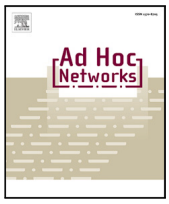
Ad Hoc Networks 177 (2025) 103912

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc



Optimizing load-balanced resource allocation in next-generation mobile networks: A parallelized multi-objective approach

Jesús Calle-Cancho [a], Jesús Galeano-Brajones [b,*], David Cortés-Polo [a], Javier Carmona-Murillo [b], Francisco Luna-Valero [c,d]

[a] Dpto. de Ingeniería de Sistemas Informáticos y Telemáticos, Universidad de Extremadura, Escuela Politécnica, Cáceres, 10003, Spain
[b] Dpto. de Ingeniería de Sistemas Informáticos y Telemáticos, Universidad de Extremadura, Centro Universitario de Mérida, Mérida, 06800, Spain
[c] ITIS Software, Universidad de Málaga, Edificio de Investigación Ada Byron, Málaga, 29071, Spain
[d] Dpto. de Lenguajes y Ciencias de la Computación, Universidad de Málaga, E.T.S.I. Informática, Málaga, 29071, Spain

ARTICLE INFO

Keywords:
Network resource allocation
Beyond 5G
6G
Multi-objective optimization
MOEA

ABSTRACT

The rapid evolution of mobile communications, driven by the proliferation of mobile devices and data-intensive applications, has driven an unprecedented increase in data traffic, pushing the current network infrastructure to its limits. In Beyond 5G and future 6G networks, minimizing network latency is crucial to support next-generation applications, such as immersive media, autonomous systems, and critical real-time services, all of which demand ultra-low latency and high reliability. In Multi-access Edge Computing environments, where future 6G networks will be deployed, efficient allocation of virtual base stations to the access network in dense environments will be essential to optimize performance and maintain quality of service. This efficient allocation will be key to effectively addressing the challenges present in these settings. This paper addresses this problem through a parallelized multi-objective evolutionary algorithm that simultaneously optimizes signaling delay, data plane overhead, and load balancing. By leveraging a Pareto-based approach, we provide a set of optimal trade-offs that enhance network adaptability and efficiency beyond traditional single-objective methods. Moreover, we introduce a novel metric inspired by the Sharpe ratio to evaluate the efficiency of load distribution across the network. Experimental results in various network topologies show that our approach significantly enhances network performance, achieving reductions in data plane overhead of up to 51.5% and 77.9% in signaling delay compared to a state-of-the-art solution based on a specialized heuristic. By providing a set of non-dominated solutions, our approach enables network operators to select configurations that best meet specific quality of service requirements and service priorities, thereby improving network adaptability and resilience under varying conditions.

1. Introduction

Mobile communications have undergone substantial transformation in recent years, largely due to the widespread adoption of mobile devices that produce unprecedented data traffic volumes. A recent Ericsson forecast anticipates that global mobile subscriptions will reach 8.6 billion by the close of 2028 [1]. Moreover, mobile data traffic worldwide is projected to increase nearly fourfold from 2023 to 2028, reaching about 53 exabytes per month by that time. This rapid growth is fueled by the continuous expansion of 5G networks and the rise of data-heavy applications. 5G represented a significant leap in the evolution of mobile networks, capable of supporting higher data transmission speeds, enabling rapid analysis and management of massive data flows, and improving the utilization of network resources with greater efficiency than that of previous generations. However, mobile communication technologies must adapt even further to support this increased demand and ensure that new services and applications cater to users' needs. These capabilities address a wide range of applications, from data-intensive streaming services to critical infrastructure systems that demand near-zero latency [2]. Advanced 5G and future 6G technologies are being designed to address the complex requirements of ultra-dense deployments, which require stringent levels of latency and reliability.

From a network architecture perspective, 3GPP has made substantial efforts to advance the standard, laying the foundation for what is known as Advanced 5G (3GPP Release-18) [3]. This release introduces features that provide greater flexibility and efficiency to the

* Corresponding author.
E-mail address: jgaleanobra@unex.es (J. Galeano-Brajones).

https://doi.org/10.1016/j.adhoc.2025.103912
Received 14 November 2024; Received in revised form 20 March 2025; Accepted 10 May 2025
Available online 27 May 2025

network and examines the evolution of network architectures to optimize deployments. Key focus areas include energy efficiency, coverage, mobility support, and positioning, among others [4]. This evolution is expected to pave the way for the sixth generation of mobile networks (6G), designed to meet demands for higher data rates, lower latencies, and greater reliability across ultra-dense networks supporting applications such as immersive media, autonomous systems, and the Internet of Everything (IoE) [5,6].

In previous generations, network architecture relied on centralized mobility management, with mobility anchors handling data forwarding and signaling. Although effective for moderate traffic loads, this approach posed challenges such as non-optimal routing, scalability issues, and reliance on centralized components, creating single points of failure. The high demands of 5G required a shift to more flexible, decentralized architectures [7]. In response, 5G introduced a service-based architecture that positions mobility anchors closer to end-users, reducing bottlenecks and enabling a *flatter* network structure. The primary mobility functions managed by the Access Mobility Function (AMF) and Session Management Function (SMF), enhance the control plane's efficiency, while the User Plane Function (UPF) improves data delivery by placing the closer to users, particularly through integration with Multi-access Edge Computing (MEC) at the network edge [8]. This integration minimizes latency by reducing the distance between users and computational resources, which is essential to maintain uninterrupted service during user movement.

As future 6G networks further intensify these demands, effective allocation of virtual base stations (vBS) to access network nodes in MEC environments will become critical to balancing load, reducing signaling overhead, and minimizing latency in dynamic, high-density environments. vBS, or software-based representations of traditional base stations, offer the flexibility to adjust to evolving user densities and mobility patterns, thus ensuring efficient resource utilization and enhanced quality of service (QoS) [9,10]. A vBS is essentially composed of two main components: the Remote Radio Head (RRH) and the Baseband Unit (BBU). The RRH is responsible for handling radio frequency functions, transmitting and receiving signals over the air to connect with mobile users. BBU processes the baseband signals and manages control functions. In modern network architectures, these components are often separated, allowing the BBU to be centralized in a datacenters while the RRH remains at the site [11]. In MEC, this separation enables virtualization, where multiple BBUs can support several RRHs in a scalable and flexible configuration, optimizing network resources [12]. By formulating this allocation as an optimization problem, network operators can address the unique challenges of next-generation mobility management and support the diverse requirements of future ultra-connected 6G applications.

In this paper, we address the challenge of vBS-to-access network allocation in highly dense 5G and future 6G networks by leveraging a parallel multi-objective evolutionary algorithm (MOEA) [13,14] which is capable of running on thousands of processing elements (cores). Unlike traditional heuristic approaches, which prioritize fast but potentially suboptimal solutions, our method simultaneously optimizes signaling delay, data plane overhead, and load balancing to ensure a more efficient and scalable network configuration. Network operators commonly use historical demand data and predictive models to anticipate network traffic and mobility patterns, ensuring optimal infrastructure configuration. By integrating a multi-objective optimization approach, our method enhances planning by simultaneously minimizing signaling delay, reducing data plane overhead, and improving load balancing. This approach generates a diverse set of non-dominated solutions using a Pareto-based strategy, providing network operators with optimal trade-offs that enhance adaptability and efficiency beyond traditional single-objective methods [15]. While computationally demanding, our parallelized implementation enables efficient exploration of the solution space, making it a practical and effective tool for next-generation network planning. There exist alternative optimization techniques, such as surrogate-assisted approaches [13], which aim to reduce computational costs by approximating objective functions. However, these surrogate methods often incur significant approximation errors, making them unsuitable for accurate evaluation in highly complex and resource-intensive scenarios such as ours. Furthermore, linear programming-based optimization methods are exact and guarantee optimal solutions for smaller scenarios but become computationally infeasible as the search space grows, making them impractical for the large-scale network configurations addressed in this study.

This work builds upon a previous approach presented in [10], which introduced a heuristic-based allocation algorithm, named LNA (Link-Network Assignment), focused on minimizing signaling delay and data plane overhead in 5G networks. In that study, LNA achieved the best results compared to widely used heuristics, demonstrating its superiority among traditional approaches. Therefore, in this paper, we focus on comparing our newly proposed method against LNA, as it represents the strongest existing alternative. Our method extends the previous optimization by simultaneously addressing three objectives: minimizing signaling delay, minimizing data plane overhead, and maximizing load balance across the network. This focused comparison allows us to highlight the effectiveness of our novel approach in addressing resource allocation challenges in next-generation networks.

A key contribution of this paper is the introduction of a metric inspired on the Sharpe ratio [16], a well-known measure of the excess return per unit of risk of a financial investment, as a novel metric to measure the efficiency of network load distribution, ensuring a more balanced allocation of vBS to routers. To ensure a comprehensive evaluation, we tested the proposed method across three network topologies (low-connected, hybrid-connected, and high-connected) each with 30 distinct scenarios. These topologies vary in router connectivity levels, allowing us to assess the robustness of the MOEA under different network configurations.

The rest of this paper is organized as follows. Section 2 presents the system model and formulates the optimization problem, detailing the network architecture and the objectives to be optimized. Section 3 describes the methodology, including the simulation setup, network topologies used for experimental evaluation, evaluation metrics and the MOEA-based framework. Section 4 provides a comparative analysis between the heuristic LNA algorithm from previous work and the proposed MOEA approach, assessing their performance from both a multi-objective optimization perspective and a performance metrics perspective. Finally, Section 5 concludes the paper, summarizing the findings and outlining potential directions for future research.

2. System model and problem formulation

This section presents the system model on which the multi-objective optimization problem is defined to minimize the impact of vBS allocation to the edge nodes of the access network, while maintaining network performance.

2.1. Mobility domain and access network

Let the access network be represented as an undirected graph $G = (V, E)$, where V denotes the set of nodes (including routers and vBS) and E represents the edges (communication links) between these nodes. Let $K \subset V$ be the set of access routers that provide connectivity to a set of vBS B. Each vBS is represented by b_i with $i \in \{1, 2, \ldots, |B|\}$, where $|B|$ is the total number of vBS. Each vBS consists of a conjunction of an RRH and a vBBU.

The set of vBS B provides full coverage to a geographical area of interest, where each location in this area is defined by the coordinates $\{L_{b_i}\}_{i=1}^{|B|}$, with $L_{b_i} \in \mathbb{R}^2$ representing the position of each vBS. Note that the position of each vBS is actually the position of its corresponding RRH, as the virtual base-band units (vBBUs) are co-located with the access routers. This setup ensures that the entire domain is efficiently

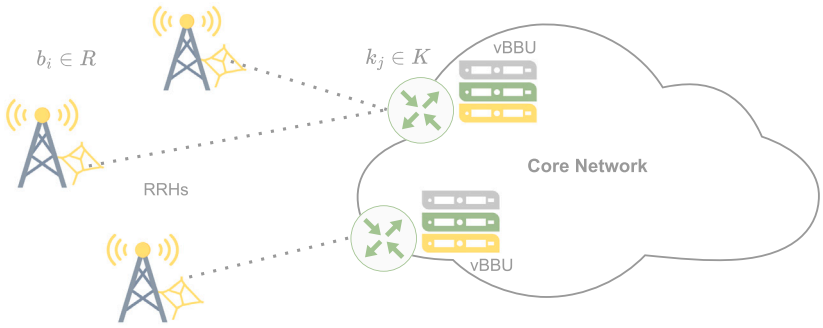


Fig. 1. RRH allocation to edge access routers deployed on core network.

served by the available infrastructure. The goal of this network design is to optimize the allocation of vBS to routers, thereby minimizing operational costs and achieving a balanced load distribution across the network.

Fig. 1 illustrates the structure of the allocation problem, which defines the placement of vBS units and their possible connections to the access network. In this figure, the core network serves as a centralized system that coordinates communication between the vBS and access routers, enabling seamless connectivity for mobile users throughout the coverage area. The links between vBS and routers are shown as potential paths that can be optimized based on the objectives of minimizing signaling delays, reducing data plane overhead, and balancing the load across network nodes.

2.2. Virtual base station allocation and mobile nodes support

Each access router k_j, with $j \in \{1, 2, \ldots, |K|\}$, manages a specific subset of vBS denoted by $B_j \subset B$. These routers function as the first level of communication within the network domain, ensuring that data and control packets are efficiently routed. Additionally, vBBUs are co-located with the access routers to facilitate processing and resource management.

The assignment of vBS to routers is represented by the vector $\vec{a} = \langle a_1, a_2, \ldots, a_{|B|} \rangle$, where each element $a_m \in \{1, 2, \ldots, |K|\}$ specifies the access node to which vBS b_m is allocated. This ensures that each vBS is connected to exactly one router and, conversely, one router may have assigned any number of vBSs, even none. That is the reason because it is important to include a load balancing mechanism. For example, if $\vec{a} = \langle 2, 1, 3, \ldots, 2 \rangle$, then vBS b_1 is assigned to router 2, b_2 to router 1, b_3 to router 3, and so on. This vector-based representation provides a compact format for defining and manipulating assignments within the optimization process.

The first objective corresponds to the signaling delay $D_{signaling}$, which reflects the overhead due to the signaling messages required for mobility management when handovers occur. It is defined as:

$$\min_{\vec{a}} D_{signaling} = \sum_{m=1}^{|B|} f_s(m, a_m), \tag{1}$$

where $f_s(m, a_m)$ is the signaling delay associated with assigning the vBS m to the router a_m normalized by the network capacity.

The second objective to minimize is the overhead of the packet delivery in the data plane O_{data}, which quantifies the resources needed to transmit the data packets from the vBS to the end users:

$$\min_{\vec{a}} O_{data} = \sum_{m=1}^{|B|} f_d(m, a_m), \tag{2}$$

where $f_d(m, a_m)$ indicates the cost of forwarding packets when the vBS m is assigned to the router a_m.

Finally, to achieve a balanced load across all access routers, we maximize the metric inspired by the Sharpe ratio [16]. Originally developed in Finance, the Sharpe ratio measures the risk-adjusted return of an investment. The Sharpe Ratio S is defined as:

$$S = \frac{E[P] - P_f}{\sigma} \tag{3}$$

where:

- $E[P]$ represents the mean expected performance.
- P_f is the baseline or risk-free performance.
- σ denotes the performance volatility, specifically the standard deviation of performance.

If the risk-free performance P_f is not considered (which will be explained later), the Sharpe Ratio can be computed as:

$$S = \frac{\mu}{\sigma} \tag{4}$$

where μ is the mean return and σ is the standard deviation of the returns.

Specifically, in this work, we perform the mapping of the original Sharpe ratio components to our load distribution metric as follows. The expected performance ($E[P]$) is translated into our optimization problem as the mean number of virtual base stations assigned per router in the network. As for the risk-free performance (P_f), a baseline return used in financial applications to evaluate excess returns, does not have a direct counterpart in our network load balancing problem. Instead, we assess the relative efficiency of load distribution without defining a strict baseline. Thus, in our adaptation, P_f is implicitly considered to be zero, meaning that the balance metric is evaluated purely based on the observed deviation from a uniform distribution. Moreover, since the Sharpe ratio is used to compare solutions, P_f is a constant term (the same for all solutions in the population) that can be eliminated from the comparisons. Finally, the performance volatility (σ) is mapped to the standard deviation of the vBS assignment across routers in our network. In this context, we adapt the Sharpe ratio to quantify the efficiency of load distribution among access routers, representing it as the ratio of mean load to the standard deviation of load across routers. A higher ratio indicates a more balanced and efficient load distribution. To our knowledge, this is the first time that this metric has been used to optimize a network management problem. However, it has been used in other areas, such as operational research [17] and portfolio optimization [18].

The total number of vBS assigned to the router r is given by:

$$|B_r^{\vec{a}}| = \sum_{m=1}^{|B|} \delta(a_m, r), \forall\ a_m \in \vec{a} \tag{5}$$

where $\delta(a_m, r)$ is an indicator function defined as:

$$\delta(a_m, r) = \begin{cases} 1 & \text{if vBS } m \text{ is assigned to router } r, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The mean number of vBS assigned per router, $\mu_B^{\vec{a}}$, is calculated as:

$$\mu_B^{\vec{a}} = \frac{1}{|K|} \sum_{r=1}^{|K|} |B_r^{\vec{a}}|, \tag{7}$$

which represents the average load per router.

The standard deviation of the load distribution across routers, which measures the variation in the number of vBS assigned to each router, is given by:

$$\sigma_B^{\vec{a}} = \sqrt{\frac{1}{|K|-1}\sum_{r=1}^{|K|}\left(|B_r^{\vec{a}}| - \mu_B^{\vec{a}}\right)^2}. \tag{8}$$

Thus, the third objective to maximize the metric inspired by the Sharpe ratio $S_{balance}$ is defined as:

$$\max_{\vec{a}} S_{balance} = \frac{\mu_B^{\vec{a}}}{\sigma_B^{\vec{a}}}. \tag{9}$$

Maximizing this ratio encourages a uniform load distribution, reducing the likelihood of overloading specific routers and promoting efficient resource utilization across the network.

2.3. Constraints

The optimization problem is subject to the following constraints:

1. **Router allocation constraint.** Decision variables a_m are integers that indicate the router to which each vBS m is assigned, ensuring that each vBS is associated with exactly one router:

$$a_m \in 1, 2, \ldots, |K|, \quad \forall m \in B. \tag{10}$$

2. **Load balancing constraint.** The standard deviation of the number of vBS assigned to the access nodes must not exceed a specified threshold $\sigma_{\max}$, with the goal of achieving the most balanced load distribution possible:

$$\sigma_B \leq \sigma_{\max}. \tag{11}$$

2.4. Multi-objective optimization problem

The complete multi-objective optimization problem can be summarized as follows:

$$\min_{\vec{a}} D_{signaling} = \sum_{m=1}^{|B|} f_s(m, a_m), \tag{12}$$

$$\min_{\vec{a}} O_{data} = \sum_{m=1}^{|B|} f_d(m, a_m), \tag{13}$$

$$\max_{\vec{a}} S_{balance} = \frac{\mu_B^{\vec{a}}}{\sigma_B^{\vec{a}}}, \tag{14}$$

subject to (15)

$$a_m \in \{1, 2, \ldots, |K|\}, \quad \forall\, m \in B, \tag{16}$$

$$\sigma_B \leq \sigma_{\max}. \tag{17}$$

3. Methodology

In this section, the simulation setup is described first, detailing the experimental conditions and parameters selected to ensure reliable and reproducible results. Next, the network topologies used to assess the adaptability of the algorithm across varying connectivity levels are explained. Following this, we define the metrics employed to evaluate the quality of solutions in terms of signaling delay, data plane overhead and load balance. Finally, the asynchronous distributed steady-state MOEA is presented, with details on the distributed framework and parameters to optimize the multi-objective problem efficiently.

3.1. Simulation setup

The simulation scenario consists of a square region with an area of 10×10 km^2, where vBS are distributed following a Poisson Point Process (PPP), suitable for modeling random distributions in dense environments. The intensity of the PPP (λ_{vBS}) corresponds to the average number of vBS (N_{vBS}) per unit area (A) [19], calculated as $\lambda_{vBS} = N_{vBS}/A$. In next-generation mobile networks, cell densification is commonly modeled in this way, enabling the construction of ultra-dense networks. This approach allows for more efficient utilization of spatial resources and enhances network capacity by deploying a high density of vBS in limited areas [20,21]. Additionally, the coverage areas of the vBS are modeled as a Poisson–Voronoi tessellation in a two-dimensional plane, where each mobile user connects to the nearest vBS. Moreover, the value of $\sigma_{\max}$ is set to 2 to ensure a fair and meaningful comparison with the heuristic-based approach while simultaneously allowing the multi-objective evolutionary algorithm to explore a broader search space.

User mobility is modeled using the Random Waypoint approach, with velocities uniformly distributed between 1 and 20 m/s. In each simulation, mobile users move across the mobility domain, connecting to different vBS following the mobility model specified in [22]. These users manage a set of sessions throughout the simulation. The incoming sessions for each user are assumed to follow a Poisson process with an average arrival rate of $\lambda = 0.01$. The duration of the session is modeled as an exponentially distributed variable with a parameter $\mu = 10$ [23]. Furthermore, the flow rate requirement for each session varies between 1500 Kbps and 10 Mbps (e.g. video streams) [24].

3.2. Network topologies

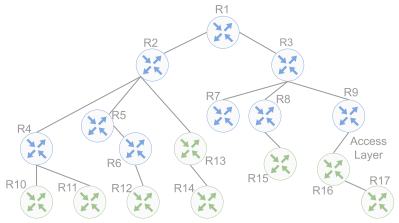
Network topology plays a crucial role in influencing performance metrics, particularly in heterogeneous environments. To evaluate the adaptability of the proposed approach, we consider three types of network topologies with varying levels of connectivity [25], as illustrated in Fig. 2. Fig. 2(a) is a low-connected topology in which routers do not connect to other routers at the same level; instead, they connect to routers at lower levels in the hierarchy. Fig. 2(b) is a hybrid-connected topology in which mobile nodes will traverse both the connected and sparse areas of the network. In the highly connected area of this topology, there are direct connections between the routers of the backbone layer and the access layer. Finally, Fig. 2(c) shows the high-connected topology where there is hierarchical connectivity between routers, along with connections between routers at the same level in the hierarchy.

3.3. Performance metrics

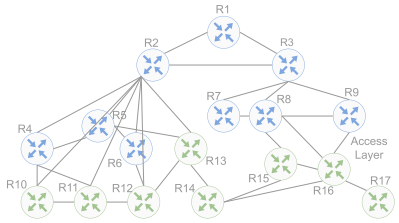
The performance metrics used to evaluate the efficiency of the MOEA and compare it to the LNA algorithm are selected to capture critical aspects of control plane and data plane performance. These metrics have been widely used in prior analyses and evaluations of mobility solutions [26,27], allowing for the assessment of overall mobile network performance.

3.3.1. Signaling delay

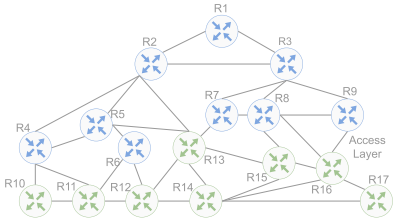
Efficient service provisioning and optimized use of network resources in next-generation networks require addressing challenges beyond the data link layer, extending into layer 3 (L3) management protocols. Mobility management mechanisms ensure that mobile users remain reachable and can maintain ongoing communication as they roam across different networks. In densified network environments, signaling plays a crucial role in performance, particularly as high-speed mobile nodes experience frequent handovers, resulting in a high signaling load due to the short cell radius [26].



(a) Low-connected network topology



(b) Hybrid-connected network topology



(c) High-connected network topology

Fig. 2. Network topologies used in the experimental evaluation.

Maintaining established communications during the movement of a mobile node across the network involves L3 handover processes, which require the exchange of control messages between various network entities. These control messages introduce delays that can undermine the low-latency goals of 5G. In a first-come, first-served transmission setup, signaling packets must wait until all preceding packets have been sent, increasing overall delay. Therefore, in this article, signaling delay is considered an important metric for evaluating the control plane, omitting the propagation latency of the transmission medium, as these two values are assumed to be identical across the scenarios we compare. This approach aims to ensure the fairest possible comparison. For an association between a vBS m and a router a_m, the signaling delay of an allocation is defined as $f_s(m, a_m)$, as shown in Eq. (18). This function represents the signaling cost associated with allocating resources from the vBS m to the router a_m and depends on the specific parameters of this association. This metric depends on the size of signaling messages (s_u), the number of anchors ($N_A(m, a_m)$) with active sessions to a particular mobile node, and the available bandwidth ($BW(m, a_m)$) in the transmission medium. The parameter values used in the simulations were obtained from [10] to ensure a fair comparison with the baseline approach.

$$f_s(m, a_m) = 2s_u \frac{1 + N_A(m, a_m)}{BW(m, a_m)}. \tag{18}$$

3.3.2. Data plane overhead

In the data plane, one of the metrics that significantly impacts overall network performance is overhead. In addition to the signaling load associated with mobility management, data packets also need to be transmitted to the mobile node. This value is influenced by the average size of data messages, which is multiplied by the number of hops required to reach the mobile node. Data plane overhead is particularly relevant for applications with high data demands, such as video streaming, where network efficiency directly affects user experience.

To ensure a fair comparison between MOEA and LNA methods, we adopt this metric as defined in [27,28]. For each allocation between a vBS m and a router a_m, the data plane overhead is defined as shown in Eq. (19):

$$f_d(m, a_m) = (p_n \; P_c^d(m, a_m) + p_h \; P_c^i(m, a_m)) \; N_{p/s}, \tag{19}$$

where $N_{p/s}$ represents the transmission rate of packets per active flow. Furthermore, p_n and p_h are the probabilities that a traffic flow is new or remains open after a handover, respectively, and $P_c^d(m, a_m)$ and $P_c^i(m, a_m)$ represent the overhead for direct and indirect modes, respectively [29]. Just as with the previous metric, the simulation parameters were sourced from [10] to facilitate an equitable comparison against the baseline approach.

3.4. Asynchronous distributed steady-state MOEA

A multi-objective evolutionary algorithm (MOEA) has been employed, specifically an asynchronous distributed steady-state implementation of the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [30]. MOEAs are widely used to optimize problems with multiple conflicting objectives, as they are capable of approximating the Pareto front (a set of optimal solutions where no objective can be improved without worsening another) in one single run [31,32]. In this case, our goal is to optimize three objectives: signaling delay, data plane overhead, and the metric inspired by the Sharpe ratio, which balances the load distribution of vBS among routers, as described in Section 2.

The optimization process begins with the MOEA generating an initial population of candidate solutions, each representing a different allocation of vBS to access nodes. These solutions are evaluated to obtain the objective values. Then, standard evolutionary operators from the literature are applied to iteratively refine the population and explore the search space, enhancing the trade-offs between the three objectives. The evolutionary cycle continues until the stopping condition is reached, which, based on preliminary tests, ensures that the MOEA has stopped making further improvements in approximating the Pareto front. In this work, we opted for a steady-state MOEA due to its replacement strategy, which supports an arbitrarily large number of worker nodes, allowing greater flexibility in a distributed environment. Unlike generational models, which update the entire population simultaneously, the steady-state approach replaces only a few individuals at a time, facilitating parallel evaluations without being limited by the population size.

The diagram in Fig. 3 presents the flow of a classical evolutionary algorithm but incorporates two key features: (i) the evaluation of candidate solutions is delegated to worker nodes, independent of the node running the algorithm, to handle the hardware and temporal resource demands of each evaluation; and (ii) after the initial population is created, the algorithm operates asynchronously, receiving evaluated solutions from the worker nodes as they become available. This allows

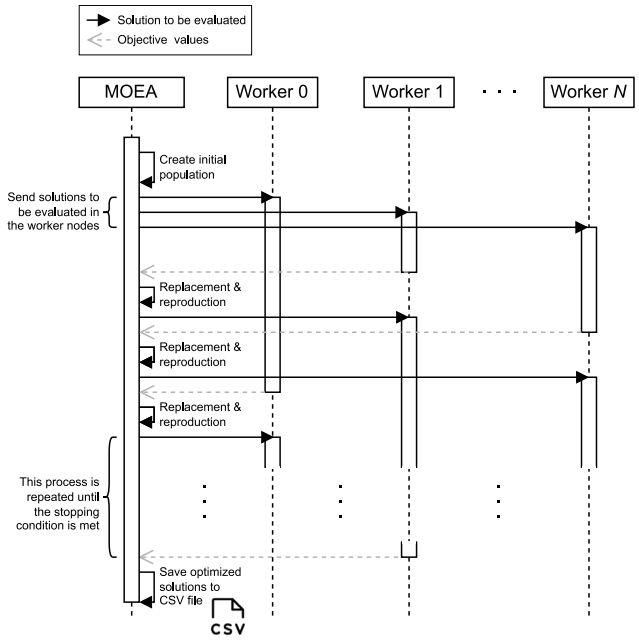


Fig. 3. Overview of the algorithmic flow in the proposed MOEA-based approach.

the algorithm to perform replacement and reproduction continuously, integrating evaluated solutions into the current population without waiting for the entire population to be evaluated.

Evaluating each solution in this optimization problem requires running computationally intensive simulations to obtain objective values, with each evaluation taking approximately 30 s. Given the number of candidate solutions required by the MOEA to sample the search space composed of all the possible vBs-to-access node assignments, a sequential approach would be infeasible. To address this, we implemented a distributed master/worker setup in which a master node manages the evolutionary process and delegates the evaluation of solutions to multiple worker nodes operating in parallel [14]. Each worker is responsible for computing the objective functions ($D_{signaling}$, O_{data}, and $S_{balance}$) for the assigned solutions using discrete-event network simulations implemented in Python, leveraging libraries such as NetworkX and Scipy. This parallelized approach improves scalability and ensures a more accurate assessment across network scenarios.

Specifically, LNA requires approximately 30 s to generate a solution; however, it does not guarantee that this solution is optimal, as demonstrated in the present study. Moreover, the solution obtained with LNA does not allow for the selection among multiple alternatives, which can be a crucial aspect during the network deployment. In contrast, MOEA not only provides an optimal solution but also generates alternative trade-off solutions that balance different objectives. This flexibility enables decision-makers to choose the most suitable configuration based on specific requirements or constraints at any given moment, significantly enhancing the adaptability and robustness of the network deployment process. Therefore, while our MOEA-based algorithm involves a higher computational cost, its ability to find optimal solutions makes it more suitable for network planning applications where solution quality is the top priority. Furthermore, our approach leverages parallelism, allowing us to obtain results in reasonable execution periods.

To significantly reduce computation times and make optimization viable for realistic scenarios, we employ a parallel approach that uses the computational capabilities of the Picasso supercomputer.[1] Given that each individual solution evaluation takes approximately 30 s, the sequential execution of the full optimization would be extremely time-consuming. In fact, a single complete optimization run without parallelization would take around 40,000 min (approximately 27.8 days). In contrast, our parallelized implementation drastically reduces computation times, achieving reductions from 50% with only two worker nodes to 99.8% with 500 workers, as shown in Table 1. The scalability of our parallelized solution allows further reduction in computational time by incorporating more workers, though incremental gains become marginal beyond a certain threshold due to diminishing returns of parallel efficiency.

Table 1
Computational time and time reduction for different number of workers.

Workers	Computation time [minutes]	Time reduction [%]
Non-parallelized	39,630	00.0
2	19,815	50.0
10	3,963	90.0
100	396	99.0
500	79	99.8

Given the stochastic nature of evolutionary algorithms, we performed 30 independent runs for each topology described in Section 3.1. Each run used a different random seed to explore variations in the topologies, ensuring robustness in the results. The final solution for each topology was selected based on the highest $S_{balance}$, that is, prioritizing the solutions with the highest balance of vBS assigned to access nodes. This approach also allowed a fair comparison with the LNA heuristic algorithm [10] used in a previous study, ensuring that the solutions generated by MOEA maintained a high degree of consistency across different scenarios.

4. Performance evaluation

This section presents a comprehensive evaluation of the proposed approach. In the first subsection, we analyze the results from the perspective of the optimization process itself, focusing on the performance of the MOEA in terms of convergence and solution quality with respect to LNA. Then, the second subsection examines the network efficiency of the two algorithms, MOEA and LNA, specifically analyzing the two primary objectives of the optimization problem: data plane overhead (O_{data}) and signaling delay ($D_{signaling}$). This comparison provides a detailed look at how each approach addresses these critical aspects of network efficiency.

It is important to note that although MOEA optimizes a problem with three objectives, the results analysis focuses mainly on $D_{signaling}$ and O_{data}, as $S_{balance}$ is designed primarily to achieve load distribution balance. This focus allows us to directly evaluate the impact of the MOEA on the efficiency of the network in terms of signaling delay and data plane overhead, while the balancing objective acts to ensure a more evenly distributed load across the network nodes. For access to raw results and 3D visualizations of the Pareto front approximations, please refer to the publicly available repository on GitHub.[2]

4.1. Multi-objective optimization analysis

First, we analyze the results from the perspective of multi-objective optimization. The outcome of a MOEA is a set of non-dominated solutions known as an approximation to the Pareto front. In the context of multi-objective optimization, a solution is considered non-dominated if there is no other solution that performs better in one objective without worsening in at least one other objective. The Pareto front, therefore,

[1] https://www.scbi.uma.es/web/.

[2] https://github.com/galeanobra/ResourceAllocationMOEA.

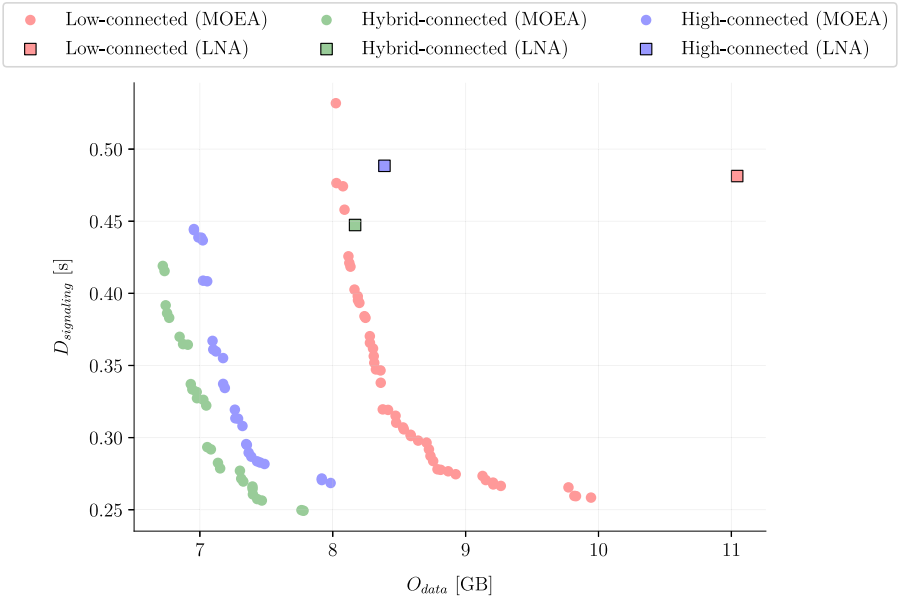


Fig. 4. Attainment surfaces achieved by the MOEA (∘) and median solutions obtained by LNA (□) for the three topologies.

represents the set of optimal trade-offs between the competing objectives, providing decision-makers with a range of solutions that balance the different performance metrics.

To evaluate the median performance of the MOEA across all problem instances and facilitate a comparison with the heuristic-based LNA approach, we use attainment surfaces. The Empirical Attainment Function (EAF) [33] is a graphical tool used to examine the Pareto front approximations generated by multi-objective optimization algorithms. Specifically, the EAF visualizes the expected performance and variability of the fronts across multiple iterations of the MOEA. In simpler terms, the 50%-attainment surface serves a role similar to the median in single-objective optimization, providing a visual representation of the typical performance of the algorithm. This allows us to compare the overall behavior of the MOEA with other approaches, such as LNA, and observe the median effectiveness of the MOEA across different scenarios.

Next, we proceed by analyzing the attainment surfaces of the MOEA and comparing them to the median performance of LNA across all problem instances, categorized by the three topologies described in Section 3.2. This comparison allows us to evaluate the effectiveness of the MOEA in optimizing the data plane overhead and signaling delay under different network configurations.

Fig. 4 illustrates the comparison between MOEA and LNA in terms of data plane overhead and signaling delay for each topology. From an optimization standpoint, the attainment surfaces of the MOEA reveal a broader range of non-dominated solutions, demonstrating the algorithm's capacity to explore a diverse set of trade-offs between data plane overhead and signaling delay. This multi-objective optimization approach allows network operators to choose the solution from the Pareto front approximation that best aligns with their specific QoS requirements and service needs, enhancing the network's adaptability. Furthermore, it is notable that the vast majority of solutions on the MOEA's Pareto front approximation outperform LNA in both objectives, providing a clear advantage over the heuristic-based method across different topologies.

The inclusion of the auxiliary objective inspired by the Sharpe ratio, $S_{balance}$, has a notable impact on the load distribution of the vBS assignments in the network. Fig. 5 illustrates the progression of $S_{balance}$ throughout the evolutionary cycle, with measurements taken every 5000 evaluations up to the stopping criterion at 100,000 evaluations. As the MOEA progresses, the increasing size of the boxes in the boxplots reflects the expanding variability within the population. This variability stabilizes around 75,000 evaluations, indicating that the algorithm is effectively exploring trade-offs across the spectrum of objectives, including different degrees of load balancing.

By approximately 85,000 evaluations, the $S_{balance}$ metric shows signs of convergence, supporting our decission of 100,000 evaluations as the stopping condition for the MOEA. This convergence ensures that the algorithm has reached a stable balance between all objectives. The radar plot in Fig. 6 provides an overview of the average values of the three objectives over the evolution of the optimization process. In this plot, each objective is normalized to ensure that improvements bring the corresponding vertex closer to the edge circumference. For the objectives $D_{signaling}$ and O_{data}, both of which are minimized, a closer position to the outer edge represents a reduction in objective values. Conversely, for $S_{balance}$, which is to be maximized, proximity to the outer edge correspond to an increase in value. With this in mind, the plot shows that $D_{signaling}$ is the first objective to improve, followed by O_{data}, while $S_{balance}$ progresses more slowly and requires more evaluations to reach best levels of balance. This trend indicates that the MOEA initially focuses on reducing O_{data} and $D_{signaling}$, with load balancing gradually improving as the optimization progresses. The steady improvement in $S_{balance}$ highlights the algorithm's ability to refine the load distribution incrementally, ultimately achieving a balanced performance across all objectives.

4.2. Performance comparison between MOEA and LNA

With the optimization results established, we proceed to analyze the performance metrics that evaluate the efficiency of the proposed approaches. Using the results of the LNA from previous work as a baseline, we compare them against those achieved by the MOEA. This comparison enables us to assess the advantages of the multi-objective optimization approach, particularly in terms of signaling delay and data plane overhead across the different topologies described in 3.2. By examining these metrics, we aim to highlight the improvements in overall network performance by the MOEA over the heuristic-based LNA.

Figs. 7 and 8 present a comparative analysis of the data plane overhead (O_{data}) and signaling delay ($D_{signaling}$) metrics across each topology (low-connected, hybrid-connected, and high-connected). Each figure shows a violin plot, where the width of each distribution reflects

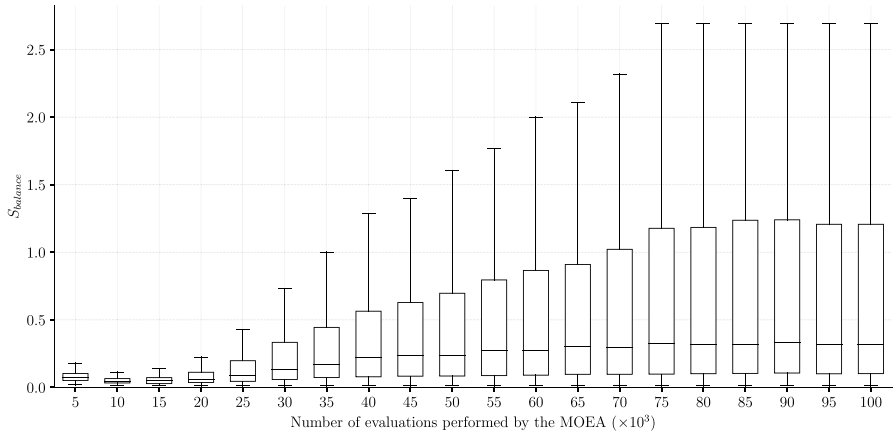


Fig. 5. Evolution of the metric inspired by the Sharpe ratio throughout the number of evaluations performed by the MOEA.

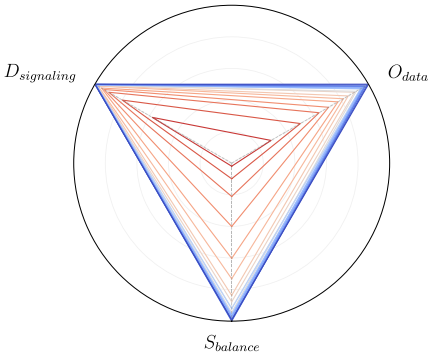


Fig. 6. Evolution of the three objectives ($D_{signaling}$, O_{data}, and $S_{balance}$) throughout the optimization process, from 5000 evaluations (red) to 100,000 evaluations (blue).

the variability of the metric values across different instances. The blue distribution represents the performance of the MOEA, based on the solutions with the highest $S_{balance}$ from the Pareto front approximation to ensure a fair comparison. In contrast, the orange distribution corresponds to the single solution generated by the heuristic-based LNA approach for each instance. The white line within each distribution indicates the median value, providing a central reference point, while the dotted gray lines represent the first and third quartiles, showing the range within which 50% of the data points lie. A narrower, lower distribution indicates a more consistent and efficient performance, as lower values correspond to reduced data plane overhead or signaling delay, respectively. By visualizing these distributions, we can assess not only the average performance but also the variability and robustness of each method across different network conditions.

In terms of O_{data} (Fig. 7), the LNA shows narrower distributions across all topologies, indicating more consistent performance with less variability. However, despite this consistency, the values achieved by LNA are consistently higher than those of the MOEA across all topologies. The MOEA, while exhibiting slightly wider distributions, consistently yields lower O_{data} values. Notably, the first quartile of the MOEA's distribution is below the third quartile of the LNA's distribution in every topology, underscoring a significant improvement. This improvement is particularly pronounced in the low-connected and hybrid-connected topologies, where the flexibility of the multi-objective approach in the MOEA allows it to explore a broader range of trade-offs, effectively minimizing data transmission costs compared to the heuristic approach of the LNA. The ability of the MOEA to generate a spectrum of solutions with lower overhead, despite slightly higher variability, demonstrates its adaptability to different network conditions, offering network operators the potential for reduced transmission costs in practical deployments.

For $D_{signaling}$ (Fig. 8), the MOEA exhibits narrower distributions in the low-connected and high-connected topologies, suggesting more consistent low-latency performance in these configurations. This consistency in $D_{signaling}$ is crucial for applications where maintaining low latency is essential for ensuring QoS. In the hybrid-connected topology, however, the LNA achieves a slightly narrower distribution, indicating more stable results in this specific setup. Nevertheless, the median and overall distribution of $D_{signaling}$ values for the MOEA are consistently lower than those for LNA across all topologies. Importantly, the first quartile of the MOEA's $D_{signaling}$ remains below the third quartile of the LNA's in every topology, highlighting a substantial reduction in latency. This reduction is particularly beneficial in highly connected networks, where the complex inter-router links can increase $D_{signaling}$ in less optimized configurations. The MOEA's superior performance in achieving lower $D_{signaling}$ across varied topologies demonstrates its effectiveness in handling handover events more efficiently, which is essential for dense and dynamic network environments.

Finally, in terms of percentage improvements, the MOEA achieves an average reduction in O_{data} of 34.1% in the low-connected topology, 21.5% in the hybrid-connected topology, and 21.6% in the high-connected topology. The best improvement was observed in the low-connected topology at 51.5%, while the smallest improvement was 0.99% in the hybrid-connected topology. For $D_{signaling}$, the MOEA's average improvement across topologies further emphasizes its effectiveness, with reductions of 53.9% in the low-connected topology, 48.9% in the hybrid-connected topology, and 48.5% in the high-connected topology. The most significant improvement was 77.9% in the low-connected topology, while the lowest was 3.8% in the high-connected topology. These substantial reductions in $D_{signaling}$ and O_{data} across different topologies demonstrate the MOEA's adaptability to varying network conditions, especially considering that in all instances of the three topologies it has superior performance.

5. Conclusions and future work

In this paper, we presented a parallelized multi-objective approach to address the complex challenge of vBS-to-access network allocation, aimed at enhancing resource deployment and mobility management in dense 5G and future 6G networks. By leveraging an asynchronous distributed steady-state multi-objective evolutionary algorithm, our approach simultaneously optimizes signaling delay and data plane overhead, while incorporating an auxiliary objective inspired by the Sharpe ratio to ensure load balancing across network nodes. This integration introduces a novel contribution to the field of communication networks, as the balance metric, adapted from the financial domain, provides an effective means to distribute resources more equitably, thereby enhancing network performance.

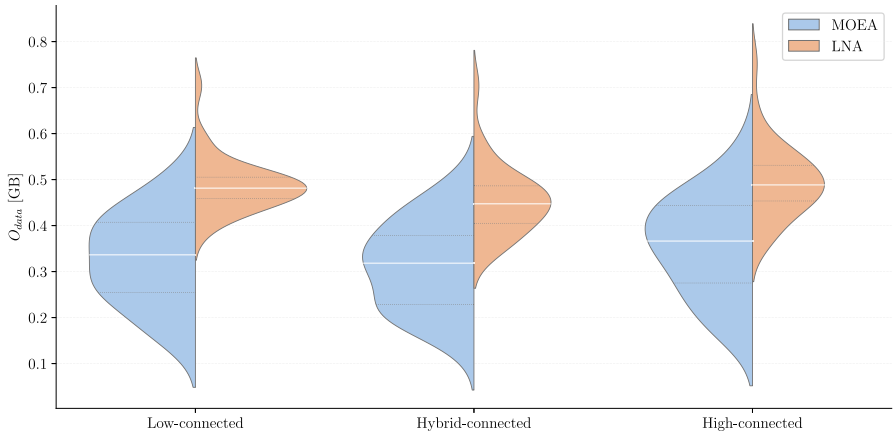


Fig. 7. Data plane overhead comparison between MOEA (blue) and LNA (orange) across different network topologies. White lines in the distributions indicates the median, and dotted gray lines mark the first and third quartiles.

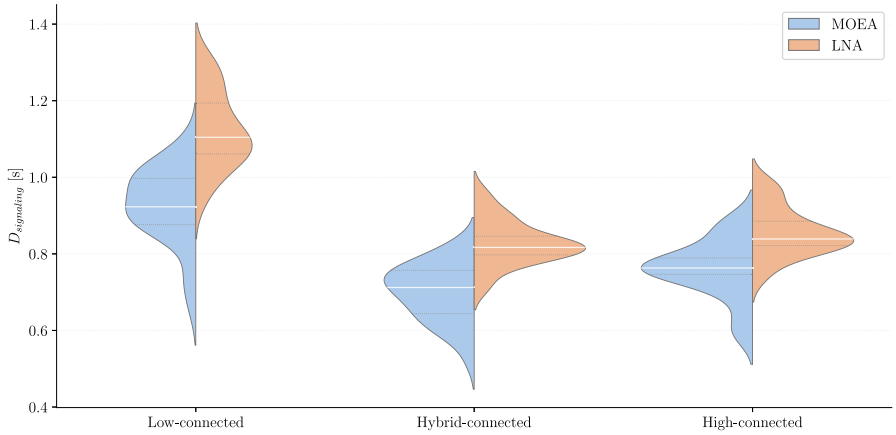


Fig. 8. Signaling delay comparison between MOEA (blue) and LNA (orange) across different network topologies. White lines in the distributions indicates the median, and dotted gray lines mark the first and third quartiles.

Our experimental results highlight the significant advantages of the MOEA over the domain-specific heuristic LNA, which has previously achieved high-quality solutions for this problem. The MOEA consistently attains lower values in both primary metrics, with improvements of up to 51.5% in data plane overhead and 77.9% in signaling delay, showcasing the effectiveness of a multi-objective optimization approach for managing mobility and resource deployment in next-generation networks. Additionally, evaluating MOEA solutions required computationally intensive, real-world network simulations. To address this, we implemented extensive parallel computation with up to 1000 computing nodes working simultaneously, which allowed us to efficiently manage the computational demands, reduce experimentation time, and ensure scalable, precise assessments of solution quality.

Although the MOEA has demonstrated superior performance, several promising directions remain for future work. One of them involves integrating additional performance metrics, such as energy consumption and service continuity, to contribute to more sustainable network designs. In this regard, optimizing cell switch-off decisions at the physical level of base stations could further enhance power efficiency without compromising network performance. Furthermore, we plan to incorporate additional communication performance metrics, such as latencies and throughput, to provide a more comprehensive evaluation of network efficiency under different traffic conditions. Additionally, extending the approach to support heterogeneous network infrastructures, including hybrid 5G/6G and Wi-Fi networks, could improve the applicability and robustness of the proposed solution, making it more adaptable to the diverse conditions of next-generation mobile networks.

Acronyms

Acronym	Description
LNA	Link-Network Assignment
MOEA	Multi-Objective Evolutionary Algorithm
NSGA-II	Non-dominated Sorting Genetic Algorithm II
vBS	Virtual Base Station
RRH	Remote Radio Head
BBU	Baseband Unit
MEC	Multi-access Edge Computing
QoS	Quality of Service
UPF	User Plane Function
AMF	Access Mobility Function
SMF	Session Management Function
3GPP	3rd Generation Partnership Project
6G	Sixth Generation Mobile Network
5G	Fifth Generation Mobile Network
IoE	Internet of Everything
PPP	Poisson Point Process
DMM	Distributed Mobility Management

CRediT authorship contribution statement

Jesús Calle-Cancho: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jesús Galeano-Brajones:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **David Cortés-Polo:** Writing – review & editing, Project administration, Methodology, Investigation. **Javier Carmona-Murillo:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Francisco Luna-Valero:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Science, Innovation and Universities, and the State Research Agency (MICIU/AEI/10.13039/501100011033) under grant PID2023-151462OB-I00, co-funded by the European Regional Development Fund (ERDF); by the European Union NextGenerationEU/PRTR (MICIU/AEI/10.13039/501100011033) under grant TED2021-131699B-I00; and by the 2025 Research and Transfer Plan of University of Extremadura under grant AV-05.

Data availability

The data is available in the manuscript through a GitHub link.

References

[1] Ericsson, Ericsson mobility report: Mobile traffic forecast, 2024, https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-traffic-forecast. (Accessed 31 October 2024).
[2] M.E. Haque, F. Tariq, M.R. Khandaker, K.-K. Wong, Y. Zhang, A survey of scheduling in 5G URLLC and outlook for emerging 6G systems, IEEE Access 11 (2023) 34372–34396.
[3] 3GPP-RP-213468, Summary for RAN Rel-18 package, 2021, (Accessed 31 October 2024).
[4] X. Lin, An overview of 5G advanced evolution in 3GPP release 18, IEEE Commun. Stand. Mag. 6 (3) (2022) 77–83.
[5] M.S. Akbar, Z. Hussain, M. Ikram, Q.Z. Sheng, S. Mukhopadhyay, On challenges of sixth-generation (6G) wireless networks: A comprehensive survey of requirements, applications, and security issues, J. Netw. Comput. Appl. (2024) 104040.
[6] Ö. Bulakçı, X. Li, M. Gramaglia, A. Gavras, M. Uusitalo, P. Rugeland, M. Boldi, Towards Sustainable and Trustworthy 6G: Challenges, Enablers, and Architectural Design, Boston-Delft, 2023.
[7] M. Khaturia, N. Sharma, J. Choi, A. Nigam, D. Kim, Service-based architecture evolution: Towards enhanced signaling in beyond 5G/6G networks, in: 2024 IEEE Wireless Communications and Networking Conference, WCNC, IEEE, 2024, pp. 1–6.
[8] I. Leyva-Pupo, C. Cervelló-Pastor, An intelligent scheduling for 5G user plane function placement and chaining reconfiguration, Comput. Netw. 237 (2023) 110037.
[9] J. Martín-Pérez, L. Cominardi, C.J. Bernardos, A. de la Oliva, A. Azcorra, Modeling mobile edge computing deployments for low latency multimedia services, IEEE Trans. Broadcast. 65 (2) (2019) 464–474.
[10] J. Calle-Cancho, J. Carmona-Murillo, J.-L. González-Sánchez, D. Cortés-Polo, A novel link-network assignment to improve the performance of mobility management protocols in future mobile networks, Wirel. Commun. Mob. Comput. 2022 (1) (2022) 7061588.
[11] M. Kassi, S. Hamouda, RAN virtualization: How hard is it to fully achieve? IEEE Access 12 (2024) 38030–38047.
[12] A. Vlahov, J. Smith, M. Johnson, T. Brown, R. Garcia, Virtualized, Open and Intelligent: The Evolution of the Radio Access Network, Taylor & Francis, 2023.
[13] K.H. Rahi, H.K. Singh, T. Ray, A steady-state algorithm for solving expensive multiobjective optimization problems with nonparallelizable evaluations, IEEE Trans. Evol. Comput. 27 (5) (2022) 1544–1558.
[14] Y.-J. Gong, W.-N. Chen, Z.-H. Zhan, J. Zhang, Y. Li, Q. Zhang, J.-J. Li, Distributed evolutionary algorithms and their models: A survey of the state-of-the-art, Appl. Soft Comput. 34 (2015) 286–300.
[15] C.A. Coello Coello, Evolutionary multi-objective optimization: A historical view of the field, IEEE Comput. Intell. Mag. 1 (1) (2006) 28–36.
[16] W.F. Sharpe, The sharpe ratio, J. Portf. Manag. 21 (1) (1994) 49–58.
[17] Z. Zhou, Z. Song, T. Ren, L. Yu, Two-stage portfolio optimization integrating optimal sharp ratio measure and ensemble learning, IEEE Access 11 (2022) 1654–1670.
[18] A.P. Guerreiro, C.M. Fonseca, An analysis of the hypervolume sharpe-ratio indicator, European J. Oper. Res. 283 (2) (2020) 614–629.
[19] M. Di Renzo, A. Zappone, T.T. Lam, M. Debbah, System-level modeling and optimization of the energy efficiency in cellular networks—A stochastic geometry framework, IEEE Trans. Wirel. Commun. 17 (4) (2018) 2539–2556.
[20] R. Wang, M.A. Kishk, M.-S. Alouini, Resident population density-inspired deployment of K-Tier aerial cellular network, IEEE Trans. Wirel. Commun. 22 (11) (2023) 7989–8002.
[21] K.M.S. Huq, J. Rodriguez, I.E. Otung, 3D network modeling for THz-enabled ultra-fast dense networks: A 6G perspective, IEEE Commun. Stand. Mag. 5 (2) (2021) 84–90.
[22] E. Gures, I. Shayea, M. Sheikh, M. Ergen, A.A. El-Saleh, Adaptive cell selection algorithm for balancing cell loads in 5G heterogeneous networks, Alex. Eng. J. 72 (2023) 621–634.
[23] M. Murtadha, N. Noordin, B. Ali, F. Hashim, Design and evaluation of distributed and dynamic mobility management approach based on PMIPv6 and MIH protocols, Wirel. Netw. (2015) 1–17.
[24] G. Zheng, C. Wang, V. Friderikos, M. Dohler, High mobility multi modal E-health services, in: IEEE International Conference on Communications, ICC, 2018, pp. 1–7.
[25] J. Carmona-Murillo, V. Friderikos, J. González-Sánchez, A hybrid DMM solution and trade-off analysis for future wireless networks, Comput. Netw. 133 (2018) 17–32.
[26] E.M.O. Fafolahan, S. Pierre, A seamless mobility management protocol in 5G locator identificator split dense small cells, IEEE Trans. Mob. Comput. 19 (8) (2020) 1745–1759.
[27] A. Abrar, A.S. Che Mohamed Arif, K. Mohd Zaini, M.H. Omar, Y. Meng, Advancing producer mobility management in named data networking: A comprehensive analytical model, J. King Saud Univ. - Comput. Inf. Sci. 36 (4) (2024) 102045.
[28] F. Giust, C.J. Bernardos, A.D.L. Oliva, Analytic evaluation and experimental validation of a network-based IPv6 distributed mobility management solution, IEEE Trans. Mob. Comput. 13 (11) (2014) 2484–2497.
[29] H. Ali-Ahmad, M. Ouzzif, P. Bertin, X. Lagrange, Performance analysis on network-based distributed mobility management, Wirel. Pers. Commun. 74 (4) (2014) 1245–1263.
[30] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.
[31] K. Deb, Multi-objective optimisation using evolutionary algorithms: an introduction, in: Multi-Objective Evolutionary Optimisation for Product Design and Manufacturing, Springer, 2011, pp. 3–34.
[32] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P.N. Suganthan, Q. Zhang, Multiobjective evolutionary algorithms: A survey of the state of the art, Swarm Evol. Comput. 1 (1) (2011) 32–49.
[33] J. Knowles, A summary-attainment-surface plotting method for visualizing the performance of stochastic multiobjective optimizers, in: 5th International Conference on Intelligent Systems Design and Applications, ISDA'05, IEEE, 2005, pp. 552–557.