UNICO I+D Project 6G-SORUS-RAN

SORUS-RAN-A2.3-E1 (E11)

Algorithms for network optimization: selection and design

Abstract

This deliverable presents an in-depth analysis of the state of the art in resource allocation and energy-aware design for virtualized Radio Access Networks (vRAN). We review and categorize existing approaches, focusing on AI-driven optimization techniques, mobility modeling, and energy-efficient infrastructure management, highlighting their strengths, limitations, and open challenges. Building upon this analysis, we outline three key research directions: (i) the development of cost-aware autoscaling algorithms for reliable and energy-efficient vRAN server farms; (ii) the design of privacy-preserving generative models for mobility-driven resource planning; and (iii) the evaluation of throughput gains achievable through dynamic spatial and temporal vRAN reconfiguration. Together, these directions provide a foundation for advancing intelligent, adaptive, and energy-efficient vRAN solutions that integrate algorithmic innovation, privacy-aware modeling, and performance evaluation.

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

Document properties

Document number	SORUS-RAN-A2.3-E1 (E11)
Document title	Algorithms for network optimization: selection and design
Document responsible	Marco Gramaglia
Document editor	Juan Manuel Montes
Editorial team	Marco Gramaglia, Juan Manuel Montes
Target dissemination level	Public
Status of the document	Final
Version	1.0
Delivery date	31-12-2023
Actual delivery date	31-12-2023

Production properties

Reviewers	Pablo Serrano, María Molina

Disclaimer

This document has been produced in the context of the SORUS-RAN Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme.

Contents

List of Figures

List of Acronyms

5GC: 5G Core

AI: Artificial Intelligence

API: Application Prorgamming Interface

ARIMA: Autoregressive Integrated Moving Average

B5G: Beyond 5G

DT: Digital Twin

IoT: Internet of Things

LSTM: Long Short Term Memory

ML: Machine Learning

NR: New Radio

RU: Radio Unit

SFC: Service Function Chain

SDN: Software Defined Network

SQL: Structured Query Language

PNF: Physical Network Function

VLAN: Virtual Local Area Network

VNF: Virtual Network Function

Resumen Ejecutivo

Este entregable presenta un análisis exhaustivo del estado del arte en las redes de acceso radio virtualizadas (vRAN), abordando los aspectos clave que determinan su rendimiento, eficiencia y viabilidad operativa. A medida que las redes móviles evolucionan hacia arquitecturas flexibles, definidas por software y basadas en la nube, las soluciones vRAN se consolidan como una tecnología fundamental para satisfacer la creciente demanda de redes escalables y rentables. No obstante, la implantación y gestión de sistemas vRAN conllevan nuevos desafíos que deben ser abordados para garantizar su funcionamiento óptimo.

En este documento, se revisan y analizan las soluciones y avances más relevantes en cinco áreas fundamentales: asignación de recursos, evaluación del rendimiento, eficiencia energética, tolerancia a fallos y fiabilidad. Se examina cómo las estrategias avanzadas de asignación de recursos pueden mejorar la eficiencia de las infraestructuras virtualizadas, garantizando al mismo tiempo la calidad del servicio. Asimismo, se ofrece una visión general de las metodologías actuales para la evaluación del rendimiento de vRAN, destacando las métricas y herramientas utilizadas tanto en la industria como en el ámbito académico. Además, se analiza la importancia de los mecanismos y diseños orientados a la eficiencia energética, esenciales para reducir los costes operativos y el impacto medioambiental en despliegues a gran escala. El entregable también explora técnicas destinadas a reforzar la tolerancia a fallos y la fiabilidad de los sistemas vRAN, asegurando la continuidad del servicio ante posibles fallos de hardware o software.

Mediante la recopilación y el análisis de los avances más recientes en estas áreas, este entregable proporciona una base sólida para identificar futuras líneas de investigación y orientar el desarrollo de soluciones vRAN de próxima generación.

Executive Summary

This deliverable presents a comprehensive analysis of the current state of the art in virtualized Radio Access Networks (vRAN), focusing on the key aspects that define their performance, efficiency, and operational viability. As mobile networks evolve towards flexible, software-defined, and cloud-native architectures, vRAN emerges as a cornerstone technology to meet the growing demand for scalable and cost-effective network solutions. However, the deployment and management of vRAN systems introduce new challenges that must be addressed to ensure their optimal performance.

In this document, we review and analyze existing solutions and research efforts across five critical areas: resource allocation, performance evaluation, energy efficiency, fault tolerance, and reliability. We examine how advanced resource allocation strategies can improve the efficiency of virtualized infrastructures while ensuring service quality. We also provide an overview of current methodologies for evaluating vRAN performance, highlighting the metrics and tools commonly used in the industry and academia. Additionally, we assess the importance of energy-efficient designs and mechanisms, which are essential to reduce operational costs and environmental impact in large-scale deployments. The deliverable further explores techniques to enhance the fault tolerance and reliability of vRAN systems, ensuring continuous service delivery in the presence of hardware or software failures.

By compiling and discussing the latest advancements in these areas, this deliverable provides a solid foundation for identifying future research directions and guiding the development of next-generation vRAN solutions.

1. Introduction

This deliverable investigates the algorithmic challenges and current solutions in the virtualization of Beyond 5G (B5G) networks. In this context, network virtualization is an essential framework that enables flexible and efficient network resource management, which is critical to meet the demands of increasingly diverse digital services.

There exist several fundamental aspects of network virtualization, including:

- Resource Optimization: Effective use of network resources is a major priority. Approaches in this area focus on dynamically adjusting resources based on demand to ensure optimal performance, often using data-driven strategies to predict usage patterns and allocate resources accordingly.
- Network Flexibility: Flexibility is at the core of virtualized networks, allowing network functions to be deployed independently of specific hardware. This enables faster adaptation to changing network demands, which is particularly useful in applications like real-time gaming and augmented reality (AR), where low latency is essential.
- Evolving Network Architectures: Modern architectures are built on cloud-based principles, allowing network functions to operate across distributed data centers. This adaptability supports applications requiring rapid response times, like connected vehicles and IoT ecosystems, enhancing the capacity of networks to meet varying service requirements.
- Network Slicing for Specialized Services: Network slicing allows the network to be divided into virtual segments, each tailored to specific application needs. In B5G networks, slices are often dedicated to services with stringent performance needs, such as low-latency communication for remote surgery or high-bandwidth streams for HD video.
- Interoperability and Efficiency: Virtualized networks aim to work seamlessly with existing technologies while adopting efficient protocols for diverse device requirements. Techniques such as centralized traffic control are essential to optimize the user experience and ensure smooth integration across different technologies.
- Energy Efficiency: With the expansion of network infrastructure, energy-efficient operations have become crucial. Protocols that manage network resources based on demand, along with predictive strategies for usage patterns, contribute to reducing energy consumption in large-scale deployments.

The push for virtualization reflects a broader transformation in telecommunications, driven by a rising demand for adaptable and reliable wireless services. The flexibility gained by virtualization supports a range of applications, from high-definition streaming to IoT, autonomous vehicles, and beyond, all of which require unique levels of bandwidth, latency, and reliability. In the context of 5G and B5G, virtualization represents a shift towards a dynamic, software-driven infrastructure that can adapt to future demands. As emerging technologies develop, this virtualized foundation will support both

current applications and new advancements, ultimately fostering a more resilient, adaptable network environment.

2. Background

Understanding the virtualization of B5G networks necessitates a thorough examination of the technologies and architectures that underpin modern wireless communication systems. This section provides a foundational overview of the key components that have paved the way for network virtualization. We begin by exploring traditional network solutions, highlighting their limitations in the context of today's dynamic service demands. This sets the stage for discussing **Virtualized Radio Access Networks (vRAN)** and their architectures, which represent a significant shift towards more flexible and scalable network designs. Additionally, we delve into **Network Function Virtualization (NFV)**, a crucial technology that decouples network functions from proprietary hardware, enabling more agile deployment and management of network services. By examining these elements, we establish a comprehensive background that informs the subsequent analysis of virtualization challenges and solutions in B5G networks.

2.1. vRAN arquitecture

The architecture of vRAN is composed of distinct layers and components that work in harmony to enable flexibility, scalability, and efficient management of network resources. Unlike traditional RAN architectures, vRAN separates hardware and software functions, allowing centralized and distributed processing for enhanced performance. Here, we break down the core elements of vRAN and their roles within the network.

Key Components of vRAN Architecture

1. **Radio Units (RUs)**: The RUs are the physical components at the edge of the network responsible for all radio frequency (RF) operations, including transmission, reception, and beamforming. The RUs house antennas and other RF equipment and are positioned directly at the cell sites, near the antennas. This direct proximity ensures efficient signal transmission and reception, maximizing signal strength and reducing path loss. The RUs convert analog radio signals to digital for processing by the DUs, and vice versa for transmission, while supporting functions such as MIMO (Multiple Input Multiple Output) to improve capacity and coverage.
2. **Distributed Unit (DU)**: Positioned closer to the cell site, the DU manages real-time, lower-level baseband processing tasks, including encoding, decoding, modulation, and demodulation. It interfaces directly with the Radio RUs and executes lower-layer protocol functions, such as those for the RLC, MAC, and parts of the physical layer. The DU's proximity to the RUs minimizes latency, which is crucial for real-time tasks like beamforming and signal processing for multiple access. DUs are often deployed across edge servers in distributed data centers, enabling them to efficiently handle high volumes of traffic.
3. **Central Unit (CU)**: The CU is responsible for high-level network protocols and logical processing. It oversees session management, mobility management, and aggregation of traffic from multiple DUs. It acts as the command center for coordinating resources and

managing the upper layer of the protocol stack (RRC, PDCP, and SDAP), enabling seamless handovers, resource allocation, and traffic control. CUs are typically hosted in regional data centers to centralize high-level control while managing a range of DUs, optimizing resource use and allowing network operators to scale coverage efficiently.

4. **Fronthaul**: The fronthaul is the communication link between the RUs and DUs, transmitting digitized radio signals for processing. The fronthaul must provide high bandwidth and low latency connectivity to accommodate the large data volumes and strict timing requirements of modern wireless networks. Typical implementations use technologies such as CPRI (Common Public Radio Interface) or eCPRI, which support the efficient transport of baseband data. For instance, eCPRI provides more flexibility in split processing and reduced bandwidth requirements compared to traditional CPRI, making it well-suited for dense urban environments with high traffic demands.

The distribution of these components is strategic. **CUs** are generally hosted in centralized or regional data centers, from where they can oversee multiple DUs. **DUs** are deployed closer to **RUs** to minimize latency in processing and ensure rapid response times. **RUs** are positioned directly at cell sites, close to antennas, for efficient signal transmission and reception. This arrangement optimizes both resource allocation and data flow throughout the network.

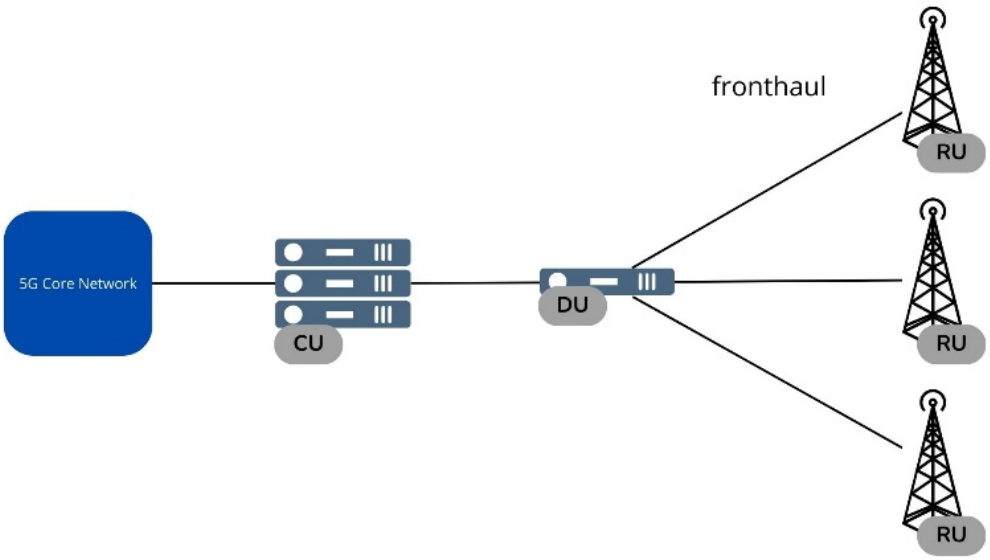

FIGURE 1: SIMPLIFIED VISION OF THE VRAN ARCHITECTURE.

The operation of vRAN involves several distinct processes that together ensure efficient and flexible network performance. Key operational steps include:

1. **Signal Processing**: The RU receives analog radio signals, which it converts to digital if necessary. Initial processing, such as filtering and amplification, takes place at the RU before sending the signal over the fronthaul to the DU for further processing. The conversion from

analog to digital format allows precise signal manipulation and is critical for supporting high-speed, high-capacity wireless communications.

2. **Baseband Processing**: Once signals reach the DU, they undergo intensive baseband processing tasks, including error correction, modulation/demodulation, and multiple access processing (e.g., OFDMA for 5G). DUs are capable of pooling resources across multiple RUs, optimizing processing power by sharing computational resources and balancing load demands. This setup allows for dynamic scaling to accommodate fluctuations in traffic and ensures efficient bandwidth utilization, which is essential for high-performance applications.
3. **Resource Management**: The CU orchestrates resource allocation across the network, managing both DUs and RUs. By leveraging virtualization, the CU dynamically allocates resources based on real-time demand and network conditions, ensuring that each component performs optimally. For example, if one DU experiences peak traffic, resources can be shifted to support it, balancing the load and reducing potential bottlenecks. This adaptive resource management helps operators maximize network efficiency and reliability, especially during peak usage periods.
4. **Network Slicing and Management**: One of the standout capabilities of vRAN is its support for network slicing. The CU enables the creation of virtual network slices, each tailored to the requirements of specific applications or services. For instance, one slice may be optimized for IoT with low power consumption, another for enhanced Mobile Broadband (eMBB) with high throughput for video streaming, and yet another for Ultra-Reliable Low Latency Communications (URLLC) for applications like autonomous driving. Network slicing ensures that each service receives the appropriate quality of service (QoS) and priority, enabling operators to support a wide range of applications within a single infrastructure.

2.2. Network Function Virtualization (NFV)

Network Function Virtualization, NFVs, is a key enabler in the modernization of network infrastructures, particularly in the deployment of vRAN architectures. NFV transforms the traditional, hardware-based network by decoupling network functions from proprietary hardware and implementing them as software instances on general-purpose servers. This shift enables greater flexibility, scalability, and efficiency by allowing network operators to deploy and manage functions dynamically based on real-time demand. In a vRAN setup, NFV is crucial to delivering adaptable network services that can support the diverse needs of B5G applications.

NFV architecture typically includes three main components: Virtual VNFs, the NFV Infrastructure (NFVI), and the Management and Orchestration (MANO) framework. Together, these components support the modular deployment, orchestration, and scaling of virtualized network functions across vRAN's distributed architecture.

Virtual Network Functions (VNFs)

VNFs represent the individual network functions traditionally performed by specialized hardware, now implemented as software processes. In the context of vRAN, VNFs may include baseband processing functions (for encoding, decoding, modulation, and demodulation), mobility management, and load balancing.

For instance, the DU in vRAN could be virtualized as a VNF handling intensive real-time baseband processing tasks. This flexibility allows multiple DUs to be deployed on general-purpose servers close to the RUs, which enhances scalability and allows processing resources to be shared and dynamically allocated based on traffic conditions. Similarly, virtualizing the CU as a VNF enables the centralized control of network slices and resource management, scaling these functions as user demand fluctuates. A VNF might manage handovers between cells by dynamically allocating resources to optimize for user mobility. This approach ensures seamless connectivity for users moving through high-demand areas, like city centers, without requiring additional dedicated hardware.

NFV Infrastructure (NFVI)

NFVI provides the underlying hardware and virtualization layer that supports the deployment of VNFs. This includes **compute, storage, and networking resources** spread across centralized data centers and distributed edge sites. NFVI relies on virtualization technologies, such as hypervisors and container platforms, to manage resource allocation and ensure efficient deployment of VNFs across the infrastructure.

Within a vRAN setup, the NFVI enables operators to position resources where they are most effective. For instance, DUs can be deployed on edge servers to minimize latency for real-time processing, while CUs can be in regional data centers to handle more computationally intensive tasks. Suppose a sudden surge in demand occurs at a particular cell site. The NFVI, through virtualization, allows for additional DUs to be instantiated on nearby edge servers, boosting capacity and maintaining service quality without the need to physically provision new hardware.

Management and Orchestration (MANO)

The MANO framework oversees the lifecycle of VNFs, handling tasks such as deployment, scaling, fault management, and resource orchestration. MANO ensures that VNFs operate in concert with one another, dynamically adjusting resources across the network to meet real-time demand and operational requirements.

In vRAN, MANO is essential for managing the complex relationships between CUs, DUs, and RUs. It coordinates VNFs to maintain seamless service continuity, enabling network slices to be created, modified, and removed based on service needs. For instance, MANO might instantiate a new slice dedicated to a particular application, such as eMBB, which requires high throughput and low latency, and release the resources once demand diminishes. If a critical latency-sensitive application like

URLLC is deployed, MANO can adjust the resources allocated to that slice, reducing latency at the DU and prioritizing connections at the CU. This allows the network to respond rapidly to changes, such as a peak in URLLC traffic, ensuring that essential services receive the necessary resources.

Benefits

By integrating NFV, vRAN networks gain several advantages:

1. **Scalability and Flexibility**: NFV enables vRAN to dynamically allocate resources across CUs, DUs, and RUs, adapting to fluctuations in traffic patterns and supporting services like network slicing for customized service delivery. For instance, as user traffic rises in specific areas (e.g., near stadiums during events), new DUs can be deployed virtually to manage the surge, with the VNFs spun up or down as needed.
2. **Resource Efficiency**: With NFV, network resources are pooled and shared, maximizing their use and minimizing idle resources. This pooling allows for **multi-tenancy**, where resources can be split between services, such as IoT, which requires low power, and eMBB for high-speed data, within a single infrastructure. For example, a single DU may simultaneously serve multiple RUs to manage various service types in one region.
3. **Reduced Operational Costs**: Virtualizing network functions reduces the reliance on proprietary hardware, enabling operators to deploy standardized, general-purpose hardware. This reduction in hardware dependency decreases CAPEX and OPEX, making it feasible to expand the network to meet growing demand while maintaining budgetary constraints.
4. **Rapid Service Deployment**: NFV allows operators to launch new services quickly without physically modifying the infrastructure. This capability is crucial for supporting emerging B5G applications that require unique network configurations, such as autonomous vehicles or remote healthcare. For instance, a network operator could launch a low-latency slice dedicated to URLLC applications without altering the physical infrastructure.
5. **Improved Fault Tolerance**: With NFV, network functions can be migrated or scaled to different parts of the network in response to hardware failures or overloads. If a DU VNF experiences a fault, MANO can automatically re-deploy it to a different edge server, ensuring service continuity. Additionally, VNFs can be backed up in multiple locations, reducing the risk of service disruptions.
6. **Enhanced Innovation**: The NFV framework fosters a modular approach, allowing new VNFs to be added and tested without impacting existing network functions. This modularity encourages the development and deployment of innovative network services, such as machine learning-driven analytics for traffic forecasting or security VNFs to detect and respond to cyber threats in real-time.

2.3. Challenges

The widespread adoption of virtualization across various domains, including computing environments and wireless networks, do not come without issues. These challenges stem from the need to efficiently manage virtualized resources, ensure performance and reliability, and provide security and isolation among virtual entities. Addressing these challenges is crucial for maximizing the benefits of virtualization while minimizing potential drawbacks.

The main challenges tackled are, among others:

1. **Resource Allocation and Scheduling**: Efficiently allocating and scheduling computing resources such as CPU time, memory, and network bandwidth among multiple virtual machines or network functions is a significant challenge [GB2016]. Algorithms must balance the demand for resources from virtual entities against the available supply, while also considering the priorities and performance requirements of different tasks. This involves complex optimization problems that must account for the dynamic nature of workloads and the heterogeneity of underlying hardware.
2. **Performance Isolation**: Ensuring that the activities of one virtual entity do not adversely affect the performance of others is critical. However, achieving strict performance isolation is challenging due to shared underlying hardware resources [SC2009]. Algorithmic solutions are required to enforce fair sharing of resources and prevent any single virtual machine or function from monopolizing resources to the detriment of others. This challenge is compounded by the need to maintain high utilization of physical resources to reap the cost benefits of virtualization.
3. **Energy Efficiency**: With the increasing emphasis on sustainability and cost-saving, optimizing the energy consumption of virtualized infrastructures is a paramount concern. Algorithms need to not only allocate resources efficiently but also minimize energy usage without compromising service quality. This involves dynamically adjusting the operational states of physical resources based on the workload and employing techniques like server consolidation to reduce idle power consumption. [PBS2023]
4. **Fault Tolerance and Reliability**: Ensuring high availability and reliability in a virtualized environment requires sophisticated algorithms for fault detection, recovery, and migration of virtual entities. This includes detecting hardware or software failures, seamlessly migrating virtual machines or functions to healthy resources, and managing state consistency across migrations. The challenge is to perform these operations with minimal impact on performance and service continuity. [PBS2023]
5. **Security and Privacy**: Virtualization introduces new security challenges, particularly related to multi-tenancy and the shared use of physical resources. Algorithms are needed to ensure data isolation and privacy, protect against attacks that exploit the shared infrastructure, and manage secure access to virtualized resources. This includes techniques for secure virtual machine introspection, encryption of data in motion and at rest, and anomaly detection within virtualized environments [MBG2024].

Addressing these challenges requires a combination of cutting-edge approaches, advanced optimization techniques, and leveraging complex machine learning and artificial intelligence methodologies.

3. Traditional solutions

In vRAN architectures, a variety of algorithmic strategies are utilized to handle diverse tasks ranging from resource allocation to signal processing and user scheduling. These algorithms can be broadly classified into several categories based on their function and the specific challenges they address in the vRAN context.

- Baseline methods in the context of resource allocation, energy efficiency, and performance isolation for vRANs typically feature straightforward, less dynamic approaches compared to modern, sophisticated algorithms. For resource allocation, methods like static allocation and round-robin scheduling are commonly used. Static allocation assigns resources based on predicted maximum demands, often resulting in inefficiencies due to either resource wastage or inadequate provisioning. Round-robin scheduling, distributing resources in a fixed cyclic order, doesn't account for the variable demands and priorities, leading to suboptimal network performance.
- Performance isolation is another critical area in vRAN management, traditionally handled by methods such as resource partitioning, where resources are divided statically among services or tenants. This method guarantees that activities in one partition don’t affect others but can lead to underutilization or shortages due to its inflexible nature. Virtual Machines (VMs) and containers also serve as baseline techniques, providing strong isolation at the cost of increased resource overhead, especially with VMs which duplicate entire operating systems for each instance. Additionally, implementing Quality of Service (QoS) rules can manage resource access and network traffic based on predefined policies, but often lacks dynamic adaptability to changing network conditions.
- Greedy algorithms, threshold-based systems, and simple machine learning models like linear regression represent other baseline approaches. Greedy algorithms optimize immediate costs without considering long-term effects, while threshold-based methods activate resources based on predefined usage metrics, responding to changes with potential delays. Rule-based systems operate under static rules that do not adapt unless manually revised, and simple machine learning models, although slightly more adaptive, usually fail to handle the complex, real-time demands of modern vRAN environments effectively. Compared to these baseline methods, more advanced techniques such as Deep Reinforcement Learning, or multi-agent systems offer significant improvements by dynamically optimizing resource allocation and energy usage to accommodate the evolving network landscape.

4. Algorithms for vRAN optimization

In this section we present a detailed summary of the different solutions for addressing the challenges of virtualization presented in section 2, these solutions encompass various multidisciplinary strategies and approaches and aim to tackle the complexities of resource allocation, performance isolation, energy efficiency, network function placement, fault tolerance, and security within virtualized environments. As we will see, some of these challenges can be jointly addressed.

Resource allocation and scheduling		
Reference	Methods	Summary
[BGF2020]	Deep Convolutional Neural Networks	The paper discusses the dynamic management of network resources in multi-tenant mobile networks through network slicing, focusing on precise capacity allocation to address future, variable service demands. It highlights the economic consequences of both overprovisioning and underprovisioning resources, emphasizing the need for optimized resource distribution that effectively balances cost and performance. To solve these issues, the authors propose DeepCog, a deep convolutional neural network architecture inspired by image processing, equipped with a cost-aware capacity forecasting loss function. This model enables network operators to make resource allocation decisions that minimize management costs and maximize revenue. The effectiveness of DeepCog is validated by extensive real-world data tests from a metropolitan mobile network, demonstrating a potential reduction in resource management costs by over 50%.
[JHD2019]	Graph Neural Networks and Deep Reinforcement Learning	The paper explores resource allocation for NFV using a Deep Reinforcement Learning (DRL) approach, mindful of network topology to handle the unpredictable fluctuations in network traffic and requests. This study focuses on the Service Function Chains (SFC) in NFV applications, which require dynamic scaling due to varying demands. To address these issues, the authors introduce an Asynchronous DRL-enhanced Graph Neural Network (GNN) model designed for topology-aware VNF resource prediction. This model not only models the netwoSrk topology using GNN but also uses DRL agents to

		adaptively learn and apply resource allocation policies. The effectiveness of this approach is demonstrated through simulations, showing significant improvements in prediction accuracy and proactive scaling, thus optimizing resource utilization, reducing costs, and enhancing service reliability without sacrificing efficiency.
[ZWQ2013]	Heuristic algorithm	The paper discusses the optimization and scaling of resources in cloud computing environments based on varying business needs. This scaling is targeted at addressing server underutilization in data centers and meeting the scalability and flexibility demands of different applications. The paper emphasizes the need to strike a balance between system performance, resource utilization, and energy efficiency. To address these challenges, the authors propose a novel method centered around "skewness," a metric for assessing the imbalance in the multidimensional resource utilization on servers. By minimizing skewness, the system can co-locate workloads with varying resource demands, in a more efficient way. The efficacy of this approach is validated through simulations and experimental results, demonstrating enhancements in resource efficiency, system performance, and significant energy savings.
[AGG2019]	Autoencoder and Deep Deterministic Policy Gradient with actor-critic	The work examines the challenges associated with dynamically managing computing and radio resources in vRANs. To tackle these challenges, the is proposed vrAIn, a system that incorporates two advanced deep learning strategies: an autoencoder and a Deep Deterministic Policy Gradient (DDPG) algorithm. The autoencoder simplifies high-dimensional context data, such as traffic flows and channel quality, into a more manageable latent representation. Simultaneously, the DDPG algorithm, leveraging an actor-critic neural network architecture, utilizes this processed data to make precise resource control decisions aimed at optimizing computing capacity and enhancing Quality of QoS. The effectiveness of vrAIn is proven through rigorous testing on an open-source LTE stack and various simulation platforms, showcasing substantial improvements in vRAN

		resource management. This includes up to 30% savings in computing capacity, a 25% better probability of meeting QoS targets compared to static methods.
[AGC2023]	Deep Deterministic Policy Gradient (DDPG) and k-nn.	The paper introduces the ATHENA framework, which leverages machine learning for radio resource scheduling to optimize network performance in vRANs. The framework dynamically adjusts resource allocations based on real-time network conditions. ATHENA leverages contextual data, such as the signal-to-noise ratio (SNR) of user equipment and the current levels of system congestion, to make informed resource scheduling decisions. This includes the dynamic allocation of radio resources like Physical Resource Blocks (PRBs) and Modulation and Coding Schemes (MCS). Moreover, the framework incorporates a machine reasoning (MR) component to analyze and interpret the decisions made by the ML model, ensuring that resource allocation is both effective and justifiable. The practical implementation of ATHENA in a real software-based vRAN environment shows it significantly outperforming standard radio resource controllers, demonstrating its effectiveness and potential for broader application.

These works collectively address the challenges of resource allocation and scheduling in dynamic and complex network environments, offering innovative solutions through deep learning, machine learning, and optimization techniques. Despite their different focuses, several common themes emerge, particularly the emphasis on optimizing resource efficiency while balancing performance, scalability, and cost.

The first paper delves into dynamic network slicing management with DeepCog, a deep convolutional neural network architecture designed for cost-aware capacity forecasting. By leveraging real-time data, DeepCog optimizes resource allocation to reduce overprovisioning and underprovisioning costs, demonstrating a 50% reduction in resource management costs in real-world tests. This focus on economic efficiency is a notable strength of DeepCog, setting it apart from approaches that primarily target technical efficiency. Comparatively, the second paper on NFV resource allocation takes a more topology-aware approach through DRL-enhanced Graph GNN. While both papers use deep learning, the GNN-based model explicitly models network topology, making it particularly adept at handling SFC in NFV, where network architecture plays a crucial role in resource management.

The Asynchronous DRL-enhanced GNN offers significant improvements in prediction accuracy and resource utilization, like DeepCog's optimization goals, but with a deeper focus on topology-specific challenges. Where DeepCog excels in cost and performance trade-offs, the DRL-GNN model stands out in its ability to proactively scale resources based on dynamic network conditions, ensuring efficiency without compromising service reliability.

Shifting from mobile networks to cloud environments, the third paper introduces a skewness-based method to optimize server utilization in data centers. While this approach also aims to balance resource utilization, it diverges from the deep learning-heavy approaches of DeepCog and the DRL-GNN model by using a metric-driven methodology. By minimizing resource imbalances, or skewness, the system efficiently co-locates workloads with varying demands, enhancing both energy efficiency and system performance. The focus on energy savings, while not as pronounced in other papers, highlights a crucial aspect of resource allocation, particularly in cloud computing environments where server underutilization can lead to significant inefficiencies.

In contrast, the fourth paper addresses resource management challenges in vRANs with vrAIn, which combines an autoencoder for dimensionality reduction with a DDPG algorithm for resource control decisions. While vrAIn shares the deep learning-driven approach of DeepCog and the DRL-GNN model, it focuses on optimizing computing capacity and enhancing QoS, specifically in vRAN environments. The use of an autoencoder to process high-dimensional data, such as traffic flows and channel quality, adds a layer of sophistication not present in the earlier methods, particularly in its ability to handle large-scale, complex data in real-time.

Both vrAIn and DeepCog share the goal of maximizing resource utilization while minimizing costs, but vrAIn's unique combination of an autoencoder and DDPG algorithm allows it to excel in complex, dynamic vRAN environments. This is further highlighted by vrAIn's success in reducing computing capacity by 30% and improving QoS targets by 25%, positioning it as a powerful solution for optimizing both computational and radio resources in real-time.

Similarly, the final paper introduces ATHENA, a machine learning-based framework for radio resource scheduling in vRANs. Like vrAIn, ATHENA focuses on real-time network conditions, dynamically adjusting resource allocations based on SNR and system congestion. However, ATHENA distinguishes itself by incorporating MR to ensure that the decisions made by the ML models are interpretable and justifiable. This transparency is a key differentiator, as other models like vrAIn and DeepCog focus primarily on performance optimization without addressing the interpretability of the model's decisions.

In comparison, ATHENA's integration of MR offers a higher level of decision-making transparency, which can be crucial in environments where operators need to understand the rationale behind resource allocation. While both vrAIn and ATHENA are geared towards optimizing vRANs, ATHENA's focus on interpretable machine learning offers a distinct advantage in scenarios where justifiable decision-making is as important as performance metrics.

Performance Isolation		
Reference	Methods	Summary
[YSH2022]	Deep Reinforcement Learning and Bidirectional LSTM.	The paper introduces the EE-DRL-RA method for optimizing resource allocation in 5G networks, which effectively combines DRL to enhance energy efficiency and service quality in network slicing. Utilizing models like A3C and SBiLSTM, this method dynamically adjusts resource allocations based on real-time network conditions and predictive analytics. Empirical results demonstrate significant improvements over existing methods, with enhanced convergence speed, reduced computational complexity, and better management of user service levels. This performance enhancement ensures optimal network utilization and robust isolation between network slices.
[YYS2023]	Autoencoder arquitecture with LSTM units.	The paper examines performance isolation issues in the control plane of virtualized software-defined networks (SDN), where existing network hypervisors struggle to maintain isolation between control channels. This lack of isolation leads to increased end-to-end control latency as more virtual switches are added, which can severely impair critical network functions such as routing in data centers. To address these challenges, the authors developed "Meteor," a network hypervisor that leverages a LSTM (Long Short-Term Memory) autoencoder to predict the control traffic for each virtual switch based on past patterns. This predictive modelling enables dynamic and tailored resource allocation, ensuring that the control traffic of one virtual switch does not adversely affect others. Meteor's performance was rigorously tested, showing a significant improvement in control message processing

		speed by up to 12.7 times and reducing end-to-end control latency by as much as 73.7%.
[ZYC2023]	Digital Twin (LSTM and Convolutional layers) enhanced Deep Reinforcement Learning	The paper introduces a digital twin-enhanced DRL framework to optimize resource management in network slicing, addressing the challenges of dynamic resource allocation in 5G networks. By simulating real network conditions using a digital twin, the framework reduces dependency on real-time data, enabling DRL agents to train more efficiently. The approach significantly outperforms traditional DRL methods in simulations, demonstrating faster convergence, improved computational efficiency, and better quality of service. This innovative use of digital twins in network management offers a scalable and adaptable solution, showcasing potential for future enhancements in wireless network technologies.
[SSC2017]	Holt Winters Method and geometric knapsack problem	The paper delves into the optimization of resource allocation in 5G networks using network slicing. The challenge lies in developing new resource allocation algorithms capable of handling various Service Level Agreements (SLAs) while maximizing network utilization. To effectively manage these challenges, the paper outlines the design of three crucial components of network slicing: traffic analysis and prediction for each network slice, admission control for slice requests, and adaptive correction based on observed deviations from forecasted loads. These elements are essential for efficiently managing network resources, leveraging traffic multiplexing gains across slices, and adapting to fluctuating traffic conditions. The authors demonstrate that their approach not only enhances system utilization but also illustrates a trade-off between conservative and aggressive forecasting strategies. They argue that striking the right

		balance is key to minimizing the risk of SLA violations while optimizing the utilization of network resources.
[MGF2018]	Emprirical Caractherization	The paper investigates the optimization of network slicing in 5G networks, where virtual network instances are tailored for specialized services on shared infrastructure. The focus is on understanding the trade-off between fully dedicated resources, which ensure service customization, and dynamic resource sharing, which enhances efficiency and cost-effectiveness. Using real-world data from a live mobile network, the authors analyze the efficiency gap caused by static resource allocation strategies across different network layers, from radio access to the core. They also explore the benefits of dynamic resource orchestration, showing how resource allocation can be adjusted over different timescales. The findings provide insights into optimizing network slicing, balancing service customization with efficiency, and improving the effectiveness of resource management algorithms.

The papers collectively tackle the core challenge of performance isolation in network slicing, each offering distinct methodologies to optimize resource allocation while maintaining robust service quality and efficiency. Although their approaches vary, the comparisons reveal overlapping themes and novel insights.

The first paper introduces the EE-DRL-RA method, focusing on energy efficiency and dynamic resource allocation using DRL techniques like A3C and SBiLSTM. Its strength lies in its dual focus on enhancing both energy efficiency and service quality, particularly emphasizing faster convergence and reduced computational complexity. This method ensures strong performance isolation by optimizing resource distribution across network slices, reducing interference, and ensuring optimal user service levels. Comparatively, the paper that examines Meteor, a network hypervisor, highlights performance isolation in the control plane of SDNs, addressing the challenge of maintaining isolation between control channels, an aspect not deeply touched upon by the EE-DRL-RA approach. Meteor also leverages LSTM autoencoders, but its focus is more on minimizing control latency in virtualized environments, which offers a novel perspective on control traffic isolation. Meteor's results, with

improvements up to 12.7 times in processing speed and a 73.7% reduction in latency, emphasize its ability to isolate performance by reducing cross-slice interference in control traffic.

While both papers apply machine learning for performance optimization, they focus on different areas: EE-DRL-RA on energy and resource allocation, and Meteor on control traffic. The integration of machine learning (DRL in EE-DRL-RA and LSTM in Meteor) as a predictive tool across these works shows a converging trend towards AI-driven resource management in 5G.

The third paper introduces a digital twin-enhanced DRL framework, another DRL-based approach that simulates real-time network conditions to reduce dependency on live data. This framework not only accelerates learning but also enhances scalability, addressing a gap that traditional DRL methods and static resource allocation strategies fail to cover. The digital twin's use in simulation allows for greater flexibility in training, which directly contrasts with the EE-DRL-RA's focus on real-time data-driven adjustments. Both methods excel at improving computational efficiency and resource management, but the digital twin approach shows greater adaptability by reducing real-time data reliance, potentially offering more robust performance isolation under fluctuating network conditions.

In contrast, the papers that investigate resource optimization in 5G slicing, including one focusing on the trade-off between dedicated resources and dynamic resource sharing, and another emphasizing forecasting strategies to balance SLA adherence and efficiency, delve into broader system-level challenges. These papers provide insights into balancing service customization with overall network efficiency. The analysis of dynamic resource orchestration across various timescales aligns closely with the digital twin approach, as both explore the benefits of flexible, adaptive resource management. However, the latter papers emphasize traffic multiplexing and predictive analytics to forecast resource needs, positioning their approaches as conservative in their predictions compared to the aggressive real-time adjustments found in EE-DRL-RA and the digital twin framework.

Lastly, the comparative trade-offs between aggressive and conservative forecasting in these papers reflect a broader discussion across the works. Whether it is balancing energy efficiency with service levels in EE-DRL-RA, or managing control traffic isolation in Meteor, each solution must navigate the risks of under-provisioning (leading to SLA violations) versus over-provisioning (leading to inefficiencies). The papers show that dynamic, predictive models like DRL and LSTM autoencoders are increasingly critical in striking this balance, offering varied solutions based on the specific challenges of their respective network domains.

Energy efficiency		
Reference	Methods	Summary
[PBS2023]	Adaptive Algorithm for Auto Scaling	The paper delves into optimizing scalability in server farms with an emphasis on balancing high reliability needed for 5G networks against reducing energy consumption. The authors

		introduce a novel auto-scaling method, A3S (Adaptive Algorithm for Auto Scaling), which accounts for server fallibility and activation delays. Utilizing control theory, A3S dynamically adjusts system parameters to optimally balance energy use and reliability. Performance evaluations via simulations show A3S outperforming contemporary reinforcement learning approaches, effectively converging to optimal states while maintaining stability and adapting well to real traffic patterns, making it a practical solution for energy-efficient service
[HLA2024]	Digital Twin and Neural Network for classification	The paper approaches the challenge of minimizing energy consumption in vRANs through the strategic allocation of Last-Level Cache (LLC) resources using the MemorAI framework. This research addresses the challenge of excessive energy usage due to non-isolated access to cache memory resources, a common issue in vRAN platforms known as the noisy neighbour problem. By employing techniques like cache memory isolation and employing a digital twin alongside a neural network classifier, MemorAI facilitates a more precise allocation of LLC resources. The results from deploying MemorAI show that it not only reduces the operational energy costs by optimizing LLC allocation in response to dynamic system demands but also achieves nearly optimal performance. This solution offers a robust and adaptable method for improving energy efficiency in the increasingly complex vRAN environment.
[ALG2024]		The paper presents an efficient energy management for vRANs using a multi-agent contextual bandit algorithm, ECORAN, which utilizes mean-field theory to handle the high variability and scale of demands in 5G network environments. The authors deploy ECORAN in a real O-RAN system and assesses its

		performance with data from a production mobile network. The results demonstrate that ECORAN significantly reduces energy consumption by up to 40% compared to traditional methods, while maintaining the necessary reliability. This optimization is achieved through a novel threshold-based offloading rule that adjusts the computational load between software and hardware accelerators in near-real-time, providing a scalable and effective solution to the challenges of modern network management.

The papers collectively address the pressing challenge of energy efficiency in 5G networks and virtualized environments, offering various innovative solutions that strike a balance between maintaining high performance and minimizing energy consumption. By comparing these approaches, we gain insights into their unique contributions and the varied techniques they employ to optimize energy use.

The first paper delves into the scalability and energy optimization of server farms, particularly within the context of 5G networks, with its introduction of the A3S. A3S stands out by incorporating control theory to dynamically adjust server parameters, balancing energy consumption with the high reliability demands of 5G. What sets A3S apart is its ability to adapt to real-time traffic while managing server fallibility and activation delays, an often-overlooked challenge in auto-scaling methods. Its performance evaluations show it outperforming contemporary reinforcement learning models by maintaining stability and adapting efficiently, making it a practical solution for energy-efficient service.

In contrast, the second paper tackles energy efficiency in virtualized Radio Access Networks (vRANs) by focusing on the LLC with its MemorAI framework. Where A3S focuses on server scalability and reliability, MemorAI zeroes in on cache resource allocation to reduce energy waste caused by the noisy neighbor problem—common in vRANs. By employing cache memory isolation and leveraging a digital twin and neural network classifier, MemorAI is able to precisely allocate LLC resources, achieving near-optimal energy savings and performance in vRAN systems. This method of addressing resource contention directly contrasts with A3S's broader focus on server management, showing how targeted resource optimization (LLC) can lead to substantial energy savings in specific network layers.

While both A3S and MemorAI offer adaptive methods to manage energy consumption, their focus areas diverge—A3S on balancing large-scale server management and MemorAI on micro-level cache allocation. This contrast highlights the different levels at which energy efficiency can be achieved, whether by managing overall server farm activity or by isolating specific resources like cache memory.

The third paper introduces ECORAN, a multi-agent contextual bandit algorithm designed to optimize energy consumption in vRANs using a real-world O-RAN system. Like MemorAI, ECORAN focuses on improving vRAN energy efficiency, but it takes a more system-wide approach by managing computational offloading between software and hardware accelerators. What differentiates ECORAN from both MemorAI and A3S is its use of a threshold-based offloading rule that adjusts the computational load in near real-time, dynamically balancing energy use across different processing units. This multi-agent strategy enables ECORAN to handle high demand variability in 5G networks while achieving up to 40% energy savings, a figure that surpasses both A3S and MemorAI in terms of overall energy reduction.

The comparison between ECORAN and MemorAI is particularly interesting because both target energy efficiency in vRAN environments but through different mechanisms. While MemorAI addresses cache isolation to reduce energy waste, ECORAN applies a broader method by optimizing the computational load across the entire vRAN system. ECORAN's threshold-based offloading provides a more flexible approach, enabling it to dynamically adapt to fluctuating network demands, which makes it better suited for large-scale, real-time 5G operations. MemorAI's strength, however, lies in its precision and ability to optimize LLC resources, showing that both targeted and system-wide approaches are valuable depending on the specific vRAN challenges being addressed.

Fault tolerance and Reliability		
Reference	Methods	Summary
[PGB2023]	Trace driven analysis	The paper addresses the optimization of server farms NFV contexts, particularly for B5G networks, with a focus on balancing high reliability and low energy consumption. It explores the complexities of auto-scaling, considering server fallibility and boot-up delays, to maintain an optimal number of active servers for reliable service without wasting resources. The research involves trace-driven simulations using real data to evaluate different auto-scaling strategies, comparing the use of a few highly reliable blade servers against a larger number of less reliable nano servers. The findings indicate a trade-off where nano servers, despite their greater energy efficiency, lead to a more dynamic and potentially disruptive operation with frequent task migrations and server activations or deactivations, which could affect hardware longevity and complicate management of the system.

5. Future Research Directions

Building upon our comprehensive analysis of the state of the art in resource allocation and energy-aware vRAN design, we identified key challenges and promising research directions that are critical to advancing the efficiency, adaptability, and sustainability of next-generation virtualized RANs. Our review highlighted the growing importance of intelligent resource management algorithms to handle increasingly complex and heterogeneous workloads, the pressing need to integrate energy-awareness into vRAN operations, and the potential of data-driven approaches to inform proactive decision-making. Moreover, we observed that existing works often treat these aspects in isolation, neglecting the interplay between infrastructure optimization, user behavior modeling, and dynamic reconfiguration strategies. To address these gaps and further explore the algorithmic and performance dimensions of vRAN evolution, we focus our next steps on three interconnected areas: (1) developing cost-aware autoscaling algorithms for reliable and energy-efficient server farm operations within vRAN environments; (2) designing privacy-preserving mobility modeling techniques to support mobility-driven resource allocation without compromising user confidentiality; and (3) evaluating the performance limits of dynamic vRAN reconfiguration to quantify the trade-offs and benefits of adapting network configurations in response to real-time user distributions and demand fluctuations. These complementary directions build on our SOTA findings and aim to provide a deeper understanding of how algorithmic innovation, privacy-aware modeling, and performance evaluation can jointly inform the design of more intelligent, efficient, and adaptive virtualized RANs.

Looking forward, subsequent deliverables will expand this research into three interconnected areas within the vRAN context:

1. **Cost-aware design of reliable and energy-efficient vRAN server farms:** We will investigate optimization algorithms for autoscaling virtualized infrastructure, combining queuing-theoretic models to ensure service reliability with cost models that capture both capital and operational expenditures. This work will provide a systematic framework for selecting the optimal server types and scaling levels needed to meet stringent performance guarantees in vRAN deployments, while simultaneously reducing operational costs and improving energy efficiency.
2. **Privacy-preserving mobility modeling for vRAN resource planning:** Building upon the state of the art we identified, we will develop a generative transformer-based model designed to synthesize realistic spatiotemporal mobility traces while preserving user privacy through differential privacy mechanisms. This approach will enable operators to leverage mobility-driven resource optimization and demand forecasting in vRAN environments without exposing identifiable user data, thereby ensuring compliance with privacy requirements while retaining the analytical value necessary for effective network management.
3. **Performance limits of dynamic vRAN reconfiguration:** We will examine the potential throughput gains achievable through spatial and temporal reconfiguration of vRAN cells,

using real network data to quantify performance improvements over static configurations. This evaluation will be conducted under the assumption that the number of connected users and their associated cell locations are known—critical information for accurately allocating resources and defining energy management policies. By incorporating this mobility-aware perspective, we aim to assess both the practical benefits and the operational boundaries of dynamic vRAN reconfiguration.

6. Summary and Conclusions

Our analysis of state-of-the-art vRAN solutions reveals a focused application of algorithms designed to enhance resource allocation, efficiency, and flexibility in network operations. Notably, Deep Reinforcement Learning (DRL) algorithms are used to optimize dynamic resource management, enabling CUs to allocate resources based on real-time demand across Distributed Units (DUs) and Radio Units (RUs). Neural networks (NNs) are also employed for tasks such as traffic prediction and signal processing, allowing the network to adapt rapidly to changing conditions and user mobility patterns. Other approaches leverage traditional optimization algorithms for scheduling and load balancing, helping to meet the diverse requirements of applications like IoT, enhanced Mobile Broadband (eMBB), and Ultra-Reliable Low Latency Communications (URLLC) through network slicing.

Despite these advancements, challenges remain, especially in coordinating these algorithms to handle high-bandwidth, low-latency demands on the fronthaul and ensuring seamless integration of Virtual Network Functions (VNFs). Algorithms must be finely tuned to balance computational load across the infrastructure and to maintain performance during peak traffic. While current solutions have demonstrated promising results, further improvements in algorithmic efficiency and scalability are needed to address these issues, especially in dense urban areas.

References

[GB2016] J. Gil Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," in *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518-532, Sept. 2016, doi: 10.1109/TNSM.2016.2598420.

[SC2009] G. Somani and S. Chaudhary, "Application Performance Isolation in Virtualization," *2009 IEEE International Conference on Cloud Computing*, Bangalore, India, 2009, pp. 41-48, doi: 10.1109/CLOUD.2009.78.

[PBS2023] J. Perez-Valero, A. Banchs, P. Serrano, J. Ortín, J. Garcia-Reinoso and X. Costa-Pérez, "Energy-Aware Adaptive Scaling of Server Farms for NFV With Reliability Requirements," in IEEE Transactions on Mobile Computing, vol. 23, no. 5, pp. 4273-4284, May 2024, doi: 10.1109/TMC.2023.3288604.

[MBG2024] M. Milani, D. Bega, M. Gramaglia, P. Serrano and C. Mannweiler, "ATELIER: Service Tailored and Limited-Trust Network Analytics Using Cooperative Learning," in IEEE Open Journal of the Communications Society, vol. 5, pp. 3315-3330, 2024, doi: 10.1109/OJCOMS.2024.3401746.

[BGF2019] D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, "DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning," IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, Paris, France, 2019, pp. 280-288, doi: 10.1109/INFOCOM.2019.8737488.

[JHD2019] N. Jalodia, S. Henna and A. Davy, "Deep Reinforcement Learning for Topology-Aware VNF Resource Prediction in NFV Environments," *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, Dallas, TX, USA, 2019, pp. 1-5, doi: 10.1109/NFV-SDN47374.2019.9040154.

[ZWQ2013] Z. Xiao, W. Song and Q. Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1107-1117, June 2013, doi: 10.1109/TPDS.2012.283.

[AGG2019] J. A. Ayala-Romero, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Pérez, A. Banchs and J. J. Alcaraz, "vrAIn: Deep Learning Based Orchestration for Computing and Radio Resources in vRANs," in *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2652-2670, 1 July 2022, doi: 10.1109/TMC.2020.3043100.

[AGC2023] N. Apostolakis, M. Gramaglia, L. E. Chatzieleftheriou, T. Subramanya, A. Banchs and H. Sanneck, "ATHENA: Machine Learning and Reasoning for Radio Resources Scheduling in vRAN Systems," in *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, pp. 263-279, Feb. 2024, doi: 10.1109/JSAC.2023.3336155.

[YSH2022] Y. Azimi, S. Yousefi, H. Kalbkhani and T. Kunz, "Energy-Efficient Deep Reinforcement Learning Assisted Resource Allocation for 5G-RAN Slicing," in IEEE Transactions on Vehicular Technology, vol. 71, no. 1, pp. 856-871, Jan. 2022, doi: 10.1109/TVT.2021.3128513.

[YYS2023] Y. Yoo, G. Yang, C. Shin, J. Lee and C. Yoo, "Control Channel Isolation in SDN Virtualization: A Machine Learning Approach," 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Bangalore, India, 2023, pp. 273-285, doi: 10.1109/CCGrid57682.2023.00034.

[ZYC2023] Zhang, Z., Huang, Y., Zhang, C., Zheng, Q., Yang, L., & You, X. (2023). Digital Twin-Enhanced Deep Reinforcement Learning for Resource Management in Networks Slicing. IEEE Transactions on Communications, 72, 6209-6224. https://doi.org/10.48550/arXiv.2311.16876

[SSC2017] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, GA, USA, 2017, pp. 1-9, doi: 10.1109/INFOCOM.2017.8057230.

[MGF2018] Cristina Marquez, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. 2018. How Should I Slice My Network? A Multi-Service Empirical Evaluation of Resource Sharing Efficiency. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18). Association for Computing Machinery, New York, NY, USA, 191–206. https://doi.org/10.1145/3241539.3241567

[PGB2023] Jesus Perez-Valero, Jaime Garcia-Reinoso, Albert Banchs, Pablo Serrano, Jorge Ortin, Xavier Costa-Perez, Performance trade-offs of auto scaling schemes for NFV with reliability requirements, Computer Communications, Volume 212, 2023, Pages 251-261, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2023.10.001.

[HLA2024] Hidalgo, E. S., Lozano, J. X. S., Ayala-Romero, J. A., Garcia-Saavedra, A., Li, X., & Costa-Perez, X. (2024, May). MemorAI: Energy-Efficient Last-Level Cache Memory Optimization for Virtualized RANs. In 2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN) (pp. 25-30). IEEE.

[ALG2024] J. A. Ayala-Romero, L. Lo Schiavo, A. Garcia-Saavedra and X. Costa-Perez, "Mean-Field Multi-Agent Contextual Bandit for Energy-Efficient Resource Allocation in vRANs," IEEE INFOCOM 2024 - IEEE Conference on Computer Communications, Vancouver, BC, Canada, 2024, pp. 911-920, doi: 10.1109/INFOCOM52122.2024.10621197.