latency. The separation of components that run on CU and DU can vary according to Quality of Service (QoS) needs. The benefits of this orientation make it possible to choose the disaggregation of radio functions according to the RAN implementation scenarios, restrictions, and foreseen services. For example, specific service requirements are needed, such as low latency, high transfer rate, specific user density, and load demand for a given geographic area. Furthermore, it presents the need to interoperate with heterogeneous Crosshaul networks with different performance levels (AL-DULAIMI; WANG; I, 2018; CARDOSO et al., 2020; BERTENYI et al., 2018; MARSCH et al., 2018).

The disaggregation of radio functions provides flexible guidance, both by demand requirements and network constraints, to determine the composition of the radio functions' protocols. For example, in previous LTE technology, RAN has a protocol stack composed of five main protocols: Physical Layer (PHY), Media Access Control (MAC), Radio Link Control (RLC), Packet Data Control Protocol (PDCP), and Radio Resource Control (RRC). Additionally, the 3GPP recommendation in Release 15, the inclusion of the Service Data Adaptation Protocol (SDAP) provided for in the protocol stack is also detailed below (3GPP, 2017b):

- PHY: is responsible for processing digital and analog communication signals between UEs and BSs and is based on adaptive modulation and coding techniques;

- MAC: provides low-level physical layer control, particularly when scheduling data transmissions between UE and BS. In addition to handling the mapping and selection of logical channels and transport channels, measurement based on traffic reports, and error correction through the Hybrid Automatic Repeat Request (HARQ);

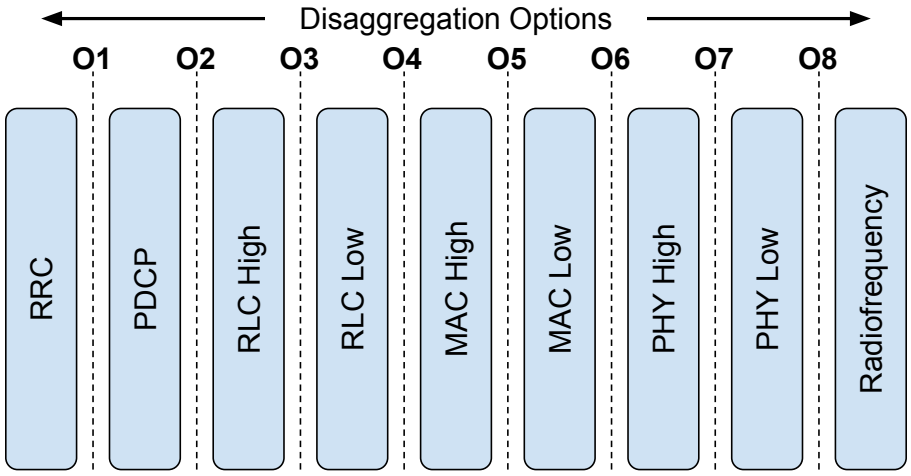
- RLC: guarantees the reliable delivery of data streams that need to reach the MAC layer intact. It is responsible for retransmission segmentation and handling using theAutomatic Repeat Request (ARQ);

- PDCP: performs higher-level transport functions related to IP datagram compression, security (encryption and information security), and virtual radio tunnels with UEs;

- RRC: is in charge of all signaling and control of UE with BS. This task includes establishing, reconfiguring, and updating radio bearers, mobility connection processes, admission control, location monitoring, and power control;

- SDAP: maps the interaction between a stream's packet with associated QoS and the data plane's radio bearer (specific to the UE payload), efficiently tagging the user's data packets;

Figure 5 – Protocol Stack Disaggregation Options.

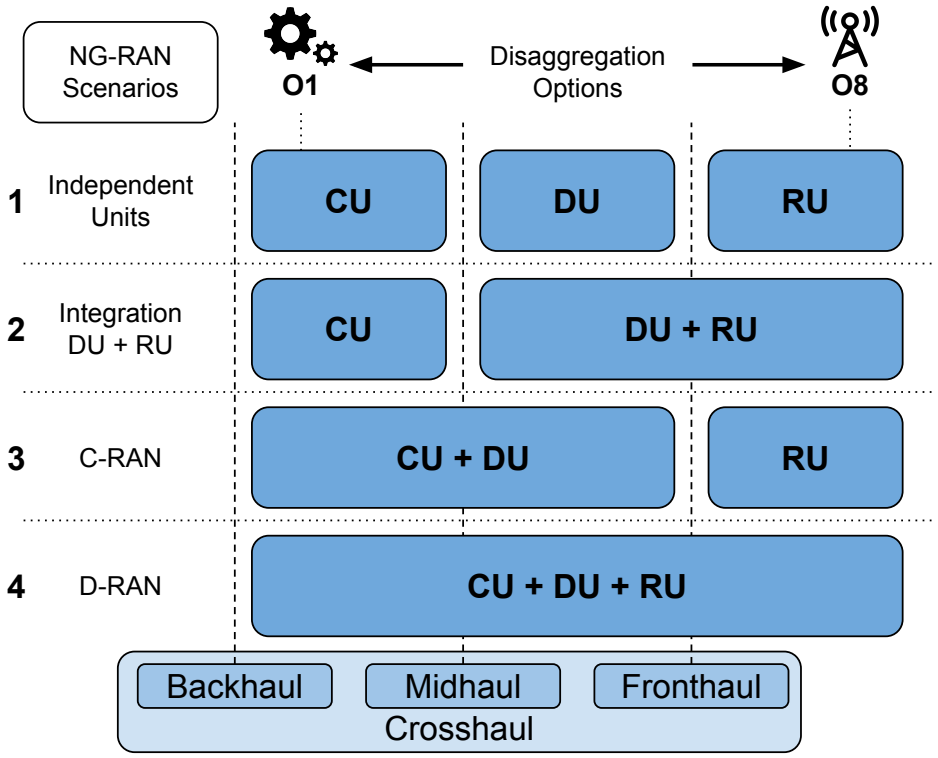


Based on (3GPP, 2017b).

Based on the vNG-RAN protocol stack, eight disaggregation options are supported, from option 1 (O1) to option 8 (O8), as shown in Figure 5. The convention of options is given by boundary between different protocols and by the intra-division of the RLC, MAC, and PHY protocols, established in low and high. All boundaries between protocols are likely to be treated as an option, except for SDAP. For example, the RRC and PDCP protocols do not have intra-divisions (ITU-T, 2018; 3GPP, 2017b).

Although vNG-RAN has three radio units (CU, DU, and RU), a maximum of two unbundling is supported, combined with the eight unbundling options, in the literature, it is called three independent units. However, single-point or no-disaggregation scenarios are also supported. One-point disaggregation is supported in two scenarios. The first is RU and CU integration, and the second is C-RAN. Finally, D-RAN operates with all units in a single device and the same physical environment without disaggregation. Figure 6 shows the detailed scenarios, combining the disaggregation options and points of the crosshaul network (ITU-T, 2018; MARSCH et al., 2018).

Figure 6 – Disaggregation Scenarios in Crosshaul.



Based on (ITU-T, 2018).

2.1.3 Crosshaul Transport Networks

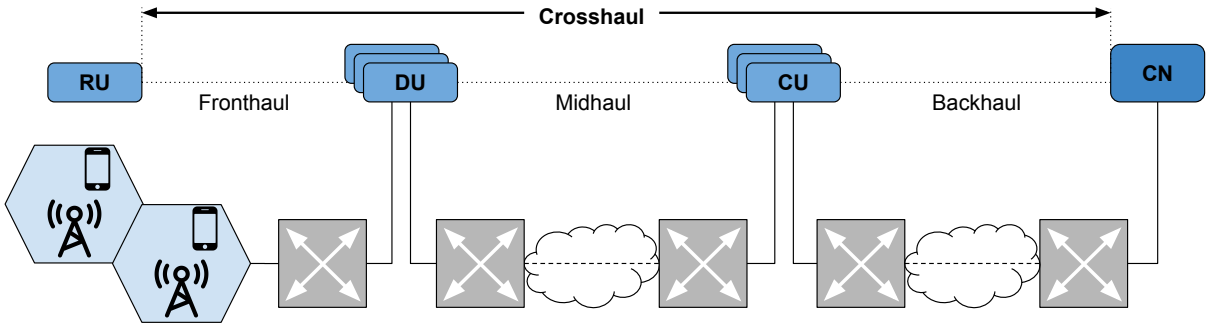
The disaggregation of RAN undergoes strong synergy with the Crosshaul network and its networks (BH, MH, and FH), which is directly associated with the 6G. Figure 7 shows the synergy between the RAN disaggregation and the Crosshaul network, detailing the conceptual positioning, the cloud in MH, and BH is to illustrate conceptually which no has restrictions for other topologies (LARSEN et al., 2018; ITU-T, 2018).

The BH network is the precursor of transport networks to serve mobile networks and was introduced with LTE technology to provide communication between BS and CN. As a result, it is positioned between CU and CN in the vNG-RAN architecture, as shown in Figure 7.

Concerning the disaggregation proposed in vNG-RAN, the network's positioning remains unchanged. However, the flexible positioning of CU can lead to traffic, with characteristics of the BH network, closer to the edge or core of the network (FIORANI et al., 2015; ITU-T, 2019).

The MH was developed to be applied in the new vNG-RAN architecture, positioned between the CU and DU nodes. Therefore, it is used when DU and CU are on different sites, as shown in Figure 7. Finally, the FH network was introduced with the C-RAN architecture to decouple the RRH and BBU nodes. Therefore, the FH network connects the RRHs to a centralized pool of BBUs. However, for the vNG-RAN architecture, the FH network is mainly positioned among the lowest protocols in the protocol stack, among the disaggregation options that make up the communication between RU and DU (PHY and MAC protocols) (FIORANI et al., 2015; ITU-T, 2018). In the following subsection, we discuss crosshaul network requirements accordingly with disaggregation options.

Figure 7 – Crosshaul Architecture.



Based on (ITU-T, 2018).

2.1.4 Disaggregation Requirements of Next Generation Access Radio Networks

Regardless of the Crosshaul network, bandwidth and latency metrics are the main ones regarding network requirements. Table 1 presents the reference values for bandwidth and latency required for the interaction between each protocol split option. These values are a 3GPP reference for radio channels with 100MHz bandwidth, 32 antennas, eight layers of Multiple-Input, Multiple-Output (MIMO), and Quadrature Amplitude Modulation (QAM) of 256 symbols (3GPP, 2017b).

For bandwidth, Table 1 presents Downlink (DL) and Uplink (UL), showing asymmetry between them, in addition to demanding very high rates from Crosshaul in options O7 and O8. Concerning latency, the metric requires extremely low measures from the Crosshaul network to meet the disaggregation, which measures less than 1ms between options O4 and O8. Such requests for latency occur because of the characteristics of the disaggregation options and the new services for fifth-generation networks. The support for high data rate and restricted delay

Table 1 – Bandwidth and latency requirements for splits options (3GPP, 2017a).

Disaggregation Options	Functional Division	Approximate Bandwidth	One way max latency
Option 1	RRC-PDCP	DL: 4Gb/s - UL: 3Gb/s	10ms
Option 2	PDCP-High RLC	DL: 4Gb/s - UL: 3Gb/s	1.5ms∼10ms
Option 3	Intra RLC	DL: 4Gb/s - UL: 3Gb/s	1.5ms∼10ms
Option 4	Low RLC-High MAC	DL: 4Gb/s - UL: 3Gb/s	∼0.1ms
Option 5	Intra MAC	DL: 4Gb/s - UL: 3Gb/s	<1ms
Option 6	Low MAC-High PHY	DL: 4Gb/s - UL: 5Gb/s	0.250ms
Option 7	Intra PHY	DL: 10.1∼86.1Gb/s UL: 16.6∼86.1Gb/s	0.250ms
Option 8	Low PHY-RF	DL: 157.3Gb/s UL: 157.3Gb/s	0.250ms

services should not exceed 4ms and 0.5ms, respectively. Moreover, 3GPP recommendations for end-to-end latency (UE to CN) specify a latency of 20ms for signaling Control Plane (CP) and a range between 1≈4ms for the User Plane (UP) payload within mobile networks (ITU-T, 2018; 3GPP, 2017a).

Although each disaggregation option has a bandwidth requirement specified, the BH and MH networks are responsible for traffic aggregation due to the grouping of different RAN nodes. Therefore, both BH and MH are specified to support hundreds of Gbit/s. Differently, BH aggregates CU nodes (last nodes before CN) and MH aggregates DU nodes (which are aggregated by CU nodes). As for the network topology, International Telecommunication Union - Telecommunication Standardization (ITU-T) is currently a point-to-point topology that is the basis for the FH and does not foresee traffic aggregation in the upper layers. Finally, for both layers MH and BH, ITU-T predicts a ring or tree topology (AL-DULAIMI; WANG; I, 2018; ITU-T, 2018, 2019). In the following subsection, we discuss the main accelerator of disaggregation, Virtualization of Mobile Networks.

2.1.5 Virtualization of RAN

Decomposing radio functions into non-monolithic components and aggregating functions to improve radio performance and processing are vital to leveraging 6G mobile networks. In this sense, concepts based on software are proposed natively for evolution, mainly virtualization. Immersed in this concept, each vNG-RAN unit (CU, DU, and RU) is considered a virtualized network function (Virtualized Central Unit (vCU), Virtualized Distributed Unit (vDU), and Virtualized Radio Unit (vRU)) composed of disaggregated radio functions. These functions can run on up to three different computational resources. Furthermore, each computing device has a set of RAN protocols running on them (whether symmetrical or asymmetrical). Although RU can be (vRU) and support the sharing of computational resources, the RF protocol does

not support this purpose. However, to achieve the desired advance, the vRAN must be carried out in such a way as to satisfy the conditions imposed by the disaggregation of the vNG-RAN architecture (BERTENYI et al., 2018; MARSCH et al., 2018).

The ETSI standardization organization is currently the primary reference in virtualized mobile networks, including RAN. It defines virtualization as "the elimination of the dependency between network functions and hardware," which works from the continuous development of the Network Function Virtualization (NFV) concept and architecture (ETSI, 2013; FULBER-GARCIA et al., 2023). To offer the opportunity to implement network functions in software that can be instantiated in different locations and hardware in the network. This network allows the creation of VNFs separate from the underlying proprietary hardware and transfer them to a shared computing resources infrastructure based on software. Notably, the industry's O-RAN initiatives are aligned with the ETSI (GUERZONI et al., 2012; ETSI, 2013; BERNARDOS et al., 2019; BONATI et al., 2020).

2.2 O-RAN Architecture

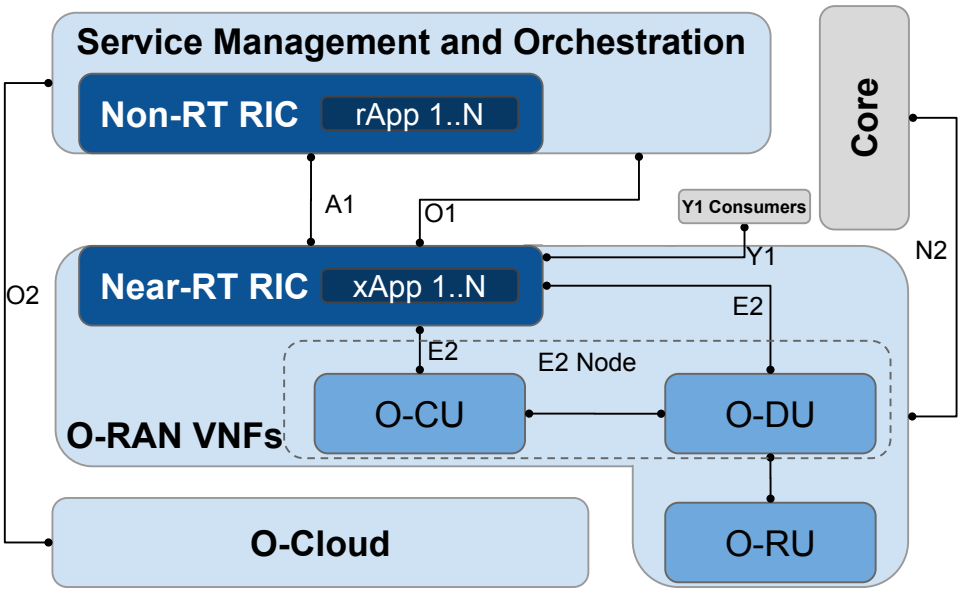
The O-RAN architecture is proposed and developed by the O-RAN Alliance. This organization is a worldwide effort to reach new levels of openness in vNG-RANs. Initially launched by five major mobile carriers a couple of years ago and supported by over 160 companies (including 24 mobile operators across four continents), it represents an outstanding example of how mobile network operators and suppliers worldwide can constructively collaborate to define novel technical standards. O-RAN is a significant carrier-led effort to define the vNG-RANs for multi-vendor deployments. It aims to disrupt the vRAN ecosystem by breaking vendors' lock-in and opening up a market that a few players have traditionally dominated (GARCIA-SAAVEDRA et al., 2021).

Previous RAN innovations, such as C-RAN, brought operational efficiencies, but earlier advances did not free operators from vendor lock-in. O-RAN leads to cloud savings and competition at RAN by enabling an open, multi-vendor RAN ecosystem. Open interface standards allow third-party products to communicate with the leading provider's RAN infrastructure. Network operators can opt for the lower-cost third-party product running on generic hardware. As network operators focus on transitioning to a vRAN architecture for 5G, open RAN interfaces can minimize the cost of deploying the new 5G technology.

The two main components introduced by O-RAN are the Non-RT RIC and the Near-RT RIC. The first component, described in Subsection 2.2.3, is hosted by the SMO framework. The SMO, described in Section 2.2.1, consolidates several orchestration and management services, which may go beyond pure RAN management, such as 3GPP Next Generation Core (NG-Core) management or end-to-end network slice management. The second component, described in Subsection 2.2.2, is colocated with 3GPP gNB functions, namely, O-RAN-compliant O-CU and O-DU (Figure 8). This component can be fully decoupled over one layer if latency con-

straints are respected. The RIC Non Real-Time (Non-RT) layer performs operations, including policy management and analytics. However, the RIC Near Real-Time (Near-RT) layer performs time-sensitive functions, e.g., load balancing, handover, and interference detection (GARCIA-SAAVEDRA et al., 2021; O-RAN Alliance, 2023a).

Figure 8 – O-RAN Architecture.



Based on (O-RAN Alliance, 2023a).

2.2.1 Service Management and Orchestration

This section addresses the role of the SMO in O-RAN architecture, specifically focusing on its functionalities related to RAN management. Within a Service Provider's Network, multiple management domains such as RAN, Core, Transport, and End-to-End Slice Management exist. SMO is a critical component that handles the RAN management in the O-RAN architecture.

SMO plays a pivotal role in the O-RAN architecture by offering three key capabilities to support RAN management. First, it provides Fault, Configuration, Accounting, Performance, and Security (FCAPS) Support by establishing an interface to O-RAN Network Functions, allowing for comprehensive Fault, Configuration, Accounting, Performance, and Security management. Second, it incorporates a Non-RT RIC designed to optimize RAN operations. Finally, SMO is responsible for O-Cloud Management, where it orchestrates, manages, and oversees the workflow of cloud resources within the O-RAN ecosystem. These functionalities collectively contribute to RAN's robust and efficient management within the O-RAN architecture.

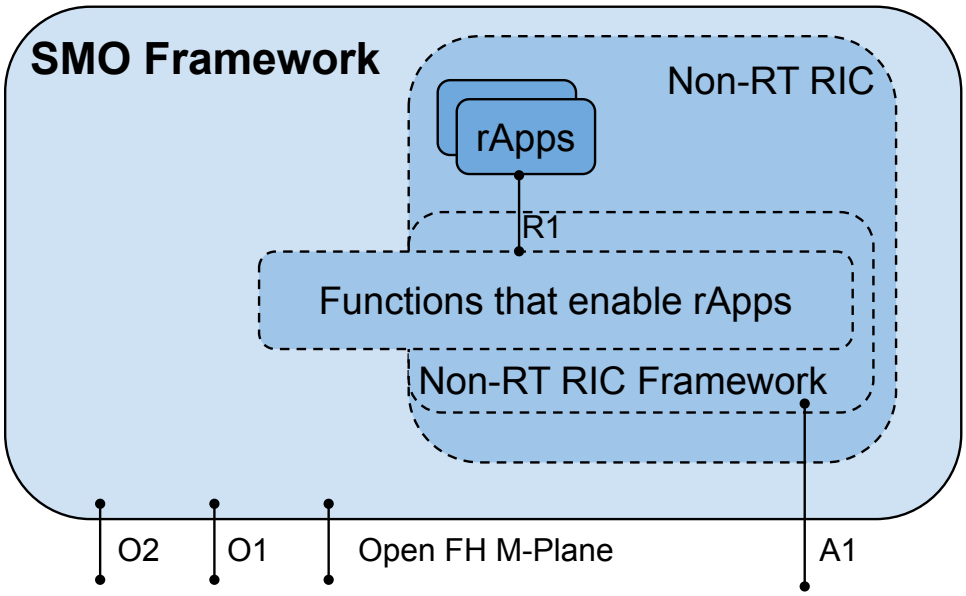
In the O-RAN architecture, the SMO is a pivotal component that communicates through four critical interfaces to manage various functionalities. The A1 Interface is a conduit between the Non-RT RIC within the SMO and the Near-RT RIC, facilitating RAN optimization. The O1

Interface is strategically designed to liaise with O-RAN Network Functions, chiefly for FCAPS management tasks. In a hybrid deployment model, the Open Fronthaul M-plane Interface establishes connectivity between the SMO and O-RU, providing additional support for FCAPS functions. Finally, the O2 Interface bridges the SMO and the O-Cloud, efficiently managing platform resources and workloads.

The functionalities of SMO can be taxonomically categorized into three distinct types. First, functionalities are anchored to the Non-RT RIC Framework, exemplified by the A1 and R1 interfaces. These functionalities are inherently tied to the framework's architecture and objectives. Second, we have the O-RAN SMO Framework Anchored Functionality, which operates independently of the Non-RT RIC Framework. This category includes interfaces such as O1, Open Fronthaul M-plane, and O2, which are fundamental for operations, such as FCAPS management and cloud orchestration. Lastly, a set of Non-anchored Functionalities exists that may or may not have a relationship with the Non-RT RIC Framework, offering additional flexibility in the SMO's operational scope. These categories delineate the varying degrees of dependency and integration within the SMO's architecture.

The SMO architecture does not mandate a strict interface specification with the Non-RT RIC Framework. This specification allows distinct SMO deployments to delineate or eliminate architectural boundaries with Non-RT RIC. Such flexibility is vividly depicted in Figure 9. It highlights that functionalities empowering rApps can be sourced from either the Non-RT RIC or the SMO framework, emphasizing the adaptability intrinsic to the SMO design.

Figure 9 – Exposure of SMO and Non-RT RIC Framework Services.



Adapted from (O-RAN Alliance, 2023a).

The O2-related functions offer a suite of services linked to O2. The A1 termination acts as a bridge, enabling the Non-RT RIC framework and the Near-RT RIC to communicate through

the A1 interface. Similarly, the O1 termination grants the SMO framework the capability to interface with the Near-RT RIC/E2 nodes via the O1 channel. The Open fronthaul M-plane termination is pivotal for the seamless exchange of messages between the SMO framework and the O-RUs, facilitated through the Open fronthaul M-plane interface. Furthermore, the O2 termination ensures that the SMO framework can effectively interact with the O-Cloud using the O2 conduit.

2.2.2 Non-RT RIC

The Non-RT RIC is a pivotal component within the O-RAN architecture, specifically located internally in the SMO. It furnishes the A1 interface to its counterpart, the Near-RT RIC. The foremost aim of the Non-RT RIC is to bolster intelligent RAN optimization, offering policy-based guidance, overseeing ML model management, and supplying enrichment information to the Near-RT RIC. This concerted effort allows the RAN to fine-tune operations, such as RRM, under specific conditions. Notably, while the Near-RT RIC operates within a timespan of up to 1 second, the Non-RT RIC functions in intervals exceeding this, handling tasks that do not necessitate real-time responsiveness.

An intrinsic capability of the Non-RT RIC is its adeptness at utilizing data analytics and AI/ML training and inference mechanisms. With these tools, the Non-RT RIC can discern the most appropriate RAN optimization actions. To facilitate this, it harnesses SMO services, encompassing data collection and provisioning services of O-RAN nodes. Furthermore, it interacts with the O1 and O2 interfaces to further its optimization endeavors.

The Non-RT RIC is bifurcated into two primary sub-functions. First, the Non-RT RIC Framework is an intrinsic functionality within the SMO Framework. It serves as the logical termination point for the A1 interface and proactively unveils the necessary services to the rApps via its R1 interface. Second, Non-RT RIC Applications (rApps) are modular applications that capitalize on the capabilities presented by the Non-RT RIC Framework. The primary role of rApps is to facilitate RAN optimization, among other tasks. Through the R1 interface, rApps can access many services that empower them to glean information and initiate actions. This information includes, but is not limited to, policies, re-configurations, and interactions with the A1, O1, O2, and Open FH M-Plane-related services (O-RAN Alliance, 2024a)

2.2.3 Near-RT RIC

The Near-RT RIC is a crucial functional component that enables rapid control and optimization of E2 Node functionalities and resources. The rapid control and optimization are achieved by employing fine-grained data collection and executing actions via the E2 interface, with control loops operating in 10 milliseconds to 1 second (O-RAN Alliance, 2023a).

Within the Near-RT RIC, one or multiple xApps can be hosted. This xApps interact with

the E2 interface to gather near-real-time data, either at the individual UE level or at a broader cellular level. The gathered information is leveraged to provide value-added services. The control that the Near-RT RIC exerts over E2Ns is guided by policies and enrichment data obtained through the A1 interface from the Non-RT RIC. The Near-RT RIC generates RAN analytics based on this incoming data, which is then accessible via the Y1 interface.

The division of RRM functions between the Near-RT RIC and the E2 Node depends on the capabilities of the E2 Node. These capabilities are explicitly listed and described through the E2 Service Model. This model outlines which functions within the E2 Node can be directly influenced by the Near-RT RIC, thereby establishing a function-specific split in RRM responsibilities between the two components. For functions detailed in the E2 Service Model, the Near-RT RIC can monitor, suspend, override, or control the behavior of the E2 Node according to pre-defined policies.

In the unlikely event of a Near-RT RIC failure, it is important to note that the E2N can continue to provide core services. However, there may be a temporary interruption in the availability of specific value-added services that rely on the Near-RT RIC for their functionality. This architecture ensures a balanced yet flexible allocation of responsibilities for optimizing network performance and service delivery.

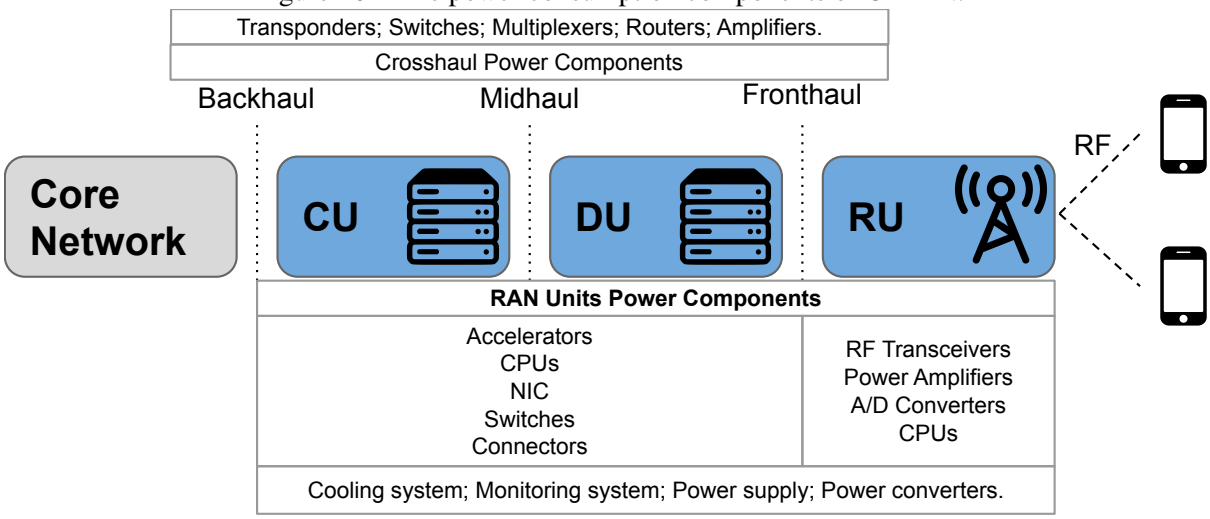
2.3 RAN Power Management

With the adoption of O-RAN, there is a reliance on Commercial Off-The-Shelf (COTS) hardware, including General-Purpose Processor (GPP) servers, and transforming network functions into virtualized software applications. Consequently, a substantial portion of O-RAN's energy utilization is attributed to data processing or computational tasks conducted in cloud or edge servers, specifically within the CUs and DUs. Radio Frequency (RF) functionalities and power amplification tasks heighten this consumption within the RUs. Additionally, the energy expended for data transmission across the BH, MH, and FH networks is another critical factor that merits consideration.

One of the principal use cases for O-RAN is energy savings. The framework introduced in this thesis uses an energy-saving use case to validate its functionality. Given the increasing energy demands of modern network infrastructures and the critical need for sustainable practices, energy efficiency within the RAN is a vital area of focus.

In the ensuing section, we aim to delineate the primary energy consumption elements pertinent to the radio and transport networks within the O-RAN architecture. These energy consumption components in O-RAN are depicted in Figure 10.

Figure 10 – The power consumption components of O-RAN.



2.3.1 Energy Utilization in O-RAN Components

The cumulative energy usage of a radio network encompasses the consumption by both the hardware and software components integral to the functioning of CUs, DUs, and RUs. The positioning of these units across network nodes is a significant determinant of the network's overall energy footprint. O-RAN architecture, which leverages network virtualization, manifests most CUs, DUs, and certain RU segments through virtual machines hosted on Commercial Off-The-Shelf (COTS) servers. Consequently, most of the radio unit's energy drain is attributed to the computational load of GPPs(AZARIAH et al., 2022). Core elements of the radio network include Central Processing Units (CPUs), specialized accelerators, and Network Interface Cards (NICs), which are incorporated into the servers housing the CUs and DUs. Moreover, the RU has RF transceivers and power amplifiers. Other shared infrastructural elements, such as cooling systems, monitoring apparatus, alarm systems, and power supply and conversion mechanisms, also play a role(ZORELLO et al., 2022). The chosen functional split further influences the radio network's energy consumption, as a more centralized setup at the CU level can reduce the RU's energy needs despite potential increases in latency and transport network complexities. This choice necessitates a balance between EE and other QoS parameters, such as latency and throughput, a common objective in optimization endeavors.

Constituted by FH, MH, and BH segments along with their respective switches, the energy consumption within a transport network is dictated by factors such as the employed technology, network design or topology, capacity demands, and the interconnectivity between CUs, DUs, and RUs. Key elements contributing to this network's energy usage include switches, transponders, and multiplexers. Diverse transport technologies such as Point-To-Point (P2P) fibre, Passive Optical Network (PON), microwave radio, Coarse Wavelength Division Multiplexing (CWDM), and Ethernet are examples in this context. Furthermore, the energy demands of the transport network fluctuate based on the implemented split option. The research docu-

mented in (LARSEN et al., 2019) examined the impact of the transport network on the total energy usage of a C-RAN with three different split options: 6, 7, and 8. The findings indicated that the transport network accounted for approximately 2%, 30%, and 60% of the total energy consumption for each split option. Therefore, while lower functional splits offer the benefits of centralization, they can inadvertently lead to elevated energy demands due to the increased capacity requirements imposed on the transport network (FIORANI et al., 2016).

2.3.2 Energy Efficiency Techniques in O-RAN

This section addresses the categorization of EE techniques identified in existing research into three groups. Initially, techniques focus on dynamically allocating resources and positioning network functions between CUs and DUs. The second category explores the strategic positioning of DUs and CUs across physical network nodes and their associations with users. The third group encompasses indirect EE optimization strategies that, while not explicitly targeting EE as a performance metric, optimize other parameters such as computational and routing costs, which beneficially influence O-RAN's EE (ABUBAKAR et al., 2023).

Dynamic Resource Allocation and Network Function Placement (DRAandNFP): Investigations in this domain concentrate on the distribution of processing tasks between CUs and DUs through dynamic network function placement. The objective is to centralize more network functions within a select number of DUs and CUs, allowing the deactivation of inactive virtual machines that host these units. Given that the power consumption of CUs and DUs correlates with their processing activity, minimizing the number of active units—while maintaining satisfactory QoS benchmarks such as latency and packet delivery ratio—can significantly enhance O-RAN's energy efficiency.

Dynamic Centralized and Decentralized Processing and User Association (DCDPandUA): This approach investigates the optimal implementation of CUs and DUs within a limited number of network nodes or physical machines. The aim is to permit the deactivation of inactive nodes or machines, thereby conserving energy through more efficient computational resource usage and centralized processing.

Indirect Optimization of Energy Efficiency (IOEE): This approach considers studies where the primary objective was not EE of O-RAN, but rather other aspects such as computational and routing costs. These factors directly impact O-RAN's energy consumption and are crucial for its indirect optimization.

2.4 Summarizing

In this chapter, we introduced the critical aspects of vNG-RAN and their architecture, emphasizing the importance of disaggregation and virtualization in enhancing network capabilities for 5G and beyond. We explored the evolution of RAN from D-RAN to C-RAN and finally to vNG-RAN, highlighting the architectural progress and the challenges encountered at each

stage. The discussion on O-RAN architecture provided insights into its components, such as the Non-RT RIC and Near-RT RIC, and their roles in improving network operations and efficiency. We also examined the crucial aspect of power management within the RAN, focusing on energy consumption components and strategies for optimizing energy efficiency. In the next chapter, *Related Work*, we present a comprehensive review of current research in O-RAN, addressing specific, pivotal questions central to developing and optimizing these networks in the context of 6G advancements. Moreover, we discuss the foundational understanding established in this chapter, enabling a deeper exploration of the advancements in radio access networks.

3 RELATED WORK

This chapter presents a comprehensive review of current research in O-RAN, addressing specific, pivotal questions central to developing and optimizing these networks in the context of 6G advancements since 2021 because that is the year that start consistent works related with O-RAN. Section 3.1 focuses on integrating O-RAN components to efficiently address dynamic user demands, a key challenge in the evolution of 6G networks, and reviews various strategies in resource allocation, data traffic management, and network orchestration, as summarized in Table 2.

In contrast, Section 3.2 introduces various use cases within the O-RAN framework, exploring how advancements in Massive MIMO, network slicing, and energy-saving strategies address the demands of future telecommunication networks. Following this, Section 3.3, evaluates orchestration strategies in O-RAN architectures, highlighting the importance of practical evaluation methods and criteria such as scalability, flexibility, and adaptability in response to fluctuating user demands.

The final section, 3.4, discusses strategies for adequate disaggregation of Near-RT RIC functions in O-RAN, crucial for managing fluctuating user demands and ensuring network scalability. Each section builds upon the existing body of knowledge in O-RAN research, contributing to a deeper understanding of how these networks can be optimized for the rapidly evolving demands of modern telecommunications.

3.1 Efficient O-RAN Integration for Dynamic User Demands

Table 2 – Related Work with Dynamic User Demands

Works	Category	Objective	Experimentation	xApp	rApp	Near-RT	Non-RT	SMO	Use Case Energy Saving
(ORHAN et al., 2021)	Data Traffic Management	A connection management as a combinatorial graph optimization problem	Theoretical	✓	✗	✓	✗	✗	✗
(YOO et al., 2022)		Load balancing	Emulation: SD-RAN	✓	✗	✓	✗	✗	✗
(KOUCHAKI et al., 2022)		QoE maximization	Emulation: OSC	✓	✗	✓	✗	✗	✗
(HUANG et al., 2022)		Throughput maximization	Theoretical	✓	✗	✓	✗	✗	✗
(AGARWAL et al., 2023)		QoE enhancement function	Simulation	✓	✗	✓	✗	✗	✗
(LACAVA et al., 2023)		Traffic Steering intelligent handover	Simulation: ns-O-RAN	✓	✗	✓	✗	✗	✗
(ALAVIRAD et al., 2023)		Admission control of UEs	Simulation: ns3 LTE	✓	✗	✓	✓	✗	✗
(THALIATH et al., 2022)		Ensure SLA at network slicing	Emulation: OSC and ONAP	✓	✓	✓	✓	✓	✗
(BONATI et al., 2023)		Spectrum sharing framework	Emulation: OSC	✓	✓	✓	✓	✓	✗
(SOHAIB et al., 2024)		Green resource allocation	Emulation: Simulation	✓	✓	✓	✓	✗	✓
(MUNGARI, 2021)	Resource Allocation	Radio resource management	Emulation: OSC and OAI	✓	✗	✓	✗	✗	✗
(VILA et al., 2022)		Capacity Sharing	Simulation: ONAP	✗	✓	✓	✓	✓	✗
(D'ORO et al., 2022)		Radio resource management	Emulation OSC	✗	✓	✓	✓	✓	✗
(KASULURU et al., 2023)		Radio resource management	Simulation	✗	✓	✗	✗	✗	✗
This Work		Dymanic resource management	Emulation and Simulation	✓	✓	✓	✓	✓	✓

This section comprehensively examines how various O-RAN components can be efficiently integrated to respond to dynamic user demands, a crucial challenge in the evolution of 6G networks. Reflected in the research summarized in Table 2, the focus is on innovative strategies in resource allocation, data traffic management, and adaptive network orchestration to ensure responsiveness and efficiency. The discussion navigates through the latest advancements and methodologies that enable O-RAN to adapt in real-time to changing network demands, maintaining performance and user experience. By dissecting key studies and their contributions to O-RAN technology, this section elucidates the multifaceted approach necessary for the dynamic and efficient operation of future telecommunication networks.

In terms of resource allocation, studies such as (MUNGARI, 2021), (VILA et al., 2022), and (KASULURU et al., 2023) highlight the need for RRM. These approaches often involve the dynamic allocation of network resources, such as bandwidth and power, to meet varying user demands efficiently. This could include allocating more resources to areas with high demand or redistributing resources during periods of low usage to maintain optimal network performance. Additionally, the integration of green resource allocation frameworks leveraging Deep Reinforced Learning (DRL) for online resource allocation decisions in cloud-native O-RAN networks, as discussed in (SOHAIB et al., 2024), can further enhance the efficiency of dynamic user demand management.

For data traffic management, articles such as (ORHAN et al., 2021), (YOO et al., 2022), and (KOUCHAKI et al., 2022) explore strategies for optimizing user-cell association and load balancing. These articles involve directing data traffic to prevent congestion and ensure consistent service quality across network segments. Techniques such as traffic steering and intelligent handover, as examined in (LACAVA et al., 2023), play a critical role in managing data flows within the network to adapt to changing user patterns and demands.

The orchestration of network components is another crucial aspect. Works such as (D'ORO et al., 2022) and (BONATI et al., 2023) discuss the orchestration frameworks within O-RAN that automate the decision-making process for resource allocation and traffic management. These frameworks must dynamically adapt to changing network conditions, making real-time decisions to optimize network performance. Moreover, the integration of Near-RT RIC and Non-RT RIC, as seen in several studies (VILA et al., 2022), (D'ORO et al., 2022), (ALAVIRAD et al., 2023), (THALIATH et al., 2022), (BONATI et al., 2023), is pivotal. Near-RT RIC enables quick response to changing network conditions, while Non-RT RIC provides longer-term network planning and optimization strategies. The synergy between these two components is vital for managing dynamic user demands efficiently.

In conclusion, this work proposes that the efficient integration of O-RAN components to address dynamic user demands can be achieved through a multifaceted approach that includes adaptive resource allocation, effective data traffic management, and intelligent orchestration. By adopting these strategies, O-RAN can dynamically respond to varying user demands in real time, ensuring optimal network performance and user experience. Specifically, the work

outlines how adaptive RRM, user-cell association, load balancing, and the orchestration of Near-RT RIC and Non-RT RIC components work in synergy to maintain network efficiency. This integrated approach answers the central question of how O-RAN components can be harmonized to efficiently manage dynamic user demands, thereby contributing to the advancement of future telecommunication networks.

3.2 Use Case to Dynamic User Demand in 6G Networks through O-RAN

This section introduces diverse use cases within the O-RAN framework, explicitly targeting the dynamic user demands anticipated in 6G networks. It focuses on how these use cases, including advancements in Massive MIMO, network slicing, and energy-saving strategies, are crucial to addressing the challenges and demands of future telecommunications networks. The section discusses the practical applications and implications of these technologies, as detailed in subsections 3.2.1 and 3.2.2, and examines how they contribute to the adaptability and efficiency of 6G networks in response to fluctuating user needs. The synthesis of these use cases, along with the research works summarized in Table 3, offers a comprehensive view of the evolving role of O-RAN in meeting the dynamic requirements of next-generation networks.

3.2.1 O-RAN Use Cases

This section discusses various use cases within the O-RAN framework, focusing on advancements in Massive MIMO, network slicing, and energy-saving strategies to address key challenges in contemporary telecommunications networks. A comprehensive study was conducted on potential use cases, and the energy-saving use case was ultimately chosen for prototyping and validating the architecture/platform. This decision was based on the critical importance of energy efficiency in modern network operations and the significant potential for operational cost savings and environmental benefits.

Massive MIMO in O-RAN: The O-RAN Working Group 1 document on Massive MIMO (mMIMO) use cases emphasizes pre-normative development phases, including advancements in beam management and user experience optimization. Key areas include Grid of Beams (GoB) optimization, adaptive beam shaping, Mobility Robustness Optimization (Mobility Robustness Optimization (MRO)), and Non-GoB optimization strategies. Layer 1 and 2 beam management optimizations, such as Downlink and Uplink transmit power optimization and Multi-User MIMO co-scheduling, are also detailed, highlighting the ongoing efforts to enhance mMIMO capabilities within O-RAN (O-RAN Alliance, 2023e).

Network Slicing within O-RAN: The network slicing use cases in the O-RAN architecture, as outlined in the O-RAN Working Group 1 document on Slicing Architecture, reflect the evolving needs of telecommunications networks. This includes managing NSI and NSSI in alignment with the 3GPP Slice Management framework. The document discusses the creation, activation,

deactivation, modification, and termination processes of NSI and NSSI, indicating the direction for future network efficiency and flexibility enhancements (O-RAN Alliance, 2023f).

Energy Saving Strategies in O-RAN: The O-RAN Working Group 1 document on Network Energy Saving Use Cases explores strategies such as Carrier and Cell Switch Off/On, RF Channel Reconfiguration and Advanced Sleep Modes. These techniques reduce energy consumption in cellular networks, particularly during low-demand periods. They are managed through intelligent automation in the RIC and the SMO. Implementing these energy-saving strategies is crucial for balancing operational efficiency with energy conservation, demonstrating O-RAN's commitment to sustainable network operations (O-RAN Alliance, 2023g).

3.2.2 Energy Saver works with O-RAN

Table 3 – Related Work with Energy Saving in O-RAN

Works	Objective	Experimentation	xApp	rApp	Near-RT	Non-RT	SMO	Dynamic Demand
(AYALA-ROMERO et al., 2021)	Orchestration framework managing resources of base stations and mobile video analytics	Testbed	✓	✓	✓	✓	✗	✓
(ABEDIN et al., 2022)	Optimize elastic O-RAN slicing in a dynamic IIoT	Simulation	✗	✗	✓	✓	✗	✓
(HAMMAMI; NGUYEN, 2022)	Compare performance of on-policy and off-policy deep reinforcement learning models for resource allocation	Simulation	✗	✗	✓	✓	✓	✓
(WANG et al., 2022)	Design computation offloading strategy for O-RAN based IoT	Simulation	✗	✗	✓	✓	✗	✗
(KALNTIS; IOSIFIDIS, 2022)	Optimize performance and energy consumption of vBS	Testbed	✗	✓	✓	✓	✓	✓
(MAI et al., 2023)	Optimize energy efficiency in C-RAN by jointing resource allocation for delay-aware traffic	Simulation	✗	✗	✓	✓	✗	✓
(WANG et al., 2023)	Minimize energy consumption of IoT devices in O-RAN based IoT systems	Simulation	✗	✗	✓	✓	✓	✓
(AMIRI et al., 2023)	Propose energy-efficient dynamic VNF splitting method	Simulation	✗	✗	✗	✗	✗	✓
(SALVAT et al., 2023)	Showcase design principle of an O-RAN compliant testbed	Testbed	✗	✗	✓	✓	✓	✓
(LO SCHIAVO et al., 2024)	Enhance cost and energy efficiency in vRANs by leveraging shared pools of heterogeneous processors among DUs	Testbed	✗	✗	✓	✓	✗	✓
This Work	Dynamic resource management	Emulation and Simulation	✓	✓	✓	✓	✓	✓

Table 3 summarizes critical research efforts on energy savings in O-RAN environments, outlining objectives, experimentation methods, and involvement of various O-RAN components such as xApp, rApp, Near-RT RIC, Non-RT RIC, and SMO. This overview highlights how these works contribute to energy efficiency in O-RAN, particularly in addressing fluctuating user demands, a critical aspect for developing 6G networks. It is worth noting that the authors focusing on the Near-RT RIC but not proposing any xApp are primarily engaging in a conceptual architectural alignment within the O-RAN framework. This approach underlines the significance of foundational O-RAN elements in fostering a scalable and flexible energy-

efficient architecture without necessarily introducing new applications or functionalities. Such research emphasizes the importance of leveraging existing O-RAN components and their interplay to enhance network energy savings, showcasing a foundational approach towards achieving sustainability in next-generation networks.

The studies by (ABEDIN et al., 2022; WANG et al., 2023) emphasize the importance of dynamic resource allocation in O-RAN environments. (ABEDIN et al., 2022) focus on O-RAN slicing for the Industrial Internet of Things (IIoT), proposing a distributed matching game and deep reinforcement learning to optimize slicing for better performance. Similarly, (MAI et al., 2023) presents a groundbreaking approach by optimizing energy efficiency through joint resource allocation for delay-aware traffic in C-RAN systems, a key consideration for O-RAN. Their methodology, articulated through rigorous simulations, showcases a delicate balance between energy efficiency and QoS within the constraints of fronthaul link capacities. Uniquely, this study harnesses an iterative-based optimization algorithm integrated with Convolutional Neural Networks (CNN) for controlling RRH transmit power, illustrating the potential of merging computational intelligence with network resource management for energy conservation. This endeavor distinctly forgoes the direct application of xApps and rApps, instead relying on Near-RT and Non-RT RIC functionalities, and marks a significant step towards employing foundational O-RAN elements in fostering energy-efficient architectures.

In the recent study (LO SCHIAVO et al., 2024), the researchers introduce CloudRIC, an innovative framework aiming to optimize cost and energy efficiency within O-RAN architectures through shared heterogeneous computing resources. This approach is validated through testbed evaluations, demonstrating its potential to significantly reduce energy consumption and operational costs in vRANs. Notably, the study underlines the significance of leveraging Near-RT and Non-RT RIC functionalities, emphasizing a foundational approach towards achieving sustainability in next-generation networks without necessarily introducing new applications or functionalities. However, (WANG et al., 2023) proposes a computation offloading strategy in O-RAN-based IoT systems to minimize energy consumption while ensuring QoS, employing a Successive Convex Approximation (SCA) algorithm to solve a non-convex optimization problem.

The works (AMIRI et al., 2023; SALVAT et al., 2023) contribute to this field by proposing energy-efficient RAN virtualization and testbed implementation methods, respectively. (AMIRI et al., 2023) utilize Deep Reinforcement Learning for a dynamic VNF splitting method in O-RAN, aiming to adapt to varying traffic conditions for energy efficiency. (SALVAT et al., 2023) present an O-RAN-compliant testbed for ML research, providing valuable datasets for analyzing computing usage, energy consumption, and performance in vRAN deployments.

(AYALA-ROMERO et al., 2021; HAMMAMI; NGUYEN, 2022) explore the application of learning algorithms for resource management in O-RAN. (AYALA-ROMERO et al., 2021) propose a Bayesian learning framework for energy-aware control in mobile edge Artificial Intelligence (AI) services. In contrast, (HAMMAMI; NGUYEN, 2022) compares on-policy and

off-policy deep reinforcement learning models for resource allocation in O-RAN, particularly for real-time video surveillance applications. (WANG et al., 2022; KALNTIS; IOSIFIDIS, 2022) also contribute to the computation offloading and resource optimization theme in O-RAN. (WANG et al., 2022) propose a strategy to reduce energy consumption while meeting latency requirements in O-RAN based IoT systems. (KALNTIS; IOSIFIDIS, 2022) introduce an online learning algorithm for optimizing the performance and energy consumption of virtualized base stations (vBS) in O-RAN systems, focusing on efficient scheduling policies.

These works identify energy-savings strategies within the O-RAN framework as the specific use case that most accurately represents fluctuating user demand in 6G networks. By leveraging techniques such as Carrier and Cell Switch Off/On, RF Channel Reconfiguration, and Sleep Modes, O-RAN can dynamically adapt to changing user demands, optimizing energy efficiency and operational performance. This adaptability is crucial for managing the variability in user activity, which is a hallmark of 6G networks. Each of these works addressed aspects of fluctuating user demand in 6G networks, focusing on energy efficiency, resource management, and the adaptability of network infrastructure to dynamic conditions. Collectively, they provide insights into the evolving landscape of O-RAN and its potential role in shaping the future of 6G networks. The implementation of these strategies through intelligent automation in the RIC and SMO demonstrates how O-RAN can effectively respond to real-time fluctuations in user demand, ensuring both sustainability and efficiency. Therefore, the energy-saving strategies detailed in this thesis answer the question by providing a practical and impactful example of how O-RAN can manage dynamic user demands in future telecommunication networks.

3.3 Assessing O-RAN Orchestration for Fluctuating User Demands

The dynamic landscape of vNG-RAN necessitates a comprehensive evaluation of orchestration strategies that are responsive to variable user demands, encompassing the performance and scalability of NFV and O-RAN architectures. Metrics are crucial in this context, serving as standard quantities for network performance and reliability assessments, and are vital for the QoS in mobile network environments.

Specific metrics recommended for nodes in NFVI include CPU utilization, network parameters such as counter measurement time, and memory occupancy measured in Bytes. The 3GPP further expands this metric spectrum for 6G networks, categorizing them into two domains: those dependent on software implementations of VNFs and those collected from the infrastructure hosting these VNFs. This approach includes metrics for network elements such as gNB, core network components, and standard metrics applicable to all VNFs.

Practical evaluation of orchestration strategies in O-RAN architectures requires qualitative and quantitative assessments, particularly in response to fluctuating user demands. Key criteria include scalability, flexibility, and rapid adaptability to changing network conditions. Evaluation methods often involve advanced simulation tools and real-time monitoring systems that

track key performance indicators such as latency, throughput, and resource utilization. These indicators provide insights into the orchestration strategy's effectiveness in managing network loads, allocating resources, and maintaining QoS under varying conditions. Integrating AI and ML algorithms in the evaluation process further enhances predictive capabilities and proactive adjustments, solidifying the orchestration strategy's responsiveness to user demands and network dynamics.

In this section, we address the orchestration strategies for O-RAN in response to the dynamic demands of vNG-RAN. Subsection 3.3.1 presents recent advances in VNF orchestration, emphasizing the integration of AI and ML for effective network management. Subsection 3.3.2 explores O-RAN's performance specifications, highlighting their role in shaping robust and efficient RAN infrastructures. This section, as a whole, offers a comprehensive overview of the strategies and standards driving the evolution of RAN technologies in the era of 5G and beyond, focusing on adaptability, scalability, and efficiency.

3.3.1 VNF orchestration works aligned with O-RAN

This section presents the recent advancements in VNF orchestration within the O-RAN framework, as highlighted in the comprehensive Table 4. The table methodically categorizes significant works in this domain, detailing their objectives, experimentation methodologies, and the integration of various O-RAN components such as xApps, rApps, Near-RT RIC and Non-RT RIC, SMO, and dynamic demand management. This collation facilitates an in-depth understanding of the diverse and innovative strategies employed to optimize and enhance the efficiency of 6G O-RAN architectures, particularly highlighting the role of optimization techniques in proactive network management and resource allocation.

The paper by (BRUNO et al., 2024a) explores the deployment of a disaggregated Near-RT RIC on a distributed cloud infrastructure. This work introduces an optimization model to minimize placement costs while meeting latency requirements and presents performance evaluations comparing distributed and monolithic strategies. The findings indicate significant cost savings and efficiency improvements in cloud-native environments, particularly relevant for 6G networks. This study contributes valuable insights into the practical implementation and benefits of disaggregated RIC architectures in modern RAN deployments.

The collection of works (HOJEIJ et al., 2023; KAZEMIFARD et al., 2021; MORAIS et al., 2023; DUONG et al., 2022) represents a significant stride in optimizing 5G O-RAN architectures. These articles collectively address the challenges of efficient function placement and resource allocation in O-RAN networks. They introduce innovative methodologies, such as Integer Linear Programming, Recurrent Neural Network models, gradient-based algorithms, and column generation techniques. These approaches enhance network capabilities, ensure robust deployments, and manage dynamic demands. They also address optimizing the user admittance ratio, minimizing network delays, and maximizing network availability, highlighting O-RAN

Table 4 – Related Work with RAN orchestration in O-RAN.

Works	Objective	Experimentation	xApp	rApp	Near-RT	Non-RT	SMO	Dynamic Demand
(BRUNO et al., 2024a)	Evaluate the deployment of a disaggregated near-RT RIC on a distributed cloud infrastructure.	Practical	✓	✗	✓	✗	✗	✓
(HOJEIJ et al., 2023)	Optimize the placement of O-CU and O-DU.	Simulation	✓	✗	✓	✓	✗	✓
(KAZEMIFARD et al., 2021)	Mathematical model for the problem of CNF placement and resource allocation.	Simulation	✗	✗	✓	✓	✓	✓
(MORAIS et al., 2023)	Propose orchestrator that supports the dynamic placement of RAN functions.	Practical	✗	✗	✓	✓	✓	✓
(DUONG et al., 2022)	Propose a decomposition model and an efficient column generation algorithm to maximize the yearly availability of O-RAN VNFs.	Simulation	✗	✗	✓	✓	✗	✓
(ALI; JAMMAL, 2023)	Implement a proactive elastic orchestration framework for dynamic resource allocation using ML and RL.	Simulation	✗	✗	✓	✓	✓	✓
(SARIKAYA et al., 2021)	Optimize the placement of RAN slices in a multi-tier 5G O-RAN architecture.	Simulation	✗	✗	✗	✗	✗	✓
(NAHUM et al., 2020)	Present a low-cost 5G mobile network testbed with a virtualized and orchestrated structure using containers, focusing on integration with AI.	Practical	✗	✗	✗	✗	✓	✓
This Work	**Dynamic resource management**	**Emulation and Simulation**	✓	✓	✓	✓	✓	✓

systems' evolving complexity and potential.

The research by (ALI; JAMMAL, 2023) and (NAHUM et al., 2020) focuses on the integration of AI and ML in 5G networks. These articles demonstrate the application of AI in forecasting traffic patterns and managing network resources. The proposed AI-driven frameworks and testbeds are pivotal in enhancing the adaptability and efficiency of network operations. This work represents a paradigm shift in 5G networking. AI plays a crucial role in proactive network management, optimizing VNF placements, and facilitating real-world testing scenarios to drive forward the capabilities of 5G networks.

Study (SARIKAYA et al., 2021) presents a unique perspective on the deployment strategies within 5G networks, specifically focusing on the strategic placement of RAN slices. The research comprehensively analyzes multi-tier O-RAN architectures, emphasizing the importance of flexible, functional splits. This approach significantly enhances network resource utilization and overall network performance. The study's advanced modelling techniques provide a deeper understanding of the operational dynamics in 5G networks and offer innovative solutions for optimizing network slice deployments.

3.3.2 Performance discussion based on O-RAN specifications

The O-RAN architecture is crucial for the future of telecommunications, especially in the era of 5G and beyond. O-RAN Alliance's specifications are vital in establishing robust, efficient, scalable network infrastructures. These specifications address various aspects of RAN performance, emphasizing the importance of latency, data processing capabilities, and interoperability

among network components. This subsection explores the specific performance-related aspects of O-RAN specifications, underlining their importance in advancing RAN technologies. Table 5 provides a summarized overview of these key performance metrics.

Table 5 – Summary of Key Performance Metrics in O-RAN Specifications

Specification	**Key Focus Area**	**Performance Metrics**
O-RAN.WG4.CUS	Fronthaul Interfaces	Latency < 100 microseconds, Jitter, Synchronization
O-RAN.WG2.A1UCR	A1 Interface	Data processing speeds up to 10 Gbps
O-RAN.WG3.E2GAP	E2 Interface	Response times < 10 milliseconds
O-RAN.WG5.C.1	Open Interfaces	Data rates of several Gbps, Low latency
O-RAN.WG6.ORCH-USE-CASES	O-Cloud	Robust computing, Efficient storage, Strong networking
O-RAN.WG11.SECURITY	Security	High network availability, Consistent performance

In the O-RAN Fronthaul Control, User and Synchronization Plane Specification (O-RAN Alliance, 2023h), critical performance metrics for open fronthaul interfaces are thoroughly outlined. The document focuses on latency, essential for timely data transmission between the RU and DU, targeting a latency below 100 microseconds for certain data types. Moreover, it addresses jitter, affecting real-time data transmission quality and synchronization, which is crucial for consistent performance across distributed RAN components. These interfaces are designed to handle high data rates and low latency requirements, vital for 5G and subsequent network generations.

The O-RAN A1 interface Use Cases and Requirements (O-RAN Alliance, 2023i) specification addresses the performance requirements of the A1 interface, linking the Non-RT RIC and Near-RT RIC. It highlights the need for efficient data processing capabilities, essential for complex policy decisions and AI/ML model integration, and aims to support data processing speeds up to 10 gigabits per second. This specification ensures timely and effective RAN optimization strategies, significantly enhancing RAN performance.

O-RAN E2 General Aspects and Principles (O-RAN Alliance, 2023d) focuses on the performance of the E2 interface. It underscores the need for rapid data processing and minimal latency in real-time control and optimization of RAN resources. The specification sets stringent response time requirements, aiming to keep them under 10 milliseconds. This response time is crucial for the swift and efficient exchange of control messages and commands between the Near-RT RIC and RAN nodes.

The O-RAN New Radio (NR) C-plane profile (O-RAN Alliance, 2023j) provides comprehensive performance guidelines for open interfaces such as F1, W1, X2, and Xn. The emphasis is on data transport and signalling efficiency, supporting high data throughput and low latency communications, with these interfaces designed to handle data rates of several gigabits per second. O-RAN Cloudification and Orchestration Use Cases and Requirements for O-RAN

Virtualized RAN (O-RAN Alliance, 2023k) discusses the performance of the O-Cloud, emphasizing robust computing power, efficient storage solutions, and strong networking capabilities. These elements are critical for the scalability and adaptability of cloud-based RAN functions, ensuring effective response to varying network demands without compromising performance.

Lastly, the O-RAN Security Aspects in O-RAN Specifications (O-RAN Alliance, 2023l), while primarily focusing on security, indirectly influences performance by stressing the importance of resilient and secure network operations. It explains how robust security measures are crucial for maintaining high network availability and consistent performance, ensuring the O-RAN system's reliability and efficiency in the face of security threats.

This work identifies and proposes specific criteria and methods to effectively evaluate orchestration strategies in response to fluctuating user demands within the O-RAN framework. These criteria include scalability, flexibility, rapid adaptability to changing network conditions, and adherence to O-RAN specifications, such as throughput and resource utilization. The proposed orchestration strategy can dynamically optimize resources, ensuring efficient and agile network management. This approach provides a comprehensive evaluation framework that addresses the complexities and variability of modern telecommunication networks, demonstrating how O-RAN can effectively respond to fluctuating user demands and maintain optimal performance. Therefore, the thesis answers the question by detailing a robust set of criteria and methods for assessing the efficacy of orchestration strategies in managing dynamic user demands in future 6G networks.

3.4 Dynamic Near-RT RIC Disaggregation Strategies in O-RAN for Variable User Demands

In exploring strategies for the adequate disaggregation of Near-RT RIC functions in O-RAN architectures, it is crucial to consider the dynamic nature of user demands and scalability requirements. The literature offers various approaches, each with unique advantages and challenges.

The centralized and distributed models for Near-RT RIC implementation are examined in (DRYJAńSKI; KLIKS, 2022), highlighting the trade-offs between unified decision-making and specialized optimization. The centralized model is excellent in global optimization but may struggle with scalability in large RANs. At the same time, the distributed approach, managing each E2 node type (O-CU, O-DU or O-RAN eNB) individually, offers specialized optimization but may lack a holistic network view.

Integrating Federated Learning (FL) into the RIC architecture for 5G slicing services is explored in (SINGH; KHOA NGUYEN, 2022), suggesting an adaptive learning model responsive to fluctuating demands. However, the framework's implementation within RIC remains underexplored. Similarly, (HUFF; HILTUNEN; DUARTE, 2021) focuses on fault tolerance strategies in RIC, offering a RAN Intelligent Control Fault Tolerance (RFT) library for distributed RIC

but does not address component placement in disaggregated infrastructures.

(D'ORO et al., 2022) introduces the OrchestRAN framework for Non-RT RIC, promoting intelligent orchestration for varying demands. Although it assumes a complete Near-RT RIC instance for each E2 node cluster, its scalability in large networks might be constrained. Meanwhile, (SCHMIDT; IRAZABAL; NIKAEIN, 2021) proposes FlexRIC, a service-oriented controller with a centralized, modular design, emphasizing extensibility and a minimal footprint for rapid adaptation to changing demands.

The disaggregation of the traditional RAN control plane is advocated in (BALASUBRAMANIAN et al., 2021), promoting a Near-RT RIC platform that separates control and data planes for enhanced network intelligence and adaptability using AI-driven applications. Additionally, the development of xApps and rApps, as demonstrated in (CAO et al., 2021), (CAO et al., 2022), and (JOHNSON; MAAS; MERWE, 2021), shows potential in intelligent user access control and O-RAN slicing to manage variable demands.

Table 6 – Summary of related work.

Works	Placement type*	Problem formulation	Real-world experiments	Approach
(DRYJAńSKI; KLIKS, 2022)	C/D_1	✗	✗	Conceptual
(D'ORO et al., 2022)	D_1	✓	✗	Federated Learning
(SINGH; KHOA NGUYEN, 2022)	C/D_1	✗	✓	Fault tolerance
(D'ORO et al., 2022)	C/D_1	✓	✓	Orchestration framework
(HUFF; HILTUNEN; DUARTE, 2021)	C	✗	✓	Architectural
(SCHMIDT; IRAZABAL; NIKAEIN, 2021)	C/D_1/D_2	✗	✗	Architectural
RIC-O (BRUNO et al., 2023b; ALMEIDA et al., 2023)	C/D_1/D_2	✓	✓	Optimal flexible placement

*Placement type: C - Centralized D_1 - Distributed D_2 - Disaggregated

3.5 Summarizing

In this chapter, we have conducted an extensive review of the current research landscape in O-RAN, focusing on pivotal questions integral to the development and optimization of these networks in the context of 6G advancements. The chapter began with a detailed examination of efficient O-RAN integration strategies for dynamic user demands, emphasizing resource allocation, data traffic management, and network orchestration. This was followed by an exploration of various use cases within the O-RAN framework, particularly those addressing the demands of future telecommunication networks through advancements in Massive MIMO, network slicing, and energy-saving strategies. We then assessed the orchestration strategies in O-RAN architectures, highlighting key evaluation criteria and methods necessary for practical implementations. Finally, we discussed strategies for disaggregating Near-RT RIC functions, crucial for managing fluctuating user demands and ensuring network scalability.

Each section has built upon the existing body of knowledge, contributing to a deeper understanding of how O-RAN can be optimized for the rapidly evolving demands of modern telecommunications. The insights and findings presented in this chapter set the stage for the next chapter, where we will introduce the proposed framework for adaptive network management in 6G O-RAN. This framework addresses dynamic user demands, focusing on energy efficiency and

leveraging the proposed architectural components and strategies. We discuss the design decisions, architectural components, and methodologies that underpin this innovative approach. In the next chapter, *Framework Architecture and Use Case*, we provide a comprehensive overview of our proposed solution, detailing how it integrates with the O-RAN architecture to enhance network performance and sustainability.

4 FRAMEWORK ARCHITECTURE AND USE CASE

This chapter introduces the foundational architectural components and use case scenarios that form the backbone of the adaptive network management framework for 6G O-RAN. By focusing on dynamic user demand and energy efficiency in our use case, we outline key project decisions and elaborate on the synergistic interaction between various components within the O-RAN architecture.

Section 4.1 details the strategic project decisions to enhance energy efficiency and optimize network performance. These decisions include the deployment of Application on Non-Real-Time RAN Intelligent Controllers (rApps) for radio power optimization, the integration of architectural components such as SMO, Near-RT RIC, Non-RT RIC, xApps, and rApps, and the implementation of adaptive energy management and enhanced monitoring capabilities.

In Section 4.2, we present a comprehensive overview of the core architectural components that constitute the adaptive network management framework for RAN. This section underscores the roles of the SMO, Near-RT RIC, and Non-RT RIC, detailing their individual functionalities and integrated operations within the larger network framework.

The specialized applications developed as a use case for our framework are discussed in Section 4.3. Here, we explore the role of rApps, such as the rApp Energy Savings, in optimizing energy consumption and managing network resources. We also introduce proprietary xApp like the "xApp Handover," which enhances energy efficiency across the RAN.

4.1 Project Decisions

The project decisions for the Adaptive Network Management in 6G O-RAN framework, focusing on dynamic user demand that supports different use cases, including energy efficiency, are grounded in strategic considerations within the O-RAN architecture and operational paradigms. Key decisions include:

4.1.1 Architectural Components

This subsection outlines the key architectural components utilized in the O-RAN framework, emphasizing their roles and the rationale behind their integration.

Utilization of rApp for Optimization: The architecture leverages an rApp for radio power optimization and handover policy management through O1 and A1 interfaces. This promotes energy efficiency and a seamless user experience, aligning with the intelligent, near-real-time network management principles advocated by the Hybrid Management Plane Model from the O-RAN Open Fronthaul Interfaces WG Management Plane Specification (O-RAN Alliance, 2024b).

Synergy Between Architectural Components: The integration of SMO, Near-RT RIC,

Non-RT RIC, xApps, and rApps is emphasized to ensure that energy-saving policies are based on comprehensive data analysis and effectively implemented. This integration optimizes network performance while minimizing energy consumption, fully aligned with O-RAN architecture.

Adaptive Energy Management: Incorporating adaptive mechanisms for energy management allows the network to respond dynamically to changes in user demand and network load. This approach balances energy efficiency with network performance by optimizing energy consumption strategies.

Enhanced Monitoring for Real-Time Decision Making: xApp monitoring capabilities provide granular network insights, enabling data-driven decision-making processes. This monitoring is crucial for optimizing network configurations and effectively implementing our energy-saving policies.

R1 Interface: Due to the low maturity of the Non-RT RIC's integration with the SMO at the implementation time, we decided not to use the R1 interface for data management and exposure services. However, we utilized the R1 interface for service management and exposure services related to A1. This selective use of R1 allowed us to leverage its capabilities where it was mature enough to be effective while avoiding areas where it was not fully developed.

4.1.2 Implementation Decisions

This subsection details the specific modifications and implementation decisions made to meet the project's requirements and optimize the deployment within the existing framework.

Modified Components for Specific Requirements: To meet specific deployment requirements, several components have been modified:

- *VespaMgr on the Near-RT RIC:* The default VespaMgr configurations from the Open Radio Access Network Software Community (OSC) were inadequate. Modifications include increased flexibility for dynamic configuration changes and enhanced data handling capacity. For instance, the VNF Event Streaming (VES) collector Uniform Resource Locator (URL) is now configurable through ConfigMap files, and the supported maximum size of Hypertext Transfer Protocol (HTTP) requests was increased to accommodate larger data volumes.

- *A1 Mediator on the Near-RT RIC:* The current Non-RT RIC release only supports version 1 of the A1 Application Programming Interface (API), which contains a bug causing policy instances to be sent without the subId. Our modifications ensure policy instances include the necessary ID for proper functionality and integration.

- *Persistent Storage in Kubernetes (K8s):* K8s lacks a default storage class for enabling pods to use persistent storage. The recommended solution is the Local Path Provisioner

from Rancher, which simplifies deployment and configuration of persistent storage for K8s pods, ensuring reliable and consistent storage management.

- *Policy Management Service:* Integrating the Near-RT RIC with the Non-RT RIC required modifications to the Policy Management Service to locate and connect to the A1 Mediator's address, ensuring seamless communication and integration between these components.

- *Customized SMO Components:* The SMO components from the OSC are deprecated and do not support all message types defined by the VES API. To address this, we developed custom VesCollector and InfluxDBConnector components, following the patterns of the O-RAN SMO. Additionally, we use Kafka, Prometheus-Blackbox-Exporter, InfluxDB, and Chronograf, which are tailored to meet our operational needs. A tailored recipe or an override YAML Ain't Markup Language (YAML) with these components can be created for streamlined deployment and integration.

4.2 Architectural Components

The architectural components illustrated in Figure 11 are fundamental to the adaptive network management framework for RANs. This architecture includes three primary components: the SMO, the Near-RT RIC, and the Non-RT RIC. The SMO, incorporating components such as the Data Lake, Data River, and A1 Policy Management System, oversees the higher-level management functions. The Near-RT RIC, deployed on the Open Cloud (O-Cloud), manages real-time operations through components such as xApp Monitoring, ensuring quick handovers and efficient monitoring of the network state. The Near-RT RIC interfaces directly with E2Ns, which can be dynamically turned on or set to standby mode by the RF Environment Manager to optimize resource usage.

This section presents the core architectural components that constitute the backbone of our Adaptive Network Management for RANs. This exploration is critical for understanding the synergistic interplay between various elements and their collective contribution to the system's efficiency, resilience, and adaptability. Our focus is on the individual functionalities of each component and their integrated operations within the larger network framework. Key components, as illustrated in Figure 11, include the SMO, Near-RT RIC, and the Non-RT RIC, each playing a distinct yet interconnected role in the network's architecture.

4.2.1 Service Management and Orchestration

In the quest to address the limitations of the pre-existing SMO framework provided by the OSC, we embarked on the development of an optimized SMO, with a concentrated focus on the O1 interface for VES. This subsection delineates the methodological advancements

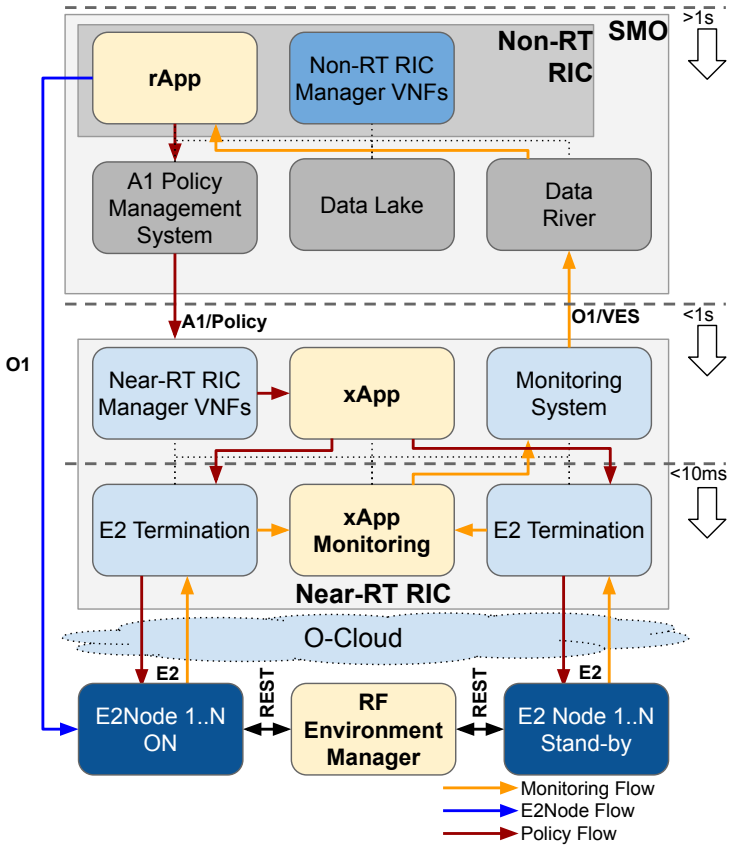


Figure 11 – Adaptive network management framework architecture.

and deployment strategies instrumental in our refined SMO implementation. Initially, the VES Prometheus Adapter (VESPA) agent within the Near-RT RIC underwent comprehensive modifications. These adaptations were critical in ensuring compatibility with the O1 interface specifications, enabling efficient communication and data handling (O-RAN Alliance, 2023m). Subsequently, implementing a VES collector provided a centralized aggregation point for event data streaming from the VESPA. The collector's role was pivotal in the filtering and preprocessing of data before its dissemination to the data streaming infrastructure.

The culmination of our SMO architecture saw the integration of rApps, designed to interface with the data lake or data river, contingent upon their application-specific requirements. This final integration stage underscored the versatility of our SMO framework, allowing for dynamic data consumption — either in real-time through the data river or via historical analysis from the data lake, as depicted in Figure 11.

Our SMO framework incorporates a critical Monitoring Flow, which gathers comprehensive data, including computer resources in O-Cloud, throughput, UE numbers, energy efficiency, and E2N metrics. The flow begins with data collection at the E2N level, utilizing E2AP to transmit data to xApp Monitoring.

4.2.2 Non-RT RIC

The Non-RT RIC plays a pivotal role in the O-RAN architecture by enabling intelligent management and orchestration of RAN elements over a longer time scale. In our implementation, we utilized the Non-RT RIC from the OSC, which provides a robust framework for integrating and managing RAN components. This subsection elucidates the architectural components and functional integration of the Non-RT RIC, as illustrated in Figure 11.

The Non-RT RIC encompasses several key elements. The rApps, Non-RT RIC Applications, are responsible for executing non-real-time data analysis and policy management tasks. These applications interact with other Non-RT RIC components to ensure optimal RAN performance and resource allocation. The Non-RT RIC Manager VNFs manage the various aspects of the Non-RT RIC, providing an interface between the rApps and the underlying infrastructure. The A1 Interface (A1) Policy Management System facilitates communication between the Non-RT RIC and Near-RT RIC, enabling the enforcement of policy decisions derived from rApp analytics. The Data Lake stores historical data, whereas the Data River handles real-time data streams. These components ensure that rApps have access to comprehensive datasets for analysis and decision-making.

In our implementation, the Non-RT RIC is integrated with the SMO framework, enhancing its capability to manage network resources efficiently. The Non-RT RIC leverages the A1 interface to communicate policy directives to the Near-RT RIC, thereby optimizing RAN operations. The data flow within the Non-RT RIC begins with collecting and aggregating data from the network via the Data River. This real-time data is then analyzed by rApps, which generates actionable insights and policies. These policies are communicated to the Near-RT RIC through the A1 interface, ensuring that the network operates in accordance with the determined strategies.

The integration of the Non-RT RIC within the broader SMO framework underscores its critical role in facilitating advanced RAN management functionalities. By leveraging the extensive data processing and analytical capabilities of the Non-RT RIC, our architecture ensures robust and adaptive network performance, aligning with the dynamic requirements of modern wireless communication systems.

4.2.3 Near-RT RIC

The Near-RT RIC is essential for managing and optimizing RAN functions on a near-real-time basis. This subsection details the architectural components and functional roles of the Near-RT RIC, as depicted in Figure 11.

The Near-RT RIC architecture comprises several critical elements. The Near-RT RIC Manager VNFs oversee the management of various near-real-time tasks, ensuring that the RIC operates efficiently. The xApps are specialized applications designed to perform specific tasks such

as monitoring and control. These xApps interact with both the Near-RT RIC Manager VNFs and the monitoring system to ensure optimal network performance.

The E2 Termination (E2T) components facilitate the communication between the Near-RT RIC and the underlying radio network nodes (E2Ns). This communication is crucial for the real-time monitoring and controlling of the RAN elements. The xApp Monitoring collects data from the E2T components, providing insights into network performance and health. The monitoring system within the Near-RT RIC is designed to continuously gather and analyze data from various RAN components. This system utilizes xApps to perform specific monitoring tasks, such as tracking network latency, throughput, and error rates. By analyzing this data, the monitoring system can detect anomalies and potential performance issues, allowing for proactive management and optimization of the RAN. Furthermore, the monitoring system provides real-time feedback to the Near-RT RIC Manager and to the Non-RT RIC using O1/VES interface, enabling dynamic adjustments to network configurations and ensuring sustained optimal performance.

In our implementation, the Near-RT RIC is integrated with the O-Cloud, enhancing its capability to manage computing resources effectively. The data flow within the Near-RT RIC starts with the E2Ns, which can be in either active or standby mode, as managed by the RF Environment Manager. The active E2Ns transmit data to the xApp Monitoring system via the E2T components. This data is then analyzed by the xApps, which provides actionable insights for real-time network optimization.

4.2.3.1 xApp Monitoring

In the complex RAN ecosystem, our proposed xApp Monitoring is a critical component, focusing on collecting and analyzing various performance metrics. This monitoring function is essential for assessing the RAN's real-time operational efficiency and responsiveness. A primary role of xApp Monitoring involves capturing data about UEs connections like the radio power of UE-connected antennae and of UE neighbor antennas. Specifically, the xApp Monitoring captures key metrics such as Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), Signal-to-Interference-plus-Noise Ratio (SINR), and Received Signal Strength Indicator (RSSI) for each connected gNB and neighboring UEs. These metrics are crucial for evaluating the quality of the radio link and the network's overall performance.

Beyond UEs measurement, xApp Monitoring also encompasses gathering data related to EE, network throughput, and radio power for each E2N. This comprehensive data collection is vital for operating the Energy Saver rApp of our use case. The integration and processing of this information by the Near-RT RIC and its transfer through the O1 Interface (O1)/VES interface to the Non-RT RIC equips the Energy Saver rApp with essential insights. These insights are critical for optimizing network energy consumption and promoting sustainable and efficient network operation.

Additionally, xApp Monitoring extends its capabilities to the O-Cloud, which monitors computing resource utilization, including CPU and memory usage, and the latency between computing nodes. This monitoring aspect is increasingly important in modern cloud-reliant RAN architectures, where efficient management of computational resources directly impacts network performance and user experience. The data gathered by xApp Monitoring is thus pivotal in maintaining optimal cloud infrastructure operations, ensuring that the RAN system remains robust and adaptive to varying network demands and conditions.

The integration of the Near-RT RIC within the broader SMO framework underscores its pivotal role in ensuring efficient RAN operations. By leveraging the real-time data processing and analytical capabilities of the Near-RT RIC, our architecture achieves a high degree of adaptability and performance, meeting the dynamic requirements of modern wireless communication systems.

4.3 Energy Savings Use Case

This section discusses specialized applications, known as rApps, such as the **rApp Energy Savings**, which are integral to our innovative approach. These applications, designed for specific tasks within the RAN infrastructure, contribute significantly to optimizing various aspects such as energy consumption, data flow management, and network configuration. Their roles and interactions, as outlined in this section, are essential for a comprehensive understanding of the architectural framework and its operational dynamics.

4.3.1 Non-RT RIC

We have deployed the Non-RT RIC from the OSC (ALLIANCE, 2023). Within the Non-RT RIC, we have integrated one rApp developed in-house: **rApp Energy Savings**. This rApp enables the Non-RT RIC to perform sophisticated, non-real-time decision-making and policy setting, informed by extensive data analysis and insight extraction. The rApp Energy Savings is engineered to enhance energy consumption efficiency across the RAN by leveraging predictive analytics, which utilizes historical data to formulate energy-saving policies.

rApp Energy Savings

The proposed rApp Energy Savings is designed to optimize energy utilization across the network. It functions by collecting and analyzing a wealth of data about energy efficiency, network throughput, and user connectivity metrics for each E2N. This crucial data is provided through the SMO framework, with the xApp Monitoring system playing a vital role in its acquisition. Utilizing this data, the rApp Energy Savings employs optimization techniques to formulate comprehensive policies that determine the operational states for each E2N. These policies indicate

which E2Ns should be active and which can be temporarily deactivated, thereby minimizing energy consumption without compromising network performance and user experience. The resulting policy, a product of intricate data analysis and optimization is then relayed to the xApp Handover for implementation.

In this work, we analyze a highly dynamic network scenario in a sports stadium, where the number of users and communication demands can change over time, requiring the network to adapt its resource allocation based on the current demand. We consider a set $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$ of time steps, where at each time step, the number of active users to the network and their requirements, as well as channel quality, can vary. To model the RAN environment, we define a set $\mathcal{U}^t = \{u_1, u_2, \ldots, u_{|\mathcal{U}^t|}\}$ representing the set of users at the sports stadium at time step $t \in \mathcal{T}$. Each user $u_i \in \mathcal{U}^t$ is characterized by their throughput demand $\lambda^t_{u_i} \in \mathbb{R}_{\geq 0}$ at time step $t \in \mathcal{T}$, measured in bits per second (bps). Additionally, we consider a set $\mathcal{R} = \{r_1, r_2, \ldots, r_{|\mathcal{R}|}\}$ of O-RUs, where each O-RU $r_i \in \mathcal{R}$ is characterized by its maximum transmission power capacity $\gamma_{r_i} \in \mathbb{R}_{\geq 0}$ and maximum bandwidth capacity $\rho_{r_i} \in \mathbb{R}_{\geq 0}$. Figures 1 and 2 illustrate the network behavior at two different time steps, highlighting scenarios of higher demand and lower demand in the sports stadium network.

The optimization framework incorporates triggers based on the fluctuation of UE connections, introducing a dynamic aspect to the energy-saver feature. Specifically, a trigger mechanism is activated in response to a 20% variation in UE connections over a predefined interval, mirroring a significant change in network usage. This fluctuation may either represent an upsurge, necessitating a scaled augmentation of resources to adeptly handle the increased load without detracting from energy efficiency, or a downturn, indicating a reduced network load. In the latter scenario, the system is optimized to scale down operations, thereby conserving energy while ensuring the network's integrity and performance are not compromised.

Furthermore, including a synchronous timer-based trigger offers an additional layer of flexibility. This trigger, configured to activate at fixed intervals, complements the user variation-based trigger by providing a regular, time-based assessment of the network's energy efficiency needs. This approach maximizes energy savings without sacrificing service quality, even without significant changes in UE connections.

Building on the flexibility provided by these dynamic triggers, the optimization approach determines the activation or deactivation of RUs based on coverage overlap, user count, data throughput, and energy efficiency. We aim to minimize the transmission power of O-RUs while accommodating all users and meeting their demands. We define four decision variables to represent this goal:

- $\mathbf{x}^t_{u_i,r_j} = \{0, 1\}$, which defines if the user $u_i \in \mathcal{U}^t$ is associated with O-RU $r_j \in \mathcal{R}$ at time step $t \in \mathcal{T}$.

- $\mathbf{y}^t_{u_i,r_j} \in \mathbb{R}_{\geq 0}$, which defines the bandwidth allocated for the user $u_i \in \mathcal{U}^t$ in O-RU $r_j \in \mathcal{R}$ at time step $t \in \mathcal{T}$.

- $\mathbf{w}_{r_j}^t \in \mathbb{R}_{\geq 0}$, which represents the transmission power selected for the O-RU $r_j \in \mathcal{R}$ at time step $t \in \mathcal{T}$.
- $\mathbf{z}_{r_j}^t = \{0, 1\}$, which represents whether the O-RU $r_j \in \mathcal{R}$ is activated or not, at the time step $t \in \mathcal{T}$.

We formulate our objective function as:

$$\textit{minimize} \qquad \left(\sum_{t \in \mathcal{T}} \sum_{r_j \in \mathcal{R}} \frac{\mathbf{w}_{r_j}^t}{\eta} + \theta_{r_j}^{RF} \right), \tag{4.1}$$

where $\theta_{r_j}^{RF}$ represents the static power consumption of O-RU $r_j \in \mathcal{R}$. This static power consumption is an inherent part of the O-RU's operation, independent of the dynamic power adjustments based on network demands. In the following, we present the problem constraints.

Every user $u_i \in \mathcal{U}^t$ must be associated with one O-RU $r_j \in \mathcal{R}$ at each time step $t \in \mathcal{T}$:

$$\sum_{r_j \in \mathcal{R}} \mathbf{x}_{u_i, r_j}^t = 1, \qquad \forall u_i \in \mathcal{U}^t, t \in \mathcal{T}. \tag{4.2}$$

The transmission power $\mathbf{w}_{r_j}^t$ assigned to O-RU $r_j \in \mathcal{R}$ at time step $t \in \mathcal{T}$ must be positive and respect its maximum power capacity γ_{r_j}:

$$0 \leq \mathbf{w}_{r_j}^t \leq \gamma_{r_i}, \qquad \forall r_j \in \mathcal{R}, t \in \mathcal{T}. \tag{4.3}$$

If the O-RU $r_j \in \mathcal{R}$ is inactive ($\mathbf{z}_{r_j}^t = 0$) at time step $t \in \mathcal{T}$, its transmission power $\mathbf{w}_{r_j}^t$ must be zero:

$$\mathbf{w}_{r_j}^t \leq \mathbf{z}_{r_j}^t \gamma_{r_j}, \qquad \forall r_j \in \mathcal{R}, t \in \mathcal{T}. \tag{4.4}$$

If the O-RU $r_j \in \mathcal{R}$ is active ($\mathbf{z}_{r_j}^t = 1$), its transmission power $\mathbf{w}_{r_j}^t$ must be positive and greater than zero:

$$\mathbf{z}_{r_j}^t \epsilon \leq \mathbf{w}_{r_j}^t, \qquad \forall r_j \in \mathcal{R}, t \in \mathcal{T}, \tag{4.5}$$

where ϵ is a small positive constant.

The total bandwidth $\sum_{u_i \in \mathcal{U}} \mathbf{y}_{u_i, r_j}^t$ used by an O-RU $r_j \in \mathcal{R}$ that is active ($\mathbf{z}_{r_j}^t = 1$) must not exceed its maximum bandwidth capacity ρ_{r_j} can be expressed as:

$$\sum_{u_i \in \mathcal{U}} \mathbf{y}_{u_i, r_j}^t \leq \rho_{r_i} \mathbf{z}_{r_j}^t \qquad \forall r_j \in \mathcal{R}, t \in \mathcal{T}. \tag{4.6}$$

To ensure that the allocated bandwidth $\mathbf{y}_{u_i, r_j}^t$ meets the throughput demand $\gamma_{u_i}^t$ of every user $u_i \in \mathcal{U}^t$ at each time step t, we employed the Shannon's capacity equation to design the last constraint:

$$\sum_{r_j \in G} \left(\mathbf{y}_{u_i, r_j}^t \cdot \log \left(1 + \frac{S}{N} \right) \right) \geq \gamma_{u_i}^t \forall r_j \in \mathcal{R}, u_i \in \mathcal{U}^t, t \in \mathcal{T}, \tag{4.7}$$

where $\frac{S}{N}$ is the Signal-to-Noise Ratio (SNR), which can be calculated for a user $u_i \in \mathcal{U}^t$ associated with an O-RU $r_j \in \mathcal{R}$ given the noise and interference ratio σ^2 and the channel gain $\beta(u_i, r_j)$, as follows:

$$\frac{S}{N} = \frac{\beta(u_i, r_j)\mathbf{w}_{r_j}^t}{\sigma^2}. \tag{4.8}$$

The formulation presented is a Mixed Integer Linear Programming (MILP) problem, which is well-known to be NP-hard. The presence of binary variables introduces combinatorial behavior, leading to significant computational complexity. Our MILP problem involves continuous variables, i.e., transmission power and bandwidth allocations, and binary variables, i.e., those indicating whether a user is associated with a specific O-RU or an O-RU is active.

The dynamic nature of the triggers and the optimization approach highlights the importance of real-time data processing and adaptive policy adjustments in achieving energy-efficient RAN operations. This mechanism ensures that the network remains robust and adaptive to varying demands, promoting sustainable and efficient management.

4.3.2 Near-RT RIC

We use the Near-RT RIC from the OSC, exploiting its advanced capabilities to orchestrate and optimize RAN operations (ALLIANCE, 2023). Leveraging the robust architecture of the Near-RT RIC, our study introduces two proprietary xApps developed in-house: the "xApp Handover" and the "xApp Monitoring." These applications, designed to interface seamlessly with RAN elements via the standardized E2 protocol, allow for unprecedented granular control over RAN behaviors and policies. The "xApp Handover" aims to optimize energy consumption across the RAN, while the "xApp Monitoring" provides enhanced real-time analytics and monitoring capabilities.

xApp Handover

The proposed xApp Handover is designed to enhance network energy efficiency. This xApp operates under the guidance of specific energy-saving policies from the rApp Energy Savings. The core function of the xApp Handover is to dynamically adjust the network's operational state in response to these policies. When the policy dictates the activation of new E2Ns (gNBs), the xApp Handover facilitates this process by activating the required E2Ns. This activation is not just a matter of powering on additional nodes; it also involves network load balancing. The xApp achieves this by managing the number connected in each E2N.

4.3.3 Components Interaction Overview

The adaptive network management framework for 6G O-RAN incorporates design choices that ensure dynamic user demand satisfaction while maintaining energy efficiency. These design choices are depicted in Figure 12, which illustrates the sequence of operations within the network management framework.

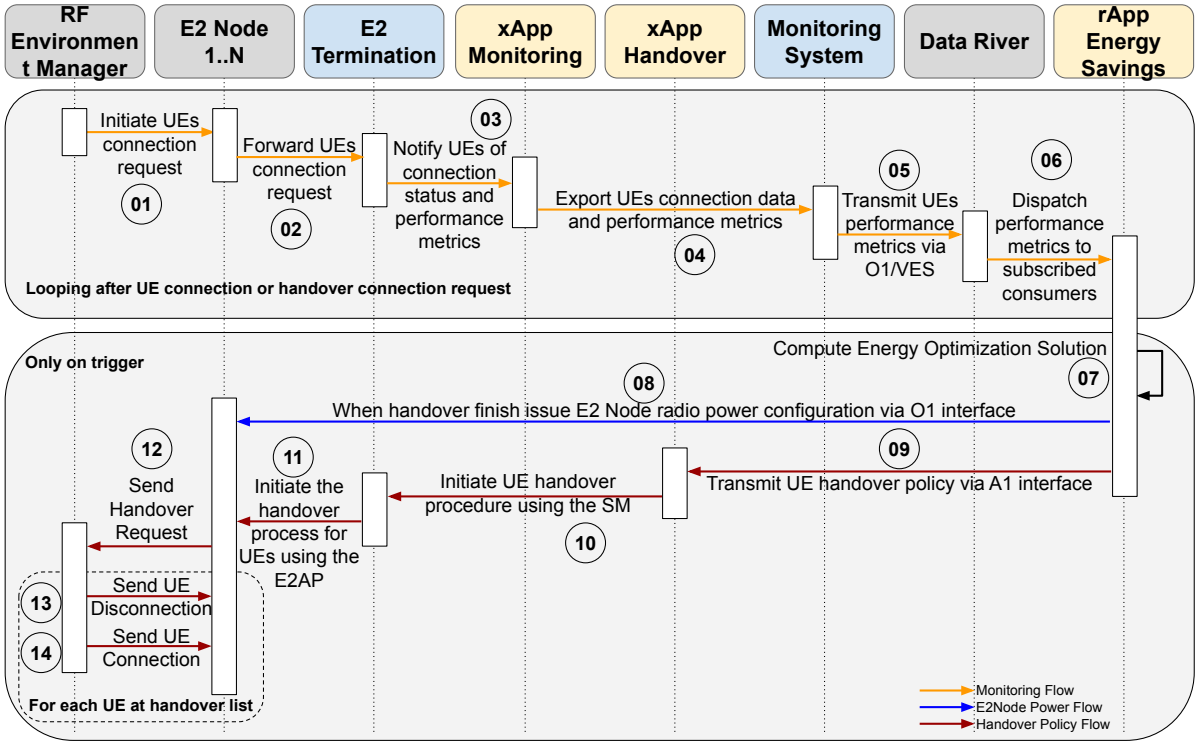


Figure 12 – Prototype Sequence.

Figure 12 illustrates the interactions between components in managing UEs. The process begins with the UE Manager initiating a UE connection request (01), which is forwarded by the E2N (02). Upon successful connection, the xApp Monitoring component notifies the UEs of their connection status and performance metrics (03). These metrics are exported to the Monitoring System (04) and transmitted to the SMO via the O1/VES interface (05). The Data River component of the SMO dispatches these performance metrics to subscribed consumers (06). Concurrently, the rApp Energy Savings computes the energy optimization solution (07) and issues the E2N radio power configuration via the O1 interface (08). The UE handover policy is transmitted through the A1 interface (09). The UE handover procedure, using the State Machine (SM), is initiated (10), triggering the handover process for UEs using the E2AP. A handover request is sent upon a specific trigger (11), and upon completion, UE disconnection (12) and re-connection (13) sequences are initiated. Finally, the E2N adjusts the radio power configuration post-handover (14), completing the process. Steps 01 to 06 run on every UE connection or handover process; steps 07 to 14 run only on specific triggers.

4.4 Summarizing

This chapter has explored pioneering strategies and architectural innovations designed to advance network management within the 6G O-RAN paradigm. Emphasizing the dynamic features of O-RAN's open and intelligent architecture, we detailed the deployment of rApps and xApps for optimizing E2N radio power management and handover policies. This approach addresses the crucial requirements for energy efficiency and enhanced user experiences.

We provided an in-depth overview of the architectural components essential for this framework, including the SMO, Near-RT RIC, and Non-RT RIC, and discussed the integration of specialized applications such as the rApp Energy Savings. Key project decisions were outlined, focusing on the utilization of rApps for optimization, the synergy between architectural components, adaptive energy management, and enhanced monitoring for real-time decision-making. Additionally, we examined the modifications made to various components to meet specific deployment requirements.

This comprehensive approach sets the stage for the next chapter, which addresses implementing and evaluating our proposed adaptive network management framework. The forthcoming chapter provides a detailed analysis of the performance and efficiency of the proposed solutions, further demonstrating their potential to meet the dynamic demands of future 6G networks.

5 EVALUATION METHODOLOGY

The chapter on evaluation methodology focuses on the systematic assessment of optimization strategies within the O-RAN framework, specifically aimed at enhancing the performance, scalability, and efficiency of RAN operations. The chapter is meticulously structured, beginning with an extensive overview of the experimental environment in Section 5.1. This section provides a detailed architectural description within the O-RAN framework, emphasizing the collaborative integration of the Non-RT RIC and the Near-RT RIC. Additionally, it highlights the crucial role of the SMO framework and discusses significant enhancements in E2N simulation capabilities alongside the introduction of the UE Manager. These elements collectively lay the groundwork for a comprehensive experimental exploration to effectively simulate real-world network scenarios and behaviors.

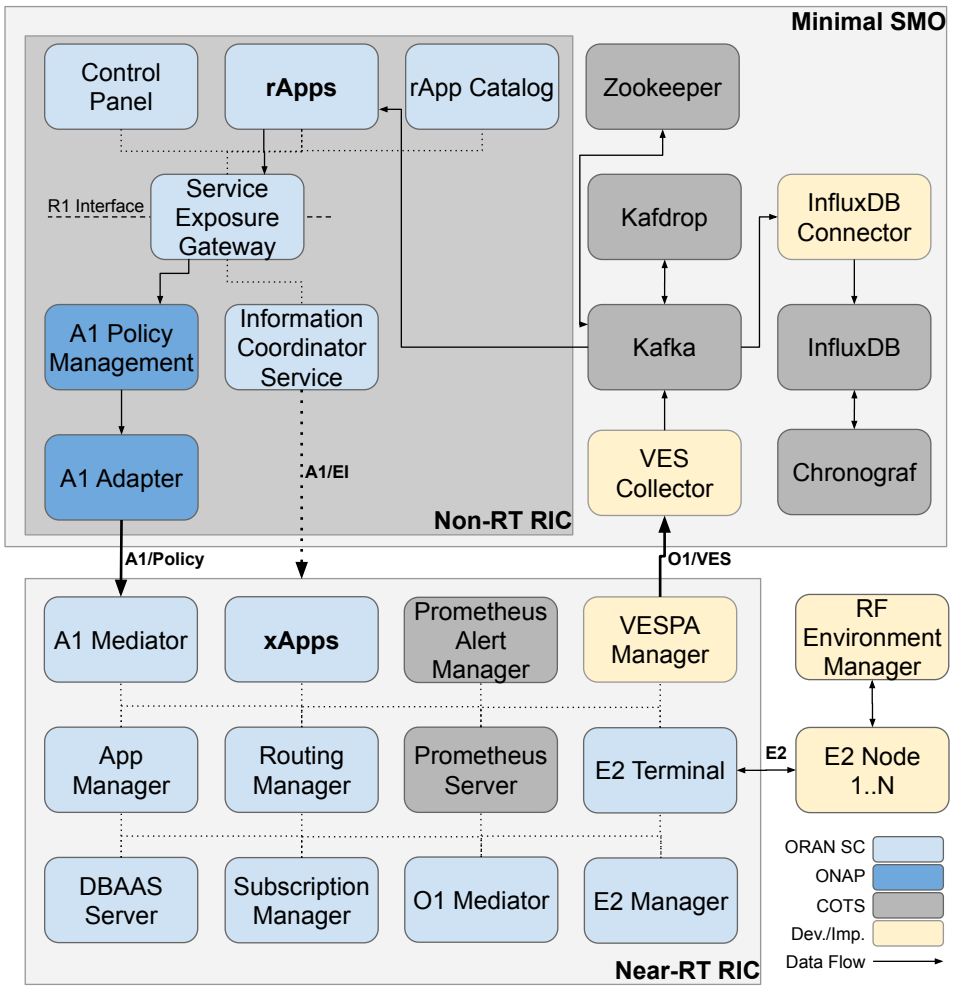
Building upon the experimental setup, Section 5.2 introduces a specific use-case scenario designed to test the framework's operational efficacy within a dynamic and high-demand environment, such as a soccer stadium during a match, representing a challenging high-density network environment. The narrative then progresses to Section 5.3, where a series of experiments are detailed. These experiments are crucial for evaluating the network's different operational and management facets, focusing on energy consumption, optimization response time, and computational resource utilization. Through these systematic assessments, the chapter provides a nuanced understanding of the potential advantages and inherent challenges of integrating optimization methodologies within RAN operations. This exploration is particularly pertinent for navigating and overcoming the complexities of managing network operations in environments that pose significant demands on the infrastructure.

5.1 Experimental Environment

This study presents an experimental environment architected within the O-RAN framework, aiming to explore integrating optimization methodologies in RAN operations for enhanced performance, scalability, and efficiency. Our experimental setup, illustrated in Figure 13, encompasses a multifaceted approach, leveraging the interaction between the Non-RT RIC and the Near-RT RIC, as outlined in Sections 5.1.1 and 5.1.2 respectively. Additionally, we explore the pivotal role of the SMO framework in orchestrating operations across RAN components, detailed in Section 5.1.3. This investigation extends to the E2N enhancements for simulation capabilities, covered in Section 5.1.4, facilitating a deeper understanding of network behaviours and application performance. Moreover, we introduce the UE Manager in Section 5.1.5, a tool designed to instantiate and manage UEs within our experimental RAN setup, offering insights into connection quality metrics and handover processes. Together, these components form the backbone of our experimental exploration, aiming to address current limitations and unlock future potentials within the O-RAN ecosystem.

We conducted the experiments in a K8s cluster environment, version 1.21, comprising four worker nodes, each configured as a Virtual Machine (VM) with four vCPUs, 8 GB RAM, and 50 GB virtual disk space. These worker nodes were managed by a master node on a VMs equipped with eight vCPUs, 16 GB RAM, and 100 GB virtual disk space. All VMs operated on Ubuntu 22.04 and were hosted on a DELL PowerEdge M610 server featuring four Intel Xeon X5660 processors and 192 GB RAM, utilizing VMware ESXi 6.7 as the hypervisor.

Figure 13 – Experimental Environment Components



Based on (ALLIANCE, 2023).

5.1.1 Non-RT RIC

Several components interact to enhance and manage network operations in the depicted architecture of a Non-RT RIC as part of an O-RAN system. The Service Exposure Gateway is an intermediary, facilitating secure and structured communication between RAN applications and the services within the Non-RT RIC. The Information Coordinator Service centralizes data collection, aggregation, and dissemination from various sources, providing a hub for information

essential to other components or applications within the Non-RT RIC.

The A1 Policy Management component oversees policies that govern the control and optimization of RAN resources and services. It achieves this by interfacing with the A1 Adapter, which acts as a conduit, converting A1 protocol messages into a form intelligible to internal components of the Non-RT RIC. Alongside, an A1 Mediator likely facilitates communication and policy implementation between the A1 Policy Management and applications designed to enact these policies, such as xApps.

Further, the architecture includes rApps, which are sophisticated RAN applications that optimize the network by applying heuristics or techniques such as AI or ML to refine network performance and resource management. The rApp Catalog serves as a repository, maintaining a list of available rApps, and enables network operators to identify and integrate new rApps into the RIC platform.

These components collectively deliver dynamic optimizations and resource management, marking a significant evolution from the traditional RAN architectures, as they enable a more adaptable and scalable network control.

5.1.2 Near-RT RIC

The Near-RT RIC within the O-RAN architecture embodies the cutting edge in orchestrating and controlling RAN elements on a real-time basis, from sub-seconds to seconds. It is instrumental in enabling instantaneous decision-making and control, critical for maintaining the efficiency and responsiveness of the RAN. At its core, the E2N interfaces directly with RAN nodes, backed by the E2T that manages low-level protocol intricacies, providing a seamless and efficient conduit for monitoring and management tasks.

Key to orchestrating the Near-RT RIC's functions are the E2 Manager and Subscription Manager. The former oversees session management, ensuring robust connectivity between the Near-RT RIC and the E2N—vital for maintaining control loops—while the latter manages information flows crucial for real-time updates on network performance. This symbiosis of management ensures that the RIC is always abreast of the network state, ready to respond to dynamic changes.

Central to the efficacy of the Near-RT RIC are specialized xApps, namely the xApp handover and xApp monitoring, as described in Chapter 4. These applications are paramount, with the former optimizing real-time handover decisions to enhance connectivity and user experience and the latter dedicated to meticulously monitoring RAN performance. Leveraging the E2 interface, these xApps utilize real-time data to implement advanced control and optimization strategies, showcasing the Near-RT RIC's capabilities in ensuring an optimally performing and reliable RAN.

The Prometheus Server and Database as a Service (DBaaS) Server support the operation and management of these xApps. The Prometheus Server and the Prometheus Alert Manager form

the backbone of the monitoring and alerting framework within the Near-RT RIC, collecting metrics and managing alerts to address potential issues preemptively. Meanwhile, the DBaaS Server facilitates essential database services, enabling efficient data management crucial for the xApps' operations.

In response to the limitations identified in the OSC SMO framework, significant enhancements have been made to the VESPA Manager within the Near-RT RIC. This strategic improvement specifically targets the efficient integration and management of event streams, a necessity given the critical roles played by the xApp handover and xApp monitoring. By fortifying the VESPA Manager, the Near-RT RIC not only overcomes the identified framework limitations but also underscores its pivotal role in delivering real-time, intelligent control over the RAN, ensuring a network environment that is both high-performing and dependable.

5.1.3 Minimal SMO

The Service Management and Orchestration (SMO) framework is an integral part of the O-RAN architecture, overseeing the operation and management of the RAN and RIC components. It adopts a holistic approach to service management across the radio access network, ensuring efficient and cohesive operations.

A key feature within the SMO framework is the Minimal SMO, designed for smaller or simpler deployment scenarios. It comprises essential elements that enable it to manage network operations effectively, even on a scaled-down level. Among these components are the Zookeeper, which provides centralized configuration and synchronization services, and Kafdrop, an open-source interface for managing Kafka topics and consumer groups. Kafka serves as the critical messaging infrastructure, allowing asynchronous communication across O-RAN components, while the InfluxDB Connector facilitates the integration with InfluxDB for time-series data management.

Adopting these tools and components, such as InfluxDB for time-series data storage and analysis and Chronograf for data visualization, allows for efficient network performance monitoring and analysis. The VES Collector is pivotal in aggregating data from the VNF Event Stream for analytical and monitoring purposes.

In response to the limitations identified in the standard OSC SMO framework, we developed specific enhancements, particularly in VES collection and InfluxDB connectivity. By focusing on these enhancements, we address the gaps in the standard framework and provide a more robust network management and orchestration solution. While we developed these specific components in-house to meet our precise requirements, other Minimal SMO elements, such as Zookeeper, Kafka, and Chronograf, were utilized off the shelf. This hybrid approach allows us to leverage the strengths of existing solutions while customizing key aspects to overcome the inherent limitations of the SMO framework as defined by the OSC. Through this strategic combination of in-house development and external tools, we ensure that our deployment of

the Minimal SMO complies with O-RAN standards and is enhanced to address our network operations' specific needs and challenges.

5.1.4 Extended E2N

The evolution of the E2sim tool, as delineated in our previous works (ALMEIDA et al., 2023; BRUNO et al., 2023b), augments the O-RAN ecosystem's simulation framework by integrating advanced functionalities. These include the simulation of UE behaviors, their radio metrics, and handover processes, accessible through a Representational State Transfer (REST) API. This iteration signifies a leap forward in simulation capabilities, offering a nuanced emulation environment for developing, testing, and optimizing RIC applications.

The tool's capacity for simulating UE interactions and radio metrics within the RAN ecosystem allows for an in-depth evaluation of RIC applications, including xApps and rApps. It explicitly addresses scenarios related to connectivity management, radio resource allocation, and handover execution. This enhanced feature set facilitates a comprehensive analysis of application performance and network resource utilization, thus informing better decision-making in application development and network management.

Moreover, the tool's advanced simulation capabilities bolster interoperability testing between RIC applications and the E2 interface. By accurately mimicking complex interactions among E2Ns, UEs, and network infrastructure, the tool ensures comprehensive application functionality validation under varied operational conditions, including dynamic handover processes and fluctuating radio environments.

The enhanced simulation environment unlocks new possibilities for research and development, particularly in examining complex network behaviors such as network density effects on RAN performance. It opens avenues for investigating sophisticated ML algorithms for predictive handover, resource optimization, and overall network performance improvements.

Additionally, the expanded simulation capabilities render E2sim a more potent educational tool, enabling learners to explore the operational challenges of RAN management, understand handover mechanisms in depth, and evaluate the effectiveness of RIC applications in optimizing network and UE performance metrics.

In conclusion, the enhancements introduced in the latest version of E2sim, encompassing the simulation of UE behavior, radio metrics analysis, and handover processes through a REST API, represent a significant progression in simulation tools for RIC application development, testing, and optimization. This advancement is pivotal in fostering intelligent, efficient, open radio access networks within the O-RAN framework.

5.1.5 RF Environment Manager

Within the ambit of our experimental framework, the RF Environment Manager emerges as a pivotal component, architected explicitly for the instantiation of UEs on E2Ns amidst our meticulously simulated RAN milieu. This module's essence lies in its ability to simulate the air interface interaction among user devices and antennas. This facilitates a granular analysis of vital connection quality metrics that are pivotal for evaluating RAN performance and efficiency. Notably, the RF Environment Manager is ingeniously designed to encompass handover support, a critical functionality that simulates UE transition across different cells or E2Ns—a cornerstone for appraising the efficacy and reliability of handover protocols within the network infrastructure.

Leveraging Python for its development, the RF Environment Manager, with vectorization at its core, to markedly enhance the efficiency of RAN parameter computations. This strategic approach permits the rapid calculation of parameters, numbering in the millions, thus positioning the RF Environment Manager as an indispensable tool for the stress testing of xApps, alongside evaluating their adeptness in managing real-world operational loads.

The RAN parameter calculations are underpinned by a rigorous process that meticulously measures the power received by user devices from a static antenna. The underlying model for these calculations is comprehensive, encapsulating a wide array of variables including, but not limited to, antenna coordinates and gain, signal frequency, bandwidth, and power emission, thereby accurately portraying the connection quality dynamics between cell antennas and user devices.

In an innovative leap, the RF Environment Manager computes and conveys critical metrics such as RSRP, RSRQ, Channel Quality Index (CQI), SINR, and Block Error Rate (BLER) to an instance of E2sim. Further augmenting its capabilities, the RF Environment Manager facilitates the simulation of handover events, enabling an in-depth exploration of handover procedures and their consequent impact on connection quality and network performance. E2sim, serving as a virtual E2N, is instrumental in relaying updates on the network's state and handover events to Near-RT RICs through the E2 interface, with these interactions streamlined via a REST API and the data formatted in JavaScript Object Notation (JSON) for optimal integration.

The equation governing the power received by each UE is articulated through the Friis transmission equation as follows:

$$p_r = p_t \cdot g_t \cdot g_r \cdot \left(\frac{c}{4\pi f_t \cdot \|l_t - l_r\|_2} \right)^2 \tag{5.1}$$

Here, p_r delineates the power received, p_t the power transmitted, with g_t and g_r representing the gains of the transmitter and receiver, respectively, c symbolizing the speed of light, f_t the frequency of transmission, and $\|l_t - l_r\|_2$ the Euclidean distance between the transmitter and receiver.

The path loss, $L_{t,r}$, is subsequently calculated as:

$$L_{t,r} = 10 \cdot \gamma \cdot \log_{10}\left(\frac{p_t}{p_r}\right) + \chi_\sigma \tag{5.2}$$

where γ epitomizes the environmental loss factor and χ_σ is a random variable that simulates shadowing effects, sourced from a normal distribution with standard deviation σ.

The Signal-to-Noise Ratio (SNR) equation is presented as:

$$SNR_{t,r} = 10\log_{10}\left(\frac{p_r}{o_r}\right) \tag{5.3}$$

with o_r designating the noise power, which is calculated as the thermal noise within the receiver antenna.

The meticulous incorporation of RSRP and RSRQ calculations, alongside the innovative inclusion of handover support in the RF Environment Manager, not only exemplifies the module's adeptness in simulating dynamic RAN environments but also underscores its critical role in the holistic development and optimization of RAN management strategies and xApps within the O-RAN framework.

5.2 Implementing energy savings use case

This section meticulously delineates a use case scenario to validate our experimental framework's operational efficacy and advanced capabilities within a dynamic and demanding environment. The selected use case is emblematic of high-density, high-demand network environments, exemplified explicitly by a soccer stadium during a match, where the connectivity demands and dynamic user engagement present a rigorous testbed for our O-RAN-based solutions.

Table 7 – Experiment Parameters

Simulation Parameter	Value
Frequency (f_t)	6 channels starting at 7125 MHz
Bandwidth (b_t)	100 MHz
Antenna Gain (g_t)	8 dB
User Device Gain (g_r)	2 dB
Power Emitted (p_t)	1 W
Channel Numerology (n_t)	4
Number of User Devices	10,000
Number of Cells	6
Environmental Loss Factor (γ)	1
Shadowing Standard Deviation (σ)	7.9

The configuration meticulously emulates cellular traffic within a soccer stadium setting, where both UEs and antennas are statically positioned. Antennas are strategically deployed on either side of the field, elevated 10 meters above ground level, and equally spaced to optimize

coverage. Conversely, UEs are randomly distributed across the grandstands surrounding the field and positioned at distances ranging from 5 to 47 meters from the perimeter. The grandstands' architecture, sloping away from the field at a 25-degree angle and incorporating a 2-meter step midway, enhances the realism of the stadium environment. The parameters outlined in Table 7 comprehensively detail the experiment's setup.

To accurately assess the effectiveness of our optimization, we establish critical performance metrics, including network throughput and EE. A simulation environment developed in Python3 models a realistic network scenario integrating multiple E2Ns and UEs. Enhancements in realism and relevance are achieved by incorporating insights from previous studies, which offer valuable perspectives on crowd dynamics and density patterns. These insights are pivotal for accurately simulating UE connections under varying scenarios. By integrating these dynamics, we craft scenarios that closely mimic real-world network conditions, featuring diverse user densities, mobility patterns, and traffic distributions. This simulation scenario ensures a comprehensive evaluation of network performance in a soccer stadium context.

The device association mechanism prioritizes connection attempts with each cell based on the CQI and RSRQ, continuing until either a timeout occurs or a response is received. Denied connection attempts prompt the UE to progress to the next cell in the list sequentially. Measured critical metrics during the experiment encompass the latency of association requests, the average and standard deviation of connected devices per cell, the mean and standard deviation of association attempts, the total count of association timeouts, and the cumulative duration required for the network to achieve its intended objectives (intent settle). Following network intent stabilization, the average SNR is calculated, serving as a crucial indicator of the QoS.

To augment our analysis on energy consumption dynamics within C-RANs, (MAI et al., 2023) provides insights into optimizing computing and power resources. Strategies including sleeping techniques and beamformer optimization are crucial for enhancing EE while adhering to QoS requirements. Considering not just the BBU pool but also RRHs and FH links emphasizes the need for comprehensive strategies to diminish energy consumption and bolster efficiency in next-generation wireless networks. This holistic approach is essential in our experiment, facilitating detailed energy utilization assessment in high-density, high-demand settings, such as a soccer stadium during a match.

Furthermore, integrating throughput requirements is imperative, particularly for applications such as video streaming, which constitutes a significant data traffic portion in high-density scenarios such as a soccer stadium during a match. Information from (VdoCipher, 2023) highlights that video streaming bandwidth requirements fluctuate substantially based on video quality, ranging from 0.5 Mbps for low-resolution videos to 25 Mbps for ultra-high-definition content. These requirements are vital for ensuring optimal user experience, directly influencing the network's load and demand. In our experiment's context, understanding these bandwidth necessities allows for a more precise simulation of network conditions, thereby accurately evaluating our network's capacity to deliver high-quality video content to many users concurrently. This

consideration is paramount for affirming the efficacy of our O-RAN-based solutions in meeting the QoS demands of video streaming in densely populated environments.

5.3 Energy savings experiments

This section details the experimental methodologies and setups used to evaluate the energy consumption, application performance, and end-to-end optimization within the 6G O-RAN framework. The experiments are designed to replicate realistic, high-density network conditions, such as those encountered in large-scale events or urban environments, to provide comprehensive insights into the network's performance and efficiency. The key focus areas include the analysis of energy consumption, Section 5.3.1, the evaluation of various xApps and rApp, Section 5.3.2, the assessment of end-to-end energy optimization processes, Section 5.3.3, and the overall analysis of resource utilization and energy efficiency, Section 5.3.4. Each subsection outlines the configurations, procedures, and metrics used to ensure rigorous and reliable results, which are crucial for optimizing network management and achieving sustainable operations in dynamic O-RAN environments.

5.3.1 Energy Consumption Analysis

The evaluation methodology for analyzing energy consumption within the O-RAN framework involves a meticulously designed experimental environment to simulate the high-density network conditions encountered during large-scale events, such as those in a soccer stadium. The simulation setup includes the following key components and procedures:

- **Network Configuration:** The network environment consists of multiple E2Ns distributed to cover the stadium area. The initial setup includes 17 E2Ns, each capable of dynamic activation and deactivation based on the network load. This configuration ensures the simulation of realistic network conditions.

- **User Equipment (UE):** The number of UEs varies from 16 to 1024 to simulate different spectator densities. Each UE is configured to generate traffic patterns typical of users in a stadium, including voice calls, video streaming, and social media usage. This variation helps in assessing the impact of user density on energy consumption.

- **Energy Savings rApp:** The energy savings rApp is deployed to manage the signal power of each E2N. It dynamically adjusts the signal power based on real-time demand, aiming to optimize energy consumption without compromising QoS. This implementation is critical for evaluating the effectiveness of energy-saving strategies.

- **Baseline Scenario:** For comparison, a baseline scenario is established where all E2Ns operate at 100% signal power continuously, regardless of the actual network demand.

This scenario serves as a benchmark to evaluate the effectiveness of the energy savings rApp, highlighting potential energy savings.

- **Simulation Parameters:** The simulation employs a set of predefined parameters, including antenna characteristics, signal propagation models, and traffic patterns. These parameters are aligned with typical stadium environments to ensure realistic and relevant results, thereby enhancing the applicability of the findings.

- **Data Collection:** Throughout the simulation, key metrics such as total energy consumption, mean signal power per antenna, and QoS indicators are continuously monitored and recorded. Each experiment is repeated ten times to gather sufficient data for statistical analysis, including calculating averages and standard deviations. This rigorous data collection ensures the reliability of the results.

By meticulously configuring and controlling these elements, the evaluation methodology provides a robust framework for assessing the energy consumption and efficiency of the O-RAN framework under varying network loads. The dynamic adjustment capabilities of the energy savings rApp are tested rigorously to assess their impact on energy usage and network performance, particularly during high-demand periods such as spectator entry in a stadium. This environment ensures that the results are reliable and applicable to real-world scenarios, thereby providing valuable insights into the potential benefits and limitations of deploying optimization methodologies in RAN operations.

5.3.2 Apps Evaluation

The evaluation methodology for assessing the rApp Energy Savings, xApp Monitoring, and xApp Handover within the O-RAN framework involves meticulously designed experiments replicating high-density network conditions and various load scenarios. The key components and procedures of this methodology are outlined below:

- **Network Setup:** The network configuration includes multiple E2Ns strategically deployed to provide comprehensive coverage of a simulated stadium environment. For rApp Energy Savings, the setup includes 17 E2Ns, while for xApp Handover, it involves up to 16 E2Ns. These E2Ns can be dynamically activated or deactivated based on network load, ensuring the simulation of realistic network conditions.

- **UE:** The number of UEs is varied from 16 to 1024 for both rApp Energy Savings and xApp Handover and from 4 to 256 for xApp Monitoring to simulate different user densities. Each UE generates realistic traffic patterns, including voice calls, video streaming, and social media interactions, emulating the behavior of users in high-density scenarios. The maximum number of users evaluated is 1024 because the time required for the rApp

Energy Saver to run its optimization increases significantly with a higher number of users. This limit ensures that the optimization processes remain efficient and manageable within the constraints of the system's performance capabilities.

- **rApp Energy Savings Implementation:** The rApp Energy Savings mechanism is designed to manage the signal power of E2Ns. This application dynamically adjusts signal power in response to real-time demand to optimize energy consumption while maintaining QoS.

- **xApp Monitoring Implementation:** Initially, a single xApp is deployed to monitor network activities. This xApp collects and processes data related to network performance, such as CPU and memory usage, and the time required to complete monitoring tasks. Additional xApp instances are deployed as needed to assess scalability.

- **xApp Handover Implementation:** The xApp responsible for managing handovers handles the transition of UEs between E2Ns. This application monitors signal strength, user mobility, and network load to execute handovers efficiently, minimizing service disruption and maintaining QoS.

- **Baseline Scenarios:** For comparative purposes, baseline scenarios are established: all E2Ns operate at 100% signal power continuously for rApp Energy Savings, the network operates under low load conditions with minimal UEs for xApp Monitoring, and handovers are managed without xApp optimizations for xApp Handover. These baselines serve as benchmarks to evaluate the efficiency and performance improvements of the rApp and xApps.

- **Simulation Parameters:** The simulations use predefined parameters, such as antenna characteristics, signal propagation models, and user traffic patterns, to reflect real-world conditions. These parameters ensure that the simulation results are realistic and applicable.

- **Data Collection:** Key performance metrics, such as CPU and memory usage, and the time required to resolve network demands and execute handovers, are monitored and recorded throughout the experiments. Each experiment is conducted multiple times to gather statistical data, including averages and standard deviations, ensuring robust analysis.

This evaluation methodology provides a thorough framework for assessing the rApp Energy Savings' capability to manage energy consumption and resource utilization and the performance and scalability of the xApp Monitoring and xApp Handover mechanisms. By varying the number of UEs and closely monitoring the impact on CPU and memory usage, as well as the time required to address network demands and execute handovers, these experiments yield valuable

insights into the efficiency and scalability of these applications. This information is crucial for network operators to optimize resource allocation, improve performance, and ensure sustainable and reliable service delivery in dynamic and high-demand 6G O-RAN environments.

5.3.3 End-to-end Energy Optimization Looping

The evaluation methodology for assessing end-to-end energy optimization looping within the 6G O-RAN adaptive network management framework is designed to replicate realistic operational conditions. This setup ensures a comprehensive analysis of the start and end times for various components involved in the network management processes. The key components and procedures of this methodology are outlined below:

- **User Equipment (UE):** The number of UEs varies from 16 to 1024 to simulate different user density scenarios. Each UE is programmed to generate typical traffic patterns, including voice, data, and multimedia services, to create a realistic load on the network.
- **Component Functions:** The key components involved in the energy optimization looping process include:
  - **RF Environment Manager:** Manages the radio frequency environment, handling connections and handovers between E2Ns.
  - **E2N Nodes:** Provide network coverage and handle user connections.
  - **xApp Monitoring:** Monitors network performance metrics.
  - **Prometheus:** Collects and stores time-series data.
  - **Vespa Manager and Ves Collector:** Manage data collection and aggregation.
  - **Kafka:** Facilitates data streaming and processing.
  - **rApp Energy Savings:** Optimizes energy consumption by dynamically adjusting network configurations.
  - **xApp Handover:** Manages user handovers between E2Ns.
- **Simulation Parameters:** The simulation employs a comprehensive set of parameters, including antenna characteristics and signal propagation models. These parameters are chosen to accurately reflect real-world network conditions, ensuring the experimental results' validity.
- **Data Collection:** Key performance metrics such as the start and end times of each component's tasks, CPU and memory usage, and the time required to complete each task are monitored and recorded. Each experiment is repeated multiple times to gather sufficient data for statistical analysis, including averages and standard deviations. This rigorous data collection ensures the reliability and robustness of the results.

This evaluation methodology is designed to thoroughly assess the efficiency and performance of various components within the 6G O-RAN adaptive network management framework. By simulating different levels of user density and monitoring the start and end times of each component's tasks, the methodology evaluates how well the network can handle varying loads and maintain optimal performance. The collected data on task processing times, resource usage, and variability offers valuable insights into the efficiency of the energy optimization processes and the scalability of the network management system. These insights are crucial for network operators to optimize resource allocation, improve system performance, and ensure sustainable operations in dynamic and high-demand O-RAN environments.

5.3.4 Overall Analysis of Resource Utilization and Energy Efficiency

The evaluation methodology for the overall analysis of resource utilization and EE within the 6G O-RAN framework focuses on providing a comprehensive understanding of network performance under varying UE densities. The key components and procedures of this methodology are outlined below:

- **UE** The number of UEs ranges from 16 to 1024 to simulate different user density scenarios. Data points for 16 UEs were excluded due to normalization issues.
- **Performance Metrics:** The key metrics evaluated include:
  - **rApp CPU:** Measures the CPU usage for rApp functionalities, reflecting the processing demands.
  - **rApp Memory:** Assesses memory consumption for rApp operations, indicating the data handling requirements.
  - **xApp Handover CPU:** Evaluates CPU usage for xApp handover management, highlighting the computational intensity of handling user transitions.
  - **xApp Handover Memory:** Measures memory usage for xApp handover processes.
  - **xApp Monitoring CPU:** Tracks CPU usage for xApp monitoring tasks, crucial for real-time network adjustments.
  - **Energy Consumed:** Quantifies the network's total energy consumption under different load conditions.
  - **E2Ns ON:** Counts the number of active E2Ns, indicating the scalability of the network infrastructure.
- **Normalization and Visualization:** All performance metrics are normalized to facilitate comparative analysis. A radar plot (Figure 30) is utilized to visualize these normalized metrics, providing a clear comparative view of resource utilization and EE across different UE densities.

- **Trend Analysis:** The methodology involves analyzing trends in the data to understand the implications for network management:
  - **Resource Utilization:** Trends in rApp and xApp CPU and memory usage are examined to identify scalability challenges.
  - **Energy Efficiency:** Energy consumption trends are analyzed to assess the network's EE.
- **Simulation Parameters:** The simulation includes parameters such as antenna characteristics and signal propagation models to accurately reflect real-world network conditions, ensuring the validity of the results.
- **Data Collection:** Key metrics, including CPU and memory usage, energy consumption, and the number of active E2Ns, are collected. Each experiment is repeated multiple times to gather sufficient data for statistical analysis, including averages and standard deviations. This rigorous data collection ensures the reliability and robustness of the results.

This evaluation methodology provides a detailed assessment of resource utilization and energy efficiency within the 6G O-RAN framework. The methodology evaluates the network's ability to handle loads and maintain optimal performance by analyzing normalized performance metrics across varying UE densities. The insights gained from this analysis are crucial for developing strategies to optimize resource allocation, improve system performance, and ensure sustainable network operations in dynamic and high-demand environments.

5.4 Summarizing

In this chapter, we presented a comprehensive evaluation methodology designed to systematically assess the optimization strategies within the O-RAN framework, focusing on enhancing the performance, scalability, and efficiency of RAN operations. We began with an extensive overview of the experimental environment, detailing the collaborative integration of the Non-RT RIC, Near-RT RIC, and SMO framework, along with significant enhancements in E2N simulation capabilities and the introduction of the UE Manager. This setup provided a robust foundation for our experimental exploration.

We then introduced a specific use-case scenario designed to test the framework's operational efficacy within a dynamic and high-demand environment, such as a soccer stadium during a match. This use case was followed by detailing the series of experiments conducted to evaluate the network's various operational and management aspects, focusing on energy consumption, optimization response time, and computational resource utilization. Through these systematic assessments, we aimed to gain a nuanced understanding of the potential advantages

and challenges of integrating optimization methodologies within RAN operations, particularly in high-density, high-demand environments.

The insights and findings from this chapter pave the way for the next chapter, where we discuss the results and analysis of our experiments. In the upcoming chapter, *Results and Analysis*, we provide a detailed examination of the data collected, interpret the performance metrics, and evaluate the effectiveness of the proposed optimization strategies. Next chapter includes an in-depth analysis of energy consumption patterns, resource utilization, and overall network efficiency, offering a clear perspective on the practical implications and benefits of our adaptive network management framework in real-world scenarios.

6 RESULTS

This chapter presents the results derived from the studies detailed in Sections 6.1, 6.2, and 6.3. These foundational studies provide essential insights into monitoring strategies, the optimization of RICs, and their deployment. Moreover, they serve as the cornerstone for the results presented in Section 6.4, which showcases the energy-saving use case within the adaptive network management framework.

The research detailed in (BRUNO et al., 2023a) facilitated the acquisition of critical skills in designing and implementing complex empirical studies, including developing a cloud-native Beyond 5G (B5G) system testbed. This study involved deploying and orchestrating microservices on a K8s cluster, enhancing the understanding of cloud-native architectures and container orchestration. Integrating open-source observability tools, such as Prometheus, the Elasticsearch, Fluentd and, Kibana (EFK) stack, and Grafana, was crucial for system performance monitoring and anomaly diagnosis. Additionally, experience with COTS ML solutions for anomaly detection was gained, which included preprocessing log data and interpreting anomaly detection results. These activities deepened technical expertise and improved problem-solving and analytical skills in real-time network management and fault diagnosis in next-generation telecommunication systems. The primary results are presented in Section 6.1.

Research efforts (BRUNO et al., 2023b; ALMEIDA et al., 2023) also significantly enhanced the understanding of RANs architectures, particularly the O-RAN Alliance's open specifications. Practical knowledge was gained in disaggregating and distributing Near-RT RICs components to meet stringent latency requirements. Skills in edge and cloud computing were developed, optimizing the deployment of critical RAN components across the cloud-edge continuum for balanced performance and cost efficiency. Mastery in using K8s for managing complex network functions and applications was achieved, alongside problem-solving skills in formulating and applying heuristic and optimal resource allocation strategies in dynamic network environments. Hands-on experiments and analytical modeling provided insights into next-generation wireless networks' performance evaluation and scalability challenges. These findings are discussed in Section 6.2.

The research conducted in (BRUNO et al., 2024a) advanced proficiency in designing and implementing empirical studies, particularly for benchmarking and evaluating disaggregated Near-RT RIC deployment on distributed cloud infrastructures. Skills in optimization modeling for efficient component placement in cloud-native environments and latency-sensitive control loops were developed. The study involved managing geographically dispersed cloud sites, analyzing network latencies and costs, and improving technical writing and presentation skills suitable for high-impact scientific journals. The main results of this study are detailed in Section 6.3.

The final section, 6.4, presents the results of energy-saving strategies implemented within the adaptive network management framework for 6G O-RAN environments. This section

specifically evaluates the chosen use case to assess the proposed architecture, focusing on the rApp Energy Savings, xApp Handover, and xApp Monitoring in terms of CPU and memory usage and the response time to varying network demands. The results highlight the system's capability to optimize energy consumption and performance, ensuring efficient network operations in real-world 6G scenarios. This comprehensive evaluation underscores the practical applicability and impact of the research, emphasizing resource allocation and processing efficiency in modern telecommunication systems.

The integration and synthesis of these skills and findings culminate in the proposed adaptive framework, validated through the EE-saving strategies discussed in Section 6.4. This demonstrates the practical applicability and significance of the research, underscoring the importance of resource allocation and processing efficiency in contemporary telecommunication systems.

6.1 Observability

This section systematically presents the outcomes of our empirical investigation, segmented into three comprehensive subsections for enhanced clarity and focus. Subsection 6.1.1 evaluates the effectiveness of traditional cloud-native tools in detecting anomalies within B5G system. Subsection 6.1.2 demonstrates the insights gained from leveraging Elasticsearch's integrated ML features for system anomaly detection.

6.1.1 Anomaly Detection Using Observability Tools

We aim to detect anomalies using observability tools. In this context, we realize two experiments. First, the B5G system under study is subjected to a resource provisioning failure during a connectivity test using a ping tool between a UE and the data network. Second, we submit the deployed B5G system to a resource provisioning failure during a test between a UE and the data network, stressing the environment using the iPerf tool (BRUNO et al., 2023a).

Connectivity Test

In the first test scenario, the B5G system under study was subjected to a resource provisioning failure during a connectivity test between a UE and the data network. This test aimed to verify if metrics and logs extracted from the system can detect the injected failure. The following steps were performed to achieve the goal: the amount of CPU provisioned for the Pod that ran the CN services was undersized. This Pod received only half (0.25 Millicores) of the resources needed for its regular operation (0.5 Millicores). However, the memory had sufficient resources for all Pods, and RAN and CN Pods were started (E1). After the initialization of the system, the Ping tool injected probes into the data session established between UE and the data network, using UE as the probe source (E2). Finally, the generation of probes was ended

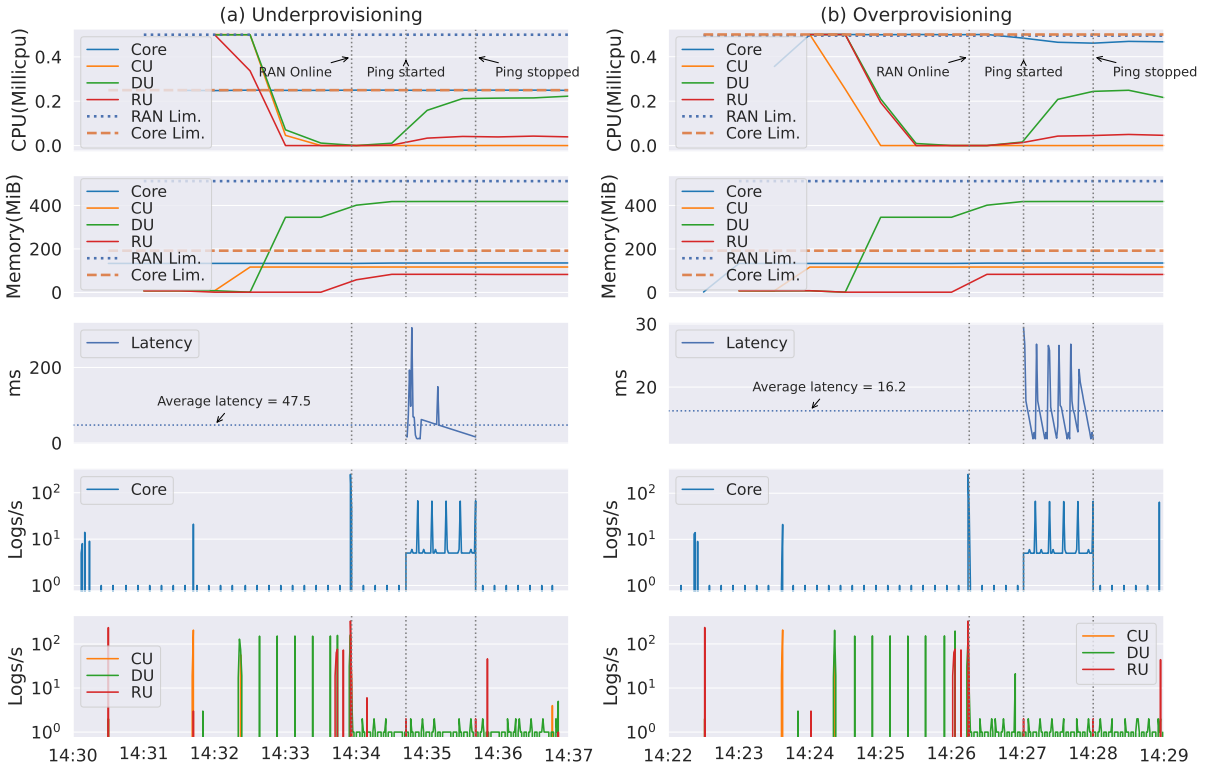


Figure 14 – Results for the connectivity test scenario. The first column (a) represents the results of the undersized feature test, while the second column (b) illustrates the results of the oversized resources test (BRUNO et al., 2023a).

(E3). The metrics analyzed in this scenario were the CPU and memory consumption of the Pods corresponding to the RAN and CN services and the average latency of the probes sent. In addition, logs extracted from Pods running these services were analyzed for the number of messages written per unit of time.

In this test, Figure 14 (a) shows the test results with resources undersized, while Figure 14 (b) illustrates the results of a similar test but where resources are overprovisioned. In the figures, the X-axis represents time throughout the tests. The vertical dotted lines on each graph represent the times when E1, E2, and E3 occurred in both tests. The first and second lines of Figure 14 represent the CPU and memory consumption of the Pods observed throughout the test. The "RAN Lim" and "Core Lim" represent the maximum provisioned resource values. The third line represents the observed latency for the probes sent throughout the tests and the average value obtained. The fourth and fifth lines represent the write rate in the CN and RAN service logs. It is observed that the CPU consumption in the CN Services Pod remains at the maximum limit throughout the test with undersized. Moreover, it is observed that the average latency of probes sent by Ping in this test is three times higher than the test with overprovisioning. These observations, in turn, are sufficient to identify and diagnose the problem of resources undersized in the CN services. However, as shown in the figures, the behavior of the logs throughout the two tests is very similar and does not contribute to detecting the anomaly. Therefore, we can conclude that, in this scenario, the collection and analysis of metrics are sufficient to detect the

injected abnormality.

Performance Test

In this evaluation, we submitted the deployed B5G system to a resource provisioning failure during a test between a UE and the data network. This test aimed to verify whether the extracted metrics and logs analysis were suitable for detecting the injected fault. To this end, we underestimated the CPU provisioned for the CN pod, receiving only a fraction (1 Millicore) of the resource required for its regular operation (4 Millicores). However, the memory was provisioned with sufficient resources for all pods. Therefore, we performed the following steps. We started RAN and CN pods (E1). After initializing the system, we generated traffic between UE and the data network using the Iperf tool (E2). However, after the first minute of traffic generation, we observed that the data session established between UE and the data network was abnormally terminated (E3). Since the data session was abnormally terminated, we finished traffic generation (E4). The metrics analyzed in this scenario were the CPU and memory consumption of the RAN and CN pods and the throughput obtained between UE and the data network.

We also analyzed the logs extracted from these pods. On the one hand, Figure 15 (a) shows the results of the experiment where the CPU was underprovisioned. On the other hand, Figure 15 (b) illustrates the results for a similar test with the CPU being overprovisioned. In both figures, the X-axis represents the time throughout the tests. The vertical dotted lines on each graphic in both figures represent the times when events (E1), (E2), (E3), and (E4) occurred. However, as the test with CPU overprovisioning did not present errors, event (E3) (abnormality) was not observed for this experiment. The first and second graphics in both figures represent the observed CPU and memory consumption of the pods during the experiments. The ”RAN Lim” and ”Core Lim” in these graphics represent the maximum resources provisioned for RAN and CN pods. The third graphic represents the throughput observed throughout the tests, while the fourth and fifth graphics represent the writing rate in the collected logs. Finally, the sixth graphic represents the moments when the DU log emitted an anomalous message. This event was not observed in the test with CPU overprovisioning. As traffic generation starts, we can observe that both experiments increased CPU consumption in RAN pods. Furthermore, the throughput dropped to zero after the UE data session ended abnormally in the test with CPU underprovisioning. However, the DU log emitted an abnormal message before this moment, as shown in the last graphic of Figure 15 (a). This message was issued repeatedly before the throughput dropped to zero. Therefore, standard cloud-native tools can detect the injected anomaly using metrics and logs from the deployed system. Indeed, log observation can detect the anomaly before it is captured by metric observation in this case.

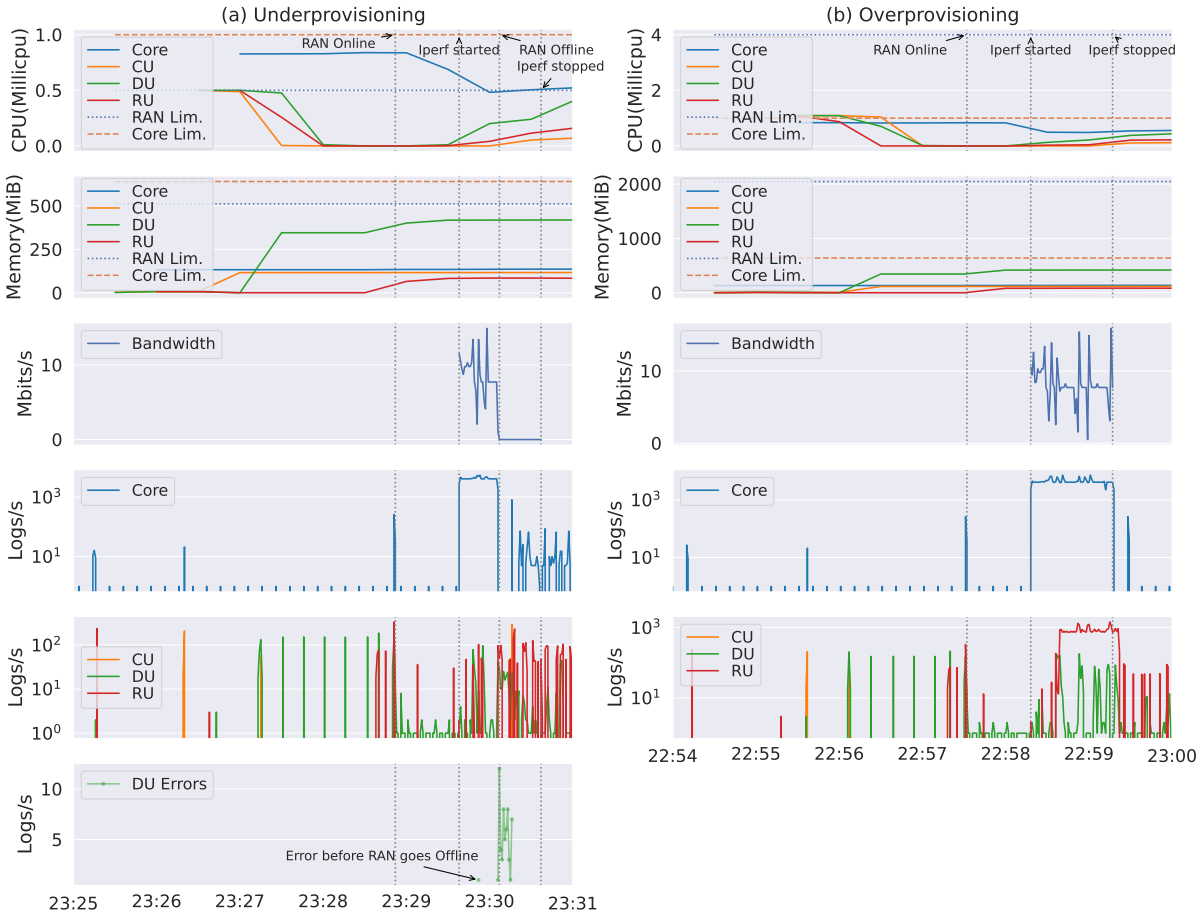


Figure 15 – Results for the performance test scenario. The first column (a) represents the test results with resource underprovisioning. The second column (b) illustrates the test results with resource overprovisioning (BRUNO et al., 2023a).

6.1.2 Anomaly Detection using ML

We used the logs extracted in the experiments described in Subsection 6.1.1 to evaluate whether the built-in ElasticSearch ML can detect the injected fault. To this end, we created an anomaly detection job in ElasticSearch to analyze the collected logs separately, i.e., creating one log for each pod service (CU, DU, RU, and CN). This analysis was achieved by partitioning the centralized dataset per pod name. In ElasticSearch, an anomaly detection job can be different. We used the Categorisation anomaly detection to look for categories that rarely occur in time. ML feed and processing were performed every second for each log. Figure 16 shows the screenshot of a Kibana dashboard where detected anomalies for each log are displayed for visualization. We can see that the abnormal messages emitted by the DU log were also captured by the ElasticSearch ML built-in, as illustrated by orange arrows 1, 2, 3, and 4 in Figure 16. Therefore, we argue that the deployed B5G system's logs are suitable for processing by a COTS ML solution.

This work is crucial in addressing the research question,"How can components of an open

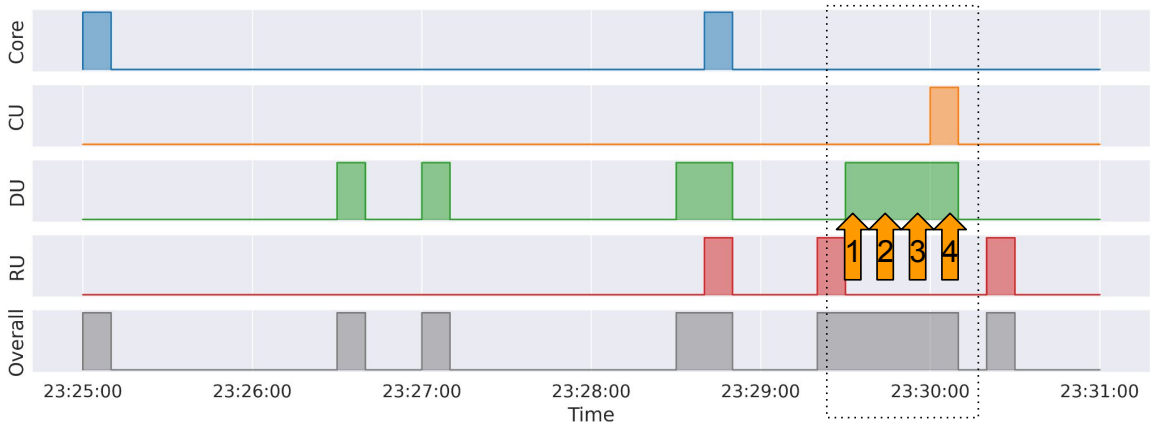


Figure 16 – Anomaly detection results from ElasticSearch (BRUNO et al., 2023a).

radio access network be integrated to address dynamic user demands efficiently?" The findings from our anomaly detection experiments not only validate the effectiveness of the proposed architecture but also provide a foundational basis for the monitoring aspects of the chosen use case. The insights gained from this study emphasize the role of observability tools and machine learning in enhancing the adaptive capabilities of modern telecommunication systems.

In summary, we conclude that the deployed B5G system was able to make its internal state observable by using standard cloud-native observability tools and a COTS ML solution. We highlight that message logs in free5GC and OpenAirInterface need to be better structured and intelligible, making log analysis difficult, especially when using visual inspection. Therefore, there is room for improving observability in such platforms.

6.2 Radio Access Network Intelligent Controller Orchestrator

This section presents a performance evaluation of RAN Intelligent Controller Orchestrator (RIC-O) from a practical perspective, employing real-world experiments. This study addresses the research question, "What strategies can effectively disaggregate Near-RT RIC functions to manage fluctuating user demands?" By focusing on real-world scenarios, we provide empirical evidence that supports the effectiveness of the proposed orchestration strategy. This evaluation also contributes significantly to developing and validating our architecture and the chosen use case. We present the results related to the effective operation of RIC-O in a practical environment using our placement solutions integrated with a real-world Near-RT RIC (OSC) bring a traditional orchestration system (K8s) (ALMEIDA et al., 2023; BRUNO et al., 2023b).

Next, we describe experiments run in a scenario with five CNs, which are virtual machines (VMs) with the following configuration: 4 Virtual Central Processing Units (vCPUs), 8 GB Random Access Memory (RAM), and 50 GB of the virtual disk. One CN represents the cloud node (i.e., c_0), and the others represent the edge computing nodes (i.e., $c_m \in C$). These CNs are worker nodes in a K8s cluster managed by a master node running a sixth VM with the

following configuration: eight vCPUs, 16 GB RAM, and 100 GB of the virtual disk. All VMs are hosted on a DELL PowerEdge M610 server with four Intel Xeon X5660 processors and 192 GB RAM, which runs VMware ESXi 6.7 as the hypervisor. Additional details about the software tools employed in the experiments are available in the public repository of this article. We also employed an E2 Simulator (E2Sim) to represent four E2 nodes that the Near-RT RIC must serve.

To illustrate the orchestration capabilities of RIC-O, we designed two scenarios in which the latency-sensitive control loop is disrupted and show how our proposal brings the Near-RT RIC back to normal operation. In the first scenario, RIC-O must deal with a sudden and high increase in the path latency used to serve a certain E2 node. The second scenario is more challenging because RIC-O needs to deal with a CN that becomes unavailable, i.e., any latency-sensitive control loop involving this CN disappears, since the Near-RT RIC components running in it suddenly become inaccessible.

In a real-world RAN, the latency between a pair of nodes may change, for example, due to a (re)route decision in the underlying network. Since our underlay network matches the overlay one, we emulate the sudden increase in the latency between the E2 node and its corresponding E2T by reconfiguring the latency in the virtual link connecting these nodes. Figure 17 illustrates the main events occurring along the time in this scenario. This figure shows the status of the control loop between each E2 node and its corresponding xApp. In addition, the figure presents the CPU utilization of some essential software components (i.e., `RIC-O Deployer`, `RIC-O Optimizer`, xApps, and `E2T`), which helps keep track of the actions performed by RIC-O.

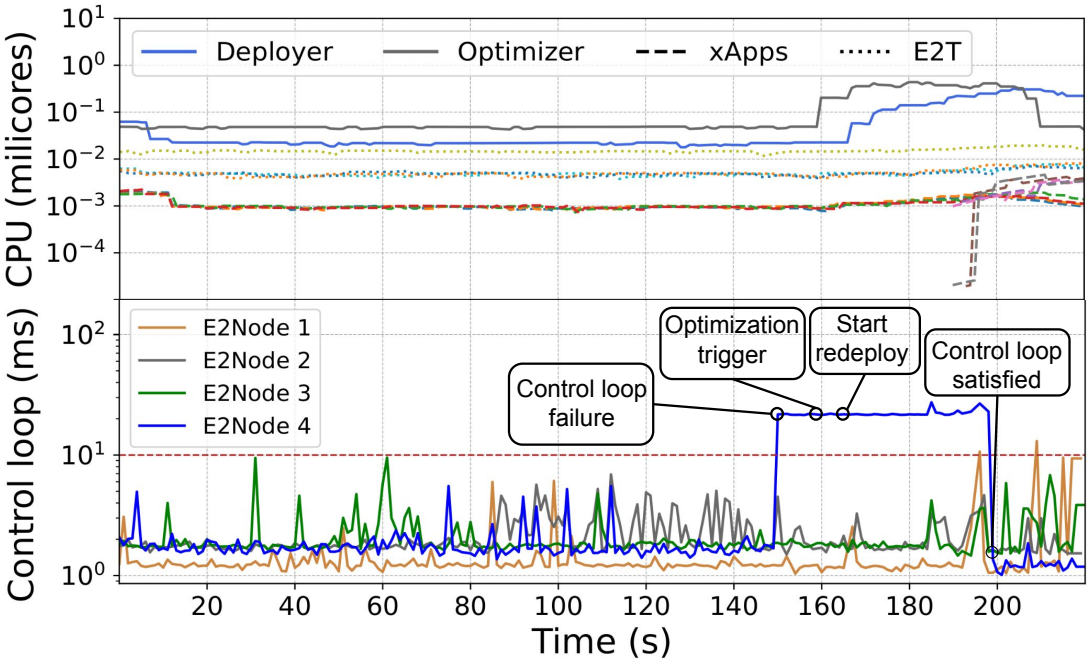


Figure 17 – Reaction to a sudden violation of the latency-sensitive control loop requirements (ALMEIDA et al., 2023).

As illustrated in Figure 17, the first scenario is initially in a fully operational state, and the

latency-sensitive control loop of each E2 node is satisfied by the Near-RT RIC thanks to the initial orchestration defined by RIC-O. Therefore, at time instant 150 s, the latency of the control loop from E2 node 4 suddenly increases and remains persistently above 10 ms, as indicated by event *Control loop failure* in the figure. After 10 seconds, the `Monitoring System` considers that the event is a consistent control loop violation and notifies the `RIC-O Optimizer` to compute a new solution, at time instant 160 s, as indicated by the event *Optimization trigger*. The heuristic strategy of the `RIC-O Optimizer` quickly finds a solution and requests `RIC-O Deployer` to apply this new placement nearly 5 seconds later, as indicated by the event *Start redeploy*. The `RIC-O Optimizer` keeps running the optimal strategy thread. Finally, the redeploy of the Near-RT RIC components and reconfiguration of E2 nodes completes at time instant 200 s, as indicated by event *Control loop satisfied*, when the latency-sensitive control loop is again limited to 10 ms.

Figure 17, and Figure 18 show a few measurements of the control loop that go above 10 ms. This behavior is related to the underlying operating system and virtualization platform (i.e., hypervisor). A tight threshold of 10 ms is on such a sensitive scale that even a traditional process scheduler may sometimes cause a small variation. Since fine-tuning those systems is out of the scope of this work, we configured the `Monitoring System` to report only persistent violations of the 10 ms threshold in latency-sensitive control loops.

The second scenario, where a CN suddenly crashes, represents a software or hardware failure or network outage. It is emulated by abruptly forcing a shutdown of the VM running the CN. Figure 18 illustrates the main events occurring along the time in this scenario. Similar to Figure 17, Figure 18 shows the status of the control loop between each E2 node and its corresponding xApp. However, we have not identified relevant information that justified presenting measurements related to Near-RT RIC nor RIC-O components.

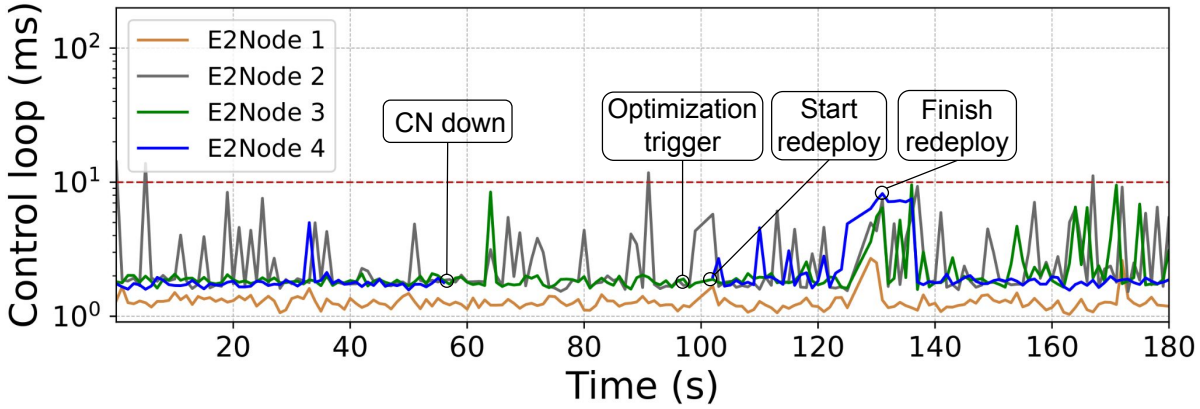


Figure 18 – Reaction to sudden unavailability of the CN under use (ALMEIDA et al., 2023).

As illustrated in Figure 18, the second scenario also starts from a fully operational state, where the latency-sensitive control loops of all E2 nodes are satisfied by the Near-RT RIC, thanks to the initial orchestration defined by RIC-O. The CN running Near-RT RIC components responsible for serving the E2 node 4 suddenly become unavailable, as indicated by the

event *CN down* in the figure. In this case, the latency-sensitive control loop of this E2 node is disrupted, i.e., no more control loop measurements exist. After 50 seconds, the `Monitoring System` detects the problem and notifies the `RIC-O Optimizer` to compute a new solution, as indicated by the event *Optimization trigger*. This time interval for reporting the problem may seem long, but it is the default K8s policy for detecting worker node unavailability. After nearly 5 seconds, the heuristic strategy of `RIC-O Optimizer` finds a solution and requests `RIC-O Deployer` to apply the new placement, as indicated by the event *Start redeploy*. Finally, the redeployment of the Near-RT RIC components finishes, and the latency-sensitive control loop of E2 node 4 is reestablished at time instant 130 s, as indicated by the event *Finish redeploy*.

In summary, the experiments and evaluations presented in this section provide comprehensive insights into the criteria and methods for evaluating the proposed orchestration strategy in response to fluctuating user demands. The successful handling of dynamic scenarios validates the robustness of RIC-O and demonstrates its practical applicability. This study has been instrumental in refining our architecture and developing the monitoring aspects of the chosen use case, thereby advancing the overall research objectives.

6.3 Disaggregated RIC Cloud Benchmark

This section evaluates placement strategies for the Near-RT RIC in a cloud-native environment. This study addresses the research question, “How can components of an open radio access network be integrated to address dynamic user demands efficiently?” By systematically examining various deployment strategies, we provide insights into how configurations impact key performance metrics such as setup and registration latency, deployment time, and resource consumption. The findings from this evaluation are crucial for refining our architecture and advancing the practical implementation of our chosen use case. We investigate three Near-RT RIC deployment strategies: MC, MD, and Dis.

We evaluate the deployment time of RIC Manager components (1) and analyze three distinct Near-RT RIC infrastructure deployment strategies: MC, MD, and Dis. These strategies are assessed for their impact on E2 Setup latency (2), deployment time of Near-RT RIC control loop functions (3), and control loop latency between the E2N, the xApp via the E2 interface (4) and discussion about consume resources (5). The MC strategy involves co-locating all components at “nv-central”, showcasing the advantages of resource centralization. In contrast, the MD configuration situates the RIC Manager, xApp, and E2T at “nv-central” while distributing E2Ns across metropolitan and internal sites. This setup allows for an exploration of how a centralized Near-RT RIC deployment affects control loop latency when E2Ns are geographically dispersed. Lastly, the Dis strategy meticulously distributes the components: positioning the RIC Manager at “nv-central” and deploying xApp instances, E2Ts, and E2Ns across metropolitan and internal sites, enabling a comprehensive examination of distributed deployment impacts on network performance.

We utilized the Near-RT RIC from OSC release G to implement the experiments as well as improved versions of the bouncer xApp[1] and the E2Sim[2] provided originally by OSC. The E2Sim is the application that simulates E2Ns.

6.3.1 Deployment Time of Near-RT RIC Manager

The deployment time of various components in a Near-RT RIC Manager plays a critical role in determining the system's responsiveness and readiness. We conducted a series of experiments to measure the initialization times for each RIC component that composes the RIC Manager as a whole. The results, illustrated in Figure 19, show significant disparities in the initialization times among different components. For instance, the Database-as-a-Service (dbaas-server) takes notably longer to initialize (∼28 s) compared to other components such as the alarmmanager, o1mediator, and vespamgr, which initialize in under 2 s. The total time required for the entire Near-RT RIC Manager to become functional is ∼51 s. This time is cumulative and includes the instantiation of individual components and the interactions between them to become ready for the Manager to operate. Understanding these initialization times is pivotal for improving the RIC Manager performance, allowing researchers to identify bottlenecks and focus on specific components that require optimization, thereby improving the overall deployment process.

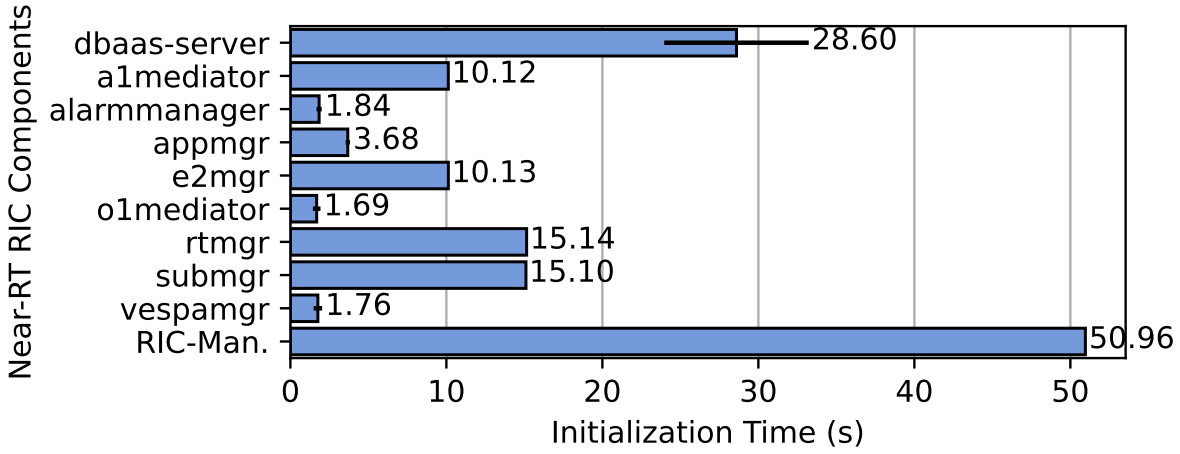


Figure 19 – Deployment times of Near-RT RIC Manager components (BRUNO et al., 2024a).

The observed deployment times pose significant challenges for meeting the stringent latency requirements of upcoming B5G networks. Such latencies can compromise results, particularly for real-time applications, impacting the network scalability and adaptability. Several optimization strategies could be explored to improve these component deployments, including ML algorithms for smarter deployments, task parallelization, and resource pre-allocation strategies. Addressing these challenges is paramount for leveraging the full potential of B5G networks, especially in use cases demanding ultra-reliable and low-latency communications.

[1]Source for the improved bouncer code can be found at: https://github.com/LABORA-INF-UFG/bouncer-rc
[2]Source code for the improved E2Sim can be accessed at: https://github.com/LABORA-INF-UFG/e2sim

The Near-RT RIC Manager test demonstrates unique deployment times attributed to the co-location of all components. Distinct from typical setups where components are centrally located for lower latency sensitivity, this configuration does not represent a full Near-RT RIC instance but focuses on control plane components. Consequently, standard deployment comparisons, MC, MD, and Dis, are not directly applicable due to these specific architectural differences.

6.3.2 E2 Node Setup Time

This subsection provides a comprehensive evaluation of the time required to set up an E2N in the Near-RT RIC across different data center infrastructures (i.e., MC, MD, and Dis). Figure 20 shows the sequence of steps required to perform the E2 Setup procedure that establishes the E2N capabilities and services with the E2 Manager in the Near-RT RIC (O-RAN Alliance, 2023b).

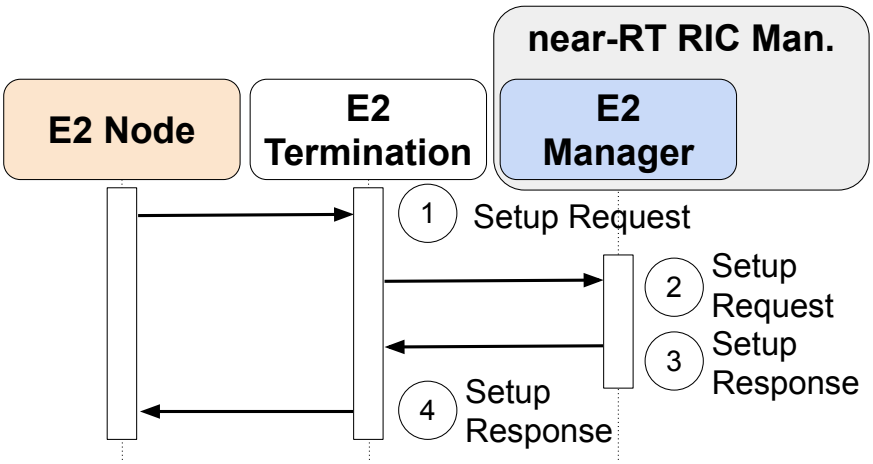


Figure 20 – Sequence of E2N setup procedure (BRUNO et al., 2024a).

The E2 setup between the E2N and the Near-RT RIC can be summarized in four key steps. First, the E2N initiates a connection by sending an E2 Setup Request message to the E2T component over the E2 interface, typically using the Stream Control Transmission Protocol (SCTP) protocol (1). Next, the E2T verifies the connection eligibility and forwards this request to the E2 Manager through RAN Intelligent Controller Message Router (RMR) for further processing (2). Upon successful validation, the E2 Manager then sends an E2 Setup Response message back to the E2T using RMR (3). Finally, the E2T forwards this E2 Setup Response message to the E2N (4). Such a response message confirms the successful establishment of the E2 connection.

We evaluated each deployment strategy of the data center infrastructures regarding the time required to set up an E2 connection. Table 8 shows the measurements of this evaluation. Despite the streamlined and centralized architecture of the MC deployment, it remarkably outperforms other configurations regarding E2N setup time. The MC deployment required only 27 ms to complete the E2 Setup procedure, highlighting the benefits of centralizing resources to achieve lower setup times.

Table 8 – E2 Node setup latency (BRUNO et al., 2024a).

Infra.	RIC Man.	E2T	E2Sim	Latency(ms)
MC	nv-central	nv-central	nv-central	**27**
MD	nv-central	nv-central	mia-metropolitan	**112**
Dis[1]	nv-central	mia-metropolitan	mia-metropolitan	**42**
MD	nv-central	nv-central	mia-internal	**169**
Dis[2]	nv-central	mia-internal	mia-internal	**73**
MD	nv-central	nv-central	nyc-metropolitan	**47**
Dis[1]	nv-central	nyc-metropolitan	nyc-metropolitan	**30**
MD	nv-central	nv-central	nyc-internal	**73**
Dis[2]	nv-central	nyc-internal	nyc-internal	**38**

The MD strategy shows variable latency levels depending on the E2Sim deployment, ranging from 47 ms in the "nyc-metropolitan" to 169 ms in the "mia-internal" configurations. This strategy indicates that while MD allows for some resource centralization, it can result in higher latencies when E2Ns are geographically dispersed or deployed at internal data centers.

The Dis deployment distributes RIC Manager, E2T, and E2Sim across various locations. We identified two variations in our dataset: Dis[1], where E2T and E2Sim share the same metropolitan data center, while RIC Manager is placed in the central data center; and Dis[2], which places E2T and E2Sim in the same internal data center, while the central data center hosts the RIC Manager. The Dis strategy results in E2 Setup procedures ranging from 30 ms in a "nyc-metropolitan" environment to 73 ms in a "mia-internal" environment. These results indicate that the Dis strategy generally offers better latency when compared to the MD configuration in similar environments.

The evaluation of E2 Setup latency across MC, MD, and Dis configurations shows key trade-offs between centralization and latency. The MC strategy outperforms in terms of latency, making it suitable for near-real-time RANs tasks with strict low latency requirements to set up E2 connections. Conversely, the MD deployment can be problematic due to its variable latency levels in scenarios that require consistently low-latency for E2 Setup procedures. The Dis setup offers a balanced latency profile, often outperforming MD. Therefore, for latency-critical E2 Setup scenarios, the choice of infrastructure should align with application requirements, node distribution, and network constraints.

6.3.3 Deployment Time of Near-RT RIC Control Loop Functions

In the Near-RT RIC, xApp deployment is a multifaceted process demanding several complex steps for successful registration and event subscription from the RAN. Figure 21 shows the sequence of steps to register an xApp in the Near-RT RIC and subscribe for events from the RAN. Initially, the xApp triggers the registration protocol by dispatching a request to the xApp

Manager (1), which replies with a registration status response (2). After successful registration, the xApp initiates a subscription to specific RAN events by sending a REST subscription request to the Subscription Manager (3). The Subscription Manager processes this request and returns a REST response (4) in case of merging subscriptions from the RAN. Next, the Subscription Manager sends a Route Create request to the Routing Manager to establish suitable routes for RAN event notifications (5). Upon processing the Route Create request, the Routing Manager refreshes the routes for the xApp and E2T (6, 7). The Subscription Manager then drives a RIC Subscription Request to the E2T (8), which encapsulates the payload in an E2 Subscription Request and forwards it to the E2 Node (E2N) (9). The E2N, in turn, processes the message and sends the corresponding response to the E2T with the outcome of the subscription status (10). The E2T then sends a RIC Subscription Response to the Subscription Manager to notify the E2N subscription status (11). Finally, the Subscription Manager dispatches a REST Subscription Notification to the xApp containing the overall outcome of the subscription process (12). All these steps must be completed to enable the xApp to receive notifications from and control the RAN and exchange messages with other Near-RT RIC components.

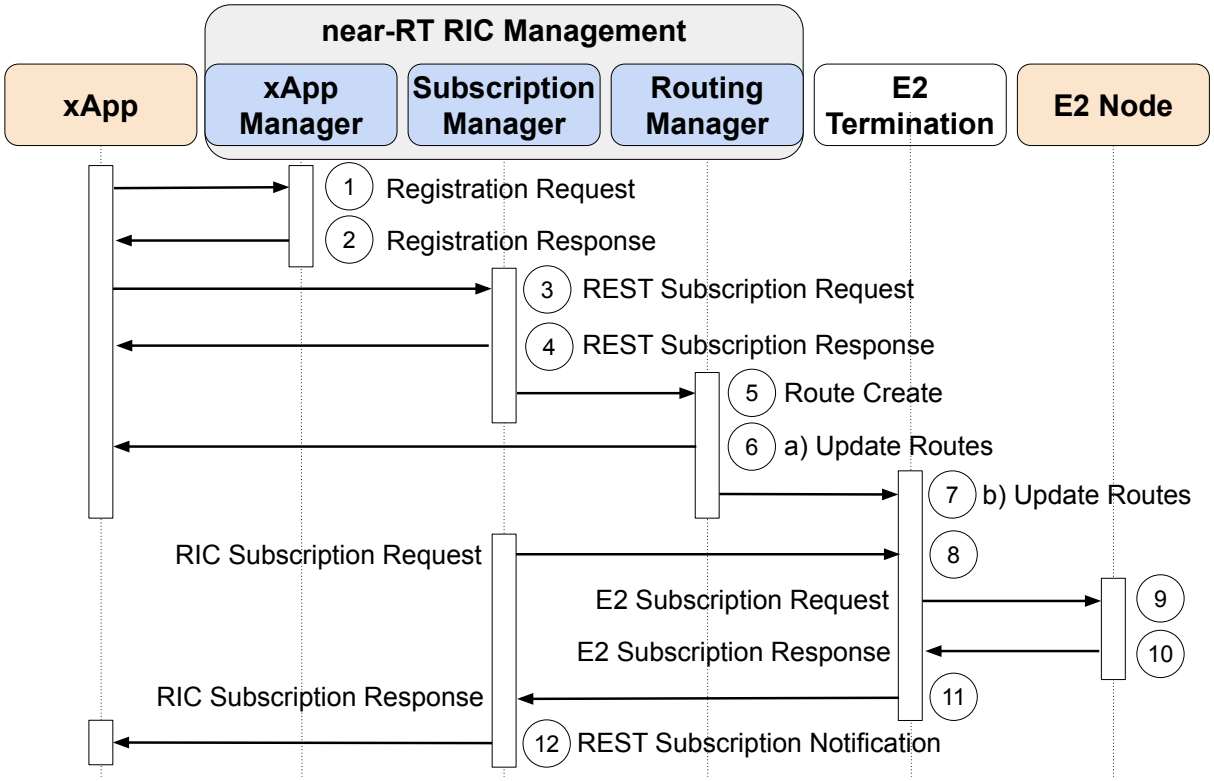


Figure 21 – Sequence of steps for xApp registration and E2 subscription (BRUNO et al., 2024a).

The optimization is triggered when the latency of a control loop from an E2N increases and remains persistently above the 10 ms threshold, indicating a consistent control loop violation. This scenario prompts the orchestrator to compute a new solution and apply new placements. In this orchestration, new xApps are deployed to meet queue latency requirements, and those not meeting these requirements are removed after all new xApps are operational, ensuring minimal

impact on the services provided by existing xApps.

The assessment of deployment times for Near-RT RIC Control Loop Functions is based on the cloud latency impact at each Node Site. For example, a RIC Manager located at Node 1 represents a MC. However, when this RIC Manager is used at Node 1 and xApp and E2T are co-located in another Node, it is characterized as Dis. The concept of a MD deployment was not deemed viable in our analysis due to the challenges in maintaining a communication link between all sites and Node 1. This behavior is particularly noticeable in the best scenario, such as the RIC Manager located in the New York City Metropolitan area, where the latency can reach almost 10 ms during certain times of the day, making deployment nearly impractical. Similarly, for other sites, the communication latency between the E2N and xApp often exceeds the maximum 10 ms threshold allowed by the Near-RT RIC, rendering this deployment approach unsuitable.

Our study evaluated the startup times of E2T and xApp components within the Near-RT RIC framework of B5G networks. We established their readiness times by deploying these components on a distributed K8s cluster and performing a series of trials. Emphasis was placed on the initialization phase of the OSC E2T component, specifically from pod instantiation to achieving full operational capability, including the K8s pod's application startup. Notably, the initial setup period of the E2T was omitted from our analysis due to its early development stage and pending code refinement. The extended readiness duration observed is pivotal for real-time applications, such as autonomous vehicles and telemedicine, which rely on millisecond-level response times. The latency introduced by this component can cause the control loop to surpass the acceptable threshold time by over three minutes, a substantial deviation from the norms suitable for these applications. These findings indicate that the current deployment time of the OSC E2T is not yet aligned with the stringent requirements of B5G networks, highlighting a critical need for specific performance improvements and optimizations.

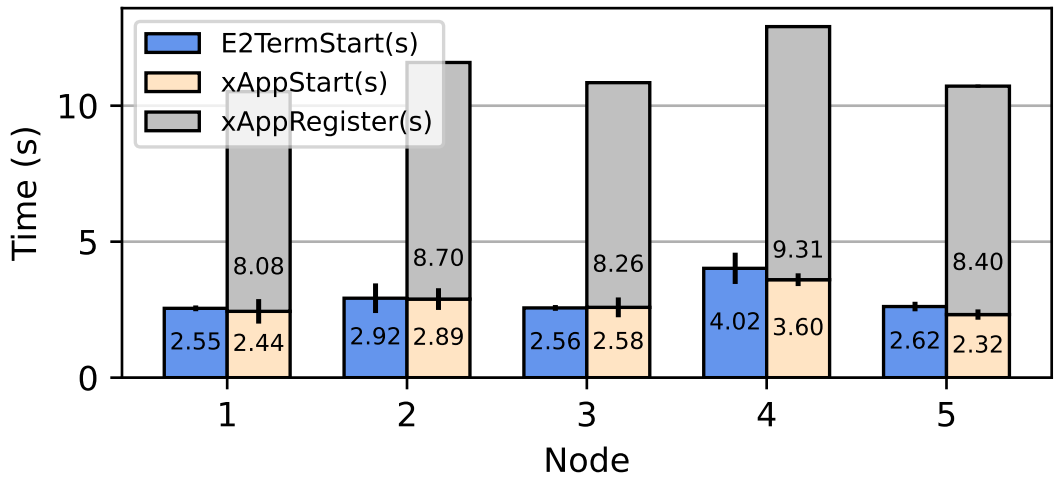


Figure 22 – Initialization time of the Near-RT RIC control looping (BRUNO et al., 2024a).

Figure 22 shows the time to initialize the components required to enable the Near-RT RIC control loop for RAN events. The experiment has been conducted across five distinct nodes distributed in different sites: Node 1 (nv-central) is located in the Central site, Node 2 (mia-metropolitan) and Node 3 (nyc-metropolitan) are located in Metropolitan sites, while Node 4 (mia-internal) and Node 5 (nyc-internal) reside in Internal sites. Results show that Node 4 takes all measured parameters' most extended initialization times. This node has the highest E2T startup time of $\sim$4 s, the highest xApp registration procedure with $\sim$9 s, and the topmost xApp total times of $\sim$12 s, mainly because of the link latency between "nv-central" and "mia-internal" sites.

Moreover, Node 1, situated in the Central site, lasts the lowest time to start the xApp, requiring $\sim$10 s, which indicates the highest efficiency in terms of initialization. Analyzing these results, we identified actionable areas for deployment optimization, such as the E2T initialization and xApp registration times. For example, optimizing the E2T initialization sequence could mitigate the bottleneck caused by its startup time for control loop deployments. Similarly, high xApp registration times suggest a need for protocol optimization or management adjustments. Such targeted refinements are crucial for adhering to the stringent latency standards of B5G networks, especially for latency-sensitive applications.

6.3.4 Control Loop Latency

This experiment evaluated the influence of the MC, MD, and Dis deployment strategies on the latency of the E2N-xApp-E2N control loop. The main objective of this experiment is to analyze the effect of the placement of the Near-RT RIC components across different locations, particularly the RIC Manager, xApp, and E2T. We also evaluated the placement of the E2N. Figure 23 shows the control loop latency for the MC, MD, and Dis deployment strategies. We can observe that the MC deployment indicates optimal control loop latency. It can introduce budget constraints due to the requirement for a dedicated Near-RT RIC instance per E2N. The MD strategy presented sub-optimal results, largely due to the link quality susceptible to latency fluctuations between Metropolitan, Internal, and Cloud sites. The Dis infrastructure has shown consistent results due to its decentralized nature and independence on the cloud site, resulting in improved performance stability.

We highlight that MC and Dis deployments demonstrated satisfactory performance, each with its trade-offs. The MC offers low-latency performance but incurs economic challenges due to the need for a dedicated Near-RT RIC instance at each node. The more cost-effective MD setup experiences an average control looping latency of 52 ms for E2N 4, as illustrated in Figure 23. These technical considerations must be weighed against the specific requirements and constraints of real-world applications. For example, telemedicine and autonomous vehicles would benefit from the low-latency capabilities of the MC strategy, albeit at a potentially higher cost. At the same time, smart city applications might find the Dis setup more appropriate due

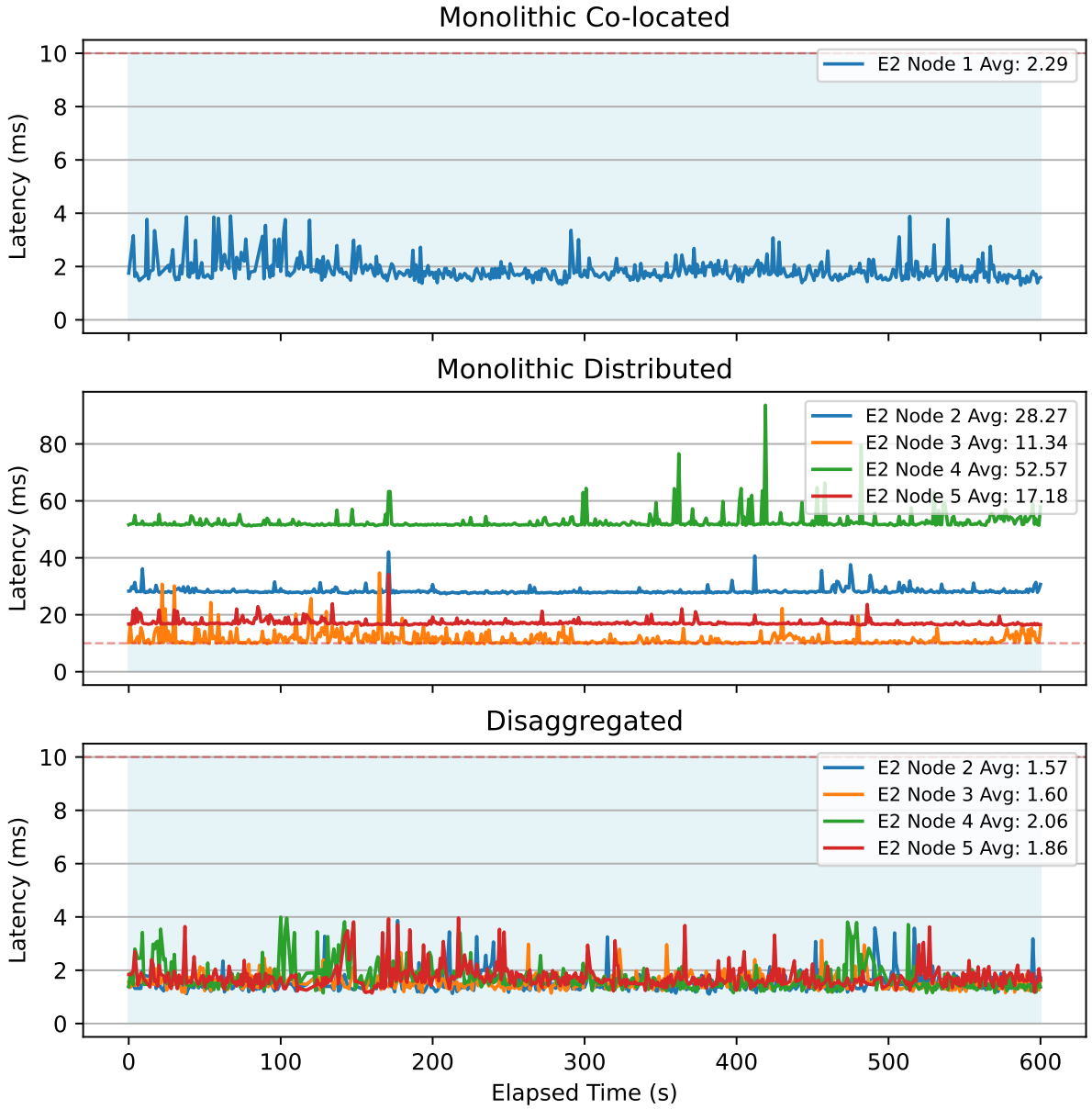


Figure 23 – Control loop latency for MC, MD, and Dis deployments (BRUNO et al., 2024a).

to its latency, cost, and reliability balance.

6.3.5 Nodes Resources

Optimal resource utilization within K8s environments is a critical factor in successfully deploying cloud-native networks, particularly as we transition towards B5G. This evaluation focuses on CPU and memory usage in MC, MD, and Dis architectures while considering the role of latency as a secondary yet important factor. As illustrated in Figure 24, resource consumption is categorized into three distinct segments: K8s overhead, which encompasses the intrinsic resource utilization of K8s; ricplt, which accounts for the resources consumed by the Near-RT RIC components in conjunction with the E2N; and ricxapp, which details the resource

usage attributable to xApps. Node 1 represents Central Cloud and Node N represents other Nodes for each deployment.

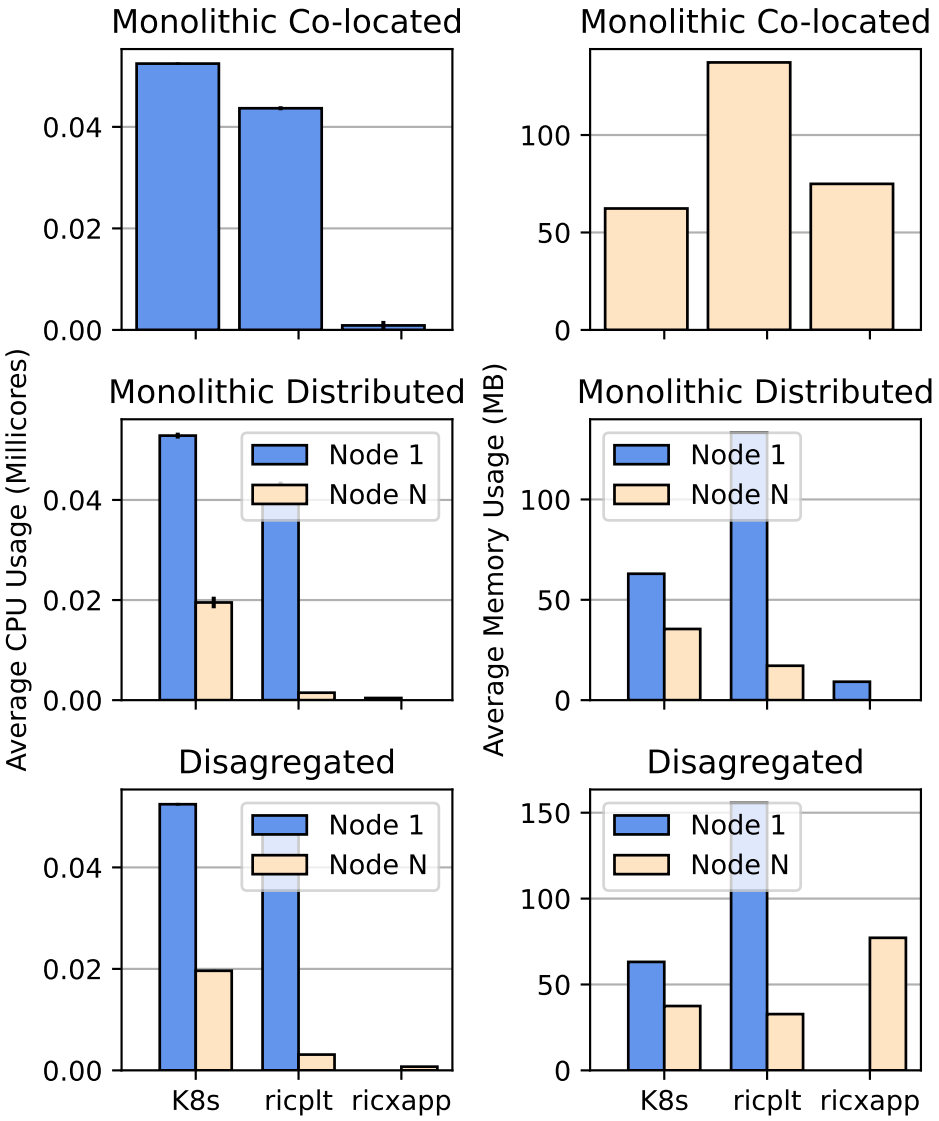


Figure 24 – Nodes Resources (BRUNO et al., 2024a).

In the MC model, all components are centralized on the cost-effective Node 1. This setup ensures a unified resource pool, simplifies management, reduces costs, and boasts the advantage of low setup latency. However, the centralized nature of this model could lead to potential challenges in resource contention, especially in scenarios demanding high resilience. The MD model differentiates itself by strategically distributing its components. Node 1 remains the primary hub for most operations, maintaining resource efficiency and cost-effectiveness. In contrast, Node N is designated to run only the E2N, which allows for specialized resource allocation but introduces higher latency, evident in the 38 ms to 169 ms range. This separation of duties between the nodes can optimize the network's overall performance by balancing the load, but it may also lead to increased complexity and cost.

The Dis configuration further refines the distribution of tasks. In this case, Node N extends its responsibilities to include the E2N and the execution of xApps. This dual role necessitates a broader allocation of resources, thus incurring higher costs and complexity. The resource utilization in this architecture is more dynamic, with setup latencies of 42 ms for Dis^1 and 73 ms for Dis^2 configurations, reflecting the intricate nature of this distributed system.

In summary, the architectural selection for an O-RAN network must be a deliberate process that weighs resource utilization and efficiency against cost and latency implications. While MC provides a singular, cost-effective platform with low latency, it may not satisfy the scalability and availability requirements of advanced B5G applications. MD and Dis offer more distributed and flexible arrangements, with Node N's role expanding from hosting E2N in MD to running both E2N and xApps in Dis, each with its associated latency and resource deployment trade-offs. Decision-makers must carefully evaluate these factors to ensure that the network architecture aligns with the performance and cost objectives of their B5G deployments.

This comprehensive evaluation of disaggregated RIC cloud benchmarks provides valuable answers to the research question on integrating open radio access network components to meet dynamic user demands efficiently. The insights gained from comparing different deployment strategies have significantly contributed to the development and optimization of our architecture and the practical implementation of the chosen use case. By understanding the trade-offs between centralization, latency, and resource allocation, we can better design and deploy O-RAN systems that are both efficient and scalable, ensuring they meet the evolving demands of modern telecommunication networks.

6.4 Use Case Energy Saving Results

This section analyzes the efficiency and performance of adaptive network management in a 6G O-RAN environment, focusing on energy savings. These results evaluate the use case chosen to validate the proposed architecture. Subsection 6.4.1 examines energy consumption, detailing how the system dynamically manages network resources by activating E2Ns as needed based on the number of UEs. Figure 25 shows the relationship between UEs and active E2Ns, highlighting the system's ability to minimize energy wastage by deactivating unnecessary nodes under lower loads.

Subsection 6.4.2 discusses the rApp Energy Savings results, emphasizing the impact of increasing UEs on CPU and memory usage and the time required to solve network demands as illustrated in Figure 26. The analysis shows that the system efficiently allocates CPU and memory resources to handle higher user loads, ensuring optimized performance. Similarly, Subsection 6.4.4 and Subsection 6.4.3 evaluate the xApp handover and monitoring performance, respectively, focusing on CPU and memory usage and the time required for handover and monitoring activities as depicted in Figures 28 and 27.

Subsection 6.4.5 presents an end-to-end energy optimization looping analysis. Figure 29 de-

tails various components' start and end times within the adaptive network management framework. This analysis offers insights into the efficiency and performance of each component under varying loads, demonstrating the system's scalability and ability to maintain low energy consumption while ensuring high network performance. Lastly, Subsection 6.4.6 provides an overall analysis of resource utilization and energy efficiency. Figure 30 presents a radar plot of normalized resource utilization and EE metrics across different UEs densities. This subsection discusses the implications of these results for adaptive network management in 6G O-RAN, focusing on the trade-offs between computational resources and EE.

6.4.1 Analysis of Energy Consumption

The results illustrated in Figure 25 show the number of E2Ns that are either ON or OFF as a function of the number of UEs. The corresponding energy consumption and consumed per UE are also depicted. The bottom graph in Figure 25 presents the relationship between the number of E2Ns active or inactive across varying UE counts. Initially, with 16 UEs, only one E2N is active while 16 E2Ns remain inactive. This trend continues up to 64 UEs, with a single active E2N. As the number of UEs increases, the number of active E2Ns rises to accommodate the increased load, with two E2Ns active at 128 UEs, five at 256 UEs, nine at 512 UEs, and all 17 E2Ns active at 1024 UEs.

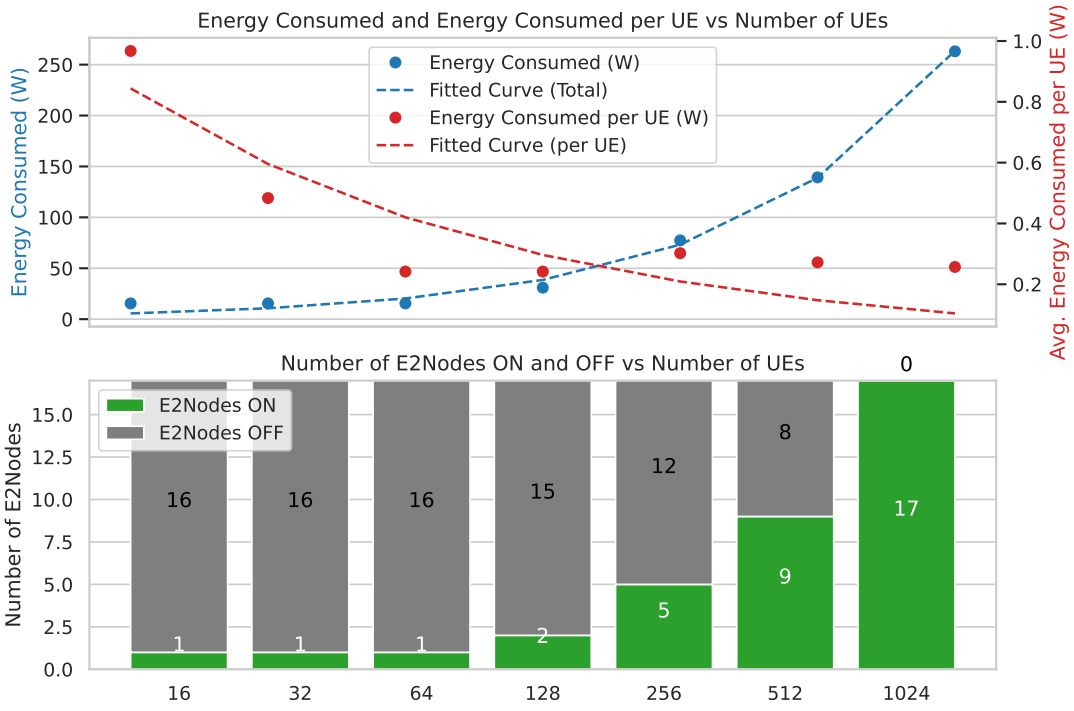


Figure 25 – Energy Consumption Results.

This adaptive activation of E2Ns demonstrates the efficiency of the rApp Energy Savings in dynamically managing network resources. The system activates additional E2Ns only when

necessary, thus minimizing energy wastage by deactivating nodes that are not needed under lower loads. The top graph in Figure 25 illustrates the total energy consumption and the energy consumed per UE as the number of UEs increases. The total energy consumed rises significantly with the number of UEs, from 15.4757 W at 16 UEs to 263.0869 W at 1024 UEs. This increase reflects the higher computational and processing demands as more UEs connect to the network.

Conversely, the energy consumed per UE decreases as the number of UEs increases, starting from 0.9672 W per UE at 16 UEs and dropping to 0.2569 W per UE at 1024 UEs. This decline indicates that the system's energy efficiency improves with a higher number of connected UEs, likely due to the more efficient utilization of network resources and the spreading of fixed energy costs over a larger number of UEs. The results underscore the effectiveness of the rApp Energy Saver in optimizing energy usage within the network. The system can maintain low energy consumption levels while ensuring network performance by dynamically adjusting the number of active E2Ns based on real-time demand. Reducing energy consumed per UE as the number of UEs increases highlights the system's scalability and efficiency in handling high-density scenarios.

6.4.2 rApp Energy Savings

Resource utilization and performance analysis reveal critical network behavior insights under varying user loads. The results are presented in Figure 26, which illustrates the impact of increasing UEs on CPU and memory usage and the time required to solve network demands.

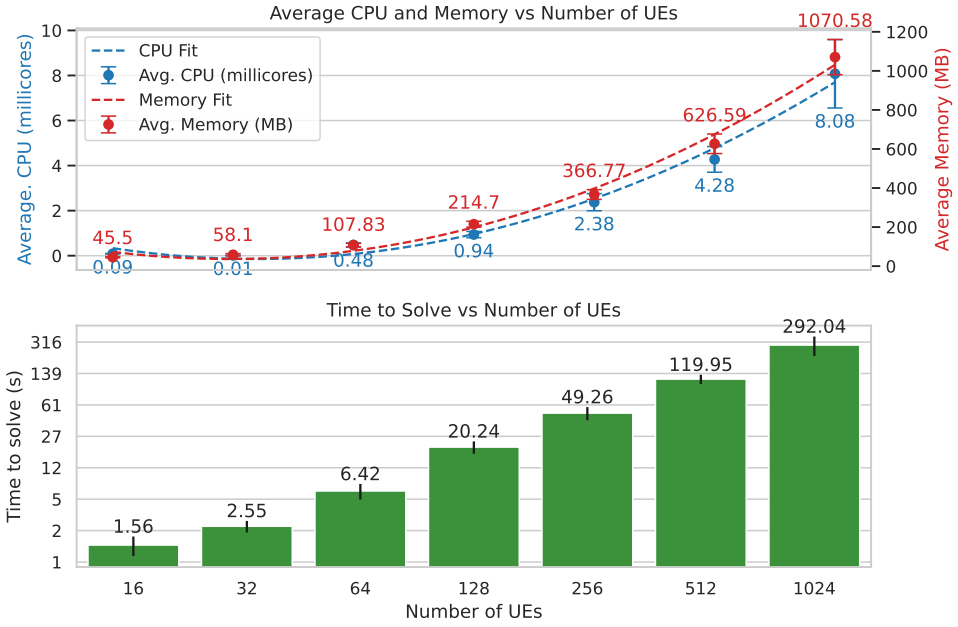


Figure 26 – rApp Energy Savings Results.

The CPU usage metrics indicate a significant increase in the number of UEs. The average CPU usage rises from 0.09 millicores for 16 UEs to 8.08 millicores for 1024 UEs, reflecting the

escalating computational demands. This trend highlights the need for efficient CPU allocation mechanisms to handle higher user loads.

Memory usage also escalates with the increase in UEs. The average memory usage grows from 45.5 MB with 16 UEs to 1070.58 MB with 1024 UEs, indicating the system's capability to manage substantial data storage and processing requirements. This significant rise underscores the necessity for robust memory management strategies in high-demand scenarios. The time required to solve network demands is another critical performance metric. The time increases from 1.56 seconds for 16 UEs to 292.04 seconds for 1024 UEs, illustrating that higher user loads impose greater computational burdens, leading to longer processing times. The standard deviation of the time to solve also increases, from 0.39 seconds for 16 UEs to 73.01 seconds for 1024 UEs, reflecting greater inconsistency in solving times as the number of UEs increases.

The resource utilization and performance analysis indicate that rApp Energy Savings is proficient at scaling to meet higher demands by efficiently allocating resources. However, the increased number of UEs significantly impacts CPU and memory usage and the time required to address network demands. These findings underscore the importance of optimizing resource allocation and processing efficiency in adaptive network management systems, especially in dynamic user demands in 6G O-RAN environments. Understanding these trends enables network operators to better plan and implement strategies to enhance performance and ensure sustainable operations as user demands grow.

6.4.3 xApp Monitoring

This subsection examines the performance metrics related to xApp monitoring, focusing on CPU and memory usage and the time required for monitoring activities. The data provides insights into how the system scales with increasing UEs. The average CPU usage during xApp monitoring gradually increases as the number of UEs grows, as shown in Figure 27. Specifically, the average CPU usage starts at 0.00095 millicores for 4 UEs and increases to 0.013 millicores for 256 UEs. This trend indicates that as more UEs are monitored, the computational demand on the CPU rises, though the increase remains relatively modest. The standard deviation of CPU usage fluctuates, with values ranging from 0.00034 to 0.0041 millicores. These variations suggest that while the average CPU usage grows steadily, there are fluctuations in demand likely due to the dynamic nature of monitoring activities.

Memory usage metrics provide a further understanding of the system's resource requirements. The average memory usage increases from 13.18 MB for 4 UEs to 13.85 MB for 256 UEs. This upward trend indicates a proportional increase in memory demands with more UEs. The standard deviation of memory usage ranges from 0.423 MB to 0.78 MB, reflecting variability in memory consumption. This variability can be attributed to differing monitoring scenarios and the associated data-handling requirements.

The time required to complete xApp monitoring activities, precisely the time to scrape data,

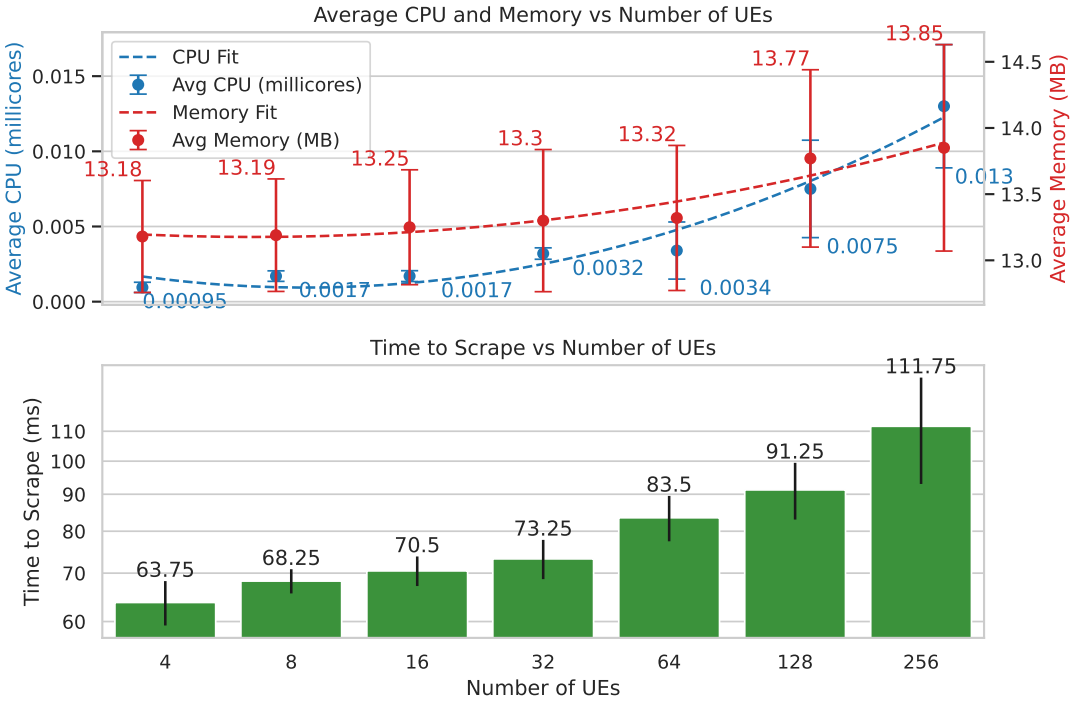


Figure 27 – xApp Monitoring.

is a critical performance metric. The average time increases from 0.064 seconds for four UEs to 0.112 seconds for 256 UEs. This significant rise highlights the growing computational and data-handling complexity as more UEs are managed. The standard deviation of the time to complete monitoring also increases, from 0.0045 seconds for 4 UEs to 0.0187 seconds for 256 UEs. This increase in variability suggests that the monitoring process becomes less predictable as the number of UEs increases, potentially due to more complex network conditions and higher data traffic.

It is essential to highlight that for this test, only one xApp was used for monitoring. To provide scalability, subsequent tests employed one xApp monitoring instance for each E2N. This approach ensures that as the number of E2Ns increases, the system can effectively scale its monitoring capabilities without significantly impacting performance metrics.

In summary, the xApp monitoring performance analysis indicates that CPU and memory usage rise proportionally as the number of UEs increases. The time required to scrape data during monitoring activities also increases significantly, highlighting the need for efficient resource management and optimization strategies to effectively handle larger UE volumes. Network operators must understand these performance trends to optimize monitoring processes and ensure seamless service delivery in 6G O-RAN environments.

6.4.4 xApp Handover

This subsection examines the performance metrics related to the xApp handover process, focusing on CPU and memory usage and the time required to execute handovers. The data provides insights into how the system scales with increasing UEs. The average CPU usage during xApp handovers gradually increases as the number of UEs grows, as shown in Figure 28. Specifically, the average CPU usage starts at 0.00073 millicores for 16 UEs and increases to 0.00109 millicores for 1024 UEs. This trend indicates that as more UEs are handled, the computational demand on the CPU rises, though the increase remains relatively modest. The standard deviation of CPU usage fluctuates, with values ranging from 0.00019 to 0.00041 millicores. These variations suggest that while the average CPU usage grows steadily, there are fluctuations in demand likely due to the dynamic nature of handover events.

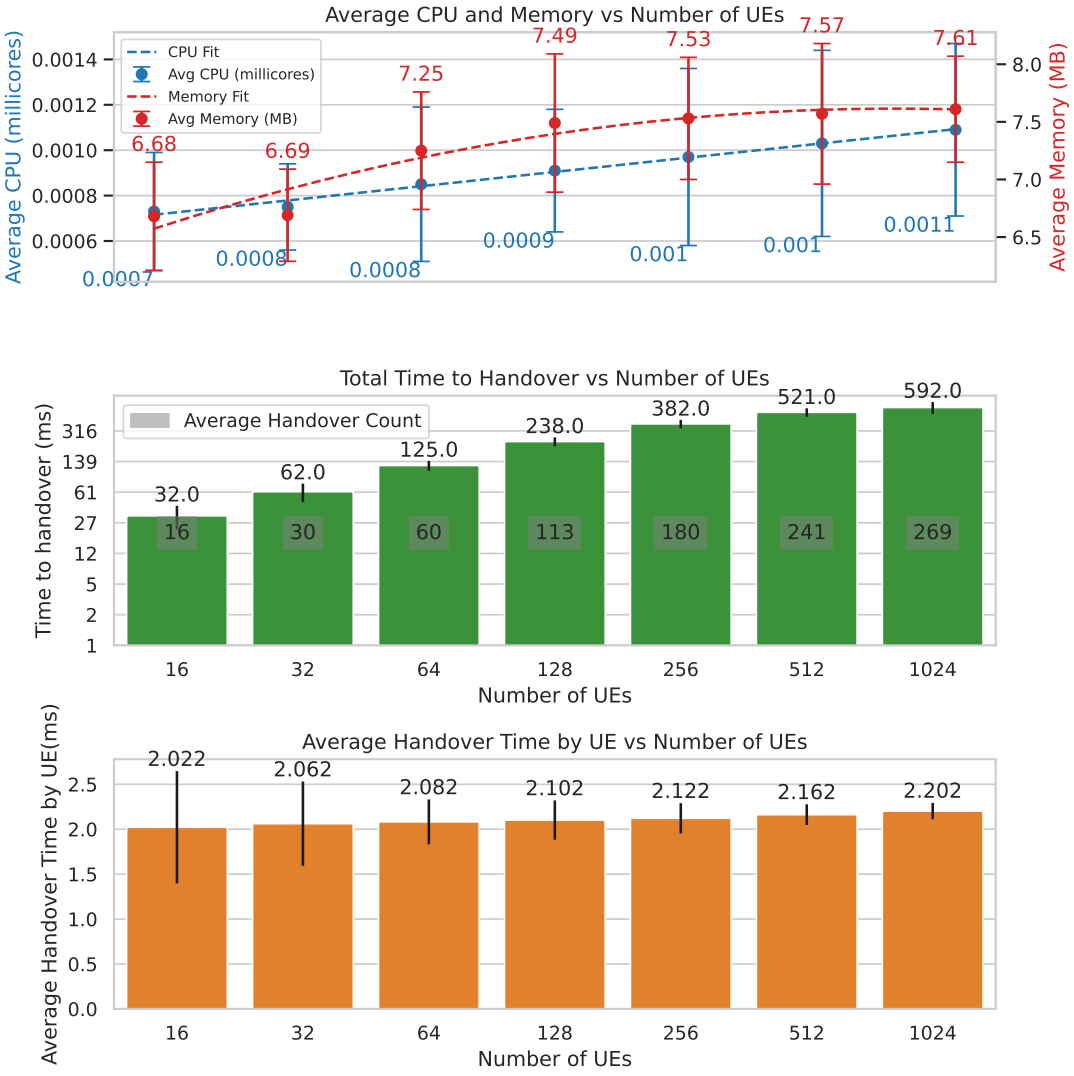


Figure 28 – xApp Handover Results.

Memory usage metrics provide a further understanding of the system's resource requirements. The average memory usage increases from 6.68 MB for 16 UEs to 7.61 MB for 1024

UEs. This upward trend indicates a proportional increase in memory demands with more UEs. The standard deviation of memory usage ranges from 0.4 MB to 0.61 MB, reflecting variability in memory consumption. This variability can be attributed to differing handover scenarios and the associated data-handling requirements.

A critical performance metric is the time required to complete xApp handovers and perform the UEs handovers. The average time increases from 32 milliseconds for 16 UEs to 592 milliseconds for 1024 UEs, highlighting the growing computational and data-handling complexity as more UEs are managed. The standard deviation of the time to complete handovers also rises, from 10 milliseconds for 16 UEs to 93 milliseconds for 1024 UEs, suggesting that the handover process becomes less predictable with an increasing number of UEs, potentially due to more complex network conditions and higher data traffic. However, the handover time per UE remains relatively stable as the number of UEs increases, starting at 2.022 milliseconds for 16 UEs and slightly increasing to 2.202 milliseconds for 1024 UEs. This trend indicates that while the total handover time increases significantly, the efficiency per UE remains consistent, highlighting the robustness of the handover process.

The requirement for handovers decreases with an increase in the number of UEs and the corresponding connection of more E2Ns. Notably, the number of handovers does not increase linearly with the number of UEs. As the number of UEs increases, the number of E2Ns also rises, which reduces the possibility of handovers. This reduction is because, with more E2Ns, the network can distribute the load more evenly, minimizing the need for handovers. Therefore, while the total handover count grows with the number of UEs, the rate of increase diminishes due to the additional E2Ns mitigating congestion and redistributing traffic efficiently.

In summary, the xApp handover performance analysis indicates that CPU and memory usage rise proportionally as the number of UEs increases. The time required to execute handovers increases significantly, highlighting the need for efficient resource management and optimization strategies to effectively handle larger UE volumes. Despite the rise in total handover time, the handover time per UE remains relatively stable, indicating consistent efficiency. Network operators must understand these performance trends to optimize handover processes and ensure seamless service delivery in 6G O-RAN environments.

6.4.5 End-to-end Energy Optimization Looping

This section presents the analysis of the start and end times for various components within the 6G O-RAN adaptive network management framework, as depicted in Figure 29. The data evaluates the time required by different components to complete their tasks, offering insights into the efficiency and performance of each component. Updated results based on new data and attached figures are discussed.

The analysis of component times involves several key processes: RF Environment Manager Connection, E2N Connection, xApp Monitoring, Prometheus, Vespa Manager, Ves Collector,

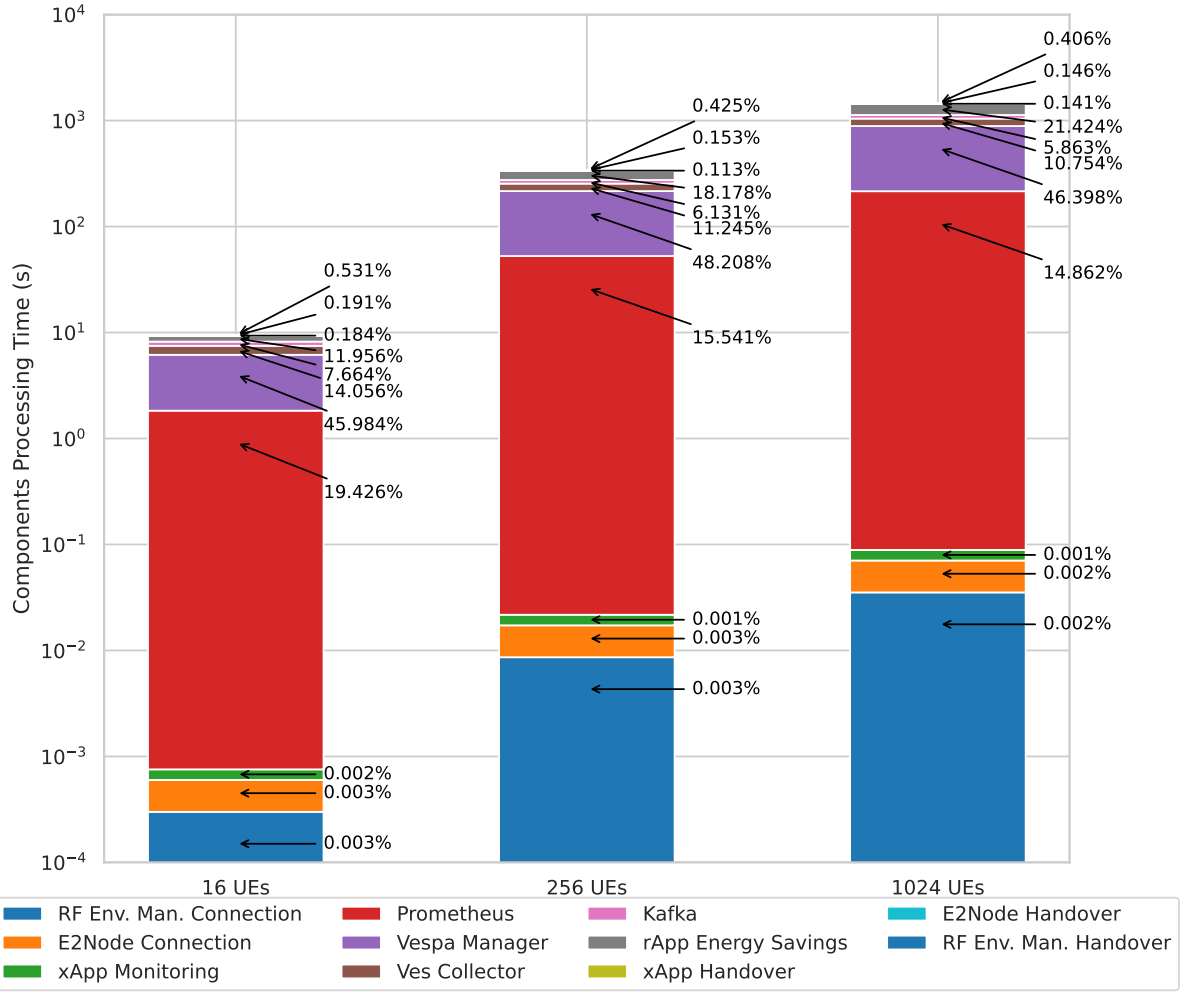


Figure 29 – Processing time end-to-end with detailed components Start and End Times for 16, 256, and 1024 UEs.

Kafka, rApp Energy Savings, xApp Handover, and E2N Handover. Each process is evaluated based on the mean time taken and the standard deviation, which provides an understanding of the variability and consistency of the process times. This experiment considers only the time to process the tasks, excluding network delays or external factors.

For 16 UEs, the RF Environment Manager and E2N components show minimal time requirements for both connection and handover tasks. The RF Environment Manager Connection process has a mean time of 0.0003 seconds and a standard deviation of 0.00001 seconds, while the E2N Connection process has identical metrics. For handover tasks, the RF Environment Manager Handover process has a mean time of 0.05 seconds and a standard deviation of 0.005 seconds, and the E2N Handover process shows a mean time of 0.018 seconds with a standard deviation of 0.002 seconds. These processes are highly efficient, with negligible variability, indicating consistent performance across different executions.

The xApp Monitoring process has a mean time of 0.000155 seconds and a standard deviation of 0.000167 seconds. Although the meantime is low, the standard deviation suggests higher variability than the mean, indicating occasional fluctuations in the time required for monitor-

ing activities. Prometheus shows a significant increase in time, with a mean of 1.83 seconds and a standard deviation of 0.2 seconds. This process is more time-consuming than the initial connections, and the moderate standard deviation indicates some variability in performance.

The Vespa Manager process has the highest mean time at 4.332 seconds and a standard deviation of 0.3 seconds. This high mean time suggests that the Vespa Manager is a critical component significantly impacting the overall processing time, with moderate variability. The Ves Collector process has a mean time of 1.324148 seconds and a standard deviation of 0.05 seconds. This process is relatively stable, with a small standard deviation indicating consistent performance. Kafka shows a mean time of 0.722 seconds and a standard deviation of 0.1 seconds. This process is moderately time-consuming with a noticeable level of variability. The rApp process has a mean time of 1.12632 seconds and a standard deviation of 0.39 seconds. The higher standard deviation indicates significant variability in the time required by this process.

For 1024 UEs, the mean and standard deviation times increase significantly across all processes. The RF Environment Manager Connection and E2N Connection processes have mean times of 0.035 seconds and a standard deviation of 0.001 seconds. The RF Environment Manager Handover and E2N Handover processes show mean times of 5.881 seconds and 2.117 seconds, respectively, with higher standard deviations, indicating increased variability and longer processing times under higher user loads.

Prometheus and Vespa Manager are the most time-consuming components, with mean times of 215.246 seconds and 672 seconds, respectively. Ves Collector and Kafka also show significant increases, with mean times of 155.747 seconds and 84.922 seconds. The rApp process has a mean time of 310.292 seconds with a high standard deviation of 73.01 seconds, indicating substantial variability.

In conclusion, the analysis reveals that the Vespa Manager is consistently the most time-consuming process, followed by Prometheus and Ves Collector. The connection and handover processes for RF Environment Manager and E2N are the least time-consuming, demonstrating high efficiency and low variability. Notably, the rApp process, despite showing moderate mean processing times, exhibits significant variability, which could impact the overall performance under fluctuating user demands. These insights are crucial for optimizing resource allocation and improving the performance of the adaptive network management framework in 6G O-RAN environments. Further attention to the variability in rApp processing times could lead to more consistent and reliable performance in energy optimization tasks.

6.4.6 Overall Analysis of Resource Utilization and Energy Efficiency

The radar plot presented in Figure 30 provides a comparative visualization of various performance metrics, including CPU usage, memory consumption, processing time, and EE, across different UE densities. The data points with 16 UEs were removed from the figure because they exhibited irrelevant values due to normalization. This section discusses the implications

of these results for adaptive network management in 6G O-RAN, with a focus on the trade-offs between computational resources and EE.

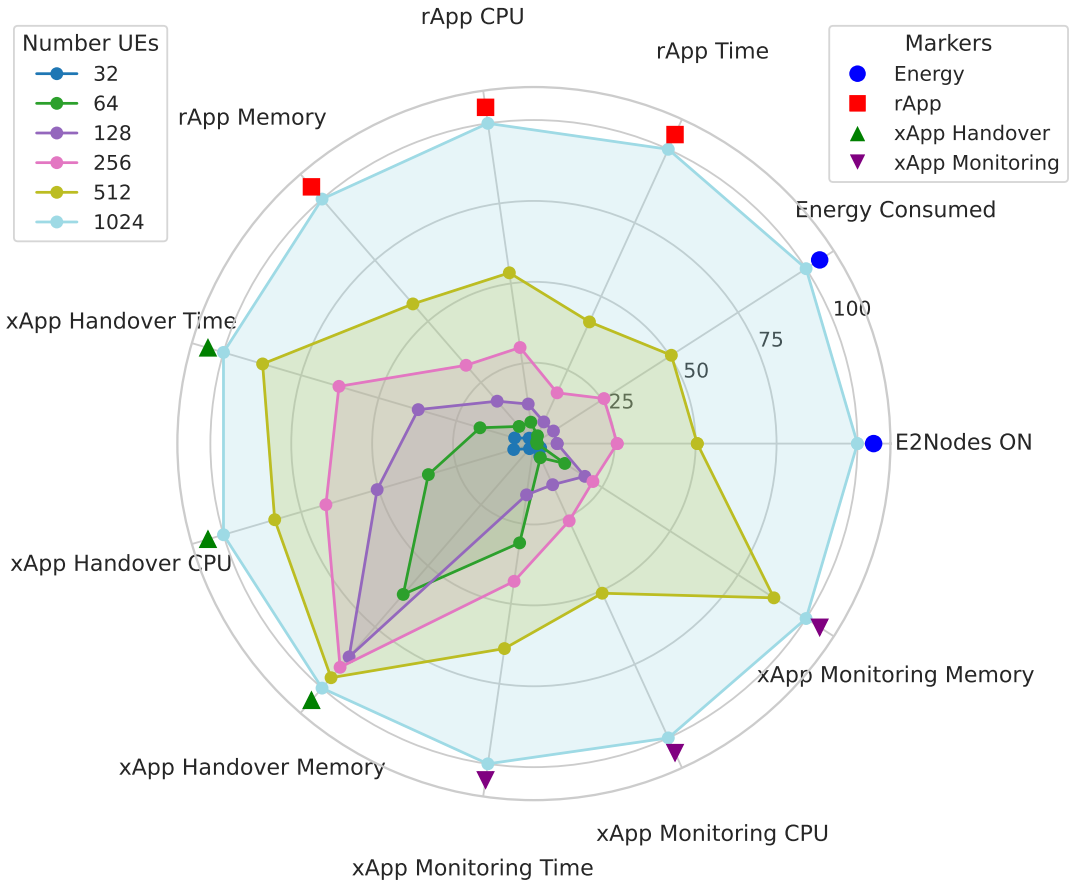


Figure 30 – Radar plot of normalized resource utilization and EE metrics across different UEs densities.

The data points in the radar plot are normalized values for different metrics collected under varying UE densities. Key metrics such as rApp CPU, rApp Memory, xApp Handover CPU, xApp Handover Memory, and xApp Monitoring CPU highlight the resource demands for rApp and xApp functionalities.

The rApp CPU usage shows a significant increase with higher UE densities, starting from a normalized value of approximately 1.92 at 16 UEs to 100 at 1024 UEs. This trend indicates the rApp's substantial processing demands as the network scales. Similarly, rApp memory consumption follows a rising trajectory, reflecting the increased memory requirements to handle more extensive data processing and storage needs for higher UE densities.

xApp Handover CPU and Memory also display a steep rise, particularly notable at higher UE densities. For instance, xApp Handover CPU utilization jumps from a normalized value of 1 at 16 UEs to 100 at 1024 UEs, underscoring the computational intensity involved in managing handovers efficiently. The xApp Monitoring CPU metric indicates the rising computational overhead for continuous network monitoring and real-time adjustments, crucial for maintaining service quality and reliability.

The plot includes energy-related metrics such as Energy Consumed and E2Ns ON, which are critical for evaluating the EE of the network under different load conditions. Energy consumption is relatively stable at lower UE densities but increases sharply beyond 128 UEs, indicating that the system's EE diminishes as the network load escalates. At the highest density (1024 UEs), the energy consumed reaches a normalized value of 100, emphasizing the need for efficient energy management strategies in high-demand scenarios.

The number of active E2Ns remains low for smaller UE counts but escalates significantly at higher densities. This metric highlights the scalability of the network infrastructure and the corresponding energy overhead associated with maintaining a larger number of active nodes to support increased user demands.

The analysis of the radar plot indicates several critical insights for the adaptive management of 6G O-RAN systems. The substantial rise in CPU and memory usage for both rApp and xApp functions with increasing UE densities points to the scalability challenges inherent in managing a large number of users. Efficient resource allocation and load-balancing mechanisms are essential to handle these demands without compromising performance.

The sharp increase in energy consumption at higher UE densities underscores the importance of implementing energy-efficient protocols and adaptive power management strategies. Techniques such as dynamic scaling of E2Ns and optimized handover management can play a vital role in mitigating energy overheads. Balancing the computational resource requirements with EE is crucial for sustainable network operations. The data suggests that proactive optimization of resource utilization, such as predictive scaling and intelligent orchestration of network functions, can help achieve this balance.

In conclusion, the radar plot provides a comprehensive view of the interplay between computational resources and EE in a 6G O-RAN environment. By analyzing these metrics, network operators can develop strategies to enhance scalability, optimize resource usage, and improve overall EE, thereby ensuring robust and sustainable network performance in the face of dynamic user demands.

The insights gained from this section help answer the research questions by demonstrating how adaptive network management techniques, such as dynamic activation of E2Ns, efficient CPU and memory allocation, and robust monitoring and handover strategies, address dynamic user demands in 6G O-RAN. Specifically, the detailed analysis of energy consumption, resource utilization, and performance metrics provides a clear understanding of how the proposed architecture can efficiently manage fluctuating user loads. This empirical evidence supports the effectiveness of the orchestration strategy in maintaining high network performance and energy efficiency, thereby validating the integration and disaggregation approaches for Near-RT RIC functions in managing user demands.

6.5 Summarizing

This chapter has presented a detailed analysis of the results derived from the studies conducted during this thesis, specifically focusing on the effectiveness of monitoring strategies, the optimization of RICs, and their deployment in 6G O-RAN environments. Key findings from these studies, published in (BRUNO et al., 2023a,b; ALMEIDA et al., 2023; BRUNO et al., 2024a), provide essential insights into the practical application of these technologies and their impact on network performance and efficiency. The analysis of energy consumption, resource utilization, and performance metrics for rApp Energy Savings, xApp Handover, and xApp Monitoring demonstrated the system's capability to optimize energy usage and maintain high performance under varying user loads. The radar plot in Subsection 6.4.6 provided a visual comparison of normalized resource utilization and EE metrics across different UE densities, offering valuable insights into the scalability and efficiency of the proposed framework.

The integration and synthesis of these findings culminate in the proposed adaptive network management framework, validated through the energy-saving strategies discussed in this chapter. This comprehensive evaluation underscores the practical applicability and significance of the research, highlighting the importance of optimizing resource allocation and processing efficiency in modern telecommunication systems. In the next chapter, we discuss the implications of these findings for future research and practical implementations, offering recommendations for further improving the scalability, efficiency, and performance of 6G O-RAN networks. This discussion provides a roadmap for advancing the state of the art in adaptive network management, ensuring robust and sustainable network operations in the face of dynamic user demands and evolving technological landscapes.

7 CONCLUSION

The research presented in this thesis has explored the adaptive network management framework for 6G O-RAN, specifically focusing on addressing dynamic user demands and enhancing EE. The proposed framework integrates several critical components to ensure optimal network performance and resource utilization, including the Near-RT RIC, the SMO, and specialized applications such rApps and xApps.

This research aimed to answer the following primary question: *How can RAN be orchestrated optimally to accommodate fluctuating user demands?* To address this, the study was guided by several sub-questions:

SRQ1: How can components of an open radio access network be integrated to address dynamic user demands efficiently?

The proposed integration strategies involving adaptive resource allocation, effective data traffic management, and intelligent orchestration demonstrated significant improvements in network responsiveness and resource utilization. This was achieved through the synergy of Near-RT RIC, Non-RT RIC, and the SMO framework.

SRQ2: What specific use case most accurately represents fluctuating user demand in 6G networks?

The use case of energy-saving strategies within the O-RAN framework, such as Carrier and Cell Switch Off/On, RF Channel Reconfiguration, and Sleep Modes, were identified and validated. These strategies enabled dynamic adaptation to changing user demands, optimizing energy efficiency and operational performance, which are crucial for managing the variability in user activity typical of 6G networks.

SRQ3: What strategies can effectively disaggregate Near-RT RIC functions to manage fluctuating user demands?

The research identified and evaluated effective strategies for disaggregating Near-RT RIC functions, including implementing centralized and distributed models. These approaches ensured scalability, rapid adaptability, and efficient resource management in response to dynamic user demands in 6G networks.

SRQ4: What criteria and methods effectively evaluate the proposed orchestration strategy in response to fluctuating user demands?

The proposed orchestration strategy was evaluated based on criteria such as scalability, flexibility, and rapid adaptability to changing network conditions. Advanced simulation tools and real-time monitoring systems were used to track key performance indicators like latency, throughput, and resource utilization.

A significant contribution of this work is the architecture presented in Figure 11 (referred to as the framework). Based on this architecture, various use cases can be explored, such as energy savings. The development and implementation of the rApp Energy Savings, xApp Handover, and xApp Monitoring serve as substantial contributions within a specific use case to exercise

the architecture. The empirical results demonstrate that these applications effectively manage CPU and memory usage, optimize handover processes, and maintain consistent performance across network scenarios. Specifically, the rApp Energy Savings markedly improved in EE by dynamically adjusting the number of active E2N based on user demand, thereby reducing overall energy consumption without compromising service quality.

In conclusion, this thesis underscores the importance of adaptive resource allocation, effective data traffic management, and intelligent orchestration in modern telecommunication networks. The findings provide a robust foundation for future advancements in 6G O-RAN environments, paving the way for more efficient, flexible, and sustainable network operations.

7.1 Contributions

This work presents significant advancements in adaptive network management for 6G O-RAN, focusing on EE and dynamic user demand adaptation. Our contributions are as follows:

- **Adaptive Network Management Framework:** We propose a versatile and scalable framework that fully aligns with O-RAN standards, integrating real-time analytics to manage network configurations dynamically. This framework facilitates adaptive control over network resources, ensuring optimal performance under varying conditions.
- **Theoretical Formulation and Optimization:** A comprehensive mathematical model is developed to minimize energy consumption across RAN nodes while optimizing the placement of Near-RT RIC instances. This model supports real-time decision-making, contributing to the sustainability and cost-effectiveness of future RAN deployments.
- **Energy-Efficiency Techniques:** Leveraging O-RAN's open architecture, we introduce novel techniques for energy optimization in RAN nodes. These techniques reduce power consumption and maintain or enhance service quality, showcasing a practical approach to greener network operations.
- **Interoperability and Standards Compliance:** Our evaluation demonstrates the proposed solutions' compliance with existing 3GPP and O-RAN standards, ensuring that our contributions can seamlessly be integrated into current and future RAN architectures without requiring extensive modifications.
- **Open Source Contributions:** To foster further research and development in the field, we are releasing our codebase and datasets as open-source resources. This initiative provides a foundation for future studies and innovations in adaptive network management within the O-RAN ecosystem.

These contributions collectively advance the state-of-the-art adaptive network management for 6G O-RAN, offering practical, theoretical, and methodological insights that pave the way for more sustainable, efficient, and adaptable wireless networks.

7.2 Limitations

The devised approach exhibits efficacy predominantly in regions with dense antenna coverage. This condition predicates its application mainly to areas where multiple antennas densely populate a confined vicinity, excluding sparser networks often found in remote or developing regions. Moreover, the integration of the proposed system is confined to the RIC as specified by the OSC. This limitation narrows its adaptability and potential applicability across different segments of the RAN ecosystem, thus restricting its widespread deployment.

A critical limitation of this study is the reliance on simulated environments for evaluating the E2N and UEs. The absence of empirical validation using actual RU and real-world scenarios may limit the findings' applicability. While simulations provide valuable insights, real-world deployments are essential for comprehensively assessing the system's performance and efficiency.

The limitations identified in this study, such as the reliance on simulated environments and the need for dense antenna coverage, highlight the necessity for further research. Future work should focus on extending the framework's applicability to real-world deployments, particularly in rural and developing regions with sparse antenna coverage. Additionally, expanding the integration of the framework to other segments of the RAN ecosystem, such as Flexible RAN Intelligent Controller (FlexRIC) and Software-Defined Radio Access Network (SD-RAN), could enhance its versatility and effectiveness.

These constraints highlight the need for further research and development to overcome the current limitations, thereby extending the applicability and enhancing the effectiveness of adaptive network management strategies in future 6G O-RAN architectures.

7.3 Future Work

Addressing the limitations identified in this study opens several avenues for future research. Extending the proposed framework's operational viability beyond regions with dense antenna coverage represents a significant area of exploration. Future works may focus on developing adaptive algorithms that efficiently manage network resources in areas with sparse antenna coverage, thus broadening the framework's applicability to rural or developing regions.

Additionally, the current system's confinement to the software of OSC suggests a more flexible and comprehensive integration strategy. Future research could explore extending compatibility to encompass other projects within the RAN ecosystem, such as FlexRIC and SD-RAN. This approach would enhance the system's adaptability and potential to contribute to the O-RAN community's broader objectives.

The reliance on simulated environments for system evaluation presents another critical area for future work. To bridge the gap between theoretical efficacy and real-world applicability, it is imperative to conduct extensive field trials using actual RU and real network environments.

Such empirical validations would provide valuable insights into the system's performance under operational conditions, identifying potential areas for optimization and refinement.

Moreover, integrating AI and ML techniques offers a promising avenue to enhance the system's decision-making processes, particularly in dynamic network conditions. Applying AI/ML could provide more sophisticated energy management, resource allocation, and user connectivity solutions, further improving network efficiency and user experience. With this approach, it will be feasible to increase the number of UEs at experiments.

Given the evolving nature of cellular networks and the transition towards 6G, future work should also consider the implications of emerging technologies and standards. Investigating how the proposed framework can be adapted or extended to support next-generation network features and requirements will be crucial for maintaining its relevance and effectiveness amidst technological advancements.

An essential next step is the integration of rApp and SMO through the R1 interface. This integration will significantly enhance the system's data collection, enabling more efficient management of network resources and improved service delivery. Future research should focus on developing and testing this integration to evaluate its impact on network performance and reliability. The proposed framework can be refined and expanded through targeted research and development efforts addressing these areas, ultimately realizing more adaptive, efficient, and inclusive next-generation cellular networks.

7.4 Publications

Our series of scholarly contributions commenced with an investigation into the observability of cloud-native 5G systems, as delineated in our 2022 Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) presentation, *"An Empirical Investigation into Observability in Cloud-Native 5G Systems"*(RODRIGUES et al., 2022) (refer to Figure31). This foundational study underscored the importance of tools such as Prometheus and Grafana for enhancing system reliability through meticulous analysis of metrics and logs, laying the groundwork for our proficiency in system observability.

That same year, we extended our research to the efficient management of network resources, culminating in the publication of *"Admission Control for Network Slicing Aware of Network and Processing Resources"* (LIMA et al., 2022). This work introduced algorithms for admission control within network slicing, striving for a harmonious balance between network capabilities and processing resources to uphold Service Level Agreements (SLAs).

Our exploration further ventured into the integration of Unmanned Aerial Vehicles (UAVs) with AI/ML algorithms under the O-RAN architecture, as demonstrated in *"Improved Support for UAV-Based Computer Vision Applications in Search and Rescue Operations via RIC"* (MACEDO et al., 2022), presented at Simpósio Brasileiro de Telecomunicações (SBRT). This pivotal study advanced the practical application of theoretical insights to enhance emergency response strate-

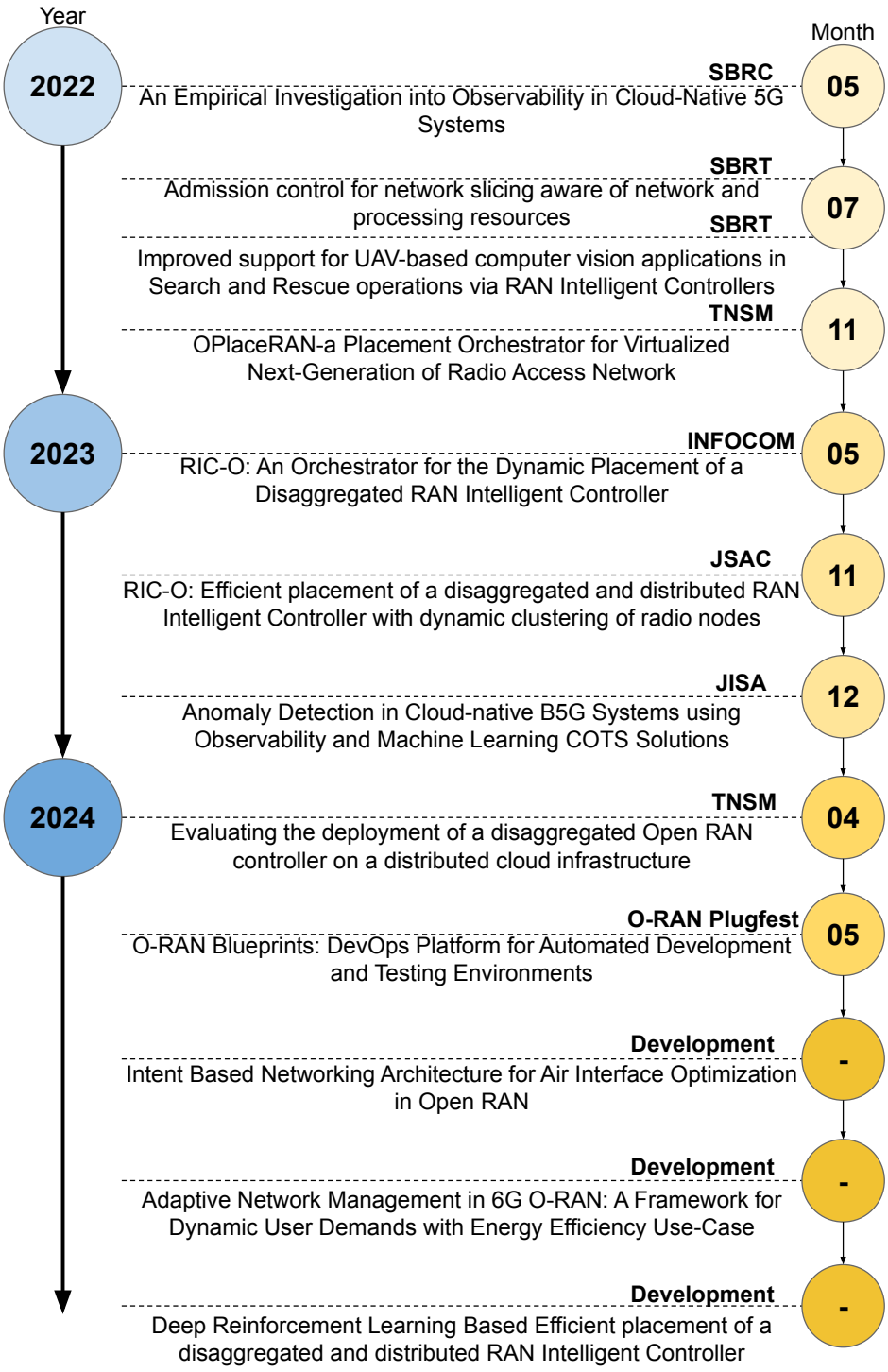


Figure 31 – Timeline of Publications.

gies.

At IEEE Transactions on Network and Service Management (TNSM), our publication, *"Orchestrator Placement for Radio Access Network (OPlaceRAN): A Placement Orchestrator for Virtualized Next-Generation of RAN"* (MORAIS et al., 2023), expanded our expertise into containerization and K8s, focusing on orchestrating components of the Near-RT RIC.

The evolution of our research is further evidenced by *"RIC-O: An Orchestrator for the*

Dynamic Placement of a Disaggregated RIC" (BRUNO et al., 2023b), presented at IEEE International Conference on Computer Communications (INFOCOM), and its subsequent work, *"RIC-O: Efficient Placement of a Disaggregated and Distributed RIC With Dynamic Clustering of Radio Nodes"* (ALMEIDA et al., 2023) published in IEEE Journal on Selected Areas in Communications (JSAC). These studies introduced the orchestration within the O-RAN framework, illustrating our command over cloud and edge computing, containerization, K8s, and network monitoring. They signify our contribution towards optimizing network performance and reliability through sophisticated orchestration strategies.

Our work in Journal of Internet Services and Applications (JISA), *"Anomaly Detection in Cloud-native B5G Systems using Observability and ML COTS Solutions"* (BRUNO et al., 2023a), advanced our application of ML techniques, employing COTS solutions for managing network health complexities in advanced systems.

A comprehensive review in TNSM, *"Evaluating the Deployment of a Disaggregated O-RAN Controller on a Distributed Cloud Infrastructure"* (BRUNO et al., 2024a), explored cloud-native deployment strategies, optimizing Near-RT RIC deployment within cloud infrastructures.

Our latest contribution, *"O-RAN Blueprints: DevOps Platform for Automated Development and Testing Environments"*, presented at the O-RAN Global PlugFest Spring 2024 (BRUNO et al., 2024b), showcases our practical implementation of advanced network management techniques, further solidifying our position at the forefront of 6G O-RAN research. The same components developed in this work are utilized in the blueprint, and the use case presented aligns with the use case detailed in this thesis.

7.4.1 Works in Development

Currently under development is *"Intent-Based Networking Architecture for Air Interface Optimization in O-RAN"*, focusing on network optimization and air interface refinement to innovate networking architectures further.

Additionally, we are developing *"Adaptive Network Management in 6G O-RAN: A Framework for Dynamic User Demands with EE Use-Case"*, aiming to pioneer in adaptive network management for the forthcoming 6G infrastructure, the work of this thesis.

Progress is also being made on *"Deep Reinforcement Learning Based Efficient Placement of a Disaggregated and Distributed RIC"*, leveraging deep reinforcement learning for dynamic network management, an extension of RIC-O (ALMEIDA et al., 2023) using AI. This study underscores our advancement in ML techniques for telecommunications infrastructure efficiency.

REFERENCES

3GPP. **Service Requirements for the 5G System, document TS 22.261 v16.0.0**. [S.l.]: 3^rd Generation Partnership Project (3GPP), 2017. Technical Specification (TS). (22.261).

3GPP. **Study on new radio access technology; radio access architecture and interfaces (release 14)**. [S.l.]: 3^rd Generation Partnership Project (3GPP), 2017. Technical Recommendation (TR). (38.801).

3GPP. **Technical specification group services and system aspects (release 15)**. [S.l.]: $3^{r\ d}$ Generation Partnership Project (3GPP), 2019. Release 15 Description. (21.915).

ABEDIN, S. F. et al. Elastic o-ran slicing for industrial monitoring and control: a distributed matching game and deep reinforcement learning approach. **IEEE Transactions on Vehicular Technology**, [S.l.], v. 71, p. 10808–10822, 10 2022.

ABUBAKAR, A. I. et al. Energy efficiency of open radio access network: a survey. In: IEEE 97TH VEHICULAR TECHNOLOGY CONFERENCE, 2023., 2023. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2023. v. 2023-June.

AGARWAL, B. et al. Qoe-driven optimization in 5g o-ran-enabled hetnets for enhanced video service quality. **IEEE Communications Magazine**, [S.l.], v. 61, p. 56–62, 1 2023.

AGRAWAL, R. et al. Cloud RAN challenges and solutions. **Annals of Telecommunications**, [S.l.], v. 72, n. 7-8, p. 387–400, 2017.

AL-DULAIMI, A.; WANG, X.; I, C.-L. **5g networks**: fundamental requirements, enabling technologies, and operations management. [S.l.]: John Wiley & Sons, 2018.

ALAVIRAD, M. et al. O-ran architecture, interfaces, and standardization: study and application to user intelligent admission control. **Frontiers in Communications and Networks**, [S.l.], v. 4, 3 2023.

ALI, K.; JAMMAL, M. Proactive vnf scaling and placement in 5g o-ran using ml. **IEEE Transactions on Network and Service Management**, [S.l.], 2023.

ALLIANCE, O.-R. **O-RAN Software Community (SC)**. Accessed on May 28, 2024.

ALMEIDA, G. M. et al. RIC-O Efficient placement of a disaggregated and distributed RAN Intelligent Controller with dynamic clustering of radio nodes. **IEEE Journal on Selected Areas in Communications**, [S.l.], p. 1–1, 2023.

AMIRI, E. et al. Energy-aware dynamic vnf splitting in o-ran using deep reinforcement learning. **IEEE Wireless Communications Letters**, [S.l.], 11 2023.

AYALA-ROMERO, J. A. et al. Edgebol: automating energy-savings for mobile edge ai. In: CONEXT 2021 - PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON EMERGING NETWORKING EXPERIMENTS AND TECHNOLOGIES, 2021. **Anais...** Association for Computing Machinery: Inc, 2021. p. 397–410.

AZARIAH, W. et al. A survey on open radio access networks: challenges, research directions, and open source approaches. **arXiv Preprint**, [S.l.], 2022.

BALASUBRAMANIAN, B. et al. RIC A RAN Intelligent Controller Platform for AI-Enabled Cellular Networks. **IEEE Internet Computing**, [S.l.], v. 25, n. 2, p. 7–17, 2021.

BERNARDOS, C. J. et al. **Network virtualization research challenges**. [S.l.]: IETF, 2019.

BERTENYI, B. et al. Ng radio access network (ng-ran). **Journal of ICT Standardization**, [S.l.], v. 6, n. 1, p. 59–76, 2018.

BESHLEY, M. et al. Energy-efficient qoe-driven radio resource management method for 5g and beyond networks. **IEEE Access**, [S.l.], v. 10, p. 131691–131710, 2022.

BONATI, L. et al. Open, programmable, and virtualized 5g networks: state-of-the-art and the road ahead. **arXiv preprint arXiv:2005.10027**, [S.l.], 2020.

BONATI, L. et al. Neutran: an open ran neutral host architecture for zero-touch ran and spectrum sharing. **IEEE Transactions on Mobile Computing**, [S.l.], p. 1–14, 2023.

BRUNO, G. Z. et al. Anomaly detection in cloudnative b5g systems using observability and machine learning cots solutions. **Journal of Internet Services and Applications**, [S.l.], v. 14, p. 189–199, 1 2023.

BRUNO, G. Z. et al. RIC-O An Orchestrator for the Dynamic Placement of a Disaggregated RAN Intelligent Controller. In: IEEE INFOCOM 2023 - IEEE CONFERENCE ON COMPUTER COMMUNICATIONS WORKSHOPS (INFOCOM WKSHPS), 2023. **Anais...** [S.l.: s.n.], 2023. p. 1–2.

BRUNO, G. Z. et al. Evaluating the deployment of a disaggregated open ran controller on a distributed cloud infrastructure. **IEEE Transactions on Network and Service Management**, [S.l.], p. 1–1, 2024.

BRUNO, G. Z. et al. **O-ran blueprints**: devops platform for automated development and testing environments. 2024.

CAO, Y. et al. Federated Deep Reinforcement Learning for User Access Control in Open Radio Access Networks. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, 2021. **Anais...** [S.l.: s.n.], 2021. p. 1–6.

CAO, Y. et al. User Access Control in Open Radio Access Networks A Federated Deep Reinforcement Learning Approach. **IEEE Transactions on Wireless Communications**, [S.l.], v. 21, n. 6, p. 3721–3736, 2022.

CARDOSO, K. V. et al. A softwarized perspective of the 5G networks. **arXiv Preprint**, [S.l.], 2020.

CHEN, T. et al. Software defined mobile networks: concept, survey, and research directions. **IEEE Communications Magazine**, [S.l.], v. 53, n. 11, p. 126–133, 2015.

D'ORO, S. et al. Orchestran: network automation through orchestrated intelligence in the open ran. In: IEEE INFOCOM PROCEEDINGS, 2022., 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. v. 2022-May, p. 270–279.

DRYJAńSKI, M.; KLIKS, A. **The O-RAN Whitepaper 2022 RAN Intelligent Controller, xApps and rApps**. [S.l.]: RIMEDO Labs, 2022.

DUONG, Q. H. et al. A column generation algorithm for dedicated-protection o-ran vnf deployment. In: INTERNATIONAL WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, IWCMC 2022, 2022., 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. p. 1206–1211.

ETSI, I. **Network functions virtualisation (nfv)-network operator perspectives on industry progress**. [S.l.: s.n.], 2013.

FIORANI, M. et al. Transport abstraction models for an SDN-controlled centralized RAN. **IEEE Communications Letters**, [S.l.], v. 19, n. 8, p. 1406–1409, 2015.

FIORANI, M. et al. Modeling energy performance of c-ran with optical transport in 5g network scenarios. **Journal of Optical Communications and Networking**, [S.l.], v. 8, 2016.

FULBER-GARCIA, V. et al. An etsi-compliant architecture for the element management system: the key for holistic nfv management. In: INTERNATIONAL CONFERENCE ON NETWORK AND SERVICE MANAGEMENT (CNSM), 2023., 2023. **Anais...** [S.l.: s.n.], 2023. p. 1–9.

GARCIA-SAAVEDRA, A. et al. O-ran: disrupting the virtualized ran ecosystem. **IEEE Communications Standards Magazine**, [S.l.], n. September, p. 1–8, 2021.

GUERZONI, R. et al. Network functions virtualisation: an introduction, benefits, enablers, challenges and call for action, introductory white paper. In: SDN AND OPENFLOW WORLD CONGRESS, 2012. **Anais...** [S.l.: s.n.], 2012. v. 1, p. 5–7.

HABIBI, M. et al. A comprehensive survey of ran architectures toward 5G mobile communication system. **IEEE Access**, [S.l.], v. 7, p. 70371–70421, 2019.

HAMMAMI, N.; NGUYEN, K. K. On-policy vs. off-policy deep reinforcement learning for resource allocation in open radio access network. In: IEEE WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, WCNC, 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. v. 2022-April, p. 1461–1466.

HOJEIJ, H. et al. Dynamic placement of o-cu and o-du functionalities in open-ran architecture. In: ANNUAL IEEE COMMUNICATIONS SOCIETY CONFERENCE ON SENSOR, MESH AND AD HOC COMMUNICATIONS AND NETWORKS WORKSHOPS, 2023. **Anais...** IEEE Computer Society, 2023. v. 2023-September, p. 330–338.

HUANG, Y. C. et al. Universal vertical applications adaptation for open ran: a deep reinforcement learning approach. In: INTERNATIONAL SYMPOSIUM ON WIRELESS PERSONAL MULTIMEDIA COMMUNICATIONS, WPMC, 2022. **Anais...** IEEE Computer Society, 2022. v. 2022-October, p. 92–97.

HUFF, A.; HILTUNEN, M.; DUARTE, E. P. RFT, Scalable and Fault-Tolerant Microservices for the O-RAN Control Plane. In: IFIP/IEEE INTERNATIONAL SYMPOSIUM ON INTEGRATED NETWORK MANAGEMENT (IM), 2021. **Anais...** [S.l.: s.n.], 2021. p. 402–409.

ITU-T, G.-T. **Transport network support of imt-2020/5G**. [S.l.]: ITU-T, 2018.

ITU-T, G.-T. **Application of optical transport network recommendations to 5G transport - serir g supplement 67**. [S.l.]: ITU-T, 2019.

ITU-T, S. S.; OF. **ITU-T focus group IMT-2020 deliverables**. [S.l.: s.n.], 2017.

JAAFARI, M. E.; CHUBERRE, N. Guest editorial ijscn special issue on 3gpp ntn standards for future satellite communications. **International Journal of Satellite Communications and Networking**, [S.l.], v. 41, p. 217–219, 2023.

JOHNSON, D.; MAAS, D.; MERWE, J. V. D. NexRAN Closed-loop RAN slicing in POWDER -A top-to-bottom open-source open-RAN use case. In: WORKSHOP WIRELESS NET. TESTBEDS, EXPERIMENTAL EVALUATION, 15., 2021. **Anais...** [S.l.: s.n.], 2021. p. 17–23.

KALNTIS, M.; IOSIFIDIS, G. Energy-aware scheduling of virtualized base stations in o-ran with online learning. In: IEEE GLOBAL COMMUNICATIONS CONFERENCE, GLOBECOM 2022 - PROCEEDINGS, 2022., 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. p. 6048–6054.

KASULURU, V. et al. On the use of probabilistic forecasting for network analysis in open ran. In: IEEE INTERNATIONAL MEDITERRANEAN CONFERENCE ON COMMUNICATIONS AND NETWORKING, MEDITCOM 2023, 2023., 2023. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2023. p. 258–263.

KATZ, M.; MATINMIKKO-BLUE, M.; LATVA-AHO, M. 6genesis flagship program: building the bridges towards 6g-enabled wireless smart society and ecosystem. In: IEEE 10TH LATIN-AMERICAN CONFERENCE ON COMMUNICATIONS (LATINCOM), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 1–9.

KAZEMIFARD, N. et al. Minimum delay function placement and resource allocation for open ran (o-ran) 5g networks. **Computer Networks**, [S.l.], v. 188, 4 2021.

KOUCHAKI, M. et al. Actor-critic network for o-ran resource allocation: xapp design, deployment, and analysis. In: IEEE GLOBECOM WORKSHOPS, GC WKSHPS 2022 - PROCEEDINGS, 2022., 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. p. 968–973.

LACAVA, A. et al. Programmable and customized intelligence for traffic steering in 5g networks using open ran architectures. **IEEE Transactions on Mobile Computing**, [S.l.], 2023.

LARSEN, L. M. et al. A survey of the functional splits proposed for 5G mobile crosshaul networks. **IEEE Communications Surveys & Tutorials**, [S.l.], v. 21, n. 1, p. 146–172, 2018.

LARSEN, L. M. et al. A survey of the functional splits proposed for 5g mobile crosshaul networks. **IEEE Communications Surveys and Tutorials**, [S.l.], v. 21, p. 146–172, 1 2019.

LIMA, H. V. de et al. Controle de admissao para network slicing ciente de recursos de rede e de processamento. In: XL SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES. SOCIEDADE BRASILEIRA DE TELECOMUNICAÇÕES, 2022. **Anais...** [S.l.: s.n.], 2022.

LO SCHIAVO, L. et al. Cloudric: open radio access network (o-ran) virtualization with shared heterogeneous computing. In: ACM INTERNATIONAL CONFERENCE ON MOBILE COMPUTING AND NETWORKING, 2024. **Anais...** [S.l.: s.n.], 2024.

MACEDO, C. J. et al. Improved support for uav-based computer vision applications in search and rescue operations via ran intelligent controllers. In: XL SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES. SOCIEDADE BRASILEIRA DE TELECOMUNICAÇÕES, 2022. **Anais...** [S.l.: s.n.], 2022.

MAI, Z. et al. An energy efficiency optimization jointing resource allocation for delay-aware traffic in fronthaul constrained c-ran. **Wireless Networks**, [S.l.], v. 29, p. 353–368, 1 2023.

MARSCH, P. et al. **5G system design**: architectural and functional considerations and long term research. [S.l.]: John Wiley & Sons, 2018.

MORAIS, F. Z. et al. Oplaceran - a placement orchestrator for virtualized next-generation of radio access network. **IEEE Transactions on Network and Service Management**, [S.l.], v. 20, p. 3274–3288, 9 2023.

MUNGARI, F. An rl approach for radio resource management in the o-ran architecture. In: ANNUAL IEEE COMMUNICATIONS SOCIETY CONFERENCE ON SENSOR, MESH AND AD HOC COMMUNICATIONS AND NETWORKS WORKSHOPS, 2021. **Anais...** IEEE Computer Society, 2021. v. 2021-July.

NAHUM, C. V. et al. Testbed for 5g connected artificial intelligence on virtualized networks. **IEEE Access**, [S.l.], v. 8, p. 223202–223213, 2020.

O-RAN Alliance. **O-ran**: towards an open and smart ran. [S.l.: s.n.], 2018. White Paper.

O-RAN Alliance. **O-ran architecture description - v09.00**. [S.l.: s.n.], 2023. White Paper.

O-RAN Alliance. **O-RAN Near-RT RIC Architecture 5.0**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG3.RICARCH-R003-v05.00).

O-RAN Alliance. **O-RAN Non-RT RIC Architecture 4.0**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG2.Non-RT-RIC-ARCH-R003-v04.00).

O-RAN Alliance. **O-RAN E2 General Aspects and Principles**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG3.E2GAP-R003).

O-RAN Alliance. **O-ran massive mimo use cases**. [S.l.: s.n.], 2023. White Paper. (O-RAN.WG1.MMIMO-USE-CASES-TR-v01.00).

O-RAN Alliance. **O-ran slicing architecture**. [S.l.: s.n.], 2023. White Paper. (O-RAN.WG1.Slicing-Architecture-R003-v11.00).

O-RAN Alliance. **O-ran network energy saving use cases**. [S.l.: s.n.], 2023. White Paper. (O-RAN.WG1.NESUC-R003-v02.00).

O-RAN Alliance. **O-RAN Fronthaul Control, User and Synchronization Plane Specification**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG4.CUS.0-R003-v13.00).

O-RAN Alliance. **O-RAN A1 interface Use Cases and Requirements**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG2.A1UCR-R003).

O-RAN Alliance. **O-RAN NR C-plane profile**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG5.C.1-R003).

O-RAN Alliance. **O-RAN Cloudification and Orchestration Use Cases and Requirements for O-RAN Virtualized RAN**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG6.ORCH-USE-CASES-R003).

O-RAN Alliance. **O-RAN Security Aspects in O-RAN Specifications**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG11.Security-Requirements-Specification.O-R003).

O-RAN Alliance. **O-RAN O1 Interface Specification for Near Real Time RAN Intelligent Controller 1.0**. [S.l.]: O-RAN Alliance, 2023. (O-RAN.WG3.O1-Interface-for-Near-RT-RIC-R003-v01.00).

O-RAN Alliance. **R1 interface**: general aspects and principles. [S.l.: s.n.], 2024. White Paper. (O-RAN.WG2.R1GAP-R003-v08.00).

O-RAN Alliance. **Open fronthaul interfaces wg management plane specification**. [S.l.: s.n.], 2024. White Paper. (O-RAN.WG4.MP.0-R003-v14.00).

ORHAN, O. et al. Connection management xapp for o-ran ric: a graph neural network and reinforcement learning approach. In: IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS, ICMLA 2021, 20., 2021. **Proceedings...** Institute of Electrical and Electronics Engineers Inc., 2021. p. 936–941.

PLIATSIOS, D. et al. Realizing 5G vision through Cloud RAN: technologies, challenges, and trends. **EURASIP Journal on Wireless Communications and Networking**, [S.l.], n. 1, p. 136, 2018.

RODRIGUES, K. B. C. et al. Uma investigação empírica sobre observabilidade em sistemas 5g nativos de nuvem. In: XL SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS, 2022. **Anais...** [S.l.: s.n.], 2022. p. 252–265.

SABELLA, D. et al. Designing the 5g network infrastructure: a flexible and reconfigurable architecture based on context and content information. **Eurasip Journal on Wireless Communications and Networking**, [S.l.], v. 2018, 12 2018.

SALVAT, J. X. et al. Open radio access networks (o-ran) experimentation platform: design and datasets. **IEEE Communications Magazine**, [S.l.], v. 61, p. 138–144, 9 2023.

SARIKAYA, E. et al. Placement of 5g ran slices in multi-tier o-ran 5g networks with flexible functional splits. In: INTERNATIONAL CONFERENCE ON NETWORK AND SERVICE MANAGEMENT: SMART MANAGEMENT FOR FUTURE NETWORKS AND SERVICES, CNSM 2021, 2021., 2021. **Proceedings...** Institute of Electrical and Electronics Engineers Inc., 2021. p. 274–282.

SCHMIDT, R.; IRAZABAL, M.; NIKAEIN, N. FlexRIC, an SDK for next-generation SD-RANs. In: INTERNATIONAL CONFERENCE ON EMERGING NETWORKING EXPERIMENTS AND TECHNOLOGIES (CONEXT), 17., 2021. **Anais...** [S.l.: s.n.], 2021. p. 411–425.

SIDDIQI, M. A.; YU, H.; JOUNG, J. 5g ultra-reliable low-latency communication implementation challenges and operational issues with iot devices. **Electronics (Switzerland)**, [S.l.], v. 8, 9 2019.

SINGH, A. K.; KHOA NGUYEN, K. Joint Selection of Local Trainers and Resource Allocation for Federated Learning in Open RAN Intelligent Controllers. In: IEEE WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE (WCNC), 2022. **Anais...** [S.l.: s.n.], 2022. p. 1874–1879.

SOHAIB, R. et al. Green resource allocation in cloud-native o-ran enabled small cell networks. **arXiv preprint arXiv:2407.11563**, [S.l.], 2024.

TATARIA, H. et al. 6g wireless systems: vision, requirements, challenges, insights, and opportunities. **Proceedings of the IEEE**, [S.l.], v. 109, p. 1166–1199, 7 2021.

THALIATH, J. et al. Predictive closed-loop service automation in o-ran based network slicing. **IEEE Communications Standards Magazine**, [S.l.], v. 6, p. 8–14, 9 2022.

TOMKOS, I. et al. Toward the 6g network era: opportunities and challenges. **IT Professional**, [S.l.], v. 22, p. 34–38, 1 2020.

TSUKAMOTO, Y. et al. Experimental evaluation of ran slicing architecture with flexibly located functional components of base station according to diverse 5g services. **IEEE Access**, [S.l.], v. 7, p. 76470–76479, 2019.

VdoCipher. **Video bandwidth requirements explained**. Accessed: 2024-03-21, `https://www.vdocipher.com/blog/video-bandwidth-explanation`.

VILA, I. et al. On the implementation of a reinforcement learning-based capacity sharing algorithm in o-ran. In: IEEE GLOBECOM WORKSHOPS, GC WKSHPS 2022 - PROCEEDINGS, 2022., 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. p. 208–214.

WANG, L. et al. Energy conserved computation offloading for o-ran based iot systems. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, 2022. **Anais...** Institute of Electrical and Electronics Engineers Inc., 2022. v. 2022-May, p. 4043–4048.

WANG, L. et al. Minimizing energy consumption of iot devices for o-ran based iot systems. **Energy Reports**, [S.l.], v. 9, p. 379–388, 11 2023.

YOO, H. M. et al. Load balancing algorithm running on open ran ric. In: INTERNATIONAL CONFERENCE ON ICT CONVERGENCE, 2022. **Anais...** IEEE Computer Society, 2022. v. 2022-October, p. 1226–1228.

YOUSAF, F. Z. et al. Manoaas: a multi-tenant nfv mano for 5g network slices. **IEEE Communications Magazine**, [S.l.], v. 57, p. 103–109, 5 2019.

ZORELLO, L. M. M. et al. Power-efficient baseband-function placement in latency-constrained 5g metro access. **IEEE Transactions on Green Communications and Networking**, [S.l.], v. 6, p. 1683–1696, 9 2022.