Virtual Network Function Placement and Routing: Formulations and Solutions

Received 26 November 2025, accepted 18 December 2025, date of publication 12 January 2026, date of current version 15 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3651743

RESEARCH ARTICLE

Virtual Network Function Placement and Routing: Formulations and Solutions

RAFAEL FOGAROLLI VIEIRA[1], MATHEUS GABRIEL GOMES PANTOJA[1], CARLOS NATALINO[2], (Senior Member, IEEE), AND DIEGO LISBOA CARDOSO[1]

[1]Post-Graduate Program in Electrical Engineering, Federal University of Pará, Belém, Pará 66075-110, Brazil
[2]Department of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

Corresponding author: Matheus Gabriel Gomes Pantoja (matheus.pantoja@itec.ufpa.br)

The Article Processing Charge for the publication of this research was funded by the Coordination for the Improvement of Higher Education Personnel-CAPES (ROR identifier: 00x0ma614). For open access purposes, the authos have assigned the Creative Commons CC BY license to any accepted version of the article.

ABSTRACT The Virtual Network Function Placement and Routing Problem (VNFP-RP) represents a fundamental challenge for achieving efficiency and scalability in Beyond Fifth-Generation (B5G) and Sixth-Generation (6G) networks. It encompasses both the placement of Virtual Network Functions (VNFs) on available computational resources and the routing of traffic through them. These routes must follow the order defined by Service Function Chains (SFCs) within Network Function Virtualization (NFV)-enabled environments. In this work, we present a complete and reproducible implementation of an established Integer Linear Programming (ILP) model adapted from the literature, integrated into a flexible and publicly available experimental testbed for the research community. This testbed includes multiple network topologies, workload generators, and execution scripts, enabling fair benchmarking, systematic evaluations, and future extensions. Using four real-world topologies from the Survivable Network Design Library (SNDLib) under varying workload levels, we optimize the VNFP-RP under four distinct objective functions: node energy consumption, number of active nodes, number of allocated VNFs, and aggregate system latency. The results reveal substantial differences in computational behavior, with execution times ranging from under 0.1 seconds to more than 6 hours, driven by differences in objective functions and network topologies. These findings highlight how modeling choices impact both runtime and placement structure, providing valuable insights into the trade-offs among efficiency, responsiveness, and resource utilization. Ultimately, this study emphasizes the importance of flexible and adaptive strategies to guide decision-making in the design and operation of B5G and 6G networks.

INDEX TERMS Virtual network function placement and routing, network function virtualization, service function chaining, integer linear programming, sixth-generation.

I. INTRODUCTION

Network Function Virtualization (NFV) has fundamentally transformed network service deployment by replacing traditional hardware-based appliances with software-based Virtual Network Functions (VNFs) [1], [2], [3]. Leveraging advances in cloud computing and virtualization technologies, NFV enables network functions—such as firewalls (FW), traffic monitors (TM), and network address translation (NAT)—to be deployed on general-purpose hardware, such as x86 servers. This decoupling of network functions from proprietary hardware appliances enables substantial reductions in both capital (CAPEX) and operational (OPEX) expenditures [4], [5]. Furthermore, NFV remains pivotal in the evolution toward Beyond Fifth-Generation (B5G) and Sixth-Generation (6G) networks, supporting open and modular architectures such as the Open Radio Access Network (O-RAN). In this context, components such as the virtual O-RAN Distributed Unit (vO-DU), virtual O-RAN Central Unit (vO-CU), and RAN Intelligent Controller (RIC) operate on common hardware and can be orchestrated in a unified manner. This integration promotes flexibility, scalability, and

The associate editor coordinating the review of this manuscript and approving it for publication was Tiago Cruz.

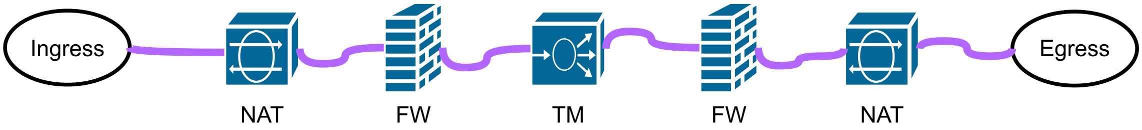

FIGURE 1. Service Function Chain (SFC) – The figure presents an example of an SFC designed to enforce specific policies on network traffic. The data flow enters the system and is routed through an ordered sequence of Virtualized Network Functions (VNFs), including Network Address Translation (NAT), Firewall (FW), and Traffic Monitoring (TM). The execution order of these functions ensures the correct application of security requirements, traffic management policies, and address translation throughout the traffic path within the network.

continuous innovation while optimizing resource utilization and reducing costs [5], [6], [7], [8], [9].

Network services are structured as ordered sequences of VNFs, referred to as Service Function Chains (SFCs), which are designed to meet the specific requirements of each application. These functions process incoming data packets and can classify or manipulate traffic flows based on their characteristics [10], [11]. For example, a network operator may configure Voice over IP (VoIP) traffic to traverse an SFC composed of $NAT \rightarrow FW \rightarrow TM \rightarrow FW \rightarrow NAT$, as illustrated in Figure 1. Virtualization enables VNFs to be instantiated at any time and location, provided that resource and connectivity constraints are satisfied, thereby offering enhanced flexibility in service provisioning [7]. However, this flexibility also introduces challenges related to the placement and routing of VNFs across the physical network [12], [13], [14].

The demand for high-performance services—such as autonomous vehicles, remote surgeries, and augmented reality applications, which require high availability, responsiveness, and ultra-low latency—has grown significantly. This trend, combined with an ever-increasing number of users, has intensified the consumption of network infrastructure resources, particularly in B5G and 6G networks [14], [15], [16]. In this context, the efficient placement of VNFs onto available physical resources, coupled with the appropriate routing of service requests through the underlying physical network, has become a critical factor in the success of NFV [17], [18]. To address these challenges, service providers must adopt strategies that carefully balance diverse and often conflicting requirements, such as minimizing the number of active nodes, reducing latency, and controlling operational costs, particularly energy consumption [6], [19], [20], [21], [22]. An effective VNF placement scheme must achieve a trade-off among these objectives while respecting the capacity constraints of the physical network and ensuring compliance with Service Level Agreements (SLAs) [19], [23]. This set of decisions constitutes the Virtual Network Function Placement and Routing Problem (VNFP-RP), which is recognized as an NP-hard (non-deterministic polynomial-time) problem due to its combinatorial nature [12], [14].

The VNFP-RP is widely recognized as one of the most complex challenges within the NFV domain and plays a central role in ensuring efficient operation in B5G and 6G networks [4], [5], [6]. In this study, we address the VNFP-RP with an emphasis on two fundamental aspects: (i) the optimal placement of VNFs on physical nodes, and (ii) the routing of service demands along paths that comply with latency constraints and satisfy the functional requirements of the applications. To tackle this problem, we implement an Integer Linear Programming (ILP) model adapted from the formulations presented in [4], [12], and [24]. The model optimizes the VNFP-RP under four distinct objective functions: (a) node energy consumption (NEC), (b) number of active nodes (NAN), (c) number of allocated VNFs (NAV), and (d) aggregate system latency (ASL). These objectives represent distinct operational priorities, ranging from energy sustainability and infrastructure efficiency to Quality of Service (QoS) and Quality of Experience (QoE) for latency-sensitive applications. They highlight the importance of assessing the differing impacts associated with each objective function to enable more robust and well-grounded planning decisions.

In this context, the main contributions of this study are threefold: (i) the release a publicly available and fully reproducible testbed—including multiple topologies, demand generators, and execution scripts—for fair comparison, systematic evaluation, and future extensions; (ii) the formalization of the VNFP-RP through an SFC-aware ILP model that jointly decides placement and routing under resource and capacity constraints; and (iii) a comparative analysis of the four objective functions—NEC, NAN, NAV, and ASL—evaluated individually to quantify their specific trade-offs and to assess their respective impacts on network performance. These contributions lay a solid foundation for fostering reproducible research in this domain and for selecting optimization strategies tailored to specific operational requirements. The remainder of this paper is organized as follows. Section II surveys the related work; Section III formalizes the VNFP-RP and presents its mathematical formulation; Section IV discusses the experimental results and comparative findings; and Section V concludes the paper.

II. RELATED WORK

The VNFP-RP, increasingly driven by the widespread adoption of SFCs across diverse application domains, has attracted significant attention from both academia and industry, thereby motivating systematic investigations into strategies for the optimal placement and routing of VNFs [7],

[14]. Recent studies have proposed various formulations of the VNFP-RP, aiming to define placement strategies under distinct system models and operational constraints, as well as optimized routing approaches that ensure the correct delivery of services along the specified SFCs. These works differ in the modeling methodologies adopted, the objective functions prioritized, and the effectiveness of their proposed solutions. Among the most frequently addressed objectives are (i) the minimization of active resources, such as nodes and links [28]; (ii) the reduction of end-to-end latency [29]; and (iii) the minimization of energy consumption [12]. The VNFP-RP has attracted extensive attention in the literature due to its practical relevance for the efficient orchestration of VNFs in B5G and 6G networks enabled by NFV technologies, as well as the inherent complexity of obtaining optimal solutions.

In advanced softwarized network environments, energy efficiency has become a central criterion in VNF orchestration, prompting a growing body of research focused on sustainability. In [22], the authors investigate the sustainable provisioning of SFCs in green mobile edge networks, prioritizing solar-powered servers and dynamically adapting routing according to energy availability and wireless link conditions. An energy-aware VNF placement strategy for wireless mesh networks deployed in emergency scenarios is proposed in [21]. Their multi-objective model aims to extend network lifetime by utilizing such networks as a resilient communication backbone in the event of infrastructure failures. A heuristic based on the Blocking Islands paradigm to reduce energy consumption while satisfying end-to-end latency constraints is proposed in [4]. The proposed approach demonstrates superior scalability and reduced execution time compared to traditional techniques and ILP-based methods. In [5], three heuristics leveraging network state reduction are introduced to minimize energy consumption without compromising SFC acceptance rates.

The stringent latency requirements of emerging applications has driven the development of solutions focused on minimizing end-to-end latency in VNF placement and routing. A mixed-integer linear programming formulation is presented in [29], selecting paths with minimal installation and node activation costs to reduce overall network latency. This work is extended in [23] with two alternative formulations for the VNFP-RP: (i) a Path Formulation, which explicitly incorporates latency constraints and predefined paths, and (ii) an approach focused on optimized function deployment across physical nodes. In [13], a model incorporating edge computing support is proposed, combining a Markov decision process with Q-learning-based algorithms to reduce service rejection rates and optimize real-time resource allocation. Approximation algorithms that jointly optimize resource allocation and latency are introduced in [25], reducing provisioning costs across edge and cloud environments. A heuristic approach based on particle swarm optimization is adopted in [20] to simultaneously minimize overall network latency and the number of active nodes while ensuring service connectivity along the SFCs.

In addition to latency and energy efficiency, other objectives have been extensively investigated in the VNFP-RP literature, including bandwidth reduction, network resilience enhancement, and optimization of computational resource allocation. These efforts aim to improve overall network efficiency while maintaining QoS despite infrastructure constraints and increasing user demand. A heuristic for VNF allocation is proposed in [28] to jointly minimize link utilization and computational resources consumption in the physical infrastructure. A joint optimization algorithm focused on load balancing is introduced in [6], improving VNF distribution and reducing request blocking rates. To improve bandwidth efficiency, an allocation strategy is presented in [26] that reuses existing VNF instances and prioritizes adjacent deployment of new instances to minimize bandwidth consumption during SFC routing. A probabilistic VNF allocation mechanism is proposed in [27], combining dynamic SFC formation based on Dijkstra's algorithm to reduce SFC failures and minimize the latency associated with service composition. The variety of objectives addressed in these studies reflects the multifaceted complexity of the VNFP-RP and underscores the need for optimization strategies capable of balancing competing performance criteria.

Although the literature explores the VNFP-RP through a variety of strategies and objectives, service providers still face the challenge of balancing conflicting goals that impact both network efficiency and user experience (QoS/QoE). However, the majority of existing studies focus either on specific heuristics or on ILP formulations, often without providing open and reproducible implementations. Moreover,

TABLE 1. Comparison of related works based to the considered criteria. A green check mark (✓) denotes that the corresponding aspect is addressed, whereas a red cross (✗) indicates its absence.

	References											
	[4]	[5]	[6]	[13]	[20]	[21]	[22]	[23]	[25]	[26]	[27]	**Our Work**
Placement	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
VNF Chain	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓
Power Consumption	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓
Instance Sharing	✓	✓	✓	✗	✗	✓	✗	✓	✗	✓	✓	✓
Delay	✓	✗	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓
ILP Formulation	✓	✓	✗	✗	✓	✓	✗	✓	✓	✗	✗	✓
Open Source	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

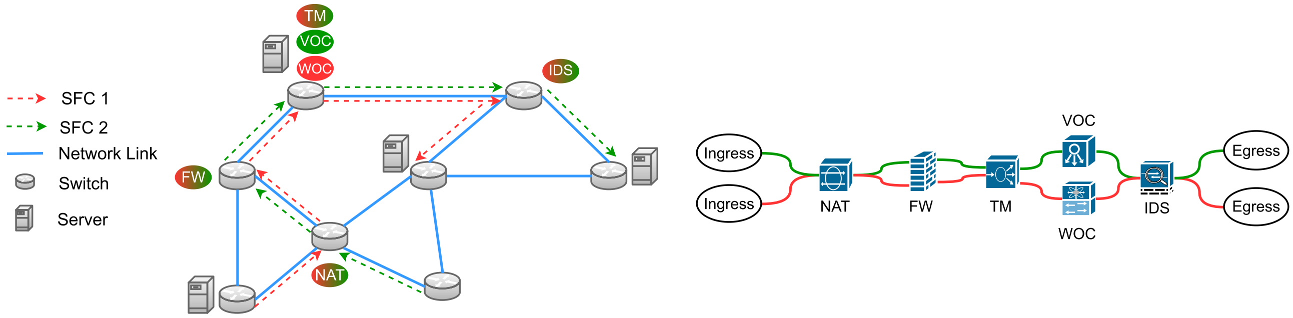

FIGURE 2. Service Function Chains (SFCs) with Virtual Network Functions (VNFs) — The figure illustrates two service chains (SFC 1 and SFC 2), each composed of VNFs such as Network Address Translation (NAT), Firewall (FW), Traffic Monitoring (TM), WAN Optimization (WOC), Intrusion Detection System (IDS), and Video Optimization (VOC). The traffic is processed and routed through these VNFs according to the forwarding policies defined by the service.

systematic comparisons across different objective functions remain limited in the current VNFP-RP literature. This gap underscores the need for reproducible baselines and comprehensive evaluations that enable fair benchmarking and inform decision-making across diverse operational scenarios. Table 1 summarizes the related works reviewed in this sections, outlining their main characteristics.

III. VIRTUAL NETWORK FUNCTION PLACEMENT AND ROUTING PROBLEM

The VNFP-RP involves efficiently placing VNFs across a distributed physical infrastructure and defining feasiable routes for traffic flows between these functions in accordance with the pre-specified SFCs [12], [23]. Solving the VNFP-RP entails jointly deciding: (i) which physical nodes will host the VNFs, subject to computational capacity constraints, and (ii) which paths data flows will traverse, in compliance with topological and datarate limitations. The problem's complexity arises from heterogeneous service requirements, limited network resources, concurrent optimization goals, and the need to address operational criteria such as energy consumption, resource utilization, and latency [11], [23], [30], [31], [32]. Figure 2 illustrates this process, highlighting the routing of traffic between physical nodes and the VNFs required by each request. The following constraints define the conditions for resource allocation and routing, guaranteeing the technical feasibility and consistency of the generated solutions.

- **Node Capacity Constraints:** The placement of VNFs on each physical node must comply with its resource capacity, ensuring that the assigned functions do not exceed available limits—such as Central Processing Unit (CPU), memory, and storage capacities—and preserving infrastructure performance and network stability.
- **Function Capacity Constraints:** The volume of data processed by each VNF instance must not exceed its maximum processing capacity, ensuring stable operation and preventing overload that could compromise QoS and QoE.
- **Service Function Chaining Constraints:** Each service demand must be routed along a path that respects the VNF order defined by its SFC, ensuring that traffic is processed in the correct sequence. The path must be elementary (i.e., cycle-free), and the end-to-end latency must not exceed the maximum threshold defined for each demand.

In the following, we present the formalization of the VNFP-RP model, which is adapted and extended from the ILP formulations proposed in [4], [12], and [24], and therefore maintains the same computational complexity characteristics of these formulations. This general modeling approach is applicable to a broad range of scenarios, including the deployment of virtualized O-RAN components [33], [34] and the optimization of dynamic software-defined satellite–terrestrial integrated networks [35], [36], providing flexibility to accommodate diverse objective functions.

A. SYSTEM MODEL

The telecommunications network is formally modeled as a bidirectional physical graph $G(N, L)$, where N denotes the set of nodes and L represents the set of links (or arcs) connecting them. In this model, each node $i \in N$ is assumed to be equipped with hardware resources capable of processing, hosting, or switching VNFs in the physical network. This abstraction not only simplifies the representation of the infrastructure but also reduces the graph size, thereby lowering computational complexity. Each node $i \in N$ provides a set of resources R, typically including CPU, memory, and storage capacities. For each link $(i,j) \in L$, $r_{(i,j)}$ denotes the available data rate, whereas $d_{(i,j)}$ represents the link delay, which is assumed to be proportional to the physical distance between nodes.

Let G denote the set of demands, where each demand $g \in G$ is represented as a tuple:

$$g(v_{(s,g)}, v_{(d,g)}, d_g, r_g, c_g) \tag{1}$$

TABLE 2. Main notations of the sets, parameters, and variables used in the VNFP-RP formulation.

Sets	Description	
N	Set of nodes	
L	Set of links	
R	Set of resources	
G	Set of demands	
F	Set of VNFs	
S	Set of services	
E_g	Set of virtual links	
V_g	Set of virtual nodes	
Parameters	**Description**	
$v_{(s,g)}$	Source node of demand g	
$v_{(d,g)}$	Destination node of demand g	
d_g	Maximum latency of demand g	
r_g	Data rate of demand g	
c_g	Ordered set of VNFs of demand g	
p_i^{idle}	Idle power of node i	
p_i^{max}	Maximum power of node i	
d_f	Processing delay of VNF f	
r_f	Processing capacity of VNF f	
$\phi_{f,r}$	Resource required of VNF f	
$c_{r,i}^T$	Processing Capacity of node i	
$r_{(i,j)}$	Available data rate of physical link (i,j)	
$d_{(i,j)}$	Latency of link (i,j)	
M	Big positive number	
Variables	**Description**	**Type**
x_i	1, if node i is active and hosts a VNF; 0, otherwise	Binary
y_i	1, if node i is active and used to switch a demand; 0, otherwise	Binary
$l_{(i,j)}$	1, if link (i,j) is active; 0, otherwise	Binary
$w_{g,(f,f')}^{(i,j)}$	1, if the virtual link (f,f') corresponding to demand g is mapped onto physical link (i,j); 0, otherwise	Binary
$u_{g,f,p,i}$	1, if demand g is processed by VNF f allocated at node i and occupying position p in the SFC; 0, otherwise	Binary
$z_{f,i}$	Number of instances of VNFs f placed at node i	Non Negative Integers

$v_{(s,g)}$ and $v_{(d,g)}$ denote the source and destination nodes of demand g, respectively. dg denotes the maximum end-to-end latency allowed for the service associated with demand g, whereas r_g represents the data rate required to serve that demand. The SFC of each demand g is defined as an ordered set of VNFs, denoted by $c_g = f_1 \to f_2 \to \cdots \to f_p$. This VNF chain is modeled as a virtual graph $G(V_g, E_g)$, where V_g denotes the set of virtual nodes and E_g represents the set of virtual links corresponding to demand g. Each virtual node represents either a VNF belonging to the set F, or one of the source or destination terminals of the demand. For each function $f \in F$, r_f denotes the maximum processing capacity available for f, whereas d_f represents the latency introduced by that function into the service. The virtual links connect the virtual nodes following the sequence defined by the SFC. Thus, the virtual network $G(V_g, E_g)$ must be embedded onto the physical network $G(N, L)$, ensuring that all requirements of demand g are satisfied. In this mapping, the virtual source and destination nodes are embedded on the corresponding nodes, thus preserving the end-to-end semantics of each demand. Each function f can be shared among multiple demands, provided that its maximum data rate capacity (expressed in megabits per second, Mbps) is not exceeded. Before presenting the mathematical formulation of the VNFP-RP, we introduce the notations used throughout this paper, including the sets, parameters, and variables summarized in Table 2.

B. CONSTRAINTS

The following constraints formalize the VNFP-RP, delimiting the feasible solution space in accordance to technical and operational requirements. Each constraint is mathematically formulated to ensure solution feasibility, addressing processing capacity, resource utilization efficiency, and compliance with QoS requirements.

$$\sum_{i \in N} u_{g,f,p,i} = 1, \forall g \in G, \forall f \in c_g, \forall p \in c_g \tag{2}$$

Equation (2) ensures that, for each function f defined in the SFC of demand g, a single node i is selected to host it at position p. This constraint guarantees the complete and ordered mapping of the SFC on the physical infrastructure, preserving the functional sequence defined for the service.

$$u_{g,f,p,i} \leq z_{f,i}, \forall i \in N, \forall f \in F, \forall p \in c_g, \forall g \in G \tag{3}$$

Equation (3) ensures that a function f can be assigned to demand g on node i only if at least one instance of f has been previously allocated on that node. This constraint guarantees consistency between the processing decision and the availability of the function within the physical infrastructure.

$$\sum_{f \in F} \phi_{f,r} z_{f,i} \leq C_{r,i}^T, \forall i \in N, \forall r \in R \tag{4}$$

Equation (4) imposes that, for each node i and each resource type r, the total consumption resulting from VNF allocation—calculated as the sum of the individual demands $\phi_{f,r}$ multiplied by the number of instances $z_{f,i}$—does not exceed the available capacity $C_{r,i}^T$. This constraint ensures that the utilization of resources remains within the operational limits of each node.

$$\sum_{g \in G} \sum_{p \in c_g} r_g u_{g,f,p,i} \leq r_f z_{f,i}, \forall i \in N, \forall f \in F \tag{5}$$

Equation (5) defines the data rate capacity constraint for each function f on each node i. The left-hand side represents the total traffic volume assigned to function f on node i, considering all demands and positions within the chain. This value must not exceed the product of the maximum data rate r_f and the number of instances $z_{f,i}$ allocated on node i. This constraint ensures that no function instance is overloaded beyond its nominal processing capacity.

$$\sum_{(f,f') \in E_g} \sum_{g \in G} r_g w_{g,(f,f')}^{(i,j)} \leq r_{(i,j)}, \forall (i,j) \in L \tag{6}$$

Equation (6) imposes that the sum of the data rates required by all demands routed through link (i,j) does not exceed its available capacity $r_{(i,j)}$. This constraint ensures compliance with link capacity limits, preventing overload and guaranteeing routing feasibility.

$$u_{g,f_s,p_f,v_{(s,g)}} = 1, \quad \forall g \in G \tag{7}$$

$$u_{g,f_d,p_l,v_{(d,g)}} = 1, \quad \forall g \in G \tag{8}$$

Equations (7) and (8) establish that, for each demand g, the initial VNF (f_s) must be allocated on the source node $v_{(s,g)}$, and the final VNF (f_d) on the destination node $v_{(d,g)}$. These constraints ensure the correct embedding of the SFC terminal points into the physical network. In this model, f_s and f_d denote the starting and ending functions of the processing chain for all demands. Similarly, p_f and p_l specify the positions of the first and last functions in the SFC associated with each demand.

$$\sum_{(i,j)\in L} w^{(i,j)}_{g,(f,f')} - \sum_{(j,i)\in L} w^{(j,i)}_{g,(f,f')} = u_{g,f,p_1,i} - u_{g,f',p_2,i},$$
$$\forall i \in N, \forall (f,f') \in V_g, \forall p_1 \in c_g, \forall p_2 \in c_g, \forall g \in G \tag{9}$$

Equation (9) enforces the flow conservation constraint for traffic between two consecutive VNFs (f, f') in the SFC associated with demand g. This equation ensures that, for each node i, the difference between the outgoing and incoming links used to route the traffic of (f, f') is consistent with the presence of functions f or f' on that node. When node i hosts function f (at position p_1), it acts as the source of the flow; when it hosts function f' (at position p_2), it acts as the destination. In all other cases—i.e., when node i does not host either function—the traffic must be forwarded in transit, with the amount of incoming traffic equal to that of outgoing traffic. This constraint is essential to ensure that routing between VNFs follows continuous and feasible paths within the physical network.

$$\sum_{j\in N} \sum_{(f,f')\in E_g} w^{(i,j)}_{g,(f,f')} \leq 1, \quad \forall i \in N, \ \forall g \in G \tag{10}$$

$$\sum_{i\in N} \sum_{(f,f')\in E_g} w^{(i,j)}_{g,(f,f')} \leq 1, \quad \forall j \in N, \ \forall g \in G \tag{11}$$

Equations (10) and (11) impose node traversal uniqueness constraints on the routing paths associated with demand g. Equation (10) limits the outgoing traffic from node i to at most one link for each pair of consecutive VNFs (f, f') in the SFC of demand g. Equation (11) restricts the incoming traffic to node j under the same conditions. Together, these constraints prevent the same node from being traversed multiple times by the traffic of a given demand, ensuring that the path remains elementary (i.e., cycle-free). This condition is essential to guarantee routing integrity, ensuring compliance with the SFC and structure and preventing unnecessary node overload.

$$latency^p = \sum_{i\in N} \sum_{f\in F} \sum_{p\in c_g} d_f u_{g,f,p,i}, \quad \forall g \in G \tag{12}$$

$$latency^t = \sum_{(i,j)\in L} \sum_{(f,f')\in E_g} d_{(i,j)} w^{(i,j)}_{g,(f,f')}, \quad \forall g \in G \tag{13}$$

$$latency^p + latency^t \leq d_g, \quad \forall g \in G \tag{14}$$

Equation (14) imposes the end-to-end latency constraint for each demand g, ensuring that the aggregated processing and transmission delays along the SFC does not exceed the maximum threshold d_g specified by the service. This constraint consists of two components: (i) the sum of the processing latencies d_f introduced by each VNF allocated to serve demand g, as defined in (12); and (ii) the sum of the transmission latencies $d_{(i,j)}$ associated with the links used to route the virtual links specified by the SFC, as defined in (13). By ensuring that the total delay remains within the latency budget defined for the service, this constraint is essential for maintaining the QoS and QoE of latency-sensitive applications, including real-time services.

$$\sum_{p\in c_g} u_{g,f,p,i} \leq z_{f,i}, \quad \forall i \in N, f \in F, g \in G \tag{15}$$

Equation (15) enforces consistency between the use of VNFs and the number of instances allocated to each node. Specifically, it ensures that a given function f on node i is used in at most one position of the SFC associated with demand g. This prevents a VNF from being redundantly assigned to multiple positions of the same demand's SFC on a single node. Moreover, it guarantees that the number of times function f is used by demand g does not exceed the number of instances $z_{f,i}$ allocated on node i, thereby ensuring mapping feasibility and coherent use of infrastructure resources.

$$\sum_{(f,f')\in E_g} \sum_{g\in G} w^{(i,j)}_{g,(f,f')} \leq l_{(i,j)} \times M, \forall i \in N, \quad \forall j \in N \tag{16}$$

$$\sum_{j\in N} l_{(i,j)} + l_{(j,i)} \leq y_i \times M, \quad \forall i \in N \tag{17}$$

$$\sum_{f\in F} z_{f,i} \leq x_i \times M, \quad \forall i \in N \tag{18}$$

Finally, we introduce three binary indicator variables that specify whether each node or link is online or offline. These variables are used in Equations (16), (17), and (18), which impose activation constraints on links and on nodes responsible for switching or processing, respectively. The constant M represents a sufficiently large number that enables or disables these constraints depending on the activation state of each component.

- Equation (16) enforces that a link (i, j) may be used for routing only when it is activated.
- Equation (17) enforces that node i may perform switching tasks only when it is active.
- Equation (18) enforces that VNFs can be placed to node i only if the node is active.

C. OBJECTIVE FUNCTIONS

The VNFP-RP formulation can be driven by different objective functions, depending on the network's operational goals. In this work, we consider four distinct objective functions, each addressing a specific aspect of VNF orchestration: energy consumption [4], [12], QoS and QoE [37], infrastructure utilization [38], and placement complexity [29]. These objectives are analyzed individually, enabling a comparative evaluation of the trade-offs involved. The following subsections detail each objective function and its corresponding mathematical formulation.

1) MINIMIZATION OF NODES ENERGY CONSUMPTION

Among the different components that contribute to server energy consumption, the CPU is the dominant one, exceeding the impact of memory, storage, and network interfaces [39]. Studies indicate that server power usage varies proportionally with workload, ranging from a baseline value at idle state to a peak value under full utilization [40], [41]. To model this behavior, the energy consumption of a node $i \in N$ is defined as a function of the CPU usage imposed by the VNFs allocated to its resources. The energy consumption of node i is defined by:

$$Power_i = P_i^{idle} + (P_i^{max} - P_i^{idle}) \times \theta_{cpu}, \quad \forall\, i \in N \tag{19}$$

P_i^{idle} represents the idle-state power consumption (at 0% CPU utilization). P_i^{max} corresponds to the power consumption under full load (100% utilization). θ_{cpu} is the CPU utilization rate, defined as the ratio between the total CPU used and the total available CPU capacity. θ_{cpu} is defined by:

$$\theta_{cpu} = \phi_{f,r} / c^T_{r,i} \tag{20}$$

Let r as the type of computational resource (e.g., CPU). Assuming that CPU utilization is proportional to the aggregate computational demand of the allocated VNFs, the objective function that minimizes the total energy consumption of all processing nodes is defined as follows:"

$$\text{Minimize} \sum_{i\in N}(P_i^{idle} x_i + (P_i^{max} - P_i^{idle}) \sum_{f\in F}\sum_{r\in R} \frac{\phi_{f,r}}{c^T_{r,i}} z_{f,i}) \tag{21}$$

The objective function aims to minimize the total energy consumption of all processing nodes, thereby reducing the overall energy footprint of the physical network. This approach promotes energy-efficient VNF placement while satisfying the computational and operational requirements of each demand, contributing to more sustainable use of computational resources.

2) MINIMIZATION OF AGGREGATE SYSTEM LATENCY

Minimizing latency is a critical objective in B5G and 6G networks, especially for applications that require high responsiveness and have stringent delay constraints. In this context, latency is modeled as the transmission delay incurred along the links traversed by the SFCs. The objective function minimizes the aggregate system latency, defined as:

$$\text{Minimize} \sum_{(i,j)\in N} \sum_{(f,f')\in F} \sum_{g\in G} w^{(i,j)}_{g,(f,f')} \times d_{(i,j)} \tag{22}$$

Equation (22) computes the total transmission delay incurred on the links (i, j) used to route the virtual links (f, f') specified by the SFC of demand g. Minimizing this metric ensures compliance with QoS and QoE requirements, especially in critical scenarios and for latency-sensitive applications.

3) MINIMIZATION OF THE NUMBER OF ACTIVE NODES

Optimizing the utilization of physical resources is a fundamental strategy for ensuring efficient operation of the physical network. Minimizing the number of active nodes reduces overall energy consumption, thereby lowering operational expenditures and promoting more sustainable network operation. In addition, reducing device usage mitigates hardware wear, prolongs the lifespan of physical components, and decreases the frequency of maintenance requirements.

$$\text{Minimize} \sum_{i\in N} x_i \tag{23}$$

Equation (23) minimizes the total number of active nodes assigned to processing tasks (x_i)). This approach enables more scalable and efficient network operation while ensuring that service demands are satisfied through the intelligent allocation of available physical resources.

4) MINIMIZATION OF THE NUMBER OF ALLOCATED VNFs

Minimizing the number of allocated VNF instances aims to limit their dispersion across the network, thereby concentrating processing in fewer locations and reducing overall management complexity. This approach contributes to conserving computational resources and simplifying infrastructure operation.

$$\sum_{f\in F} \sum_{i\in N} z_{f,i} \tag{24}$$

Equation (24) minimizes the total number of VNF instances, promoting more compact placements and reducing the spatial dispersion of VNFs along the SFC execution.

IV. PERFORMANCE EVALUATION

This section presents the experimental analysis of the ILP model introduced in Section III, evaluating the impact of the different objective functions on the number of allocated VNFs, overall energy consumption, average per-demand latency, and the computational effort required to solve the instances. The model was implemented in Python 3.12 [42] using the Pyomo library [43] and executed on a machine equipped with an AMD Ryzen 5 8600G processor (6 cores, 12 threads, up to 5.0GHz), 16GB of DDR5 RAM at 5200MHz, running Ubuntu 20.04 LTS [44]. The CPLEX solver was used to solve the optimization models [45]. Each execution was limited to 21,600 seconds. The instances exhibited large variations in initial optimality gaps: dense scenarios—such as the New York topology under high-demand conditions—often began above 90% and required several hours to close the gap. Given this behavior, the selected time limit offered a balanced compromise that constrained computational effort without compromising solution quality.

To capture the variability of service requests, 50 independent instances were generated for each scenario, following the methodological practices adopted in [4], [12], and [46]. In each instance, the service demands (source, destination, and associated services) were randomized while keeping the network topology and resource capacities fixed. The reported results correspond to the average values obtained across all instances of each scenario. To ensure experimental reproducibility, all random components of the instance generator—including the selection of sources, destinations, SFCs, and service demands—were produced using a fixed random seed. Moreover, the same demand sets were used for all four objective functions, guaranteeing that the comparison among objectives is performed under strictly equivalent conditions. This design eliminates variability caused by input randomness and ensures that differences in the results arise solely from the optimization criteria, not from distinct traffic scenarios. Table 3 summarizes all experimental settings. The full implementation of the model is publicly available at [47].

TABLE 3. Summary of the experimental environment.

Component	Specification
Processor	AMD Ryzen 5 8600G
Memory	16GB DDR5
Operating System	Ubuntu 20.04 LTS
Programming Language	Python 3.12
Modeling Framework	Pyomo 2024
Solver	IBM CPLEX
Execution Time Limit	21,600 seconds per instance
Number of Instances	50 per scenario
Random Seed	Fixed across all experiments

A. PARAMETER SETTINGS

In telecommunication networks, each service type is supported by a specific SFC. In this study, four types of traffic flows are considered: Video Streaming (VS), Web Services (WS), Online Gaming (OG), and VoIP, as summarized in Table 4. For each service, the following parameters were specified: (i) the ordered sequence of required VNFs (the SFC), (ii) the required data rate, (iii) the end-to-end latency, and (iv) the traffic distribution percentage, based on [4], [12], and [46]. Six VNFs were considered in this work, as presented in Table 5: WAN Optimization Controller (WOC), Video Optimization Controller (VOC), Intrusion Detection System (IDS), NAT, FW, and TM. For each function, the required CPU capacity and data rate are specified. The processing delay is fixed at 10ms for all functions [4]. Sets of service demands were generated by randomly selecting source–destination node pairs across the network, ensuring that $v_{(s,g)} \neq v_{(d,g)}$ for each demand g. Each demand is associated with a service type, following the percentage distribution indicated in Table 4.

For the model analysis, four topologies from Survivable Network Design Library (SNDLib) were used [48]: (i) DFN-BWIN, with 10 nodes and 45 links; (ii) Abilene, with 12 nodes and 15 links; (iii) Polska, with 12 nodes and 18 links; and (iv) New York, with 16 nodes and 51 links. These network topologies provide structural diversity and representativeness for evaluating the VNFP-RP. DFN-BWIN and New York exhibit denser connectivity, enabling the analysis of placement and routing decisions under higher path redundancy. In contrast, Abilene and Polska are sparser networks, highlighting scenarios in which routing alternatives are limited and placement decisions become more constrained. Taken together, the four topologies span a broad spectrum of sizes, link densities, and connectivity patterns, allowing the model to be assessed under heterogeneous and realistic backbone conditions. Each link was configured with a capacity of 10$Gbps$ (gigabits per second) [12]. The propagation delay of each link was estimated based on its geographical length, computed using the Haversine formula and assuming fiber-optic transmission with an approximate speed of light of $2 \times 10^5 km/ms$ [49].

The processing node energy consumption model is described in Subsection III-C1. For the simulations, two sets of energy parameters were considered to account for different power profiles. The first set corresponds to an idle-state and maximum power consumption of $(p_i^{idle}, p_i^{max}) = (46.8\,\text{W}, 210\,\text{W})$, while the second set adopts $(p_i^{idle}, p_i^{max}) = (69.3\,\text{W}, 367\,\text{W})$. These parameter pairs were used to assess the sensitivity of the model to variations in processing node energy characteristics. Each substrate node is equipped with a processor featuring 64 CPU cores, according to the technical specifications reported in [50].

For the experiments, a total of 120 scenarios were evaluated. These scenarios result from combining the four selected SNDLib topologies with traffic loads ranging from 1 to 30 service demands per processing node. For each scenario, all structural and resource parameters of the topology remain fixed, while only the number of generated service demands varies. This setup enables a systematic assessment of the VNFP-RP model under heterogeneous network structures and progressively increasing load conditions. The four objective functions considered in this study—NEC, NAN, NAV, and ASL—are optimized independently, resulting in four separate ILP formulations. Due to the NP-hard nature of the VNFP-RP and the complexity of solving ILP-based models, the simulations were conducted on small-sized topologies to ensure computational tractability and allow the full execution of all instances within the specified time limit.

B. NUMBER OF ALLOCATED VNFs

Figure 3 illustrates the evolution of the total number of allocated VNFs as the number of demands increases across

TABLE 4. Characteristics of the considered service types, including their respective Service Function Chains (SFCs), required data rate, allowed end-to-end latency, and share of total traffic. The Virtual Network Functions (VNFs) used in the SFCs include: NAT (Network Address Translation), FW (Firewall), TM (Traffic Monitor), WOC (WAN Optimization Controller), VOC (Video Optimization Controller), and IDS (Intrusion Detection System) [12], [46].

Service Type	Web Service	VoIP	Video Streaming	Online Gaming
VNF Chain	NAT-FW-TM-WOC-IDS	NAT-FW-TM-FW-NAT	NAT-FW-TM-VOC-IDS	NAT-FW-VOC-WOC-IDS
Data rate	100 kbps	64 kbps	4 Mbps	50 kbps
Latency	500 ms	100 ms	100 ms	60 ms
% Traffic	18.2%	11.8%	69.9%	0.1%

TABLE 5. Parameters used for the Virtual Network Functions (VNFs), including CPU requirements and maximum supported data rate [12], [46].

VNFs	CPU	Data rate
NAT	2	500 Mbps
FW	8	400 Mbps
TM	1	200 Mbps
WOC	2	580 Mbps
VOC	2	300 Mbps
IDS	8	600 Mbps

the four evaluated topologies, considering only the ASL objective function. The ASL objective consistently resulted in the highest number of VNF allocations, exceeding the number of allocated VNFs in the largest scenarios—such as those of the New York topology—by more than one hundred times, with an average increase of approximately forty times across all scenarios. This behavior is explained by its placement strategy, which enforces strict latency minimization and, as a result, favors replicating functions across geographically dispersed nodes to reduce hop counts and response times. Although effective in reducing delay, this replication amplifies resource usage and directly increases energy consumption, as more servers must remain active to support the replicated instances, further increasing the complexity of orchestration. Minor fluctuations in the number of allocated VNFs are also observed as the number of demands increases, with average variations of approximately 8% and slightly higher values in scenarios characterized by greater service diversity.

The NEC, NAN, and NAV objectives consistently maintained the minimum required set of eight VNFs in all topologies and demand scenarios, with statistically significant differences relative to ASL ($p < 2.8 \times 10^{-39}$). Although minimizing the number of allocated VNFs is not an explicit goal of the energy-oriented and node-oriented approaches, the results show that it naturally emerges as a byproduct of their optimization strategies. By prioritizing node deactivation, these objectives concentrate allocations onto a smaller subset of nodes, thereby avoiding the activation of additional nodes while preventing overutilization of those already in operation. These findings highlight the direct influence of the optimization criterion on VNF replication and infrastructure resource efficiency, reinforcing its critical role in balancing service performance and structural cost—particularly in large-scale scenarios where efficient resource utilization becomes essential.

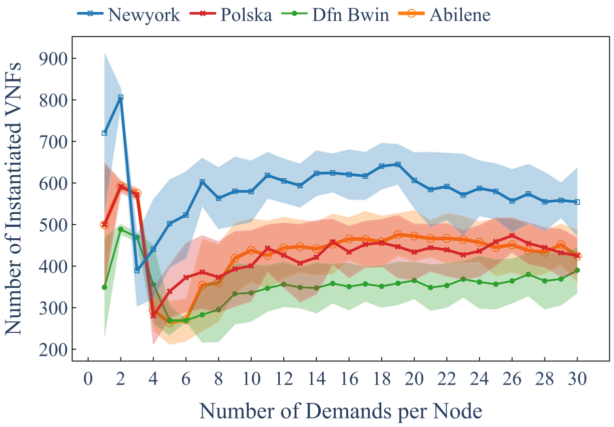

FIGURE 3. Total number of allocated Virtual Network Functions (VNFs) as a function of the number of demands per node in the Abilene, DFN-BWIN, Polska, and New York topologies. The Aggregate System Latency (ASL) objective consistently produced the highest number of instances, with variations driven by service diversity and functional chain configurations. In contrast, the Node Energy Consumption (NEC), Number of Active Nodes (NAN), and Number of Allocated VNFs (NAV) objectives always allocated the minimum required number of functions (eight VNFs) across all scenarios and are therefore not displayed in the figure. In extreme cases, the ASL allocated up to nearly two orders of magnitude more functions than the alternatives.

C. ENERGY CONSUMPTION EVALUATION

Figure 4 illustrates the total energy consumption of the nodes as the number of demands increases in the New York topology. The results are expressed as normalized percentages, enabling direct comparisons across objective functions and topologies with distinct structural characteristics. Only the results for the New York topology are reported, as they are representative of the overall behavior observed. The analysis reveals that the energy consumption curves exhibit highly similar trends across all topologies, with the maximum post-normalization variance of 5.32×10^{-7}. Furthermore, a separate sensitivity analysis using two distinct energy parameter values showed no measurable difference in the results, with a 0% variation across all scenarios.

Among the four objectives, ASL consistently produced the highest energy consumption across all load levels, reaching full (100%) utilization. This outcome follows naturally from its latency-driven placement strategy, which keeps the entire infrastructure active. In contrast, both NEC and NAN maintained energy usage at approximately 4% in all scenarios. This behavior reflects the model's preference for activating only the minimal set of nodes required for feasibility, inherently stabilizing energy consumption. Considering that an idle node consumes approximately 22% of its nominal power, restricting activation to this minimal subset substantially improves energy efficiency while preserving complete allocation coverage. The NAV objective exhibited an intermediate energy consumption profile, stabilizing at approximately 25% across the evaluated topologies. This occurs because minimizing the number of allocated functions consolidates processing on fewer nodes, though it does not explicitly limit node activation, allowing the model to maintain a larger active subset than energy-optimal solutions.

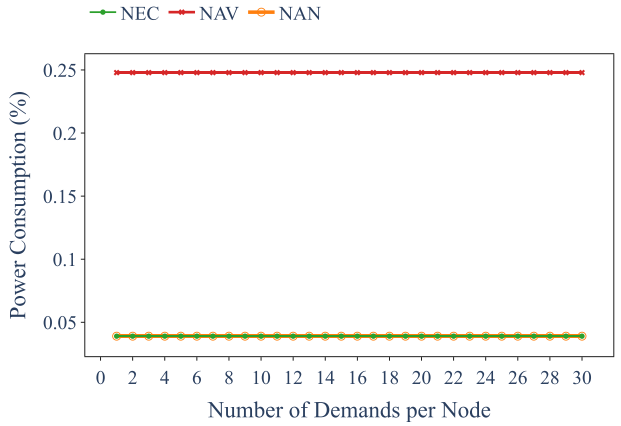

FIGURE 4. **Average normalized energy consumption (%) as a function of the number of demands per node in the New York topology. The figure compares results for three objective functions: (i) Node Energy Consumption (NEC), (ii) Number of Allocated VNFs (NAV), and (iii) Number of Active Nodes (NAN). The objective function Aggregate System Latency was omitted, since it consistently yielded 100% energy consumption across all scenarios. Similar patterns were observed in the other evaluated topologies (Abilene, Polska, and DFN-BWIN), with no evidence of infrastructure saturation even under increased load.**

Statistical validation reinforces these findings, with all pairwise comparisons yielding p-values near zero—except for NEC and NAN, whose p-value equals 1 due to their identical behavior. Overall, the comparative analysis of the four objective functions highlights clear trade-offs between performance and energy efficiency, revealing distinct operating regimes shaped by each optimization criterion. The stability observed across load increments further indicates that the previously allocated resources were sufficient to accommodate the increasing demand without requiring additional node activation or VNF instantiation, confirming that the network remained well below its saturation point under the evaluated scenarios.

D. AVERAGE LATENCY PER DEMAND

Figure 5 shows the evolution of average demand latency as the number of demands increases across the four evaluated topologies. Latency was computed as the sum of processing delays introduced by the VNFs ($10ms$) and the transmission delays incurred along the SFCs supporting each demand. This metric directly reflects QoS and QoE, making it a critical indicator of overall network performance. All demands were served within acceptable latency thresholds, ensuring compliance with QoS requirements even as the load increased.

As expected, the ASL objective achieved the lowest average response times across all scenarios and topologies. Its latency values remained tightly bounded, with a variation of only 0.05 ms, which corresponds to an extremely low variance in all topologies (e.g., 1.7×10^{-10} in DFN-BWIN and 1.8×10^{-8} in Abilene). This narrow dispersion confirms that ASL's performance is not an artifact of rounding or normalization but rather reflects a genuinely stable operational behavior. Moreover, its latency evolved linearly as the load increased, indicating a strong capacity to preserve responsiveness even under significant demand growth. This performance results from explicitly minimizing routing delay between VNFs along the SFCs—although this benefit comes at the expense of substantially higher OPEX due to the required increase in resource usage.

The NEC and NAV objectives exhibited the highest average response times, with transmission delay increasing steadily as the load grew. In the most pronounced case (NewYork topology), transmission delay rose by approximately 4.27 ms, corresponding to a relative increase of about 165% from the lowest to the highest load level. In these cases, demand forwarding is governed solely by the need to satisfy connectivity and maximum-latency constraints, without any mechanism dedicated to minimizing end-to-end delay; consequently, the system produces substantially higher response times even though all demands are successfully routed. Quantitatively, the impact of topology is clear: the Polska topology reached an average transmission delay of only 0.02 ms, whereas Abilene—despite having the same number of nodes—reached 0.1 ms, a difference of more than five times attributable exclusively to connectivity patterns. Conversely, the New York topology exhibited the largest transmission delay across all scenarios, with an average delay nearly 27 times higher than Polska and 6 times higher than Abilene. These results confirm that compact and well-connected topologies enhance temporal performance even without explicit latency-oriented optimization, whereas larger networks—despite higher link density—incur longer transit times between functions due to increased spatial dispersion and routing depth.

Overall, the results reveal a clear trade-off between transmission delay and infrastructure efficiency. Objectives that prioritize latency naturally incur higher resource usage—ASL, for example, keeps transmission delay below 0.002 ms in the DFN-BWIN topology while requiring up to four times more active processing capacity than NEC and NAV. Conversely, resource-saving strategies decrease node activation and energy consumption but lead to transmission delays up compared with the most latency-efficient configurations. Statistical analysis reinforces this distinction: for most topologies and objective-function comparisons,

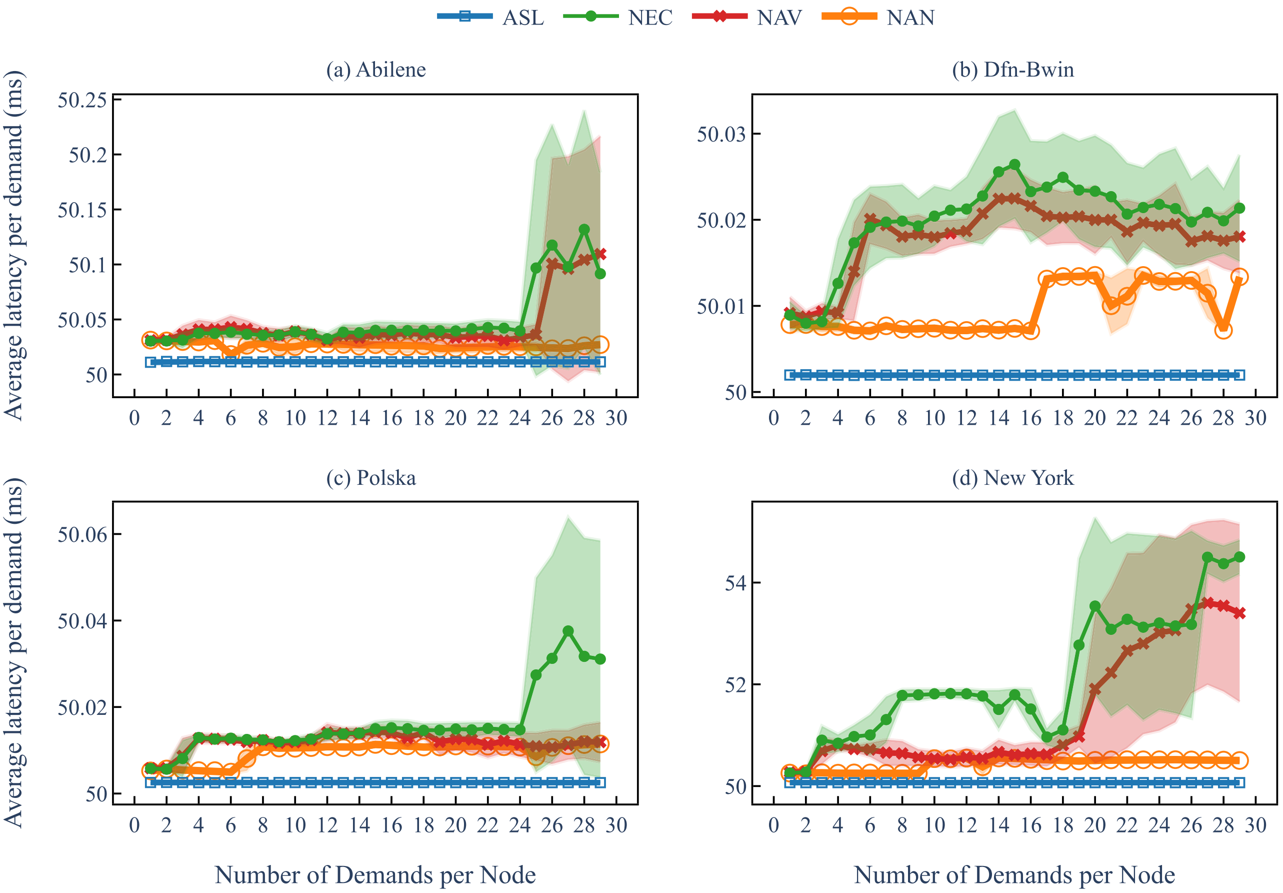

FIGURE 5. Average latency per demand as a function of the number of demands per node in: (a) Abilene, (b) DFN-BWIN, (c) Polska, and (d) New York topologies. The objective function Aggregate System Latency (ASL) consistently achieved the lowest values across all scenarios, maintaining stable behavior even under increasing load. In contrast, the Node Energy Consumption (NEN), Number of Active Nodes (NAN), and Number of Allocated VNFs (NAV) objectives exhibited higher latency levels. This outcome reflects their placement strategies, which concentrate allocations on a reduced set of nodes to optimize resource usage, but at the cost of longer service paths and, consequently, increased end-to-end latency. Nonetheless, all latency values remained within acceptable thresholds for service operation, ensuring adequate demand fulfillment.

the p-values were below 0.03, indicating significant performance differences. These findings highlight that the chosen optimization criterion directly shapes the network's operating regime, highlighting the importance of aligning the objective function with application-specific performance requirements—particularly in scenarios where ultra-low-latency operation is essential.

E. TIME COST COMPARISONS

Figure 6 presents the solution time (in seconds) required by the solver to obtain a feasible solution across all evaluated scenarios, considering only the optimization phase. As expected, the execution time increases with the number of demands, driven by the combinatorial growth of the solution space. Even for the lowest demand levels, the initial processing times are strictly greater than zero, indicating that a non-negligible computational effort is required to configure the first feasible allocations. As the demand load increases, the growth pattern diverges significantly across objective functions and network topologies, revealing critical aspects of computational complexity and underscoring the interplay between problem scale, structural constraints, and solver performance.

The evaluated topologies exhibited distinct computational behaviors determined by their structural characteristics. Abilene and Polska showed similar processing times, although Polska required, on average, about 12.5% more time for most objectives. The exception was the NAV objective, in which Polska was approximately 22% faster. This superior performance results from its higher connectivity, which increases the number of viable paths and facilitates satisfying service chains with the smallest possible number of allocated VNFs. Since the NAV objective does not aim to minimize latency, energy consumption, or server usage, the solution process focuses primarily on finding any feasible route that

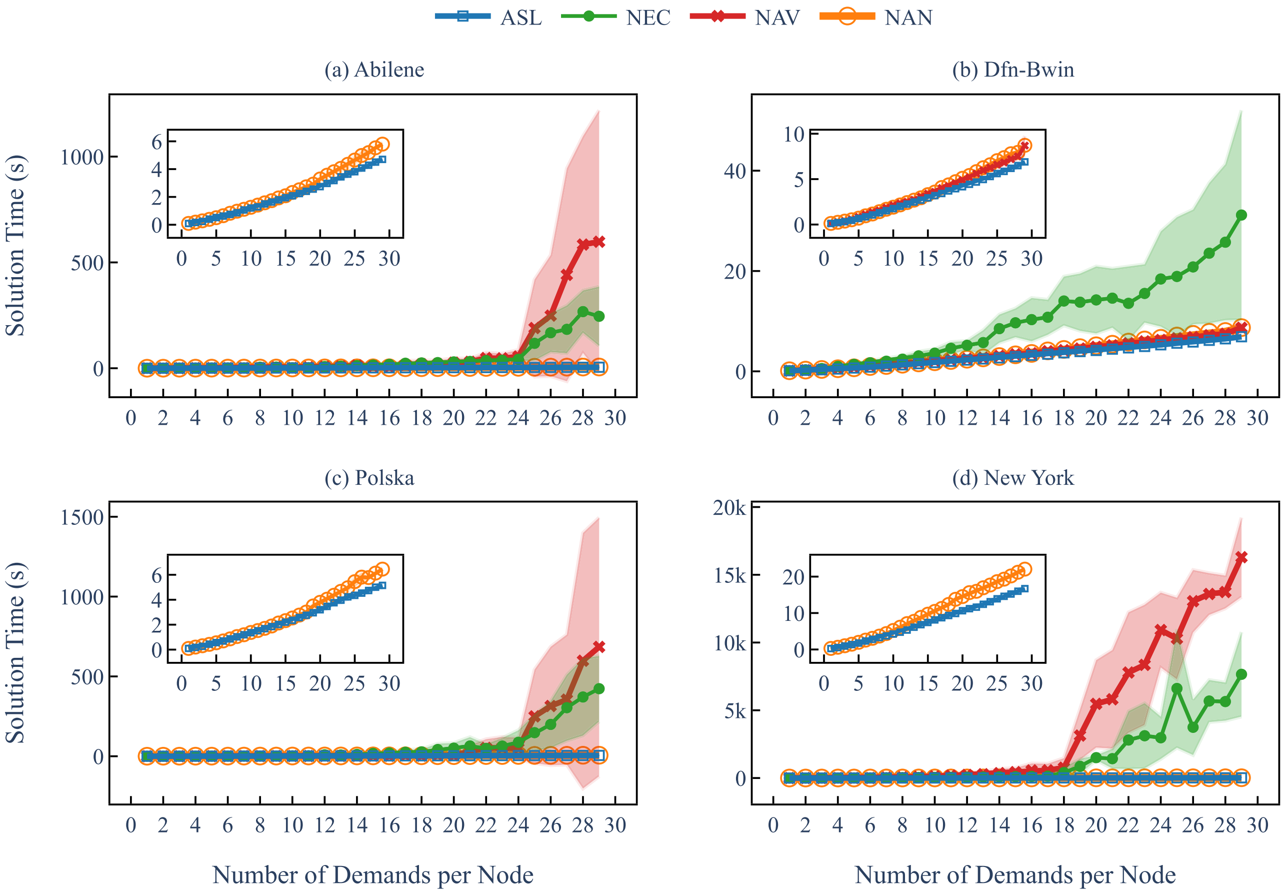

FIGURE 6. Solution time (seconds) required by the solver as a function of the number of demands per node in the topologies: (a) Abilene, (b) DFN-BWIN, (c) Polska, and (d) New York. Each plot reports the four objective functions (Nodes Energy Consumption (NEN), Number of Active Nodes (NAN), Number of Allocated VNFs (NAV), and Aggregate System Latency (ASL)), with insets zooming into the low-load region to show that initial solution times are strictly greater than zero. The highest computational costs occur for NEN and NAV, particularly in the New York topology, where the solver often approached the time limit of 21.600 seconds. In contrast, ASL and NAN exhibit considerably lower solution times, reflecting reduced structural complexity and more tractable solution spaces. The DFN-BWIN topology, being the smallest, consistently yielded the lowest solution times, highlighting the strong impact of network size on computational effort.

reduces redundant allocations. In a denser topology, this search is subject to fewer constraints and can therefore be solved more quickly. Topology DFN-BWIN consistently yielded the lowest solving times, with average values remaining below 37 s even in the most demanding scenarios. It achieved the best performance for the NEC and NAV objectives compared to Abilene and Polska, while presenting worse performance for NAN and ASL relative to these same topologies. This contrasting outcome reflects how DFN-BWIN's dense connectivity favors objectives focused on energy-efficient and allocation-efficient placements, while being less advantageous for objectives that are more sensitive to node activation or path length. This behavior follows directly from DFN-BWIN's structural properties. Its high connectivity favors NEC and NAV, as the abundance of alternative paths allows the solver to quickly identify energy-efficient and allocation-efficient configurations. In contrast, NAN and ASL are more sensitive to node activation and path length. In these objectives, DFN-BWIN's dense link structure expands the solution space, requiring the solver to evaluate more possibilities, which results in longer solving times. Finally, the New York topology was the most computationally demanding, with several instances approaching the $21,600$-second time limit. This outcome reflects the significantly higher complexity of solving VNFP-RP in larger and structurally intricate networks, where the expanded search space imposes a substantial computational burden on the solver. Overall, the results show that the tractability of the VNFP-RP depends not only on network size but also on structural configuration. Larger topologies significantly expand the solution space, increasing computational effort, while higher link density can either facilitate or hinder the search process depending on the objective. Consequently, connectivity and structural complexity determine how quickly the solver converges when finding optimal placements and routes.

TABLE 6. Mean execution times (in seconds) with their corresponding confidence intervals (95%), for each evaluated topology (Abilene, Dfn-Bwin, Polska, and New York) and each objective function (NAN, NAV, NEC, and ASL).

	NAN	NAV	NEC	ASL
Abilene	2.56 ± 0.69	100.61 ± 72.75	53.85 ± 32.24	2.19 ± 0.56
Dfn-Bwin	3.89 ± 1.05	3.90 ± 1.06	11.05 ± 3.60	3.22 ± 0.81
Polska	2.88 ± 0.79	103.22 ± 77.40	77.62 ± 47.48	2.48 ± 0.62
New York	10.55 ± 2.72	4244.33 ± 2115.02	1708.46 ± 952.54	8.01 ± 1.99

Beyond the influence of topology, the results show that scalability is strongly driven by the computational complexity inherent to each objective function. Objectives such as NAN and ASL consistently yield lower solving times, as their optimization criteria restrict the search space more directly—either by minimizing node activations or by favoring shorter routes—allowing the solver to converge more quickly. In contrast, NEC and NAV introduce broader combinatorial dependencies, as they require simultaneously exploring energy-efficient configurations or minimizing VNF deployments across many possible routing alternatives. These objectives generate substantially larger feasible regions, leading to higher solving times regardless of the underlying topology. This pattern indicates that scalability in the VNFP-RP is shaped by the combined influence of both objective-driven search space complexity and topological structure. While each objective imposes distinct computational challenges, the underlying connectivity and size of the network further modulate how easily feasible configurations can be found. Therefore, aligning the chosen optimization objective with the structural characteristics of the target network—according to the service provider's performance goals—can lead to more efficient resolutions and better overall performance. Table 6 summarizes these findings by presenting the mean execution times (in seconds) and their corresponding confidence intervals (95%) for each evaluated topology. This aggregated view reinforces the distinct computational profiles observed and highlights how objective complexity and topological structure jointly influence solver performance.

F. DISCUSSION HIGHLIGHTS

The consolidated analysis highlights the influences of the objective function and network topology on both performance metrics and computational tractability in the VNFP-RP. Problem complexity proved to be highly sensitive to the chosen objective: functions imposing multiple simultaneous constraints—such as NEN and NAV—consistently resulted in the longest solving times, whereas simpler objectives, like ASL and NAN, converged significantly faster. Topology played a decisive role in the tractability of the VNFP-RP. Smaller networks were generally easier to solve, as the search space remained limited. At the same time, higher connectivity density facilitated the discovery of feasible placements by providing more routing alternatives. However, dense connectivity in large-scale networks, such as New York, significantly increased computational requirements and resource consumption compared to smaller dense topologies, such as DFN-BWIN. These results demonstrates that the complexity of solving the VNFP-RP is not determined by size or density alone, but by the interplay between both factors, with direct implications for the design of efficient placement and routing strategies.

The objective functions focused on reducing overall resource usage — such as NEN and NAN— demonstrated excellent performance in allocation compaction, maintaining the minimum number of allocated VNFs and significantly reducing physical infrastructure utilization. However, these approaches differed notably in processing time and latency levels. The function NAN achieved average resolution times around 100 times shorter than the NEN, while yielding very similar energy outcomes. This difference can be attributed to the additional complexity introduced by the NEN function, which requires more precise decisions to balance nodes activation with utilization levels, thereby enlarging the search space and computational cost — particularly in larger or sparsely connected topologies. In contrast, the function ASL achieved the lowest average response times per demand, with stable behavior even under high load conditions. This performance, however, came at the cost of excessive VNF replication and a substantial increase in energy consumption. The results thus reveal a clear trade-off between service responsiveness and the efficient use of computational and energy resources.

The results reinforce that effectively solving the VNFP-RP requires a context-aware approach. The absence of a superior objective function underscores the need for adaptive or multi-objective mechanisms that allow dynamic reconfiguration of priorities based on operational requirements, infrastructure constraints, or performance targets. Strategies that integrate dimensions such as energy efficiency, compact plecement, and responsiveness provide greater robustness across diverse scenarios and topologies, while contributing to more sustainable, scalable, and practically deployable network systems.

V. CONCLUSION

In this work, we addressed the VNFP-RP, a problem that is fundamental to the efficiency and performance of NFV-enabled B5G and 6G networks. To this end, we implemented an ILP model and conducted a comparative evaluation of four objective functions under different performance metrics, including energy efficiency and processing time. The results reveal a clear trade-off among the analyzed objectives, each associated with distinct operational implications. The ASL objective delivered the best response times—up to four times faster than NEC—making it suitable for time-sensitive applications such as remote surgeries, cloud gaming, and live video streaming; however, this responsiveness came at the cost of substantial energy consumption and intensive VNF replication. Conversely, the NEC objective proved most effective in reducing energy consumption—achieving over 50% savings compared to latency-oriented objectives—but was also the most computationally demanding, frequently approaching the solving time limit. The NAN objective offered a balanced compromise, combining low energy usage with feasible solving times, and thus emerges as an attractive alternative for scenarios that require efficiency without imposing severe computational overhead. These insights emphasize that the choice of objective function should be guided by the priorities of each environment: networks oriented toward sustainability and infrastructure consolidation—such as data centers and cloud providers—may benefit from resource-centric objectives, whereas latency-critical IoT or real-time communication applications may favor latency minimization despite the associated increase in energy consumption. To foster reproducibility, fair benchmarking, and further research, we make our ILP-based testbed publicly available on GitHub [47]. Ultimately, this study underscores the importance of adopting flexible and potentially multi-objective approaches for the VNFP-RP, capable of dynamically adapting allocation strategies to operational demands, infrastructure constraints, and performance goals.

REFERENCES

[1] M. Chiosi, D. Clarke, P. Willis, A. Reid, J. Feger, M. Bugenhagen, W. Khan, M. Fargano, C. Cui, and H. Deng, "Network functions virtualisation: An introduction, benefits, enablers, challenges and call for action," in *Proc. SDN OpenFlow World Congr.*, vol. 48, 2012, pp. 1–16.

[2] European Telecommunications Standards Institute (ETSI). (2021). *Network Functions Virtualisation (NFV) Release 4*. European Telecommunications Standards Institute. Accessed: Nov. 1, 2023. [Online]. Available: https://www.etsi.org/technologies/nfv

[3] Z. Zhang and C. Wang, "Service function chain migration: A survey," *Computers*, vol. 14, no. 6, p. 203, May 2025.

[4] A. Varasteh, M. De Andrade, C. M. Machuca, L. Wosinska, and W. Kellerer, "Power-aware virtual network function placement and routing using an abstraction technique," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[5] R. Lin, L. He, S. Luo, and M. Zukerman, "Energy-aware service function chaining embedding in NFV networks," *IEEE Trans. Services Comput.*, vol. 16, no. 2, pp. 1158–1171, Mar. 2023.

[6] W. Wu, Y. Li, L. Chen, B. Zhang, W. Wang, Y. Zhao, and J. Zhang, "Service function chain mapping based on joint load balancing in computing power network," in *Proc. Opto-Electron. Commun. Conf. (OECC)*, Jul. 2023, pp. 1–4.

[7] W. Khemili, J. E. Hajlaoui, and M. N. Omri, "Optimizing resource and power consumption in a cloud environment via consolidation and placement investigation: A survey," *Eng. Appl. Artif. Intell.*, vol. 141, Feb. 2025, Art. no. 109818.

[8] B. Brik, H. Chergui, L. Zanzi, F. Devoti, A. Ksentini, M. S. Siddiqui, X. Costa-Pérez, and C. Verikoukis, "Explainable AI in 6G O-RAN: A tutorial and survey on architecture, use cases, challenges, and future research," *IEEE Commun. Surveys Tut.*, vol. 27, no. 5, pp. 2826–2859, Oct. 2025.

[9] V. H. L. Lopes, G. M. Almeida, A. Klautau, and K. V. Cardoso, "O-RAN-oriented approach for dynamic VNF placement focused on interference mitigation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2024, pp. 5479–5484.

[10] K. Gray and T. D. Nadeau, *Network Function Virtualization*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[11] M. Y. Saidi, I. A. Ikhelef, S. Li, and K. Chen, "Constrained routing in multi-partite graph to solve VNF placement and chaining problem," *J. Netw. Comput. Appl.*, vol. 230, Oct. 2024, Art. no. 103931.

[12] A. Varasteh, B. Madiwalar, A. Van Bemten, W. Kellerer, and C. Mas-Machuca, "Holu: Power-aware and delay-constrained VNF placement and chaining," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1524–1539, Jun. 2021.

[13] C. R. de Mendoza, B. Bakhshi, E. Zeydan, and J. Mangues-Bafalluy, "Near optimal VNF placement in edge-enabled 6G networks," in *Proc. 25th Conf. Innov. Clouds, Internet Netw. (ICIN)*, Mar. 2022, pp. 136–140.

[14] J. Munshi, S. Sultana, M. J. Hassan, P. Roy, M. A. Razzaque, A. Alelaiwi, M. Z. Uddin, and M. M. Hassan, "Attention model-driven MADDPG algorithm for delay and cost-aware placement of service function chains in 5G," *Ad Hoc Netw.*, vol. 173, Jun. 2025, Art. no. 103806.

[15] M. S. Akbar, Z. Hussain, M. Ikram, Q. Z. Sheng, and S. C. Mukhopadhyay, "On challenges of sixth-generation (6G) wireless networks: A comprehensive survey of requirements, applications, and security issues," *J. Netw. Comput. Appl.*, vol. 233, Jan. 2025, Art. no. 104040.

[16] M. N. A. Siddiky, M. E. Rahman, M. S. Uzzal, and H. M. D. Kabir, "A comprehensive exploration of 6G wireless communication technologies," *Computers*, vol. 14, no. 1, p. 15, Jan. 2025.

[17] M. Scatá, A. La Corte, A. Marotta, F. Graziosi, and D. Cassioli, "A complex network and evolutionary game theory framework for 6G function placement," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2926–2941, 2024.

[18] G. Meihui, "Models and methods for network function virtualization (NFV) architectures," Ph.D. dissertation, Université de Lorraine, Nancy, France, 2019.

[19] B. K. Umrao and D. K. Yadav, "Placement of virtual network functions for network services," *Int. J. Netw. Manage.*, vol. 33, no. 6, p. 2232, Nov. 2023.

[20] I. E. Said, L. Sayad, and D. Aissani, "Placement optimization of virtual network functions in a cloud computing environment," *J. Netw. Syst. Manage.*, vol. 32, no. 2, p. 39, Apr. 2024.

[21] A. P. Tchinda, B. Shala, A. Lehmann, B. Ghita, D. Walker, and U. Trick, "Energy-efficient placement of virtual network functions in a wireless mesh network," *IEEE Access*, vol. 12, pp. 64807–64822, 2024.

[22] J. Liang, S. Huang, Y. Qiu, L. Liu, F. Aziz, and M. Chen, "Sustainable virtual network function placement and traffic routing for green mobile edge networks," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 4, pp. 1450–1465, Dec. 2024.

[23] A. Mouaci, É. Gourdin, I. Ljubić, and N. Perrot, "Two extended formulations for the virtual network function placement and routing problem," *Networks*, vol. 82, no. 1, pp. 32–51, Jul. 2023.

[24] B. Zhang, Q. Fan, X. Zhang, Z. Fu, S. Wang, J. Li, and Q. Xiong, "A survey of VNF forwarding graph embedding in B5G/6G networks," *Wireless Netw.*, vol. 30, no. 5, pp. 3735–3758, Jul. 2024.

[25] Y. Mao, X. Shang, Y. Liu, and Y. Yang, "Joint virtual network function placement and flow routing in edge-cloud continuum," *IEEE Trans. Comput.*, vol. 73, no. 3, pp. 872–886, Mar. 2024.

[26] B. Xia, C. Li, Z. Zhou, and J. Liu, "Research on deployment method of service function chain based on network function virtualization in distribution communication network," in *Proc. IEEE 6th Inf. Technol.,Netw.,Electron. Autom. Control Conf. (ITNEC)*, vol. 6, Feb. 2023, pp. 1410–1414.

[27] A. Laxmana, P. Nagaraja, and S. U. Nagara Vinayaka, "Probabilistic service function placement and dynamic service function chain formation," *IEEE Access*, vol. 12, pp. 121261–121268, 2024.

[28] B. Addis, D. Belabed, M. Bouet, and S. Secci, "Virtual network functions placement and routing optimization," in *Proc. IEEE 4th Int. Conf. Cloud Netw. (CloudNet)*, Oct. 2015, pp. 171–177.

[29] A. Mouaci, É. Gourdin, I. LjubiC, and N. Perrot, "Virtual network functions placement and routing problem: Path formulation," in *Proc. IFIP Netw. Conf. (Networking)*, Jun. 2020, pp. 55–63.

[30] H. Tran Huy, N. T. Tam, H. T. T. Binh, and L. T. Vinh, "Two-stage metaheuristic for reliable and balanced network function virtualization-enabled networks," *Soft Comput.*, vol. 28, nos. 13–14, pp. 8259–8277, Jul. 2024.

[31] J. Xu and H. Hu, "Research on deployment of service function chains under hybrid network functions," in *Proc. 5th Int. Seminar Artif. Intell., Netw. Inf. Technol. (AINIT)*, Mar. 2024, pp. 1412–1417.

[32] X. Zhang, Z. Xu, L. Fan, S. Yu, and Y. Qu, "Near-optimal energy-efficient algorithm for virtual network function placement," *IEEE Trans. Cloud Comput.*, vol. 10, no. 1, pp. 553–567, Jan. 2022.

[33] M. Mehrabi, A. K. Boroujeni, V. Latzko, and S. Köpsell, "Towards deploying secure and highly available O-RAN components," in *Proc. IEEE Future Netw. World Forum (FNWF)*, Oct. 2024, pp. 736–741.

[34] K. Ali and M. Jammal, "Proactive VNF scaling and placement in 5G O-RAN using ML," *IEEE Trans. Netw. Service Manage.*, vol. 21, no. 1, pp. 174–186, Feb. 2024.

[35] S. Yuan, Y. Sun, and M. Peng, "Joint network function placement and routing optimization in dynamic software-defined satellite-terrestrial integrated networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 5172–5186, May 2024.

[36] Z. Jia, M. Sheng, J. Li, D. Zhou, and Z. Han, "VNF-based service provision in software defined LEO satellite networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6139–6153, Sep. 2021.

[37] L. Yala, P. A. Frangoudis, and A. Ksentini, "Latency and availability driven VNF placement in a MEC-NFV environment," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[38] P. K. Thiruvasagam, A. Chakraborty, and C. S. R. Murthy, "Latency-aware and survivable mapping of VNFs in 5G network edge cloud," in *Proc. 17th Int. Conf. Design Reliable Commun. Netw. (DRCN)*, Apr. 2021, pp. 1–8.

[39] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Gener. Comput. Syst.*, vol. 28, no. 5, pp. 755–768, May 2012.

[40] Y. C. Lee and A. Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems," *J. Supercomput.*, vol. 60, no. 2, pp. 268–280, May 2012.

[41] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," *ACM SIGOPS Operating Syst. Rev.*, vol. 35, no. 5, pp. 103–116, Dec. 2001.

[42] Python Software Foundation. (2025). *Python Language Reference, Versão 3.12*. Accessed: Jul. 23, 2025. [Online]. Available: https://python.org/

[43] P. D. Team. (2024). *Pyomo—Python Optimization Modeling Objects*. Accessed: Jul. 23, 2025. [Online]. Available: https://www.pyomo.org/

[44] Canonical. (2025). *Ubuntu: Open Source Operating System*. Accessed: Jul. 23, 2025. [Online]. Available: https://ubuntu.com/

[45] IBM Corporation. (2024). *IBM ILOG CPLEX Optimization Studio*. [Online]. Available: https://www.ibm.com/products/ilog-cplex-optimization-studio

[46] N. Huin, A. Tomassilli, F. Giroire, and B. Jaumard, "Energy-efficient service function chain provisioning," *J. Opt. Commun. Netw.*, vol. 10, no. 3, pp. 114–124, Mar. 2018.

[47] R. F. Vieira. (2025). *Virtual Network Function Placement and Routing*. Accessed: Aug. 14, 2025. [Online]. Available: https://github.com/Fogarolli/Virtual-Network-Function-Placement-and-Routing

[48] S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "SNDlib 1.0—Survivable network design library," *Networks*, vol. 55, no. 3, pp. 276–286, May 2010. [Online]. Available: http://www3.interscience.wiley.com/journal/122653325/abstract

[49] V. Myasnikov, "Efficient light coupling and propagation in fiber optic systems," *Technobius Phys.*, vol. 2, no. 3, p. 0017, Sep. 2024.

[50] Standard Performance Evaluation Corporation. (2024). *Spec Benchmarks*. Accessed: Jul. 23, 2025. [Online]. Available: https://www.spec.org/benchmarks.html

RAFAEL FOGAROLLI VIEIRA received the bachelor's degree in computer engineering from Faculdade Estácio de Belém, the first specialization degree in artificial intelligence and machine learning from the Pontifical Catholic University of Minas Gerais (PUC Minas), the second specialization degree in artificial intelligence applied to industry from the SENAI School of Technology, and the master's degree in electrical engineering with an emphasis in applied computing from the Federal University of Pará (UFPA). He is currently an Assistant Professor I with the Production Engineering Program, State University of Amapá (UEAP), and a member of the Operational Research Laboratory (LPO), UFPA. His research interests include evolutionary computing, machine learning, telecommunications networks, and computer networks.

MATHEUS GABRIEL GOMES PANTOJA received the bachelor's degree in computer engineering from the Federal University of Pará (UFPA), in 2025, where he is currently pursuing the master's degree. He was a Scientific Initiation Scholar through the Institutional Program of Scientific Initiation Scholarships (PIBIC) with UFPA. He is also a member of the Operational Research Laboratory (LPO). His research interests include evolutionary computing, machine learning, telecommunications networks, and computer networks.

CARLOS NATALINO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Federal University of Pará, Brazil, in 2016. He is currently a Researcher with the Optical Networks Unit, Chalmers University of Technology. His research focuses on network automation and on the challenges and opportunities for application of machine learning in the network automation context. Over the past years, he has been researching how to leverage machine learning for optical network design and operation, in problems such as resource efficiency (e.g., spectrum) and physical layer security. He has been involved in several national and international projects funded by research bodies in EU, Sweden, and Brazil. He is a member of Optica.

DIEGO LISBOA CARDOSO received the bachelor's degree in computer science from the University of Amazônia, in 2002, and the master's and Ph.D. degrees in electrical engineering from the Federal University of Pará (UFPA), in 2005 and 2010, respectively. He completed a postdoctoral fellowship with the Royal Institute of Technology of Sweden (KTH). He was the Director of Technology with the State Department of Education of the Government of the State of Pará (2008–2009). He was the Rector of Undergraduate Education with the Federal University of South and South-East Pará (UNIFESSPA), in 2014. He is currently an Associate Professor with the School of Computer Engineering and Telecommunications, UFPA, and the Graduate Program in Electrical Engineering (PPGEE). He has experience in computer science and computer engineering, with an emphasis on performance evaluation. His research interests include digital TV, access technologies, Markovian performance and simulation models, applied computational intelligence, and optimization techniques.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614