IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 9, 2021, accepted May 29, 2021, date of publication June 16, 2021, date of current version July 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3089782

A Cooperative Online Learning-Based Load Balancing Scheme for Maximizing QoS Satisfaction in Dense HetNets

HYUNGWOO CHOI[1], (Student Member, IEEE), TAEHWA KIM[1], (Graduate Student Member, IEEE), HONG-SHIK PARK[2], (Member, IEEE), AND JUN KYUN CHOI[2], (Senior Member, IEEE)

[1]School of Information and Communication Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea
[2]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea

Corresponding author: Jun Kyun Choi (jkchoi59@kaist.edu)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant through the Korea Government (MSIT), Development of Autonomous Collaborative Swarm Intelligence Technologies for Disposable IoT Devices, under Grant 2018-0-00691.

ABSTRACT This paper proposes a cooperative multi-agent online reinforcement learning-based (COMORL) bias offset (BO) control scheme for cell range expansion (CRE) in dense heterogeneous networks (HetNets). The proposed COMORL scheme controls BOs for CRE to maximize the number of user equipments (UEs) that satisfy their quality of service (QoS) requirements, especially in terms of delay and data rates. For this purpose, we developed a QoS satisfaction indicator that measures a violation of delay requirements by considering both QoS requirements and signal-to-interference-plus-noise ratio (SINR). In addition, we formulated a Markov decision process (MDP) model that is solved with a cooperative multi-agent online reinforcement learning algorithm. The proposed COMORL scheme maximizes the global utility for load-coupled base stations. Our simulation results verify the proposed COMORL scheme's effectiveness in terms of throughput, delay satisfaction ratio, and fairness. Specifically, we verify that the proposed COMORL scheme achieves a maximum of approximately 27% and 30% improvement of the delay satisfaction ratio, which is how many UEs satisfy their delay requirement among all of the UEs in a serving BS under medium and full traffic loads, respectively, in a dynamic scenario in comparison to the max-SINR scheme.

INDEX TERMS Hetnets, cell range expansion, load balancing, QoS, cooperative multi-agent reinforcement learning.

I. INTRODUCTION

HetNets are one of the key enabling techniques for fifth-generation (5G) mobile networks. The employment of small cells is an inevitable trend because increasing the node deployment density is the only feasible way to improve spectral efficiency. In HetNets, however, the user equipment (UE) tends to connect to a macro base station (MBS) even though UEs are closer to a small base station (SBS) because the transmission power of an MBS is much larger than that of an SBS. The difference in the transmission power level between MBSs and SBSs causes low utilization of small cells.

Hence, enhanced inter-cell interference coordination (eICIC) has been standardized by the 3rd Generation Partnership Project (3GPP) release 10 to offload the traffic of UEs from MBSs to SBSs [1]. In eICIC, the almost blank subframe (ABS) technique prevents an adjacent MBS from interfering with cell-edge UEs in SBSs by muting subframes of the MBS while UEs in the SBSs transmit their data. Moreover, SBSs expand their coverage by adding a bias offset (BO) to the reference signal receive power (RSRP) of SBSs with the cell range expansion (CRE) technique so that more UEs near the cell edge can connect to SBSs.

In dense HetNets, CRE is a good alternative for user association (UA) in that it is a simple and effective technique for load balancing. Also, CRE can achieve near-optimal load-aware performance [2]. Load balancing in HetNets is closely related to UA, which determines which BS serves a particular UE. The problem of UA naturally falls into the scope of integer programming, of which the computational

The associate editor coordinating the review of this manuscript and approving it for publication was Chao-Yang Chen.

complexity is known as NP-complete because UA is the problem of mapping between BSs and UEs [3]. Most studies on UA have formulated the problem with stochastic geometry, game theory, and combinatorial optimization as surveyed in [4], and they can achieve optimal load balancing. However, the computational complexity becomes much higher as the number of network nodes increases in dense HetNets. Therefore, we focus on CRE, which is a more pragmatic approach than the previous UA schemes because of its operational simplicity for load balancing in dense HetNets.

To exploit small cells efficiently, it is essential to determine the appropriate BO of SBSs for offloading the traffic from MBSs to SBSs. Also, care is needed when choosing BOs in dense HetNets. The coverage adjustment of SBSs in the middle of dense HetNets has more impact on adjacent SBSs than in conventional HetNets because the average distance between SBSs decreases as the networks become denser. In other words, they are in load-coupled relations. For example, SBS(1) expands its coverage with a BO to offload UE(1) traffic from an MBS, as in Fig. 1. Meanwhile, the coverage expansion of SBS(1) also affects UE(2), which has a prior association with SBS(2). Consequently, SBS(2) has to adjust its BO to maintain its cell traffic load. SBSs have to choose their BOs by considering the traffic offload from not only an MBS but also neighboring SBSs. Therefore, it is necessary to consider cooperation between SBSs to configure the optimal BOs. SBSs must interact with their neighbors when deciding how much to expand their coverage with BO. However, early-stage studies on the optimal BO for CRE did not consider network densification sufficiently [5]–[9]. In this regard, we propose a BO control scheme based on cooperative multi-agent reinforcement learning (RL).

Determining the optimal BOs of SBSs with an RL-based approach enables self-organizing features. The self-organizing network (SON), which monitors changes in the environment and reconfigures network parameters, is essential as the number of network elements increases in dense HetNets, which are hard to configure manually. RL is well suited for enabling SONs because it adapts to changing environments by interacting and learning. It has been studied for many algorithms in various areas, as surveyed in [10]. Although RL helps to enable SONs, RL's feature based on a trial and error scheme may harm the network performance. To minimize performance loss when applying RL-based algorithms to our proposed scheme, we introduce a handover strategy to reduce unnecessarily frequent handovers.

In [7], UE determines the BOs for SBSs in a distributed manner using RL to minimize outages. It is still valid in dense HetNet because UEs only have to consider nearby SBSs because learning is done from the UE side. However, it lacks consideration of UE's quality of service (QoS) requirement because the minimization of outages was the primary consideration. Especially in the circumstances in which the emerging services in 5G mobile networks such as virtual reality, augmented reality, health-care, entertainment, intelligent transportation, and factory automation, have various

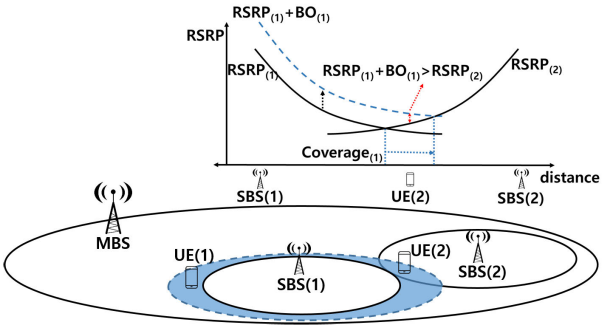


FIGURE 1. An example of cell range expansion in dense HetNets.

QoS requirements in terms of delay and data rates [11], a model that considers various individual QoS requirements is needed. Therefore, we define an indicator for our proposed scheme to measure QoS satisfaction according to individual QoS requirements.

Besides, some studies have attempted to optimize BO while considering other performance metrics together. In [12], a coordinated CRE for mobility management strategy was introduced, which considers the cell load, QoS of UEs, and inter-cell interference with the maximum throughput scheduling technique. However, the handovers between SBSs in dense HetNets was not appropriately handled. In [13], the authors proposed a BO-optimization algorithm based on Gibbs-sampling. However, in dense HetNets, the computational complexity becomes a critical issue as the number of SBSs increases. It is necessary to share a more substantial volume of information on UEs and SBSs and search all possible combinations of BOs for the SBSs to find the optimal BOs in dense HetNets.

There have been studies based on stochastic geometry to find the optimal BOs [14]–[16]. These approaches can provide useful analytical insight into how to set BOs. However, they cannot cope with the network dynamics in real networks, such as ever-changing wireless channel quality and UE mobility. Moreover, it is difficult for them to consider the load-coupled relations between SBSs in dense HetNets.

The main contributions of the paper are listed as follows:

- This paper proposes a cooperative multi-agent online reinforcement learning-based (COMORL) BO control scheme for CRE in dense HetNets to overcome the limitations of existing approaches. The proposed COMORL scheme aims to maximize the number of UEs that satisfy their QoS requirements.
- To achieve this, firstly, a QoS satisfaction indicator (QSI) is newly defined to evaluate the QoS satisfaction of the UE considering the statistical delay requirement, data rates, and SINR of the UE. Then, a utility function is defined using the statistics of the QSI for BSs. Also, a new handover strategy based on the QSI is introduced to minimize the side effects of RL-based algorithms.
- To determine the optimal BO for maximizing QoS satisfaction while considering the situation of neighboring

BSs, we newly formulate a Markov decision process (MDP) model for our COMORL scheme and propose a cooperative multi-agent online reinforcement learning algorithm based on a message-passing approach.
- Through extensive simulations, the proposed COMORL scheme achieves notable performance improvement in terms of throughput, delay satisfaction ratio, and fairness. In particular, the proposed COMORL scheme achieves a maximum of approximately 27% and 30% improvement of the delay satisfaction ratio under medium and full traffic loads, respectively, in a dynamic scenario, compared to the max-SINR scheme (without cell range expansion).

The rest of the paper is organized as follows. Section II briefly reviews related works. Section III explains the effective capacity (EC) link-layer model used in our proposed scheme. In Section IV, we present the COMORL scheme, which consists of a CRE execution module and a cooperative multi-agent Q-learning module. Section V details the performance evaluation of the proposed COMORL scheme. Finally, the conclusions are drawn in Section VI.

II. RELATED WORKS

A mobility load-balancing (MLB) and CRE have a similar principle in that they adjust the cell offset for load balancing. MLB was introduced as an example of a SON to re-distribute the load by optimizing cell reselection/handover parameters [17]. In [18], the authors proposed an adaptive MLB algorithm in small-cell networks. They adopted an adaptive threshold to find overloaded cells and restricted the load released from the overloaded cells. The load is effectively distributed by estimating the load status of currently overloaded cells and candidate target cells. In [19], the authors proposed a two-layer MLB architecture in which the top layer dynamically groups SBSs into self-organized clusters according to their historical loads and the bottom layer balances the intra-cluster load distribution using a deep-reinforcement learning-based algorithm. For stability, they designed a mechanism that works online for system control and learns policies offline. In [20], a cluster-based load-balancing algorithm was proposed for ultra-dense heterogeneous networks. Further improvement in network performance was achieved in comparison to previous MLB schemes by constructing clusters dynamically and performing load balancing locally. Thus unnecessary MLB operations across the networks are avoided as well. However, these studies measured the cell load with physical resource block utilization, taking into account only the data rate requirements of UEs, and lacking consideration of delay QoS requirements.

Ever since eICIC was standardized by 3GPP release 10, there have been many studies to optimize BOs for CRE. We will review the most recent and relevant studies in this section. In [7], UE learns its BO for nearby BSs that minimizes the number of outage UEs by using the Q-learning algorithm. The states for Q-learning are defined as the received powers of the pilot signals from BSs, and the actions for Q-learning are defined as BOs. The cost for Q-learning is the number of UEs that cannot get radio service. Thus, each UE can learn the BOs for the received power of nearby BSs by Q-learning. However, this approach does not consider UE's QoS requirement. Moreover, learning on the UE-side can be a burden for mobile devices.

In [12], a coordinated CRE for mobility management strategy was introduced. It analytically computes the joint optimal BOs at SBSs and MBSs. A combined objective function was designed, which considers multiple factors, such as the cell load, QoS of UEs, and interference limitation, and utility functions for MBSs and SBSs were formulated from the combined objective function. Also, they introduced the maximum throughput scheduling technique. However, they did not sufficiently consider dense HetNet environments.

In [13], the authors proposed BO adjusting algorithms based on Gibbs-sampling. The introduced various versions of the algorithm: centralized, distributed, and central-aided distributed. The centralized algorithm has to know all information at the central processor, resulting in significant message exchange overhead and computational complexity. The distributed algorithm was proposed to deal with the complexity problem. To further reduce both the message exchange overhead and the time complexity, they proposed the central-aided distributed algorithm with a graph-coloring-based clustering method. However, in Gibbs sampling, it is still a burden to search all possible combinations of BOs for the SBSs to find the optimal BOs in dense HetNets.

In [21], a particle swarm optimization (PSO) algorithm was utilized to find optimal BOs for all SBSs. The authors formulated a user association problem to maximize the number of UEs with fulfilled downlink requirements and the number of BSs associated with users. At first, the particles are randomly scattered in the search space, and then they move according to PSO update equations until the end condition. With the proposed PSO algorithm, it can expect to balance the network load without the resolution of the combinatorial optimization problem.

In recent years, stochastic geometry has attracted attention for the analysis of the optimal BO for CRE. In [14], the authors jointly optimized the BO and the density of SBSs to maximize energy efficiency (EE). They analytically derived the closed-form expression of the network EE as a function of the density of SBSs and BO based on stochastic geometry theory. Also, the joint optimization algorithm was introduced for joint optimization of the density of SBSs and BO. In [15], a CRE-based association for massive multiple-input-multiple-output (MIMO) HetNets was proposed. The key feature is that they used the user's long-term perceived rate instead of the instantaneous rate to capture both the multi-antenna mode and the BS load. Also, the proposed closed-form bias factors can approximate the expected long-term rate instead of evaluating the bias factors by simulations. In [16], the authors optimized local delay and

TABLE 1. Brief description of acronyms.

Notation	Description
$\mathcal{B}$	Set of BSs
$\mathcal{U}_i$	Set of UEs served by BS i
δ_i	Bias offset of BS i
p_i^{rsrp}	RSRP of BS i
$SINR_{ik}$	SINR of UE k served by BS i
$R_{i,k}$	Service rate of UE k at BS i
θ_k	QoS exponent of UE k
λ_k	Data rate of UE k
D_k^{max}	Maximum delay bound of UE k
ε_k	Delay violation probability of UE k
$C_{ik,j}^M$	Measured capacity of UE k at BS i for j-th frame
$C_{ik,j}^E(\cdot)$	Effective capacity of UE k at BS i for j-th frame
f_{ik}	QoS satisfaction indicator of UE k at BS i
$\mu_i(\delta_i)$	Mean of QoS satisfaction indicator of BS i
$\sigma_i^2(\delta_i)$	Variance of QoS satisfaction indicator of BS i
$\psi_i(\delta_i)$	Utility function of BS i
$\mathbf{\Psi}(\mathbf{\Delta})$	Global utility
$\mathbf{\Delta}$	Vector of BOs for all BSs
$\mathbf{\Delta}^*$	Otimal BO vector
I_{μ_i}	Index for mean of BS i
$I_{\sigma_i^2}$	Index for variance of BS i
$\nu_{ij}(\delta_j)$	Message value from BS i to BS j
$g_i(\delta_i)$	Contribution of BS i to the global Q-funcion

EE with CRE. The analysis based on stochastic geometry can provide useful analytical insight into how to set BOs. However, these studies cannot cope with network dynamics in real networks, such as ever-changing wireless channel quality and UE mobility.

III. EFFECTIVE CAPACITY MODEL

We consider a downlink (DL) of two-tier HetNets, in which a single macrocell is overlaid with small cells. Let $\mathcal{B} = \{b_i\}, i = 0, 1, \ldots, B$ denote a set of BSs in which b_0 is an MBS and is overlaid with B SBSs, and let $\mathcal{U}_i$ denote a set of UEs served by BS i. The UEs select the serving BS b_i^{serve} based on the RSRP p_i^{rsrp} and the BO δ_i of BS i as

$$b_i^{serve} = \underset{i\in\mathcal{B}}{\text{argmax}}(p_i^{rsrp} + \delta_i). \tag{1}$$

Thus, the UE association and the traffic load of each BS depend on the BO δ_i.

We assume Rayleigh block fading for the subchannels and orthogonal frequency-division multiple access (OFDMA) systems. The SINR of UE k served by the BS i can be given by

$$SINR_{ik} = \frac{p_i \cdot g_{ik}}{\sum_{h\in\mathcal{B},h\neq i} p_h \cdot g_{hk} + n_0}, \tag{2}$$

where p_i is the transmission power of BS i, g_{ik} is the channel gain from BS i to UE k, and n_0 is the noise power. The service rate R_{ik} of the UE k, which can be supported by the associated BS i in a frame can be given as [22]

$$R_{ik} = \frac{WT_f}{MN}\sum_{n=1}^{N} s_{k,n}\log_2(1 + SINR_{ik}), \tag{3}$$

where W is the spectral bandwidth, T_f is the frame duration, M is the number of slots, N is the number of subchannels, and $s_{k,n}$ is the service indicator that denotes whether UE k is served in subchannel n.

We use the EC model as a criterion to judge whether the delay QoS requirement of a UE can be satisfied or not in an associated BS. The EC model was initially introduced in [23] as a link-layer channel model that characterizes wireless channels in terms of the statistical delay requirement. It measures the maximum constant arrival rate under the delay-bound violation probability constraint for a given channel capacity. The EC model helps translate both SINR and delay bounds parameters into throughput.

The EC of UE k served by BS i is given by [22]

$$C_{ik}^E(SINR_{ik}, \theta_k) = -\frac{1}{\theta_k T_f}\log(\mathbb{E}\{e^{-\theta_k R_{ik}}\}), \tag{4}$$

where θ_k is the QoS exponent of UE k. Here, θ_k is expressed by a QoS triplet $(\lambda_k, D_k^{max}, \varepsilon_k)$, which represents the traffic attribute of UE with a traffic arrival rate λ_k, a maximum delay bound D_k^{max}, and a delay violation probability ε_k. It can be calculated as $\theta_k = -\log\varepsilon_k/\lambda_k D_k^{max}$ from the approximation of delay bound violation probability

$$Pr\{D_k(\infty) > D_k^{max}\} \leq \varepsilon_k \approx e^{-\theta_k\lambda D_{max}}, \tag{5}$$

where $D_k(\infty)$ is the steady-state delay experienced by the traffic flow of UE k.

A large QoS exponent means a stringent QoS requirement, so the arrival rate of UE that a wireless channel can support decreases. In other words, EC decreases with increasing QoS exponent θ_k and vice versa. Thus, EC is upper bounded by the average service rate (Shannon capacity) as $\theta \to 0$, and lower bounded by the minimum service rate 0 as $\theta \to \infty$.

In the remainder of this paper, we assume that BSs have the QoS triplets of UEs in advance through the connection establishment process.

IV. PROPOSED BIAS CONTROL SCHEME

In this section, we propose a cooperative multi-agent online reinforcement learning-based (COMORL) scheme for BO control to maximize the number of UEs that satisfy their delay QoS requirement. If an SBS adjusts its BO, it influences the number of associated UEs and other BSs in dense HetNets. This kind of system is called a multi-agent system, where one agent's behavior affects the other agents' actions. They should interact with each other and learn the behavior from the environment to solve the problem; this is called multi-agent learning.

A. OVERVIEW OF THE PROPOSED SCHEME

The proposed scheme utilizes multi-agent cooperative RL. To utilize RL, we have to model the problem using the MDP model so that RL works with defined states, actions, and rewards. In addition, because we consider a multi-agent system, we need a tool by which RL can determine the optimal BO considering neighboring BSs decisions. First, a coordination graph (CG) [24] is a useful tool to show the relation for cooperation in a graph. The CG starts from the fact that an

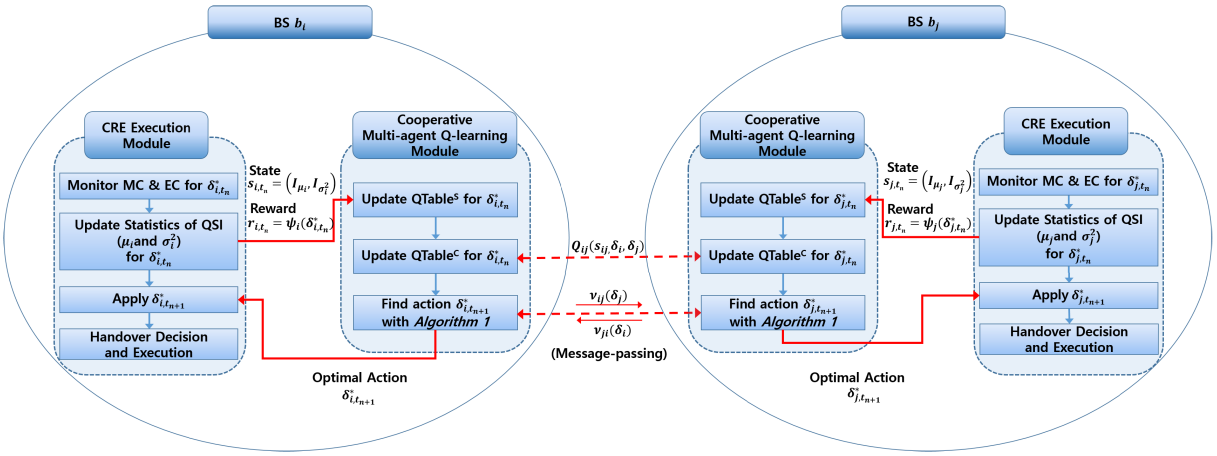


FIGURE 2. Overall process for a cycle t_n of the proposed bias control scheme.

agent does not cooperate with all agents in a multi-agent system. Instead, it only cooperates with adjacent agents (namely sparsity). Also, this fact can make a large problem be divided into small problems. One of these approaches using the CG is sparse cooperative Q-learning [25], which decomposes the problem into smaller ones. We adopt edge-based decomposition for the proposed scheme.

Meanwhile, the BSs (agents) operate $QTable^S$ for their own and $QTable^C$ for cooperation. They determine the optimal BO with Q-values in QTables while considering the BOs of adjacent BSs. Thus, BSs determine their best BO with neighbors by the message-passing algorithm. The messages include the best outcome for neighbors' BOs. After the message exchanges converge or a timer expires, BSs have the optimal BO. If BSs apply the optimal BO, handover can be carried out. However, because the BOs keep changing during the learning process, unnecessary handovers may be triggered, which results in performance degradation. Therefore, we introduce our handover strategy, which limits the handover of UEs satisfying their QoS requirements.

The proposed COMORL scheme consists of a CRE execution (CE) module and a cooperative multi-agent Q-learning (CMQ) module. The CE module applies the optimal BOs and monitors the QSI of UEs in the cell. Also, the CE module passes states and rewards for the applied BO to the CMQ module. Meanwhile, the CMQ module conducts Q-learning-related tasks. It determines the optimal BO and passes it to the CE module. In addition, it manages the Q-table and infers the optimal BO through cooperative Q-learning based on message passing between adjacent BSs. Each BS maintains the $QTable^S$ for itself and the $QTable^C$ for coordination of its neighbor BSs. The $QTable^S$ includes states, actions, and Q-values for its own, and the $QTable^C$ has the joint states, actions, and Q-values of its neighbors. The proposed scheme periodically iterates to find the optimal BO $\delta^*_{t_n}$ and applies it. The overall operation process for adjusting the BOs of BSs is illustrated in Fig. 2, and the following subsection presents details of each module. Note that we use RL and Q-learning interchangeably because Q-learning is a representative model-free reinforcement learning algorithm.

B. CRE EXECUTION MODULE

The CE module describes the status of the BS with the QSI of each serving UE, which is defined as follows.

Definition: Consider that BS i serves UE k and measures a measured capacity (MC) $C^M_{ik,j}$ and the EC $C^E_{ik,j}(\cdot)$ of UE k at j-th frame during m frames. Then, the QSI is defined as

$$f_{ik} = \frac{1}{m}\sum_{j=1}^{m} \frac{C^M_{ik,j}}{C^E_{ik,j}(SINR_{ik,j}, \theta_{ik})}, \tag{6}$$

where $SINR_{ik,j}$ is the SINR of UE K at the j-th frame of BS k, and θ_{ik} is the QoS exponent of UE k at BS i. MC is the actual average service rate of UE k during a measured period measured by the serving BS, and EC means the minimum required service rate to guarantee the statistical delay QoS requirement of UE k. Thus, the QSI indicates whether the UE's delay QoS requirement is guaranteed or not at the serving BS. Also, the proposition related to the QSI is presented as follows.

Proposition: If BS i has provided enough capacity for UE k to satisfy its delay QoS requirement during m frames, then $f_{ik} \geq 1$; if not, $f_{ik} < 1$.

In terms of the queuing theory, if the service rate is higher than the minimum service rate for guaranteeing the delay requirement, it is clear that the queuing delay is less than the delay requirement.

The QSI can support various emerging services with various delay and data rate requirements such as virtual reality, health care, factory automation, and smart grid because the QoS exponent of EC is determined by the QoS triplet. Moreover, if we are only concerned about delay, it would be better to directly monitor the packet queuing delay. However, using the QSI, we can estimate the guarantee of delay requirement by considering both the SINR and the degree of resource competition according to the density of UE, so it is possible to understand the offloading effect according to a BO.

The CE module works periodically. It monitors MC and EC during a cycle after the optimal BO is applied and the handover is done. We suppose that each BS has the statistics of the QSI according to the BO δ_i for the set $\mathcal{U}_i$ of UEs at BS i as

$$\mu_i(\delta_i) = \frac{1}{|\mathcal{U}_i|} \sum_{k \in \mathcal{U}_i} f_{ik}(\delta_i), \tag{7}$$

$$\sigma_i^2(\delta_i) = \frac{1}{|\mathcal{U}_i|} \sum_{k \in \mathcal{U}_i} \{f_{ik}(\delta_i) - \mu_i(\delta_i)\}^2, \tag{8}$$

where $\mu_i(\delta_i)$ is the mean, and $\sigma_i^2(\delta_i)$ is the variance of QSI for BO δ_i of BS i, and $|\mathcal{U}_i|$ is the cardinality of the set $\mathcal{U}_i$. These statistics consist of the state of BSs for the MDP model presented in the following subsection. In addition, we define the utility of BS i with the statistics of the QSI as described below.

Definition: Consider that the BS i applies the BO δ_i, and $\mu_i(\delta_i)$ and $\sigma_i^2(\delta_i)$, the statistics of the QSI for $\mathcal{U}_i$ is given. Then the utility function of BS i is defined as

$$\psi_i(\delta_i) = \mu_i(\delta_i) - \frac{\rho}{2}\sigma_i^2(\delta_i) \quad \text{for} \quad 0 \le i \le B, \tag{9}$$

where $0 \le \delta_i \le \delta^{\max}$, and $\rho > 0$ is the risk aversion parameter of the conventional mean-variance utility function, which is one of the tenets in rational decision making under risk [26]. Here, ρ determines the sensitivity of the QSI's variance to the utility. As ρ increases, the rate of decrease in utility by variance increases, and vice versa.

Note that $\psi_i(\delta_i)$ is the function of the BO δ_i because the UEs' association and the statistics of BS i change as δ_i changes.

To evaluate the global utility and the contribution of each BS to it, the global utility is defined in an additive manner as follows.

Definition: Consider that the vector of BOs for all BSs denoted by $\Delta = (\delta_0, \delta_1, \ldots, \delta_B)$ is given, and each BS applies its BO δ_i. The global utility of the entire system is defined as

$$\Psi(\Delta) = \sum_{i \in \mathcal{B}} \psi_i(\delta_i). \tag{10}$$

Theorem (Existence of Δ^*): There is an optimal BO vector $\Delta^* = (\delta_0^*, \delta_1^*, \ldots, \delta_B^*)$ that maximizes the global utility (10) of the system,

$$\Delta^* = \underset{\Delta}{\operatorname{argmax}}\, \Psi(\Delta). \tag{11}$$

Proof: Actually, the utility function (9) can transform to the type of $\psi_i(x_i, y_i) = x_i - \frac{\rho}{2}y_i^2$. Then, the Hessian matrix of the function is given by $\mathbf{H}_{\psi_i}(x_i, y_i) = \begin{pmatrix} 0 & 0 \\ 0 & -\rho \end{pmatrix}$. Because the first principle minors are 0 and $-\rho$, which are ≤ 0, and the second principle minor is 0, the function $\psi_i(x_i, y_i)$ is concave on $\mathbb{R}^2$. Furthermore, because the sum of concave functions is itself concave, the global utility function $\Psi(\Delta)$ is also concave. Hence, there exist a maximum value of $\Psi(\Delta)$ and the arguments that maximize the function. ■

If the CMQ module determines the optimal BO vector, the CE module applies the BO for each SBS and executes the handover process. However, the CE module may produce unnecessarily frequent handovers, which result in the degradation of throughput and delay performance when the BO is applied in every iteration of learning. Therefore, we adopt a handover strategy to minimize the side effects. The CE module decides whether to invoke handover or not to prevent frequent handovers during operation. If there is a BS with better signal strength than the current serving BS for a UE by (1), the CE module executes handover only when the current serving BS cannot guarantee the delay QoS requirement of a UE, $f_{ik} < 1$ from (6). Limiting handover based on the delay satisfaction of UEs can reduce performance degradation that can occur with a trial and error scheme.

C. COOPERATIVE MULTI-AGENT Q-LEARNING MODULE

To find the optimal BO vector (11), we formulate the problem as an MDP model with the previously defined QSI and utility function to apply RL. Then we propose a message-passing-based cooperative multi-agent Q-learning algorithm, which can find the optimal BO vector (11) while considering the influence on neighboring BSs for each BS.

1) MARKOV DECISION PROCESS MODEL

We model the RL problem in the form of an MDP, a mathematical framework for modeling decision making. It consists of 4-tuple (S, A, P, R), where S is the finite set of states, A is the finite set of actions, P is the state transition probability, and R is the reward function.

1) State: the state of BS i is the joint vector of I_μ and I_σ

$$s_i = \left(I_{\mu_i}, I_{\sigma_i^2}\right), \tag{12}$$

where

$$I_{\mu_i} = \begin{cases} 0, & \text{if } \mu_i < 1 - \omega_\mu \\ 1, & \text{if } 1 - \omega_\mu \le \mu_i \le 1 + \omega_\mu \\ 2, & \text{if } \mu_i > 1 + \omega_\mu \end{cases},$$

$$I_{\sigma_i^2} = \begin{cases} 0, & \text{if } \sigma_i^2 \le \omega_{\sigma^2} \\ 1, & \text{if } \sigma_i^2 > \omega_{\sigma^2} \end{cases},$$

and ω_μ and ω_{σ^2} are the tuning parameters for determining the optimality of BSs. Here, I_{μ_i} indicates the average delay satisfaction level of UEs at the BS i. Note that $I_{\mu_i} = 0$ means that UEs at the BS i receive poor service on average, whereas $I_{\mu_i} = 2$ means that UEs at the BS i receive good service, and $I_{\mu_i} = 1$ means that UEs receive adequate service. The possible factors affecting I_{μ_i} are the SINR of UEs or UEs' density at the BS i, which may cause resource competition. Here, $I_{\sigma_i^2}$ represents the index of the balance, which indicates whether the load is balanced or not in terms of the variance of QSI. The variance of QSI indicates the level of fairness among UEs in the cell. If all of the UEs in the cell satisfy their QoS requirements at a similar

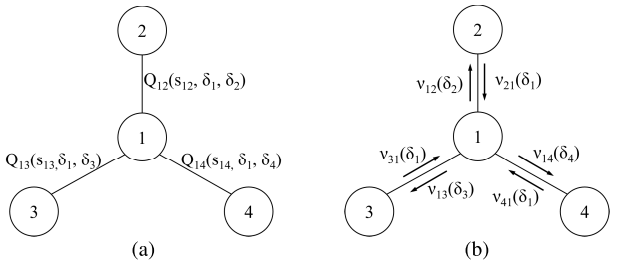


FIGURE 3. Examples of (a) edge-based decomposition (b) message v_{ij}.

level, the variance of QSI approaches zero. Thus we conceive $I_{\sigma_i^2} = 0$ as the more balanced state. Thus, the combination of I_{μ_i} and $I_{\sigma_i^2}$ can represent a total of 6 states of the BS i, and the state $(1, 0)$ is the most well-balanced state.

2) Action: the action of BS i is defined as the n-step BOs $\delta_i = \{\delta_i^0, \delta_i^1, \ldots, \delta_i^n\}$. For example, if the BO of the BS i varies from 0 dB to 16 dB, then $\delta_i = \{0, 1, \ldots, 16\}$. Note that $\delta_0 = \{0\}$ because MBS does not adjust its BO. We use actions and BOs interchangeably in the rest of the paper.
3) State Transition Probability: state transition probabilities are difficult to model because they depend on various factors in real environment dynamics, such as UE mobility, channel states, and data rates. However, RL can find a solution for the MDP without explicitly specifying the state transition probability by a trial-and-error approach. Therefore, we adopt the RL-based algorithm to solve this MDP, and we explain it in detail in the next subsection.
4) Reward: the reward of BS i, $R_i(\mathbf{S}, \Delta)$ is determined by the utility function (9) that is achieved after all BSs apply their actions Δ in the global states $\mathbf{S} = (s_0, s_1, \ldots, s_B)$.

2) MESSAGE-PASSING-BASED COOPERATIVE Q-LEARNING ALGORITHM

Our approach is motivated by sparse cooperative Q-learning (SparseQ) [25], which approximates the global Q-function into a linear combination of local Q-functions. The CG takes advantage of the sparsity that only a few agents depend on each other in many multi-agent problems [24]. Thus, the problems can be decomposed into simpler subproblems with the CG. In this regard, based on the fact that the influence of adjusting the BO of BSs is limited to adjacent BSs, the CG can represent BSs' interdependency where the vertices of the CG stand for BSs, and the edge means the dependency between adjacent BSs. Thus, with a given CG, we can decompose the global Q-function into the local Q-function, and it enables the BSs to update their actions and rewards locally while considering the action of neighboring BSs jointly.

There are two ways to perform decomposition, namely, agent-based and edge-based decomposition [25]. A agent-based decomposition must consider the dependency for all actions of connected vertices in CG. Thus, the representation of the local Q-function in agent-based decomposition becomes more complex as the number of neighbors increases. In contrast, the local Q-function depends on only the actions of two agents (vertices) over the edge in edge-based decomposition, and it scales linearly in the number of neighbors. Therefore, it is suitable for dense small cells that may have many neighbors.

The CMQ module manages Q-Tables: the $QTable^S$ is used for the local state and action, whereas the $QTable^C$ is used for the joint state and action with neighboring BSs. For the $QTable^S$, the Q-function of BS i is defined and updated as

$$Q_i(s_i, \delta_i) \leftarrow Q_i(s_i, \delta_i) + \alpha\{R_i(\mathbf{S}, \Delta) + \gamma Q_i(s_i', \delta_i^*) - Q_i(s_i, \delta_i)\}, \quad (13)$$

where $R_i(\mathbf{S}, \Delta)$ is the reward of BS i, α is the learning rate, γ is the discount factor of Q-learning and $Q_i(s_i', \delta_i^*)$ is the Q-function for the optimal action δ_i^* in the next state s'.

If a coordination graph $G = (V, E)$ with $|V|$ vertices and $|E|$ edges that represents the cooperative relations between BSs is given, an edge-based local Q-function can decompose the global Q-function as

$$Q(\mathbf{S}, \Delta) = \sum_{(i,j)\in E} Q_{ij}(s_{ij}, \delta_i, \delta_j), \quad (14)$$

where the local Q-function $Q_{ij}(s_{ij}, \delta_i, \delta_j)$ is defined on the edge $(i, j) \in E$ and depends on the actions (δ_i, δ_j) of BSs i and j as in Fig. 3 (a), and $s_{ij} \subseteq (s_i \cup s_j)$ is the subset of the state of BS i and j.

To compute the local Q-function Q_{ij}, we assume that the contribution of Q_i and Q_j to Q_{ij} is proportional to their number of neighbors $|\Gamma(i)|$ and $|\Gamma(j)|$,

$$Q_{ij}(s_{ij}, \delta_i, \delta_j) = \frac{Q_i(s_i, \delta_i)}{|\Gamma(i)|} + \frac{Q_j(s_j, \delta_j)}{|\Gamma(j)|}. \quad (15)$$

Then, the update equation of Q_{ij} can be given by

$$Q_{ij}(s_{ij}, \delta_i, \delta_j) \leftarrow Q_{ij}(s_{ij}, \delta_i, \delta_j) + \alpha\left\{\frac{R_i(\mathbf{S}, \Delta)}{|\Gamma(i)|} + \frac{R_j(\mathbf{S}, \Delta)}{|\Gamma(j)|} + \gamma Q_{ij}(s_{ij}', \delta_i^*, \delta_j^*) - Q_{ij}(s_{ij}, \delta_i, \delta_j)\right\}, \quad (16)$$

where $Q_{ij}(s_{ij}', \delta_i^*, \delta_j^*)$ is the Q-function for the optimal joint action of δ_i^* and δ_j^* in the next state s_{ij}'. After the CE module updates the statics of QSI for the BO δ_i^* and shares them with neighboring BSs, the CMQ module updates the $QTable^S$ and $QTable^C$.

To find the optimal joint action of BSs, we propose a hybrid message-passing-based algorithm based on the max-product algorithm for finding the maximum a posteriori (MAP) [27]. The CMQ module in each BSs must decide the optimal BO with consideration of its $QTable^S$ and $QTable^C$ as well as neighboring BSs' $QTable^S$ and $QTable^C$. Thus, it finds the optimal joint action Δ^* that maximizes (14) by repeatedly

Algorithm 1 Find Optimal BOs Using Hybrid Message-Passing Based Algorithm

Require: Coordination Graph $G = (V, E)$, $QTable_i^S$, $QTable_i^C$ for BSs
Send an initial message value $\nu_{ij}(\delta_j)$ to all neighbors
while *Timer* $<$ *Timeout* **do**
  if Receive a message value $\nu_{ji}(\delta_i)$ **then**
    for Neighbor $j \in \Gamma(i)$ in CG **do**
      Compute $\nu_{ij}(\delta_j) \cdots (17)$
      if $\nu_{ij}(\delta_j) \neq$ previously sent message value **then**
        Send $\nu_{ij}(\delta_j)$
      end if
    end for
    $\delta_i' = \text{argmax}_{\delta_i} g_i(\delta_i) \cdots (18)$
  end if
end while
Every SBSs send the $g_i(\delta_i')$ to the MBS
MBS sends $Q(\Delta') = \frac{1}{2(|B|+1)} \sum_{i=0}^{B} g_i(\delta_i')$ to SBSs
if $Q(\Delta') > Q^{prev}(\Delta^*)$ **then**
  $\delta_i^* = \delta_i'$, $Q^{prev}(\Delta^*) = Q(\Delta')$
end if

sending a message value $\nu_{ij}(\delta_j)$ to its neighbors as in Fig. 3(b). For simplicity, we omit the states of BSs:

$$\nu_{ij}(\delta_j) = \max_{\delta_i} \left\{ Q_i(\delta_i) + Q_{ij}(\delta_i, \delta_j) + \sum_{h \in \Gamma(i) \setminus j} \nu_{hi}(\delta_i) \right\} - c_{ij}, \quad (17)$$

where $\Gamma(i) \setminus j$ is the set of neighbors except j, and to prevent divergence in the case of a cyclic graph, the normalization parameter is defined as $c_{ij} = \frac{1}{|\Gamma(i)|} \sum_{k \in \Gamma(i)} \nu_{ik}(\delta_k)$, which is the average value of ν_{ik}. The contribution of BS i to the global Q-function is defined by

$$g_i(\delta_i) = Q_i(\delta_i) + \sum_{j \in \Gamma(i)} \nu_{ji}(\delta_i), \quad (18)$$

where $\Gamma(i)$ is the set of neighbors of BS i, and $\nu_{ji}(\delta_i)$ represents the sum of the local payoff that BS j receives from the neighbors except BS i when BS i selects the action δ_i. After exchanging the message values, each BS i selects the action δ_i that maximizes (18).

The detailed algorithm is shown in Algorithm 1. We assume that the coordination graph of BSs is known in advance. The CMQ module sets the timeout to be shorter than the iteration cycle time of the proposed COMORL scheme. Until a timeout occurs, it continues exchanging message values (17) with neighboring BSs, whose relations are represented by the coordinated graph G. After an initial message value $\nu_{ij}(\delta_j)$ is sent to neighbors on G, successive message exchanges are invoked, and this continues until a timeout occurs. When BSs receive a message value $\nu_{ji}(\delta_i)$, they should send a message value $\nu_{ij}(\delta_j)$, which is the response to the neighbor BS's action δ_j to all neighbor BSs only when the calculated $\nu_{ij}(\delta_j)$ is not the same as the previously sent message value. In addition, when a BS receives a new message value, the BS always records the optimal action that maximizes (18) temporally. BSs do not know the global payoff directly because they send and receive messages locally. Therefore, when the timer expires, BSs share their contribution to the global payoff through MBS and select the optimal action by checking the global payoff's enhancement.

The hybrid message-passing-based algorithm benefits from the hierarchical structure of HetNets in that it sends messages to adjacent BSs directly in a distributed manner and shares the global Q-function through MBS in a centralized manner. Furthermore, it is scalable because it only considers the actions received from neighboring BSs via messages, not all possible combinations of neighboring BSs' actions.

D. COMPLEXITY ANALYSIS

The sample complexity (or sample efficiency) is the time required to find an approximately optimal policy or the number of samples that the RL algorithm needs to learn a target function successfully. Generally, a strategy for managing the tradeoff between exploration and exploitation determines the sample complexity. The sample complexity of the proposed COMORL scheme follows that of model-free Q-value iteration (QVI) because the COMORL scheme works based on model-free QVI. The sample complexity of model-free QVI is known as $\widetilde{O}(\frac{SA}{(1-\gamma)^5\eta^2} ln \frac{1}{\xi})$, where S and A are the numbers of states and actions of the MDP, respectively; γ is the discount factor; and η and ξ are accuracy parameters [28].

The space complexity, that is, the amount of memory used by the algorithm, of general model-free RL is $O(SAH)$, where S, A, H are the numbers of states, actions, and steps of the MDP [29]. The proposed COMORL scheme exploits the joint Q-function, which requires additional memory space proportional to the joint S and A for neighboring BSs. Thus, its space complexity is $O(S^3A^3HN)$ per BS, where N is the number of a BS's neighbors.

The message overhead is another important concern because the CMQ module works based on the message-passing scheme. First, the CMQ module exchanges its utility (9) with neighboring BSs to update the $QTable^C$. If the size of the message is p bytes for sharing the utility value, the message overhead for updating the $QTable^C$ becomes $(k \cdot p)^{|V|}$ bytes, where k is the average degree of a given CG $G = (V, E)$, and $|V|$ is the number of vertices. Second, if the CMQ module sends a message value (17) to the neighboring BSs, it induces successive message exchanges. The total size of the induced messages is $q \cdot (|V| - 1)^k$ bytes, where q is the size of the message value in bytes. Generally, the total size of induced messages is smaller than $q \cdot (|V| - 1)^k$ bytes because the CMQ module sends a message value only when the message is not the same as the previously sent message.

V. SIMULATION RESULTS

A. SIMULATION SETUP

In this section, we present the evaluation of the performance of the proposed COMORL scheme using the LTE model in the NS3 network simulator [30] in two scenarios: static

and dynamic. In the static scenario, it was assumed that the source nodes generate traffic at a constant bit rate (CBR), and UEs have no mobility. The UEs receive data at CBRs, namely, about 137, 218, and 275 kbps under medium, heavy, and full traffic loads, respectively. COMORL was compared with other algorithms: max-SINR, max-SINR with fixed-bias ('FB'), UE-side Q-learning ('UEQ') [7], and the optimal BO vector ('OPT'). For comparison with the OPT, we found the optimal BO vector of (11) by an exhaustive search algorithm before the simulation began with known parameters, such as the positions of BSs and UEs, the SINR of UE receiving from BSs, and the bitrates of traffic per UE. Then, we applied it from the beginning of the simulation in the static scenario.

The dynamic scenario is more similar to the actual network situation. UEs' mobility follows the two-dimensional random walk model, and UEs move around at a speed of 3km/h and received data at a random rate, which follows an exponential distribution with averages of 126, 200, and 252 kbps under medium, heavy, and full traffic loads, respectively. The packets arrive at the network according to the Poisson process, where the mean arrival rates were set to the average data rates divided by the packet size. We compared COMORL with max-SINR, FB, and UEQ except for OPT. We calculated the optimal bias offsets of 'OPT' with the QSI, which was measured by the actual throughput and effective capacity. In the static scenario, the networks were stable (no mobility, CBR traffic), and all parameters for the simulation were prior knowledge. Thus, at the beginning of the simulation, the optimal bias offsets for 'OPT' could be found with an exhaustive search. However, in the dynamic scenario where the networks' condition kept changing, it was impossible to get the optimal bias offsets.

In both scenarios, we deployed four SBSs over one MBS, and a total of 200 UEs were distributed in the network. Also, 1/3 of UEs were uniformly distributed over the coverage area of MBS, and 2/3 of UEs were uniformly distributed over the coverage area of each SBS. We varied the number of UEs among SBSs, which was randomly generated from the normal distribution $\mathcal{N}(33, 53)$, to induce load imbalance between SBSs. We assumed that all BSs had an X2-interface between them, and the control channel was ideal, which means there was no loss of control messages. The event A3 [31], at which the RSRP of neighbors becomes better than that of the serving cell, triggers the X2-based handover. The almost subframe blank (ABS) ratio was set to 50%. We simulated three different traffic load cases: medium, heavy, and full traffic loads, which were around 50%, 70%, and 100% of the system capacity, respectively. The bit rate of traffic per UE was increased according to the traffic load cases because the number of UEs was fixed for every case. The results from multiple simulations were averaged. Table 2 lists the detailed simulation parameters [32].

TABLE 2. Simulation parameters.

Parameters	Value
Carrier frequency	2 GHz
System bandwidth	10 MHz
Subframe duration	1 ms
The number of resource block	50
Transmission power	MBS: 46 dBm, SBS: 30 dBm
Cell radius	MBS: 512 m SBS: 180 m
Feasible bias offset set	{0,1,2,...,16}dB
Distance b/w MBS and SBS	300 m
Min. distance b/w MBS and UE	35 m
Min. distance b/w SBS and UE	10 m
Path loss model	$46.67 + 30\log_{10}(R)$ dB (R m)
Scheduling algorithm	Round Robin (RR)
Packet size	300 Bytes
Delay QoS requirement	150ms
Number of UEs	200
UE mobility speed	3 km/h
Learning rate	0.8
Discount factor	0.8
Exploration rate	1.0 diminishing with rates of 0.99

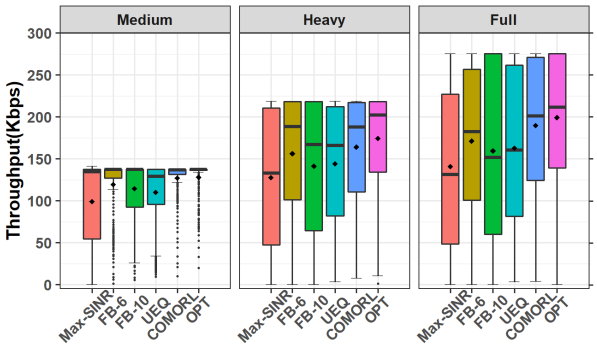


FIGURE 4. Average throughput of UEs in the static scenario.

B. THROUGHPUT

Fig. 4 presents the throughput distribution of UEs for the six schemes for the three different traffic load cases in the static scenario. The throughput distribution tends to spread as the network traffic load increases. This is because network resources become scarce when the network traffic load increases while each traffic source is generating traffic at the same rate. The big dots in the boxplot show the average throughput of UEs. COMORL shows an enhancement in average UE throughput in comparison to the other algorithms by improving the 50th percentile throughput. In particular, the cell-edge throughput performance is shown in Fig. 5. We assumed the 10th percentile throughput as the cell-edge throughput. Remarkably, the average throughput of COMORL approaches that of OPT under a medium traffic load. The cell-edge throughput under a medium traffic load is higher than that in the other traffic load cases because there are enough network resources even though the SINR of UEs at the cell-edge is low. The gap between COMORL and OPT in the cell-edge throughput widens when traffic load becomes heavy because UE handover has more impact on system performance under a heavier traffic load than in the low traffic load. UEs at cell-edge perform handovers in the process to find the optimal BO set in COMORL, whereas there is no handover in OPT because the optimal BO vector was calculated before the beginning of simulations.

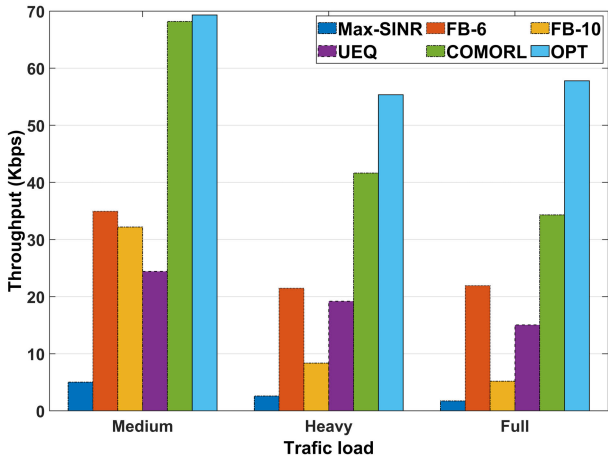


FIGURE 5. Cell edge UE throughput in the static scenario.

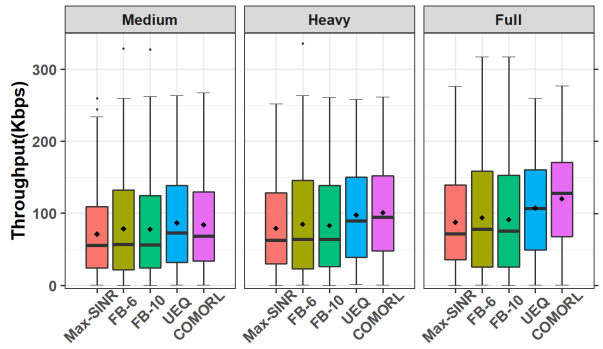


FIGURE 6. Average throughput of UEs in the dynamic scenario.

Fig. 6 presents the throughput distribution of UEs for the five schemes in the three different traffic load cases in the dynamic scenarios. COMORL shows an improvement in terms of the average throughput as the traffic load level increases. Under a medium traffic load, COMORL offloads a limited number of UEs due to its handover strategy, which restricts the handover of UEs that meets their delay requirements because there are enough network resources available to guarantee an individual UE's delay requirement. Therefore, UEQ shows a slightly better average throughput than COMORL under a the medium traffic load because COMORL roughly balances it. However, as the traffic load increases, the network resources become scarce, and accordingly, COMORL performs stricter load balancing to assure UEs' delay requirement. This results in remarkable improvement in the 50th percentile throughput and the average throughput. Also, cell edge throughput in Fig. 7 shows similar trends with average throughput. Without BO adjustment as max-SINR, most UEs at the cell edge attach to MBS and suffer from low SINR and high competition for network resources resulting in throughput degradation. The adaptive adjustment of BO in UEQ and COMORL improves cell edge and 50th percentile throughput by offloading UEs to nearby BSs, which results in average throughput enhancement, whereas the fixed BO in FB shows no meaningful throughput improvement in the dynamic scenario.

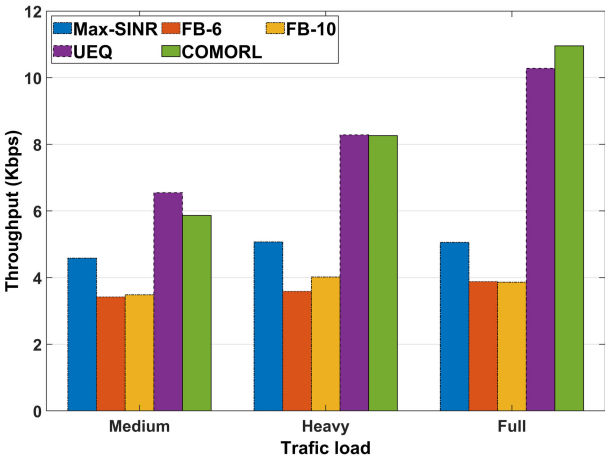


FIGURE 7. Cell edge UE throughput in the dynamic scenario.

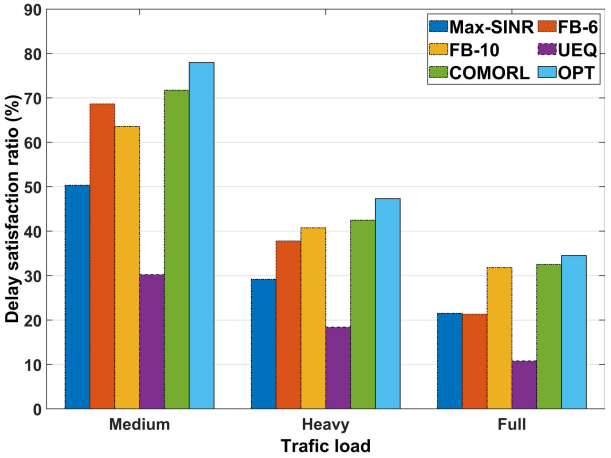


FIGURE 8. Average delay satisfaction ratio in the static scenario.

C. DELAY SATISFACTION RATIO

The delay satisfaction ratio is defined as the number of delay-guaranteed UEs over the total number of UEs in a BS. In this simulation, the UEs' delay requirement was set to 150 ms for a general video service [33]. Fig. 8 shows the delay satisfaction ratio in the static scenario. UEQ showed much lower performance than the other schemes in terms of delay satisfaction. This is mainly attributed to the trial-and-error scheme, which may produce ping-pong handovers. Most UEs at the cell edge may experience this phenomenon in adjusting the BO, and they show lower throughput and higher delay. COMORL reduces the number of unnecessary handovers dramatically by limiting handovers of UEs that satisfy their delay requirements. COMORL shows a slight improvement in the delay satisfaction ratio compared to FB-6 or FB-10 in the static scenario. This result is attributed to the fact that CBR traffic has a constant inter-arrival time, which increases the probability of network congestion. Nevertheless, COMORL shows a remarkable throughput improvement while maintaining an improved delay satisfaction ratio in comparison to other algorithms.

Fig. 9 illustrates the satisfaction ratio in the dynamic scenario. COMORL improves it about 27p.p. and 30p.p.

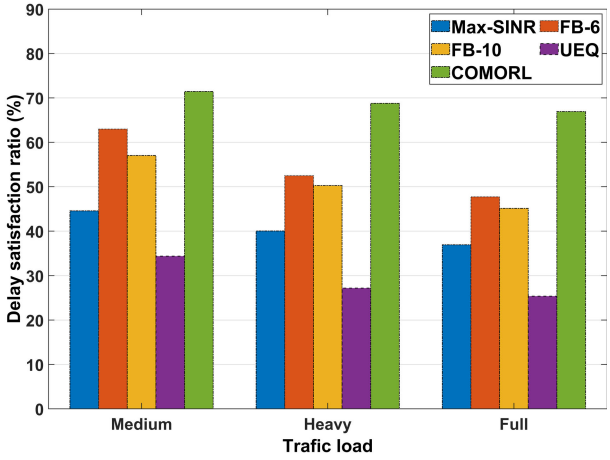


FIGURE 9. Average delay satisfaction ratio in the dynamic scenario.

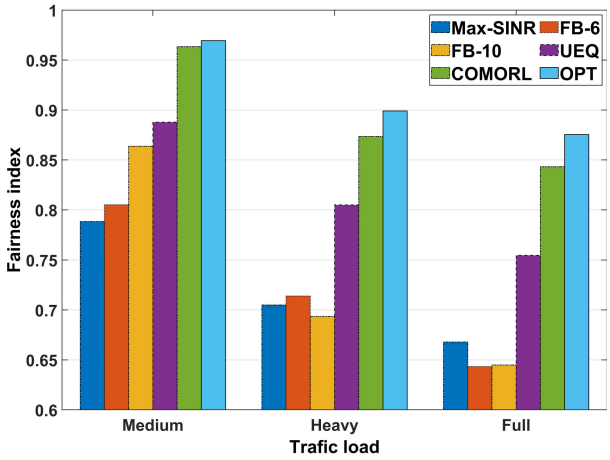


FIGURE 10. Jain's fairness index in the static scenario.

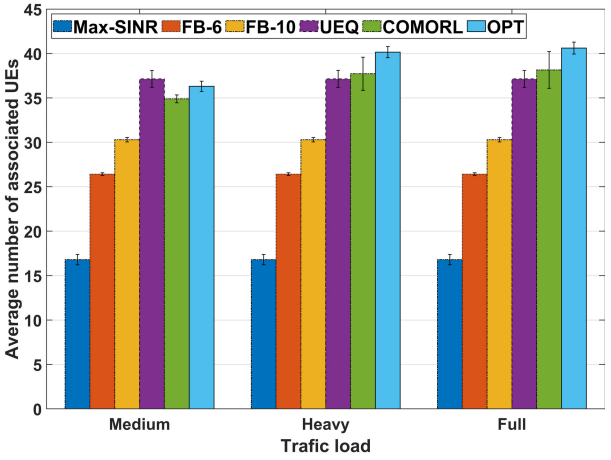


FIGURE 11. Average associated number of UEs at SBSs in the static scenario.

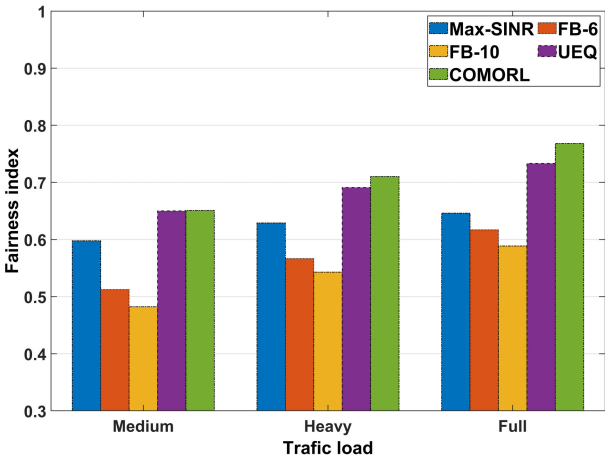


FIGURE 12. Jain's fairness index in the dynamic scenario.

compared to max-SINR under medium and full traffic loads, respectively. FB-6 and FB-10 meet the UE's delay requirements quite well under a medium traffic load because there are available network resources, but they degrade further when the network traffic load becomes heavy. COMORL still maintains UEs' delay satisfaction ratio over 65% even when the network traffic load is full, while those of the other algorithms drop to less than 50%.

D. FAIRNESS

Fig. 10 compares the fairness index of UEs' throughput by using Jain's fairness index (JFI) [34] in the static scenario. The high JFI of UEs means that all UEs receive at the same data rates. In this case, network resources are used efficiently.

Uneven distribution of UEs degrades the JFI in max-SINR because UEs in hotspot areas suffer from competition for scarce resources, while there are extra resources in sparse areas. Load balancing through the adjustment of BOs of SBS in UEQ and COMORL results in better throughput fairness by enhancing cell-edge throughput, as shown in Fig. 4 and 5. In addition, Fig. 11 presents the average number of UEs associated with SBSs in the static scenario, and the error bars in the figure represent the standard deviation of the number of associated UEs to SBSs for four SBSs.

COMORL tends to offload more traffic to SBSs as the traffic load increases, whereas max-SINR, FB, and UEQ keep the same UE association regardless of traffic load. For load balancing, it makes sense to offload traffic to SBSs depending on the traffic load level, which results in better fairness.

The JFI in the dynamic scenario is shown in Fig. 12. From the medium traffic load to the full traffic load, the JFI is improved from 0.05 to 0.1 by COMORL compared to max-SINR. In the dynamic scenario, the overall fairness index appears lower than that in the static scenario because the originally generated data rates for each UE are different. However, we can compare the JFI with that of max-SINR. The JFI of FB-6 and FB-10, which is lower than that of max-SINR, shows that the uniform application of BOs without consideration of each BS-specific situation worsen fairness. UEQ also shows similar enhancement of the JFI under a medium traffic load, but COMORL's JFI performance gets better as the traffic load increases. Besides, the average number of UEs associated with SBSs in COMORL increases, and the standard deviation of the number of connected UEs

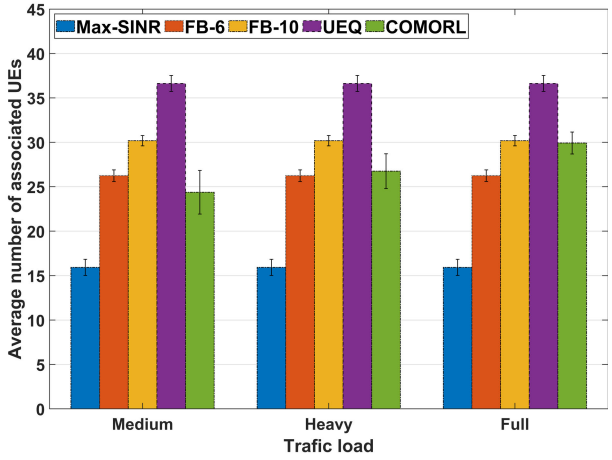


FIGURE 13. Average associated number of UEs at SBS in the dynamic scenario.

to SBS for four SBSs becomes smaller with the network traffic load, as shown in Fig. 13. As the network's traffic increases, the network's free resources decrease, so more UEs are offloaded to SBS, and load balancing is done more tightly to ensure the UEs' delay requirements in COMORL. Accordingly, COMORL further improves fairness under a higher traffic load.

In the static scenario, we observed that the performance of COMORL nearly approaches that of OPT. Some the performance losses are inevitable, and the difference in performance between COMORL and OPT increases as the traffic load increases because RL works in a trial and error manner.

Nevertheless, COMORL shows outstanding performance in terms of throughput, delay satisfaction ratio, and fairness in comparison to the other algorithms. The simulation results in the dynamic scenario showed that COMORL outperforms the other algorithms in terms of throughput, delay satisfaction ratio, and fairness. In particular, COMORL shows outstanding improvement in the delay satisfaction ratio, and the improvement becomes better as the traffic load in the network increases.

VI. CONCLUSION

In this work, a COMORL scheme was proposed for CRE in dense HetNets. The proposed COMORL scheme dynamically adjusts the BOs of load-coupled BSs by learning cooperatively from neighbor BSs to maximize the delay satisfaction ratio. To achieve this, an MDP model was developed with a proposed QoS satisfaction indicator and utility function. Also, a cooperative multi-agent Q-learning algorithm based on a message-passing approach was proposed to find the BO of BSs in the networks, which maximizes the global utility of the networks. Through extensive simulations, we found that the proposed COMORL scheme achieves a maximum of approximately 27% and 30% improvement of the delay satisfaction ratio under medium and full traffic loads, respectively, in a dynamic scenario in comparison to the max-SINR scheme (without CRE).

REFERENCES

[1] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 10)*, document 3GPP, (TS) 36.902, Version 10.12.0, Dec. 2014.

[2] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[3] H. Boostanimehr and V. K. Bhargava, "Unified and distributed QoS-driven cell association algorithms in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1650–1662, Mar. 2015.

[4] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.

[5] S. Mishra, A. Sengupta, and C. S. R. Murthy, "Enhancing the performance of HetNets via linear regression estimation of range expansion bias," in *Proc. 19th IEEE Int. Conf. Netw. (ICON)*, Dec. 2013.

[6] T. Koizumi and K. Higuchi, "Simple decentralized cell association method for heterogeneous networks in fading channel," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*, Sep. 2013.

[7] T. Kudo and T. Ohtsuki, "Cell range expansion using distributed Q-learning in heterogeneous networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2013, no. 1, p. 61, Dec. 2013.

[8] K. Yamamoto and T. Ohtsuki, "Parameter optimization using local search for CRE and eICIC in heterogeneous network," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2014, pp. 1536–1540.

[9] X. Gu, X. Deng, Q. Li, L. Zhang, and W. Li, "Capacity analysis and optimization in heterogeneous network with adaptive cell range control," *Int. J. Antennas Propag.*, vol. 2014, pp. 1–10, Apr. 2014.

[10] M. Qin, Q. Yang, N. Cheng, J. Li, W. Wu, R. R. Rao, and X. Shen, "Learning-aided multiple time-scale SON function coordination in ultra-dense small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2080–2092, Apr. 2019.

[11] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098–3130, May 2018.

[12] E. Rakotomanana and F. Gagnon, "Optimum biasing for cell load balancing under QoS and interference management in HetNets," *IEEE Access*, vol. 4, pp. 5196–5208, 2016.

[13] H. Jiang, Z. Pan, N. Liu, X. You, and T. Deng, "Gibbs-sampling-based CRE bias optimization algorithm for ultradense networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1334–1350, Feb. 2017.

[14] Y. Sun, W. Xia, S. Zhang, Y. Wu, T. Wang, and Y. Fang, "Energy efficient pico cell range expansion and density joint optimization for heterogeneous networks with eICIC," *Sensors*, vol. 18, no. 3, p. 762, Mar. 2018.

[15] G. Hattab and D. Cabric, "Rate-based cell range expansion for downlink massive MIMO heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 296–299, Jun. 2018.

[16] X. Dong, F.-C. Zheng, X. Zhu, and J. Luo, "HetNets with range expansion: Local delay and energy efficiency optimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6147–6150, Jun. 2019.

[17] *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-Configuring and Self-Optimizing Network (SON); Use Cases and Solutions*, document 3GPP, (TS) 36.902, Version 9.3.1, May 2011.

[18] M. M. Hasan, S. Kwon, and J.-H. Na, "Adaptive mobility load balancing algorithm for LTE small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2205–2217, Apr. 2018.

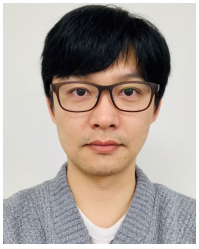
[19] Y. Xu, W. Xu, Z. Wang, J. Lin, and S. Cui, "Load balancing for ultradense networks: A deep reinforcement learning-based approach," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9399–9412, Dec. 2019.

[20] M. M. Hasan and S. Kwon, "Cluster-based load balancing algorithm for ultra-dense heterogeneous networks," *IEEE Access*, vol. 8, pp. 2153–2162, 2020.

[21] H. P. Kuribayashi, M. A. De Souza, D. De Azevedo Gomes, K. Da Costa Silva, M. S. Da Silva, J. C. W. A. Costa, and C. R. L. Francês, "Particle swarm-based cell range expansion for heterogeneous mobile networks," *IEEE Access*, vol. 8, pp. 37021–37034, 2020.

[22] S.-W. Ahn, H. Wang, and D. Hong, "Throughput–delay tradeoff of proportional fair scheduling in OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 9, pp. 4620–4626, Nov. 2011.

[23] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 630–643, May 2003.
[24] C. Guestrin, D. Koller, and R. Parr, "Multiagent planning with factored MDPs," in *Proc. NIPS*, vol. 1, 2001, pp. 1523–1530.
[25] J. R. Kok and N. Vlassis, "Collaborative multiagent reinforcement learning by payoff propagation," *J. Mach. Learn. Res.*, vol. 7, pp. 1789–1828, Sep. 2006.
[26] L. G. Epstein, "Decreasing risk aversion and mean-variance analysis," *Econometrica*, vol. 53, no. 4, p. 945, Jul. 1985.
[27] M. Wainwright, T. Jaakkola, and A. Willsky, "Tree consistency and bounds on the performance of the max-product algorithm and its generalizations," *Statist. Comput.*, vol. 14, no. 2, pp. 143–166, Apr. 2004.
[28] M. Azar, R. Munos, M. Ghavamzadaeh, and H. Kappen, "Speedy Q-learning," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. S. Zemel, and P. Bartlett, Eds. 2011, pp. 2411–2419.
[29] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 4863–4873.
[30] *NS3. The NS-3 Network Simulator*. [Online]. Available: http://www.nsnam.org
[31] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*, document 3GPP, (TS) 36.331, Version 14.2.2, Apr. 2017.
[32] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, document 3GPP, (TS) 36.902, Version 9.2.0, Mar. 2017.
[33] R. A. Cacheda, D. C. García, A. Cuevas, F. J. G. Casta no, J. H. Sánchez, G. Koltsidas, V. Mancuso, J. I. M. Novella, S. Oh, and A. Pantò, "QoS requirements for multimedia services," in *Resource Management in Satellite Networks*, G. Giambene, Ed. Boston, MA, USA: Springer, Jan. 2007, pp. 67–94.
[34] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," 1998, *arXiv:cs/9809099*. [Online]. Available: https://arxiv.org/abs/cs/9809099



HYUNGWOO CHOI (Student Member, IEEE) received the B.S. degree from Chungnam National University, Daejeon, South Korea, in 2005, and the M.S. degree in information and communication engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2007, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include traffic engineering, resource management, the Internet of Things, and machine learning for networking.

TAEHWA KIM (Graduate Student Member, IEEE) received the B.S. degree from Jeonbuk National University, Jeonju, South Korea, in 2005, and the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2007, all in information and communication engineering, where she is currently pursuing the Ph.D. degree in information and communication engineering. Her research interests include network coding and video streaming protocols, the Internet of Things, and machine learning for networking.

HONG-SHIK PARK (Member, IEEE) received the B.S. degree from Seoul National University, Seoul, South Korea, in 1977, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1986 and 1995, respectively, all in electrical engineering. In 1977, he joined the Electronics and Telecommunications Research Institute (ETRI) and was involved in the development of the TDX digital switching system family, including TDX-1, TDX-1A, TDX-1B, TDX10, and ATM switching systems. In 1998, he moved to Information and Communications University, Daejeon, as a Faculty Member. From 2004 to 2012, he was the Director of the BcN Engineering Research Center sponsored by KEIT, South Korea. He is currently a Professor with the School of Electrical and Electronics Engineering, KAIST. His research interests include network architectures and protocols, traffic engineering, and performance analysis of telecommunication systems. He is a member of the Institute of Electronics Engineers of Korea (IEEK) and the Korea Institute of Communication Science (KICS).

JUN KYUN CHOI (Senior Member, IEEE) received the B.Sc. (Eng.) degree in electronics engineering from Seoul National University, Seoul, South Korea, in 1982, and the M.Sc. (Eng.) and Ph.D. degrees in electronics engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1985 and 1988, respectively. From June 1986 to December 1997, he was with the Electronics and Telecommunication Research Institute (ETRI). In January 1998, he joined the Information and Communications University (ICU), Daejeon, South Korea, as a Professor. In 2009, he moved to the Korea Advanced Institute of Science and Technology (KAIST) as a Professor. He is an Executive Member of the Institute of Electronics Engineers of Korea (IEEK), an Editor Board of Member of the Korea Information Processing Society (KIPS), and a Life Member of the Korea Institute of Communication Science (KICS).