UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

WILLIAM TEIXEIRA PIRES JUNIOR

Energy-aware approaches to resource allocation in open radio access networks

Goiânia
2026

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

William Teixeira Pires Junior

3. Título do trabalho

Energy-aware approaches to resource allocation in open radio access networks

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Kleber Vieira Cardoso**, **Professor do Magistério Superior**, em 27/02/2026, às 11:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **William Teixeira Pires Junior**, **Discente**, em 27/02/2026, às 11:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6011958** e o código CRC **5CCB3EC8**.

Referência: Processo nº 23070.066221/2025-89 SEI nº 6011958

WILLIAM TEIXEIRA PIRES JUNIOR

Energy-aware approaches to resource allocation in open radio access networks

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Informática, da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação.

Linha de Pesquisa: Sistemas de Computação.

Orientador: Prof. Dr. Kleber Vieira Cardoso

Co-Orientador: Prof. Dr. Leizer de Lima Pinto

Goiânia
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Pires Junior, William Teixeira
Energy-aware approaches to resource allocation in open radio access networks [manuscrito] = Abordagens Sensíveis à energia para alocação de recursos em redes de acesso por rádio abertas / William Teixeira Pires Junior. - 2026.
LXXVIII, 78 f.: 2026

Orientador: Prof. Dr. Kleber Vieira Cardoso; co-orientador: Dr. Leizer de Lima Pinto
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2026.
Bibliografia.
Inclui: siglas, simbolos, tabelas, algoritmos, grafico, lista de figuras, lista de tabelas.

1. Mobile Network. 2. Radio Access Network. 3. Energy-efficiency. 4. Resource Orchestration. 5. Mathematical Programming.

I. Cardoso, Kleber Vieira, orient. II. Pinto, Leizer de Lima, co-orient. III. Título.
CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **2** da sessão de Defesa de Dissertação de **William Teixeira Pires Junior**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos nove dias do mês de fevereiro de dois mil e vinte e seis, a partir das nove horas, na sala 257 do INF, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Energy-aware approaches to resource allocation in open radio access networks"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Kleber Vieira Cardoso (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Leizer de Lima Pinto (INF/UFG), coorientador; Professor Doutor Aldebaro Barreto da Rocha Klautau Júnior (UFPA), membro titular externo; e Professor Doutor Humberto José Longo (INF/UFG), membro titular interno. A participação do professor Aldebaro Barreto da Rocha Klautau Júnior ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Kleber Vieira Cardoso, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos nove dias do mês de fevereiro de dois mil e vinte e seis.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Humberto Jose Longo**, **Professor do Magistério Superior**, em 09/02/2026, às 11:21, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Kleber Vieira Cardoso**, **Professor do Magistério Superior**, em 09/02/2026, às 11:23, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Leizer De Lima Pinto**, **Professor do Magistério Superior**, em 09/02/2026, às 11:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **William Teixeira Pires Junior**, **Discente**, em 09/02/2026, às 11:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Aldebaro Barreto da Rocha Klautau Junior**, **Usuário Externo**, em 09/02/2026, às 17:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Referência: Processo nº 23070.066221/2025-89 SEI nº 5909429

Resumo

Pires-Jr, William Teixeira. **Energy-aware approaches to resource allocation in open radio access networks**. Goiânia, 2026. 78p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

Abordagens Sensíveis à energia para alocação de recursos em redes de acesso por rádio abertas A evolução das redes de comunicação móvel rumo ao 5G e além busca atender às demandas por alta vazão de dados, baixa latência e eficiência energética. As Redes Abertas de Acesso por Rádio (O-RAN), com sua arquitetura desagregada e virtualizada, oferecem flexibilidade e interoperabilidade, mas também impõem desafios significativos em termos de consumo de energia. Este trabalho aborda a alocação de recursos com consciência energética em O-RAN, com foco no posicionamento de Funções de Rede Virtualizadas (VNF) e na associação de equipamentos de usuário. Propomos um modelo de Programação Linear Inteira Mista que considera conjuntamente os custos energéticos dos equipamentos de rádio, rede de transporte e migração de VNFs, ao mesmo tempo em que suporta opções flexíveis de divisão funcional da pilha de protocolos da RAN e decisões de roteamento, complementado por uma heurística para escalabilidade. Utilizando geradores sintéticos de carga e topologia, avaliamos cenários variados e mostramos que topologias hierárquicas alcançam até 15% mais centralização e reduzem o consumo de energia em cerca de 28% em comparação com topologias amplamente adotadas atualmente. Além disso, a otimização conjunta do posicionamento de VNFs e da associação de equipamentos de usuário possibilita uma melhoria significativa da eficiência energética ao desativar completamente estações base em momentos de baixa carga. Uma abordagem disjunta para o problema é capaz de resolver instâncias maiores e alcançar soluções próximas ao ótimo, superando aquelas baseadas apenas na associação com máxima Relação Sinal-Ruído (SNR). Apesar das soluções ótimas não atenderem o tempo de resposta exigido em implantações práticas, elas oferecem um resultado de referência robusto para a avaliação da rede e de abordagens não ótimas para o problema.

Palavras–chave

Rede Móvel, Rede de Acesso por Rádio, Eficiência Energética, Orquestração de Recursos, Programação Matemática.

Abstract

Pires-Jr, William Teixeira. **Energy-aware approaches to resource allocation in open radio access networks**. Goiânia, 2026. 78p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

The evolution of mobile communication networks toward 5G and beyond seeks to meet the demands for high throughput, low latency, and energy efficiency. Open Radio Access Networks (O-RAN), with their disaggregated and virtualized architecture, provide flexibility and interoperability but also pose significant challenges in terms of energy consumption. This work addresses energy-aware resource allocation in O-RAN, focusing on Virtualized Network Function (VNF) placement and user equipment association. We propose a Mixed Integer Linear Programming model that jointly considers radio equipment, transport network, and VNF migration energy costs while supporting flexible functional split options and routing decisions, complemented by a heuristic for scalability. Using synthetic load and topology generators, we evaluate diverse scenarios and show that hierarchical topologies achieve up to 15% more centralization and reduce energy consumption by about 28% compared to current widely adopted topologies. Additionally, joint optimization of VNF placement and user equipment association enables significant energy savings by disabling entire base stations during moments of low load. A disjoint approach to the problem is able to solve larger instances of the problem and achieves solutions close to optimal while surpassing a maximum-Signal to Noise Ratio (SNR) solution. Despite the optimal solutions being unable to meet the stringent response time required in practical deployments, they present a robust baseline for the evaluation of both the network and non-optimal approaches to the problem.

Keywords

Mobile network, radio access network, energy efficiency, resource orchestration, mathematical programming.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

B&B Branch-and-bound
BBU Baseband Unit
BS Base Station
CoMP Coordinated Multi-Point
C-RAN Centralized Radio Access Network
CR Computing Resource
CU Centralized Unit
DWDM Dense Wavelength Division Multiplexing
D-RAN Distributed Radio Access Network
DRL Deep Reinforcement Learning
DU Distributed Unit
eMBB Enhanced Mobile Broadband
Gbps Gigabits Per Second
GOPS Giga Operations Per Second
GPP General Purpose Processor
ICT Information and Communication Technology
ITU International Telecommunication Union
LP Linear Programming
ILP Integer Linear Programming
MILP Mixed Integer Linear Programming
MINLP Mixed Integer Nonlinear Programming
MIMO Multiple Input Multiple Output
MIP Mixed Integer Programming
mMTC Massive Machine Type Communications
NFV Network Function Virtualization
NLP Nonlinear Programming
OAI OpenAirInterface
OFDM Orthogonal Frequency Division Multiplexing
OFDMA Orthogonal Frequency Division Multiple Access
O-RAN Open Radio Access Network

PUE Power Usage Effectiveness
QoS Quality of Service
RAN Radio Access Network
RB Resource Block
RF Radio Frequency
RIC Radio Access Network Intelligent Controller
RNN Recurrent Neural Network
RRH Remote Radio Head
RU Radio Unit
SDN Software Defines Network
SINR Signal to Interference plus Noise Ratio
SNR Signal to Noise Ratio
UE User Equipment
URLLC Ultra Reliable Low Latency Communication
vBBU virtualized Baseband Unit
vCU virtual Centralized Unit
vDU virtual Distributed Unit
VM Virtual Machine
VNF Virtualized Network Function
vRAN Virtualized Radio Access Network

CHAPTER 1

Introduction

Energy efficiency has become one of the most pressing challenges in the evolution of mobile communication networks. While each new generation of mobile systems has introduced innovations to support emerging use cases, the environmental and economic impacts of their operation have grown substantially. Today, the Information and Communication Technology (ICT) sector accounts for around 3.5–4.0% of global electricity consumption, with telecommunication infrastructure alone responsible for about 1% of the total. Moreover, energy costs represent up to 40% of operators' expenditures [16]. Within this context, the Radio Access Network (RAN) is particularly critical, responsible for 75–80% of the total energy used by mobile network infrastructure, while nearly 95% of CO_2 emissions associated with network equipment occur during its operation [40]. These statistics highlight the need for energy-aware strategies that can ensure the sustainability of future mobile networks.

The Open Radio Access Network (O-RAN) paradigm offers unprecedented flexibility by disaggregating the RAN protocol stack into Virtualized Network Functions (VNFs) executed on General Purpose Processors (GPPs). This flexibility enables dynamic orchestration and cost reduction; however, it also introduces complex decisions regarding VNF placement and traffic routing. Each of these decisions directly influences the energy footprint of the network. Exploring energy efficiency in this context is therefore essential to conform with the International Telecommunication Union (ITU) directive of aligning the network's evolution with the Sustainable Development Goals (SDGs) established by the United Nations [22].

This dissertation investigates energy-aware approaches to resource allocation in open and virtualized RANs, with a major focus on understanding the trade-offs between optimality, scalability, and efficiency. The research is guided by the following questions:

1. **Which scale can be achieved when optimally solving the problem of VNF placement for minimum energy consumption when considering the problem in its most complete form?**

The first research question is motivated by the observation that existing approaches to energy efficiency in Virtualized Radio Access Network (vRAN) infrastructure frequently simplify the problem formulation to achieve tractability. Common simplifications include neglecting the energy consumption of the transport network, ignoring the overhead associated with VNF migration, or restricting the choice of functional split to a single option. Other works resort to heuristic or machine-learning-based methods to address scalability concerns. However, these approaches often lack a quantitative evaluation of the trade-off between scalability and solution optimality. As a result, it remains unclear to what extent the complete energy-aware VNF placement problem can be solved optimally when all relevant energy consumption components are considered.

In the following chapters, we show that the intractability of optimally solving the problem reported in the literature is a consequence of formulation choices. By applying a simple linearization technique, the problem can be expressed as a Mixed Integer Linear Programming (MILP) model, enabling the use of off-the-shelf solvers for efficient resolution. The results show that instances of considerable size compared to those reported in the literature can be solved optimally within practical time limits. Furthermore, the analysis reveals that energy consumption from computing resources, the transport network, and VNF migration each has the potential to influence the optimal solution, reinforcing the need for a comprehensive formulation. The objectives derived from this research question are:

- Formulate a flexible and comprehensive mathematical programming model to minimize the energy consumption of vRAN systems.
- Evaluate how the solution time for optimization models scales with the size of instances of the problem.

2. **How different RAN topologies and different user equipment network usage profiles affect the energy consumption of the RAN infrastructure?**

The second research question addresses an open gap in the literature concerning the analysis of how different network scenarios influence the energy consumption of network infrastructures [2]. Existing studies predominantly focus on static deployments (with fixed placements for some VNFs) and simplified topologies, such as single-cloud architectures or hexagonal Base Station (BS) distributions, which may not capture the heterogeneity of real-world networks or the full flexibility proposed for 5G/6G networks. To overcome these limitations, a flexible formulation capable of representing arbitrary network scenarios is required. Such formulation enables a comprehensive evaluation of the interactions between User Equipment (UE) traffic, usage profile, network topology, and VNF placement decisions.

Through an evaluation of multiple RAN topologies and heterogeneous UE usage profiles, this dissertation shows that architectural choices play a decisive role in characterizing energy efficiency. For instance, hierarchical topologies can achieve significantly lower energy consumption than ring topologies by enabling more efficient aggregation of VNFs and better utilization of computing resources. At the same time, high-throughput usage profiles tend to amplify transport network energy costs, reducing the benefits of aggressive centralization and favoring more distributed deployments. Conversely, under lower load conditions, greater centralization often leads to energy savings, provided that routing and functional split decisions are jointly optimized. The following objectives were based on this research question:

- Identify different network topology structures and UE network usage profiles for evaluation.
- Quantify the energy consumption for different RAN topologies under similar loads.
- Evaluate the relationship between UE network usage profiles and energy optimized decision trends.

3. **Regarding the problems of UE-BS association and VNF placement. Is there any benefits in adopting a heuristic solution instead of an optimal solution of the problem in its holistic or relaxed formulation?**

The third research question is motivated by the practical limitations of deploying optimal solutions in large-scale or time-sensitive operational environments. An efficient formulation may be capable of presenting solutions in practical time for a technical evaluation of the network. However, solving the UE-BS association and holistic VNF placement problems optimally may not always be feasible within the time constraints imposed by real network deployment requirements. For this reason, heuristic solutions are frequently proposed as an alternative. Still, such heuristics are often evaluated in isolation, without a rigorous comparison against an optimal baseline.

This dissertation seeks to address this question by designing and evaluating a heuristic solution against the optimization model. The results show that the heuristic developed for VNF placement can achieve energy consumption levels lower than those of a fully centralized deployment while remaining close to the optimal solution during periods of low traffic load. Additionally, we evaluate a heuristic approach to the problem of UE-BS association that is commonly used to maximize UE signal quality. While the heuristic is capable of presenting solutions with energy consumption close to optimal in scenarios with high UE density, it may do so at the cost of a reduced admission rate. Furthermore, the optimization model remains essential for achieving significantly lower energy consumption in low UE density scenarios. Overall, these results demonstrate that

while heuristic approaches can offer a solution within an acceptable time frame for real-world network deployments, optimization-based solutions are essential for evaluating non-optimal methods and for clarifying the complex relationships between network operational decisions and energy efficiency. Accordingly, the objectives derived from this research question are:

- Design a heuristic aligned with the holistic formulation for the problem of VNF placement, aiming to minimize energy consumption.
- Quantify the energy gap between the heuristic and optimal solutions for both problems.

4. **What is the extent of the impact on energy consumption when addressing the problems of UE-BS association and VNF placement disjointly compared to the joint approach?**

The fourth research question is driven by a practice observed in the literature of addressing the problems of UE-BS association and VNF placement disjointly, without assessing the gap in solution quality in comparison to a joint approach. While a disjoint formulation reduces complexity, an optimal solution for the first problem alone can block an overall optimal solution for the complete problem. The results presented in this dissertation show that, despite the joint approach yielding lower energy consumption in the majority of the instances evaluated, the energy consumption gap of the disjoint approach is consistently lower than 1%. The objectives associated with this research question are therefore:

- Formalize both a joint and a disjoint optimization model for the problems of UE-BS association and VNF placement.
- Quantify the energy gap between the joint and disjoint approaches to the problem.

By placing energy consumption at the center of the analysis, this dissertation contributes to the design of sustainable mobile networks, offering insights that can guide both future theoretical advancements and practical orchestration strategies for real-world O-RAN deployments.

The remainder of this dissertation is organized as follows. Chapter 3 formalizes and evaluates the problem of VNF placement with the objective of minimizing energy consumption from a comprehensive perspective. Chapter 4 extends the scope of the evaluation to include the UE-BS association problem, considering both joint and disjoint optimization approaches. Finally, Chapter 5 presents the concluding remarks and discusses potential future research directions for improving energy efficiency in next generation RAN deployments.

CHAPTER 2

Background

This chapter provides the background necessary to contextualize the dissertation. It reviews the evolution of mobile networks, outlines the main characteristics of radio access network architectures, introduces orchestration problems relevant to vRAN operation, and summarizes fundamental concepts in mathematical programming.

2.1 Evolution of mobile networks

The mobile communication network has undergone continuous evolution since its initial proposal in the 1980s [5, 7]. This progression has been driven by the emergence of use cases introduced by the technological and social trends of its time. In practice, this is reflected in the adoption of new types of mobile devices and the development of software applications with increasingly stringent network requirements.

The first generation of mobile communication (1G) was designed to support analog transmission and focused on wireless voice services, with limited support for handoff. In the early 1990s, the second generation of mobile communication (2G) introduced digital transmission over the radio link, improving spectral efficiency and enabling an initial, though highly limited, form of data services. The third generation of mobile communication (3G), deployed in the early 2000s, provided proper wireless internet access for e-mail, picture messaging, mobile TV, and video conferencing services. In 2014, the fourth generation (4G) introduced enhancements in mobile broadband to enable higher data rates, supporting emerging smartphone applications such as high-quality video streaming.

We are now in the fifth generation (5G) of mobile networks. Introduced in 2020, 5G incorporated innovations designed to address three primary use cases [21]. In the Enhanced Mobile Broadband (eMBB) scenario, the goal is to meet the growing demand of users, characterized by higher device density and increased requirements for data rate in mobile applications. The second scenario, known as Ultra Reliable Low Latency Communication (URLLC), supports industrial, medical, and transportation applications that are critically sensitive to connection loss, latency, and transmission errors. Finally,

Massive Machine Type Communications (mMTC) provides support for Internet of Things (IoT) applications, involving large numbers of low-cost devices with energy constraints.

To meet these requirements, 5G incorporates several technological advancements. At the radio layer, Orthogonal Frequency Division Multiplexing (OFDM)-based techniques were extended to provide greater flexibility in waveform parametrization (i.e., numerology), enabling the network to tailor transmission parameters to different propagation conditions and service requirements [50]. The introduction of massive Multiple Input Multiple Output (MIMO) and advancements in beamforming improve spectral efficiency, reducing propagation losses at higher carrier frequencies. Additionally, the use of higher frequency bands, including the millimeter wave spectrum, expands the available bandwidth to support the targeted data rates and device densities. At the architectural level, network densification through small cells and heterogeneous deployments increases spatial reuse, but at the cost of powering more active sites and raising coordination complexity.

2.2 5G network architecture

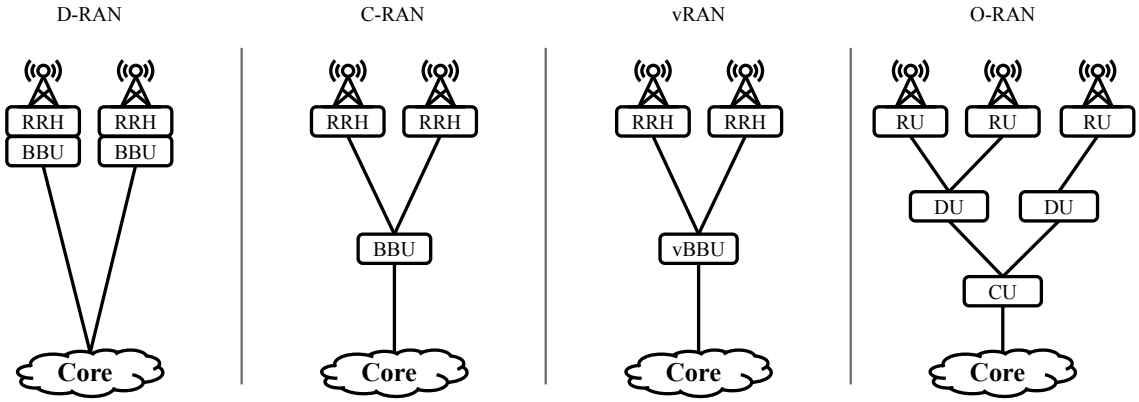


Figure 2.1: RAN architectures. Lines connects components positioned in different geographic locations.

Figure 2.1 illustrates RAN architectures. In traditional Distributed Radio Access Network (D-RAN) deployments, Baseband Units (BBUs) are located at (or near) the radio base station sites, co-located with Remote Radio Heads (RRHs), which are responsible for radio frequency functions. This approach simplifies transport requirements because only higher-layer traffic must traverse the backhaul. However, each site is dimensioned for peak load, even if average utilization is low, resulting in inefficient hardware utilization. As traffic grows and densification increases the number of sites, the inefficiency of isolated processing resources is intensified.

Centralized Radio Access Network (C-RAN) emerged as a response by pooling BBUs from multiple base stations into a shared, centralized location. The idea is to enhance hardware utilization and enable inter site coordination techniques. However, complete centralization introduces stringent transport network requirements. The link connecting radio sites to centralized BBUs must provide high capacity and low latency, which limits real deployments.

Network Function Virtualization (NFV) builds on the idea that protocol stack functions can be implemented as software instances instead of specialized hardware. In the RAN, NFV enables BBU processing to be deployed as VNFs in GPPs platforms, or virtualized Baseband Units (vBBUs), thereby defining the vRAN paradigm. This decoupling of software from the underlying hardware is motivated by improved management flexibility and potential cost reductions through infrastructure sharing. Leveraging the concepts of virtualization, O-RAN architecture introduces standardized, open interfaces for the compatibility of components among different vendors and Radio Access Network Intelligent Controllers (RICs) for network management. Furthermore, to circumvent the constraints in transport network capacity and latency requirements from a fully centralized network, O-RAN offers options to balance pooling gains against deployment feasibility [29].

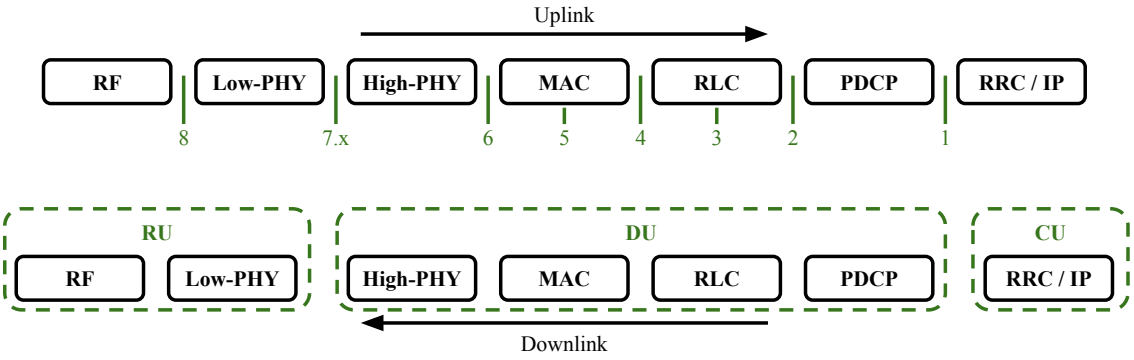


Figure 2.2: 3GPP functional split options and example where options 7.x and 1 are used to define RU, DU and CU.

These options are illustrated in Figure 2.2, splitting the RAN protocol stack into 7 VNFs: Radio Frequency (RF), Low Physical (Low-PHY), High Physical (High-PHY), Media Access Control (MAC), Radio Link Control (RLC), and Packet Data Convergence Protocol (PDCP). Following a Software Defines Network (SDN) paradigm, the higher layer is further separated into Radio Resource Control (RRC) in the control plane and Internet Protocol (IP) in the user plane. Downlink traffic destined for a UE and uplink radio signals received from a UE are processed sequentially by each VNF in the order shown.

By choosing one or two functional split options, the RAN protocol stack can be divided into up to three logical units: the Radio Unit (RU), the Distributed Unit (DU), and the Centralized Unit (CU). In this decomposition, RU is distributed among radio sites and processes VNFs closer to the RF, the DU may aggregate flows from multiple nearby RUs, and the CU may further centralize processing across multiple DUs. Furthermore, these RAN logical units are connected by a crosshaul transport network, composed of three segments. The fronthaul connects the RU to the DU, the midhaul connects the DU to the CU, and the backhaul connects the CU to the core network.

Functional splits, therefore, serve as an architectural configuration parameter that trades off centralization benefits against transport network feasibility. When higher centralization of VNFs is pursued to enable coordination or enhance the efficiency of computing resource usage, the transport network must often carry higher data rate flows and meet tighter latency targets. Conversely, when functions are pushed closer to the radio sites, transport constraints relax, but compute and energy requirements at distributed sites can increase.

2.3 Orchestration problems

Once RAN functions can be virtualized and disaggregated, effective network operation requires solving the associated orchestration problems. For example, selecting the functional split option and crosshaul transport network route for each BS defines the VNF placement problem because these decisions determine where each VNF associated with a given BS is executed and, therefore, which computing resources are used.

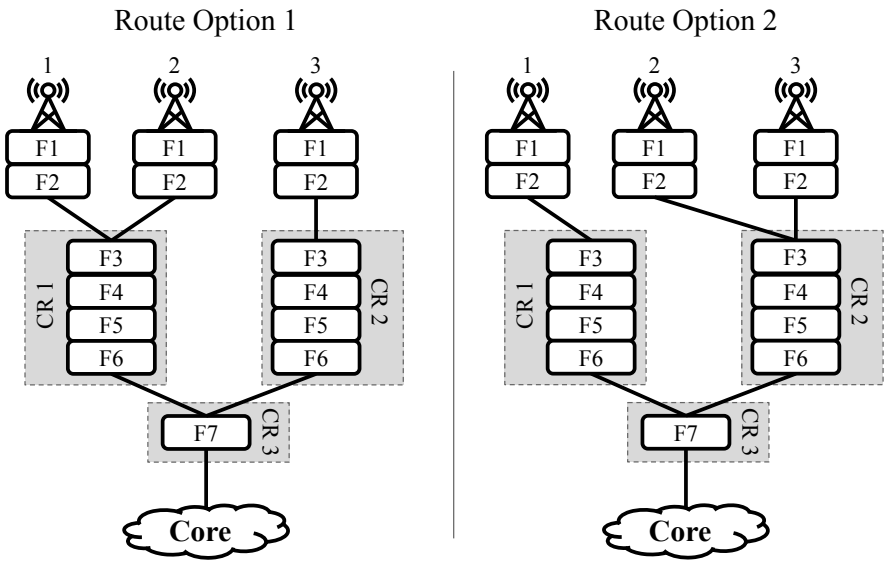


Figure 2.3: Deployment with same functional split and different routes.

Figures 2.3 and 2.4 show practical examples of how routing and split choices jointly determine VNF placement. The VNFs are indexed by number, but they correspond to the same functions described in the previous section. In Figure 2.3, VNFs 3–6 associated with BS 2 are relocated from Computing Resource (CR) 1 to CR 2 by changing the route, while keeping the same functional split options. In Figure 2.4, changing the functional split option, while preserving the routes, relocates VNF 3 for BSs 1 and 2 from CR 1 to local processing, and relocates VNF 6 for all BSs to CR 3.

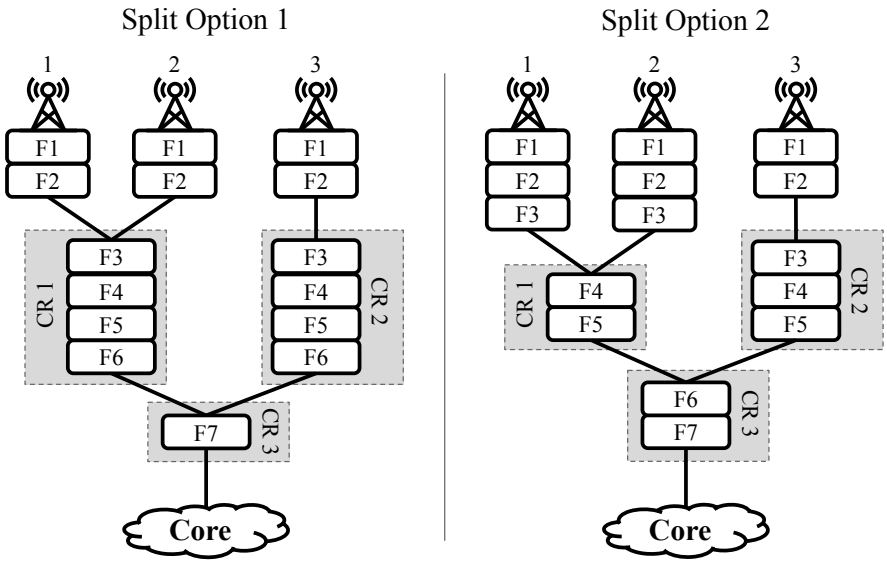


Figure 2.4: Deployment with different functional split options and same routes.

UE-BS association adds an additional layer to the problem. The choice of which BS serves each UE affects radio feasibility (e.g., signal quality and radio resource demand), which in turn determines which sites must be active and the amount of traffic generated by each RU. This traffic directly impacts transport network requirements and consequently influences VNF placement decisions. Overall, this interdependence highlights both the importance of a joint evaluation of these problems and the inherent complexity of a comprehensive analysis of the network.

Existing studies address the VNF placement problem with the goal of promoting centralization toward C-RAN-like deployments [35]. This is achieved by defining a VNF centralization expression to be maximized as the optimization objective. Regarding the UE association problem, the objective is often to maximize user admission under saturation or to maximize the quality of service delivered by the network. However, finding a solution that aims to minimize energy consumption has also been presented as a relevant approach in the literature, especially since the ITU aligned the evolution of the network with the UN's sustainable development goals [22].

Although prior work in the literature addresses these problems from the perspec-

tive of energy consumption, as will be shown in the following chapters, this dissertation aims to address important gaps that remain. An example is the lack of a comprehensive formulation that accounts for every energy consuming component and captures all the flexibility proposed by the next generation of mobile networks. Additionally, this dissertation seeks to avoid a common method observed in the literature, in which a heuristic or machine learning based solution is evaluated without a proper comparison against an exact solution.

2.4 Mathematical programming

The primary method adopted in this dissertation to address the orchestration problems is mathematical programming. In its most general form, a combinatorial optimization problem is defined by (i) a set of decision variables representing controllable choices, (ii) an objective function that quantifies the quality of a solution, and (iii) a set of constraints defined by the nature of the problem that ensures the practical feasibility of the solution.

Mathematical programming models can be classified along two main dimensions. The first concerns the linearity of the objective function and the constraints. Linear models restrict both the objective and all constraints to linear expressions, whereas nonlinear models allow nonlinear terms, representing more complex relationships. The second dimension concerns the domain of the decision variables. Continuous variables take values over a real-valued range (e.g., frequency, energy, spectral efficiency), while discrete variables represent choices among alternatives (e.g., on/off activation, assignment, routing decisions). Combining these dimensions yields the following model classes:

- **Linear Programming (LP)** – linear objective and linear constraints with continuous decision variables.
- **Integer Linear Programming (ILP)** – linear objective and linear constraints with integer (including binary) decision variables.
- **Mixed Integer Linear Programming (MILP)** – linear objective and linear constraints with both continuous and integer decision variables.
- **Nonlinear Programming (NLP)** – nonlinear objective and/or constraints with continuous decision variables.
- **Mixed Integer Nonlinear Programming (MINLP)** – nonlinear objective and/or constraints with both continuous and integer decision variables.

Solving a mathematical programming model consists of finding values for the decision variables that satisfy all constraints (i.e., a feasible solution) and optimize the objective function. A feasible solution that achieves the best objective value among all

feasible solutions is called an optimal solution. LP models, as well as some subcategories of non-LP models, can be solved in polynomial time, with the interior-point method being an example of a polynomial time algorithm for these problem classes. Therefore, when a problem can be modeled as a LP, it typically represents the most favorable scenario from a practical tractability perspective.

In contrast, ILP and MILP formulations increase the complexity of solving the optimization problem. Integrality constraints, which restrict decision variables to integer values, render the feasible region nonconvex, and LP solution methods cannot always be applied. The Branch-and-bound (B&B) technique enables the solution of these classes of problems. In this approach, the algorithm repeatedly solves linear relaxations (i.e., the original problem with integer constraints ignored), branches on fractional variables to create subproblems, and tightens relaxations via additional constraints. This branching behavior creates a tree of LP problems that grows exponentially with the number of decision variables. Consequently, solving general ILP and MILP models is NP-Hard.

In practice, despite the NP-Hardness of ILP and MILP models, B&B can solve many instances within an acceptable time delay. Therefore, the scalability of a given formulation becomes a critical factor. Scalability can often be improved by terminating the search early and returning a suboptimal solution with good enough quality. In this case, solution quality is usually assessed using the optimality gap, which quantifies the distance between the best known integer solution and the optimal solution of the linear relaxed version of the problem.

Finally, the presence of nonconvex expressions (e.g., sine, cosine, or logarithm) further increases the complexity of MINLP models. Such problems require additional features to relax the convexity, which expands the search space and leads to poorer practical scalability when compared to linear counterparts.

2.5 Conclusion

This chapter provided the background needed to support energy-aware orchestration in open and virtualized RANs. It reviewed the evolution toward 5G, introduced the innovations of vRAN and O-RAN architectures, and described the flexibility of functional split options which, together with routing decisions, determine VNF placement. To ground the methodological approach adopted in this dissertation, the chapter also summarized the fundamentals of mathematical programming.

Building on this foundation, the next chapter formalizes the VNF placement problem as a comprehensive MILP and shows how a simple linearization technique enables the solution of considerable-sized instances within practical time limits.

CHAPTER 3

VNF Placement for Optimal Energy Consumption

In this chapter, we formulate the problem of VNF placement with the objective of minimizing energy consumption as a MILP model. The aim is to represent the problem in its most comprehensive form while showing that the application of a simple linearization technique enables the solution of network instances with 50 BSs in less than one second. Furthermore, a heuristic is proposed to improve scalability and achieve reduced energy consumption when compared to a solution that maximizes centralization.

3.1 Related work

The VNF placement problem in vRAN has typically been investigated with the objective of maximizing VNF centralization; that is, increasing the number of VNFs from different RUs processed on the same CR, subject to transport network and processing capacity constraints [14, 15, 35, 37]. However, as highlighted in [41], centralization alone does not necessarily yield energy consumption gains. That work formulates energy efficiency in vRAN as a bi-objective optimization problem, explicitly capturing the tradeoff between centralization and energy consumption. While insightful, the approach does not provide a decisive solution that can be periodically deployed in dynamic network environments.

Several studies have also examined the VNF placement problem in vRAN with the goal of minimizing energy consumption. In [17], the authors propose a heuristic to address the association between CU, DU, and UE, aiming to reduce network energy usage while limiting mobile device handovers. However, their approach assumes fixed associations between DU and RU, overlooking the potential benefits of dynamically adjusting RAN split options. This limitation may lead to inefficient resource utilization, particularly during off-peak periods when a single DU could support a larger number of RUs. GreenRAN [46] formulates the VNF placement problem as a quadratic integer program to minimize energy consumption. Due to its computational complexity, the authors propose

a metaheuristic to solve a relaxed version of the problem, incorporating both VNF migration costs and the division of the RAN into up to three units. Nevertheless, their model simplifies network routing and excludes transport network energy consumption, which can lead to sub-optimal decisions in scenarios where transport costs dominate overall energy usage.

In [34], the authors propose a reinforcement learning algorithm to determine the placement of VNFs across a set of DUs with varying capacities, as well as the scheduling of radio resource blocks for each UE demand. The objective is to reduce energy consumption by switching off idle DUs while meeting heterogeneous latency requirements. However, the model does not account for the energy consumption of the transport network and considers only a single split option, C-RAN, thereby overlooking potential energy gains from more flexible functional splits. Moreover, C-RAN represents the most demanding split option to implement in practice due to its stringent transport network requirements [18], which significantly limits the applicability of the proposed solution.

In [45], the authors address the problem of selecting the RAN function split per network slice to reduce the energy consumption of CRs and the transport network. A heuristic is proposed to obtain solutions within a reasonable time; however, the routing model is oversimplified, restricting associations between RU and DU, which may prevent more energy-efficient configurations. The work in [4] tackles the assignment of VNFs split for each BS across multiple time slots using a deep reinforcement learning approach. Both [45] and [4] consider only midhaul energy consumption, neglecting the potential impact of backhaul and midhaul together. Zorello et al. [36] investigate CU and DU placement under diverse 5G service requirements for bandwidth and latency, proposing a heuristic compared against the optimal solution. Their findings highlight that routing choices strongly influence power consumption. Similarly, Malandrino et al. [33] formulates energy minimization in a C-RAN scenario, though C-RAN represents only one of several possible functional splits. In [9], the joint allocation of cloud, network, and radio resources per UE is explored in a cell-free mobile network to reduce power consumption. The authors assume a single cloud connected to all RUs, which simplifies routing but fails to capture the diversity of real-world topologies. Importantly, none of these works [4, 9, 33, 34, 36, 45] accounts for the energy overhead of VNF migration. This omission limits their applicability in dynamic environments, where even small load variations may trigger migrations whose energy costs outweigh the potential savings.

Table 3.1 provides an overview of the challenges addressed in prior studies, along with the instance sizes and response times reported. To the best of our knowledge, our approach is the first to simultaneously tackle all of these challenges, as reflected in the table. It is worth noting, however, that the NP-hard nature of the problem means

that relying on off-the-shelf solvers to obtain optimal solutions still entails exponential complexity, rendering them impractical for very large topologies. Nevertheless, as shown in Table 3.1, our MILP formulation achieves optimal solutions for reasonably large instances in significantly reduced time, offering a strong baseline against which future non-optimal strategies can be evaluated.

Table 3.1: Summary of related work in vRAN energy consumption

Article	Comprehensiveness			Flexibility		Solution	Instance Size*	Response Time
	RAN	TNet	Mig	Splitting	Routing			
[17]	●	○	●	●	○	Optimal†	50 BSs	10^4s
[46]	●	○	●	●	○	Metaheuristic	25 BSs	10^2s
[34]	●	○	○	○	○	RL	–	–
[45]	●	◐	○	●	○	Optimal†	30 Slices	10^1s
[4]	●	◐	○	●	○	DRL	–	–
[36]	●	●	○	○	●	Optimal†	10 BSs	10^3s
[33]	●	●	○	○	●	Heuristic	–	–
[9]	●	●	○	●	◐	Heuristic	–	–
Our Proposal	●	●	●	●	●	Optimal†	50 BSs	10^{-1}s

Symbols: ● Considered - ◐ Partially considered - ○ Not considered.
RAN and TNet represent the energy consumption of the radio access network and the transport network, respectively.
Mig represents the energy cost of migration.
* Instance size is related to the response time column. We omit this value when solution response time is not available.
† This work also presents a heuristic solution; however, we consider the optimal solution in comparison.

3.2 System model and problem statement

We consider a RAN topology composed of a set $\mathcal{B} = \{b_1, b_2, \ldots, b_{|\mathcal{B}|}\}$ of RUs and a set of general-purpose servers $\mathcal{H} = \{h_1, h_2, \ldots, h_{|\mathcal{H}|}\}$ capable of processing the RAN VNFs. RUs and servers are connected through a set of transport network nodes $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$. We define the set of CRs as $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ where $c_m \subseteq \mathcal{T} \cup \mathcal{H}$ represents a group of co-located switches and servers. We represent the RAN topology as a graph $G = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{H}$ denotes the vertices, where v_0 represents the core network, and $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$ represents the set of edges corresponding to the network links connecting the nodes. This graph structure is flexible enough to enable the representation of any possible RAN topology design, from a traditional distributed RAN topology to infrastructures featuring cloud and edge computing centers.

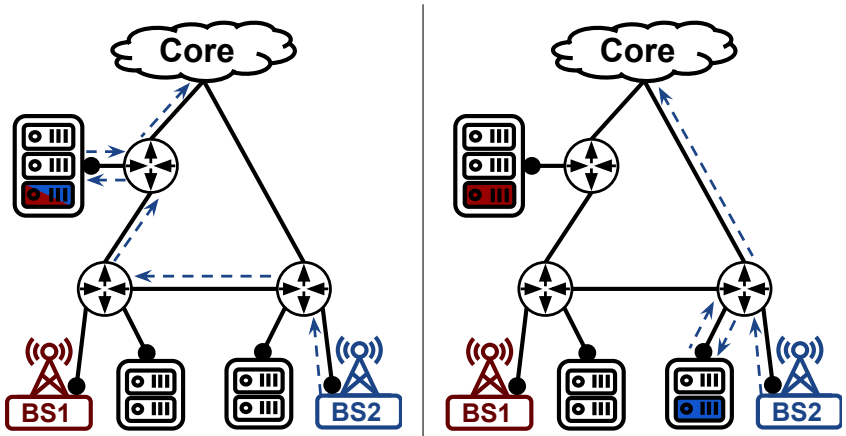
Furthermore, in practical network deployment, the graph-based structure of our system model supports the implementation of our solution as an rApp, i.e., as an application that runs in the Non-Real Time RAN Intelligent Controller in the O-RAN architecture [3]. This is particularly justified as the rApp manages the placement of

virtualized radio functions across multiple base stations, leveraging a broader network perspective [27, 32].

Routing – All data traffic originates (downlink) or terminates (uplink) at the core network v_0. The best route for a BS depends highly on the power efficiency of the transport network and processing equipment, as well as the data load generated at the BS. In a scenario where the cost of turning on a server has the greatest impact on energy consumption, a longer route, as illustrated in Figure 3.1(a), can lead to lower energy costs. Conversely, if the energy cost of the transport network is the dominant factor, a shorter route, as shown in Figure 3.1(b), can be the most beneficial choice. As the size of the topology increases, this trade-off becomes more complex and must be considered in the formulation. Without loss of generality, we consider only the downlink flow. We define $\mathcal{P}_l$ as the set of all possible paths from each RU $b_l \in \mathcal{B}$ to the core network v_0. Each path is represented as an ordered sequence of CRs connecting the RU to the core. To support different functional split options, a path is decomposed into at most three segments that connect CU, DU, and RU: p_{Bh} (backhaul), p_{Mh} (midhaul), and p_{Fh} (fronthaul). As an example, consider a path represented by the sequence of edges (e_1, e_2, e_3, e_4); Different segmentations correspond to different paths in $\mathcal{P}_l$, such as:

$$
\begin{aligned}
p_l &= p_{Bh}(e_1, e_2, e_3, e_4), \\
p_{l+1} &= p_{Bh}(e_1, e_2, e_3) + p_{Mh}(e_4), \\
p_{l+2} &= p_{Bh}(e_1) + p_{Mh}(e_2, e_3) + p_{Fh}(e_4),
\end{aligned}
$$

and so forth. Moreover, we consider a crosshaul transport network, i.e., a link $e_{ij} \in \mathcal{E}$ can serve different segment parts for different BSs.



(a) *Long route.* (b) *Short route.*

Figure 3.1: Example of different routes in the same vRAN topology.

Virtual Network Functions – For each RU $b_l \in \mathcal{B}$, our objective is to determine the best CR to deploy the RU set of RAN VNFs, denoted by $\mathcal{F} = \{f_1, f_2, f_3, f_4, f_5\}$, representing High-PHY, MAC, RLC, PDCP, and RRC functions, respectively. The VNF stack may be partitioned at most twice according to the functional split combination option $D_r \in \mathcal{D}$, which enhances the flexibility of the solution by allowing parts of the VNF stack to be deployed in different CRs. The Low-PHY function is not virtualized and is assumed to always be deployed at the RU. In our formulation, each option $D_r \in \mathcal{D}$, with $D_r \neq D_0$, represents either a single functional split or a combination of two functional split options presented in Section 2.2. The option $D_0 \in \mathcal{D}$ denotes the D-RAN configuration, in which no functional split is performed and all VNFs are placed on a server co-located with the RU.

3.2.1 VNF processing

Let the binary decision variable $x_l^{p,r}$ denote if the path $p \in \mathcal{P}_l$ was chosen for the $b_l \in \mathcal{B}$ using the functional split $D_r \in \mathcal{D}$. However, during the decision, we must ensure that the assigned servers selected by the path $p \in \mathcal{P}_l$ have the required processing capacity. To achieve this, we formulate the computing resource cost for each VNF in terms of Giga Operations Per Second (GOPS) as described in the following.

We adopt local partial minimum mean square error (LP-MMSE) as the precoding method, for which the computational cost was formulated in [9] as follows:

$$\begin{aligned} C_{precoding} = {} & \frac{N_{used}}{T_s \tau_c 10^9}\left(8N_l\tau_p^2 + 8N_l^2\left(\tau_p + Load(b_l)\right)\right) + \\ & \frac{N_{used}\tau_d}{T_s \tau_c 10^9}\left(8N_l Load(b_l)\right) + \frac{N_{used}}{T_s \tau_c 10^9}\left(8N_l Load(b_l)\right) + \\ & \frac{N_{used}}{T_s \tau_c 10^9}\left((4N_l^2 + 4N_l)\tau_p + 8N_l^2 Load(b_l) + \frac{8(N_l^3 - N_l)}{3}\right). \end{aligned} \tag{3-1}$$

In this expression, N_{used} denotes the number of used subcarriers, T_s is related to the OFDM symbol duration, and N_l is the number of antennas at RU $b_l \in \mathcal{B}$. As in [9], we assume without loss of generality that all bandwidth is allocated to downlink transmission (i.e., no uplink data transmission is considered) and that pilot symbols for OFDM channel estimation are transmitted in every coherence block. Accordingly, τ_c denotes the number of samples per coherence block, τ_p is the number of samples used for uplink training, and $\tau_d = \tau_c - \tau_p$ is the number of samples effectively available for downlink data transmission.

Next, based on the model in [8], we estimate the computational costs of OFDM modulation and of mapping modulated symbols onto resource elements (REs), respectively, as:

$$C_{modulation} = 1.3N_l\left(\frac{N_{bits}}{16}\right)^{1.2}, \tag{3-2}$$

$$C_{mapping} = 1.3 Load(b_l) \left(\frac{N_{bits}}{16}\right)^{1.2} \left(\frac{SE_0}{6}\right)^{1.5}, \tag{3-3}$$

where N_{bits} is the number of bits used for data quantization and SE_0 denotes the spectral efficiency of the channel. Therefore, the number of GOPS required for High-PHY layer processing is given by:

$$C_{HighPHY} = C_{precoding} + C_{modulation} + C_{mapping}. \tag{3-4}$$

Channel coding, system control, and data redirection to the core network operations are implemented in the higher (or superior) layers of the RAN VNF stack. Similar to [8], the number of GOPS required by all these superior layers is calculated as follows:

$$C_{SupLayers} = 1.3 Load(b_l) \left(\frac{N_{bits}}{16}\right)^{1.2} \left(\frac{SE_0}{6}\right) + 2.7\sqrt{N_l} \left(\frac{N_{bits}}{16}\right)^{0.2} + 8 Load(b_l) \left(\frac{SE_0}{6}\right). \tag{3-5}$$

To determine the processing required by each VNF individually, given the aggregated load of the superior layers $C_{SupLayers}$ from (3-5), we utilize proportions based on the CPU utilization profiles observed in the OpenAirInterface (OAI) implementation [26]. Finally, the total processing load γ_w for a given server $h_w \in \mathcal{H}$ is calculated as follows:

$$\begin{aligned} \gamma_w = \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} \Big[& F_1(h_w, D_r, p) C_{HighPHY} + F_2(h_w, D_r, p) \Big(0.4\ C_{SupLayers}\Big) \\ & + F_3(h_w, D_r, p) \Big(0.028\ C_{SupLayers}\Big) + F_4(h_w, D_r, p) \Big(0.286\ C_{SupLayers}\Big) \\ & + F_5(h_w, D_r, p) \Big(0.286\ C_{SupLayers}\Big) \Big], \end{aligned} \tag{3-6}$$

where $F_n(h_w, D_r, p)$, based on the input data, returns 1 when server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ according to functional split $D_r \in \mathcal{D}$ and route $p \in \mathcal{P}_l$; otherwise, it returns 0.

3.2.2 Optimization model

We formulate the problem of RAN VNF placement to minimize the energy consumption as a MILP model. Moreover, we decompose the energy consumption into three main components, which are detailed below.

vRAN energy consumption – The energy consumed by the general purpose servers processing the VNFs of all RUs $b_l \in \mathcal{B}$ characterizes the vRAN energy consumption. We use a traditional energy model [11] to estimate this energy consumption, in which the total

energy consumed by a server $h_w \in \mathcal{H}$ over the period T includes a static power component P_w^{idle} consumed whether the server is turned on, and a dynamic load-dependent power consumption $P_w^{busy} - P_w^{idle}$, defined as:

$$E_{vRAN} = \sum_{h_w \in \mathcal{H}} T\left[\psi_w^{on} P_w^{idle} + \left(\gamma_w / C_w^{cap}\right)\left(P_w^{busy} - P_w^{idle}\right)\right], \tag{3-7}$$

where γ_w is the total load assigned to server $h_w \in \mathcal{H}$ as calculated in (3-6), C_w^{cap} represents the number of GOPS the server $h_w \in \mathcal{H}$ can perform. ψ_w^{on} is a ceiling function that ensures that a server $h_w \in \mathcal{H}$ is counted as active if, and only if, at least one VNF of any RU is assigned to it. This ceiling function is defined as follows:

$$\psi_w^{on} = \left\lceil \sum_{f_n \in \mathcal{F}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} u_w^p M(h_w, f_n, b_l, D_r)}{|\mathcal{F}|\,|\mathcal{B}|} \right\rceil, \tag{3-8}$$

where $u_w^p \in \{0,1\}$ is based on the input data, indicating whether a server $h_w \in \mathcal{H}$ is part or not of the route $p \in \mathcal{P}_l$. The mapping function $M(h_w, f_n, b_l, D_r)$ returns 1 when the server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ from RU $b_l \in \mathcal{B}$ according to functional split $D_r \in \mathcal{D}$.

Transport network energy consumption – We consider an optical transport network dedicated to the RAN infrastructure, where the energy consumption comes from (i) Ethernet switches, allowing higher routing flexibility by enabling packet aggregation, and (ii) Dense Wavelength Division Multiplexing (DWDM) pluggable transceivers that can be directly installed in switches and in the most recent 5G RUs [43, 12]. Each link $e_{ij} \in \mathcal{E}$ is characterized by its data transmission capacity $R_{e_{ij}}^{tr}$ and power consumption $P_{e_{ij}}^{tr}$ of the transceivers at each end of the link. For each topology node $v_k \in \mathcal{V}$, the function $S(v_k) \in \{0,1\}$ indicates whether v_k is a packet switch, while $P_{v_k}^s$ represents the power consumed by each switch port. Finally, The transport network energy consumption is defined as follows:

$$E_{TNet} = \sum_{e_{ij} \in \mathcal{E}} \left[T \frac{\gamma_{e_{ij}}}{R_{e_{ij}}^{tr}} \left(2P_{e_{ij}}^{tr} + S(v_j) P_{v_j}^s + S(v_i) P_{v_i}^s\right)\right]. \tag{3-9}$$

$\gamma_{e_{ij}}$ in (3-9) represents the total throughput over link e_{ij}, which is defined as:

$$\gamma_{e_{ij}} = \sum_{D_r \in \mathcal{D}} \sum_{b_l \in B} \sum_{p \in \mathcal{P}_l} x_l^{p,r} R^l \left(y_{e_{ij}}^{p_{Bh}} \alpha_{Bh}^{r,l} + y_{e_{ij}}^{p_{Mh}} \alpha_{Mh}^{r,l} + y_{e_{ij}}^{p_{Fh}} \alpha_{Fh}^{r,l}\right), \tag{3-10}$$

where the data throughput $R^l = Load(b_l) R^{\text{dev}}$ generated by RU $b_l \in \mathcal{B}$ is a dynamic parameter that fluctuates over time, and can be estimated by the number of devices connected to the RU at a given instant and the mean throughput generated by these devices. $y_{e_{ij}}^{p_{Bh}}$, $y_{e_{ij}}^{p_{Mh}}$,

and $y_{e_{ij}}^{p_{Fh}}$ indicate if the link e_{ij} is part of the backhaul, midhaul, or fronthaul, respectively, for the path $p \in \mathcal{P}_l$. For each RU $b_l \in \mathcal{B}$, the chosen functional split $D_r \in \mathcal{D}$ increases the data rate required at the backhaul, midhaul, and fronthaul by the factors $\alpha_{Bh}^{r,l}$, $\alpha_{Mh}^{r,l}$, and $\alpha_{Fh}^{r,l}$, respectively. As the solution is expected to be effective during the time period T, we consider the estimated number of active links $\gamma_{e_{ij}} / R_{e_{ij}}^{tr}$ as a possible non-integral real value to account for the deactivation of the link during periods of inactivity.

VNF migration energy consumption – To ensure that the energy cost overhead due to migration does not outweigh the energy gains in the new solution, we use a linear approximation model based on empirical data, similar to [17, 46]. As shown in Section 3.3, VNF migration is the component with the lowest impact on total energy consumption. Therefore, despite the possible inaccuracies inherent in linear approximations, this approach has a limited impact on the quality of the solution and complies with our objective of preserving the linearity of the formulation. Considering that each VNF is processed individually in its own Virtual Machine (VM), allowing for flexible function placement, we estimate the energy cost of migrating a specific VNF $f_n \in \mathcal{F}$ as $E_{f_n} = aV_{fn} + b$, where V_{fn} is the memory volume of the VM hosting the VNF $f_n \in \mathcal{F}$, and coefficients a and b are derived from experimental observations [30], which maps the data traffic of VM migration to energy consumption. Lastly, we define the total VNF migration energy consumption by the following equation:

$$E_{Mig} = \sum_{h_w \in \mathcal{H}} \sum_{f_n \in \mathcal{F}} \sum_{D_r \in \mathcal{D}} \sum_{b_l \in \mathcal{B}} \sum_{p \in P_l} \left[1 - N(h_w, f_n, b_l)\right] x_l^{p,r} u_w^p M(h_w, f_n, b_l, D_r) E_{f_n}, \tag{3-11}$$

where $N(h_w, f_n, b_l)$ is defined over the input data, resulting in 1 when server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ from RU $b_l \in \mathcal{B}$ for the previous VNF deployment. Otherwise, it returns 0.

The objective is to minimize the total energy consumption needed to process the vRAN VNFs. This total energy consumption is impacted by the hardware processing the VNFs, the transport network connecting the CRs, and the migration of VNFs given a previous deployment. Therefore, we define the objective function as follows:

$$\text{minimize} \quad E_{vRAN} + E_{TNet} + E_{Mig} \tag{3-12a}$$

subject to

$$\sum_{c_m \in \mathcal{C}} \sum_{f_n \in \mathcal{F}} \left(\sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r) - \psi_{m,n}^{single} \right) \geq \rho^c, \tag{3-12b}$$

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} = 1, \quad \forall b_l \in \mathcal{B}, \tag{3-12c}$$

$$\gamma_{e_{ij}} \leq e_{ij}^{Cap}, \quad \forall e_{ij} \in \mathcal{E}, \tag{3-12d}$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Bh}} e_{ij}^{L} \leq \beta_{Bh}^{r}, \ \ \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{3-12e}$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Mh}} e_{ij}^{L} \leq \beta_{Mh}^{r}, \ \ \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{3-12f}$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Fh}} e_{ij}^{L} \leq \beta_{Fh}^{r}, \ \ \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{3-12g}$$

$$\gamma_w \leq C_w^{cap}, \quad \forall h_w \in \mathcal{H}, \tag{3-12h}$$

$$x_l^{p,r} \in \{0,1\}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l. \tag{3-12i}$$

We define centralization as the number of VNFs from different RUs being processed in the same CR, i.e., in servers at the same geographical location. The constraint in (3-12b) ensures a lower bound centralization ρ^c of VNFs, where $v_{m,w}^{p}$, defined over the input data, indicates whether a server $h_w \in \mathcal{H}$ is associated with CR $c_m \in \mathcal{C}$ and is part of route $p \in \mathcal{P}_l$. The term ψ_m^{single} is an expression and assures that at least two RUs $b_l \in \mathcal{B}$ must have the same type of VNF $f_n \in \mathcal{F}$ allocated to the same CR $c_m \in \mathcal{C}$ to count as centralization, which is formulated as:

$$\psi_{m,n}^{single} = \left\lceil \sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r)}{|\mathcal{F}|\,|\mathcal{B}|} \right\rceil. \tag{3-13}$$

For each RU $b_l \in \mathcal{B}$, only one combination of route $p \in \mathcal{P}_l$ and functional split $D_r \in \mathcal{D}$ must be assigned, as represented in the constraint (3-12c).

Each link $e_{ij} \in \mathcal{E}$ has a maximum data rate capacity e_{ij}^{Cap} defined by the number of fibers and data rate of transceivers composing it. This maximum capacity must not be exceeded, as represented in the constraint (3-12d). Furthermore, scenarios where the transport network is shared with other services can be considered by extending constraint (3-12d).

Depending on the functional split $D_r \in \mathcal{D}$ chosen, different latencies must be granted at *fronthaul* (β_{Fh}^{r}), *midhaul* (β_{Mh}^{r}), and *backhaul* (β_{Bh}^{r}) of path $p \in \mathcal{P}_l$. Since each link $e_{ij} \in \mathcal{E}$ incurs in delay e_{ij}^{L} according to its capacity, distance between nodes, number of hops, and packet queue size, the chosen path $p \in \mathcal{P}_l$ must ensure the latency required by functional split $D_r \in \mathcal{D}$, as defined in the constraints (3-12e) – (3-12g).

The VNFs assigned to a given server $h_w \in \mathcal{H}$ must not exceed its maximum

processing capacity C_w^{cap}, as defined in the constraint (3-12h). Finally, the constraint in (3-12i) specifies that the decision variable is binary.

3.2.3 Linearization

The problem stated in (3-12) represents an integer program formulation, which is known to be an NP-hard problem, as demonstrated by the authors of [35]. However, the ceiling function in (3-8) and (3-13) renders those constraints discontinuous. This approach makes the model nonlinear, i.e., unsupported by MILP solvers and dependent on inefficient solutions. Therefore, we linearize these constraints by introducing two new integer decision variables, y_w and y_m, for each server, $h_w \in \mathcal{H}$ and CR $c_m \in \mathcal{C}$. Regarding the former and considering the following relaxation of ψ_w^{on}:

$$\psi_w^{on'} = \sum_{f_n \in \mathcal{F}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} u_w^p M(h_w, f_n, b_l, D_r)}{|\mathcal{F}|\ |\mathcal{B}|}, \tag{3-14}$$

we are able to mimic the behavior of the ceil function by adding the following constraints:

$$y_w \geq \psi_w^{on'}, \tag{3-15}$$

$$y_w \leq \psi_w^{on'} + 1 - \epsilon, \tag{3-16}$$

where $\epsilon < 1/(|\mathcal{F}|\ |\mathcal{B}|)$ is an integrity component, granting that y_w remains equivalent to ψ_w^{on}, when $\psi_w^{on'}$ results in an integer value. Similarly, for y_m, we define $\psi_m^{single'}$ as the relaxation of ψ_m^{single}:

$$\psi_m^{single'} = \sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{x_l^{p,r} v_{m,w}^p M(h_w, f_n, b_l, D_r)}{|\mathcal{F}||\mathcal{B}|}, \tag{3-17}$$

and formulate the following constraints:

$$y_{m,n} \geq \psi_{m,n}^{single'}, \tag{3-18}$$

$$y_{m,n} \leq \psi_{m,n}^{single'} + 1 - \epsilon. \tag{3-19}$$

To conclude, after replacing ψ_w^{on} with y_w in (3-7) and ψ_m^{single} with y_m in (3-12b), we end up with:

minimize $\quad E_{vRAN} + E_{TNet} + E_{Mig}$

subject to

$$\sum_{c_m \in \mathcal{C}} \sum_{f_n \in \mathcal{F}} \left(\sum_{h_w \in \mathcal{H}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} v_{m,w}^{p} M(h_w, f_n, b_l, D_r) - y_m \right) \geq \rho^c,$$

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x_l^{p,r} = 1, \quad \forall b_l \in \mathcal{B},$$

$$\gamma_{e_{ij}} \leq e_{ij}^{Cap}, \quad \forall e_{ij} \in \mathcal{E},$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Bh}} e_{ij}^{L} \leq \beta_{Bh}^{r}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l,$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Mh}} e_{ij}^{L} \leq \beta_{Mh}^{r}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l,$$

$$\sum_{e_{ij} \in \mathcal{E}} x_l^{p,r} y_{e_{ij}}^{p_{Fh}} e_{ij}^{L} \leq \beta_{Fh}^{r}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l,$$

$$\gamma_w \leq C_w^{cap}, \quad \forall h_w \in \mathcal{H},$$

$$y_w \geq \psi_w^{on'}, \quad \forall h_w \in \mathcal{H},$$

$$y_w \leq \psi_w^{on'} + 1 - \epsilon, \quad \forall h_w \in \mathcal{H},$$

$$y_{m,n} \geq \psi_{m,n}^{single'}, \quad \forall c_m \in \mathcal{C}, f_n \in \mathcal{F},$$

$$y_{m,n} \leq \psi_{m,n}^{single'} + 1 - \epsilon, \quad \forall c_m \in \mathcal{C}, f_n \in \mathcal{F},$$

$$x_l^{p,r} \in \{0,1\}, \quad \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l,$$

$$y_w, y_{m,n} \in \mathbb{Z}, \quad \forall c_m \in \mathcal{C}, h_w \in \mathcal{H}, f_n \in \mathcal{F},$$

which represents a more efficient MILP formulation, as illustrated by the results presented in Section 3.3.

Table 3.2 summarizes all the decision variables, sets, and data parameters used throughout the formulation.

3.2.4 Heuristic

Solving an MILP formulation is known to be NP-Hard, presenting non-polynomial complexity to solve. To obtain a faster solution for instances that cannot be solved within an acceptable time using the MILP model, we propose a heuristic solution.

Algorithms 1 and 2 present the heuristic to assign a route between each BS in the topology and the core network, as well as a functional split for the association. As input, we consider the sets L_n of nodes with distance n (in number of hops) from the core, the RAN topology graph $\mathcal{G}$, the set of functional splits $\mathcal{D}^*$ excluding the no-split option D_0, the set of paths $\mathcal{P}$ from every BS until the core, and the associations $\mathcal{S}^{-1}$ currently deployed in the network. Additionally, Algorithm 2 receives the partial solution from Algorithm 1.

Table 3.2: Sets, Input Data, Decision Variables, and Expressions

	Notation	Description
Sets	$\mathcal{B}$	Set of RUs
	$\mathcal{H}$	Set of general-purpose servers
	$\mathcal{T}$	Set of transport network nodes
	$\mathcal{C}$	Set of CRs
	$\mathcal{G}$	Topology graph where $G = (\mathcal{V}, \mathcal{E})$
	$\mathcal{V}$	Set of topology nodes where $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{H}$
	$\mathcal{E}$	Set of topology links where $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$
	$\mathcal{P}_l$	Set of all paths from each RU $b_l \in \mathcal{B}$ to the core network v_0
	$\mathcal{F}$	Set of VNFs
	$\mathcal{D}$	Set of functional splits
Input Data	ρ^c	VNF centralization lower bound
	C_w^{cap}	Processing capacity of server $h_w \in \mathcal{H}$
	P_w^{busy}	Load-dependent power consumption of server $h_w \in \mathcal{H}$
	P_w^{idle}	Static power consumption of server $h_w \in \mathcal{H}$
	E_{f_n}	Memory volume of VM hosting the VNF $f_n \in \mathcal{F}$
	e_{ij}^{Cap}	Maximum link data rate
	e_{ij}^{L}	End-to-end link latency
	$R_{e_{ij}}^{tr}$	Transceiver transmission capacity
	$P_{e_{ij}}^{tr}$	Transceiver power consumption
	$P_{v_i}^{s}$	Ethernet switch port power consumption
	T	Expected solution deployment duration
	T_s	OFDM symbol duration
	N_{used}	Number of sub-carriers used
	N_{bits}	Number of bits used for data quantization
	N_l	Number of antennas in RU $b_l \in \mathcal{B}$
	u_w^p	Indicates whether a server $h_w \in \mathcal{H}$ is part of the path $p \in \mathcal{P}_l$
	$v_{m,w}^p$	Indicates whether a server $h_w \in \mathcal{H}$, part of the path $p \in \mathcal{P}_l$, is associated with CR $c_m \in \mathcal{C}$
	SE_0	Channel spectral efficiency
	τ_c	Number of samples per coherence block
	τ_p	Number of received samples during training phase
	τ_d	Number of received samples during downlink data transmission
Decision Vars. and Exprs.	$x_l^{p,r}$	Binary decision variable indicating that path $p \in \mathcal{P}_l$ was chosen for RU $b_l \in \mathcal{B}$ using functional split $D_r \in \mathcal{D}$
	y_w, y_m	Integer decision variables that linearize ψ_w^{on} and $\psi_{m,n}^{single}$
	γ_w	Processing load for server $h_w \in \mathcal{H}$
	$\gamma_{e_{i,j}}$	Total throughput over link $e_{i,j}$
	ψ_w^{on}	Indicate if server $h_w \in \mathcal{H}$ is assigned to the processing of any VNF and needs to be activated
	$\psi_{m,n}^{single}$	Indicate if at least two different RUs deploy the same VNF $f_n \in \mathcal{F}$ in CR $c_m \in \mathcal{C}$

Algorithm 1 VNF Placement For The First Level

Input : $L_n, \mathcal{G}, \mathcal{D}^*, \mathcal{P}, \mathcal{S}^{-1}$.
Output: Partial set of associations between BS, route and functional split $\mathcal{S}$.

$\mathcal{S} \leftarrow \varnothing$
Sort set L_n, for each n, by BS load
for $v_i \in L_1$ **do**
    if $v_i \notin \mathcal{B}$ **then**
        continue
    $p \leftarrow$ No-split route for v_i with the most power-efficient available server
    if *Migrating to* (v_i, p, D_0) *is worthwhile* **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, D_0)$

Algorithm 2 VNF Placement For Level 2 and Beyond

Input : $L_n, \mathcal{G}, \mathcal{D}^*, \mathcal{P}, \mathcal{S}^{-1}, \mathcal{S}$.
Output: Complete set of associations between BS, route and functional split $\mathcal{S}$.

Sort set L_n, for each n, by BS load
for $n > 1$ **do**
    for $v_i \in L_n$ **do**
        if $v_i \notin \mathcal{B}$ **then**
            continue
        $\mathcal{C}_{cands.} \leftarrow$ CRs with turned on server
        Reverse sort $\mathcal{C}_{cands.}$ by number of associated BSs
        $p_0 \leftarrow$ No-split route for v_i with the most power-efficient available server
        candAssociation $\leftarrow (v_i, p_o, D_0)$
        feasible $\leftarrow$ False
        for $c_m \in \mathcal{C}_{cands.}$ **do**
            for *route* p *that contains candidate CR* c_m **do**
                for $O_j \in \mathcal{D}^*$ **do**
                    if (v_i, p, O_j) *is unfeasible* **then**
                        continue
                    feasible $\leftarrow$ True
                    if (v_i, p, O_j) *consumes less energy than candAssociation* **then**
                        candAssociation $\leftarrow (v_i, p, O_j)$
        if *feasible* **then**
            if *Migrating to* (v_i, p, O_j) *is worthwhile* **then**
                $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, O_j)$
        else
            $c_m \leftarrow$ a CR with available server to turn on, prioritizing already active CRs
            `Assuming that the BS load in` v_i `is equivalent to the load of all` v_x `not yet associated in` L_n
            if $\exists p \in \mathcal{P}$ *and* $\exists O_r \in \mathcal{D}^*$ *such that* (v_i, p, O_r) *is feasible and consumes less energy than performing no-split* **then**
                if *Migrating to* (v_i, p, O_r) *is worthwhile* **then**
                    $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, O_r)$
            else
                $p \leftarrow$ No-split route for v_i with the most power-efficient available server
                $\mathcal{S} \leftarrow \mathcal{S} \cup (v_i, p, D_0)$
            Try to apply no-split option for the remaining nodes in distance n
            Iterate loop in step 10

Algorithm 1 creates a partial solution by assigning a route and functional split to every BS in the first hop from the core, if any, that cannot centralize their VNFs. In step 2 we sort the nodes in each level n (i.e., nodes with the same distance from the core) by BS load, so we associate the nodes with the least load first. In steps 3 to 8 a route with a no-split option is associated for BSs at level 1. To address the energy cost of VNF migration, in step 7 we evaluate whether the energy savings from the new association compensate for the energy overhead incurred by migrating from the previous solution.

For levels 2 and beyond, in Algorithm 2, we evaluate whether performing a split and using an active CR in the upper levels is more efficient than turning on a local server and opting for a no-split option, as detailed in steps 14 to 29. If none of the evaluated associations are feasible, we determine whether a new server should be activated in the upper levels or if the no-split option should be assigned to the remaining BSs in the current level, as addressed in steps 30 to 40. This decision is based on the assumption that the total load in the remaining BSs at this level is generated by the current BS being evaluated. This assumption implies that the greater the difference in efficiency between CRs and transport network links, the farther from optimal the heuristic solution is expected to be.

Complexity analysis – The proposed heuristic algorithm can find a satisfactory solution in polynomial time, as shown in Section 3.3. In Algorithm 1, step 2 sorts each hop-level set L_n by BS load, which costs $\sum_n |L_n| \log |L_n| \leq O(|\mathcal{V}| \log |\mathcal{V}|)$. This upper limit stems from the definition of L_n, where $\sum_n L_n = |\mathcal{V}| - 1$. steps 3 to 8 iterate over $L_1 < |\mathcal{V}|$ once, therefore, the overall complexity of Alg. 1 is $O(|\mathcal{V}| \log |\mathcal{V}|)$, dominated by the sorting step. For algorithm 2, step 9 performs the same sort as previously described. Steps 10 and 11 iterate over each CRs once, while step 19, in the worst case, handles all CRs again. Step 20 iterates through the routes for a given CR. To limit the number of paths and reduce the complexity of the heuristic, we use only the k-shortest paths, which gives k paths of length ℓ where each is further divided into 1 route without splits, $\ell - 1$ routes with a single split, and $\sum_{j=1}^{\ell-2} j = O(\ell^2)$ routes with two splits. In the worst case, all paths have $\ell = |\mathcal{V}| - 1$, limiting the number of iterations in step 20 to $O(k \cdot |\mathcal{V}|^3)$. Finally, steps 10 to 26 have complexity $O(k \cdot |\mathcal{V}|^5)$. The evaluation in step 33 is done similarly, resulting in the same complexity.

3.3 Evaluation

In this section, we evaluate the proposed formulation for the problem of vRAN VNF placement for energy efficiency. We introduce the method and parameters used in the evaluation, followed by a discussion of the results.

3.3.1 Method and parameters setup

Topologies – We consider the two classes of RAN topologies commonly deployed by network operators [23, 35, 38, 52]: T1, in which the nodes are connected in ring structures, and T2, in which the transport network follows a hierarchical tree structure. Both topology classes are represented by instances comprising 50 nodes. In T2, each node's network capacity and computing resources are defined according to its distance (in hops) from the core network. Since links closer to the core may be subject to higher load, they are defined with larger capacity. In T1, we consider two configurations: a High Capacity (HC) scenario, where every link utilizes 100G pluggable transceivers, and a Low Capacity (LC) scenario, with links composed of 10G pluggable transceivers. In a previous work [41], we assessed the impact of heterogeneous hardware topologies on energy consumption and observed that servers with higher energy efficiency are prioritized for activation. To preserve this behavior within our problem instances, we randomly assign the idle power consumption P_w^{idle} from 20% to 25% of the busy power consumption P_w^{busy} for each server $h_w \in \mathcal{H}$.

Paths – The number of paths from each RU to the core network can grow exponentially with the number of nodes in the topology, hindering the ability to solve instances of larger topologies. In this work, all simple paths in $\mathcal{P}_l$ are considered, preserving the full search space. While restricting the evaluation to a subset (e.g., the k-shortest paths) would reduce computational complexity, it would do so at the cost of potentially sacrificing optimality. Furthermore, the length of the path is measured by the number of hops.

Latency – The latency experienced in a given route from the BS to the core network is composed of four components: (i) optical propagation in the fiber, (ii) processing in packet switches, (iii) transmission delay, and (iv) queue delay. Once a solution is expected to remain active during a period of time T, latency constraints must be granted most of the time. To estimate an upper bound for total latency, as in [48], we consider 5 μs/Km of propagation delay, 5 μs of switch processing, a packet of 12368 bits, and an average queue size of two packets. The worst-case latency with these parameters is 26 μs, which does not preclude any evaluated functional splits.

Load Variation – For BS load, we use synthetic data generated by the Markov Chain based algorithm proposed in [42]. Since publicly available real-world data are limited to small fragments, we generated synthetic data to model realistic demand variations over time. This allows us to perform a comprehensive evaluation of our model over a 72-hour period, starting on Sunday and ending on Tuesday. The data comprises the hourly state of the network. Therefore, we always solve for a time $T = 1$ hour. However,

due to the dynamic nature of the network, T can be adjusted dynamically to suit other scenarios. For example, it is possible to actively monitor the state of a deployed network and dynamically invoke the model whenever a new solution is required.

UE Profiles – We consider four device usage profiles to analyze how different device network requirements impact the energy consumption of the vRAN. Profile P1 is based on devices requiring URLLC service and 1 Gigabits Per Second (Gbps) of data throughput. Profiles P12, P24, and P53 are all based on eMBB service, with throughput requirements of 12 Gbps, 24 Gbps, and 53 Gbps, respectively.

Functional Splits – Due to data unavailability, we evaluate only 3GPP functional splits 6 and 7.2 with a delay requirement of 250 µs. However, the model is formulated to support additional functional split options, as more data becomes available in the future. When transmitting an eMBB packet (1500 Bytes), the transport network's required bandwidth increases by 1.001 for split 6 and 7.175 for split 7.2 [44]. The bandwidth factors for URLLC packets (128 Bytes) are 1.070 for split 6 and 7.634 for split 7.2.

A transport network based on pluggable DWDM transceivers and Ethernet packet switches is considered, as in [12]. Radio and CR parameters are mainly extracted from [9]. We estimate the VM memory footprint for each VNF based on the memory values provided in [46] and the CPU core usage for each VNF presented in [35]. The parameters utilized in the model evaluation are summarized in Table 3.3.

To perform the evaluation, the proposed model was implemented in Python using the docplex library. Next, CPLEX is used to solve each instance of the problem. An instance is represented by a vRAN topology, its load in a given time, and the network usage profile of the devices. We investigate the impact of those three elements in the solution. The experiments were executed on an Intel i7-12700. Considering 50 BSs, instances of topology with low capacity links took a mean time of 300 ms, while instances of topology T1 with high capacity were solved in a mean time of 700 ms.

3.3.2 Results

Scalability – To evaluate the scalability of the solution, we consider scenarios where all CRs have a co-located RU, and the UEs present a profile P12. For each topology size, i.e., number of CRs, we evaluate 5 instances with different BS loads. The solver is configured with a time limit of 30 minutes and a relative Mixed Integer Programming (MIP) gap tolerance of 10^{-5} (slightly lower than the default value of 10^{-4}), meaning that the solver may stop early and return a solution that can be up to 0.001% worse than a possible optimal solution. Instances not solved within the time limit are not included

Table 3.3: Evaluation parameters

Parameter	Value
$\lvert\mathcal{B}\rvert$	50
ρ^c	0
C_w^{GOPS}	180
P_w^{busy}	94.8 W
P_w^{idle}	20-25% of P_w^{busy}
E_{f_s}	{1795, 242.08, 172.92, 410, 410} MB
e_{ij}^{Cap}	{100, 200, 400, 800, 1000} Gbps
e_{ij}^{L}	$(10^{-1}, 10^{-4})$ ms
$R_{e_{ij}}^{tr}$	{1, 10, 100} Gbps
$P_{e_{ij}}^{tr}$	{1.0, 2.0, 4.5} W
$P_{v_i}^{s}$	{1.0, 4.2, 14.0} W
T	3600 s
T_s	71.4 µs
N_{used}	1200
N_{bits}	12
N_l	4
SE_0	1.0 bits/s/Hz
τ_c, τ_p	192, 8

in the statistics. Figure 3.2 shows that the solution time grows exponentially with the number of nodes in the topology. The data rate capacity of the transport network also impacts the time it takes to solve an instance. The data indicates that the complexity of an instance increases with the number of nodes in a topology and with the increase of the capacity in the transport network. In addition, we observe that different BS loads may cause greater variation in the solution time, especially as the complexity of the instance increases, resulting in more instances that are not solved within the time limit, as shown by the bar charts which illustrate the stop criteria of the MIP Solver for every evaluated instance. Nonetheless, we obtain a solution for instances with 450 RUs within a time range of 2 to 16 minutes, which is nearly one order of magnitude larger than the largest instance reported as optimally solved in the literature.

To further evaluate the impact of link capacity on the absolute solution time, we present the results in Figure 3.3 for a topology with 100 CRs and a varying transport network. First, we evaluate a transport network composed entirely of 1 Gbps, 10 Gbps, and 100 Gbps transceivers. Next, in the scenario labeled “Hier.”, we assess a transport network where the data rate capacity of the transceivers decreases with the distance from

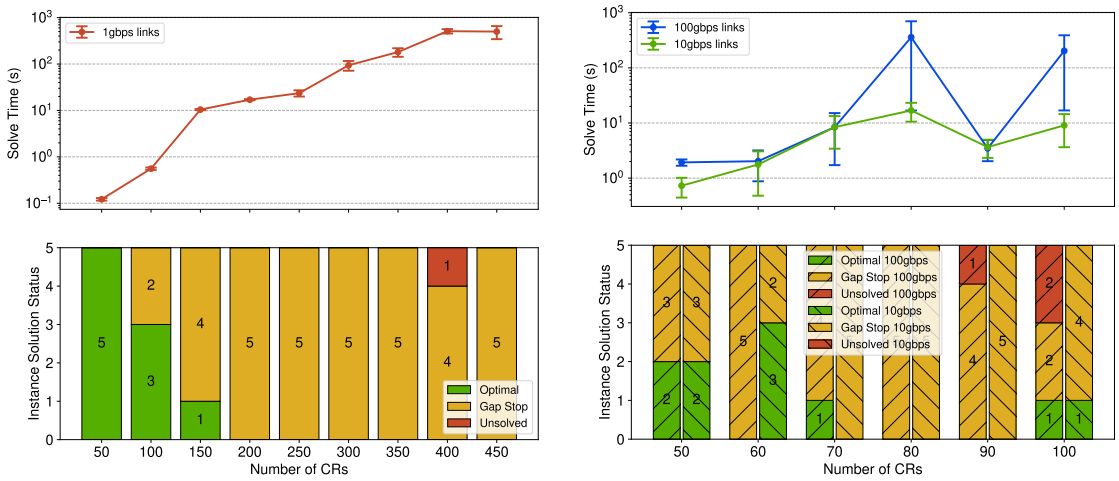


Figure 3.2: Solution time scalability for topologies with different transport network link capacities and increasing topology size. The x-axis, representing topology size, is common to all vertically aligned sub-figures.

the core. The "Inv. Hier." scenario follows the same logic; however, the transceiver data rate capacity increases with the distance from the core. Finally, in the "Random" scenario, transceivers are randomly assigned to each link. Unlike in the previous evaluation, we now maintain the same overall capacity of the transport network by reducing the number of links between CRs as we increase the data rate capacity of the transceivers. The results indicate that a transport network with lower capacity transceivers in the links closer to the core tends to present lower solution times.

The MILP formulation in Section 3.2 introduces a set of decision variables, the number of which grows combinatorially with the size of the topology. The core binary variable $x_l^{p,r}$ is defined for every RU $b_l \in \mathcal{B}$, every candidate path $p \in \mathcal{P}_l$, and every functional split $D_r \in \mathcal{D}$. Consequently, the total number of such variables is $|\mathcal{B}||\mathcal{P}_l||\mathcal{D}|$. Since the number of functional split options is bounded, $|\mathcal{D}|$ can be treated as a constant. Therefore, the growth in the number of decision variables is primarily driven by $|\mathcal{B}|$ and $|\mathcal{P}_l|$. When all simple paths are considered, the number of paths can grow factorially in dense topologies, yielding a worst-case bound $|\mathcal{P}_l| = \mathcal{O}(|\mathcal{B}|!)$. Under this assumption, the number of binary variables scales as

$$|\mathcal{B}| \cdot \mathcal{O}(|\mathcal{B}|!) = \mathcal{O}(|\mathcal{B}|!),$$

indicating that the search space is dominated by the number of candidate paths.

Comparison of different solutions – As shown in Figure 3.4, for a scenario with 50 CRs, 48 RUs, and UEs with profile P53, our optimization model presents a solution that consumes 22% less energy than C-RAN. This is because C-RAN centralizes all VNFs in a single CR continuously, which limits optimal server allocation, and results in higher

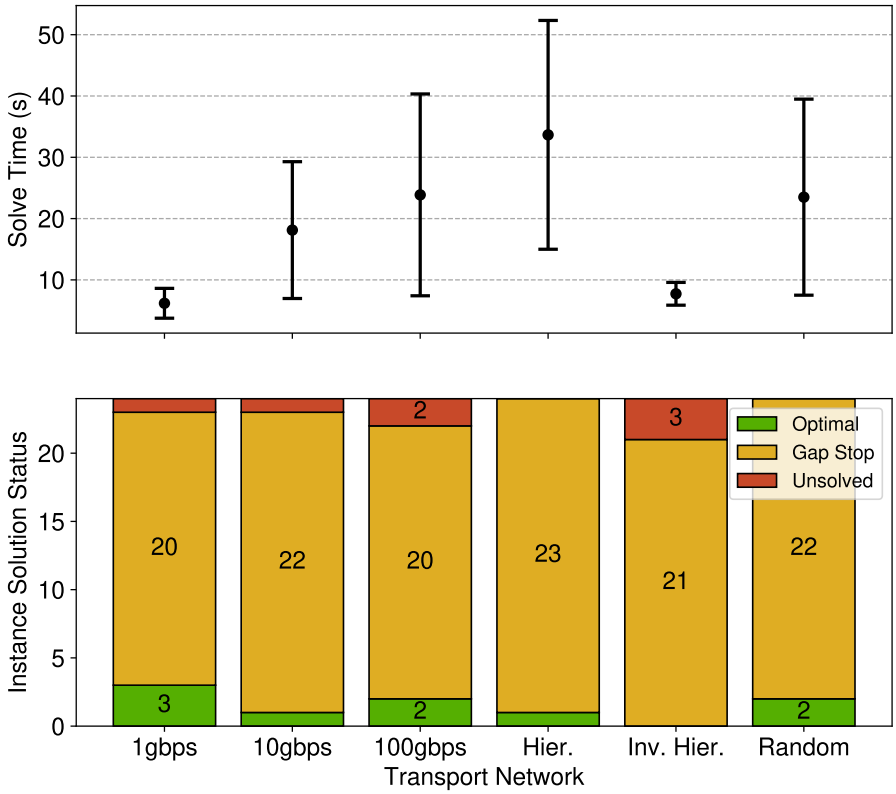


Figure 3.3: Solution time for topologies with 100 CRs and different transport network link capacities.

utilization of the transport network. If compared with D-RAN, the energy savings increase to 52%. In this case, D-RAN deploys the VNFs from each RU in a co-located server, resulting in under-utilization of active equipment. Since the heuristic solution can assign different split options according to the energy cost of each evaluated situation, it achieves a better result than C-RAN, with the optimal solution presenting 14% energy saving overall if compared to the heuristic. Given the assumptions made in the heuristic, it cannot reach the efficacy of the optimal solution. However, as shown in Section 3.2-3.2.4, the heuristic has polynomial complexity and, therefore, has better scalability than the optimal solution.

Impact of Topology Structures – Figure 3.5 shows the empirical distribution of the total energy consumption achieved in different vRAN topologies when our model is employed. Since transceivers with higher throughput capacity are more energetically efficient, topology T1 HC can achieve lower energy consumption and higher centralization rate during low BS load than T1 LC. However, none of them can surpass the energy efficiency of topology T2. Figure 3.6(a) and 3.6(b) show the energy consumption by component for each topology under different conditions, i.e., low and high loads.

Figure 3.7 shows the centralization rate achieved by each topology over time, presented as a percentage of the maximum VNF centralization possible. This maximum centralization occurs when the VNFs of all RUs, implementing functional split 7.2, are

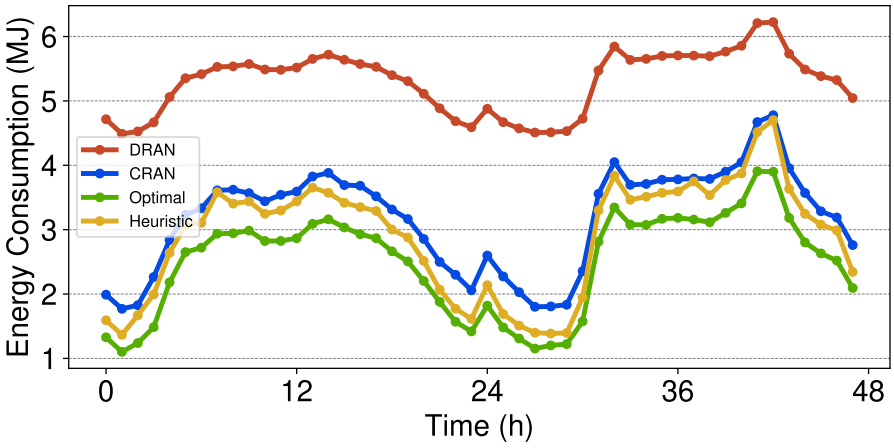


Figure 3.4: Total energy consumption achieved by different solutions.

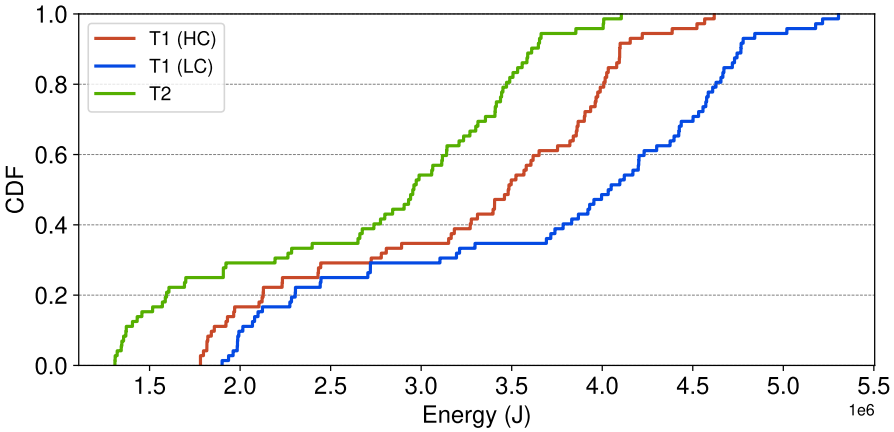
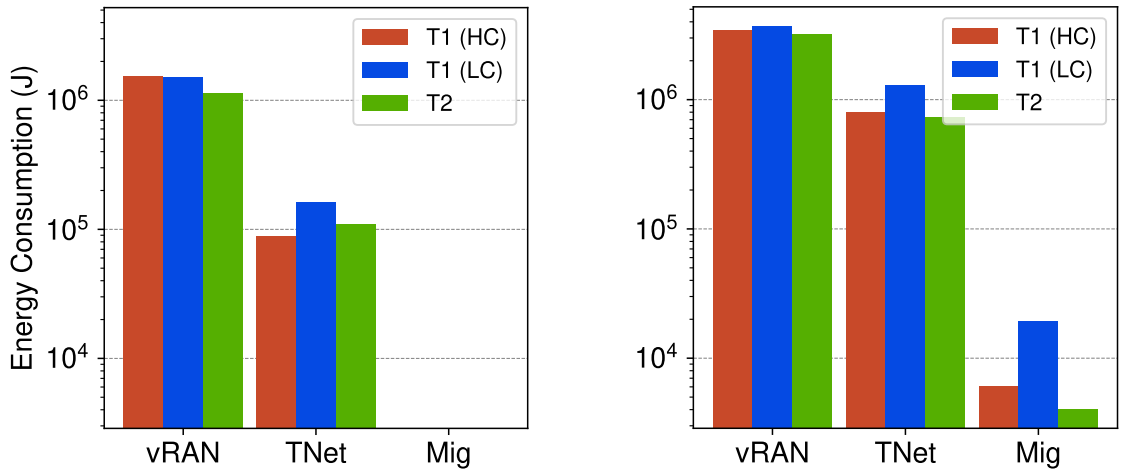


Figure 3.5: Total energy consumption for different vRAN topologies.

deployed in a single CR. We can observe that despite T2 presenting a higher energy usage of the transport network compared to T1 HC, the gains in vRAN processing allow a lower overall energy consumption. This behavior occurs due to the hierarchical organization of the infrastructure components in T2, allowing more effective usage of network equipment, which enables (i) higher centralization of VNFs with lower transport network impact and (ii) most efficient usage of the already turned-on servers in the centralization of CRs during high load. Although the energy consumption of VNF migration is formulated in the objective function, it also indirectly acts as a constraint. As BS load fluctuates over time, it prevents changes in previous associations when the energy overhead in the transport network to deploy a new solution outweighs its energy savings. Beyond that, we could not identify any relation between the number of VNF migrations and the topology structure.

We illustrate solutions during peak BS load for both topologies in Figure 3.8 and Figure 3.9. Topology T1 requires more active CRs, as illustrated by the higher number of percentage tags in the nodes, which inform the turned-on servers CPU load. In topology



(a) *Low load* (b) *High load*

Figure 3.6: Energy consumption per component: vRAN, TNet, and Mig.

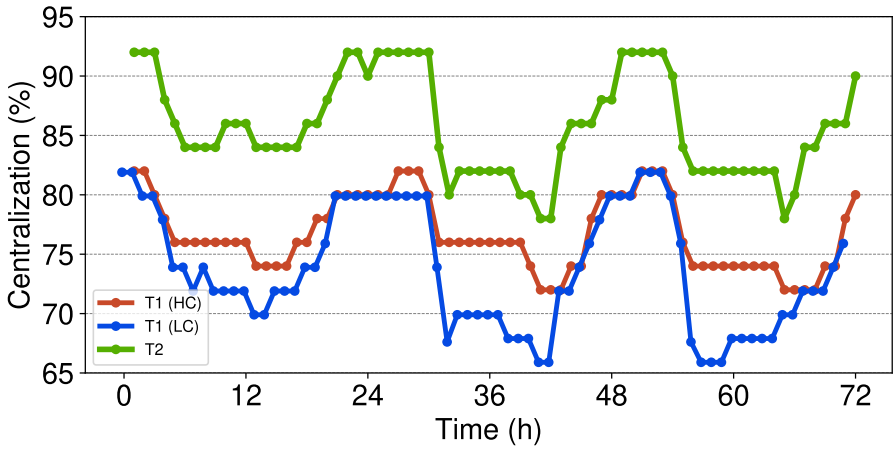


Figure 3.7: Centralization ratio achieved for profile P53.

T2, a slightly higher centralization rate is possible. However, it is not achieved because the model identifies that turning on an additional server in an active CR to aggregate the processing of a few BSs would not be worth the extra energy cost in the transport network.

Impact of Network Usage Profiles – Figure 3.10 shows an empirical distribution of the vRAN energy consumption achieved by the optimization model for different usage profiles. The throughput generated by the devices connected to the mobile network increases the total energy consumption of vRAN. This increase in consumption is more pronounced during high BS load, as illustrated in Figure 3.11(a) and Figure 3.11(b). These figures present the vRAN energy consumption broken down into components in low and high BS load moments. Those figures confirm that VNF migration is the component with the lowest impact, representing less than 2% of the total energy consumption of the

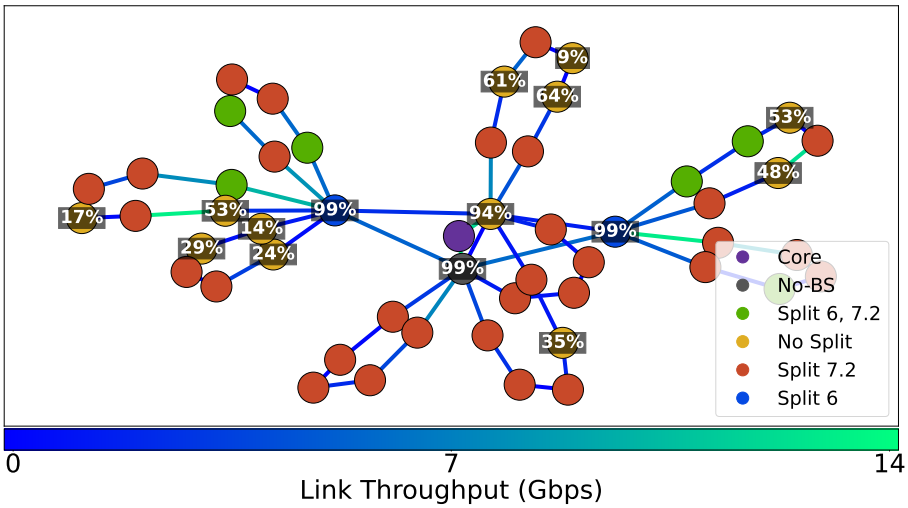


Figure 3.8: Solution for T1 HC topology during peak load.

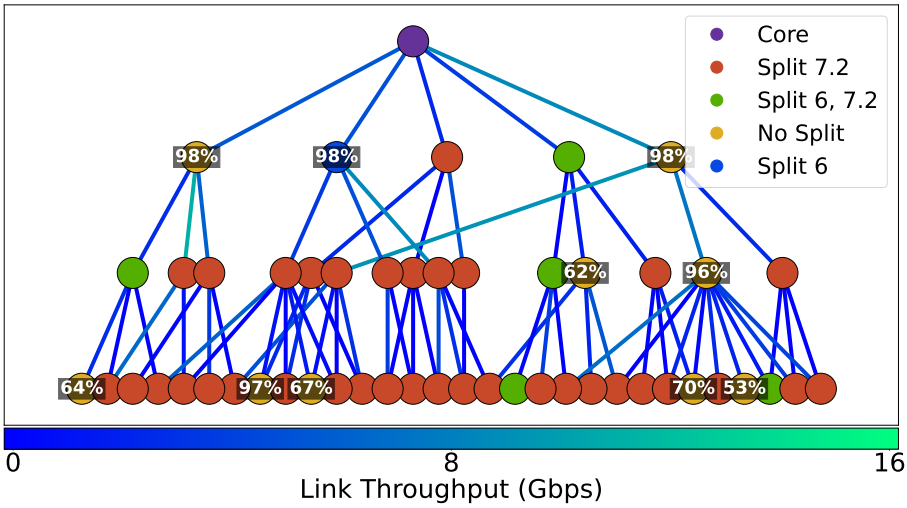


Figure 3.9: Solution for T2 topology during peak load.

topology. This result shows that using a linear approximation model for VNF migration cost does not compromise the robustness of our formulation, as the minimal gain in precision from more complex models does not justify the added complexity. Additionally, the impact of usage profiles on the energy consumed by the CRs running vRAN and VNF migration is minimal. However, the transport network energy consumption is more sensitive to the throughput generated by user devices, further aggravated by the increase in BS load.

The higher demand for the transport network also results in solutions with lower centralization rates, mainly during high BS load, as shown in Figure 3.12. Once the need for more CRs arises, given the increase in BS load with time, using local servers while performing no functional split becomes preferable to turning on more servers in centralization nodes. This behavior occurs because the energy consumption incurred by the transport network outweighs the energy gains achieved by centralizing

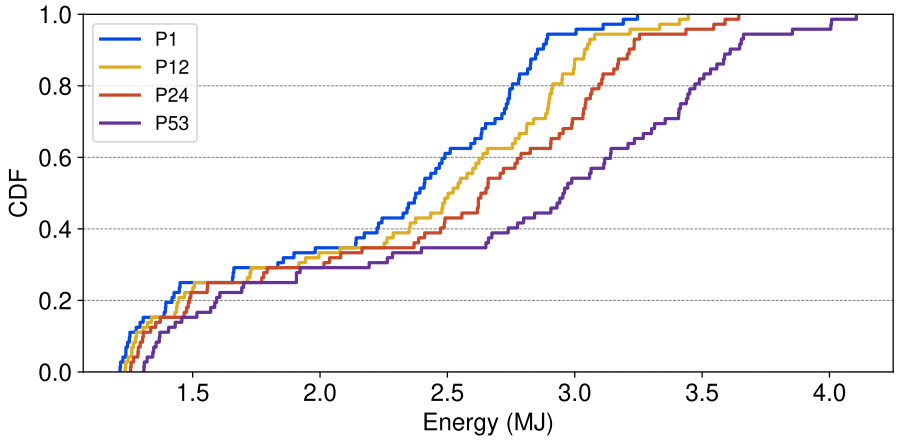
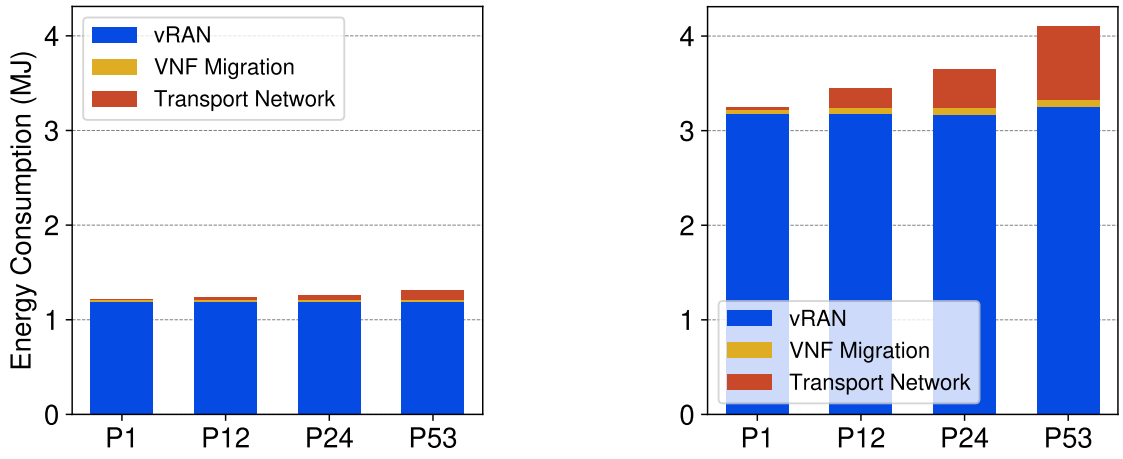


Figure 3.10: Total energy consumption for different network usage profiles.



(a) *Low BS load* (b) *High BS load*

Figure 3.11: Energy consumption per component for different net. usage profiles.

the VNFs of multiple BSs into a single node. On the other side, moments with lower BS load can achieve better energy consumption by increasing the centralization rate. In the formulation, we considered that the maximum latency required by each functional split must be granted as a hard constraint. According to the values in [1], all functional splits remain viable in the evaluated topology. However, the latency requirements of the functional splits may vary, as shown in [44]. If the maximum allowed latency is low enough to prevent the choice of some functional split options, it may limit the centralization rate, and we can expect to observe a negative impact on energy consumption.

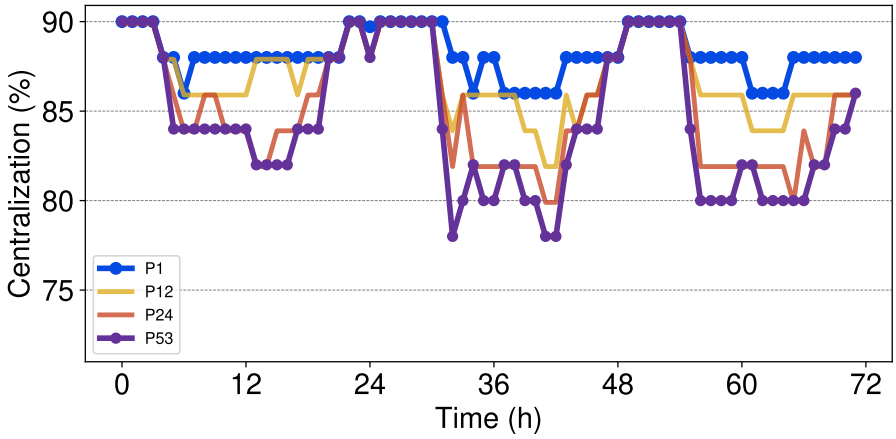


Figure 3.12: Centralization achieved for network usage profiles.

3.4 Conclusion

This chapter addressed the vRAN VNF placement problem with the objective of minimizing infrastructure energy consumption by proposing a flexible and comprehensive MILP formulation. The evaluation accounted for the energy consumption of computing resources and the transport network, as well as VNF migration overhead. Moreover, the results show that an appropriate formulation can keep the model solvable within practical time limits.

The evaluation further indicates that optimal deployment decisions are characterized by the interaction between compute and transport network costs. Increased centralization can reduce energy consumption by improving hardware utilization. However, in scenarios with high-throughput UE demand, transport network energy costs become dominant and tend to limit the degree of centralization. Finally, VNF migration remained a minor contributor in the evaluated instances (below 2%). As a next step, Chapter 4 extends this orchestration problem by incorporating UE-BS association and related radio feasibility constraints into the optimization model.

CHAPTER 4

Energy Efficient User Equipment Association

This chapter extends the formulation presented in Chapter 3 to include the problems of UE-BS association and BS transmission power configuration, in addition to the previously addressed problem of VNF placement. Our objective is to formally state the problem, investigate the scalability of an optimal joint solution, and assess the impacts of a disjoint approach.

4.1 Related work

The problem of VNF placement in 5G RAN, with the objective of minimizing energy consumption, has been widely explored in the literature [17, 34, 36, 42, 46]. These works demonstrate that energy consumption can be reduced through VNF placement optimization. Nonetheless, mobile network infrastructure is typically over-dimensioned to accommodate peak load, which creates the opportunity to achieve greater energy savings by disabling BSs during periods of low activity.

Some works, such as [25] and [49], address the problem of UE-BS association without accounting for energy costs. The former aims to maximize the network Quality of Service (QoS), while the latter seeks to achieve network load balancing. However, both objectives are inherently prone to conflict with energy efficiency goals, since offering higher QoS results in more intense resource utilization and load-balancing benefits from active and underutilized CRs.

Also within this category, [24] proposed a Deep Reinforcement Learning (DRL)-based algorithm to solve the problem of UE-BS association and vRAN placement to bi-objectively minimize end-to-end delay and computing cost. Similarly, [19] addresses the same problem but aims to maximize UE admission in overloaded scenarios. They compare the optimal solution for the joint and disjoint formulations of the problem and introduce a Recurrent Neural Network (RNN) approach for the joint problem.

In [31], the authors formulate the problem of maximizing energy efficiency through computing and cache resource allocation combined with UE-BS association in Fog RAN. Due to the complexity of the problem, a relaxed version is solved using the

Alternating Direction Method of Multipliers (ADMM). In [47], the authors address the problem of BS sleeping, user association, and UE power control to maximize energy efficiency in the fully decoupled RAN (FD-RAN). They employ the Tammer decomposition method to split the problem into two levels, and subsequently apply the many-to-many swap matching and Dinkelbach algorithms to solve each sub-problem. In [51], the multi-agent proximal policy optimization (MAPPO) technique is introduced to solve the problem of UE-BS association, advanced sleeping mode management, and antenna switching in a massive MIMO network under dynamic traffic conditions. Beyond these contributions, none of these works addresses the problems of VNF placement or crosshaul network routing.

In [13] the problem of UE-RU-DU-CU association was formulated with the aim of minimizing Power Usage Effectiveness (PUE) and maximizing power efficiency as a MINLP model. To present a solution within an acceptable time frame, they propose the use of a Quantum Genetic Algorithm. However, the energy consumption of transport network equipment is not considered.

Most solutions [25, 47, 49] address the problems of UE-BS association and vRAN placement disjointly, without evaluating the impact compared to a joint solution. As shown in [19], a degradation in solution quality is expected and must be properly quantified. Thus, in this chapter, we address the problem of jointly optimizing UE-BS association, BS transmission power allocation, VNF placement, and crosshaul transport network routing. In terms of energy consumption, our formulation accounts for the contributions of VNF processing, crosshaul transport network equipment, and the RU. Finally, we compare the optimal solution of the joint formulation with both a disjoint approach and a widely adopted solution that associates each UE with the BS offering the maximum Signal to Noise Ratio (SNR).

4.2 System model

We build on the model presented in Chapter 3 to incorporate UE-BS assignment. The RAN topology is composed of a set $\mathcal{B} = \{b_1, b_2, \ldots, b_{|\mathcal{B}|}\}$ of RUs and a set of general-purpose servers $\mathcal{H} = \{h_1, h_2, \ldots, h_{|\mathcal{H}|}\}$ capable of processing the RAN VNFs. RUs and servers are connected through a set of transport network nodes $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$. We define the set of CRs as $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ where $c_m \subseteq \mathcal{T} \cup \mathcal{H}$ represents a group of co-located switches and servers. We represent the RAN topology as a graph $G = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{H}$ denotes the vertices, where v_0 represents the core network, and $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$ represents the set of edges corresponding to the network links connecting the nodes. Furthermore, we define the set of UEs $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$.

Routing – All data traffic originates (downlink) or terminates (uplink) at the core network v_0. Without loss of generality, we consider only the downlink flow. We define $\mathcal{P}_l$ as the set of all possible paths from each RU $b_l \in \mathcal{B}$ to the core network v_0. Each path is represented as an ordered sequence of CRs connecting the RU to the core. To support different functional split options, a path is decomposed into at most three segments that connect CU, DU, and RU: p_{Bh} (backhaul), p_{Mh} (midhaul), and p_{Fh} (fronthaul). Moreover, we consider a crosshaul transport network, i.e., a link $e_{ij} \in \mathcal{E}$ can serve different segment parts for different BSs.

Virtual Network Functions – For each RU $b_l \in \mathcal{B}$, our objective is to determine the best CR to deploy the RU set of RAN VNFs, denoted by $\mathcal{F} = \{f_1, f_2, f_3, f_4, f_5\}$ representing High-PHY, MAC, RLC, PDCP, and RRC functions, respectively. The VNF stack may be partitioned up to two times according to the functional split combination option $D_r \in \mathcal{D}$, which enhances the flexibility of the solution by allowing parts of the VNF stack to be deployed in different CRs. The Low-PHY function is not virtualized and is assumed to always be deployed at the RU. In our formulation, each option $D_r \in \mathcal{D}$, with $D_r \neq D_0$, represents either a single functional split or a combination of two functional split options presented in Section 2.2. The option $D_0 \in \mathcal{D}$ denotes the D-RAN configuration, in which no functional split is performed and all VNFs are placed on a server co-located with the RU.

4.2.1 VNF processing

Let the binary decision variable $x_{l,k}^{p,r}$ indicate whether UE $k \in \mathcal{K}$ is associated with BS $b_l \in \mathcal{B}$ performing functional split $D_r \in \mathcal{D}$ on the path $p \in \mathcal{P}_l$. In addition, the binary decision variable $z_l^{p,r}$ is based on $x_{l,k}^{p,r}$ and denotes if the path $p \in \mathcal{P}_l$ was chosen for $b_l \in \mathcal{B}$ using the functional split $D_r \in \mathcal{D}$.

During the decision process, servers allocated by the choice of path $p \in \mathcal{P}_l$ must present the required processing capacity. To achieve this, we model the computing resource cost for each VNF in terms of GOPS as described below.

We adopt local partial minimum mean square error (LP-MMSE) as the precoding method, for which the computational cost was formulated in [9] as follows:

$$
\begin{aligned}
C_{precoding}^{p,r,l} = z_l^{p,r} \Bigg(& \frac{N_{used}}{T_s \tau_c 10^9} \left(8 N_l \tau_p^2 + 8 N_l^2 \tau_p\right) + \\
& \frac{N_{used}}{T_s \tau_c 10^9} \left((4 N_l^2 + 4 N_l)\tau_p + \frac{8(N_l^3 - N_l)}{3}\right)\Bigg) + \\
& \sum_{k \in \mathcal{K}} x_{l,k}^{p,r} \left(\frac{N_{used}}{T_s \tau_c 10^9} 16 N_l^2 + \frac{N_{used} \tau_d}{T_s \tau_c 10^9} 8 N_l + \frac{N_{used}}{T_s \tau_c 10^9} 8 N_l \right).
\end{aligned} \tag{4-1}
$$

In this expression, N_{used} denotes the number of used subcarriers, T_s is related to the OFDM symbol duration, and N_l is the number of antennas at RU $b_l \in \mathcal{B}$. As in [9], pilot symbols for OFDM channel estimation are assumed to be transmitted in every coherence block. Accordingly, τ_c denotes the number of samples per coherence block, τ_p is the number of samples used for uplink training, and $\tau_d = \tau_c - \tau_p$ is the number of samples effectively available for downlink data transmission.

Next, based on the model in [8], we estimate the computational costs of OFDM modulation and of mapping modulated symbols onto resource elements (REs), respectively, as:

$$C^{p,r,l}_{modulation} = 1.3 N_l \left(\frac{N_{bits}}{16}\right)^{1.2} z^{p,r}_l, \tag{4-2}$$

$$C^{p,r,l}_{mapping} = 1.3 \left(\frac{N_{bits}}{16}\right)^{1.2} \left(\frac{SE_0}{6}\right)^{1.5} \sum_{k \in \mathcal{K}} x^{p,r}_{l,k}, \tag{4-3}$$

where N_{bits} is the number of bits used for data quantization and SE_0 denotes the spectral efficiency of the channel. Therefore, the number of GOPS required for High-PHY layer processing is given by:

$$C^{p,r,l}_{HighPHY} = C^{p,r,l}_{precoding} + C^{p,r,l}_{modulation} + C^{p,r,l}_{mapping}. \tag{4-4}$$

Channel coding, system control, and data redirection to the core network operations are implemented in the higher (or superior) layers of the RAN VNF stack. Similar to [8], the number of GOPS required by all these superior layers is calculated as follows:

$$\begin{aligned} C^{p,r,l}_{SupLayers} =& 1.3 \left(\frac{N_{bits}}{16}\right)^{1.2} \left(\frac{SE_0}{6}\right) \sum_{k \in \mathcal{K}} x^{p,r}_{l,k} + \\ & 2.7 \sqrt{N_l} \left(\frac{N_{bits}}{16}\right)^{0.2} z^{p,r}_l + 8 \left(\frac{SE_0}{6}\right) \sum_{k \in \mathcal{K}} x^{p,r}_{l,k}. \end{aligned} \tag{4-5}$$

Finally, to determine the processing required by each VNF individually, given the aggregate load of the superior layers $C_{SupLayers}$ from Eq. 4-5, we utilize proportions based on the CPU utilization profiles observed in the OAI implementation [26] to define the total processing load γ_w for a given server $h_w \in \mathcal{H}$ as follows:

$$\begin{aligned} \gamma_w = \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \Bigg[& F_1(h_w, D_r, p) C^{p,r,l}_{HighPHY} + F_2(h_w, D_r, p) \left(0.4\ C^{p,r,l}_{SupLayers}\right) + \\ & F_3(h_w, D_r, p) \left(0.028\ C^{p,r,l}_{SupLayers}\right) + F_4(h_w, D_r, p) \left(0.286\ C^{p,r,l}_{SupLayers}\right) + \\ & F_5(h_w, D_r, p) \left(0.286\ C^{p,r,l}_{SupLayers}\right) \Bigg], \end{aligned} \tag{4-6}$$

where the mapping function $F_n(h_w, D_r, p)$, based on the input data, returns 1 when server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ according to functional split $D_r \in \mathcal{D}$ and route $p \in \mathcal{P}_l$; otherwise, it returns 0.

4.2.2 Signal model

The Signal to Interference plus Noise Ratio (SINR) quantifies the quality of a wireless link through the ratio between the strength of the received signal power to the combined effect of interference generated by other devices sharing the same communication channel plus background noise. For a UE $k \in \mathcal{K}$, connected to BS $b_l \in \mathcal{B}$, SINR is expressed as:

$$SINR_{l,k} = \frac{P_{rx}(l,k)}{\sum\limits_{D_r \in \mathcal{D}} \sum\limits_{p \in \mathcal{P}_l} \sum\limits_{b'_l \in \mathcal{B}^*_l} \sum\limits_{\substack{k' \in \mathcal{K}, \\ k' \neq k}} x^{p,r}_{l',k'} P_{rx}(l,k') + N_0 B}, \tag{4-7}$$

where N_0 is the background noise power spectral density, B is the bandwidth, and $\mathcal{B}^*_l$ is the set of BSs that share the same frequency channel with a given BS $b_l \in \mathcal{B}$. The power $P_{rx}(l,k)$ of the signal received at BS $b_l \in \mathcal{B}$ from UE $k \in \mathcal{K}$, is modeled as:

$$P_{rx}(l,k) = P_k + G_l + G_k - L(l,k), \tag{4-8}$$

where P_k is the transmit power of UE $k \in \mathcal{K}$, G_l and G_k represent the antenna gains at BS $b_l \in \mathcal{B}$ and UE $k \in \mathcal{K}$, respectively, and $L(l,k)$ is the path loss between BS $b_l \in \mathcal{B}$ and UE $k \in \mathcal{K}$.

The capacity of a BS employing Orthogonal Frequency Division Multiple Access (OFDMA), in terms of radio resources, is given in Resource Blocks (RBs). An RB represents the smallest unit in the time-frequency domain that can be allocated to a UE. According to the Shannon-Hartley theorem, and assuming that the spectral efficiency for a BS-UE association is uniform across all RBs, while disregarding the channel capacity used for control signaling and considering a single spatial layer per RB, the data rate per RB assigned for UE $k \in \mathcal{K}$ in BS $b_l \in \mathcal{B}$ is defined as:

$$R^{l,k}_{RB} = B_{\mathrm{RB}} \log_2(1 + SINR_{l,k}), \tag{4-9}$$

where B_{RB} is the bandwidth of a single RB, which depends on the overall channel bandwidth and the subcarrier spacing selected by the BS.

4.3 Problem statement

To formulate the problem of RAN VNF placement and UE-BS association to minimize energy consumption, we decompose the energy consumption into three main components.

vRAN energy consumption – The energy consumed by the general purpose servers processing the VNFs of all RUs $b_l \in \mathcal{B}$ characterizes the vRAN energy consumption. We use a traditional energy model [11] to estimate this energy consumption, in which the total energy consumed by a server $h_w \in \mathcal{H}$ over the period T includes a static power component P_w^{idle} consumed whether the server is turned on, and a dynamic load-dependent power consumption $P_w^{busy} - P_w^{idle}$, defined as:

$$E_{vRAN} = \sum_{h_w \in \mathcal{H}} T\left[\psi_w^{on} P_w^{idle} + \left(\gamma_w / C_w^{GOPS}\right)\left(P_w^{busy} - P_w^{idle}\right)\right], \tag{4-10}$$

where T is the expected duration of the deployment in seconds, γ_w is the total load assigned to server $h_w \in \mathcal{H}$ as calculated in Eq. 4-6, C_w^{GOPS} represents the number of GOPS the server $h_w \in \mathcal{H}$ can perform. ψ_w^{on} is a ceiling function that ensures a server $h_w \in \mathcal{H}$ is counted as active if, and only if, at least one VNF of any RU is assigned to it. This ceiling function is defined as follows:

$$\psi_w^{on} = \left\lceil \sum_{f_n \in \mathcal{F}} \sum_{b_l \in \mathcal{B}} \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \frac{z_l^{p,r} u_w^p M(h_w, f_n, b_l, D_r)}{|\mathcal{F}|\,|\mathcal{B}|} \right\rceil, \tag{4-11}$$

where $u_w^p \in \{0, 1\}$ is based on the input data, indicating whether a server $h_w \in \mathcal{H}$ is part or not of the route $p \in \mathcal{P}_l$. The mapping function $M(h_w, f_n, b_l, D_r)$ returns 1 when the server $h_w \in \mathcal{H}$ processes VNF $f_n \in \mathcal{F}$ from RU $b_l \in \mathcal{B}$ according to functional split $D_r \in \mathcal{D}$.

Radio energy consumption – Let w_l be a integer decision variable that represents the transmit power from BS $b_l \in \mathcal{B}$ in watts. Similar to the approach in [9], we employ a generic model based in [6] to represent the total energy consumption of radio equipment:

$$E_{RU} = \sum_{b_l \in \mathcal{B}} \left(\psi_l^{on} P_l^{idle} N_l + \Delta^{tr} w_l\right) \cdot T, \tag{4-12}$$

where P_l^{idle} denotes the power consumed by BS $b_l \in \mathcal{B}$ when there is no load, the slope Δ^{tr} maps transmit power to load-dependent power consumption, and ψ_l^{on} is an expression that indicates whether the BS $b_l \in \mathcal{B}$ is active, as formulated in the following equation:

$$\psi_l^{on} = \left\lceil \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \sum_{k \in \mathcal{K}} \frac{x_{l,k}^{p,r}}{|\mathcal{K}|} \right\rceil. \tag{4-13}$$

Transport network energy consumption – We consider an optical transport network

where the energy consumption arises from (i) Ethernet switches, which allow for higher routing flexibility by enabling packet aggregation, and (ii) DWDM pluggable transceivers that can be directly installed in radio devices or switches [12, 43]. Each link $e_{ij} \in \mathcal{E}$ is characterized by its data transmission capacity $R^{tr}_{e_{ij}}$ and power consumption $P^{tr}_{e_{ij}}$ of the transceivers at each end of the link. For each topology node $v_k \in \mathcal{V}$, the function $S(v_k) \in \{0,1\}$ indicates whether v_k is a packet switch, while $P^{s}_{v_k}$ represents the power consumed by each switch port. Finally, The transport network energy consumption is defined as follows:

$$E_{TNet} = \sum_{e_{ij} \in \mathcal{E}} \left[T \frac{\gamma_{e_{ij}}}{R^{tr}_{e_{ij}}} \left(2P^{tr}_{e_{ij}} + S(v_j) P^{s}_{v_j} + S(v_i) P^{s}_{v_i} \right) \right]. \tag{4-14}$$

Component $\gamma_{e_{ij}}$ in Eq. 4-14 represents the total throughput over link e_{ij}, which is defined as:

$$\gamma_{e_{ij}} = \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \sum_{p \in \mathcal{P}_l} \sum_{k \in \mathcal{K}} x^{p,r}_{l,k} R^k \left(y^{p_{Bh}}_{e_{ij}} \alpha^{r,l}_{Bh} + y^{p_{Mh}}_{e_{ij}} \alpha^{r,l}_{Mh} + y^{p_{Fh}}_{e_{ij}} \alpha^{r,l}_{Fh} \right), \tag{4-15}$$

where R^k represents the data throughput incurred from UE $k \in \mathcal{K}$. $y^{p_{Bh}}_{e_{ij}}$, $y^{p_{Mh}}_{e_{ij}}$, and $y^{p_{Fh}}_{e_{ij}}$ indicate if the link e_{ij} is part of the backhaul, midhaul, or fronthaul, respectively, for the path $p \in \mathcal{P}_l$. For each RU $b_l \in \mathcal{B}$, the chosen functional split $D_r \in \mathcal{D}$ increases the data rate required at the backhaul, midhaul, and fronthaul by the factors $\alpha^{r,l}_{Bh}$, $\alpha^{r,l}_{Mh}$, and $\alpha^{r,l}_{Fh}$, respectively.

The primary objective of this work is to minimize the overall energy consumption associated with the processing of vRAN VNFs, the operation of radio unit equipment, and the utilization of the transport network interconnecting the CRs. Therefore, we define the objective function as:

$$\text{minimize} \quad E_{total} = E_{vRAN} + E_{RU} + E_{TNet}. \tag{4-16}$$

To ensure that the association between UE and BS is feasible, we first require that the SINR for every UE $k \in \mathcal{K}$ is bounded below by the minimum ratio $SINR_{min}$ necessary for a BS to decode the received signal, as formulated in the following constraint:

$$\sum_{D_r \in \mathcal{D}} \sum_{b_l \in \mathcal{B}} \sum_{p \in \mathcal{P}_l} x^{p,r}_{l,k} SINR_{l,k} \geq 10^{SINR_{min}/10}, \quad \forall k \in \mathcal{K}, \tag{4-17}$$

next, if an UE is associated with a given BS, the power of the signal received by the UE, after attenuation, must be above the minimum threshold S_{min}, as expressed in the following constraint:

$$\frac{w_l}{L'(l,k)} \geq S_{min} \cdot \sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} x^{p,r}_{l,k}, \quad \forall b_l \in \mathcal{B}, \forall k \in \mathcal{K}, \tag{4-18}$$

where $L'(l,k)$ denotes the path loss converted into linear units, since w_l represents the transmit power in milliwatts.

For each RU $b_l \in \mathcal{B}$, at most one combination of route $p \in \mathcal{P}_l$ and functional split $D_r \in \mathcal{D}$ must be assigned, as represented in the following constraint:

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} z_l^{p,r} \leq 1, \quad \forall b_l \in \mathcal{B}, \tag{4-19}$$

Similarly, each UE $k \in \mathcal{K}$ must be associated with exactly one combination of BS $b_l \in \mathcal{B}$ and channel $ch \in \mathcal{C}_h$:

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \sum_{b_l \in \mathcal{B}} x_{l,k}^{p,r} = 1, \quad \forall k \in \mathcal{K}. \tag{4-20}$$

When a BS is active, the assigned transmission power must remain within the minimum and maximum constraints defined by its operational capacity, expressed as:

$$w_l \geq P_{min} \cdot \sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} z_l^{p,r}, \quad \forall b_l \in \mathcal{B}, \tag{4-21}$$

$$w_l \leq P_{max}, \quad \forall b_l \in \mathcal{B}. \tag{4-22}$$

Each link $e_{ij} \in \mathcal{E}$ has a maximum data rate capacity e_{ij}^{Cap} defined by the number of fibers and data rate of transceivers composing it. This maximum capacity must not be exceeded, as represented in the following constraint:

$$\gamma_{e_{ij}} \leq e_{ij}^{Cap}, \quad \forall e_{ij} \in \mathcal{E}. \tag{4-23}$$

Depending on the functional split $D_r \in \mathcal{D}$ chosen, different latencies must be granted at *fronthaul* (β_{Fh}^r), *midhaul* (β_{Mh}^r), and *backhaul* (β_{Bh}^r) of path $p \in \mathcal{P}_l$. Since each link $e_{ij} \in \mathcal{E}$ incurs in delay e_{ij}^L according to its capacity, distance between nodes, number of hops, and packet queue size, the chosen path $p \in \mathcal{P}_l$ must ensure the latency required by the functional split $D_r \in \mathcal{D}$, as defined in the following constraints:

$$\sum_{e_{ij} \in \mathcal{E}} z_l^{p,r} y_{e_{ij}}^{p_{Bh}} e_{ij}^L \leq \beta_{Bh}^r, \ \ \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{4-24}$$

$$\sum_{e_{ij} \in \mathcal{E}} z_l^{p,r} y_{e_{ij}}^{p_{Mh}} e_{ij}^L \leq \beta_{Mh}^r, \ \ \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l, \tag{4-25}$$

$$\sum_{e_{ij} \in \mathcal{E}} z_l^{p,r} y_{e_{ij}}^{p_{Fh}} e_{ij}^L \leq \beta_{Fh}^r, \ \ \forall b_l \in \mathcal{B}, D_r \in \mathcal{D}, p \in \mathcal{P}_l. \tag{4-26}$$

The VNFs assigned to a given server $h_w \in \mathcal{H}$ must not exceed its maximum

processing capacity C_w^{GOPS}, as defined by the following constraint:

$$\gamma_w \leq C_w^{GOPS}, \quad \forall h_w \in \mathcal{H}. \tag{4-27}$$

Similarly, the set of UEs associated with a specific BS $b_l \in \mathcal{B}$ must not exceed its maximum available RBs, denoted as C_l^{RB}, which is defined by the following constraint:

$$\sum_{p \in \mathcal{P}_l} \sum_{D_r \in \mathcal{D}} \sum_{k \in \mathcal{K}} \frac{x_{l,k}^{p,r} R^k}{R_{RB}^{l,k}} \leq C_l^{RB} \quad \forall b_l \in \mathcal{B}. \tag{4-28}$$

Finally, we establish the relation between the decision variables z and x, and define their binary scope:

$$z_l^{p,r} \geq \sum_{k \in \mathcal{K}} \frac{x_{l,k}^{p,r}}{|\mathcal{K}|}, \tag{4-29}$$

$$x_{l,k}^{p,r},\ z_l^{p,r} \in \{0,1\}, \tag{4-30}$$

$$w_l \in \mathbb{N}. \tag{4-31}$$

Tables 4.1 and 4.2 summarizes all the data parameters, decision variables, and sets used throughout the formulation.

4.4 Evaluation

To evaluate the proposed formulation, we implement the optimization model using the python library docplex (version 2.28.240) in conjunction with CPLEX (version 22.1.1.0) as the solver. Since constraint (4-17) renders the problem non-linear, the CPLEX Constraint Programming (CP) module is employed due to its capability of solving problems in this category. In the topology considered for input data, illustrated in Figure 4.1, the BSs were positioned within a 2 Km^2 area using Poisson-Disk sampling with a minimum separation radius of 500 meters. The transport network links were generated randomly, while the UE locations are also random and were sampled uniformly within the coverage area of the BSs.

Additionally, the Sionna simulation library [20] was employed to generate radio link input data for the instances. The Urban Macrocell (UMa) scenario was considered, using the frequency band n78 with a subcarrier spacing of 15 kHz, divided into 3 channels of 20 MHz each. Shadow fading was not considered, as it introduces stochastic variability into the propagation model. Isolating its impact on optimal decisions from that of topology architecture and UE usage profiles would require a large number of instances within each scenario, along with multiple channel simulations, to obtain statistically reliable

Table 4.1: Inputa Data

Notation	Description
C_w^{GOPS}	Processing capacity of server $h_w \in \mathcal{H}$
C_l^{RB}	RBs available in BS $b_l \in \mathcal{B}$
P_w^{busy}	Load-dependent power consumption of server $h_w \in \mathcal{H}$
P_w^{idle}	Static power consumption of server $h_w \in \mathcal{H}$
$P_{rx}(l,k)$	Power of the signal received at BS $b_l \in \mathcal{B}$ from UE $k \in \mathcal{K}$
$P_{v_i}^s$	Ethernet switch port power consumption in node $v_i \in \mathcal{V}$
$P_{e_{ij}}^{tr}$	Power consumption of a transceiver in the link $e_{ij} \in \mathcal{E}$
P_k	Transmit power of UE $k \in \mathcal{K}$
Δ^{tr}	Constant value that maps transmit power to load-dependent energy consumption of an RU
G_l, G_k	Antenna gain at BS $b_l \in \mathcal{B}$ and UE $k \in \mathcal{K}$
$L(l,k)$	Path loss between BS $b_l \in \mathcal{B}$ and UE $k \in \mathcal{K}$
e_{ij}^{Cap}	Maximum link data rate
e_{ij}^{L}	End-to-end link latency
$R_{e_{ij}}^{tr}$	Transceiver transmission capacity
R^k	Data throughput from UE $k \in \mathcal{K}$
T	Expected solution deployment duration
T_s	OFDM symbol duration
N_{used}	Number of sub-carriers used
N_{bits}	Number of bits used for data quantization
N_l	Number of antennas in RU $b_l \in \mathcal{B}$
u_w^p	Indicates whether a server $h_w \in \mathcal{H}$ is part of the path $p \in \mathcal{P}_l$
$v_{m,w}^p$	Indicates whether a server $h_w \in \mathcal{H}$, part of the path $p \in \mathcal{P}_l$, is associated with CR $c_m \in \mathcal{C}$
$SINR_{min}$	Minimum SINR threshold
S_{min}	UE radio receiver sensitivity
SE_0	Channel spectral efficiency
N_0	Background noise power spectral density
B	Bandwidth
B_{RB}	Bandwidth if a single RB
τ_c	Number of samples per coherence block
τ_p	Number of received samples during training phase
τ_d	Number of received samples during downlink data transmission

Table 4.2: Sets, Input Data, Decision Variables, and Expressions

	Notation	Description
Sets	$\mathcal{B}$	Set of RUs
	$\mathcal{H}$	Set of general-purpose servers
	$\mathcal{T}$	Set of transport network nodes
	$\mathcal{C}$	Set of CRs
	$\mathcal{G}$	Topology graph where $G = (\mathcal{V}, \mathcal{E})$
	$\mathcal{V}$	Set of topology nodes where $\mathcal{V} = \{v_0\} \cup \mathcal{B} \cup \mathcal{T} \cup \mathcal{H}$
	$\mathcal{E}$	Set of topology links where $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$
	$\mathcal{K}$	Set of UEs
	$\mathcal{P}_l$	Set of all paths from each RU $b_l \in \mathcal{B}$ to the core network v_0
	$\mathcal{F}$	Set of VNFs
	$\mathcal{D}$	Set of functional splits
Decision Vars. and Exprs.	$x_{l,k}^{p,r}$	Binary decision variable indicating that path $p \in \mathcal{P}_l$ was chosen for RU $b_l \in \mathcal{B}$ using functional split $D_r \in \mathcal{D}$ serving UE $k \in \mathcal{K}$
	$z_l^{p,r}$	Binary decision variable indicating that path $p \in \mathcal{P}_l$ was chosen for RU $b_l \in \mathcal{B}$ using functional split $D_r \in \mathcal{D}$
	w_l	Integer decision variable representing transmit power from RU $b_l \in \mathcal{B}$ in watts
	γ_w	Processing load for server $h_w \in \mathcal{H}$
	$\gamma_{e_{i,j}}$	Total throughput over link $e_{i,j}$
	ψ_w^{on}	Indicate if server $h_w \in \mathcal{H}$ is assigned to the processing of any VNF and needs to be activated
	ψ_l^{on}	Indicate if BS $b_l \in \mathcal{B}$ is active
	$\psi_{m,n}^{single}$	Indicate if at least two different RUs deploy the same VNF $f_n \in \mathcal{F}$ in CR $c_m \in \mathcal{C}$

performance estimates. By disabling this source of randomness, the evaluation focuses on deterministic topology characteristics while reducing the number of required experiments.

The BS transmitter was configured with dual polarization and 2 omnidirectional antennas, with a maximum transmit power of 33 dBm, while the UE transmitter was equipped with a single omnidirectional antenna and a maximum transmit power of 23 dBm. The remaining parameters utilized in the model evaluation are maintained as described in Chapter 3 and summarized in Table 4.3.

4.4.1 Joint versus disjoint approaches

The decision regarding the UE-BS association problem has the potential to influence subsequent VNF placement and routing decisions. To better understand this interdependence, we evaluate the joint optimization problem of UE-BS association, VNF placement, and routing, as defined in Section 4.3. For comparison, we also consider a disjoint approach in which the UE-BS association problem is solved independently, and

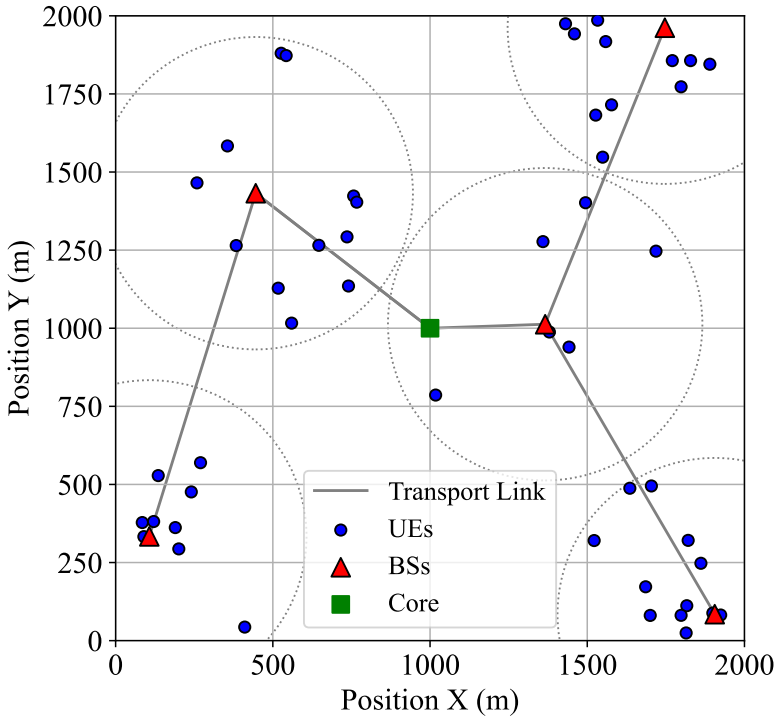


Figure 4.1: Topology used in evaluation with 5 BSs and 50 UEs.

Table 4.3: Evaluation parameters

Parameter	Value	Parameter	Value
$\|\mathcal{B}\|$	5	ρ^c	0
C_w^{GOPS}	180	P_w^{busy}	94.8 W
P_w^{idle}	20-25% of P_w^{busy}	E_{f_s}	{1795, 242.08, 172.92, 410, 410} MB
e_{ij}^{Cap}	1000 Gbps	e_{ij}^{L}	$(10^{-1}, 10^{-4})$ ms
$R_{e_{ij}}^{tr}$	100 Gbps	$P_{e_{ij}}^{tr}$	4.5 W
$P_{v_i}^{s}$	14.0 W	T	3600 s
T_s	71.4 µs	N_{used}	1200
N_{bits}	12	N_l	6
SE_0	1.0 bits/s/Hz	τ_c, τ_p	192, 8
P_k	23 dBm	G_l, G_k, X_σ	16 dB, 0 dB, 0 dB
P_l^{idle}	84 W	Δ^{tr}	2.8
G_{virt}	0.723	$SINR_{min}$	10.0 dB
S_{min}	-121.7 dBm		

its solution is then used as input for the VNF placement plus routing problem.

In Figure 4.2, we present a comparison of the overall energy consumption between joint and disjoint solutions across instances with varying numbers of UEs. The

solver was configured with a time limit of 10 minutes; therefore, the joint model was able to solve only up to the instance with 38 UEs. As the number of UEs increases, a larger set of active BSs is required to satisfy the minimum SINR required for each UE. Since the RU represents the major source of energy consumption within a BS, this increase in active BSs can be identified by jumps in energy consumption, which can be observed in instances with 2, 9, 10, and 22 UEs.

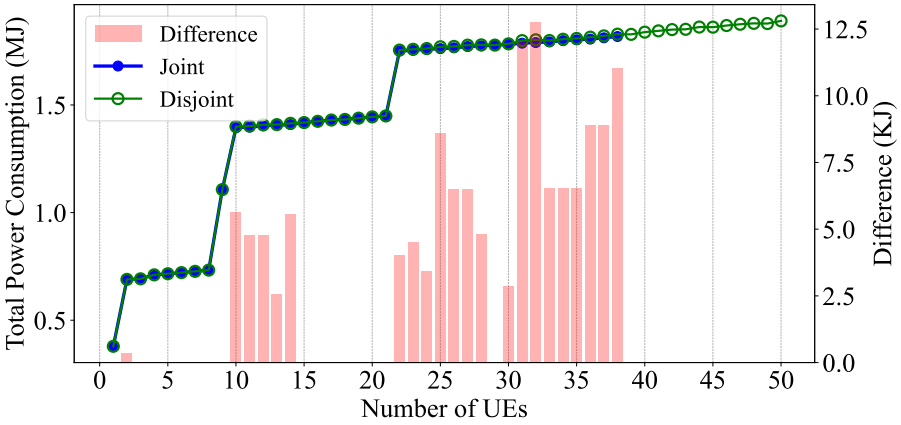


Figure 4.2: Overall energy consumption comparison between joint and disjoint solutions across instances with varying numbers of UEs. The left y-axis corresponds to the line curves, while the right y-axis corresponds to the bar values.

In some instances, the disjoint approach finds solutions with the same energy consumption as the Joint approach. However, when the disjoint approach performs worse, the difference in objective cost remains 2 to 3 orders of magnitude smaller than the total cost. To further investigate the source of this gap, Figure 4.3 decomposes the energy consumption for the instance with the largest observed difference (32 UEs). In this case, solving only the UE-BS association enables the disjoint solution to achieve a 0.1% reduction in radio power consumption. Nevertheless, the subsequent stage of VNF placement and routing increases the cost by 0.2% and 36.5%, respectively, thereby worsening the overall solution cost. This instance highlights the interdependence among the problems; yet, the difference in objective cost remains minimal, while the ability to solve larger instances suggests that the disjoint approach is a robust solution.

4.4.2 Comparison with best SNR approach

A common solution for UE-BS association currently employed in mobile networks is to associate each UE with the BS offering the highest SNR. As shown in Figure 4.4, this approach leads to a greater number of active BSs for some instances with lower UE density compared to the optimization model solution. This occurs because the

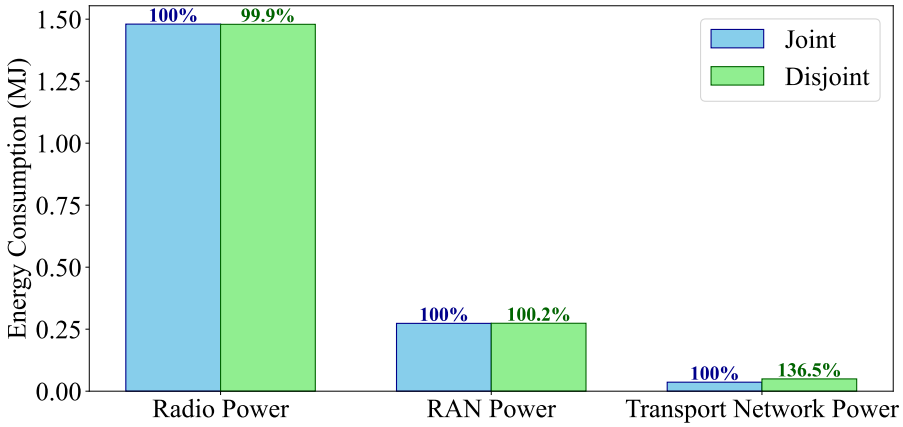


Figure 4.3: Comparison of energy consumption by component for joint and disjoint solutions of instance with 32 UEs.

BS providing the best SNR must be activated, even when the instance could be satisfied by a smaller set of BSs offering a slightly lower SNR.

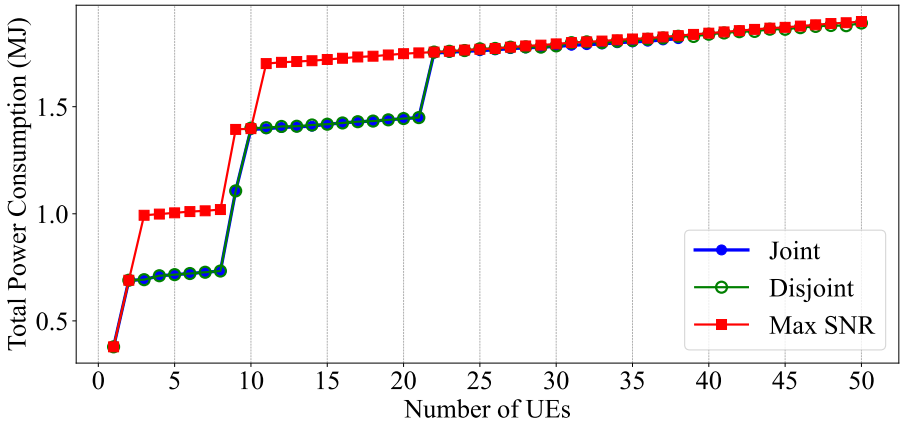


Figure 4.4: comparison of overall energy consumption between joint, disjoint and Maximum SNR solutions.

When each UE is assigned to the BS offering the highest SNR, fewer RBs are required to satisfy the UE data rate demand, as illustrated in Figure 4.5. However, because this approach disregards inter-BS interference, starting from the instance with 35 UEs, interference prevents the communication of some UEs resulting in the requirement of an unfeasible quantity of RBs.

Furthermore, beyond the reduced energy consumption achieved in instances with lower UE density, Figure 4.6 shows that, unlike the maximum SNR approach, the optimization model consistently satisfies the 10 dB of SINR requirement in instances with higher UE density, as expected, since this condition was formulated as a hard constraint. Finally, when the disjoint approach differs from the joint approach, the results indicate slightly better SINR in the disjoint approach. However, this appears to be an arbitrary

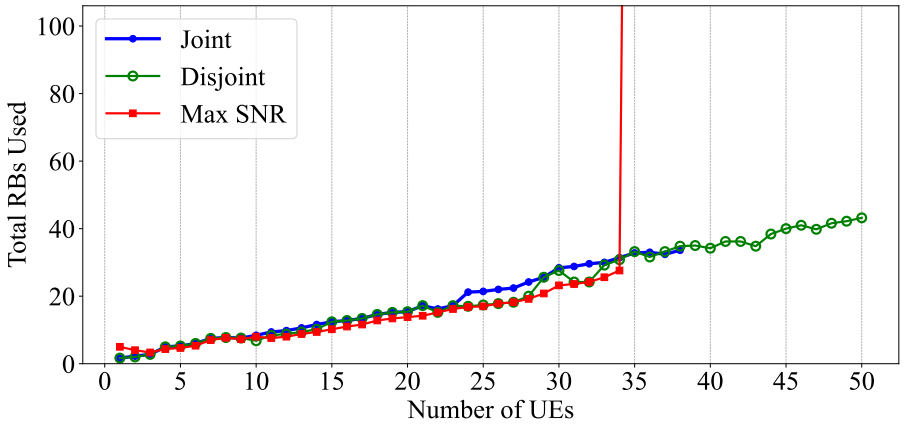


Figure 4.5: Comparison of mean RB usage between joint, disjoint and Maximum SNR solutions.

outcome in the evaluated instance, since the formulation guaranties only the minimum SINR and provides no logic that could lead to one approach presenting solutions with better SINR.

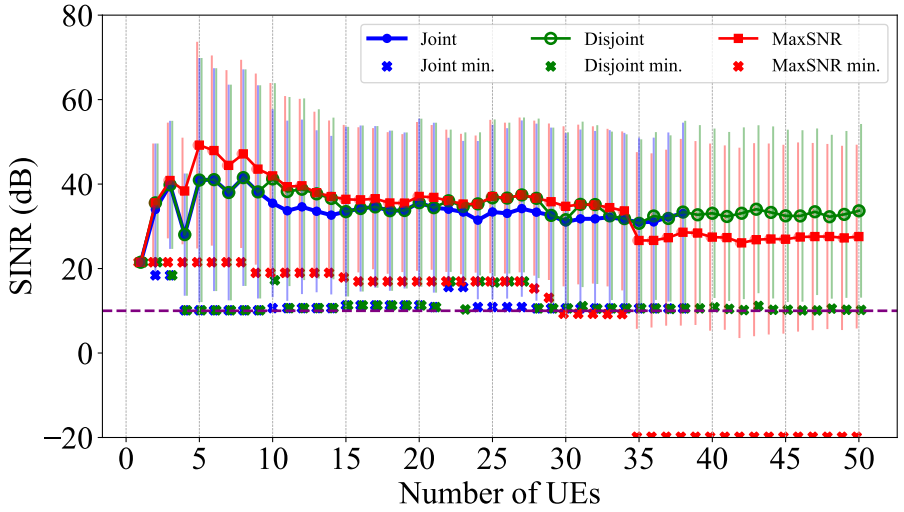


Figure 4.6: Comparison of SINR between joint, disjoint and Maximum SNR solutions.

4.4.3 Saturation scenario

Constraint (4-20) establishes that every UE is associated with exactly one BS. Under overload conditions, i.e., when aggregate UE demand exceeds network capacity, it may be infeasible to associate every device. A strategy to handle this case is to relax the aforementioned constraint to allow optional admission:

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \sum_{b_l \in \mathcal{B}} x_{l,k}^{p,r} \leq 1, \quad \forall k \in \mathcal{K}, \tag{4-32}$$

which ensures that every UE is associated with at most one BS, while permitting some UEs to remain unserved. To prioritize admitting as many UEs as possible, we include the following admission constraint:

$$\sum_{D_r \in \mathcal{D}} \sum_{p \in \mathcal{P}_l} \sum_{b_l \in \mathcal{B}} \sum_{k \in \mathcal{K}} x_{l,k}^{p,r} = |\mathcal{K}| - i, \tag{4-33}$$

where the parameter i iterates over the interval $[1, |\mathcal{K}| - 1]$ until the model becomes feasible. Since increasing i relaxes the admission constraint, the feasible region expands monotonically with i. Therefore, the first value of i that yields a feasible solution corresponds to the smallest relaxation required to satisfy the constraints, which is equivalent to admitting the maximum possible number of UEs under the model assumptions. This corresponds to a greedy strategy. However, if feasibility verification can be performed efficiently, a binary search procedure could reach the same solution with fewer iterations overall.

To evaluate these overload cases, we use the same topology illustrated in Figure 4.1 and consider two scenarios with different per-UE throughput requirements R^k. Following the observations in [51], the low throughput scenario adopts $R^k = 6.02$ MB/s, consistent with the mean throughput reported for an urban setting, while the high throughput scenario adopts $R^k = 23.72$ MB/s, corresponding to a residential setting.

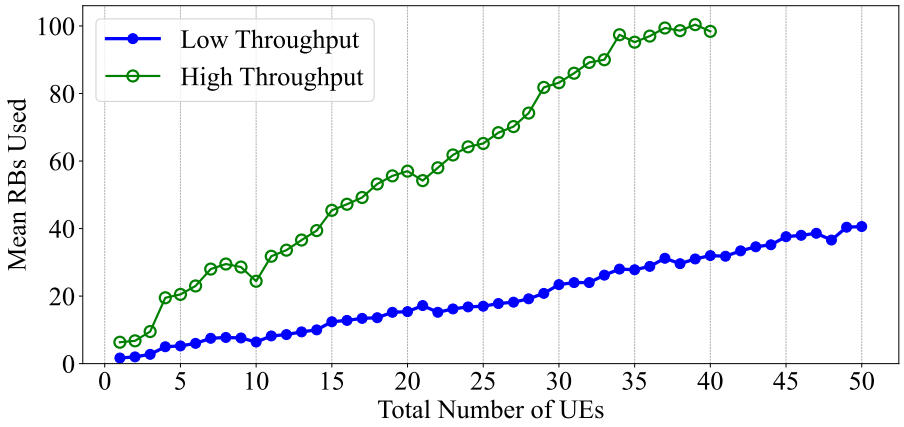


Figure 4.7: Comparison of mean RB usage between high and low UE throughput scenarios.

Figure 4.7 shows the mean RB usage per BS. As expected, ensuring higher per-UE throughput requires a larger number of RBs. As a Consequence, BSs reaches saturation with fewer associated UEs, and the admission rate drops below 100% starting at the instance with 30 UEs onward, as shown in Figure 4.8. Also, under the high throughput scenario, the disjoint optimization model could be solved only up to the instance with 40 UEs within the one-hour time limit, indicating that BS saturation is a relevant factor for the solution time.

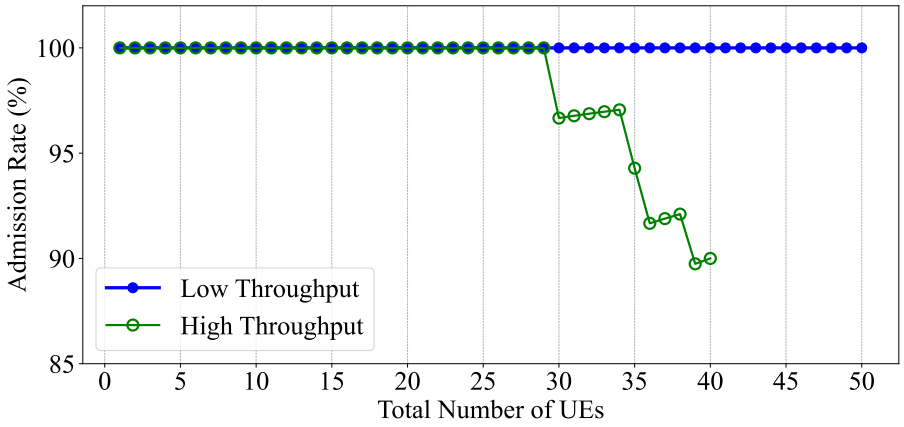


Figure 4.8: Admission rate in high and low UE throughput scenarios.

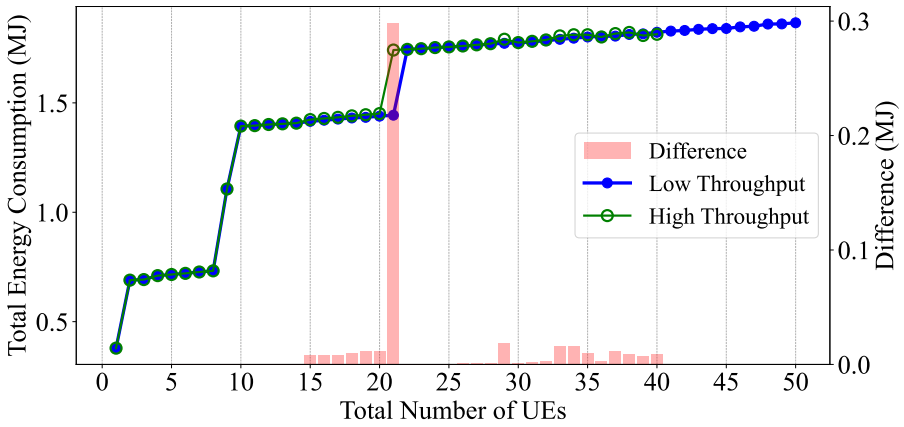


Figure 4.9: Energy efficiency achieved by high and low UE throughput scenarios.

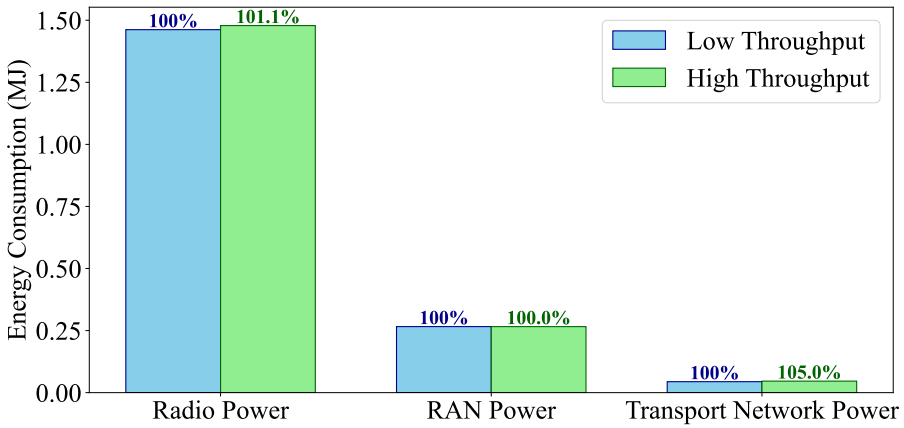


Figure 4.10: Comparison of energy consumption by component for a instance with 29 UEs in low and high throughput scenarios.

Overall, the difference in energy consumption between the two scenarios is small, with the high throughput scenario exhibiting a higher value in the majority of the instances, as depicted in Figure 4.9. The largest gap is observed for the instance with 21 UEs, where the higher traffic load triggers the activation of the fifth BS earlier. The next

largest difference occurs in the instance with 29 UEs. In this case, BSs capacity saturation forces associations under worse SNR conditions, which in turn require a higher BS transmit power configuration. At the same time, the higher traffic load increases transport network utilization. These effects are reflected in Fig 4.10, which shows a higher energy consumption in both radio and transport network equipment, whereas VNF processing energy remains unchanged across the two scenarios.

4.4.4 Energy efficiency

The optimization model formulated in this chapter aims solely to minimize energy consumption, however, it may provide some insights into energy efficiency. In this way, the metric of energy efficiency measured in bits per second per joule (bps/J) is adopted. For each evaluated instance, energy efficiency η^{EE} is computed as the ratio between the achieved aggregate throughput and the total energy consumption of the network, i.e.,

$$\eta^{\text{EE}} = \frac{\sum_{k \in \mathcal{K}} \bar{R}^k}{E_{\text{total}}}, \tag{4-34}$$

where E_{total} is the total energy consumption as computed in Equation (4-16), and $\bar{R}^k$ denotes the throughput effectively delivered to UE k (set to zero for non-admitted UEs), obtained by considering a homogeneous distribution of the RBs available in each BS between the associated UEs.

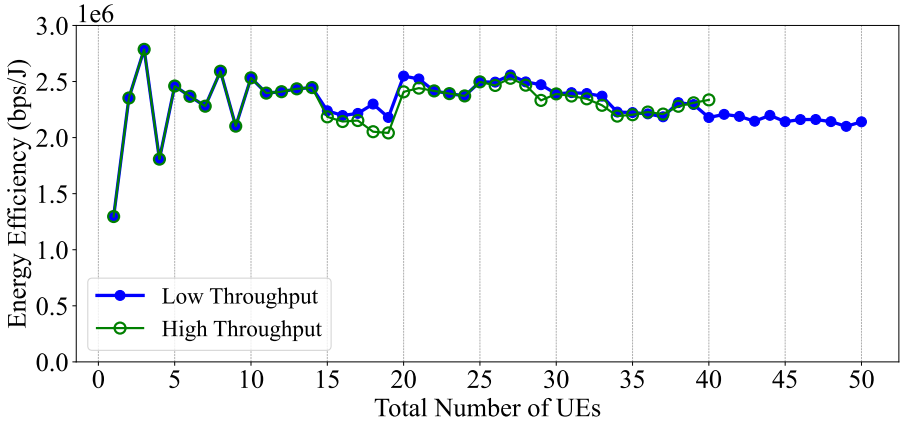


Figure 4.11: Energy efficiency achieved between scenarios with high and low UE throughput requirements.

Figure 4.11 compares energy efficiency under the low and high UE throughput scenarios. The achieved aggregate throughput is primarily determined by the number of active BSs and the mean SINR of the admitted UEs. Since these parameters present low variation between the two scenarios, the observed differences in energy efficiency are

mainly driven by variations in total energy consumption, with higher consumption leading to lower bps/J.

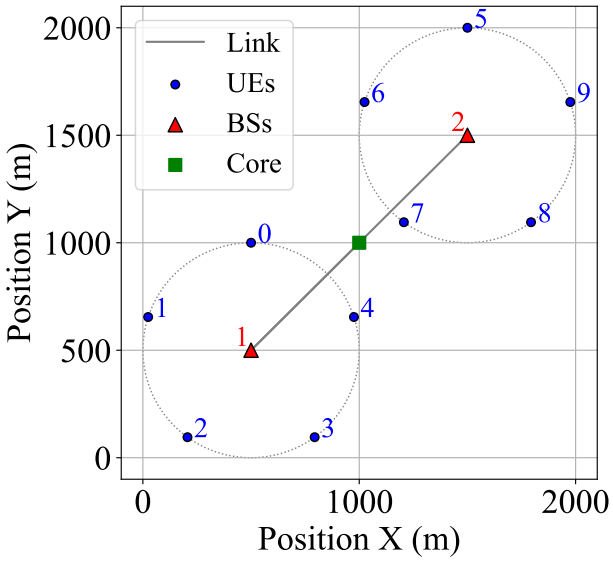


Figure 4.12: Topology used in energy efficiency evaluation with 2 BSs and 10 UEs.

To further examine this relationship, we consider the topology with two BSs presented in Figure 4.12, where five UEs are uniformly placed at a distance of 500 meters from each BS. We focus exclusively on the high throughput requirement. Additionally, as the number of UEs in the instance increases, equipments are added according to their index ordering. In this way, the first BS is saturated before adding UEs close to the second one.

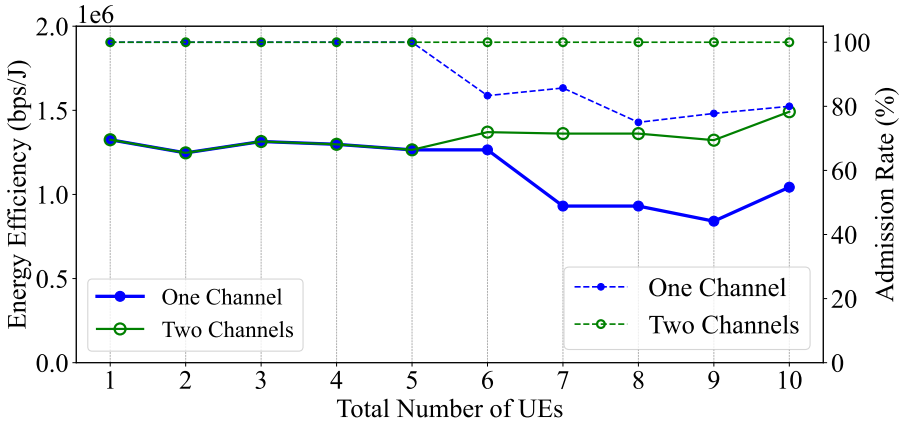


Figure 4.13: Energy efficiency achieved in a scenario with a single shared channel against a scenario with two orthogonal channels.

Figure 4.13 compares two spectrum allocation scenarios. In the first scenario, both BSs operate on the same frequency channel; in the second, each BS is assigned an orthogonal channel. When the channel is shared, admitting users at the second BS

(starting in the instance with 7 UEs) incurs a decrease in the mean SINR of the system. This degradation affects both energy efficiency (solid line) and admission rate (dashed line). In contrast, assigning orthogonal channels substantially reduces or completely eliminates inter-BS interference. The result is an increase in aggregate throughput that outweighs the additional energy consumption from activating the second BS, leading to higher energy efficiency.

Finally, these results outline that a formulation aimed at minimizing total energy consumption does not necessarily yield the most energy-efficient solution. Nevertheless, architectural decisions can significantly affect the energy efficiency achieved by the minimum-energy solution and should, therefore, be considered in mobile network design and evaluation processes.

4.5 Conclusion

This chapter extended the energy-aware orchestration problem by jointly modeling UE–BS association and RAN VNF placement. A direct comparison between joint optimization and the common practice of solving the problems disjointly was performed. The results indicate that, although the joint model tends to achieve lower total energy consumption, the disjoint approach consistently remains close (below a 1% gap) in the evaluated instances, suggesting a practical scalability–optimality trade-off.

Beyond joint versus disjoint comparison, the chapter also showed that a radio-centric heuristic (i.e., maximum-SNR association) can increase the number of active base stations and intensify inter-cell interference, which can degrade feasibility and worsen energy outcomes. Finally, minimizing total energy does not necessarily maximize energy efficiency since, by the definition of the metric, gains in throughput may outweigh increases in energy consumption. Nonetheless, the evaluation of energy efficiency remains relevant, as architectural decisions may impact the performance of the minimum-energy solution.

CHAPTER 5

Final remarks

This dissertation investigated the problem of energy-efficient resource management in virtualized and disaggregated Radio Access Networks, with a focus on the optimization of UE–BS association, BS transmission power configuration, and VNF placement. Motivated by the increasing energy footprint of mobile networks and the architectural flexibility introduced by vRAN paradigm, the research addressed the limitations of existing approaches that do not consider all these optimization decisions in a comprehensive formulation or rely on non-optimal solutions without proper assessment of the optimality gap.

The results demonstrate that energy consumption in vRAN systems is characterized by the interaction among radio equipment, computing resources, and transport networks. Consequently, an effective assessment of energy efficiency requires accounting for all these components. This dissertation shows that it is possible to optimally solve problem instances of considerable size when an efficient formulation and appropriate linearization techniques are employed.

Nevertheless, real-network deployments impose strict time constraints, under which obtaining optimal solutions may remain impractical. Even so, the optimization model is still valuable, as it provides both insights that can guide practical orchestration strategies and establishes a robust baseline for the evaluation of faster non-optimal approaches.

Moreover, the analysis revealed that transport network energy consumption can significantly influence the optimal deployment strategy, particularly under high-throughput UE usage profiles and in topologies with lower transport network efficiency. Additionally, the impact of VNF migration on overall energy consumption is minimal in the evaluated instances, practical deployments should still account for migration costs, since the energy savings achieved by a new configuration may not compensate for the energy overhead incurred during the process of migration to the new deployment. The dissertation also provided a systematic evaluation of how RAN topology design and UE traffic profiles affect energy consumption, highlighting that architectural choices are fundamental in determining whether centralization or distribution is energetically

favorable.

UE–BS association is a key component in energy-aware vRAN operation because it determines how traffic demand is distributed across base stations. This distribution affects radio equipment power consumption and the load generated both on computing resources and on the crosshaul transport network. We compared a joint approach in which association and transmission power configuration are optimized together with VNF placement to a disjoint strategy that first addresses the association problem before optimizing the remaining decisions. although it is expected that disjoint optimization can reduce solution quality by restricting the global search space, the evaluated instances presented only a small gap relative to the joint formulation, while the decoupling of the problem enabled solving larger instances.

The investigations conducted in this dissertation resulted in the publication of the article [42], in the IEEE Open Journal of the Communications Society. This work presents a comprehensive model of energy consumption in O-RAN systems and reports the findings on how diverse network scenarios affect energy consumption. In addition, a text describing our developments on UE-BS association is expected to be submitted for peer review in the near future.

To foster future research and facilitate the reproducibility of the academic work developed in the context of the published article, we make the implementation of the proposed optimization model and the datasets used in the evaluation publicly available in a repository.[1]

While this dissertation contributes insights into energy-efficient vRAN resource management, several directions remain open for future research. To meet the network requirements anticipated for 6G networks, some studies have investigated cell-free massive MIMO (CF-mMIMO) architectures [39]. In CF-mMIMO, each UE is served by a subset of BSs through joint transmission and/or reception, rather than being associated with a single base station. However, supporting CF-mMIMO is expected to require substantial changes in network infrastructure and topology, which may constrain the deployment of dynamic and flexible functional split solutions [28]. In this context, a comprehensive comparison of the energy efficiency of CF-mMIMO architectures against dynamic, flexible functional-split vRAN models remains an open research question. Additionally, extending the heuristic presented in Section 3.2.4 to support CF-mMIMO is a relevant research direction. Its evaluation on larger instances, where the solver may not obtain the optimal solution, would still allow assessing solution quality through the reported optimality gap, providing insight into the heuristic's performance.

As shown in Sub-section 3.3.2, the size of the search space for the problem

[1] https://github.com/LABORA-INF-UFG/paper-WGSCLK-2025

formulated in Section 3.2 is dominated by the number of simple paths. Future work may therefore investigate the impact on solution quality of restricting the candidate set to a reduced subset of paths, such as the shortest paths. In addition, the application of the Column Generation [10] algorithm represents a promising direction for solving larger instances more efficiently.

Regarding the UE-BS association problem, a natural extension is to incorporate temporal dynamics into the optimization framework, explicitly accounting for time-varying traffic patterns and UE mobility. This would enable the evaluation of long-term trade-offs between energy savings and VNF migration overhead. Furthermore, a systematic evaluation of energy minimization strategies based on heuristics and machine-learning approaches, in comparison to the optimal solution, is a promising research direction. In this context, optimal solutions can also be leveraged to guide intelligent model training and validation, providing high-quality supervision and robust performance baselines.

Bibliography

[1] 3GPP. **3GPP TR 38.801: Study on new radio access technology: Radio access architecture and interfaces**. Technical report, 3GPP, 2017.

[2] ABUBAKAR, A. I.; OTHERS. **Energy Efficiency of Open Radio Access Network: A Survey**. In: *Proceedings of the 97th IEEE Vehicular Technology Conference (VTC)*, p. 1–7, August 2023.

[3] ALMEIDA, G. M.; BRUNO, G. Z.; HUFF, A.; HILTUNEN, M.; DUARTE, E. P.; BOTH, C. B.; CARDOSO, K. V. **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*, 42(2):446–459, 2024.

[4] AMIRI, E.; OTHERS. **Energy-Aware Dynamic VNF Splitting in O-RAN Using Deep Reinforcement Learning**. *IEEE Wireless Communications Letters*, 12(11):1891–1895, November 2023.

[5] ARJMANDI, M. K. **5G Overview: Key Technologies**. In: Hu, F., editor, *Opportunities in 5G Networks*, p. 19–32. CRC Press, 2016.

[6] AUER, G.; GIANNINI, V.; DESSET, C.; GODOR, I.; SKILLERMARK, P.; OLSSON, M.; IMRAN, M. A.; SABELLA, D.; GONZALEZ, M. J.; BLUME, O.; FEHSKE, A. **How much energy is needed to run a wireless network?** *IEEE Wireless Communications*, 18(5):40–49, 2011.

[7] DAHLMAN, E.; PARKVALL, S.; SKÖLD, J. **Chapter 1 - what is 5g?** In: Dahlman, E.; Parkvall, S.; Sköld, J., editors, *5G NR: the Next Generation Wireless Access Technology*, p. 1–6. Academic Press, 2018.

[8] DEBAILLIE, B.; DESSET, C.; LOUAGIE, F. **A Flexible and Future-Proof Power Model for Cellular Base Stations**. In: *Proceedings of IEEE Vehicular Technology Conference (VTC Spring)*, p. 1–7, 2015.

[9] DEMIR, O. T.; OTHERS. **Cell-Free Massive MIMO in O-RAN: Energy-Aware Joint Orchestration of Cloud, Fronthaul, and Radio Resources**. *IEEE Journal on Selected Areas in Communications (JSAC)*, 42(2):356–372, January 2024.

[10] DESROSIERS, J.; LÜBBECKE, M. E. **A Primer in Column Generation**, p. 1–32. Springer US, 2005.

[11] FAN, X.; WEBER, W.-D.; BARROSO, L. A. **Power Provisioning for a Warehouse-Sized Computer**. *SIGARCH Computer Architecture News*, 35(2):13–23, June 2007.

[12] FIORANI, M.; OTHERS. **Modeling energy performance of C-RAN with optical transport in 5G network scenarios**. *Journal of Optical Communications and Networking*, 8(11):B21–B34, November 2016.

[13] GAO, D.; XIA, N.; LIU, X.; GAO, L.; WANG, D.; LIU, Y.; PENG, M. **Joint load adjustment and sleep management for virtualized gnbs in computing power networks**. *IEEE Transactions on Wireless Communications*, 24(3):2067–2082, 2025.

[14] GARCIA-SAAVEDRA, A.; OTHERS. **FluidRAN: Optimized vRAN/MEC Orchestration**. In: *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, p. 2366–2374, October 2018.

[15] GARCIA-SAAVEDRA, A.; OTHERS. **WizHaul: On the Centralization Degree of Cloud RAN Next Generation Fronthaul**. *IEEE Transactions on Mobile Computing*, 17(10):2452–2466, October 2018.

[16] GSMA. **5G energy efficiencies: green is the new black (the sequel)**. Technical report, GSMA, Nov 2023.

[17] GUPTA, H.; OTHERS. **Apt-RAN: A Flexible Split-Based 5G RAN to Minimize Energy Consumption and Handovers**. *IEEE Transactions on Network and Service Management*, 17(1):473–487, March 2020.

[18] HABIBI, M. A.; OTHERS. **A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System**. *IEEE Access*, 7:70371–70421, May 2019.

[19] HOJEIJ, H.; RICARDO, G. I.; SHARARA, M.; HOTEIT, S.; VÈQUE, V.; SECCI, S. **On flexible association and placement in disaggregated ran designs**. *Computer Communications*, 238:108166, 2025.

[20] HOYDIS, J.; CAMMERER, S.; AIT AOUDIA, F.; NIMIER-DAVID, M.; MAGGI, L.; MARCUS, G.; VEM, A.; KELLER, A. **Sionna**, 2022.

[21] ITU-R. **IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond**. Technical Report M.2083-0, International Telecommunication Union, 2015.

[22] ITU-R. **Framework and overall objectives of the future development of IMT for 2030 and beyond**. Technical Report M.2160-0, International Telecommunication Union, 2023.

[23] ITU-T, T. S. S. **Characteristics of transport networks to support IMT-2020/5G**. ITU-T, 2020.

[24] JODA, R.; PAMUKLU, T.; ITURRIA-RIVERA, P. E.; EROL-KANTARCI, M. **Deep reinforcement learning-based joint user association and cu–du placement in o-ran**. *IEEE Transactions on Network and Service Management*, 19(4):4097–4110, 2022.

[25] KARBALAEE MOTALLEB, M.; SHAH-MANSOURI, V.; PARSAEEFARD, S.; ALCARAZ LÓPEZ, O. L. **Resource allocation in an open ran system using network slicing**. *IEEE Transactions on Network and Service Management*, 20(1):471–485, 2023.

[26] KAZUNARI, K. **Approach to Commercial Use of OAI**, 2017. https://openairinterface.org/4th-openairinterface-workshop-fall-2017.

[27] L. LOPES, V. H.; ALMEIDA, G. M.; KLAUTAU, A.; CARDOSO, K. **A coverage-aware vnf placement and resource allocation approach for disaggregated vrans**. In: *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, p. 185–190, 2022.

[28] LARRAÑAGA, A.; LAGÉN, S.; FÀBREGA, J. M.; RIVAS-MOSCOSO, J. M.; FERNÁNDEZ-PALACIOS, J. P.; TOMKOS, I.; MUÑOZ, R. **Fronthaul/midhaul networks: Capacity and latency requirements imposed by 6g disaggregated rans**. *IEEE Communications Magazine*, 63(5):86–93, 2025.

[29] LARSEN, L. M. P.; CHECKO, A.; CHRISTIANSEN, H. L. **A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks**. *IEEE Communications Surveys & Tutorials*, 21(1):146–172, Firstquarter 2019.

[30] LIU, H.; OTHERS. **Performance and Energy Modeling for Live Migration of Virtual Machines**. In: *Proceedings of the 20th International Symposium on High Performance Distributed Computing*, p. 171–182, June 2011.

[31] LIU, X.; ZHANG, H.; LONG, K.; NALLANATHAN, A.; LEUNG, V. C. M. **Energy efficient user association, resource allocation and caching deployment in fog radio access networks**. *IEEE Transactions on Vehicular Technology*, 71(2):1846–1856, 2022.

[32] LOPES, V. H. L.; ALMEIDA, G. M.; KLAUTAU, A.; CARDOSO, K. V. **O-ran-oriented approach for dynamic vnf placement focused on interference mitigation**. In: *ICC 2024 - IEEE International Conference on Communications*, p. 5479–5484, 2024.

[33] MALANDRINO, F.; OTHERS. **An Optimization-Enhanced MANO for Energy-Efficient 5G Networks**. *IEEE/ACM Transactions on Networking*, 27(4):1756–1769, August 2019.

[34] MOLLAHASANI, S.; OTHERS. **Energy-Aware Dynamic DU Selection and NF Relocation in O-RAN Using Actor-Critic Learning**. *Sensors*, 22(13), July 2022.

[35] MORAIS, F. Z.; OTHERS. **PlaceRAN: Optimal Placement of Virtualized Network Functions in Beyond 5G Radio Access Networks**. *IEEE Transactions on Mobile Computing*, 22(9):5434–5448, September 2023.

[36] MOREIRA ZORELLO, L. M.; OTHERS. **Power-Efficient Baseband-Function Placement in Latency-Constrained 5G Metro Access**. *IEEE Transactions on Green Communications and Networking*, 6(3):1683–1696, September 2022.

[37] MURTI, F. W.; OTHERS. **An Optimal Deployment Framework for Multi-Cloud Virtualized Radio Access Networks**. *IEEE Transactions on Wireless Communications*, 20(4):2251–2265, April 2021.

[38] NGMN. **NGMN Overview on 5 G RAN Functional Decomposition**, 2018.

[39] NGO, H. Q.; INTERDONATO, G.; LARSSON, E. G.; CAIRE, G.; ANDREWS, J. G. **Ultra-dense cell-free massive mimo for 6g: Technical overview and open questions**. *Proceedings of the IEEE*, 112(7):805–831, 2024.

[40] NOKIA. **Reducing energy use with 5G-Advanced: The essential guide to RAN energy savings in 3GPP Release 18**. Technical report, Nokia, 2023.

[41] PIRES, W.; OTHERS. **Bi-objective Optimization for Energy Efficiency and Centralization Level in Virtualized RAN**. In: *Proceedings of IEEE International Conference on Communications (ICC)*, p. 1034–1039, August 2022.

[42] PIRES-JR, W. T.; ALMEIDA, G. M.; CORRÊA, S. L.; BOTH, C. B.; PINTO, L. L.; CARDOSO, K. V. **Optimizing energy consumption for vran placement in o-ran systems with flexible transport networks**. *IEEE Open Journal of the Communications Society*, 6:4279–4294, 2025.

[43] RAZA, M. R.; OTHERS. **Power and cost modeling for 5G transport networks**. In: *Proceedings of the 17th International Conference on Transparent Optical Networks (ICTON)*, p. 1–7, August 2015.

[44] ROCHA, F. G. C.; OTHERS. **Optimal Resource Allocation with Delay Guarantees for Network Slicing in Disaggregated RAN**, June 2023.

[45] SEN, N.; FRANKLIN A., A. **Towards Energy Efficient Functional Split and Baseband Function Placement for 5G RAN**. In: *Proceedings of International Conference on Network Softwarization (NetSoft)*, p. 237–241, July 2023.

[46] SINGH, R.; OTHERS. **Energy-Efficient Orchestration of Metro-Scale 5G Radio Access Networks**. In: *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, p. 1–10, July 2021.

[47] SUN, Y.; YU, K.; XU, Y.; ZHOU, H.; XUEMIN.; SHEN. **Flexible base station sleeping and resource cooperation enabled green fully-decoupled ran**, 2023.

[48] TESFALIDET, N.; KHOSRAVI, S. **Development of a C-RAN Fronthaul Simulator**. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2023.

[49] ZAFAR, H.; TOHIDI, E.; KASPARICK, M.; STAŃCZAK, S. **Load balancing in o-ran**. In: *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, p. 1–6, 2024.

[50] ZAIDI, A. A.; BALDEMAIR, R.; TULLBERG, H.; BJORKEGREN, H.; SUNDSTROM, L.; MEDBO, J.; KILINC, C.; DA SILVA, I. **Waveform and numerology to support 5g services and requirements**. *IEEE Communications Magazine*, 54(11):90–98, 2016.

[51] ZHANG, S.; CAI, T.; DEMIR, O. T.; CAVDAR, C. **Multi-agent rl for sleep mode and antenna configuration with user offloading under dynamic traffic in massive mimo networks**. *IEEE Transactions on Vehicular Technology*, 74(6):9734–9749, 2025.

[52] ZHANG, Z.; MARDER, A.; MOK, R.; HUFFAKER, B.; LUCKIE, M.; CLAFFY, K. C.; SCHULMAN, A. **Inferring regional access network topologies: methods and applications**. In: *Proceedings of the 21st ACM Internet Measurement Conference*, IMC '21, p. 720–738, New York, NY, USA, 2021. Association for Computing Machinery.