Energy Consumption Estimation Model as a QoS-Aware Elasticity Support Mechanism for 5G Network Slicing

Marvin A. L. Martinez[1], **José A. Suruagy**[1,2], **Marcia C. Machado**[3]
Tereza C. Carvalho[3] **Joberto S. B. Martins**[4], **Sabino R. S. Antunes**[1]
Adnei W. Donatti[3], **Sand Correa**[5], **Tiago Ferreto**[6], **Flávio de Oliveira Silva**[7]

[1] Informatics Center (CIn) – Federal University of Pernambuco (UFPE) –PE – Brazil

[2]CESAR School Recife – PE – Brazil

[3]University of São Paulo (USP) São Paulo – SP – Brazil

[4]Salvador University (UNIFACS) Salvador – BA – Brazil

[5]Universidade Federal de Goiás (UFG) Goiânia – GO – Brazil

[6]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) Porto Alegre – RS – Brazil

[7]Universidade do Minho Braga – Portugal

malm@cin.ufpe.br, jasm@cesar.school, macrismachado@usp.br
carvalho@larc.usp.br, joberto.martins@unifacs.br, srsa@cin.ufpe.br
adnei.donatti@edu.udesc.br, sand@inf.ufg.br, tiago.ferreto@pucrs.br
flavio@di.uminho.pt

Abstract. *This paper presents a dynamic energy consumption estimation model for Fifth Generation (5G) Network Slicing (NS) with support for Quality of Service (QoS)-aware elasticity decisions. The model evaluates the admission of a new virtual network function under four controlled scenarios and shows that optimized strategies reduce the additional energy required for admission. The results also identify the EDGE segment as the main point of energy variation. These findings indicate that energy-aware elasticity can improve resource use, reduce energy consumption, and help lower indirect CO_2 emissions in sliced networks.*

1. Introduction

The evolution of 5G networks has enabled the support of heterogeneous services with different latency, bandwidth, and reliability requirements. In this context, NS allows a shared infrastructure to be partitioned into multiple virtual networks or slices adapted to specific service demands. While this improves flexibility and resource management, it also raises important energy-efficiency challenges [Lorincz et al. 2024]. As 5G infrastructures support a growing number of services and virtualized functions, the demand for computational and transmission resources increases considerably, resulting in higher overall energy consumption. Despite the technological advances in 5G networks, significant gaps remain in estimating the energy consumption of NS.

In particular, existing approaches still lack an integrated model capable of estimating the energy consumption of slices across the main network segments, including Radio Access Network (RAN), Edge Cloud, Transport, and Core Network, in an end-to-end manner. Previous studies have typically analyzed these segments separately, without explicitly considering their interactions or their combined impact on the overall energy consumption of the slice [Lorincz et al. 2024, S. et al. 2023]. To address this limitation, this work develops an end-to-end energy consumption estimation model for NS to support QoS-aware elasticity decisions based on resource scaling and Virtual Network Function (VNF) migration, enabling the evaluation of energy consumption under dynamic resource allocation while preserving service requirements.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed energy consumption estimation model for NS. Section 4 describes the energy optimization mechanisms considered in this work. Section 5 discusses the experimental setting and the obtained results. Finally, Section 6 concludes the paper and outlines future work.

2. Related work

Table 1 presents a comparative overview of the main studies related to energy consumption estimation and resource management strategies in NS. The selected works are analyzed according to four key aspects: the existence of energy consumption estimation models for NS, the use of QoS consolidation mechanisms, the adoption of migration strategies. Several studies have proposed analytical models to estimate the energy consumption of NS in 5G infrastructures. The work presented in [Lorincz et al. 2024] introduces an aggregated formulation that combines the energy consumption of dedicated hardware with estimates associated with virtualized network functions. In [Masoudi et al. 2022], the authors develop an energy estimation model for slices in a Cloud Radio Access Network (C-RAN) architecture, where the total consumption is obtained from the contributions of base stations, fronthaul/midhaul, and centralized processing. The study presented in [Taskou et al. 2022] focuses on modeling the energy consumption of the core network based on the computational load associated with NS and the capacity parameters of physical servers. Finally, [Doorgakant et al. 2025] propose an end-to-end mathematical model that estimates energy consumption by considering the different domains of the virtualized 5G infrastructure.

In addition to energy estimation models, other studies investigate elasticity mechanisms to improve the energy efficiency of virtualized network infrastructures. In [Chintapalli et al. 2024], the authors analyze the energy-efficient deployment of parallelized service function chains in Network Function Virtualization (NFV)-based networks by prioritizing the reuse of already active resources while meeting end-to-end delay constraints. Similarly, [Liu et al. 2021] propose a resource reconfiguration strategy based on short-term demand prediction, where VNFs can be consolidated into fewer physical servers and migration to another server is performed only when the global energy gain exceeds the migration cost. The framework presented in [Nine et al. 2023] focuses on dynamic resource allocation using reinforcement learning techniques to promote VNF consolidation while satisfying service-level constraints. In contrast, the survey conducted in [Attaoui et al. 2023] highlights consolidation and joint placement of VNFs and CNFs as relevant strategies to improve resource utilization and energy efficiency in 5G infrastructures. Finally, [Tchinda et al. 2024] investigate energy-efficient VNF placement in wireless mesh networks, where migration strategies are used to relocate VNFs toward nodes with lower energy costs in order to balance infrastructure consumption.

Most related studies address energy estimation, consolidation, or migration separately. In contrast, the present work combines end-to-end energy consumption estimation with dynamic QoS-aware elasticity support, explicitly considering both VNF consolidation and migration. This represents the main distinction of our proposal.

Table 1. Comparative analysis of related works on energy estimation and elasticity mechanisms in Network Slicing

No	Article	Energy Estimation in Network Slicing	Consolidation (Energy Saving %)	Migration	Reference
1	A Comprehensive Overview of Network Slicing for Improving the Energy Efficiency of Fifth-Generation Networks	Yes	No	No	[Lorincz et al. 2024]
2	Energy-Optimal End-to-End Network Slicing in Cloud-Based Architecture	Yes	No	No	[Masoudi et al. 2022]
3	Minimizing Energy Consumption for End-to-End Slicing in 5G Wireless Networks and Beyond	Yes	No	No	[Taskou et al. 2022]
4	End-to-End Power Models for 5G Radio Access Network Architectures with a Perspective on 6G	Yes	No	No	[Doorgakant et al. 2025]
5	Energy efficient and delay aware deployment of parallelized service function chains in NFV-based networks	No	Yes (35%, 40%)	No	[Chintapalli et al. 2024]
6	Energy-Efficient Virtual Network Function Reconfiguration Strategy Based on Short-Term Resources Requirement Prediction	No	Yes (20%, 40%)	Yes	[Liu et al. 2021]
7	GreenNFV: Energy-Efficient Network Function Virtualization with Service Level Agreement Constraints	No	Yes (35%)	No	[Nine et al. 2023]
8	VNF and CNF Placement in 5G: Recent Advances and Future Trends	No	Yes	Yes	[Attaoui et al. 2023]
9	Energy-Efficient Placement of Virtual Network Functions in a Wireless Mesh Network	No	No	Yes	[Tchinda et al. 2024]
10	**Energy Consumption Estimation Model as a QoS-Aware Elasticity Support Mechanism for 5G Network Slicing**	**Yes**	**Yes (20%)**	**Yes**	**This work**

3. Energy Consumption Model for Network Slices

A slice is flexible and scalable, and it can be tailored to specific service demands, thereby improving service performance. A slice also enables value creation for vertical or horizontal segments, service providers, and third parties lacking physical network infrastructure by offering radio, network, and cloud resources. This architecture supports personalized network operations and elasticity in resource management, allowing slices to dynamically adapt resource allocation according to service demands and enabling true service differentiation [Fahmi and Lin 2021].

3.1. Components of a Slice

The segments play an interdependent role, contributing to a modular and scalable network capable of meeting diverse demands, ranging from latency-sensitive critical applications to services requiring high bandwidth.

- *RAN*: Connects user devices to the network, being the first access point for data transmission. It manages radio resources to meet QoS requirements such as latency, bandwidth, and reliability.

- *Edge*: Extends the cloud closer to the user, processing data in a distributed manner and reducing latency. It supports resource allocation per slice, ensuring flexibility and scalability.
- *Transport Network*: Responsible for connectivity between RAN, Edge, and core. It includes fronthaul, midhaul, and backhaul, optimizing bandwidth and latency for multiple independent applications.
- *Core*: The central segment of the network responsible for processing data and managing critical functions, such as the Evolved Packet Core (EPC) and the User Plane Function (UPF). It implements authentication, mobility, and control policies, while addressing QoS. In this work, we adopt the distinction between core and Transport for energy consumption analysis.

As mentioned before, NS are composed of virtual resources allocated in several physical resources such as computing, networking, and memory resources. The NS life-cycle includes preparation, commissioning, operation, and decommissioning phases.

3.2. Energy Consumption Estimation Model for Networking Slice

This model provides an estimation of the energy consumption of a slice, based on the four fundamental segments of its end-to-end operation: RAN, Edge, Transport, and Core segments Figure 1. Each segment is modeled to reflect the energy impact derived from the resources allocated to the slice within that specific part of the network infrastructure.

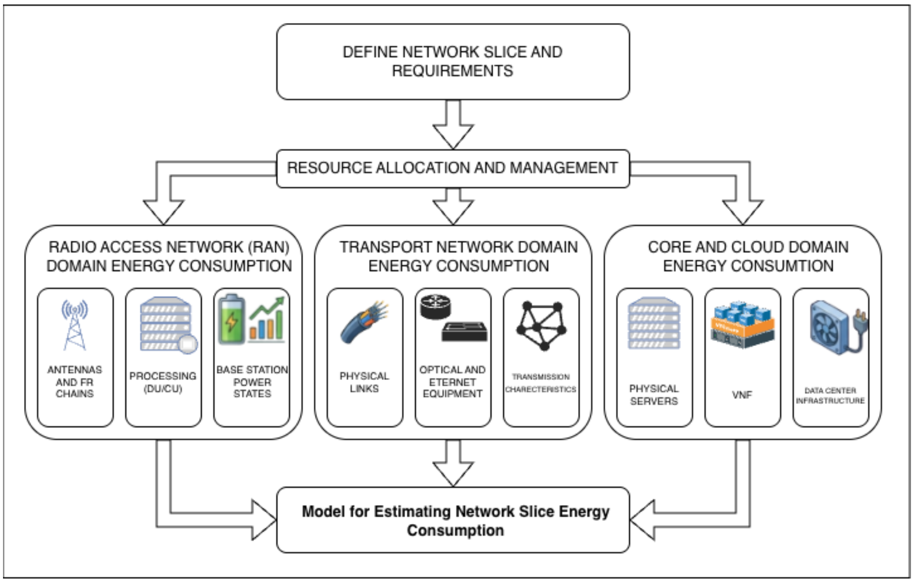

Figure 1. Proposed framework for estimating network slice energy consumption across RAN, Transport, and Core/Cloud domains.

Accordingly, the total energy consumption of a slice S is defined as the sum of the consumptions associated with that slice in the correspondent core nodes (datacenters), transport elements, edge resources and RANs of the infrastructure:

$$EC_S^{Total} = \sum_{c=1}^{C} CEC_c^s + \sum_{t=1}^{T} TEC_t^s + \sum_{e=1}^{E} EEC_e^s + \sum_{r=1}^{R} REC_r^s \tag{1}$$

In the absence of direct slice-level energy measurements, the proposed model proportionally assigns the total energy of each component to the slices according to their allocated resource share. The estimation is restricted to the network infrastructure supporting the slice and excludes external devices, such as mobile terminals or user equipment.

3.2.1. Core Network Model

For the Core Network, we estimate the energy consumption of a slice from the Data Center (DC) resources that support its operation. This includes both the Information Technology Information Technology (IT) equipment responsible for processing and storage, such as servers, Central Processing Units Central Processing Unit (CPU), memory, and storage systems, and the supporting infrastructure required to maintain the DC in operation, including cooling, internal networking devices, power supply systems, and auxiliary control components. The total DC consumed energy is proportionally assigned to each slice according to the share of computational resources used by that slice relative to the total resources used by all slices.

$$CEC_c^s = \sum_{m=1}^{M} \frac{Computational\ Resources(s, m)}{\sum_{i=1}^{S} Computational\ Resources(s, m)} \times CEC_m \quad (2)$$

In other words, the total energy consumed by a slice s in a given all DC in the network infrastructure is estimated as the sum of the energy consumed by the slice in each of its physical machines. This value is obtained as the proportion of the total energy consumed by the DC (including cooling and other overheads) according to the fraction of computational resources (e.g., processing cores) allocated to slice s with respect to the total computational resources allocated to all slices in that DC.

3.2.2. Edge Cloud Model

The Edge Cloud model estimates the energy consumption associated with distributed computing resources deployed close to the end user. In this segment, virtualized virtualized Radio Access Networks (vRANs), Open Radio Access Networks (O-RANs), and Cloud Radio Access Networks (C-RANs) support virtualized and low-latency processing near the service demand. Unlike the Core, which concentrates centralized functions, Edge represents the energy contribution of computation performed closer to users.

$$EEC_e^s = \frac{Resources(s, e)}{\sum_{i=1}^{S} Resources(i, e)} \times EEC_e \quad (3)$$

In this model, a weighted summation of the resources consumed is performed, where “Resources” represent the computational units allocated to each slice. The consumption of a slice s in a given edge site e is determined as the fraction of its allocated resources compared to the total resources used by all slices in that site. This proportional distribution provides an estimation of the energy consumption of each slice, reflecting its actual impact on the overall energy usage of the Edge infrastructure.

3.2.3. Transport Model

The Transport model estimates the energy consumption associated with traffic forwarding between the main segments of a NS. In this case, the model assumes that the dominant factor of transport energy consumption is the traffic load, expressed as the number of bits transferred.

$$TEC_t^s = \frac{bits(s,t)}{\sum_{i=1}^{S} bits(i,t)} \times TEC_t \tag{4}$$

In this model, we analyze the bit transmission inside the Crosshaul infrastructure, evaluating how slice traffic impacts energy consumption in fronthaul, midhaul, and backhaul segments. The consumption is computed based on the ratio of bits transferred by the slice relative to the total network traffic. This method allows quantifying the energy impact of each slice, considering the dynamics of resource allocation and the volume of transported data.

3.2.4. RAN Model

The RAN energy consumption model focuses on calculating the energy used by Radio Units (RUs) within a slice. These units include antennas, power amplifiers, and transmission equipment, which are considered the main energy consumers in the RAN. Other relevant elements include power supply units, control units, and cooling systems.

$$REC_r^s = \frac{Transmission(s,r)}{\sum_{i=1}^{S} Transmission(i,r)} \times REC_r \tag{5}$$

This model computes the total energy consumption of the most relevant components within the RAN allocated to a slice. The term “Transmission” in the equation represents the combined operation of individual elements needed for the slice to function. Dividing this consumption by the total transmission across all components, we estimate the specific RAN energy consumption per slice.

3.3. Elasticity

Elasticity is a key concept in virtualized and flexible computing environments, such as cloud computing, NFV, and NS [Saha et al. 2023]. It refers to the ability of a system to adjust allocated resources according to workload and service requirements, balancing performance and infrastructure efficiency [Subrahmanyam et al. 2023]. By supporting dynamic scaling and automated resource management, elasticity improves resource utilization and helps reduce energy consumption [Moreno-Vozmediano et al. 2025].

Consolidation. In NS environments, elasticity mechanisms can also support resource consolidation strategies. Consolidation consists of reallocating multiple VNFs onto a reduced number of infrastructure nodes in order to improve resource utilization and decrease infrastructure energy consumption. By concentrating VNFs on fewer nodes when workload decreases, consolidation allows idle resources to be released or powered off, contributing to more energy-efficient network operation [Subrahmanyam et al. 2023, Moreno-Vozmediano et al. 2025].

4. Energy Optimization Mechanisms in Network Slicing

The proposed mechanisms support the reconfiguration of a slice when a new high-priority VNF must be admitted in a target location with insufficient available resources. In this situation, the mechanisms evaluates whether reallocation, consolidation, or migration can accommodate the new demand with lower total end-to-end energy cost while preserving QoS. A reconfiguration action is accepted only if it is feasible and improves the energy efficiency of the infrastructure allocation.

4.1. Migration.

In virtualized environments and NS, functional elasticity includes not only dynamic resource scaling (vertical and horizontal), but also the capability to relocate network functions across infrastructure nodes when necessary. Migration enables the system to adapt to variations in workload, user mobility, or infrastructure conditions while maintaining QoS requirements [Qu et al. 2020]. By dynamically reallocating VNFs to nodes with better processing capacity or lower latency, migration contributes to improved resource utilization and supports more efficient network operation [Hu et al. 2023].

However, VNF migration may also introduce operational overhead in energy-aware decision-making. This process can involve state transfer, source-destination synchronization, routing updates, and additional traffic during relocation. Therefore, the benefits of migration should be evaluated together with its transient impact on energy consumption, network load, and QoS stability.

4.2. Network Functions virtualization in Network Slicing

These functions are responsible for providing the necessary control and data processing capabilities required by the services deployed in the slice. In the context of elasticity and resource consolidation, it is important to identify the main functions associated with each network segment. the principal VNFs considered in this work and their corresponding roles within the network architecture.

In the Core, functions correspond to the NFV-based elements of the 5G architecture, particularly the Access and Mobility Management Function (AMF), the Session Management Function (SMF), and the UPF. These functions may be deployed as virtualized instances over cloud infrastructure. Thus, consolidation in the Core is understood as the grouping or reorganization of function instances rather than as an architectural redistribution such as that observed in the RAN.

4.3. Dynamic Optimization

In the proposed mechanisms, dynamic optimization is understood as a mechanisms-guided reconfiguration process triggered when the current allocation of resources is no longer sufficient or no longer energy-efficient for slice operation. This situation may occur, for example, when a newly arriving VNF with high priority must be deployed in a specific network zone, but the available resources in that location are limited. Under such conditions, the mechanisms evaluates whether a reconfiguration action can release the required resources while preserving QoS constraints.

VNF Consolidation for Baseline Energy Savings. When the destination node or instance requires additional capacity to host the incoming demand, consolidation may be supported by vertical resource adjustment. This adjustment allows the receiving environment to absorb the additional load without compromising service continuity or QoS. However, vertical adjustment is treated here as an auxiliary enabler of consolidation rather than as the consolidation action itself. The main objective remains the same: to reduce the number of active resources required to sustain the set of VNFs while maintaining service continuity and QoS compliance.

VNF Migration. is triggered when a high-priority incoming VNF must be admitted in a target location with insufficient available resources. In this case, the proposed mechanism evaluates the relocation of an already deployed VNF to a feasible destination, considering energy reduction, resource release, and QoS preservation. Once relocated, the migrated VNF may be consolidated with active functions, reducing infrastructure cost and releasing part of the resources required for the admission of the incoming VNF.

QoS-Aware Energy Efficiency. Energy optimization in network slicing should not be driven solely by energy reduction goals, since aggressive reconfiguration may negatively affect service performance. For this reason, the proposed mechanisms considers QoS as a fundamental constraint in the optimization process. A reconfiguration action is considered feasible only when it contributes to lowering end-to-end energy consumption without violating the operational requirements of the slice.

4.4. Energy Efficiency in Network Slicing

The proposed mechanisms contributes to energy efficiency in network slicing by supporting reconfiguration decisions that reduce end-to-end infrastructure consumption while preserving QoS requirements. Since the model estimates the energy contribution of the RAN, Edge, Transport, and Core segments, it can be used to identify operating conditions in which the slice can be sustained with lower energy expenditure.

- *CO_2 Emissions Reduction.* By reducing the end-to-end energy consumption associated with slice operation, the proposed approach can also decrease the indirect CO_2 emissions linked to electricity usage. In this work, such reduction is understood as a consequence of lower infrastructure energy demand rather than as the result of a dedicated carbon-emission model.
- *Environmental Sustainability.* Beyond the reduction of indirect emissions, the proposed approach supports a more sustainable operation of the infrastructure by avoiding unnecessary resource activation, limiting overprovisioning, and favoring more efficient use of active nodes during slice operation.

5. Evaluation

The evaluation aims to analyze how different allocation strategies affect the energy cost of admitting a new VNF and the resulting end-to-end energy consumption of the slice. Four scenarios were considered for this purpose: power-on, conventional allocation, consolidation, and migration plus consolidation. The comparison was designed to verify the impact of elasticity mechanisms on admission cost and slice energy behavior. To support this analysis, 10,000 controlled simulations were performed under predefined rules of priority, resource availability, functional compatibility, and QoS preservation.

5.1. Experimental Setting

The evaluation focuses on the *EDGE* segment, which is defined in this controlled experiment as the target location for the admission of the incoming VNF. This choice was adopted to enable a controlled analysis of the local impact of reconfiguration and to verify how changes in the energy consumption of the affected segment are reflected in the end-to-end energy consumption of the slice. By concentrating the reconfiguration process on computational resources, the evaluation avoids mixing this effect with variations associated with other network segments, enabling a clearer analysis of its impact on the end-to-end energy consumption of the slice. The experiment considers the energy consumption of active infrastructure resources, including the idle power associated with resources that remain powered on, without explicitly modeling power-down states of physical equipment. Thus, the observed energy variation reflects both the additional workload introduced by the incoming VNF and the energy cost of operating active resources during slice execution.

Four scenarios were considered. In *S1_POWER_ON*, the incoming VNF requires the activation of additional resources, representing the highest admission cost condition. In *S2_NO_CONSOLIDATION*, the VNF is admitted through the conventional allocation strategy, without applying optimization mechanisms, and is therefore adopted as the baseline scenario. In *S3_CONSOLIDATION*, the incoming VNF is consolidated with a compatible function already active in the system, reducing the incremental admission cost. Finally, in *S4_PRIORITY_MIGRATE_CONSOLIDATION*, a previously allocated lower-priority VNF is reallocated to another feasible location, releasing space in the constrained EDGE segment and enabling subsequent consolidation, thus combining migration and consolidation as a joint optimization strategy.

5.2. Energy Efficiency Results

Figure 2 (a) shows a clear decreasing trend in the optimized admission energy cost from S1 to S4. The highest values are observed in S1, while S2 defines the non-optimized reference level. In contrast, S3 and S4 present progressively lower admission costs, indicating that the application of consolidation and migrate+consolidation strategies reduces the incremental energy required to accommodate the incoming VNF.

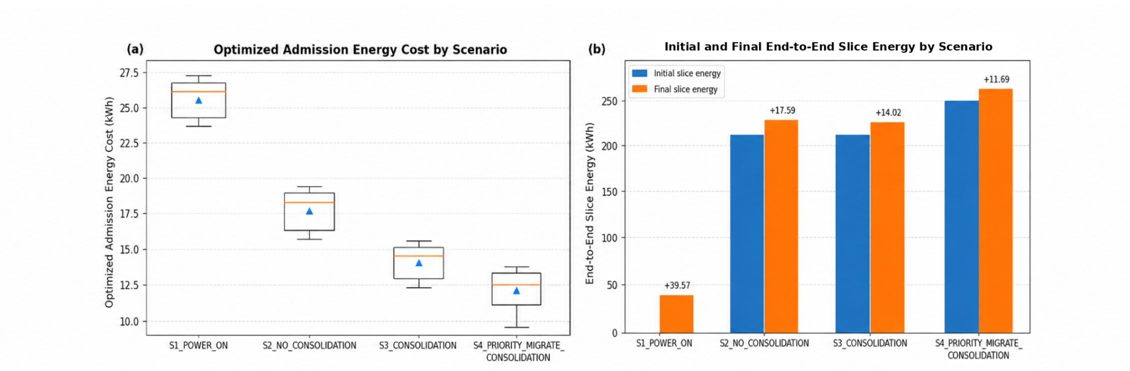

Figure 2. Optimized admission energy cost of the incoming VNF across the evaluated scenarios.

Taking S2 as the baseline scenario, the results indicate that consolidation in S3 reduces the optimized admission energy cost by approximately 20%, whereas the joint use of migration and consolidation in S4 increases this reduction to about 31%. This behavior suggests that the energy benefit is not only associated with resource sharing, but also with a more efficient redistribution of already allocated VNFs within the constrained EDGE segment.

5.3. Segment Energy Consumption Estimation

Figure 2 (b) presents the end-to-end slice energy consumption before and after the admission of a new VNF under the four evaluated scenarios. Although the comparison is expressed at the slice level, the energy increase originates in the EDGE segment, since this controlled experiment considers EDGE as the domain where the incoming VNF is installed and where the reconfiguration strategies are applied. In this sense, the proposed mechanisms makes it possible to observe how a local variation in EDGE is reflected in the total energy consumption of the slice, while still preserving the integrated estimation of the four end-to-end segments.

The results show that *S1* presents the largest energy increase, since the admission of the incoming VNF requires the activation of additional resources. In *S2*, the installation follows the conventional strategy and is therefore adopted as the reference scenario. Relative to *S2*, *S3* reduces the incremental admission cost from 17.59 to 14.02 kWh, corresponding to a reduction of approximately 20.3%. In *S4*, this cost decreases further to 11.69 kWh, representing an improvement of about 33.5% with respect to *S2*. These results indicate that consolidation improves energy efficiency, while the joint use of migration and consolidation can further reduce the energy cost of admitting the new VNF.

5.4. QoS-Aware Performance Evaluation

Figure 3 summarizes the QoS-related behavior of the evaluated scenarios using latency, jitter, and throughput as representative performance metrics. These metrics were selected because they allow evaluating whether the proposed energy-aware reconfiguration strategies preserve the main service conditions affected by resource allocation, consolidation, and migration. For the controlled evaluation, the adopted reference constraints were 50 ms for latency, 10 ms for jitter, and 35 Mbps for minimum throughput [3GPP 2024, Demichelis and Chimento 2002].

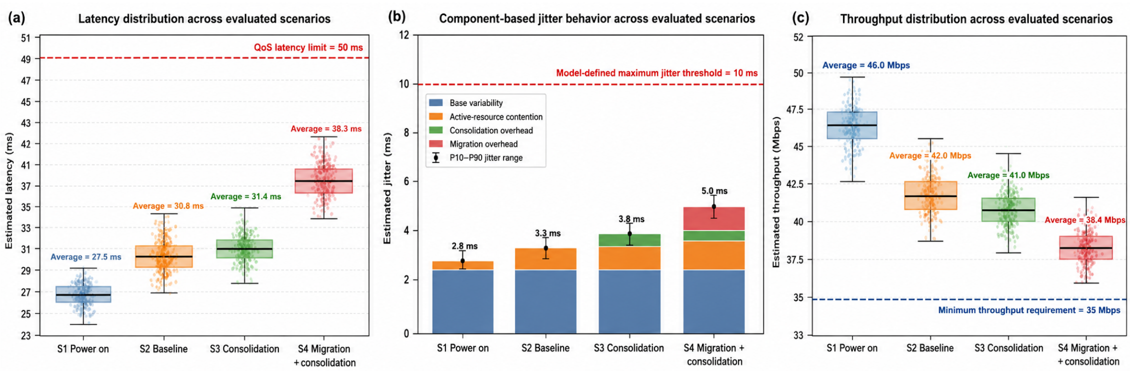

Figure 3. Latency, jitter, and throughput behavior across the evaluated energy-aware elasticity scenarios.

Overall, the results show that the proposed energy-aware elasticity strategies preserve QoS feasibility under the controlled evaluation. In Figure 3(a), latency remains relatively stable in S1, S2, and S3, since these scenarios do not require VNF migration. The higher latency observed in S4 is expected because migration introduces additional reconfiguration steps in the constrained Edge segment. Figure 3(b) shows a progressive increase in jitter as additional sources of variability are introduced, including active-resource contention, consolidation overhead, and migration overhead. Finally, Figure 3(c) shows a gradual reduction in throughput from S1 to S4, reflecting the effect of resource sharing and reconfiguration on the effective data transfer rate. Nevertheless, all scenarios remain within the adopted QoS thresholds, indicating that the proposed strategies reduce the energy impact of VNF admission without violating the QoS constraints considered in the mechanisms.

6. Conclusion and future work

This work presented an energy consumption estimation model for 5G NS with support for QoS-aware elasticity mechanisms. The mechanisms evaluates the admission of a new VNF under different reconfiguration strategies, including conventional allocation, consolidation, and migration-assisted consolidation. The results showed that the optimized strategies reduce the incremental energy cost when compared with the conventional allocation scenario, contributing to better resource utilization, lower energy consumption, and a potential reduction of indirect CO_2 emissions.

In addition to the energy analysis, QoS-related metrics such as latency, jitter, and throughput were evaluated. The results indicated that, although migration introduces a moderate performance overhead due to the reconfiguration process, all evaluated scenarios remain within the thresholds considered in the mechanisms used. This shows that energy-aware elasticity strategies can reduce energy consumption while preserving the operational feasibility of the slice.

As future work, the mechanisms should be extended to represent the energy cost of migration in greater detail, including the traffic generated during state transfer and routing adjustments. Future evaluations should also consider scenarios in which consolidation allows physical resources to be released or powered down, as well as analyses across other network segments.

Acknowledgments

The authors thank CNPq and FAPESP for their support. This work was also developed under the FAPESP/MCTIC/CGI cooperation agreement within the thematic research project 2018/23097-3 – SFI2 – Slicing Future Internet Infrastructures.

References

3GPP (2024). 5G; System Architecture for the 5G System (5GS). Technical Specification TS 23.501, 3rd Generation Partnership Project (3GPP). Version 18.5.0, Release 18.

Attaoui, W., Sabir, E., Elbiaze, H., and Guizani, M. (2023). VNF and CNF placement in 5G: Recent advances and future trends. *IEEE Transactions on Network and Service Management*, 20:4698–4733.

Chintapalli, V. R., Partani, R., Tamma, B. R., and Murthy, C. S. R. (2024). Energy efficient and delay aware deployment of parallelized service function chains in NFV-based networks. *Computer Networks*, 243:110289.

Demichelis, C. and Chimento, P. (2002). IP Packet Delay Variation Metric for IP Performance Metrics (IPPM). RFC 3393, IETF.

Doorgakant, B., Fowdur, T. P., and Akinsolu, M. O. (2025). End-to-end power models for 5G radio access network architectures with a perspective on 6G. *Mathematics 2025, Vol. 13*, 13.

Fahmi, H. Z. and Lin, F. J. (2021). NFV-enabled vertical scalability for IoT slices. *2021 22nd Asia-Pacific Network Operations and Management Symposium, APNOMS 2021*, pages 5–8.

Hu, Y., Min, G., Li, J., Li, Z., Cai, Z., and Zhang, J. (2023). VNF migration in digital twin network for NFV environment. *Electronics 2023, Vol. 12*, 12.

Liu, Y., Ran, J., Hu, H., and Tang, B. (2021). Energy-efficient virtual network function reconfiguration strategy based on short-term resources requirement prediction. *Electronics 2021, Vol. 10*, 10.

Lorincz, J., Kukuruzović, A., and Blažević, Z. (2024). A comprehensive overview of network slicing for improving the energy efficiency of fifth-generation networks. *Sensors (Basel, Switzerland)*, 24:3242.

Masoudi, M., Demir, O. T., Zander, J., and Cavdar, C. (2022). Energy-optimal end-to-end network slicing in cloud-based architecture. *IEEE Open Journal of the Communications Society*, 3:574–592.

Moreno-Vozmediano, R., Huedo, E., Montero, R. S., and Llorente, I. M. (2025). AI-driven resource allocation and auto-scaling of VNFs in Edge-5G-IoT ecosystems. *Electronics 2025, Vol. 14*, 14.

Nine, M. D. Z., Bulut, M. F., Kosar, T., and Hwang, J. (2023). Greennfv: Energy-efficient network function virtualization with service level agreement constraints. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*.

Qu, K., Zhuang, W., Ye, Q., Shen, X., Li, X., and Rao, J. (2020). Dynamic flow migration for embedded services in SDN/NFV-enabled 5G core networks. *IEEE Transactions on Communications*, 68:2394–2408.

S., S., Mishra, S., and Hota, C. (2023). Joint qos and energy-efficient resource allocation and scheduling in 5G network slicing. *Computer Communications*, 202:110–123.

Saha, N., Shahriar, N., Boutaba, R., and Saleh, A. (2023). Monarch: Network slice monitoring architecture for cloud native 5G deployments.

Subrahmanyam, V., Kumar, S., Srivastava, S., Bist, A. S., Sah, B., Pani, N. K., and Bhambu, P. (2023). Optimizing horizontal scalability in cloud computing using simulated annealing for internet of things. *Measurement: Sensors*, 28:100829.

Taskou, S. K., Rasti, M., and Nardelli, P. H. (2022). Minimizing energy consumption for end-to-end slicing in 5G wireless networks and beyond. *IEEE Wireless Communications and Networking Conference, WCNC*, 2022-April:860–865.

Tchinda, A. P., Shala, B., Lehmann, A., Ghita, B., Walker, D., and Trick, U. (2024). Energy-efficient placement of virtual network functions in a wireless mesh network. *IEEE Access*, 12:64807–64822.