



# Joint user plane function instance and base station scheduling in mobile networks

Seokwon Jang<sup>1</sup> | Namseok Ko<sup>1</sup> | Jaewook Lee<sup>2</sup>  | Yeunwoong Kyung<sup>3</sup> | Haneul Ko<sup>4</sup> 

<sup>1</sup>Mobile Core Network Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

<sup>2</sup>Department of Information and Communication Engineering, Pukyong National University, Busan, Republic of Korea

<sup>3</sup>Division of Information & Communication Engineering, Kongju National University, Cheonan, Republic of Korea

<sup>4</sup>Department of Electronic Engineering, Kyung Hee University, Yongin, Republic of Korea

## Correspondence

Haneul Ko, Department of Electronic Engineering, Kyung Hee University, Yongin, Gyeonggi, Republic of Korea.  
Email: [heko@khu.ac.kr](mailto:heko@khu.ac.kr)

## Funding information

ICT R&D Program of MSICT/IITP (2021-0-02094, International Collaborative Research on 6G Network Architecture and Core Technologies).

## Abstract

To guarantee a high data transmission rate in heterogeneous mobile networks, sufficient small base stations (SBSs) and user plane function (UPF) instances should be active. However, the excessive operation of SBSs and UPF instances can increase the operating expenditure (OPEX) for the network operator. To balance the data rate and OPEX, we propose a joint UPF instance–SBS scheduling algorithm (J-UBSA). In the proposed J-UBSA, a controller periodically determines the appropriate number of active SBSs and UPF instances based on estimated variations in session requests. The decision-making process is formulated as a constrained Markov decision process and converted into a linear programming model to obtain the optimal solution using a traditional algorithm with low complexity. Evaluation results demonstrate that J-UBSA can substantially reduce the OPEX while guaranteeing a suitable average data rate for mobile devices.

## KEYWORDS

constrained Markov decision process, joint optimization, operating expenditure, sleep scheduling, user plane function

## 1 | INTRODUCTION

To increase the data transmission rate in a radio access network by allowing the spatial reuse of frequencies [1–3], several small base stations (SBSs) should be deployed. Meanwhile, the user plane function (UPF) can be implemented as a containerized software instance [4–6], and each UPF instance can serve a certain data rate by allocating resources to the container instance. Hence, to increase the data rates using an end-to-end structure in mobile networks, the number of operating

SBSs and UPF instances must be increased. However, this can lead to a higher operating expenditure (OPEX) for network operators [5, 7, 8]. Therefore, to balance the tradeoff between data rate and OPEX, the appropriate number of active SBSs and UPF instances should be determined based on the number of mobile devices (MDs), which can be equated to the number of protocol data unit (PDU) session requests.

We propose a joint UPF instance and SBS scheduling algorithm (J-UBSA), in which the controller periodically adjusts the number of operating SBSs and UPF instances

based on estimated variations in PDU session requests. For example, if more PDU sessions are expected, the controller proactively increases the number of operating SBSs and UPF instances. However, owing to the latency in activating SBSs and/or initializing UPF instances, their quantities should be predetermined based on the estimated session requests. To minimize the OPEX while ensuring adequate data rates, we formulate a constrained Markov decision process (CMDP). To obtain the optimal stochastic policy, the formulated problem is converted into a linear programming (LP) model that can be efficiently solved using conventional low-complexity algorithms. Evaluation results demonstrate that J-UBSA can substantially reduce the OPEX while guaranteeing an adequate average data rate for MDs. In addition, J-UBSA adapts its policy according to the operating environment (i.e., minimizing the number of SBSs and UPF instances while limiting the number of MDs handled per SBS and UPF instance). The contributions of this study are as follows. (1) To the best of our knowledge, this is the first effort to jointly optimize the number of operating SBSs and UPF instances, ensuring an adequate end-to-end data rate with a reduced OPEX. (2) The joint optimal policy for operating SBSs and UPF instances has a low computational complexity, thereby facilitating the implementation of the proposed algorithm in real systems. (3) Evaluation results encompass various environments. The comprehensive evaluation provides valuable insights and practical recommendations for the design of cost-efficient and high quality-of-service (QoS)-guaranteed mobile networks.

The remainder of this paper is organized as follows. Section 2 provides a summary of related work. Section 3 introduces the proposed J-UBSA and its working principles. The CMDP is formulated in Section 4. Evaluation results are presented in Section 5. Finally, we draw conclusions in Section 6.

## 2 | RELATED WORK

Numerous studies have been focused on balancing the tradeoff between data rate and OPEX [5, 7–18]. These studies rely on either (1) base station (BS) scheduling [7–11, 13, 15] or (2) network function (NF) instance scheduling [5, 12, 14, 16–18]. Sambo and others [8] proposed a simple BS sleep scheduling algorithm with a motion sensor that detects MDs in the coverage area. Chopra [7] analyzed the tradeoff between the data rate and fairness and proposed a sleep scheduling configuration according to the number of connected MDs. Lee and others [10] studied sleep scheduling for energy-harvesting BSs considering the uncertainty of harvesting energy to reduce the OPEX. Celebi and others [11] analyzed traffic

load statistics using gamma approximation and proposed two sleep scheduling algorithms based on traffic load statistics. Luo and others [13] proposed a joint power control and sleep scheduling algorithm and formulated a Lyapunov optimization problem to minimize power consumption and delay. Lee and others [15] investigated a traffic threshold to adjust the coverage of BSs to reduce the OPEX. Femenias and others [9] analyzed the performance of BS scheduling policies considering the number and/or locations of MDs.

Nguyen and Rotter [5] proposed a reinforcement-learning-based UPF instance scheduling algorithm to minimize UPF instances while maintaining the QoS considering traffic volumes. Woo and others [14] designed a method for NF elastic scaling in which the states of the NFs are shared using a distributed shared object space. Leyva-Pupo and others [12] formulated an integer LP problem to determine the number of UPF instances and MD allocation to the UPF instances. Subramanya and Riggio [18] designed centralized and federated approaches to predict the number of required NF instances based on the expected traffic demand. Wu and others [16] developed formal abstractions for general NF packet-processing pipelines to accommodate up/down NF scaling. Lin and Leon-Garcia [17] proposed and demonstrated a prototype to monitor network performance metrics (e.g., data rate) and manage NF instances for improving the MD QoS.

Notably, the numbers of operating SBSs and UPF instances have not been optimized in existing studies.

## 3 | PROPOSED J-UBSA

Figure 1 shows the system model for this study. We consider heterogeneous wireless networks comprising one macro-BS and multiple SBSs. The BSs are interconnected with each other and clouds. To reduce the OPEX, the network operator can deactivate certain SBSs (that is, it activates the sleep mode in certain SBSs), but the macro-BS cannot be set to the sleep mode to avoid disrupting the service coverage. When several SBSs are in the sleep mode, the data rate in the radio access network may drop owing to inadequate spatial frequency reuse. Hence, to guarantee suitable data rates in the core networks, the operator should increase the number of UPF instances.<sup>1</sup> However, an excessive operation of UPF instances may increase the OPEX.

<sup>1</sup>Allocating more resources to active UPF instances may increase data rates, but this requires temporarily stopping the active UPF instances, leading to service interruptions. Therefore, we assume that the network operator adjusts the number of UPF instances to ensure the data rate while minimizing the potential impact to service continuity.

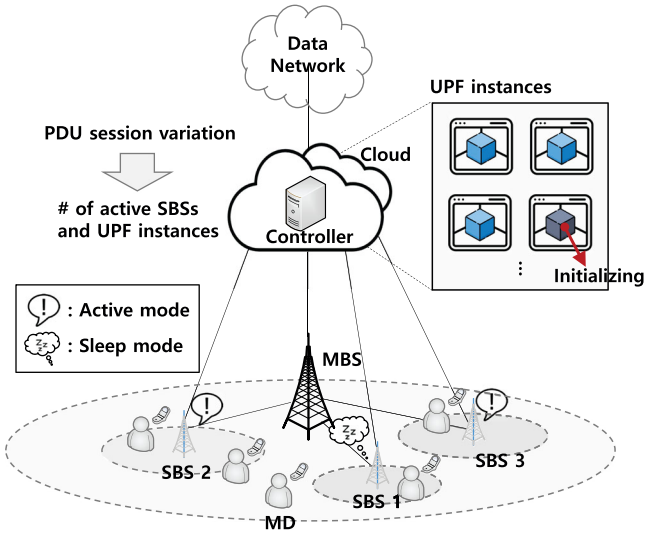


FIGURE 1 System model.

We exploit J-UBSA in the network controller to balance the tradeoff between data rate and OPEX in mobile networks. Specifically, the controller dynamically determines the number of operating SBSs and UPF instances by considering variations in PDU session requests. We formulate the CMDP model, which is elaborated in the following section, to make the optimal decisions.

## 4 | CMDP

In the CMDP model, an agent follows a series of actions to minimize (or maximize) a cost (or reward) while adhering to specific constraints [19]. The network operator makes decisions regarding the numbers of active SBSs and UPF instances over time at steps denoted as  $\mathbf{T} = 1, 2, 3, \dots$  to minimize the OPEX while ensuring that the data rate is maintained at an acceptable level. Hence, the numbers of SBSs and UPF instances should be minimized while maintaining the average number of MDs handled per SBS and UPF instance below a predetermined threshold. The notation for the CMDP model is summarized in Table 1.

### 4.1 | State space

State space  $\mathbf{S}$  can be defined as the Cartesian product of three individual state spaces, namely,  $\mathbf{B}$  (number of SBSs in active mode),  $\mathbf{U}$  (number of active UPF instances), and  $\mathbf{M}$  (number of MDs):

TABLE 1 Notation for CMDP model.

Notation	Description
$\mathbf{S}$	State space
$\mathbf{B}$	State space for number of active SBSs
$\mathbf{U}$	State space for number of active UPF instances
$\mathbf{M}$	State space for number of MDs
$N_B$	Number of SBSs
$N_U$	Maximum number of UPF instances in cloud
$N_M$	Maximum number of MDs
$\mathbf{A}$	Action space
$\mathbf{A}_B$	Action space for SBS scheduling
$\mathbf{A}_U$	Action space for UPF instance scheduling
$1/\lambda_B$	Average latency for activating SBSs
$1/\lambda_U$	Average latency for initializing UPF instances
$\alpha$	Unit cost for operating single SBS
$\beta$	Unit cost for operating single UPF instance
$\zeta_O$	Average OPEX
$\psi_B$	Average number of MDs handled per SBS
$\psi_U$	Average number of MDs handled per UPF instance

$$\mathbf{S} = \mathbf{B} \times \mathbf{U} \times \mathbf{M}. \quad (1)$$

Let  $N_B$  be the number of SBSs in the system model. Then,  $\mathbf{B}$  can be expressed as

$$\mathbf{B} = \{0, 1, 2, \dots, N_B\}. \quad (2)$$

For  $N_U$  being the maximum number of active UPF instances in the cloud,  $\mathbf{U}$  can be described as

$$\mathbf{U} = \{0, 1, 2, \dots, N_U\}. \quad (3)$$

Similarly, for  $N_M$  being the maximum number of MDs in a system model,  $\mathbf{M}$  can be described as

$$\mathbf{M} = \{0, 1, 2, \dots, N_M\}. \quad (4)$$

### 4.2 | Action space

Action space  $\mathbf{A}$  is defined as

$$\mathbf{A} = \mathbf{A}_B \times \mathbf{A}_U, \quad (5)$$

where  $\mathbf{A}_B$  and  $\mathbf{A}_U$  are the action spaces for SBS and UPF instance scheduling, respectively.

$$\mathbf{A}_B = \{0, 1, 2, \dots, N_B\}, \quad (6)$$

where  $A_B (\in \mathbf{A}_B)$  is the number of SBSs that the controller attempts to operate at the next timestep.

$$\mathbf{A}_U = \{0, 1, 2, \dots, N_U\}, \quad (7)$$

where  $A_U (\in \mathbf{A}_U)$  is the number of UPF instances that the controller attempts to operate at the next timestep.

### 4.3 | Transition probability

All states in the system have independent transitions. The state of the number of SBSs in active mode,  $B$ , is influenced by the SBS scheduling action,  $A_B$ . Similarly, the state of the number of active UPF instances,  $U$ , is affected by the UPF instance scheduling action,  $A_U$ . Therefore, the transition probability from current state  $S = [B, U, M]$  to next state  $S' = [B', U', M']$  can be described as

$$P[S'|S, A] = P[B'|B, A_B] \times P[U'|U, A_U] \times P[M'|M]. \quad (8)$$

We assume that the latency for activating the SBSs follows an exponential distribution with mean  $1/\lambda_B$ . The probability of activating the SBSs within decision period  $\tau$  is  $\lambda_B \tau$  [20]. The latency for activating SBS is needed only when the current number of SBSs in active mode,  $B$ , is less than the number of SBSs that the controller tries to operate during the next timestep,  $A_B$  (i.e., only when  $A_B > B$ ). Therefore, the corresponding transition probability can be calculated as

$$P[B'|B, A_B > B] = \begin{cases} \lambda_B \tau, & \text{if } B' = A_B \\ 1 - \lambda_B \tau, & \text{if } B' = B \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Meanwhile, SBSs can be deactivated immediately. Thus, when  $A_B \leq B$ , the corresponding transition probability can be calculated as

$$P[B'|B, A_B \leq B] = \begin{cases} 1, & \text{if } B' = A_B \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

We also assume that the latency for initializing the UPF instances follows an exponential distribution with mean  $1/\lambda_U$ . The probability of initializing the UPF instances within decision period  $\tau$  is  $\lambda_U \tau$  [20].

Therefore, the corresponding transition probability can be calculated as

$$P[U'|U, A_U > U] = \begin{cases} \lambda_U \tau, & \text{if } U' = A_U \\ 1 - \lambda_U \tau, & \text{if } U' = U \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Because UPF instances are immediately deactivated (i.e., the latency of deactivating UPF instances is zero), when the number of active UPF instances  $U$  is larger than or equal to that of SBSs that the controller attempts to operate at the next timestep,  $A_U$ , the latency for deactivating the SBS can be neglected. Thus,  $P[U'|U, A_U \leq U]$  can be expressed as

$$P[U'|U, A_U \leq U] = \begin{cases} 1, & \text{if } U' = A_U \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

Furthermore, the transition probability of  $M$  can be defined statistically as follows:

$$r(S, A) = \alpha B + \beta U, \quad (13)$$

where  $\alpha$  and  $\beta$  are the unit costs of operating a single SBS and UPF instance, respectively.

A state diagram based on the formulated CMDP model is shown in Figure 2. The number of MDs,  $M$ , remains constant, regardless of the action, to simplify the diagram. In Figure 2, each case represents the state transitions resulting from a specific action of the CMDP agent. For example, when  $A_B > B$  and  $A_U > U$  (i.e., Case 1), state  $(B; U)$  transitions to state  $(A_B; U)$  with probability  $(1 - \lambda_U \tau) \lambda_B \tau$ .

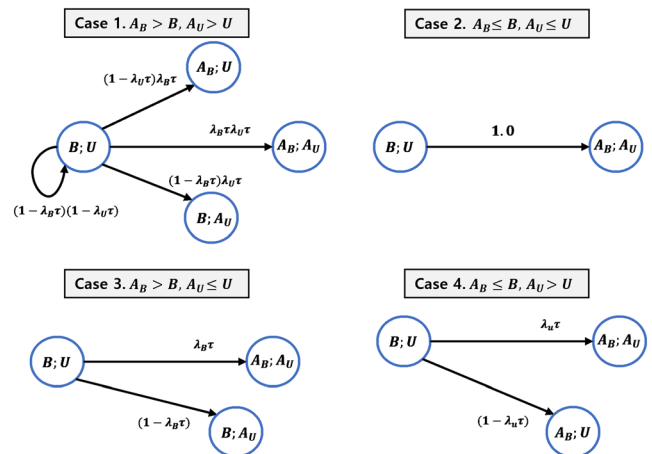


FIGURE 2 Simplified state diagram of CMDP model.

## 4.4 | Cost and constraint functions

### 4.4.1 | Cost function

We define the cost function for the OPEX, which is proportional to the number of active SBSs and UPF instances the network operator. The cost function for the OPEX,  $r(S, A)$ , is defined as.

### 4.4.2 | Constraint functions

We define constraint functions for the data rate. To deliver an adequate end-to-end data rate in mobile networks, we limit the number of MDs served by a single SBS and UPF instance to certain thresholds. We assume that MDs are uniformly distributed on the service coverage area, and their PDU sessions are uniformly distributed over the active UPF instances. Therefore, the constraint functions for the number of MDs at a single SBS and UPF instance,  $c_B(S, A)$  and  $c_U(S, A)$ , respectively, are defined as

$$c_B(S, A) = \frac{M}{B}, \quad (14)$$

$$c_U(S, A) = \frac{M}{U}. \quad (15)$$

## 4.5 | Optimization

To obtain the optimal solution in the target stochastic environment, we first define time-averaged performance metrics (i.e., OPEX, number of MDs handled per SBS, and number of MDs handled per UPF instance) using the limit supremum. In detail, the time-averaged OPEX, number of MDs per SBS, and number of MDs per UPF instance,  $\zeta_O$ ,  $\psi_B$ , and  $\psi_U$ , respectively, can be expressed as

$$\zeta_O = \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{t'}^t E[r(S_{t'}, A_{t'})], \quad (16)$$

$$\psi_B = \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{t'}^t E[c_B(S_{t'}, A_{t'})], \quad (17)$$

and

$$\psi_U = \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{t'}^t E[c_U(S_{t'}, A_{t'})], \quad (18)$$

where  $S_{t'}$  and  $A_{t'}$  are the state and chosen actions at  $t' \in \mathbf{T}$ , respectively.

With the time-averaged performance metrics, the minimization of the OPEX while maintaining the number of MDs per single SBS and UPF instance below certain levels can be formulated as follows:

$$\min_{\pi} \zeta_O, \quad (19)$$

$$\text{s.t. } \psi_B \leq \theta_B \text{ and } \psi_U \leq \theta_U, \quad (20)$$

where  $\pi$  represents the policy that determines the probability of selecting a specific action per state. The upper bounds for the average numbers of MDs per single SBS and UPF instance are denoted by  $\theta_B$  and  $\theta_U$ , respectively, which ensure a continuously adequate data rate.

To convert the CMDP model into an equivalent LP model, we introduce decision variable  $\varphi_{S,A}$ , which represents the stationary probabilities of state  $S$  and action  $A$ .<sup>2</sup> The LP model can be represented as follows:

$$\min_{\varphi_{S,A}} \sum_S \sum_A \varphi_{S,A} r(S, A), \quad (21)$$

subject to

$$\sum_S \sum_A \varphi_{S,A} c_B(S, A) \leq \theta_B, \quad (22)$$

$$\sum_S \sum_A \varphi_{S,A} c_U(S, A) \leq \theta_U, \quad (23)$$

$$\sum_A \varphi_{S',A} = \sum_S \sum_A \varphi_{S,A} P[S'|S, A], \quad (24)$$

$$\sum_S \sum_A \varphi_{S,A} = 1, \quad (25)$$

$$\varphi_{S,A} \geq 0. \quad (26)$$

The objective function described in (21) aims to minimize the average OPEX of the network operator. The constraints in (22) and (23) are aligned with those of the CMDP model stated in (20). In addition, (24) represents the constraint for the Chapman–Kolmogorov equation, whereas the probability properties are satisfied by the constraints in (25) and (26).

Once optimal solution  $\varphi_{S,A}^*$  of the LP model is obtained, we can derive the optimal stochastic policy,  $\pi^*$ ,

<sup>2</sup>Converting the CMDP model into an LP problem is a widely used general solution method [21–23].

for the CMDP model. The optimal policy represented by the optimal probability distribution allows the controller to dynamically determine the numbers of active SBSs and UPF instances.

## 5 | EVALUATION RESULTS

To demonstrate the effectiveness of the proposed J-UBSA in terms of OPEX and data rate, we devised the following comparative approaches: (1) MAX, in which the network operator activates the maximum number of SBSs and UPF instances; (2) MIN, in which the network operator activates one SBS and one UPF instance; (3) OPT-BS, in which the network operator operates the optimal number of SBSs and randomly activates UPF instances; and (4) OPT-UPF, in which the network operator operates the optimal number of UPF instances and randomly activates SBSs. The optimal number of SBSs in OPT-BS and that of UPF instances in OPT-UPF are the same as those selected by J-UBSA.

We used several performance metrics to evaluate the algorithm effectiveness, including the average OPEX,  $\zeta_O$ ; average number of MDs per SBS,  $\psi_B$ ; average number of MDs per UPF instance,  $\psi_U$ ; and satisfaction ratio (QoS) regarding the data rate,  $\psi_Q$ . Satisfaction is reduced if  $\psi_B$  and  $\psi_U$  exceed their upper bounds for the average number of MDs per SBS and UPF instance,  $\theta_B$  and  $\theta_U$ . In this case,  $\psi_Q$  is calculated as  $\min(\theta_B/\psi_B, \theta_U/\psi_U)$ . On the other hand, if  $\psi_B \leq \theta_B$  and  $\psi_U \leq \theta_U$ , no satisfaction degradation occurs, and  $\psi_Q$  equals 1.

The default parameters were set as follows. The number of SBSs,  $N_B$ , was 5, and the maximum number of active UPF instances,  $N_U$ , was 5, while the maximum number of MDs,  $N_M$ , was 15. The rates for activating SBSs,  $\lambda_B$ , and initializing UPF instances,  $\lambda_U$ , were set to 0.55 and 0.8, respectively. The unit costs for operating a

single SBS and UPF instance,  $\alpha$  and  $\beta$ , respectively, were set to 1. The upper bounds for the average number of MDs per SBS and UPF instance,  $\theta_B$  and  $\theta_U$ , were set to 3.5 and 3, respectively.

### 5.1 | Effect of $E[M]$

Figure 3 shows the effects of the average number of MDs,  $E[M]$ , on the algorithm performances. Figure 3A shows that the average OPEX,  $\zeta_O$ , obtained from J-UBSA increases with increasing  $E[M]$ . This is because numerous MDs demand activating an adequate number of SBSs and UPF instances to ensure a satisfactory data rate. In J-UBSA, the controller acknowledges this requirement and dynamically adjusts the number of active SBSs and UPF instances in proportion to the number of MDs. Specifically, by considering the variation in the number of MDs, the controller dynamically changes the number of active SBSs and UPF instances. For example, when the number of MDs is expected to increase, the controller increases the number of active SBSs and UPF instances in advance, thus increasing the OPEX for the network operator. A sufficient number of SBSs and UPF instances should be activated in advance before the number of MDs increases to seamlessly guarantee a satisfactory data rate considering the latency for activating SBSs and/or initializing UPF instances.

Figure 3B,C shows that the average numbers of MDs handled per SBS and UPF instance,  $\psi_S$  and  $\psi_U$ , of J-UBSA do not exceed upper bounds  $\theta_B$  and  $\theta_U$ , respectively (i.e., an adequate data rate can be guaranteed).

OPT-BS and OPT-UPF follow the policy of J-UBSA for the numbers of active SBSs and UPF instances, respectively. Hence, the number of active SBSs in OPT-BS (or UPF instances in OPT-UPF) increases proportionally with the average number of MDs,  $E[M]$ , thus increasing the OPEX. In contrast, the average OPEX,  $\zeta_O$ ,

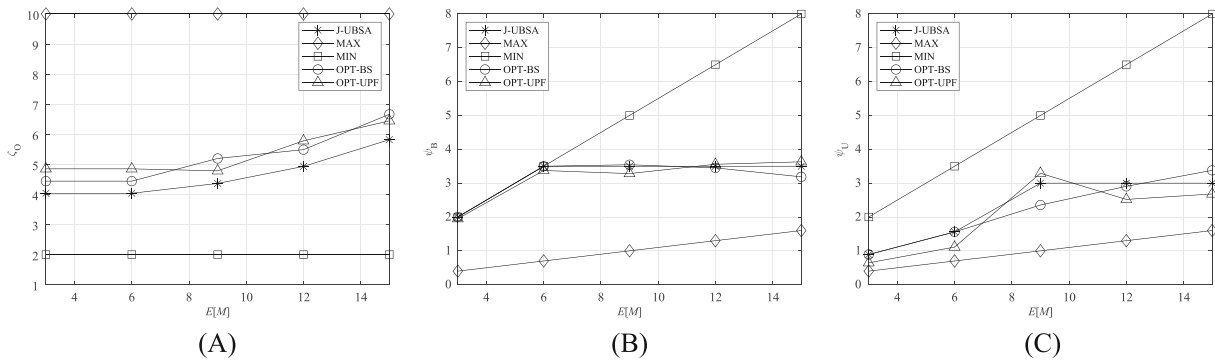
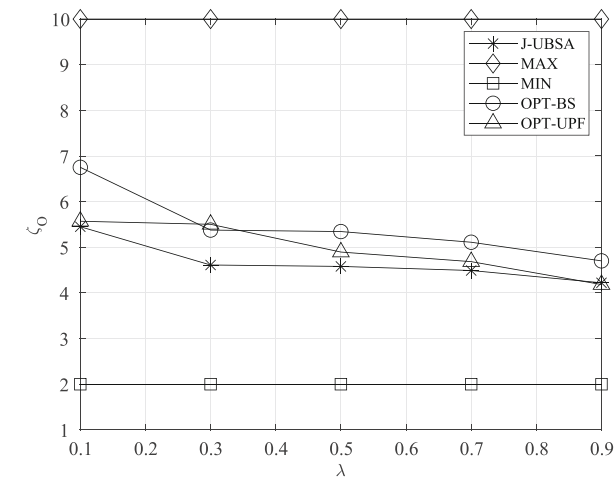


FIGURE 3 Effect of average number of MDs on average (A) OPEX, (B) number of MDs handled per SBS, and (C) number of MDs handled per UPF instance.

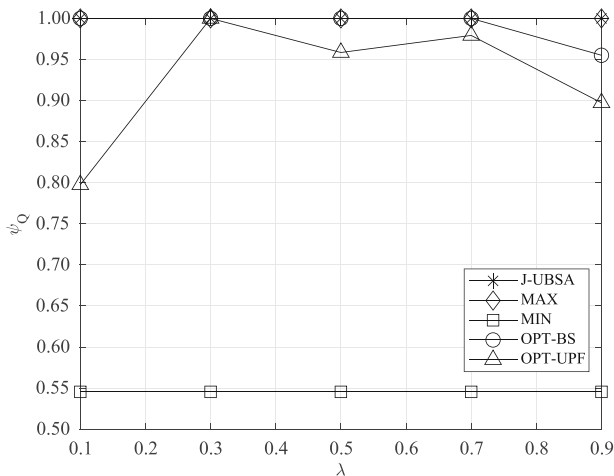
for MAX and MIN remain constant regardless of  $E[M]$ . This is because they do not adjust their policies based on the number of MDs but maintain a fixed number of active SBSs and UPF instances.

### 5.2 | Effect of $\lambda$

Figure 4A,B shows the effect of the SBS activation/UPF initialization rate,  $\lambda$ , on the average OPEX,  $\zeta_O$ , and average satisfaction ratio,  $\psi_Q$ , respectively. Figure 4A shows that  $\zeta_O$  using J-UBSA decreases with increasing  $\lambda$ , possibly because a higher  $\lambda$  enables more dynamic control of the number of SBSs and UPF instances according to the number of MDs. Hence, J-UBSA can maintain a reduce the number of active SBSs and UPF instances, thereby reducing the OPEX without compromising the



(A)



(B)

FIGURE 4 Effect of SBS/UPF instance activation rate on average (A) OPEX and (B) satisfaction ratio.

satisfaction ratio (Figure 4B). The similar OPEX trends observed for OPT-BS and OPT-UPF are due to the replication of the J-UBSA policy regarding the number of active SBSs and active UPF instances, respectively. In contrast, MAX and MIN do not consider  $\lambda$  in their operating policies, resulting in a constant OPEX irrespective of  $\lambda$ .

### 5.3 | Effect of $\alpha$

Figure 5 shows the influence of the unit cost of operating a single SBS,  $\alpha$ , on the average OPEX,  $\zeta_O$ , and other metrics. Average OPEX  $\zeta_O$  increases for all schemes as unit cost  $\alpha$  increases. However, J-UBSA consistently achieves the lowest OPEX among the compared schemes, except for MIN.

The advantage of J-UBSA can be attributed to its efficient strategy for minimizing the number of active SBSs, particularly when unit cost  $\alpha$  is high. By activating only necessary SBSs, J-UBSA optimizes the utilization of resources, leading to a reduced OPEX compared with other schemes in various cost scenarios.

### 5.4 | Effect of $\theta_B$

Figure 6 shows the effect of the upper bound on the average number of MDs handled per SBS,  $\theta_B$ . Figure 6A,B shows that J-UBSA minimizes the OPEX while ensuring that the number of MDs per SBS remains below upper bound  $\theta_B$ . This is because J-UBSA activates the necessary SBSs by considering dynamic changes in the number of MDs.

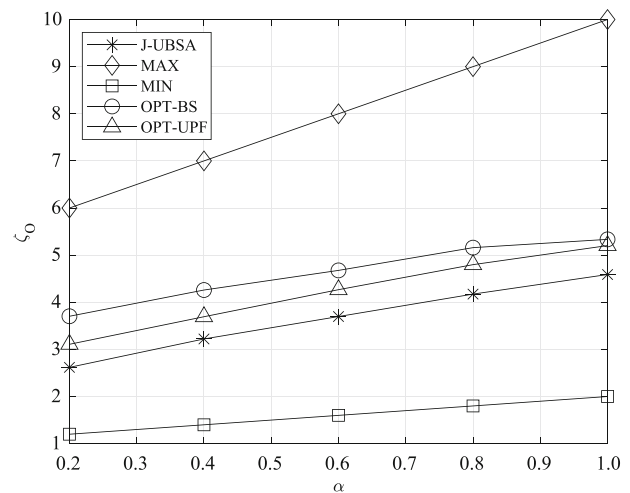
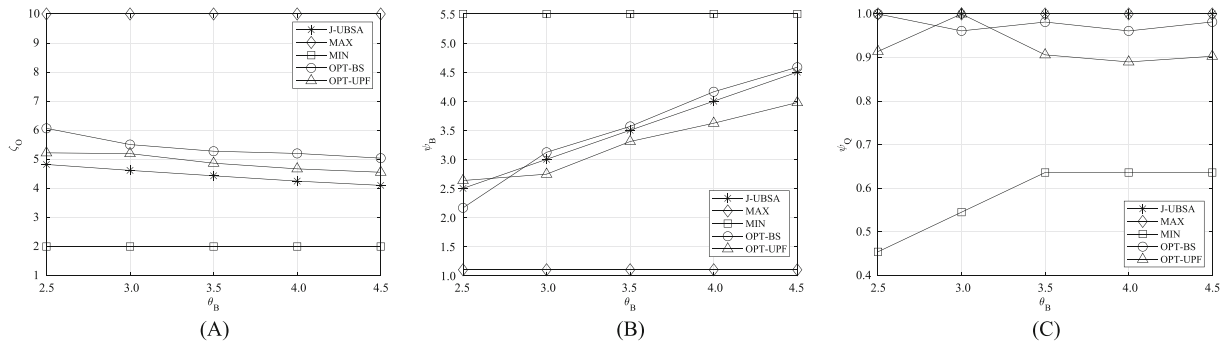


FIGURE 5 Effect of unit cost for operating a single SBS on average OPEX.



**FIGURE 6** Effect of the upper bound for average number of MDs per SBS on average (A) OPEX, (B) number of MDs per SBS, and (C) satisfaction ratio.

The comparison schemes do not consider fluctuations in the number of MDs or the latency associated with activating SBSs (or initializing UPF instances). Consequently, they fail to consistently maintain the number of MDs per SBS below  $\theta_B$  (Figure 6B). Consequently, their satisfaction ratio deviates from 1, unlike J-UBSA, which maintains a satisfaction ratio of 1 across all the scenarios, as shown in Figure 6C.

## 6 | CONCLUSION

We introduce J-UBSA to optimize the scheduling of SBSs and UPF instances based on fluctuations in PDU session requests. By formulating the problem as a CMDP and converting it into an equivalent LP model, we derive an optimal policy for SBS and UPF instance scheduling with low complexity. Evaluation results demonstrate that J-UBSA can substantially reduce the OPEX while ensuring a satisfactory data rate. In future research, we plan to extend our algorithm to consider MD locations, enabling the dynamic activation of specific SBSs based on their proximity to MDs. Such enhancement may further improve the efficiency and performance of mobile networks.

### CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

### ORCID

Jaewook Lee  <https://orcid.org/0000-0003-0422-280X>

Haneul Ko  <https://orcid.org/0000-0002-9067-445X>

### REFERENCES

1. A. Zhu, M. Ma, S. Guo, S. Yu, and L. Yi, *Adaptive multi-access algorithm for multi-service edge users in 5G ultra-dense heterogeneous networks*, IEEE Trans. Veh. Technol. **70** (2021), no. 3, 2807–2821.
2. G. Zhang, H. Zhang, Z. Han, and G. K. Karagiannidis, *Spectrum allocation and power control in full-duplex ultra-dense heterogeneous networks*, IEEE Trans. Commun. **67** (2019), no. 6, 4365–4380.
3. M. Ding, D. Lopez-Perez, H. Claussen, and M. A. Kaafar, *On the fundamental characteristics of ultra-dense small cell networks*, IEEE Netw. **32** (2018), no. 3, 90–100.
4. B. Dzogovic, B. Feng, and T. Van Do, *Building virtualized 5G networks using open source software*, (Proceedings of IEEE Symposium on Computer Applications & Industrial Electronics, Penang, Malaysia), 2018. DOI [10.1109/ISCAIE.2018.8405499](https://doi.org/10.1109/ISCAIE.2018.8405499)
5. T. D. Nguyen and C. Rotter, *Scaling UPF instances in 5G/6G core with deep reinforcement learning*, IEEE Access **9** (2021), 165892–165906.
6. C. Rotter and T. V. Do, *A queuing model for threshold-based scaling of UPF instances in 5G core*, IEEE Access **9** (2021), 81443–81453.
7. G. Chopra, *An efficient base station sleeping configuration for ultra-dense networks*, (International Conference on Emerging Smart Computing and Informatics, Pune, India), 2023. DOI [10.1109/ESCI56872.2023.10100245](https://doi.org/10.1109/ESCI56872.2023.10100245)
8. Y. A. Sambo, G. C. Valastro, G. M. Patané, M. Ozturk, S. Hussain, M. A. Imran, and D. Panno, *Motion sensor-based small cell sleep scheduling for 5G networks*, (IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, Limassol, Cyprus), 2019. DOI [10.1109/CAMAD.2019.8858435](https://doi.org/10.1109/CAMAD.2019.8858435)
9. G. Femenias, N. Lassoued, and F. Riera-Palou, *Access point switch on/off strategies for green cell-free massive MIMO networking*, IEEE Access **8** (2020), 21788–21803.
10. G. Lee, W. Saad, M. Bennis, A. Mehdodniya, and F. Adachi, *Online ski rental for ON/OFF scheduling of energy harvesting base stations*, IEEE Trans. Wireless Commun. **16** (2017), no. 5, 2976–2990.
11. H. Celebi, Y. Yapici, I. Güvenç, and H. Schulzrinne, *Load-based on/off scheduling for energy-efficient delay-tolerant 5G networks*, IEEE Trans. Green Commun. Netw. **3** (2019), no. 4, 955–970.
12. I. Leyva-Pupo, C. Cervell-Pastor, C. Anagnostopoulos, and D. P. Pezaros, *Dynamic scheduling and optimal reconfiguration of UPF placement in 5G networks*, Proceedings of ACM MSWiM 2020, Association for Computing Machinery, New York, NY, USA, 2020.



13. J. Luo, Q. Chen, and L. Tang, *Reducing power consumption by joint sleeping strategy and power control in delay-aware C-RAN*, *IEEE Access* **6** (2018), 14655–14667.
14. S. Woo, J. Sherry, S. Han, S. Moon, S. Ratnasamy, and S. Shenker, *Elastic scaling of stateful network functions*, (Proceedings of USENIX NSDI 2018, USENIX, Renton, WA, USA), 2018.
15. Y. Lee, K. Miyanabe, H. Nishiyama, N. Kato, and T. Yamada, *Threshold-based RRH switching scheme considering baseband unit aggregation for power saving in a cloud radio access network*, *IEEE Syst. J.* **13** (2019), no. 3, 2676–2687.
16. Z. Wu, Y. Zhang, W. Feng, and Z. L. Zhang, *NFlow and MVT abstractions for NFV scaling*, (Proceedings of IEEE INFOCOM 2022-IEEE Conference on Computer Communications, London, UK), 2022. DOI [10.1109/INFOCOM48880.2022.9796764](https://doi.org/10.1109/INFOCOM48880.2022.9796764)
17. T. Lin and A. Leon-Garcia, *Towards a client-centric QoS auto-scaling system*, (Proceedings of IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary), 2020. DOI [10.1109/NOMS47738.2020.9110450](https://doi.org/10.1109/NOMS47738.2020.9110450)
18. T. Subramanya and R. Riggio, *Centralized and federated learning for predictive VNF autoscaling in multi-domain 5G networks and beyond*, *IEEE Trans. Netw. Service Manag.* **18** (2021), no. 1, 63–78.
19. H. Ko, H. Lee, T. Kim, and S. Pack, *LPGA: location privacy-guaranteed offloading algorithm in cache-enabled edge clouds*, *IEEE Trans. Cloud Comput.* **10** (2022), no. 4, 2729–2738.
20. H. Ko, S. Pack, and V. C. Leung, *Spatiotemporal correlation-based environmental monitoring system in energy harvesting Internet of Things (IoT)*, *IEEE Trans. Ind. Inform.* **15** (2019), no. 5, 2958–2968.
21. H. Kang, X. Chang, J. Mišić, V. B. Mišić, J. Fan, and J. Bai, *Improving dual-UAV aided ground-UAV bi-directional communication security: joint UAV trajectory and transmit power optimization*, *IEEE Trans. Veh. Technol.* **71** (2022), no. 10, 10570–10583.
22. H. Ko, S. Pack, and V. C. Leung, *Performance optimization of serverless computing for latency-guaranteed and energy-efficient task offloading in energy harvesting industrial IoT*, *IEEE Internet Things J.* **10** (2023), no. 3, 1897–1907.
23. M. M. Moghaddam, M. H. Manshaei, M. N. Soorki, W. Saad, M. Goudarzi, and D. Niyato, *On coordination of smart grid and cooperative cloud providers*, *IEEE Syst. J.* **15** (2021), no. 1, 672–683.

## AUTHOR BIOGRAPHIES



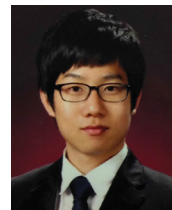
**Seokwon Jang** received the BS and PhD degrees from the School of Electrical Engineering, Korea University, Seoul, Republic of Korea, in 2015 and 2022, respectively. He is currently a senior researcher at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His research interests include mobile networking, data-center networking, programmable data planes, P4, SDN, and Net4AI.



**Namseok Ko** received the MS and PhD degrees from KAIST, Republic of Korea, in 2000 and 2015, respectively. He is currently the director of the Mobile Core Network Research Section of the Electronics and Telecommunications Research Institute (ETRI). He is also an associate professor at the Department of Information and Communication Engineering of the University of Science and Technology. He serves as the vice chair of SG11 and rapporteur of Q20 of SG13 at ITU-T. He also serves as the vice chair of the Technology Committee and the chair of the Network Technology Work Group at 6G Forum, Republic of Korea. He has participated in various research and development projects, including developments in 5G mobile core network technologies, since joining ETRI in 2000. Currently, he is leading several projects related to 6G network architecture. His research interests include 5G/6G mobile core network architecture and its enabling technologies, for example, supporting network programmability, convergence of networking and computing, and non-terrestrial networks.



**Jaewook Lee** received the BS and PhD degrees from the School of Electrical Engineering, Korea University, Seoul, Republic of Korea, in 2014 and 2021, respectively. He is currently an assistant professor at the Department of Information and Communication Engineering, Pukyong National University, Busan, Republic of Korea. From 2021 to 2023, he was a senior researcher at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His research interests include 6G mobile networks, federated learning, network automation, and time-sensitive networking.



**Yeunwoong Kyung** received the BS and PhD degrees from the School of Electrical Engineering, Korea University, Seoul, Republic of Korea, in 2011 and 2016, respectively. He is a staff engineer at Advanced CP Lab, Mobile Communications Business, Samsung Electronics, Japan. He is currently an assistant professor at the Division of Information and Communication Engineering, Kongju National University, Cheonan, Republic of Korea. His research interests include mobility management, mobile cloud computing, SDN/NFV, and IoT.



**Haneul Ko** received the BS and PhD degrees from the School of Electrical Engineering, Korea University, Seoul, Republic of Korea, in 2011 and 2016, respectively. He is currently an assistant professor at the Department of Electronic Engineering, Kyung Hee University, Yongin, Republic of Korea. From 2019 to 2022, he was an assistant professor at the Department of Computer and Information Science, Korea University, Sejong, Republic of Korea. From 2017 to 2018, he was a postdoctoral fellow at the University of British Columbia, Vancouver, Canada. From 2016 to 2017, he was a postdoctoral fellow in Mobile Networks and Communication at

Korea University, Seoul, Republic of Korea. He was the recipient of the Minister of Education Award in 2019 and IEEE ComSoc APB Outstanding Young Researcher Award in 2022. His research interests include 5G/6G networks, network automation, mobile cloud computing, SDN/NFV, and Future Internet.

**How to cite this article:** S. Jang, N. Ko, J. Lee, Y. Kyung, and H. Ko, *Joint user plane function instance and base station scheduling in mobile networks*, ETRI Journal **46** (2024), 977–986. DOI [10.4218/etrij.2023-0336](https://doi.org/10.4218/etrij.2023-0336)