Received 6 August 2024; accepted 27 August 2024. Date of publication 3 September 2024; date of current version 12 September 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3453972

Leveraging Load Balance Metrics to Unravel the Impact of Multi-Access Edge Computing Locations on Online Dynamic Network Performance

MARYAM MASOUMI[1], IGNACIO DE MIGUEL[1] (Senior Member, IEEE), FABRIZIO GABRIO BRASCA[2], NOEMÍ MERAYO[1], RUBÉN M. LORENZO[1], AND RAMÓN J. DURÁN BARROSO[1]

[1]Department of Signal and Communications Theory and Telematics Engineering, Universidad de Valladolid, 47002 Valladolid, Spain

[2]Wind Tre SpA, 20121 Milan, Italy

CORRESPONDING AUTHORS: M. MASOUMI AND R. J. DURÁN BARROSO (e-mail: maryam.masoumi@uva.es; rduran@tel.uva.es)

This work was supported in part by the EU H2020 Research and Innovation Programme (MSCA, IoTalentum, and Smart5Grid) under Grant 953442 and Grant 101016912; in part by the Consejería de Educación de la Junta de Castilla y León and the European Regional Development Fund under Grant VA231P20; and in part by the Spanish Ministry of Science of Innovation and the State Research Agency funded by MCIN/AEI/10.13039/501100011033 under Grant PID2020-112675RB-C42.

ABSTRACT Telecommunication operators are increasingly relying on Network Function Virtualization (NFV) and Multi-access Edge Computing (MEC) to support emerging 5G/6G services, which demand ultra-low latency and ultra-reliability. Employing NFV and MEC enable operators to deliver services through Service Function Chains (SFC) composed of Virtual Network Functions (VNFs) utilizing computing resources close to the end user. A critical challenge in this architecture is the efficient allocation of these resources and the strategic placement of MEC sites to host VNFs. This paper introduces, for the first time, a novel approach to efficiently determine where to locate MEC sites with the aim of optimizing dynamic performance. Instead of conducting time-consuming simulations to evaluate and compare each and every potential selection of MEC sites, we demonstrate that by quickly precomputing load balance metrics, such as the Jain fairness index (JFI), promising sets of sites can be identified. Our research shows that there is a statistically significant negative monotonic relationship between the precomputed JFI and the blocking probability when, during network operation, SFCs are dynamically established and released. Thus, by leveraging this fast identification method, network operators can focus their efforts, such as conducting detailed dynamic simulations (necessarily long and time-consuming since networks should operate with low or very low blocking ratios), solely on the most promising combinations. Therefore, this approach streamlines the process of determining the strategic location of MEC sites in a network, reducing the time required to plan and optimize the network configuration effectively.

INDEX TERMS Blocking ratio, load balance, MEC placement, network planning, protection, service function chains.

I. INTRODUCTION

IN ORDER to support emerging 5G/6G services like augmented reality, Massive Internet of Things (MIoT), and cloud gaming, network operators are under pressure to ensure unprecedented availability standards. Moreover, recently the networks have encountered a multitude of challenges as a result of the increasing demand for traffic from emerging Internet applications and the various Quality of Service (QoS) requirements [1]. To address these challenges and provide dynamic QoS to applications, technologies like Software-Defined Networking (SDN), Multi-access Edge Computing (MEC), and Network Function Virtualization (NFV) are starting to be implemented in operators' networks since they are capable of rendering networks flexible, programmable, and vendor-agnostic, as well as cost-effective [2]. The emerging paradigm of edge computing, pivotal for processing data at the network's periphery, plays a crucial role in enhancing response times,

saving bandwidth, and enhancing privacy for Internet of Things (IoT) applications. This approach is indispensable as it directly addresses the critical demands of modern network infrastructures, particularly in supporting the dynamic service requirements of IoT environments [3].

SDN is a networking framework that overcomes the limitations of conventional network infrastructures by separating the control plane and data plane from switches and routers. SDN allows network management through centralized software controllers, making it more efficient, quick, and adaptable [4]. The second aforementioned technology, MEC, is expected to have a significant effect on 5G/6G networks by satisfying the ultra-low latency requirements of specific applications and services, as well as decreasing the load on transport networks [5]. By adding Information Technology resources to edge nodes, this technology enables them to perform cloud computing tasks, which in turn allows for data processing to be conducted closer to end-users. This reduction in distance decreases latency, which is a critical performance indicator for 5G/6G networks [6]. The third mentioned technology, NFV, uses the computational resources within network nodes or data centers, consisting of CPU cores, memory, and storage. Dedicated hardware-based network functions are replaced by Virtualized Network Functions (VNFs) that are installed on these computation elements. More specifically, when combined with NFV, MEC allows for the optimal placement of the VNFs at the network edge. Additionally, NFV allows network operators to decrease their expenses in terms of capital and operation, as well as the time required to provide new services.

As part of an NFV environment, the European Telecommunications Standards Institute (ETSI) introduced a MEC reference architecture. In this architecture, MEC servers are deployed and mobile edge applications are provisioned as VNFs [7].

In an NFV-based network, multiple VNFs are linked in a predetermined sequence to establish a specific Service Function Chain (SFC) and ultimately provide a specific service. To carry out these services, it is essential to conduct SFC placement (and thus VNF placement) in the underlying physical network, taking into account various resource and service requirements [8], [9]. Ensuring the reliability and availability of services in an NFV-based MEC deployment is crucial to meet user requirements, due to potential service outages resulting from the failure of a MEC site, server, or VNF. Each of these failures disrupts the continuity of the hosted SFC [10]. To effectively address such failures, the implementation of protection plans, such as dedicated or shared backup protection strategies, becomes crucial [11].

One of the primary challenges in the networks based on NFV is effectively managing both computational and communication resources, all while accounting for numerous SFCs that provide connectivity and the required services to end-users [10]. In many studies on MEC and VNF deployment, the focus has been on offline scenarios where all service requests are assumed to be known in advance. However, realistic 5G network environments are characterized by their dynamic and unpredictable nature, where service requests are not known *a priori*, necessitating a dynamic and agile approach to service provisioning [12]. The strategic placement of MEC nodes within the network infrastructure becomes critical in this context, significantly enhancing network performance and optimizing resource utilization. By appropriately situating MEC nodes, network operators can minimize latency, reduce network congestion, and enhance overall service quality. Moreover, MEC placement enables efficient utilization of resources, facilitating the seamless delivery of latency-sensitive applications such as augmented reality, autonomous vehicles, and real-time analytics. Therefore, in the realm of VNF placement, the importance of incorporating MEC placement as a previous critical consideration cannot be overstated, since it directly influences the overall performance, responsiveness, and scalability of the network infrastructure. In addition, when the network operates dynamically, requests are served on-demand, and the blocking ratio serves as a practical benchmark for evaluating networks that cater to users with varying QoS needs. The blocking ratio measures the proportion of requests for service deployment that are denied and cannot be accommodated due to lack of available resources, and subject to the specific requirements and constraints of the service (like latency). Those resources can be computational (CPU or RAM), or network bandwidth [13].

Traditional facility location problems [14], as explored over the decades, have focused on static scenarios, where the objective is to optimize location choices for facilities based on minimizing costs such as transportation, or maximizing accessibility for consumers or users. This kind of work, thoroughly documented in various studies, has advanced through the development of numerous deterministic and probabilistic models that cater to static or slowly evolving environments. In contrast, our research specifically addresses the dynamic and rapidly changing demands of MEC networks.

In this paper, we address a critical planning challenge faced by network operators. We consider a scenario in which a network operator aims to determine where to locate a set of M MEC sites within a network composed of N nodes, with the goal of optimizing dynamic performance. The MEC sites are equipped with servers to allocate their resources to host a series of VNFs that respond to serve instantiation requests dynamically and on-demand. In particular, the objective is to select the location of these M sites to minimize the blocking probability that will be experienced during operation, when SFCs are dynamically established and released, while also meeting bandwidth and latency requirements, and ensuring network survivability. This issue represents a significant challenge and, to the best of our knowledge, has not been considered in the literature so far. Since the arrival time and requirements of SFC requests are not known, their establishment (or blocking), depends on the availability of resources in the different MEC sites (CPU, RAM) and network links (bandwidth), when a request is received. The

heterogeneity of resources involved, their distribution across different nodes, and the impact of network topologies make it challenging to develop analytical models for assessing *dynamic* performance. Therefore, in order to compare the dynamic performance (blocking probability) of different candidate sets of locations for MEC placement, the network operator should run time-consuming simulations to evaluate the different possibilities. The number of possible combinations for MEC placement is given by the binomial coefficient $\binom{N}{M}$, which can be very large for real deployments, making exhaustive dynamic simulations (to compare all the different possibilities) infeasible due to their computational intensity. It should be noted that network operators require low or very low blocking ratios, which makes each simulation necessarily long to accurately assess those values with small confidence intervals, thus being highly time-consuming. Therefore, it is essential to reduce the number of possibilities to analyze in detail only the most promising ones.

To overcome this challenge, we propose a novel and efficient strategy for the planning phase of the network to identify those promising MEC locations. Our approach leverages precomputed load balance metrics, in particular, the standard deviation (STD) and the Jain fairness index (JFI) [15], to evaluate different MEC site combinations. These metrics are quick and easy to compute, and through rigorous statistical analyses, we show that there is a negative monotonic relationship between the JFI and the blocking probability observed during dynamic network operation. This critical insight allows network operators to predict, during the initial planning stages, which MEC site combinations are likely to yield superior performance when the network operates dynamically.

Thus, by precomputing these metrics, network operators can narrow down the most promising MEC placements, focusing detailed dynamic simulations exclusively on these sets. This significantly reduces the time and computational resources required for network planning and optimization, offering a practical and scalable solution for large networks.

In summary, the key contributions are:

- We show that the blocking probability for dynamic SFC establishment depends (amongst other factors) on the number and location of MEC sites, the desired level of protection and the traffic load.
- We demonstrate there is a negative monotonic relationship between the JFI and the blocking probability observed during dynamic network operation (when only the location of a certain number of MEC sites is varied).
- We suggest a strategy that leverages precomputed metrics to identify promising locations for MEC sites.

Fig. 1 summarizes the main contributions of the paper: the demonstration of the relationship between JFI and blocking probability, and how to leverage this information to only conduct detailed simulations for those locations with a JFI value higher than a threshold.

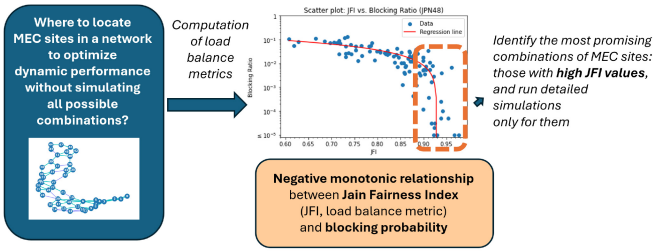


FIGURE 1. Summary of the main contributions.

The remainder of the paper is organized as follows. Section II reviews related work. Then, Section III presents the network and service chain model as well as different protection strategies and the SFC placement heuristic which is employed along the paper. Next, Section IV analyzes, through a simulation study, the impact of having different number of MEC sites and protection strategies on dynamic performance. Then, we present the main contribution of our work, which is the use of load balance metrics to find the most suitable sets of MEC locations. The metrics are defined in Section V, and simulation studies and statistical analyses are performed in Section VI. Finally, Section VII concludes the paper, and shows potential directions of future research.

This paper significantly extends two conference papers previously published by us. In [16], we presented the heuristic for SFC placement described Section III and used in the simulation studies conducted in this paper. The network and SFC models presented in that section also originate from that conference paper. However, they have been significantly enhanced by incorporating additional services with a much more realistic configuration (as chains of VNFs) and a more realistic traffic distribution. Furthermore, a more pragmatic configuration of the available resources at MEC sites and bandwidth assumptions are now considered. Finally, that conference paper presented a set of analysis that have not been included in this article, as the aim of the papers is different. In fact, the concept of "load balancing", which is a key element in this submission, was not even mentioned in that conference paper.

In the second conference paper, [17], we identified that the imbalance on the load handled by different MEC sites may translate in higher service blocking probabilities. However, this issue was only identified and not analyzed in detail (no load balance metrics were defined or used, and only one example comparing the blocking probability of three sets of MEC sites in a single network topology was provided).

This article builds upon that prior work but significantly extends those conference papers. Sections V, VI, and VII, which represent the core of this paper (proposing the use of specific load balance metrics and performing an extensive simulation and statistical analysis for different topologies), are completely new, presenting novel contributions.

II. RELATED WORK

A. THE MEC LOCATION PROBLEM

This section reviews research on the MEC location problem. Efforts to enhance network flexibility and reduce operational expenditures have led to significant research on the strategic deployment of MEC servers.

Various approaches in the literature have proposed MEC placement mechanisms, with the aim of reducing latency, through genetic algorithms, metaheuristics, and clustering techniques. For instance, Chantre and da Fonseca [10] utilized a non-dominated sorting genetic algorithm (NSGA-II) to optimize the placement of MECs to support 5G service latency and reliability requirements by determining the optimal number of instantiated slices and MECs. Similarly, Ghasemzadeh et al. [18] applied NSGA-II to improve edge server placement in MEC environments, focusing on reducing latency and balancing workloads across servers.

Thiruvasagam et al. [19] developed an algorithm to determine MEC locations by selecting base stations with high closeness centrality values, thereby minimizing delays for delay-sensitive services. Additionally, Chiha et al. [20] proposed three clustering models and evaluated them using a total cost of ownership model that considers both capital and operational expenditures. Their goal was to recommend cost-effective MEC placements that meet the latency needs of cooperative, connected, automated, and autonomous mobility services. Dash et al. [21] introduced a clustering-based Radio Access Network (RAN) association and MEC server placement model to reduce deployment costs while managing delays for delay-sensitive applications. Moreover, a clustering and nonlinear programming model was used to reduce latency and maximize profits in MEC server deployment [22].

Optimization strategies for MEC server placement also frequently focus on enhancing energy efficiency alongside other performance factors. Clustering models and hierarchical learning algorithms play a crucial role in these strategies. Hua et al. [23] leveraged integer linear programming to optimize MEC server selection by prioritizing energy efficiency, privacy protection, and load balancing. Additionally, Wu et al. [24] proposed a hierarchical learning algorithm based on a Stackelberg game, which focuses on selecting optimal MEC servers among unmanned aerial vehicle coalition heads, aiming to minimize both energy consumption and service delays. These strategies collectively enhance the energy efficiency and performance of MEC deployments, ensuring cost-effectiveness and high-quality service delivery.

On the other hand, it is worth nothing that the MEC placement problem is also similar to the Controller Placement Problem (CPP) in SDNs. This problem focuses on determining optimal (or at least near-optimal) locations for deploying controllers to effectively manage network traffic and ensure seamless operation. Therefore, insights can also be drawn from the existing literature on that topic. In particular, as described below, consideration is usually given to ensuring a balanced distribution of the traffic load among the controllers.

Thus, Lange et al. [25] focused on CPP considering crucial metrics such as latencies (both from nodes to controllers as well as among controllers), resilience to node and link failures, and load balancing. As some of these metrics are competing, the authors proposed a framework which offers operators a set of Pareto optimal placements with different trade-offs among performance metrics. Adekoya and Aneiba [26] also focused on those metrics but proposed an adapted version of an NSGA method to solve the problem, and Llerena and Gondim [27] leveraged an ant colony system metaheuristic, and validated its effectiveness in terms of controller response time, latency and load balancing, measured in terms of the JFI [15]. Clustering techniques have also been used to solve CPP efficiently [28], [29], sometimes coupled with metaheuristics as in [30]. Again, the metrics typically assessed encompass end-to-end delay, reliability, and load balancing.

A method for controller placement in software-defined IoT networks was proposed in [28]. This method segmented the network into clusters, using a multi-criteria decision-making scheme to evaluate placements based on hop count and propagation latency. The approach aimed to minimize end-to-end delay and communication overhead. However, it did not address energy efficiency and load balancing and faced challenges in real-time dynamic conditions. Ramaya and Manoharan [31] introduced a dynamic controller placement method for SDN using machine learning, leveraging traffic engineering and the K-Means++ algorithm to determine the number and location of controllers. This method improved network performance by minimizing the flow rule installation time in SDN switches and the average delay, but encountered computational complexity and relied heavily on accurate traffic prediction in dynamic environments.

Nevertheless, coming back to the MEC location problem, a few works have already considered the importance of load balancing when determining the location of MEC sites. Thus, Guo et al. [32] combined K-means clustering with mixed-integer quadratic programming to locate edge servers and allocate users, aiming to minimize service communication delay and workload imbalance. Similarly, Xu et al. [33] employed K-medoids clustering and NSGA-III to place edge servers in intelligent transportation systems, balancing latency and workload, which was measured in terms of the standard deviation. Liu et al. [34] developed a graph neural network framework for optimizing MEC server placement, focusing on balancing server workloads and minimizing session and service continuity management costs. Zhao et al. [35] combined graph partitioning and a multiple choice-upper confidence bound algorithm to optimize QoS by selecting and placing MEC servers, thereby improving transmission delay and overall service quality. Moreover, the results in [18] showed that the proposed method significantly reduces workload variance. Huang et al. [36] analyzed the deployment of edge servers and task assignment considering

distance and workload balance, focusing on minimizing the largest workload handled by any server in the network.

However, none of the previously mentioned references has evaluated the impact of load balancing on dynamic performance, specifically on blocking ratio. The reason is that in all these references the edge servers are used to process offloaded tasks (or to perform controller tasks in SDN). Therefore, load balancing in those scenarios is a way to minimize the computational latency experienced by the tasks. If a server becomes overloaded, tasks will encounter higher queueing delays before being processed, resulting in an overall increase in the total delay. In contrast, in this paper, we focus on MEC servers as providers of computing resources for the establishment of SFCs. Therefore, the impact of a MEC server encountering a higher load, i.e., handling more requests, mainly translates into a higher blocking probability due to lack of available resources, and this issue, that we identified in [17] (but did not study in detail there), has not been analyzed by any of the aforementioned references.

B. THE VNF AND SFC PLACEMENT PROBLEMS

The coupling of MEC placement and VNF placement within the context of SFC establishment is crucially significant for the overall performance. The deployment of MEC servers influences the decisions on where to place services, as these two aspects are interconnected. Over the past few years, there has been a significant amount of research done on the issue of VNF placement. As far as we know, there has not been much investigation into dynamic VNF placement in a distributed edge NFV infrastructure, even though several research projects have suggested solutions for the static placing of VNFs (e.g., [37], [38], and [39]).

Several studies have focused on reducing latency in VNF placement through various algorithms and optimization techniques. Zhang et al. [40] developed an adaptive interference-aware algorithm for VNF allocation within customized network slices to minimize interference. This approach helps in reducing latency by efficiently managing interference in network slices. In another approach, Slamnik-Kriještorac et al. [41] concentrated on automated VNF placement and migration to enhance vehicular communications. By leveraging MEC, their method aimed to minimize end-to-end latency and optimize resource utilization. However, the researchers did not explore the complexities of managing multiple resources and load balancing within the network.

Some works have focused on load balancing and resource management in VNF and service placement. Sun et al. [42] focused on optimizing resource management in network-softwarized environments. It addressed the efficient placement of VNFs, dynamic CPU allocation, and adaptive flow routing to minimize bandwidth consumption and support variable traffic loads. Nogales et al. [43] employed a single NFV orchestrator to manage VNFs across multiple sites, ensuring efficient resource allocation and adherence to key performance indicators such as deployment time. This approach facilitated seamless coordination and optimal use of resources across the network. Furthermore, Mosahebfard et al. [44] introduced an analytical framework to model admission ratio in NFV-based 5G networks, concentrating on both computational and communication resources. This framework aimed to optimize resource utilization by accurately modeling and managing the network's capacity to admit SFCs and users, thereby enhancing overall network performance. In [45], a method was presented for placing SFCs in a MEC-NFV environment that focuses on optimizing resource usage. To address the issue of reducing end-to-end latency, the paper proposed a dynamic VNF placement approach. However, [44] and [45] failed to address the survivability and resilience aspects of network components.

Studies on VNF and service placement have also concentrated on ensuring resilience. Karimzadeh-Farshbafan et al. [46] examined the simultaneous placement of VNFs and backup VNFs, proposing a strategy for sharing computational resources and link bandwidth to ensure service reliability. This method enhances network resilience by ensuring that backup resources are efficiently utilized to maintain service continuity in case of failures. In [16], we proposed a heuristic for SFC establishment with the aim of minimizing latency and providing dedicated or shared VNF protection. That heuristic will be used in the studies in this paper, and further details will be provided in Section III-D.

C. POSITIONING OUR CONTRIBUTION

In reviewing the extensive body of work on MEC placement, it is clear that our approach stands apart from the existing literature. While the referenced studies provide valuable insights into optimization strategies for MEC deployment, focusing on metrics such as latency management, load balancing, and energy efficiency, our approach is fundamentally different. We are pioneering the analysis of MEC placement by focusing on its performance in dynamic scenarios for SFC establishment. Specifically, we analyze in detail, for the first time to the best of our knowledge, the impact of load balancing among MEC locations on the blocking probability of dynamically requested SFCs.

As previously mentioned, in [17], we identified that an imbalance in the load handled by different MEC sites might translate into higher service blocking probabilities. However, this issue was only identified and not analyzed in detail there. Thus, in this paper, particularly in Sections V and VI, we comprehensively analyze this issue, demonstrating the existence of a relationship between traffic load balance and blocking probability observed during dynamic network operation. This constitutes the main contribution of our paper.

III. NETWORK, SERVICE CHAIN AND PROTECTION MODELS

The objective of this paper is to analyze how the selection of locations for MEC sites influences dynamic performance,

with the goal of identifying the MEC locations that yield superior performance in terms of blocking ratio when SFCs are dynamically established and released.

Moreover, it aims to address the challenges associated with mapping dynamic service requests onto MEC sites in 5G/6G networks with optical network backhauls. The use of virtualization at MEC sites to host VNF chains enables the handling of service requests, but it is subject to constraints on the available computational and storage resources, as well as bandwidth. The specific resource requirements of different services can result in blocking, so efficient design and operation methods should be employed. Additionally, the network may face MEC failures and interruption of multiple VNF-dependent services. To prevent these failures, it is crucial to develop a solution that considers the protection of services while mapping dynamic service requests to MEC sites. The solution should take into account the available resources, the specific resource requirements of each service, and the impact of potential failures on the operation of multiple VNFs. In particular, the dynamic resource allocation problem aims to allocate computing resources in MEC environments in order to provide reliable and efficient services to users, minimizing the average delay experienced by nodes in communicating with their associated MEC site, while also maximizing the utilization of the MEC resources. Regarding resiliency and availability of the established services, the protection methods considered include dedicated VNF backup and shared VNF backup.

Nevertheless, the final objective is to solve the MEC location problem (planning) in order to optimize network performance, specifically to minimize the ratio of non-established SFC requests during dynamic operation. As will be demonstrated by a complete statistical analysis in Section VI, the JFI is a key metric for the planning problem to identify sets of MEC locations that lead to improved dynamic performance during network operation, and to discard unpromising ones.

A. NETWORK AND SERVICE CHAIN MODELS

Let $\mathcal{G}(\mathcal{N}, \mathcal{E})$ represent an undirected topology of the physical network, where $\mathcal{N} = \{0, 1, 2, \ldots, N-1\}$ represents the set of nodes, and $\mathcal{E}$ is the set of physical links connecting these nodes in the network. On the other hand, let $\mathcal{M}$ represent the set of nodes (or sites) enabled with MEC resources ($\mathcal{M} \subset \mathcal{N}$), each having K servers offering computing capacity and thus equipped with a certain number of CPU cores and RAM.

A request requires the establishment of an SFC composed of a set of VNFs to be deployed on MEC servers. It is assumed that the entire SFC must be placed on the servers of a single MEC site with the aim of reducing latency and facilitating smooth recovery from failures. Different service types and thus different SFCs are considered. A service type is characterized by the set of VNFs composing the SFC, its bandwidth requirement, and the maximum latency. Moreover, each VNF type requires a specific amount of computing resources (CPU and RAM) to process incoming traffic. (Later, in Section IV, Tables 1 and 2 provide a specific example, which is used in the simulation studies of this paper.)

TABLE 1. Services types (based on [48]).

Service	Bandwidth per user (Mbps)	Delay (ms)	VNFs composing the service chain	Fraction of total traffic	Prob. of selection
Cloud Gaming	4	80	NAT, FW, VOC, WO, IDPS	25%	0.4048
Augmented Reality	100	10	NAT, FW, TM, VOC, IDPS	3%	0.0243
VoIP	0.064	100	NAT, FW, TM, FW, NAT	1%	0.0081
Video Streaming	4	100	NAT, FW, TM, VOC, IDPS	51%	0.4128
MIoT	1	5	NAT, FW, IDPS	17%	0.1376
Industry 4.0	70	8	NAT, FW	3%	0.0124

TABLE 2. CPU core, RAM usage and throughput for various VNFs [49].

VNF	CPU cores	RAM (GB)	Throughput
NAT	2	4	2 Gb/s
FW	2	4	2 Gb/s
IDPS	1	2	1 Gb/s
TM	1	2	1 Gb/s
VOC	2	4	2 Gb/s
WO	1	2	0.5 Gb/s

Given a request r of service type S originating at node n, the SFC establishment problem consist in determining a MEC site $m \in \mathcal{M}$ such that, at the time the request is received:

- The available CPU and RAM at site m meet or exceed the requirements of request r, which depend on the service type S (and thus on the set of VNFs composing that SFC).
- All links of the path between n and m have sufficient available bandwith to support the SFC, taking into account the number of users it can support.
- The end-to-end delay complies with the latency requirement of service type S.
- If there are several sites meeting the above requirements, the site with the lowest end-to-end delay is selected.

If the search is successful, the final step is to allocate the required resources at the MEC site and along the links of the path, and later released when the SFC is no longer needed.

Additionally, if protection against failures is to be provided, a second MEC site complying with an equivalent set of requirements must also be identified, so that a primary and a backup SFC are allocated.

In the following subsections, all these issues are described in more detail.

B. PROTECTION STRATEGIES FOR DYNAMIC SERVICE CHAINING

To ensure reliability, backup resources are allocated to protect each primary SFC. We consider three different scenarios: unprotected operation, dedicated SFC protection, and shared VNF protection.

In dedicated protection, a backup SFC is established to protect one and only one primary SFC. Shared protection,

on the other hand, is more resource-efficient, as it allows a shared backup VNF to protect multiple primary VNFs if they are located in different MEC sites. However, we assume that the entire backup SFC must be located in the servers of a single MEC site. When a request is received, resources must be reserved for both the primary and backup SFC. If resources cannot be reserved, the request is blocked. This backup approach is not only capable of handling the failure of a VNF or a single server but also provides protection in the event of failure of the entire MEC site due to different reasons such as power outages or any natural or man-made disaster.

C. REQUIREMENTS AND CONSTRAINTS FOR SFC ESTABLISHMENT

When provisioning SFC requests, there are several types of requirements and limitations that must be taken into account. These requirements are grouped into four categories:

- *Bandwidth requirement:* The bandwidth requirement states that the bandwidth (or, more precisely, the data rate) available on the network links of the paths utilized by the primary and backup SFCs must be sufficient to allow the transportation of the associated traffic.
- *Latency constraint:* This category involves the maximum end-to-end delay that can be allowed for the primary and backup SFCs, which depends on the service type. The primary end-to-end delay and the backup end-to-end delay are denoted by $d_{e2e-\text{primary}}$ and $d_{e2e-\text{backup}}$, respectively. The end-to-end delay comprises the communication delay between the source node (n) and the primary MEC site (m_p) or the backup MEC site (m_b), respectively, and the processing delay of the SFC (d_{SFC}). The first term depends upon the distance between the source node and the primary or backup MEC site (l_{n,m_p} or l_{n,m_b}), respectively, and on the propagation speed through the physical medium or group velocity v_g (e.g., 2×10^8 m/s for optical fiber). Finally, the processing delay depends on the number of VNFs composing the SFC (V) and the processing time required by each VNF composing the chain (d_VNF).

$$d_{e2e-\text{primary}} = 2d_{n,m_\text{p}} + d_{SFC} = 2\frac{l_{n,m_\text{p}}}{v_g} + Vd_\text{VNF} \quad (1)$$

$$d_{e2e-\text{backup}} = 2d_{n,m_\text{b}} + d_{SFC} = 2\frac{l_{n,m_\text{b}}}{v_g} + Vd_\text{VNF} \quad (2)$$

- *Resource requirement:* The resource requirement declares that the number of resources allocated to an SFC cannot exceed the available resources at the MEC site. If the selected site lacks adequate computational resources such as CPU cores or RAM to accommodate one or more VNFs, the request will not be allowed to proceed.
- *SFC establishment constraint:* In this research, we also consider this constraint, which asserts that each SFC must be installed on the servers of a single MEC site but without any possibility of being distributed across different MEC sites. This constraint is enforced to reduce latency and facilitate the seamless recovery of the SFC in case of failure.

If any of these constraints cannot be satisfied, the creation of an SFC is not accepted and the request will be blocked.

D. HEURISTIC FOR SFC ESTABLISHMENT

As mentioned earlier, a subset of M out of N network nodes will be chosen to host MEC resources. Along this paper, we use the same heuristic as in [16] to establish the requested SFCs. Next, we describe this heuristic for each of the three scenarios previously mentioned: unprotected, dedicated protection, and shared protection.

Independently of the protection scenario, first, for each network node n, the shortest-distance path to each MEC location is pre-computed, as these will be the paths used for SFC establishment.

In the first scenario (unprotected approach), each request is served by establishing a single (primary) SFC, as there is no backup. The establishment process involves identifying the nearest MEC site that possesses enough CPU and RAM capacity in its servers to accommodate the SFC, and ensuring there is enough available bandwidth in all the links of the pre-computed path between the source node of the request and the MEC site. If there are not enough resources, the search expands to the next nearest MEC site. If the search is successful, resources in the MEC site and traversed links are allocated. However, the request will be blocked if no resources are available in any site or other constraints (bandwidth or latency) are not met.

In the second and third scenarios, where dedicated and shared protections are implemented, the selection of primary and backup MEC sites for an SFC request is based on proximity, with the closest (available) MEC node assigned as the primary and the second closest as the backup. If there are not enough resources in a MEC site or other constraints are not fulfilled, the algorithm initiates a search for the next nearest available MEC site. In the case of shared protection, a backup VNF can protect multiple primary VNFs located in different MEC sites. The algorithm conducts a search for an available backup VNF of the same type before resorting to the creation of a new instance.

Once a service chain lifetime ends, resources allocated to primary and backup SFCs are released, except in the case of shared protection, where resources for a backup VNF are retained until they are no longer required by any SFCs.

IV. IMPACT OF THE NUMBER OF MEC SITES AND PROTECTION STRATEGIES ON DYNAMIC PERFORMANCE

In this section, simulation experiments are conducted using a Python discrete-event simulator to evaluate the performance of the different protection approaches when considering different numbers of MEC sites, M.

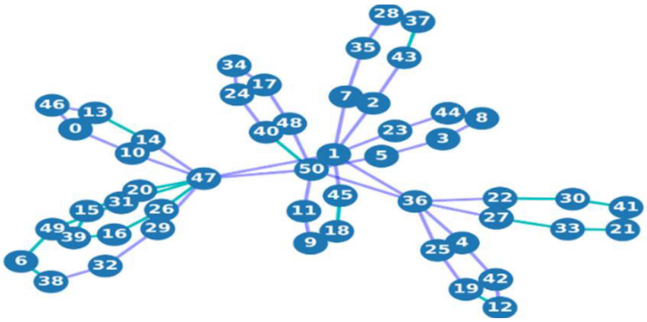


FIGURE 2. Metropolitan network topology in the northern area of Italy, from the 5G-Crosshaul project (I5GX) [47].

For the performance evaluation, a metropolitan network topology in the northern area of Italy, previously defined in the H2020 5G PPP 5G-Crosshaul project [47], is considered. This topology (I5GX, from now on) consists of 51 nodes and 61 fiber links (Fig. 2). The network topology exhibits a maximum distance of 3063.3 km between the two furthest nodes (network diameter).

The distances of the different links can be found in [47]. Among all the network nodes, a subset of M nodes are designated as MEC nodes. Initially, these M nodes are selected so that when each node ($n \in \mathcal{N}$) communicates with the closest node equipped with MEC resources ($m \in \mathcal{M}$), the average communication delay is minimized.

Six service types have been considered (cloud gaming, augmented reality, VoIP, video streaming, MIoT, and Industry 4.0), as shown in Table 1. These service types have distinct bandwidth per user and latency requirements, and their associated SFCs are made by different sequences of VNFs [48], which must be all placed within the operator's network and thus in one of the MEC sites. The SFCs themselves consist of specific VNFs, including Network Address Translation (NAT), Firewall (FW), Intrusion Detection and Prevention System (IDPS), Traffic Monitor (TM), Video Optimization Controller (VOC), and WAN Optimization (WO). The CPU core and RAM requirements as well as the maximum throughput for each VNF are detailed in Table 2.

The fraction of total traffic associated to each type of service (Table 1) has been estimated from Fig. 4 of [48]. Since different VNFs have different throughputs, and different services also have different bandwidths per user, the probability of selecting each service (also shown in Table 1) has been set accordingly to be consistent with the fraction of total traffic in the network.

The arrival of SFC requests follows a dynamic pattern, where the arrival instances are randomly generated based on a Poisson process with arrival rate λ. The lifetime of each SFC, held for a random duration, is determined using an exponential distribution with a mean of T. Furthermore, the source node for each request is randomly determined using a uniform distribution. Each request is associated to one of the six services in Table 1, with the probability shown in that table. The traffic intensity or offered load on the network (in Erlangs) is thus defined as λT. The load so defined represents the average number of service chains that would be established in the network if there were no blocking.

TABLE 3. Location of MEC sites in different cases.

Number of MEC Locations	MEC Locations	Servers per MEC	CPU cores per Server	RAM (GB) per Server
3 MECs	1, 36, 47	35	64	256
5 MECs	1, 7, 10, 36, 47	21	64	256
7 MECs	1, 7, 10, 21, 23, 36, 47	15	64	256

The objective of this section entails evaluating the performance of the heuristic for SFC placement with diverse protection scenarios and different number of MEC locations. Table 3 presents three different scenarios, each with a varying number of MEC sites (3, 5, and 7). In the cases where there are 3 or 7 MEC nodes, the locations of these nodes are chosen to minimize the average communication delay between the nodes and their closest MEC. In the scenario with 5 MEC nodes, we assume that the network undergoes an evolutionary upgrade in which two new nodes are added as an intermediate step toward the final configuration with 7 MEC nodes. Anyway, to ensure a fair comparison of the three scenarios, it should be mentioned that the same computing resources are employed in all situations. This implies that a total of 105 servers, each with 64 CPU cores and 256 GB of RAM, are allocated to each scenario. Consequently, in the 3-MEC scenario, there are 35 servers per MEC site, while the 5-MEC and 7-MEC scenarios consist of 21 and 15 servers per MEC site, respectively. All the computing resources at a MEC site are made accessible through a virtualized pool of resources.

Nodes are connected by optical fibers with a bandwidth (data rate) of 300 Gb/s, and the propagation speed (group velocity) through the fibers is 2×10^8 m/s. Additionally, the processing delay per VNF is $d_{\text{VNF}} = 0.05$ ms.

The blocking ratio metric is measured to compare the simulation performance with and without protection methods for different number of MEC sites. This metric measures the ratio of the total number of blocked SFCs due to lack of computing resources, bandwidth, or latency violations, to the total number of requests. It gives an overall picture of the system's ability to handle SFC requests and can be used to compare the effectiveness of different protection methods and the impact of different number of MEC locations. Simulations are conducted in this section considering unprotected operation, operation with dedicated backup, and operation with shared backup. For each load, a dynamic generation of 110,000 requests takes place, with the first 10,000 reserved for simulator warm-up and the remaining 100,000 used for performance analysis. The comparison of blocking ratios for different protection approaches and the use of 3, 5, and 7 MEC sites is shown in Fig. 3. Results are shown for different offered loads in steps of 50 Erlangs, along with their corresponding 95% confidence intervals. Note that some curves lack data for low blocking ratios (e.g.,

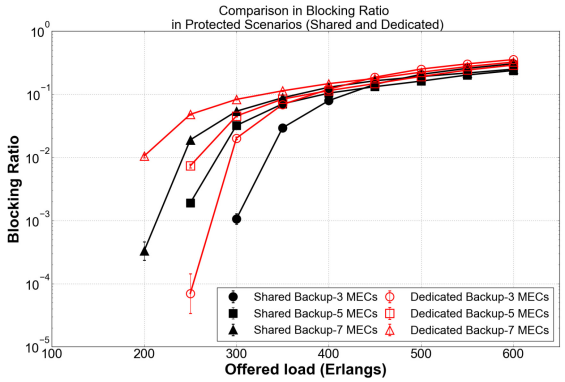


FIGURE 3. SFC blocking ratio in 3, 5 and 7-MEC sites scenarios for different protection approaches.

for loads $\leq$ 250 Erlangs for the shared backup strategy with 3 MECs). This is because no blocking events were observed in the simulations at those offered loads. Consequently, these points are not represented in the figure for the sake of clarity. This approach is also used in other figures of the paper.

Obviously, the unprotected approach exhibits lower blocking probabilities, with no blocking observed for the range of loads considered (and thus they are not represented in the figure). On the other hand, dedicated protection is the least favorable approach in terms of blocking as it necessitates resource reservation for entire backup SFCs. Shared protection shows better performance in terms of blocking ratio compared to the dedicated protection approach, since it allows for efficient resource utilization by sharing backup resources among multiple VNF instances, in contrast to dedicated protection.

Although, apparently, as shown in Fig. 3, concentrating computing resources in fewer MEC sites rather than distributing them in a higher number of MEC locations leads to lower blocking probabilities, such a conclusion cannot really be drawn from these results alone. Since all resources in a MEC site are made accessible through a virtualized pool of resources, concentrating all network resources in fewer locations enables a more efficient use of these resources, but this comes at an expense of congestion in accessing links to those sites (although no blocking events were observed due to lack of bandwidth). On the other hand, the load handled by each MEC site is not perfectly balanced, and, in fact, the amount of load imbalance varies between the different configurations of MEC sites analyzed in this section. As we will show later, this also has an impact on blocking probability, and, thus, accounts for the differences observed in the results.

Hence, the objective of the next section is to focus on the latter issue, and analyze how the location of MEC sites has an impact on load balancing (or imbalance), and ultimately on blocking ratio. In order to eliminate the influence associated with having different number of MEC sites, we will focus on scenarios with the same number of MEC sites, but located in different strategic locations, and analyze their performance in terms of blocking ratio. We consider a set of metrics that can be quickly computed and that we demonstrate to be helpful in determining the most appropriate set of MEC site locations (among several options), without the need to run a lengthy dynamic simulation.

V. LOAD BALANCE METRICS FOR MEC LOCATION PLANNING

Building upon the results from the previous section, the impact of different sets of MEC sites on the placement and hosting of SFCs and thus on the blocking ratio is examined in this part. It is crucial to analyze the reasons behind the diverse dynamic performances exhibited by different MEC locations and explore if there are *a priori* methods to determine the set of locations with the best performance (without having to rely exclusively on simulations). By investigating different location configurations (all with the same number of MEC sites), we aim to understand how the choice of MEC node locations affects the efficient deployment and execution of dynamic SFCs within the network infrastructure.

First of all, we consider a set of metrics that are easy and quick to compute yet potentially useful to compare and estimate which set of MEC site locations may lead to the best performance when the network is dynamically operated, without requiring long and time-consuming simulations. Nevertheless, in the next section, we will perform simulation experiments to determine the suitability of these metrics and to validate this assumption.

In order to compute the metrics, the load handled per server in each MEC node must be computed. This will give us a baseline understanding of the load patterns and help us identify any bottlenecks or areas of congestion. Once we have these data, we can calculate the following metrics:

- *Standard deviation (STD):* The standard deviation is a measure of how spread out the load per server is across different MEC sites. A high standard deviation indicates that there are significant differences in the SFC loads per server between MEC sites, while a low standard deviation means that the load is more evenly distributed. The method used to determine how the SFC load will be divided among the MEC sites is by calculating the standard deviation in the following manner:

$$STD = \sqrt{\frac{\sum_{m=1}^{M}(x_m - \mu)^2}{M}}, \tag{3}$$

  where x_m is the total load handled per server at MEC site m, and μ is to the average value of x_m.
- *Jain Fairness Index (JFI):* The Jain Fairness Index [15] is a positive number less than or equal to one, which quantifies the fairness of load distribution among MEC sites. A value of 1 signifies perfect fairness, indicating an equal allocation of SFC requests across all MEC

TABLE 4. Sets of MEC site locations considered in the I5GX topology. They represent all the different possibilities (with 5 sites) for evolution from 3 to 7 MEC Sites.

Location Set	MEC Locations	Servers per MEC	CPU cores per Server	RAM (GB) per Server
A	1, 7, 23, 36, 47	21	64	256
B	1, 7, 21, 36, 47	21	64	256
C	1, 21, 23, 36, 47	21	64	256
D	1, 10, 21, 36, 47	21	64	256
E	1, 10, 23, 36, 47	21	64	256
F	1, 7, 10, 36, 47	21	64	256

sites. Thus, a higher Jain Fairness Index is generally desirable. The calculation for the Jain Fairness Index is:

$$JFI = \frac{\left(\sum_{m=1}^{M} x_m\right)^2}{M \sum_{m=1}^{M} x_m^2}. \quad (4)$$

The JFI and STD metrics are indicators that determine the balance or imbalance of the SFCs to be allocated in different MEC sites. The insights gained from these metrics can be utilized to estimate dynamic performance, and thus make informed decisions regarding the selection of MEC sites without the necessity of conducting a dynamic simulation in all cases.

VI. IMPACT OF THE LOCATION OF MEC SITES ON DYNAMIC NETWORK PERFORMANCE AND RELATIONSHIP WITH LOAD BALANCE METRICS

In this section, we perform a set of simulation experiments and statistical analyses to validate the suitability of the metrics defined in the previous section to help determining which set of MEC sites provides the best dynamic performance without requiring to run a dynamic simulation.

A. ANALYSIS OF DIFFERENT SCENARIOS, WITH 5 MEC SITES, FOR THE I5GX TOPOLOGY

Firstly, we focus again on the I5GX topology (Fig. 2) and analyze different scenarios with 5 MEC sites. In particular, six distinct sets of MEC node locations are considered, as shown in Table 4. For the sake of clarity and readability, each set of locations is mapped to a letter from A to F, as shown in that table. They represent all the migration alternatives (having 5 MEC sites) from the 3-MEC towards the 7-MEC site scenario described in Table 3. Thus, MEC sites with indexes 1, 36, and 47 are present in all location sets (as they were the selected nodes in the 3-MEC scenario), and the other two sites provide all the intermediate migration alternatives towards the 7-MEC site scenario. We will compute the STD and JFI metrics for the different sets of MEC sites; then, we will perform dynamic simulations assuming a shared protection strategy, and finally, we will analyze the blocking ratio and its relationship with the load balance metrics.

We first compute the expected load to be handled per server in each MEC site. To compute that value, we consider that a node will successfully use the closest MEC site for the establishment of the primary SFC and the second closest MEC site for the establishment of the backup SFC. Obviously, those MEC sites should be at a relatively low distance from the source node to ensure that latency constraints can be met. Based on those assumptions, the load handled per server in each MEC site, x_m, is calculated. That information is then utilized to compute the STD and JFI metrics, which are represented in Fig. 4 together with the percentage of the total load (including primary and backup SFCs requests) that is handled by each MEC site. This helps in understanding how the load is distributed among the different MEC nodes.

As mentioned, we have considered six distinct sets of MEC nodes, and, as shown in Fig. 4, each set exhibits varying levels of load balance. The location sets A, B, and C display significant asymmetries and thus an imbalanced load distribution, whereas the remaining location sets demonstrate a more balanced distribution.

Moving towards scenarios E and F, the load distribution becomes increasingly homogeneous. This translates into a decrease in the STD and an increase in the JFI as we progress from scenario A to F. In particular, scenarios E and F result in the lowest STD (0.0078) and the highest JFI (0.88).

Based on the STD and JFI values, the expectation is that configurations E and F should provide the best dynamic performance. To validate that hypothesis, the blocking ratio has been subsequently evaluated by conducting simulations in a dynamic context, and the outcomes are shown in Fig. 5.

The findings indicate that as the level of load imbalance increases (high STD and low JFI), the dynamic blocking ratio is also likely to increase. Conversely, a higher JFI is generally associated with a lower blocking ratio. Therefore, by leveraging the computation of STD and JFI, we can estimate the relative dynamic performance of different sets of MEC site locations. Thus, the computation and comparison of these metrics help in determining the most suitable set of MEC sites, without the need for extensive dynamic simulations for MEC placement. The results show that the location sets E and F, having the lowest STD and highest JFI, exhibit the anticipated best behavior in terms of the blocking ratio. Nevertheless, having a higher JFI and lower STD does not always guarantee to have lower blocking probabilities as, for instance, location sets B and C have better JFI and STD metrics than location set A, but lead to a higher blocking probability. (Again, in Fig. 5 the E and F curves have no values at offered loads below 250 Erlangs because no blocking events were observed).

Although the findings suggest that a higher JFI is generally associated with a lower blocking ratio, only considering six different sets of MEC locations does not provide enough supporting evidence of this fact. For that reason, in Section VI-B we conduct an extensive analysis, considering a high number of different sets of MEC locations and different topologies, and then, in Section VI-C, also considering different traffic distributions.

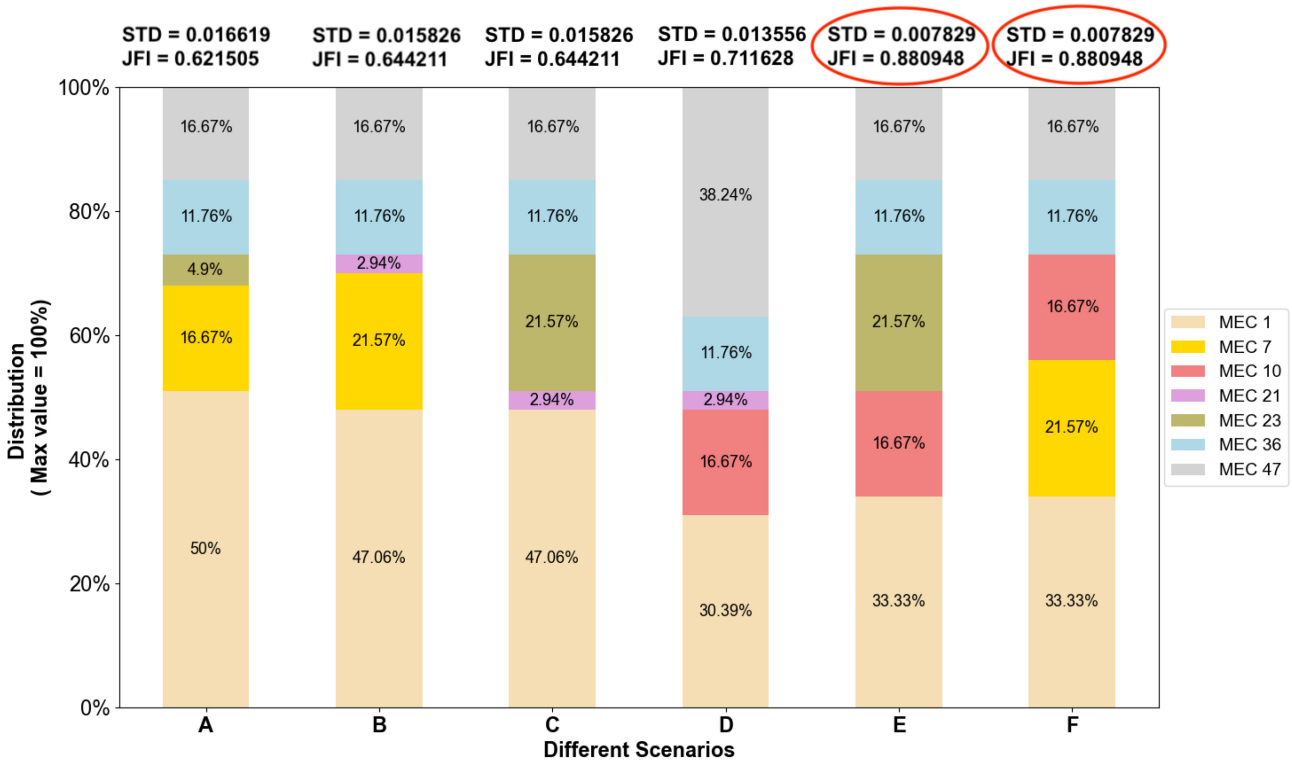


FIGURE 4. Percentage of the total load expected (i.e., pre-computed) to be handled by each MEC node. Each column represents a different set of MEC sites, as defined in Table 4.

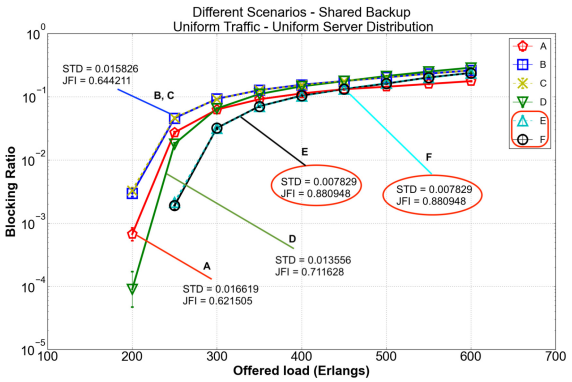


FIGURE 5. Dynamic performance for the different options of migration scenarios with 5 MEC sites. The SFC blocking ratio is shown as a function of the selected set of MEC sites (Table 4).

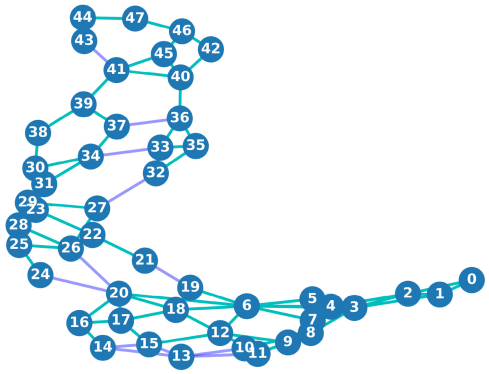


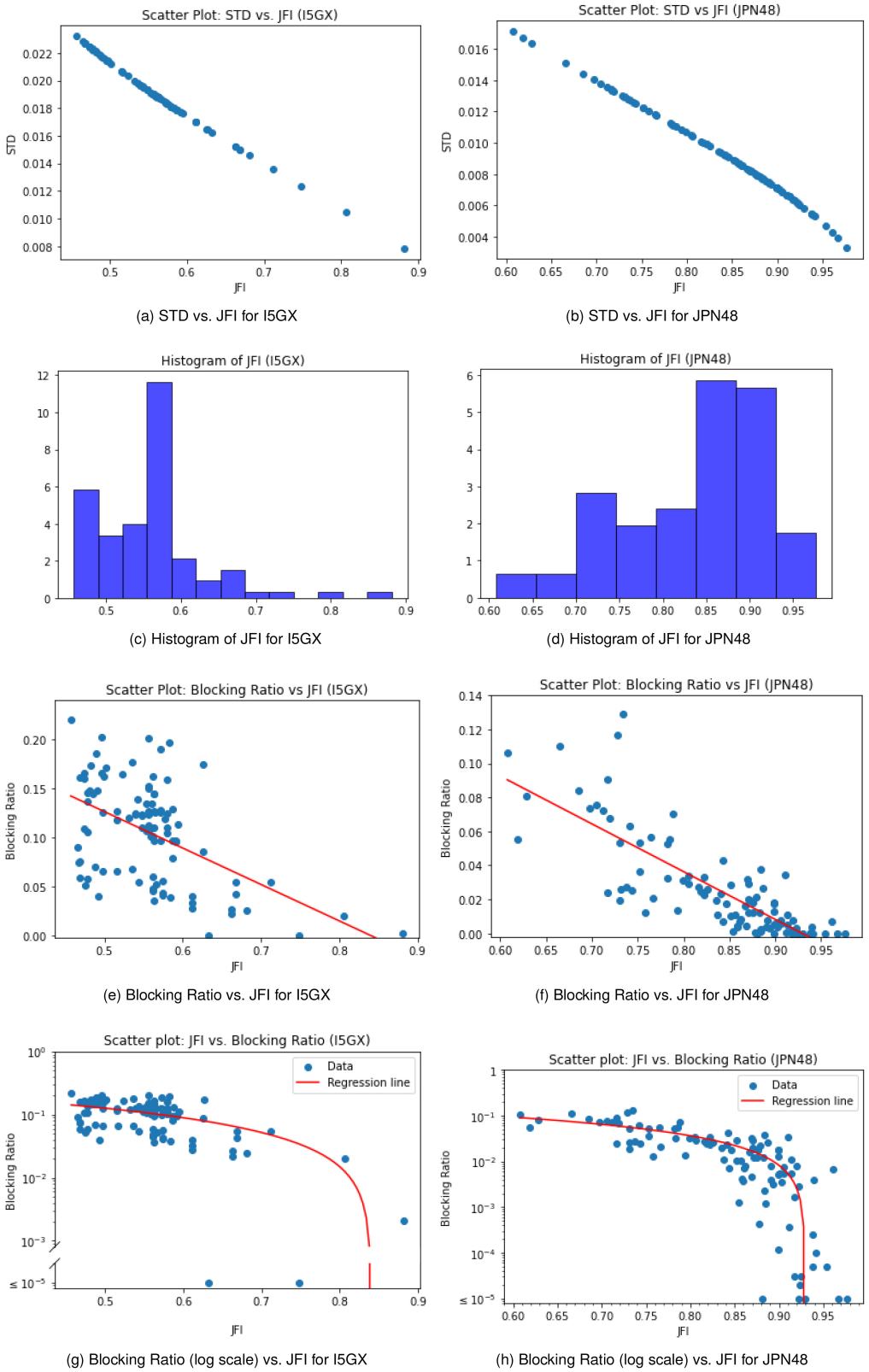
FIGURE 6. Japan Photonic Network (JPN48) [50].

B. ANALYSIS OF MULTIPLE SETS OF LOCATIONS ON DIFFERENT TOPOLOGIES

To further validate the relationship of the metrics with the dynamic performance in terms of the blocking ratio of the SFC requests, we provide a more extensive analysis of the location of MEC sites in the I5GX topology. Moreover, we also showcase the results obtained when considering an alternative topology. By presenting the outcomes for the Japan Photonic Network model (JPN48) [50] in Fig. 6, we provide evidence of the effectiveness and robustness of the metrics across varying scenarios. The network encompasses 48 nodes and 82 links. However, we consider a scaled version of this topology (dividing by 10 the length of each link) so that it has approximately the same dimensions as the I5GX network previously analyzed.

For each topology, 100 sets of 5-MEC locations have been randomly generated. For each set, the STD and JFI metrics have been computed, followed by a dynamic simulation to obtain the blocking ratio. For the I5GX topology, an offered load of 250 Erlangs has been considered, while for JPN48 the load has been set to 300 Erlangs, so that they lead to similar values of blocking probabilities.

Until this point, we have discussed two different metrics, namely STD and JFI. First, we investigate whether a statistical association exists between these two metrics. Fig. 7(a) and 7(b) depict scatter plots showing the relationship



(a) STD vs. JFI for I5GX

(b) STD vs. JFI for JPN48

(c) Histogram of JFI for I5GX

(d) Histogram of JFI for JPN48

(e) Blocking Ratio vs. JFI for I5GX

(f) Blocking Ratio vs. JFI for JPN48

(g) Blocking Ratio (log scale) vs. JFI for I5GX

(h) Blocking Ratio (log scale) vs. JFI for JPN48

FIGURE 7. **Analysis of the I5GX and JPN48 topologies.**

between the STD and the JFI metrics for the I5GX and JPN48 topologies, respectively. The visual inspection suggests a strong association between these metrics.

However, employing the usual Pearson correlation coefficient is not suitable in this case due to the non-normal distribution of the STD and JFI metrics. Fig. 7(c) and 7(d)

show the histogram for the JFI, which does not follow a normal distribution. Further analysis through Q-Q plots (not shown in the manuscript) and the Shapiro-Wilk test also support this conclusion.

Therefore, Kendall and Spearman correlation analyses emerge as more appropriate options for this case. Both Kendall's τ and Spearman's rank correlation coefficient yield values of -1.0, with p-values lower than 10^{-47}, indicating a statistically significant strong negative monotonic relationship between these metrics. Therefore, we can confidently rely on just one of these metrics. We choose to focus on JFI, given its consistent range (0 to 1), unlike STD values, which can significantly vary across different network scenarios.

The histograms presented in Fig. 7(c) and 7(d) show that, with a random selection of MEC sites, there tends to be a prevalence of low JFI values in the I5GX topology (85% of the sets exhibit JFI < 0.6), while high values are observed in the JPN48 topology (70% of the sets have JFI > 0.8). This difference is attributed to the distinct topological features of the networks. In the case of the I5GX, characterized by an interconnection of rings, a random selection of sites is more likely to result in unbalanced configurations. Conversely, the meshed structure of the JPN48 generally leads to more balanced scenarios in random selections. It is essential to note that, in practical network design, the location of MEC sites should not be arbitrary. As we will demonstrate next, the JFI metric serves as a valuable tool in discarding potentially unfavorable configurations and directing attention toward those with the potential for improved performance.

Next, Fig. 7(e) and 7(f) show the relationship between the blocking ratio and the JFI metric for the two topologies. These figures present scatter plots along with linear regression lines for each case. Fig. 7(g) and 7(h) display the same data and regression lines as before, but the y-axis is now represented in a logarithmic scale to better show configurations that approach a blocking ratio close to zero.

For the I5GX topology, Kendall's τ is -0.32 (p-value $< 3 \cdot 10^{-6}$), and the Spearman's rank correlation coefficient is -0.44 (p-value $< 6 \cdot 10^{-6}$). These values indicate a slight to moderate negative monotonic relationship between the JFI and the blocking probability, which is statistically significant. In the case of the JPN48 topology, Kendall's τ is -0.64 (p-value $< 6 \cdot 10^{-21}$), and the Spearman's rank correlation coefficient is -0.83 (p-value $< 3 \cdot 10^{-26}$). Therefore, there is clear evidence of a statistically significant moderate to strong negative monotonic relationship between the JFI and the blocking probability.

To sum up, these findings underscore that higher values of JFI generally correlate with improved dynamic performance.

C. ANALYSIS OF MULTIPLE SETS OF LOCATIONS WITH A DIFFERENT TRAFFIC LOAD DISTRIBUTION

In our previous analyses, we considered a uniform traffic distribution scenario where each node generated an equal share of the total traffic, that is, $1/N$. This means that the probability of an SFC request originating from node i in the network is $1/N$, $\forall i \in \mathcal{N}$. Although straightforward, this scenario does not accurately capture the diverse and uneven nature of real-world traffic patterns, which are influenced by varying demographic and geographic factors.

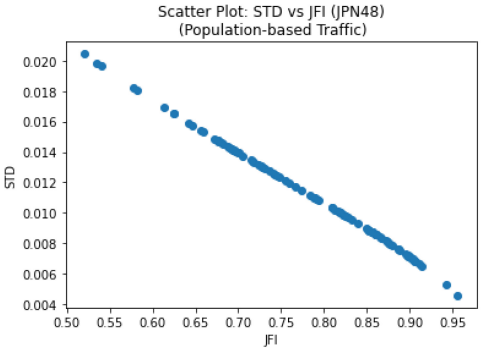
To enhance the realism and also further validate the relationship of the JFI with the blocking probability in a different scenario, we now consider a non-uniform traffic distribution model that takes into account the population density associated with each network node. This approach assumes that areas with higher populations generate more traffic. In this way, the probability of an SFC request coming from a node is equal to the portion of the total population of network users that is specifically associated with that node. Let P_i denote the population of the area associated with node i. The total population P_{total} is the sum of the populations of all those areas, $P_{\text{total}} = \sum_{i=0}^{N-1} P_i$. Thus, node i is selected as the source node for a SFC request with probability P_i/P_{total}.

Unfortunately, we have not been able to obtain population data for the I5GX topology. However, this data is available for the JPN48 topology [50], so we focus on the analysis of this topology. We have considered the same 100 sets of 5-MEC locations randomly generated in the previous subsection, but now considering the population-based traffic scenario. Therefore, the STD and JFI metrics have been computed under these new conditions, and we have also run dynamic simulations to determine the blocking ratio for each set of MEC locations, assuming an offered load of 300 Erlangs.

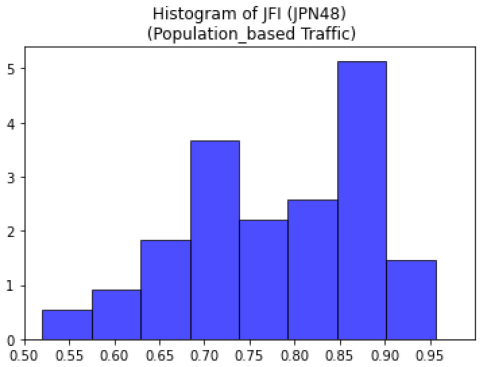
Again the STD and JFI metrics are highly correlated. Fig. 8(a) depicts the strong relationship between STD and JFI, with Kendall's τ and Spearman's rank correlation coefficients yielding values of -1.0, with p-values lower than 10^{-48}. Then, Fig. 8(b) shows the histogram for the JFI, which shows a similar but different distribution to that obtained for the uniform traffic scenario. When considering uniform traffic 70% of the sets had JFI > 0.8. However, with the population-based traffic 49% of the sets exceed that JFI value. So in this case, there are less combinations of MEC sites with highly-balanced load distributions.

The scatter plots in Fig. 8(c) and Fig. 8(d) depict the relationship between the blocking ratio and JFI. Fig. 8(c) shows the relationship on a linear scale, while the Fig. 8(d) presents the same data on a logarithmic scale. The plots show a clear negative correlation, as higher JFI values generally correspond to lower blocking ratios. This is evidenced by the linear regression line, which slopes downward, indicating that more balanced traffic distribution generally reduces blocking events.

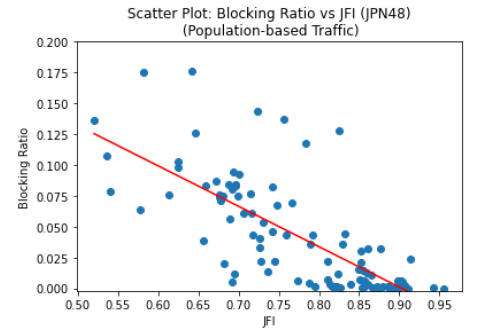
The statistical analysis further supports this observation. Kendall's τ is -0.60 (p-value $< 2 \cdot 10^{-18}$), and the Spearman's rank correlation coefficient is -0.80 (p-value $< 5 \cdot 10^{-23}$). These results reinforce the evidence that there is a statistically significant moderate to strong negative monotonic relationship between the JFI and the blocking probability.



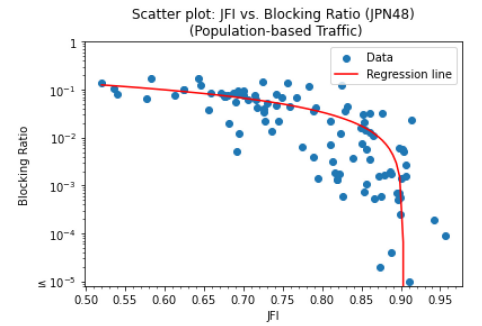
(a) STD vs. JFI for JPN48 (Population-based Traffic)



(b) Histogram of JFI for JPN48 (Population-based Traffic)



(c) Blocking Ratio vs. JFI for JPN48 (Population-based Traffic)



(d) Blocking Ratio vs. JFI for JPN48 (Population-based Traffic) (log scale)

FIGURE 8. **Analysis of the JPN48 topology in case of population-based traffic.**

In summary, these results show again that configurations with higher JFI values, which represent more evenly distributed traffic loads, tend to experience significantly lower blocking ratios. This relationship is critical for network performance optimization, as it highlights the importance of achieving fair traffic distribution to minimize blocking.

D. DISCUSSION ON RUNNING TIME AND THE APPLICATION OF THE JFI METRIC IN MEC PLACEMENT STRATEGIES

As mentioned in Section I, evaluating and comparing the dynamic performance, specifically the blocking probability, of different candidate locations for MEC placement, requires network operators to run extensive simulations.

The total required simulation time can be approximated by $t_{\text{req}} N_{\text{req}} LC$ (besides an additional initialization time), where:

- t_{req} is the average time required to process a request (checking for resource availability, reserving resources, and also including the time for releasing them when they are no longer needed),
- N_{req} is the length of the simulation, specifically the number of requests generated during the simulation,
- L is the number of different traffic loads to evaluate,
- C the number of candidate sets of MEC locations to be compared.

We have obtained the average value for t_{req} when running dynamic simulations for the IG5X and JPN48 topologies. Six and four random sets of 5-MEC sites were considered for the IG5X and JPN48 topologies, respectively, under nine different traffic loads (ranging from 200 to 600 Erlangs) and for 110,000 requests. The simulations were conducted using the previously mentioned Python-based simulator on a server equipped with an AMD EPYC Rome 7552 2.2 GHz CPU and 128 GB RAM, running Ubuntu 20.04 LTS. The average processing time per request was $t_{\text{req}} = 0.042 \pm 0.001$ seconds for the IG5X topology and 0.047 ± 0.001 seconds for the JPN48 topology.

Based on this data, assessing the performance of $C = 100$ different sets of 5-MEC sites with simulations of $N_{\text{req}} = 110,000$ requests and a single traffic load ($L = 1$) requires approximately 5 days and 8 hours for the IG5X topology, and around 5 days and 23 hours for the JPN48 topology. In contrast, initializing the network by precomputing the shortest-distance paths and computing the JFI and STD for those 100 sets takes only around 10.45 ± 0.04 seconds for the IG5X topology and 10.0 ± 0.3 seconds for the JPN48 topology.

It should be noted that if lower blocking probabilities need to be evaluated (which requires longer simulations) or a higher number of potential sets of MEC sites are analyzed, the simulation time will be even longer. For instance, the IG5X topology consists of 51 nodes, resulting in $\binom{51}{5}$ different sets of 5-MEC sites. In that case, more than 300 years would be required (without considering parallelism) to simulate all possibilities for a single traffic load with the same simulation length as before (110,000 requests).

Therefore, it is evident that a brute force approach, which evaluates all potential combinations of MEC sites through simulation, is not scalable. Although having a higher JFI

does not guarantee a lower blocking ratio when SFCs are dynamically established and released, we have demonstrated that there is a negative monotonic relationship between the JFI and the blocking probability. This implies that computing the JFI metric, which can be done quickly, allows network operators to identify promising site configurations and concentrate their efforts on conducting detailed simulations on these subsets, thereby optimizing network performance more efficiently. Consequently, only those combinations with higher JFI values should undergo detailed simulation, as we summarized in Fig. 1.

Moreover, this strategy can be combined with many proposals in the literature for MEC placement, by including an additional constraint on the JFI or incorporating it as part of the objective function to be optimized.

VII. CONCLUSION

In this paper, we have approached a critical challenge faced by network operators, which is to determine the efficient placement of MEC sites to enhance dynamic performance, specifically by minimizing the blocking ratio when SFCs are dynamically established and released during operation.

Given that the number of possible combinations to select M out of N nodes in the network for MEC placement can be very high, performing dynamic simulation to evaluate all configurations is infeasible due to the required computational intensity. To address this challenge, we have analyzed the impact of load balancing across the selected MEC sites on blocking probability. We computed load balance metrics, specifically the JFI, for a hundred of random selections of MEC sites for two different topologies, and conducted detailed simulations for each configuration to determine the corresponding blocking ratios. These analyses have revealed a statistically significant negative monotonic relationship between the JFI and the blocking probability.

Therefore, by employing the JFI as a proxy for dynamic performance, in addition to ensuring proximity of MEC sites to client nodes to meet latency requirements, network operators can estimate which combinations of MEC sites are likely to yield a lower blocking ratio without having to run time-consuming dynamic simulations to evaluate the performance of all potential alternatives. This approach allows network operators to concentrate efforts on conducting detailed dynamic simulations only for the most promising site combinations, significantly reducing the time and computational resources required for network planning and optimization. Moreover, these metrics can also be very valuable in helping to dimension network resources and in determining where to add new MEC sites in a network that currently has a number of these sites but needs to be upgraded.

As part of future work, our goal is to develop a comprehensive planning algorithm that uses the JFI metric to guide the placement of MEC sites, complemented by final detailed dynamic simulations to determine the set of MEC sites with best dynamic performance. We aim to apply this technique to both greenfield scenarios (where no MEC sites initially exist in the network) and brownfield scenarios (where additional MEC sites are to be added alongside existing ones). Moreover, we also plan to incorporate energy efficiency considerations into the MEC placement framework. Thus, exploring techniques to minimize energy consumption by intelligently placing MECs, and optimizing resource allocation are to be included in the next steps.

ACKNOWLEDGMENT

During the preparation of this work the authors used DeepL and ChatGPT in order to improve the grammar and readability of some sentences. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] J.-P. H. Asdikian, L. Askari, O. Ayoub, F. Musumeci, S. Bregni, and M. Tornatore, "Availability evaluation of service function chains under different protection schemes," in *Proc. IEEE Int. Mediterr. Conf. Commun. Netw. (MeditCom)*, 2022, pp. 244–249.

[2] S. Sharma and A. Nag, "Cognitive software defined networking and network function virtualization and applications," *Future Internet*, vol. 15, no. 2, p. 78, 2023.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[4] O. S. Al-Heety, Z. Zakaria, M. Ismail, M. M. Shakir, S. Alani, and H. Alsariera, "A comprehensive survey: Benefits, services, recent works, challenges, security, and use cases for SDN-VANET," *IEEE Access*, vol. 8, pp. 91028–91047, 2020.

[5] I. Alam et al., "A survey of network Virtualization techniques for Internet of Things using SDN and NFV," *ACM Comput. Surv.*, vol. 53, no. 2, pp. 1–40, Apr. 2020. [Online]. Available: https://doi.org/10.1145/3379444

[6] L. Ruiz et al., "Design of VNF-mapping with node protection in WDM metro networks," in *Proc. 10th EAI Int. Conf., Broadnets*, 2019, pp. 285–298.

[7] ETSI industry specification group (ISG), "Mobile edge computing (MEC); deployment of mobile edge computing in an NFV environment," ETSI, Sophia Antipolis, France, Rep. GR MEC 017 V1.1.1, Feb. 2018.

[8] Y. Yu, X. Bu, K. Yang, H. K. Nguyen, and Z. Han, "Network function virtualization resource allocation based on joint benders decomposition and ADMM," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1706–1718, Feb. 2020.

[9] X. Cheng, Y. Wu, G. Min, and A. Y. Zomaya, "Network function virtualization in dynamic networks: A stochastic perspective," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2218–2232, Oct. 2018.

[10] H. D. Chantre and N. L. S. da Fonseca, "The location problem for the provisioning of protected slices in NFV-based MEC infrastructure," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 7, pp. 1505–1514, Jul. 2020.

[11] M. Casazza, P. Fouilhoux, M. Bouet, and S. Secci, "Securing virtual network function placement with high availability guarantees," in *Proc. IFIP Netw. Conf. (IFIP Netw.) Workshops*, 2017, pp. 1–9.

[12] D. Harris and D. Raz, "Dynamic VNF placement in 5G edge nodes," in *Proc. IEEE 8th Int. Conf. Netw. Softwariz. (NetSoft)*, 2022, pp. 216–224.

[13] J. J. A. Esteves, A. Boubendir, F. Guillemin, and P. Sens, "Heuristic for edge-enabled network slicing optimization using the "power of two choices," in *Proc. 16th Int. Conf. Netw. Service Manage. (CNSM)*, 2020, pp. 1–9.

[14] H. Chan, A. Filos-Ratsikas, B. Li, M. Li, and C. Wang, "Mechanism design for facility location problems: A survey," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 4356–4365.

[15] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Eastern Res. Lab., Digit. Equip. Corp., Maynard, MA, USA, Rep. TR-301, Sep. 1984. [Online]. Available: https://www.cs.wustl.edu/ jain/papers/ftp/fairness.pdf

[16] M. Masoumi et al., "Dynamic online VNF placement with different protection schemes in a MEC environment," in *Proc. 32nd Int. Telecommun. Netw. Appl. Conf. (ITNAC)*, 2022, pp. 1–6.
[17] M. Masoumi et al., "Efficient protected VNF placement and MEC location selection for dynamic service provisioning in 5G networks," in *Proc. 20th Int. Conf. Distrib. Comput. Artif. Intell.*, 2023, pp. 448–456.
[18] A. Ghasemzadeh, H. S. Aghdasi, and S. Saeedvand, "Edge server placement and allocation optimization: A tradeoff for enhanced performance," *Clust. Comput.*, vol. 27, pp. 1–15, Feb. 2024.
[19] 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 Rel. Commun. Netw. (DRCN)*, 2021, pp. 1–8.
[20] A. Chiha, B. Denis, S. Verbrugge, and D. Colle, "Techno-economic analysis of MEC clustering models for seamless CCAM service provision," *IEEE Commun. Mag.*, vol. 61, no. 2, pp. 32–37, Feb. 2023.
[21] S. Dash et al., "CRAMP: Clustering-based RANs association and MEC placement for delay-sensitive applications," *J. Netw. Comput. Appl.*, vol. 227, Jul. 2024, Art. no. 103893.
[22] X. Zhang, Z. Li, C. Lai, and J. Zhang, "Joint edge server placement and service placement in mobile-edge computing," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11261–11274, Jul. 2022.
[23] W. Hua, Z. Zhou, and L. Huang, "Location privacy-aware offloading for MEC-enabled IoT: Optimality and heuristics," *IEEE Internet Things J.*, vol. 10, no. 21, pp. 19270–19281, Nov. 2023.
[24] Q. Wu et al., "Joint computation offloading, role, and location selection in hierarchical multicoalition UAV MEC networks: A Stackelberg game learning approach," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18293–18304, Oct. 2022.
[25] S. Lange et al., "Heuristic approaches to the controller placement problem in large scale SDN networks," *IEEE Trans. Netw. Service Manag.*, vol. 12, no. 1, pp. 4–17, Mar. 2015.
[26] O. Adekoya and A. Aneiba, "An adapted nondominated sorting genetic algorithm III (NSGA-III) with repair-based operator for solving controller placement problem in software-defined wide area networks," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 888–901, 2022.
[27] Y. P. Llerena and P. R. Gondim, "SDN-controller placement for D2D communications," *IEEE Access*, vol. 7, pp. 169745–169761, 2019.
[28] J. Ali and B.-H. Roh, "An effective approach for controller placement in software-defined Internet-of-Things (SD-IoT)," *Sensors*, vol. 22, no. 8, p. 2992, 2022.
[29] R. K. Devi and G. Murugaboopathi, "An efficient clustering and load balancing of distributed cloud data centers using graph theory," *Int. J. Commun. Syst.*, vol. 32, no. 5, p. e3896, 2019.
[30] K. Yang, D. Guo, B. Zhang, and B. Zhao, "Multi-controller placement for load balancing in SDWAN," *IEEE Access*, vol. 7, pp. 167278–167289, 2019.
[31] G. Ramya and R. Manoharan, "Prediction based dynamic controller placement in SDN," *EAI Endorsed Trans. Scal. Inf. Syst.*, vol. 8, no. 32, p. e6, 2021.
[32] Y. Guo, S. Wang, A. Zhou, J. Xu, J. Yuan, and C.-H. Hsu, "User allocation-aware edge cloud placement in mobile edge computing," *Softw., Pract. Exp.*, vol. 50, no. 5, pp. 489–502, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2685
[33] X. Xu et al., "Edge server quantification and placement for offloading social media services in industrial cognitive IoV," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2910–2918, Apr. 2021.
[34] J. Liu, Z. Xu, C. Wang, X. Liu, X. Xie, and G. Shi, "Mobility-aware MEC planning with a GNN-based graph partitioning framework," *IEEE Trans. Netw. Service Manag.*, vol. 21, no. 4, pp. 4383–4395, Aug. 2024.
[35] Z. Zhao, H. Cheng, X. Xu, and Y. Pan, "Graph partition and multiple choice-UCB based algorithms for edge server placement in MEC environment," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 4050–4061, May 2024.
[36] P.-C. Huang, T.-L. Chin, and T.-Y. Chuang, "Server placement and task allocation for load balancing in edge-computing networks," *IEEE Access*, vol. 9, pp. 138200–138208, 2021.
[37] H. Moens and F. De Turck, "VNF-P: A model for efficient placement of virtualized network functions," in *Proc. 10th Int. Conf. Netw. service Manage. (CNSM) Workshop*, 2014, pp. 418–423.
[38] M. F. Bari, S. R. Chowdhury, R. Ahmed, and R. Boutaba, "On orchestrating virtual network functions," in *Proc. 11th Int. Conf. Netw. service Manage. (CNSM)*, 2015, pp. 50–56.
[39] M. Dieye et al., "CPVNF: Cost-efficient proactive VNF placement and chaining for value-added services in content delivery networks," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 2, pp. 774–786, Jun. 2018.
[40] Q. Zhang, F. Liu, and C. Zeng, "Adaptive interference-aware VNF placement for service-customized 5G network slices," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 2449–2457.
[41] N. Slamnik-Kriještorac, H. C. C. de Resende, C. Donato, S. Latré, R. Riggio, and J. Marquez-Barja, "Leveraging mobile edge computing to improve vehicular communications," in *Proc. IEEE 17th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2020, pp. 1–4.
[42] J. Sun, F. Liu, H. Wang, and D. O. Wu, "Joint VNF placement, CPU allocation, and flow routing for traffic changes," *IEEE Internet Things J.*, vol. 10, no. 2, pp. 1208–1222, Jan. 2023.
[43] B. Nogales et al., "Integration of 5G experimentation infrastructures into a multi-site NFV ecosystem," *J. Vis. Exp.*, vol. 168, no. 168, 2021, Art. no. e61946.
[44] M. Mosahebfard, J. S. Vardakas, and C. Verikoukis, "Modelling the admission ratio in NFV-based converged optical-wireless 5G networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 12024–12038, Nov. 2021.
[45] M. Wang, B. Cheng, W. Feng, and J. Chen, "An efficient service function chain placement algorithm in a MEC-NFV environment," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.
[46] M. Karimzadeh-Farshbafan, V. Shah-Mansouri, and D. Niyato, "Reliability aware service placement using a viterbi-based algorithm," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 1, pp. 622–636, Mar. 2020.
[47] L. Contreras (Eur. Comm., Brussels, Belgium). *D1.2-Final 5G-Crosshaul System Design and Economic Analysis*. (2017). Accessed: Aug. 01, 2023. [Online]. Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b77d8419&appId=PPGMS
[48] J. M. Ziazet, B. Jaumard, H. Duong, P. Khoshabi, and E. Janulewicz, "A dynamic traffic generator for elastic 5G network slicing," in *Proc. IEEE Int. Symp. Meas. Netw. (M N)*, 2022, pp. 1–6.
[49] L. Ruiz et al., "Genetic algorithm for holistic VNF-mapping and virtual topology design," *IEEE Access*, vol. 8, pp. 55893–55904, 2020.
[50] "Japan Photonic network model (JPNM)." 2016. Accessed: Aug. 1, 2023. [Online]. Available: https://www.ieice.org/cs/pn/eng/jpnm_en.html

MARYAM MASOUMI received the B.Sc. degree in communication engineering from the Shiraz University of Technology, Iran, in 2012, and the M.Sc. degree in communication engineering from Yazd University, Yazd, Iran, in 2016. She is currently working as a Ph.D. Fellow under the framework of Marie Skłodowska-Curie Actions H2020 IoTalentum Project with the Universidad de Valladolid, Spain. Her research interests are mainly focused on network planning, resource allocation, edge computing for 5G/6G services, softwarization, and virtualization of computer networks.

IGNACIO DE MIGUEL (Senior Member, IEEE) received the degree in telecommunication engineering and the Ph.D. degree from the Universidad de Valladolid, Spain, in 1997 and 2002, respectively, where he is an Associate Professor. He has published more than 40 papers in international journals and more than 200 conference papers. His main research interests include the design, control, and performance evaluation of communication networks.

FABRIZIO GABRIO BRASCA received the Graduate degree in computer science in 2000. He is a Technology Architect with the Technology Architecture and Remedies Management Department, Wind Tre. He has been involved in several European projects and in different research programs in automotive, energy, smart-city, and industry 4.0 areas.

NOEMÍ MERAYO received the Telecommunication Engineer and Ph.D. degrees from the Universidad de Valladolid, Spain, in 2004 and 2009, respectively, where she works as a Lecturer since 2005. She has been a Visiting Research Fellow with the University of Hertfordshire, the University of Zaragoza, and the Technology University or Munich. Her research focuses on the design and performance evaluation of optical networks and the application of artificial intelligence techniques.

RUBÉN M. LORENZO received the Telecommunication Engineer and Ph.D. degrees from the Universidad de Valladolid, Spain, in 1996 and 1999, respectively, where he has been a Lecturer since 2000 and was a Junior Lecturer with Optical Communications Group from 1996 to 2000. His research interests include communication networks and multi-access edge computing.

RAMÓN J. DURÁN BARROSO received the Telecommunication Engineer and Ph.D. degrees from the University of Valladolid, Spain, in 2002 and 2008, where he is currently a Full Professor. He coordinates the H2020 MSCA IoTalentum Project. He has published more than 50 papers in JCR journals and more than 130 conference papers. His research focuses on the use of artificial intelligence techniques for the design, optimization and operation of heterogeneous networks, multi-access edge computing, and network function virtualization.