SEMARAK ILMU PUBLISHING

Semarak Engineering Journal

Journal homepage:
https://semarakilmu.my/index.php/sej/index
ISSN: 3036-0145



Traffic Scenario Evaluation of Joint MEC Server Placement and Routing in Urban and Rural Networks

Sandrine Khaled[1,*], Mohamed Shehata[1]

[1] Department of Electronics and Communications, Faculty of Engineering and Technology, Arab Academy for Science, Technology and Maritime Transport, Cairo 11799, Egypt

ARTICLE INFO

Article history:
Received 6 July 2026
Received in revised form 13 September 2026
Accepted 10 October 2026
Available online 8 November 2026

Keywords:
Mobile Edge Computing, ILP, Server Placement, Routing, Latency

ABSTRACT

Multiaccess Edge Computing (MEC) reduces the latency of modern applications by deploying edge servers close to the users, at the base stations (BSs). However, the placement of a limited number of MEC servers is a challenging problem that strongly depends on the traffic demand, which varies over the time of the day and across the deployment environment. In this paper, we formulate the joint MEC server placement and traffic routing problem as an Integer Linear Programming (ILP) model that minimizes a weighted combination of the number of activated servers and the traffic-weighted average delay, subject to server-capacity, link-capacity, and end-to-end-latency constraints. The model is evaluated on a real dataset of 554,383 sessions from the Shanghai Telecom network, using a dense urban and a sparse rural sub-network of 50 BSs each. The results reveal a constraint regime shift between the two environments: the urban placement is dominated by the server capacity, whereas the rural placement is dominated by the end-to-end latency, which leads to rural delays that are 4.7–5.8× higher than the urban ones. A cross-scenario robustness analysis shows that a static single-scenario placement either overflows the capacity at the peak (up to 248% server utilization) or wastes resources during the off-peak periods. To address this, an aggregated deployment strategy is proposed, in which a single placement is dimensioned on an aggregated demand; it remains feasible across all the traffic periods while avoiding the overflow and reducing the waste of the single-scenario designs.

1. Introduction

Multiaccess networks are experiencing sustained and rapid growth in both traffic volume and service diversity. Global mobile network data traffic (including Fixed Wireless Access (FWA)) increased by 22% between the first quarters of 2025 and 2026 [1]. Excluding FWA, global mobile data traffic reached approximately 146 EB per month by the end of 2025 and is projected to exceed 328 EB per month by 2031 [1]. This growth is accompanied by the increasing adoption of computation-intensive and delay-sensitive applications, including immersive XR and cloud gaming, connected and autonomous vehicles, industrial automation and digital twins, and AI-enabled IoT services [2-5].

* *Corresponding author.*
E-mail address: sandrinekhaled@gmail.com

https://doi.org/10.37934/sej.17.1.133149

These applications need response times of only a few milliseconds and require heavy local processing. In 5G, such stringent communication requirements are addressed through Ultra-Reliable Low-Latency Communication (URLLC), for which the minimum user-plane latency capability is 1 ms with very high reliability [6]. Such requirements can be difficult to meet by distant, centralized data centers, where the physical distance to the user adds tens to hundreds of milliseconds of delay and places a heavy load on backhaul and core-network links. This mismatch between the needs of modern applications and the limits of centralized cloud computing motivates moving computation and storage closer to the users, at the edge of the network.

Mobile Edge Computing (MEC) has been introduced as a key paradigm to meet the stringent latency requirements of these applications [7]. The main idea of MEC is to deploy edge servers within the access network, such as at base stations or aggregation nodes. That allows users to offload their computational tasks to nearby servers instead of sending them entirely to distant cloud data centers. Thanks to this proximity, the user tasks can be processed within one or a few hops. Hence, the end-to-end latency is significantly reduced [8]. In addition, processing the traffic locally relieves the backhaul and the core network and mitigates network congestion [9]. Processing data close to its source can also help preserve user privacy by reducing the amount of raw or sensitive data transmitted to the cloud [10]. These advantages enable MEC to support real-time and location-aware applications while improving service responsiveness, network efficiency, and users' Quality of Experience (QoE). For these reasons MEC is regarded as a key enabler for 5G and the upcoming 6G networks [11].

Despite these advantages, the deployment of MEC servers introduces several challenges. Unlike centralized cloud data centres, edge servers have limited computing and storage capacities [12]. Deploying a server at every Base Station (BS) can provide low latency, but it requires high capital and operational costs. On the other hand, deploying only a few servers reduces the deployment cost, but it may increase the traffic carried by the backhaul, overload the selected servers, and violate the latency requirements of the users [13]. Without appropriate workload balancing, some servers may become overloaded while others remain underutilized [14]. This raises a central question: where should a limited number of MEC servers be placed within the access network? The MEC server placement problem does not only concern the selection of the server locations, but also the assignment of each base station to an active server and the routing of its traffic through the network, while respecting the server capacity, the link capacity, and the end-to-end latency constraints. The problem is further complicated by two main factors. First, the traffic demand changes over time. Therefore, a placement that is designed for the peak demand wastes resources during the off-peak periods. Whereas a placement that is designed for the off-peak demand becomes infeasible at the peak [15]. Second, although it is recognized that the deployment context affects the placement [16], the way in which the deployment density determines the binding constraint has received little attention. Our results show that, under the evaluated configurations, server capacity is the dominant constraint in the urban network, whereas end-to-end latency is the dominant constraint in the rural network. These challenges motivate the need for joint optimization of MEC server placement and traffic routing under spatially heterogeneous and time-varying traffic demand.

To address these challenges, in this paper, we formulate the joint MEC server placement and traffic routing problem as an Integer Linear Programming (ILP) model. The model jointly optimizes the server location, the base station assignment, and the traffic routing, while satisfying the server capacity, the link capacity, and the end-to-end latency constraints. The objective is to minimize a weighted combination of the number of servers and the traffic-weighted propagation delay. The main contributions of this paper are summarized as follows: i) we formulate an ILP model that jointly determines the MEC server locations, the base station assignment, and the traffic routing within a

single optimization framework; ii) we study the effect of the temporal demand variation by considering peak, off-peak, and aggregated demand scenarios, and we quantify the overprovisioning waste and the capacity overflow that result from applying a static single-scenario placement across different time periods; iii) we conduct a comparative study between a dense urban and a sparse rural deployment, and we reveal a constraint regime shift, in which the placement is dominated by the server capacity in the urban environment and by the end-to-end latency in the rural environment; and iv) we validate the model on a real-world dataset from the Shanghai Telecom network.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the network architecture and the system model. Section 4 formulates the ILP model. Section 5 describes the urban and rural sub-network selection and experimental setup, and presents and discusses the results, including the demand scenario comparison, the alpha sweep, and the cross-scenario robustness analysis. Section 6 concludes the paper and outlines directions for future work.

2. Related Work

Several studies have investigated MEC server placement using optimization, clustering, heuristic, and learning-based approaches. Zhai *et al*., [7] formulated joint access-edge selection and service-data placement as an MINLP problem and proposed a heuristic reassignment algorithm to obtain close-to-optimal solutions. Shibata and Miyata [9] used a genetic algorithm to jointly determine edge server placement and task allocation with the objective of reducing the maximum delay. Dash *et al*., [12] proposed a clustering-based method for joint RAN association and MEC server placement to reduce deployment cost while satisfying application-delay requirements. However, these studies mainly focus on placement and assignment decisions, while routing through a capacity-constrained multi-hop backhaul network is simplified or not explicitly considered.

Other studies have jointly considered placement and routing. Li *et al*., [17] formulated the joint placement of service instances on existing edge clouds and the routing of user requests as an ILP problem under storage, computation, and communication constraints. Since the problem is NP-hard, they proposed an LP-relaxation-based randomized-rounding algorithm to obtain close-to-optimal integer solutions. However, their work does not determine the physical locations or activation of MEC servers. Forghieri *et al*., [13] formulated latency-constrained service placement and routing as an ILP problem that combines multi-commodity flows with linear latency constraints. To address the scalability limitations of the exact formulation, they proposed a network structure based heuristic and a PPO-based deep reinforcement learning method for large and dynamic instances. However, their work considers service placement on existing edge servers rather than the physical deployment of MEC servers. Meanwhile, Wang *et al*., [15] demonstrated that cellular traffic varies considerably across locations and time periods. To address such variations, Guo and Chen [8] and Li *et al*., [14] developed learning-based placement approaches that adapt to user mobility and changing workloads. The deployment environment also affects the placement problem. Anzola-Rojas *et al*., [16] studied joint fiber and MEC deployment in sparsely populated areas and explained that capacity is generally a major concern in dense environments, whereas connectivity and delay become more important in sparse environments. However, their study focuses on infrastructure deployment cost and does not consider temporal traffic mismatch.

These studies mainly develop adaptive placement policies for changing network conditions. However, they do not explicitly quantify the cross-period feasibility and resource utilization of a fixed physical MEC server placement by applying a placement optimized for one traffic period under another period. This evaluation is important because MEC server locations represent long-term infrastructure decisions and cannot be changed whenever traffic demand varies. A placement

designed only for off-peak demand may become overloaded during peak periods, whereas a peak-demand placement may result in substantial resource underutilization during off-peak periods. Therefore, cross-period evaluation is necessary to determine whether a placement can remain feasible and efficiently utilized under daily traffic variations.

In this work, we jointly optimize MEC server placement, BS assignment, and multi-hop traffic routing under server-capacity, link-capacity, and end-to-end latency constraints. Placements obtained for morning, night, and aggregated demand are cross-evaluated to quantify resource overprovisioning and server-capacity violations. Moreover, the same formulation is applied to dense urban and sparse rural configurations to compare their placement patterns and binding constraints.

3. System Model

3.1 Network Model

We consider a cellular backhaul network composed of a set of interconnected Base Stations (BSs). The network is represented as a directed graph $G = (B, A)$, where B is the set of BSs and A is the set of directed backhaul arcs. Each physical bidirectional link between two BSs u and v is represented by two directed arcs, $a = (u, v)$ and $a^{\setminus} = (u, v)$. Each arc $a \in A$ is characterized by its geographical length d_a and maximum traffic capacity C_a. Considering all the links are optical fiber, then propagation latency l_a of each arc a is the length of the arc in km divided by the speed of light. Every BS is considered a candidate location for deploying a MEC server. Therefore, the set of candidate server locations is $K = B$. The traffic generated by a BS can be processed locally when a MEC server is active at that BS. Otherwise, the traffic is routed through one or more backhaul arcs to its assigned MEC server. The network topology is assumed to be connected and remains fixed during each optimization run. The same graph representation is used for the urban and rural configurations, while the BS locations, link connectivity, and inter-BS distances differ between the two configurations.

3.2 Demand and Traffic Model

The workload d_i of each BS i is the mean number of concurrent sessions (sess/hr) during a defined time interval. Two demand scenarios are derived from the Shanghai Telecom dataset [17]-[20] (that will be discussed later in section 4) (i) Morning peak (08:00–10:00): highest demand, worst-case placement stress. (ii) Night trough (02:00–05:00): lowest demand, minimum infrastructure scenario. Table 1 reports the resulting d_i values for a representative sample of the selected urban sub-network BSs, illustrating the scale of this variation.

Table 1
Urban sub-network base station characteristics (10 of 50 BSs shown)

BS ID	Latitude	Longitude	Users	d_i morning	d_i night
16	30.718108	121.283169	13	3.384	0.042
40	30.746286	121.344028	39	10.990	10.594
84	30.824298	121.363989	38	11.668	2.544
142	30.876032	121.466173	15	9.654	2.351
992	31.182415	121.381272	54	12.986	8.158
1179	31.201965	121.452457	87	22.932	1.416
1244	31.209025	121.616514	65	28.888	8.033
1325	31.214683	121.551574	64	19.930	15.578
1859	31.247080	121.419525	76	24.240	2.921
1860	31.247155	121.231120	54	25.899	11.005

Table 1 shows a raw dataset statistic included for context only and is not used as an ILP input. The morning demand d_i (concurrent sessions) represents the mean number of simultaneously active sessions at BS i during the morning peak window (08:00–10:00), averaged across all days in the dataset; it is computed by counting how many recorded sessions overlap each clock-minute within the window and averaging the result. The night demand d_i (concurrent sessions) is defined identically but computed over the off-peak night window (02:00–05:00). Both d_i morning and d_i night serve directly as the ILP demand inputs and capture the significant temporal heterogeneity in cellular traffic that motivates scenario-specific MEC deployment.

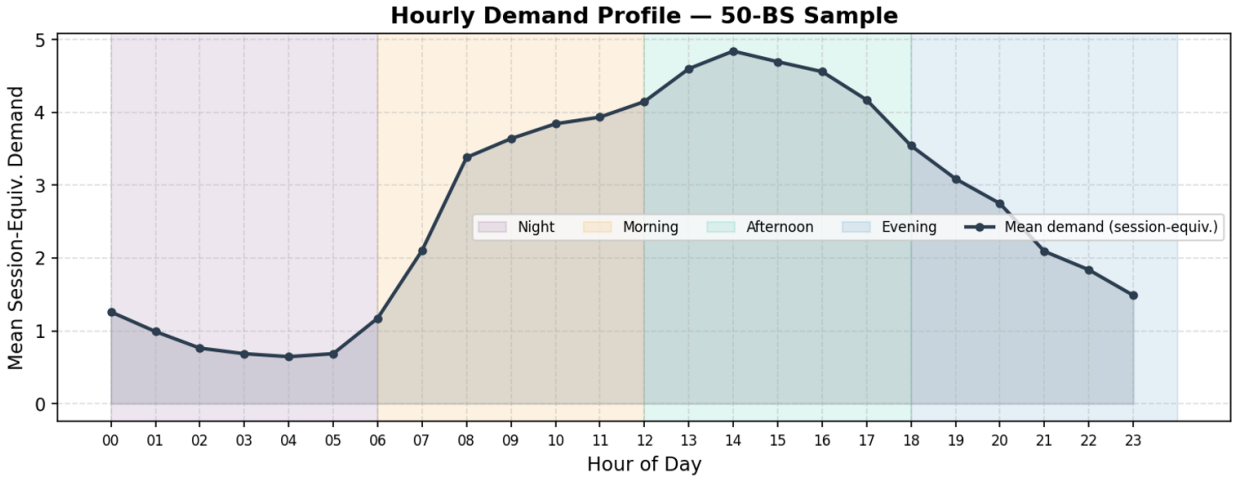
To define the ILP demand inputs, the raw Shanghai Telecom session records are aggregated into two representative scenarios based on the time-of-day traffic patterns observed in the dataset. Table 2 reports the total network demand $\Sigma_i d_i$ for each scenario across both sub-networks. Morning demand is the highest in both environments — 150.58 concurrent sessions (urban) and 220.89 (rural); while night demand drops to 72.62 and 120.86 respectively, a reduction of more than 45% in both cases. Rural demand consistently exceeds urban by approximately 47%, reflecting the larger geographic coverage and higher accumulated traffic per BS in the outer Shanghai ring.

Table 2
Demand scenarios derived from Shanghai Telecom dataset

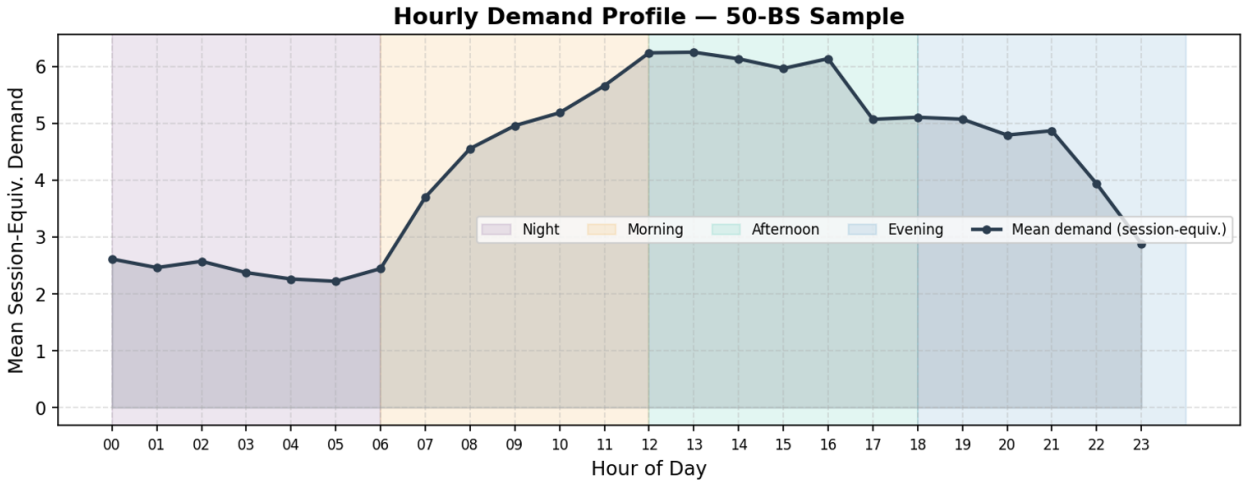
Scenario	Time Window	Urban Total Demand	Rural Total Demand
Morning	08:00–10:00	150.58	220.89
Night	02:00–05:00	72.62	120.86

The temporal variation underlying these three scenarios is visualised in Fig. 1, which plots the mean hourly demand per BS across all 24 hours of the day for both sub-networks.

Fig. 1 confirms the demand structure captured in Table 2. Both urban and rural profiles share the same temporal shape — a quiet night trough (00:00–05:59), a sharp morning ramp from 06:00, a broad afternoon plateau peaking around 13:00–14:00, and a gradual evening decline after 18:00. The rural profile (Fig. 1b) sits consistently above urban (Fig. 1a) at every hour of the day, with the gap widest at morning peak. This persistent demand heterogeneity (both across time and across environments) directly motivates the scenario-specific ILP approach: a single static placement cannot simultaneously optimise for the low, concentrated night demand and the high, spatially distributed morning demand without either over-provisioning resources or risking capacity violations.



(a) Urban hourly demand



(b) Rural hourly demand

Fig. 1. Hourly total demand across 50 BSs

4. ILP Problem Formulation

The MEC placement and routing problem is formulated as an ILP problem as follows:

- **Given**: network topology, data traffic, Maximum end-to-end latency threshold, maximum MEC server capacity, Maximum link capacity
- **Output:** MEC server locations such that each base station is connected to exactly one MEC server, routing of the traffic between the base station and MEC server
- **Objective:** minimize number of MEC servers, minimize workload weighted delay

4.1 Input Sets, Parameters and Decision Variables

4.1.1 Input Sets and Parameters

Table 3 summarizes the input sets and parameters used to formulate the optimization model.

Table 3
Input sets and parameters

Parameter	Description
V	Set of nodes (BSs and aggregation points)
B	Set of base stations in the network $i, k \in B, B \subseteq V$
K	Set of candidates MEC server locations $K = B$
A	Set of directed arcs (backhaul links).
d_i	Workload demand of BS i (sess/hr)
d_a	Length of arc a (km)
C_{server}	Maximum MEC server capacity
C_a	Arc capacity (sess/hr)
l_a	Propagation latency for arc a n ms
l_{max}	Maximum end to end latency (ms)
l_{norm}	Normalization constant for delay objective
β, α	Weights of the objective function ($\beta = \alpha - 1$)

4.1.2 Decision Variables

Table 4 defines the decision variables used.

Table 4
Decision Variables

Parameter	Value	Description
y_k	1	if node k hosts a MEC server
$x_{i,k}$	1	if cell site i is assigned to a MEC server at node k (binary)
$r_{i,k,a}$	1	if arc a is used to route the traffic between BS at node i and MEC server at node k (binary)

4.2 Objective Function

The bi-objective function minimises a weighted sum of normalised server count and workload-weighted delay:

$$min \quad \alpha \cdot \frac{\sum_{k \in K} y_k}{|K|} + \beta \cdot \frac{\sum_{i \in B} \sum_{k \in K} \sum_{a \in A} d_i \cdot l_a \cdot r_{i,k,a}}{l_{norm}} \tag{1}$$

where β = 1 − α and l_{norm} normalises the delay term to [0,1]. α = 1 optimises cost only; α = 0 optimises delay only; α = 0.5 (canonical) gives equal weight.

4.3 Constraints

C1 — Unique Assignment: Every BS is assigned to exactly one active MEC server.

$$\sum_{k \in K} x_{i,k} = 1 \qquad \forall i \in B \tag{2}$$

C2 — Activation Coupling: BS i can only be assigned to server k if server k is active.

$$x_{i,k} \leq y_k \qquad \forall i \in B, k \in K \tag{3}$$

C3 — Server Capacity: Total demand assigned to server k cannot exceed its capacity. Dominant binding constraint in urban settings.

$$\sum_{i \in B} d_i \cdot x_{i,k} \leq C_{server} \cdot y_k \qquad \forall k \in K \tag{4}$$

C4 — Flow Conservation (Routing Validity): For every BS–server pair (i,k) and every node v ∈ V, the routing variables must form a valid directed path from BS i to server k. Let $\delta^+(v)$ and $\delta^-(v)$ denote the sets of arcs leaving and entering node v

$$\sum_{a \in \delta^+(v)} r_{i,k,a} - \sum_{a \in \delta^-(v)} r_{i,k,a} = \begin{cases} x_{i,k} & \text{if } v = i \\ -x_{i,k} & \text{if } v = k \\ 0 & \text{otherwise} \end{cases} \qquad \forall i \in B, k \in K, v \in V \tag{5}$$

C5 — Link Capacity: Total traffic on each arc cannot exceed its capacity.

$$\sum_{i \in B} \sum_{k \in K} d_i \cdot r_{i,k,a} \leq C_a \qquad \forall a \in A \tag{6}$$

C6 — Latency Threshold: End-to-end path latency from BS i to server k must not exceed $L_{\max}$.

$$\sum_{a \in A} l_a \cdot r_{i,k,a} \leq L_{\max} \cdot x_{i,k} \qquad \forall i \in B, k \in K \tag{7}$$

C7 — Binary constraint: All decision variables are binary.

$$y_k, x_{i,k}, r_{i,k,a} \in \{0,1\} \tag{8}$$

5. Results and Discussion

In this section, we present a structured evaluation of the proposed ILP formulation from three different angles. First, we evaluate the performance of the ILP across the urban and the rural sub-networks in terms of the number of activated servers and the end-to-end delay. Second, we study the sensitivity of the bi-objective trade-off to the weighting parameter α. Third, we conduct a cross-scenario robustness analysis, which concludes with the introduction of an aggregated deployment strategy as a physically realizable single-design solution.

The study uses real cellular session data from the Shanghai Telecom network on 16 June 2015. After preprocessing, 554,383 sessions are retained across 2,769 BSs. Each session record contains: BS identifier, session start time, duration, and data volume. The demand d_i is computed mean number of concurrent sessions (sess/hr) during the target time interval. All ILP instances were solved using Gurobi Optimizer 11.0 on a laptop equipped with an Intel Core i7-8550U CPU at 1.80 GHz and 8 GB of RAM (2133 MHz) running a 64-bit Windows operating system. The simulation parameters are listed in Table 5 as follows

Table 5
Dataset and experimental parameters

Parameter	Value
Dataset	Shanghai Telecom, 16 June 2015
Total sessions	554,383
Total BSs	2,769
Urban sub-network	50 BSs, inner Shanghai (< 8.03 km from centroid)
Rural sub-network	50 BSs, outer Shanghai ring (> 20.39 km from centroid)
l_{max}	1.4 ms
C_{server}	20 Sess/hr

5.1 Urban and Rural for different demand scenarios

Figure 2(a) shows the number of activated servers for the morning and the night scenarios in both the urban and the rural environments. In both cases, the morning (peak) demand requires more servers than the night (off-peak) demand; however, this reduction is more pronounced in the urban environment. In the urban area, the number of servers decreases by 28% (from 18 servers in the morning to 13 servers in the night), whereas in the rural area it decreases by only 16% (from 19 to 16 servers). This behavior can be explained by the constraint that dominates the placement in each environment. In the urban area, the server capacity and the link capacity are the dominant constraints, and since both depend on the traffic demand, the placement follows the demand, and

fewer servers are needed as the demand decreases at night. While in the rural area, the end-to-end latency is the dominant constraint and the propagation delay depends on the arc length, which is constant over time. So, the number of servers changes only slightly between the morning and the night. Moreover, the rural area consistently requires more servers than the urban area, although both environments contain the same number of base stations, because the geographic sparsity of the rural area forces the ILP to activate a server near each isolated cluster of base stations regardless of the load level, which prevents the demand-driven consolidation that is achievable in the dense urban topology.

Figure 2(b) shows the traffic-weighted average path latency for the morning and the night scenarios in both environments. The rural delays are 4.7–5.8 times higher than the urban delays across both scenarios. This is a direct consequence of the longer backhaul distances between the sparsely distributed base stations in the rural area, since the propagation delay increases with the arc length. In addition, in both environments the night scenario achieves a lower delay than the morning scenario. This is because the lower demand at night relaxes the capacity constraints and gives the ILP more freedom to minimize the delay term of the objective; hence, the high-demand base stations can be assigned to their nearest servers through shorter and less congested paths. At the morning peak, on the other hand, the capacity constraints become tighter and force some base stations to be assigned to farther servers in order to balance the load, which increases the average delay.

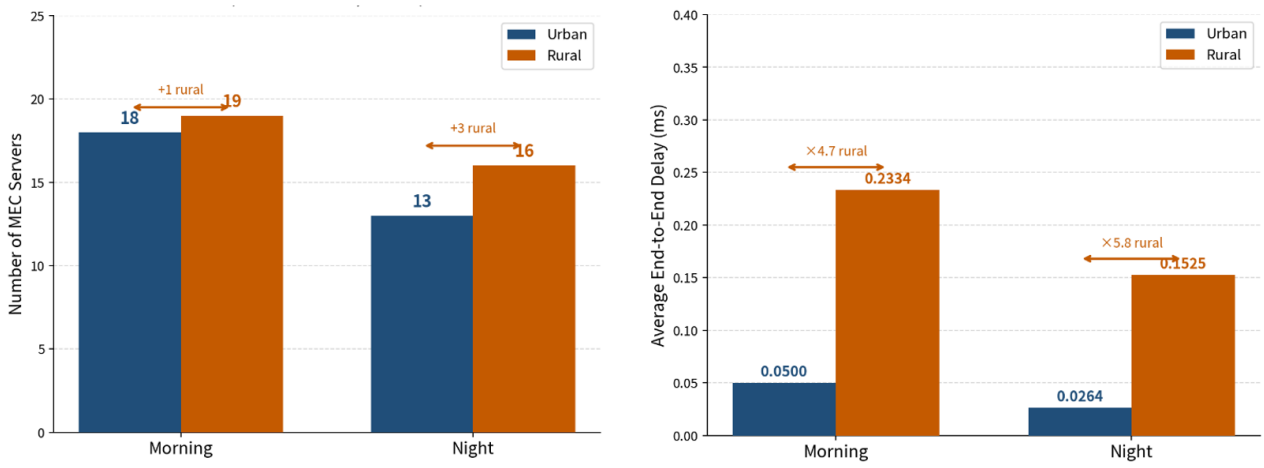
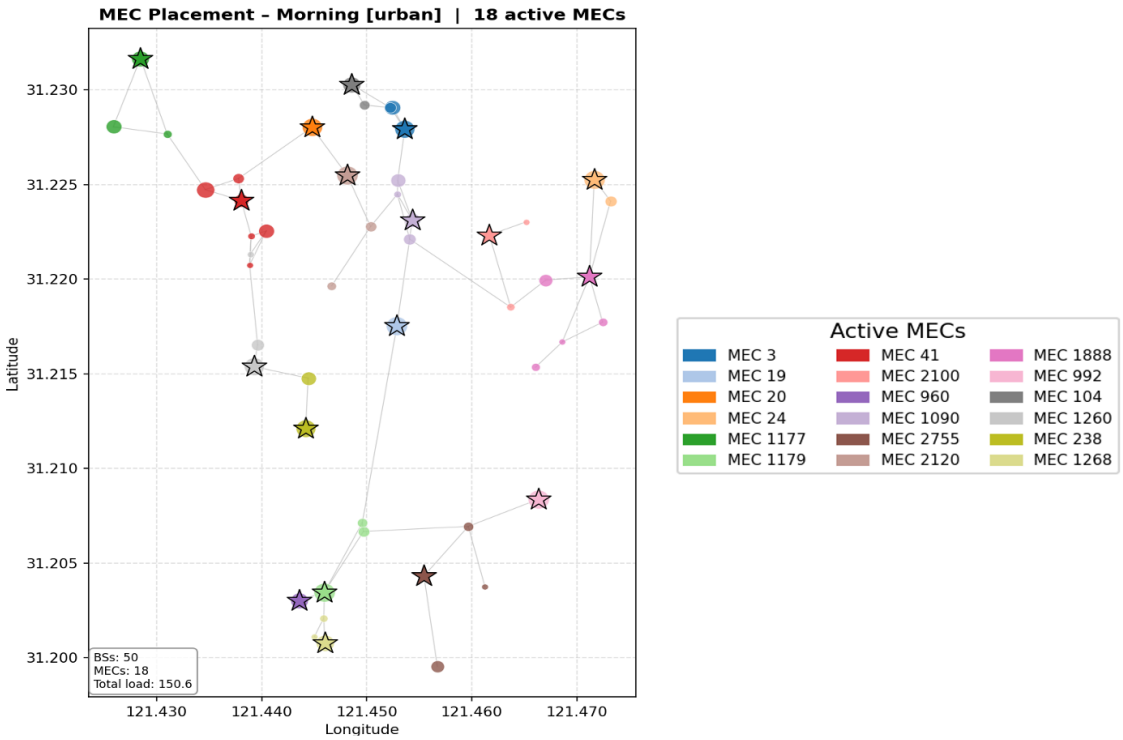
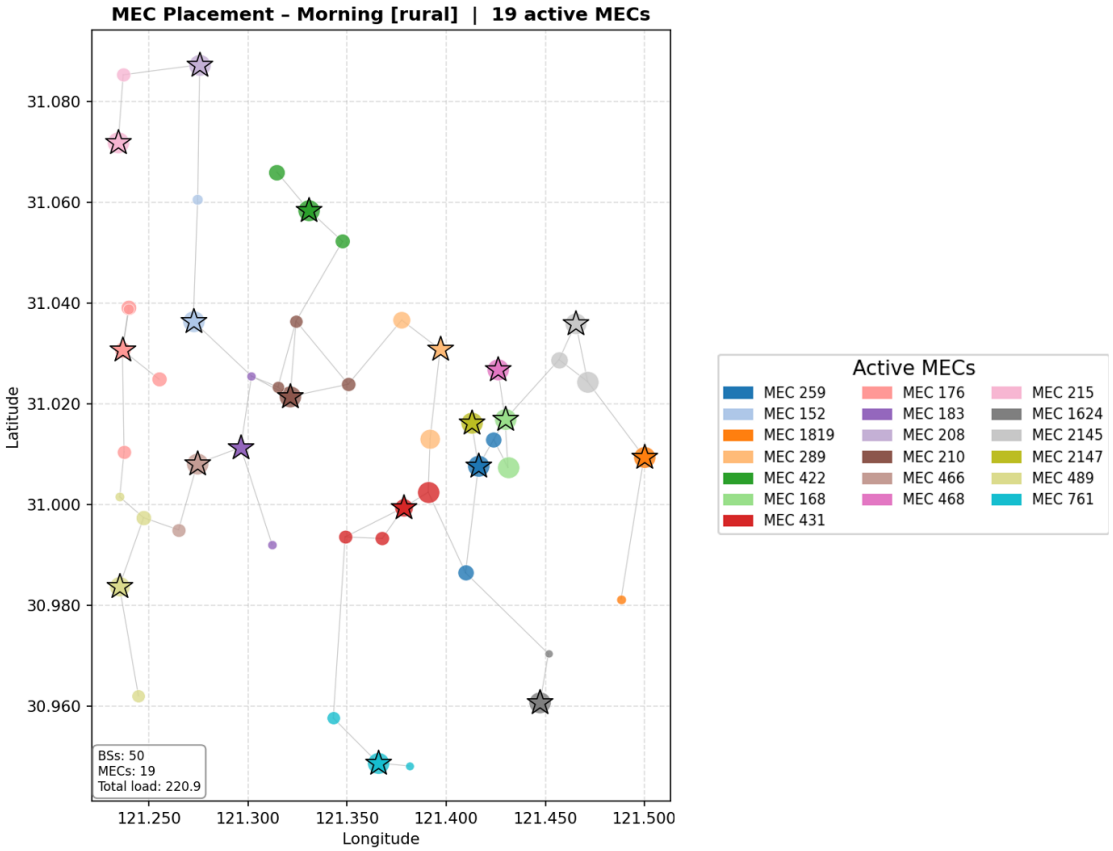


Fig. 2. ILP deployment performance for urban and rural sub-networks at morning and night scenarios (α = 0.5) (a) Number of MEC servers (Urban vs Rural, α= 0.5), (b) Traffic-weighted average path latency (Urban vs Rural, α= 0.5)

Figure 3 shows the optimal ILP placement for the morning scenario in both environments. In the urban sub-network (Fig. 3(a)), the 18 active servers are distributed across the dense and compact topology, where each server covers, on average, two to three neighboring base stations owing to the high spatial density. As a result, the routing paths are short and the resulting delay is low. In the rural sub-network (Fig. 3(b)), on the other hand, the placement is structurally different. Although it requires only one additional server (19 versus 18), the active servers are geographically far more dispersed and are spaced several kilometers apart. This is because, in the rural environment, the servers must be placed close to the isolated clusters of base stations in order to maintain the latency feasibility, rather than to consolidate the load as in the dense urban topology.



(a) Urban (Morning) 18 servers



(a) Urban (Morning) 18 servers

Fig. 3. Optimal ILP server placement — urban and rural morning

5.2 Alpha Sweep

Figure 4 illustrates the effect of the weighting parameter α on both the number of deployed MEC servers and the traffic-weighted average path latency for the four scenario–environment combinations. As α increases, the objective assigns greater weight to the minimization of the number of servers; hence, the ILP consolidates the base stations onto fewer servers in order to reduce the infrastructure cost. Consequently, the number of deployed servers decreases monotonically in all four panels of Figure 4. However, activating fewer servers means that each active server has to serve more base stations, including geographically distant ones, and therefore the average routing delay increases as α rises. This inverse relationship between the number of servers and the delay represents the fundamental cost–performance trade-off of the bi-objective formulation.

This trade-off is most pronounced in the urban night scenario (Fig. 4(b)), where the number of servers drops sharply from 20 at α = 0 to 9 at α = 1 (a 55% reduction), while the delay rises from nearly zero to 0.128 ms. This is because the low night demand gives the ILP the maximum freedom to consolidate the base stations. In the rural morning scenario (Fig. 4(c)), on the other hand, the reduction is more gradual (from 20 servers at α = 0 to 16 servers at α = 1, only 20%), since the geographic sparsity limits how aggressively the base stations can be consolidated, regardless of the value of α. Finally, the canonical operating point α = 0.5, which assigns an equal weight to both objectives, is marked by the vertical dotted line in Fig. 4.

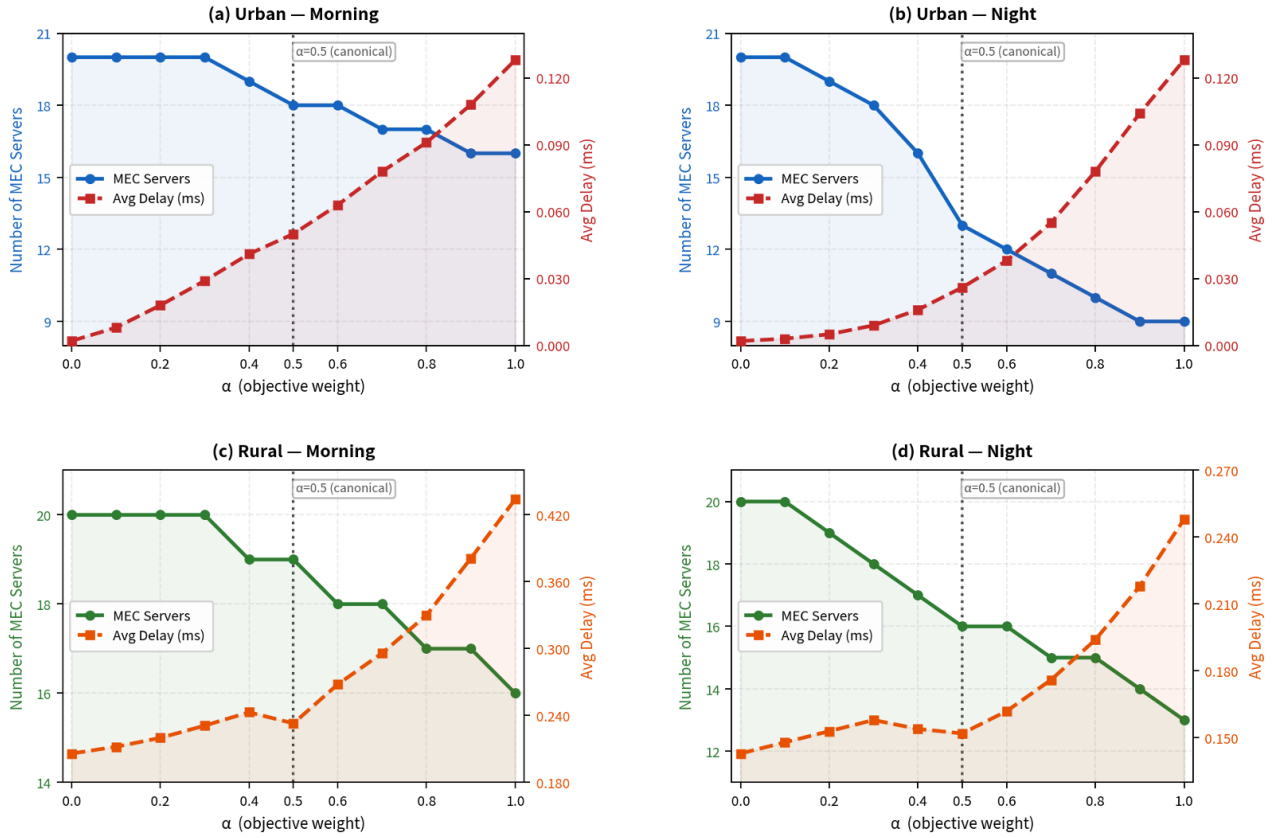
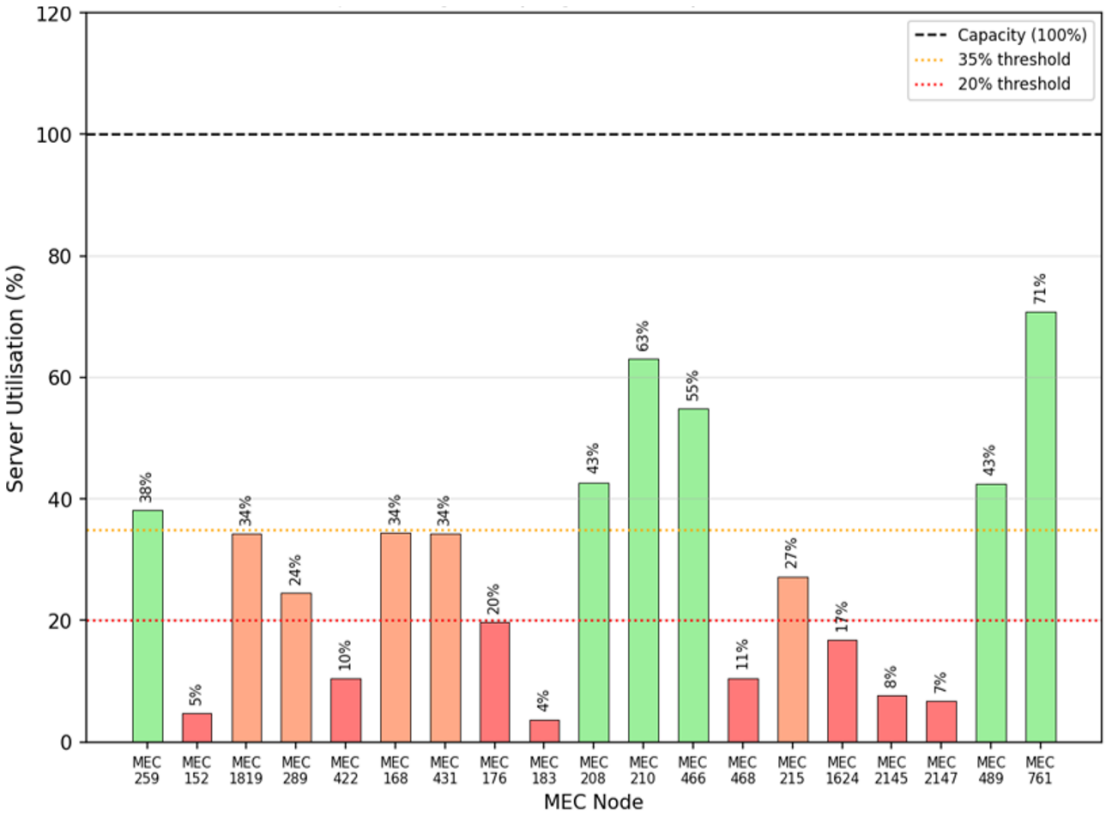


Fig. 4. Effect of the weighting parameter α on number of deployed MEC servers (left axis) and average traffic weighted average path delay (right axis) for: (a) urban morning, (b) urban night, (c) rural morning, and (d) rural night scenarios

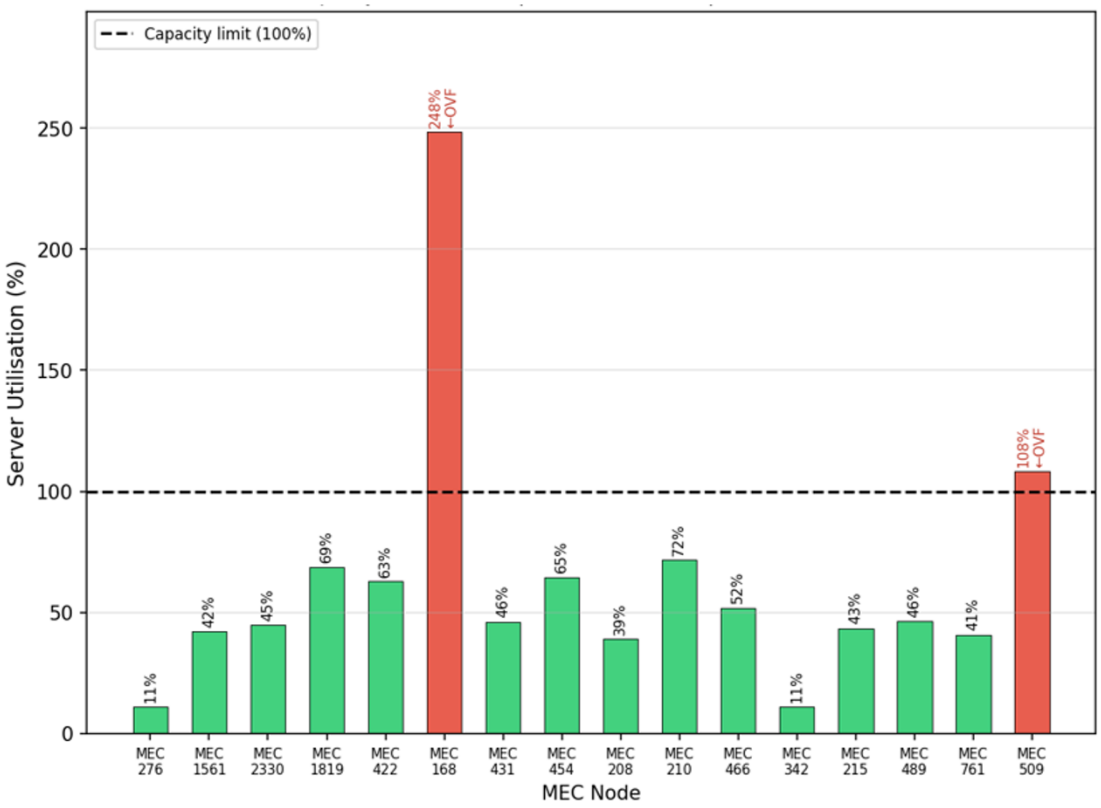
5.3 Cross Scenario Robustness

In this section, we investigate the robustness of the ILP-optimal placements when they are subjected to demand conditions other than the ones they were designed for. Since the MEC servers are physical infrastructure that cannot be relocated as the traffic varies over the day, a placement that is optimized for a single scenario may fail once the demand changes (e.g., a placement optimized for the night traffic may fail in the morning). We therefore evaluate two failure modes. The first is the capacity overflow, which occurs when an under-provisioned placement (designed for the off-peak demand of the night) is subjected to the peak demand of the morning. The second is the resource waste, which occurs when an over-provisioned placement (designed for the peak demand of the morning) remains idle during the off-peak hours of the night. In addition, we introduce and evaluate an aggregated deployment strategy, in which a single ILP placement is optimized on the aggregated demand, as a physically realizable compromise that remains feasible across all the traffic periods.

Figure 5 evaluates the two failure modes of a static placement in the rural sub-network. Fig. 5a shows the resource waste that results when the morning design is subjected to the night demand. Since the placement is dimensioned for the peak, all the 19 servers remain far below their capacity under (wasted MEC server resources) the lighter night load, and 8 of them fall below the 20% utilization threshold. In this case, no capacity violation occurs, but a large part of the deployed infrastructure remains idle (wasted MEC server resources). **Fig. 5b** shows the opposite failure mode, in which the night design is subjected to the morning demand. As the placement is under-provisioned for the peak, 2 servers exceed their capacity, with a maximum utilization of 248% at MEC 168, and the placement therefore becomes infeasible.



(a)



(b)

Fig. 5. (a) Workload of servers in rural area: Morning design with night demand, (b) Workload of servers in rural area: Night design with morning demand

Figure 6 summarizes the cross-scenario robustness of the rural placement in terms of three metrics: the feasibility, the number of overloaded servers, and the load-imbalance variance. Each design (the morning design and the night design) is tested under both the morning and the night demands. Both designs are feasible and well balanced under their own demand. When the designs are cross-evaluated, an asymmetric behavior is observed. The morning design remains feasible under the night demand, with no overloaded servers and a variance of 18.58, since it is over-provisioned for the lighter night load (with wasted server resources as shown in fig. 5a). The night design, on the other hand, becomes infeasible under the morning demand, where 2 servers are overloaded and the load-imbalance variance rises sharply to 136.82, compared to 15.86 for the morning design under the same demand. This confirms that a placement optimized for the off-peak demand cannot absorb the peak load, and a placement optimized for the peak demand wastes the MEC servers when applied to the off-peak demand. Since the MEC servers cannot be relocated as the demand varies over the day, this motivates a scenario-aware deployment that is dimensioned based on an aggregated demand. It is worth mentioning that same pattern is obtained when applying this cross-scenario to the urban area.

The aggregated demand is defined as a time-weighted combination of the peak (morning) and the off-peak (night) demands, given by

$$d_{agg}(i) = \lambda . d_{morning}(i) + (1 - \lambda) d_{night}(i) \quad (9)$$

$d_{morning}(i), d_{night}$ are the demands of base station i during the morning (peak) and the night (off-peak) intervals, respectively, and λ ∈ [0, 1] is a weighting factor that controls the bias of the aggregated demand toward the peak demand.

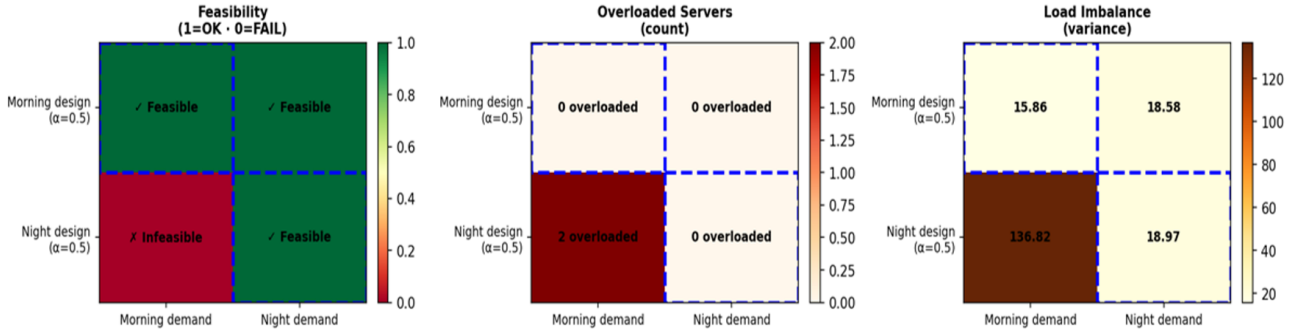


Fig. 6. Cross test heat map for rural area

To justify the choice of λ, a sensitivity analysis is carried out on the rural sub-network, in which the ILP is solved on the aggregated demand for different values of λ to obtain a single aggregated design, and each design is then tested under both the morning and the night demands. Table 6 reports the number of activated servers and the number of overflow servers for each case. It can be observed that, as λ increases, the aggregated demand is dimensioned closer to the peak; hence, the number of activated servers increases from 17 to 19, whereas the number of overflow servers under the morning demand decreases from 4 at λ = 0.45 to 0 at λ ≥ 0.65. Under the night demand, no overflow occurs for any value of λ, since the night demand is always lighter than the aggregated demand for which the design is dimensioned. Therefore, λ = 0.65 is selected, as it is the smallest weight that keeps the morning peak feasible with no overflow servers while activating the fewest number of servers, thus avoiding unnecessary resource waste during the off-peak periods.

Table 6
Sensitivity Analysis for λ

	Morning		Night	
λ	Number of MEC servers	Number of overflow servers	Number of servers	Number of overflow servers
0.45	17	4	17	0
0.55	17	3	17	0
0.65	18	0	18	0
0.75	18	0	18	0
0.85	19	0	19	0

Figure 7 and figure 8 evaluate the rural aggregated design (18 servers, α = 0.5) under the morning and the night demands, respectively. Under the morning demand (Fig. 7), the 18 servers absorb the peak load within their capacity, with an average utilization of 55.6% and a maximum of 96.5% at MEC 168, and only one server (MEC 183) falls below the 20% threshold; hence, the placement remains fully feasible at the peak. Under the night demand (Fig. 8), the placement is also feasible, but the average utilization drops to 30.4% and 7 of the 18 servers fall below the 20% threshold, which reflects a resource waste during the off-peak hours however, this waste is still lower than that incurred by the morning design when applied to the night demand (Fig. 5a). In both cases,

no server exceeds its capacity. This confirms that the aggregated design provides a physically realizable one-size-fits-all deployment: it remains feasible under both the morning and the night demands, and it avoids the capacity overflow of the night design, and less resource waste during the off-peak periods.

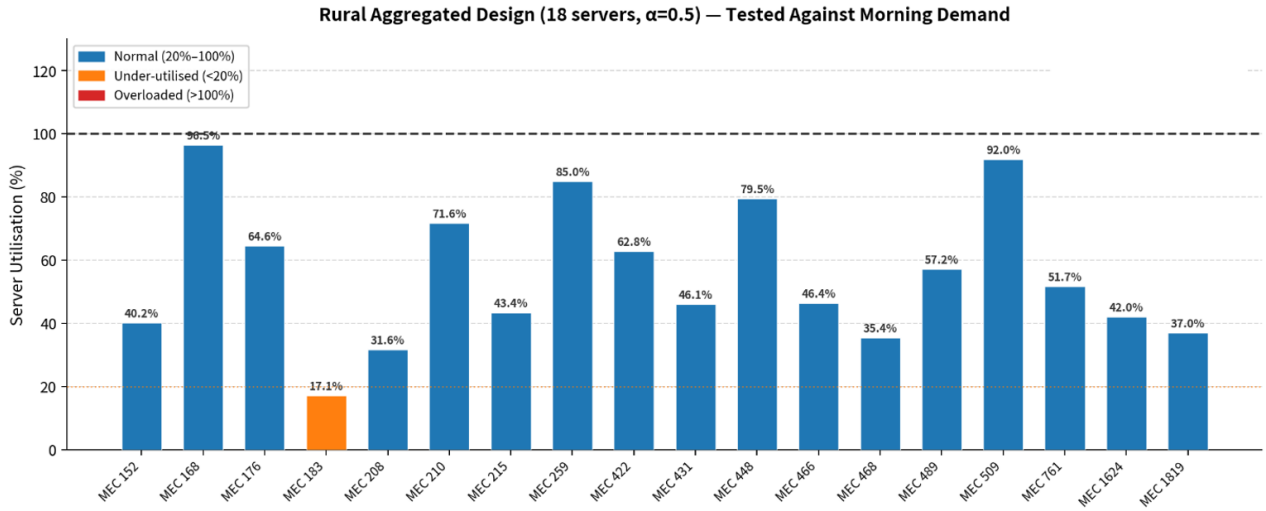


Fig. 7. Workload of servers in rural area: Aggregated design with morning demand

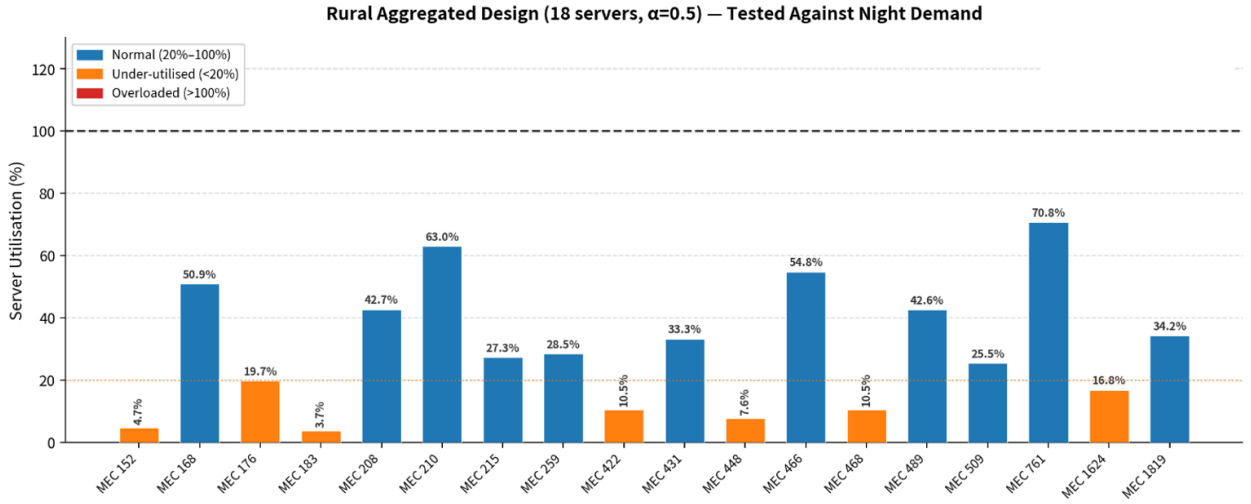


Fig. 8. Workload of servers in rural area: Aggregated design with night demand

6. Conclusion

In this paper, we investigated the joint MEC server placement and traffic routing problem and formulated it as an Integer Linear Programming (ILP) model that minimizes a weighted combination of the number of activated servers and the traffic-weighted average delay, subject to the server-capacity, the link-capacity, and the end-to-end-latency constraints. The model was evaluated on a real dataset from the Shanghai Telecom network, using a dense urban and a sparse rural sub-network. The results revealed a constraint regime shift between the two environments: the urban placement is dominated by the server capacity, whereas the rural placement is dominated by the end-to-end latency. As a result, the rural environment requires more servers and exhibits delays that

are 4.7–5.8× higher than the urban ones, and its placement is far more dispersed in order to maintain the latency feasibility of the isolated clusters of base stations. The cross-scenario robustness analysis showed that a static single-scenario placement is structurally inadequate, as it either overflows the capacity at the peak or wastes resources during the off-peak periods. To overcome this limitation, an aggregated deployment strategy was proposed, in which a single placement is dimensioned on a peak-biased aggregated demand; this placement remains feasible under both the morning and the night demands, while avoiding the overflow of the night design and reducing the waste of the morning design.

This study has several limitations that should guide future work. The evaluation is based on a single day of Shanghai Telecom traffic and two 50-BS sub-networks, which limits the generalizability of the reported constraint regime shift; broader validation across multiple dataset days, additional cities, and a wider range of network sizes is needed. The latency model considers only propagation delay (transmission, processing, queuing, and MEC service delays are not modelled) so the reported end-to-end latency should be interpreted as a lower bound rather than a complete latency figure. Similarly, the aggregation weight $\lambda = 0.65$ was selected from a limited sweep over five values and two traffic periods; a more exhaustive sensitivity analysis across additional λ values, traffic periods, and network topologies is left for future work to further validate the robustness of the aggregated deployment strategy.

Acknowledgements
This study was not supported by any grants from funding bodies in the public, private, or not-for-profit sectors.

Conflict of Interest
The authors declare no conflicts of interest

Authors Contribution
Sandrine Khaled: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft. **Mohamed Shehata**: Conceptualization, Methodology, Validation, Supervision, Writing – review & editing. All authors have read and agreed to the final version of the manuscript.

References
[1] Ericsson. 2026. Ericsson Mobility Report. June 2026.
[2] Li, Z., X. Liang, J. Liu, et al. 2025. “Optimizing Mobile-Edge Computing for Virtual Reality Rendering via UAVs: A Multiagent Deep Reinforcement Learning Approach.” IEEE Internet of Things Journal 12 (17): 35756–35772.
[3] Uddin, A., A. H. Sakr, and N. Zhang. 2026. “Counterfactual Multi-Agent DRL for Efficient Task Offloading in Vehicular Edge Computing.” IEEE Transactions on Vehicular Technology 75 (6): 9270–9281.
[4] Padmavathi, V., R. Kanimozhi, and R. Saminathan. 2025. “Digital Twin Driven Smart Factories: Real Time Physics Based Co-Simulation Using Edge A.I. and Federated Learning.” Scientific Reports 15: 43373.
[5] Li, P., X. Wang, B. Yi, et al. 2026. “Cost-Aware Routing for Computation Offloading in Knowledge-Defined AIoT.” Future Generation Computer Systems 174: 108013.
[6] International Telecommunication Union Radiocommunication Sector (ITU-R). 2017. Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s). Report ITU-R M.2410-0. November 2017.
[7] Zhai, L., X. Xu, Z. Liu, L. Wei, X. Li, et al. 2026. “Access Selection and Service Placement in Mobile Edge Computing Networks.” Computer Networks 283: 112317.
[8] Guo, Y., and C. Chen. 2025. “DRLO: Optimizing Edge Server Placement in Dynamic MEC Scenarios Using Deep Reinforcement Learning.” Computer Networks 268: 111377.
[9] Shibata, K., and S. Miyata. 2025. “Edge Server Placement and Task Allocation for Maximum Delay Reduction.” IEEE Open Journal of the Communications Society 6: 6207–6217.

[10] Elmobark, N., H. El-Ghareeb, and S. Elhishi. 2025. "BlueEdge Neural Network Approach and Its Application to Automated Data Type Classification in Mobile Edge Computing." Scientific Reports 15: 43823.
[11] Nguyen Trung, K., M.-N. Tran, and Y. Kim. 2026. "Enabling Service Continuity for Stateful Service Segmentation in Mobile Edge Computing toward 6G." IEEE Access 14: 20589–20604.
[12] Dash, S., A. U. Khan, B. Kar, et al. 2024. "CRAMP: Clustering-Based RANs Association and MEC Placement for Delay-Sensitive Applications." Journal of Network and Computer Applications 227: 103893.
[13] Forghieri, O., Y. Carlinet, E. Hyon, E. Le Pennec, and N. Perrot. 2025. "Faster Latency-Constrained Service Placement in Edge Computing with Deep Reinforcement Learning." In Proceedings of the IFIP Networking Conference Workshops (TX4NETs). Limassol, Cyprus.
[14] Li, S., Y. Zhou, B. Zhou, and Z. Wang. 2025. "Workload-Based Adaptive Decision-Making for Edge Server Layout with Deep Reinforcement Learning." Engineering Applications of Artificial Intelligence 139: 109662.
[15] Wang, X., Z. Zhou, F. Xiao, et al. 2018. "Spatio-Temporal Analysis and Prediction of Cellular Traffic in Metropolis." IEEE Transactions on Mobile Computing 18 (9): 2190–2202.
[16] Anzola-Rojas, C., I. de Miguel, J. C. Aguado, N. Merayo, P. Fernández, et al. 2024. "Joint Fiber and MEC Deployment for Sparsely Populated Areas." Journal of Optical Communications and Networking 16 (1): 45–58.
[17] Li, Y., A. Zhou, X. Ma, and S. Wang. 2022. "Profit-Aware Edge Server Placement." IEEE Internet of Things Journal 9 (1): 55–67.
[18] Wang, S., Y. Guo, N. Zhang, P. Yang, A. Zhou, et al. 2021. "Delay-Aware Microservice Coordination in Mobile Edge Computing: A Reinforcement Learning Approach." IEEE Transactions on Mobile Computing 20 (3): 939–951.
[19] Wang, S., Y. Zhao, J. Xu, J. Yuan, and C.-H. Hsu. 2019. "Edge Server Placement in Mobile Edge Computing." Journal of Parallel and Distributed Computing 127: 160–168.
[20] "Shanghai Telecom Dataset." 2014. Online dataset. http://sguangwang.com/TelecomDataset.html.