IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 29, 2021, accepted February 21, 2021, date of publication February 24, 2021, date of current version March 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062196

Dual Connectivity-Aided Proactive Handover and Resource Reservation for Mobile Users

KAIQIANG QI[1], (Student Member, IEEE), TINGTING LIU[1], (Member, IEEE), CHENYANG YANG[1], (Senior Member, IEEE), SHIQIANG SUO[2], AND YUANFANG HUANG[2]

[1]School of Electronics and Information Engineering, Beihang University, Beijing 100191, China
[2]DaTang Mobile Communications Equipment Co., Ltd., Beijing 100191, China

Corresponding author: Tingting Liu (ttliu@buaa.edu.cn)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61671036.

ABSTRACT Dual connectivity has been regarded as an effective mechanism for mobility management to reduce the handover latency and improve transmission reliability. In this article, we investigate a dual connectivity-aided proactive mobility management scheme for mobile users requesting realtime service with minimum data rate requirements. The proactive mobility management consists of a predictive resource reservation scheme and a proactive handover scheme. The predictive resource reservation scheme is optimized to improve the network throughput or user fairness with the predicted average channel gains and the predicted set of connected base stations for each user. The proactive handover scheme aims to avoid frequent and delayed handovers, and reduce the computational complexity by partitioning cell-center and cell-edge users and choosing different numbers of connected base stations for the two types of users. Simulation results show that the proposed scheme outperforms reactive counterparts, which can effectively reduce service outages meanwhile support higher network throughput when the traffic load is heavy, or improve user fairness meanwhile cause less service outages when the traffic load is light.

INDEX TERMS Dual connectivity, mobility management, proactive handover, predictive resource reservation, neural network.

I. INTRODUCTION

Mobility management plays an important role in improving the continuity, reliability and quality of services for mobile users [1]. When a user moves, the serving cells switch by handovers to keep its quality of service (QoS). If each user is connected to only one base station (BS), service interruptions occur easily during the handover process [2].

Dual connectivity is a mechanism that allows each user to connect to more than one BS simultaneously, which is effective to reduce the handover interruption time. It has been shown in [2]–[4] that with this mechanism, the handover interruption time can be decreased to approximately zero. Furthermore, dual connectivity can guarantee reliable data transmission and improve the network throughput [4]–[6]. Many existing works have considered dual connectivity for mobility management [7]–[9]. To improve the reliability of transmission for ultra-reliability low-latency communication, dual connectivity was used to transmit duplicate packets in [7]. In [8], a dual connectivity mechanism was proposed to enhance the handover performance in dense networks. A dual connectivity-aided handover framework was introduced in [9] to improve transmission reliability in millimeter-wave (mmWave) mobile networks. However, all existing research efforts focus on improving transmission or handover performance without taking into account the data rate requirements of mobile users.

With the rapid development of machine learning and mobile big data analysis, predicting the future trajectory or connected BSs for proactive handover has been shown promising in satisfying the demands of mobile users and improving the system performance in the mobile networks. Proactive handover schemes were proposed to improve the QoS of mobile users during the handover process in different scenarios [10]–[15], which harness the historical data by machine learning. In [10], mmWave beam and blockage were predicted for proactive handover using sub-6 GHz channels based on the multi-layer perceptron (MLP), which enhance the mobility and reliability in mmWave systems. In [11], a deep neural network was used to predict the next

The associate editor coordinating the review of this manuscript and approving it for publication was Emre Koyuncu.

prepared BS for proactive conditional handover to avoid wrong handover preparations in mmWave networks. In [12] and [13], two reinforcement learning-based policies were proposed to reduce unnecessary handovers in mmWave and ultra-dense networks, respectively. A hidden Markov process was adopted in [14] to predict the next connected access point for low latency mobile networks. A support vector machine-based proactive handover method was proposed in [15] to decrease the number of service interruptions and the impact of ping-pong effect in Long-Term Evolution networks. Again, all these works focus on reducing unnecessary handovers or handover latency without considering the data rate requirements.

In addition to the proactive handover, predictive resource reservation for mobile users requesting realtime service is taken into account in [16], [17]. By predicting the future handover time and the next connected BS based on future location, the bandwidth of the next BS was reserved in [16]. By estimating the time instance for each user to execute a handover along its trajectory to the destination and the available bandwidth in each cell, an efficient call admission control scheme was proposed in [17]. However, the resource reservation methods in these two works do not consider time-varying channel fading and QoS requirements. To exploit the available resource at each BS more efficiently, it is necessary to optimize the resource reservation with ensured QoS according to the predictable channel gains.

Predictive resource allocation can exploit the predictable fine-grained channel information to improve user experience or resource usage efficiency [18]–[21]. In [18], the future average channel gains were used for energy-efficient resource allocation with ensured QoS of mobile users requesting hybrid realtime and non-realtime services. In [19], the predicted average channel gains were used for interference coordination and resource allocation planning to improve the user satisfactory rate in heterogeneous networks. In [20], an energy-efficient adaptive video streaming framework was proposed by leveraging the predicted data rates in wireless networks. In [21], deep neural networks were adopted to predict the coarse-grained future information for facilitating a hierarchical and multi-timescale resource management scheme towards non-realtime services.

The aforementioned works about proactive handover and predictive resource allocation only consider single connectivity. Extending these methods into dual-connectivity cases brings new challenges. For example, dual connectivity will inevitably increase the complexity of optimizing resource allocation. So far, it is still not well-understood how to exploit dual connectivity for proactive handover and predictive resource allocation.

In this article, we investigate a proactive mobility management scheme, including both the predictive resource reservation and the proactive handover, with the aid of the dual connectivity mechanism for mobile users requesting realtime service in cellular networks.

The major contributions are summarized as follows:

- We optimize a resource reservation scheme by predicting average channel gains and the set of connected BSs. This scheme aims to maximize the weighted data rate to balance the network throughput and user fairness. A two-timescale resource management method is proposed to guarantee the minimum rate requirement of each realtime user.
- We propose a proactive handover scheme to avoid frequent and delayed handovers. By partitioning users into the cell-center and cell-edge users, respectively with single connectivity and dual connectivity, the proposed handover scheme can reduce the computational complexity for optimizing the resource reservation.
- We leverage neural network as an illustrative machine learning technique to predict the number of connected BSs and the average channel gains. Simulation results show the high accuracy of predicting these two types of information.

The remainder of this article is organized as follows. In Section II, we describe the system model. In Section III, we introduce the proposed proactive mobility management, including a predictive resource reservation scheme and a proactive handover scheme. Section IV provides simulation results. Finally, we conclude this article in Section V.

Notations: $|x|$ denotes the absolute value of x, $|\mathcal{X}|$ denotes the cardinality of a set $\mathcal{X}$, $\|\boldsymbol{x}\|$ denotes the norm of a vector $\boldsymbol{x}$, $\mathcal{X}\backslash x$ denotes the set of elements in $\mathcal{X}$ except x, and $(\cdot)^H$ denotes the conjugate transpose.

II. SYSTEM MODEL

Consider a homogeneous cellular network, which consists of multiple BSs serving mobile single-antenna users over the bandwidth of W. Denote the index set of G BSs as $\mathcal{G} = \{1, \cdots, G\}$. Each BS with maximal transmit power $P_{\max}$ is equipped with N_{tx} antennas, and connected with a central processor (CP).

Each user requests a realtime service, such as video conference and voice-over-IP. The QoS of realtime service can be characterized by the packet transmission delay exceeding a delay bound with a probability less than a threshold [18]. For example, the delay bound for voice-over-IP service cannot exceed 50 ms with a probability of 2% for radio access networks [22]. When designing resource allocation, such a QoS requirement is often simplified conservatively as a hard delay bound, i.e., a packet must be transmitted within the delay bound. Considering that the delay bound is comparable to the correlation time of channel fading, imposing an instantaneous data rate constraint can guarantee the QoS requirement. Here, we divide time into multiple frames each with duration of Δ_{f} (usually in hundred-millisecond or second level, according to the handover decision period). Each frame is divided into N_{s} time slots each with duration of $\Delta_{\text{s}} = \Delta_{\text{f}}/N_{\text{s}}$ (say 1 ms, according to the channel correlation time). To meet the QoS requirement, in each time slot, say the tth time slot of the fth

frame, the instantaneous data rate of the kth user equipment (denoted as UE_k) needs to satisfy

$$R_k^{f,t} \geq R_k^{\mathrm{th}}, \tag{1}$$

where R_k^{th} is the minimum rate requirement of UE_k.

To guarantee the constraint in (1), at the end of the $(f-1)$th frame, the CP makes a resource reservation plan for each BS serving users in each time slot of the fth frame. As shown in Figure 1, in each time slot, N_{tot} resource blocks (RBs) each with bandwidth of $W_0 = W/N_{\mathrm{tot}}$ can be used.

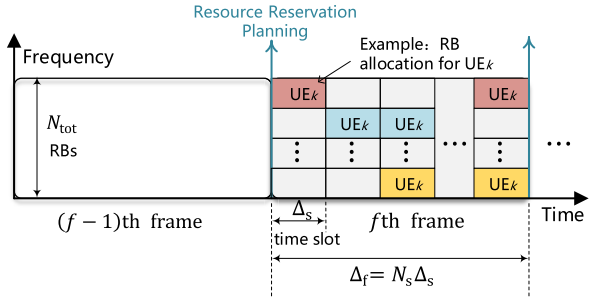


FIGURE 1. Resource reservation planning.

A. CHANNEL MODEL

Assume that the average channel gain (including pathloss and shadowing) remains constant in each frame but may vary among different frames, and the instantaneous channel gain remains constant in each time slot but changes independently among time slots.

Denote the average channel gain between UE_k and the gth BS (denoted as BS_g) in the fth frame as [21]

$$\alpha_{k,g}^f = (d_{k,g}^f)^{\beta} 10^{X_{k,g}^f/10}, \tag{2}$$

where $d_{k,g}^f$ and $X_{k,g}^f$ are the distance and shadowing between UE_k and BS_g in the fth frame, respectively, β is the pathloss exponent, and $X_{k,g}^f$ is a Gaussian random variable with zero mean and standard deviation of σ_X.

B. DUAL CONNECTIVITY

To guarantee the QoS requirement of each user, increase the reliability for connection, and reduce the handover interruption time, we consider the dual connectivity mechanism. This work can be readily extended to the multi-connectivity case.

For each user, assume that the set of connected BSs remains unchanged in each frame but may vary among frames. Let $\mathcal{C}_k^f$ denote the set of the connected BSs of UE_k in the fth frame and $|\mathcal{C}_k^f|$ denote the number of connected BSs (called the *number of connections* for short).

Due to the random arrival and departure of mobile users, the users who are connected to BSs in the network may also change among frames. Let $\mathcal{K}^f$ denote the index set of users served in the fth frame and $K^f = |\mathcal{K}^f|$ be the number of users. Then, the set of users connected to BS_g can be expressed as $\mathcal{K}_g^f = \{k | g \in \mathcal{C}_k^f, \forall k \in \mathcal{K}^f\}$.

C. RESOURCE RESERVATION PLAN

Due to the time-varying instantaneous channel gains in different time slots, the user may be served by different BSs over different RBs. To represent the resource reservation plan in the tth time slot of the fth frame, we introduce a binary variable $\rho_{k,g}^{f,t,n} \in \{0,1\}, g \in \mathcal{C}_k^f$ to denote whether BS_g allocates the nth RB in the tth time slot of the fth frame to UE_k or not, which is called *resource reservation variable*. To guarantee that the number of RBs allocated to UE_k in each time slot is no greater than the total number of available RBs, the resource reservation variables need to satisfy

$$\sum_{n=1}^{N_{\mathrm{tot}}} \rho_{k,g}^{f,t,n} \leq N_{\mathrm{tot}}. \tag{3}$$

To avoid multi-user interference, we consider linear precoder. Then, the number of users served with one RB should satisfy

$$\sum_{k \in \mathcal{K}_g^f} \rho_{k,g}^{f,t,n} \leq K_{\max}, \tag{4}$$

where $K_{\max}$ is the maximal number of users served with one RB, $K_{\max} \leq N_{\mathrm{tx}}$.

With dual connectivity, each user may simultaneously receive signals from both connected BSs. To avoid the interference among the two connected BSs, both BSs need to use different RBs to serve one user in the same time slot. Therefore, it is necessary to meet

$$\sum_{g \in \mathcal{C}_k^f} \rho_{k,g}^{f,t,n} \leq 1. \tag{5}$$

D. INSTANTANEOUS DATA RATE

When $\rho_{k,g}^{f,t,n} = 1$, i.e., BS_g allocates the nth RB in the tth time slot of the fth frame to UE_k, the instantaneous rate of UE_k served by BS_g with the nth RB can be obtained as

$$r_{k,g}^{f,t,n} = W_0 \log_2(1 + \gamma_{k,g}^{f,t,n}), \quad \forall g \in \mathcal{C}_k^f, \tag{6}$$

where

$$\gamma_{k,g}^{f,t,n} = \frac{P_{k,g}^{f,t,n} |(\mathbf{h}_{k,g}^{f,t,n})^H \mathbf{w}_{k,g}^{f,t,n}|^2}{I_{k,g}^{f,t,n} + \sigma^2} \tag{7}$$

is the corresponding instantaneous signal to interference-plus-noise ratio (SINR) [23], $P_{k,g}^{f,t,n} = \frac{P_{\max}}{\sum_{n=1}^{N_{\mathrm{tot}}} \sum_{l \in \mathcal{K}_g^f} \rho_{l,g}^{f,t,n}}$ is the transmit power allocated to UE_k by BS_g at the nth RB via equal power allocation among all the users and RBs, $\mathbf{h}_{k,g}^{f,t,n} \in \mathbb{C}^{N_{\mathrm{tx}} \times 1}$ is the instantaneous channel vector from BS_g to UE_k at the nth RB, which is a zero-mean Gaussian vector satisfying $\|\mathbf{h}_{k,g}^{f,t,n}\|^2 = N_{\mathrm{tx}} \alpha_{k,g}^f$, $\mathbf{w}_{k,g}^{f,t,n} \in \mathbb{C}^{N_{\mathrm{tx}} \times 1}$ is the precoding vector satisfying $\|\mathbf{w}_{k,g}^{f,t,n}\| = 1$, σ^2 is the noise

power, and

$$I_{k,g}^{f,t,n} = \sum_{l \in \mathcal{K}_g^f \setminus k} \rho_{l,g}^{f,t,n} P_{l,g}^{f,t,n} |(\mathbf{h}_{k,g}^{f,t,n})^H \mathbf{w}_{l,g}^{f,t,n}|^2 + \sum_{g' \in \mathcal{G} \setminus g} \sum_{l' \in \mathcal{K}_{g'}^f} \rho_{l',g'}^{f,t,n} P_{l',g'}^{f,t,n} |(\mathbf{h}_{k,g'}^{f,t,n})^H \mathbf{w}_{l',g'}^{f,t,n}|^2 \quad (8)$$

is the interference power. In (8), the first term in the right-hand side is the multi-user interference power, and the second term is the inter-cell interference power.

From (6), the instantaneous rate of UE_k provided by BS_g in the tth time slot of the fth frame can be obtained as

$$R_{k,g}^{f,t} = \sum_{n=1}^{N_{\mathrm{tot}}} \rho_{k,g}^{f,t,n} r_{k,g}^{f,t,n}. \quad (9)$$

With the aid of dual connectivity, the instantaneous rate of UE_k in the tth time slot of the fth frame is

$$R_k^{f,t} = \sum_{g \in \mathcal{C}_k^f} R_{k,g}^{f,t}. \quad (10)$$

III. PROACTIVE MOBILITY MANAGEMENT

In this section, we first optimize the proactive mobility management scheme in the next frame, including the predictive resource reservation and the proactive handover. Then, we provide two methods for predicting the number of connections of each user and the average channel gains from all BSs to each user in the next frame with neural networks. Finally, we summarize the proposed scheme.

A. PREDICTIVE RESOURCE RESERVATION

1) PROBLEM FORMULATION

At the end of the $(f-1)$th frame, given the predicted sets of connected BSs of all users in the fth frame (i.e., $\{\hat{\mathcal{C}}_k^f\}$), the CP optimizes the resource reservation variables (i.e., $\{\rho_{k,g}^{f,t,n}\}$) in all time slots of the fth frame to guarantee the QoS requirement of each user.

To exploit available resources for improving the system performance, such as network throughput or user fairness, the resource reservation plan is optimized to maximize a predicted weighted rate in the fth frame, i.e., the weighted sum of the predicted network rate and the predicted minimal rate among users. The network rate is defined as the rate averaged over the average rates of all users in the considered network to reflect the network throughput, i.e., $\sum_{k \in \mathcal{K}^f} R_k^f / K^f$, where $R_k^f = 1/N_{\mathrm{s}} \sum_{t=1}^{N_{\mathrm{s}}} R_k^{f,t}$ is the average rate of UE_k in the fth frame. The minimal rate is the defined as $\min_{k \in \mathcal{K}^f} \{R_k^f\}$ to reflect the worst user experience for achieving the max-min fairness among users [24]. Maximizing the weighted rate can allow mobile network operators to balance these two performance metrics.

The optimization problem is formulated as

$$\mathsf{P1}: \max_{\{\rho_{k,g}^{f,t,n}\}} \quad \omega \frac{1}{K^f} \sum_{k \in \mathcal{K}^f} \hat{R}_k^f + (1-\omega) \min_{k \in \mathcal{K}^f} \{\hat{R}_k^f\} \quad (11)$$

$$\text{s.t. } \hat{R}_k^{f,t} \geq R_k^{\mathrm{th}}, \quad (11a)$$

$$\rho_{k,g}^{f,t,n} \in \{0,1\},\ (3),\ (4),\ (5),\ g \in \hat{\mathcal{C}}_k^f, \quad (11b)$$

where $\omega \in [0,1]$ is the weighting coefficient and $\hat{R}_k^f$ is the planned average rate of UE_k averaged over the predicted instantaneous rates $\hat{R}_k^{f,t}$ in all the time slots of the fth frame, which depends on the resource reservation variables $\{\rho_{k,g}^{f,t,n}\}$ and the predicted instantaneous rates served by one of the BSs in $\hat{\mathcal{C}}_k^f$ with one RB (i.e., $\{\hat{r}_{k,g}^{f,t,n}\}$).

From (6) and (7), we can see that solving problem $\mathsf{P1}$ requires the CP to predict the instantaneous channel gains $\hat{\mathbf{h}}_{k,g}^{f,t,n}$ from all BSs to all users at N_{tot} RBs of N_{s} time slots to obtain $\hat{r}_{k,g}^{f,t,n}$. However, instantaneous channel gains are hard to predict beyond channel coherence time. In addition, problem $\mathsf{P1}$ is a combinatorial optimization problem, which is with high computational complexity. When adopting the branch and bound algorithm to solve this problem, the complexity in the worst case is $\mathcal{O}_{\mathsf{P1}} = \mathcal{O}\left(2^{N_{\mathrm{s}} N_{\mathrm{tot}} \sum_{k \in \mathcal{K}^f} \left|\hat{\mathcal{C}}_k^f\right|}\right)$ [25]. It is impractical to solve such a problem at the hundred-millisecond scale.

To circumvent this difficulty, we first simplify the problem to optimize the average number of RBs reserved for each user in each time slot of the next frame, where the predictable average channel gains of each user from all BSs in the next frame are used. Then, at the beginning of each time slot in the next frame, each BS allocates RBs to the connected users according to the optimized average number of reserved RBs and the estimated instantaneous channel gains.

2) PROBLEM SIMPLIFICATION

To simplify problem $\mathsf{P1}$, we first introduce an approximation of the average SINR. Consider zero-forcing precoder, then the multi-user interference in (8) can be completely avoided if $K_{\max} \leq N_{\mathrm{tx}}$. From the analysis in [19] and [26], the average SINR in the fth frame can be approximated as

$$\hat{\gamma}_{k,g}^f \approx \frac{P_{k,g}^{f,n}(N_{\mathrm{tx}} - K_{\max} + 1)\hat{\alpha}_{k,g}^f}{\sum_{g' \in \mathcal{G} \setminus g} (P_{k,g'}^{f,n} K_{\max}) \hat{\alpha}_{k,g'}^f + \sigma^2}, \quad (12)$$

where $P_{k,g}^{f,n} = P_{\max}/(K_{\max} N_{\mathrm{tot}})$ is the transmit power allocated to UE_k by BS_g at one RB, and $\hat{\alpha}_{k,g}^f$ is the average channel gain of UE_k from BS_g in the fth frame.

Then, the planned average rate of UE_k in the fth frame can be obtained as

$$\hat{R}_k^f = \sum_{g \in \hat{\mathcal{C}}_k^f} \hat{R}_{k,g}^f, \quad (13)$$

where

$$\hat{R}_{k,g}^f = \bar{m}_{k,g}^f \hat{r}_{k,g}^f = \bar{m}_{k,g}^f W_0 \log_2(1 + \hat{\gamma}_{k,g}^f) \quad (14)$$

is the planned average rate of UE_k provided by BS_g, $\hat{r}^f_{k,g}$ is the predicted average rate of UE_k served by BS_g with one RB in the fth frame, and

$$\bar{m}^f_{k,g} = \frac{1}{N_s}\sum_{t=1}^{N_s}\sum_{n=1}^{N_{tot}} \rho^{f,t,n}_{k,g} \tag{15}$$

is the average number of RBs reserved for UE_k at BS_g in each time slot of the fth frame.

The simplified problem can be formulated as

$$\mathsf{P2}: \max_{\{\bar{m}^f_{k,g}\}} \quad \omega\frac{1}{K^f}\sum_{k\in\mathcal{K}^f}\hat{R}^f_k + (1-\omega)\min_{k\in\mathcal{K}^f}\{\hat{R}^f_k\} \tag{16}$$

$$\text{s.t. } \hat{R}^f_k \geq R^{\text{th}}_k, \tag{16a}$$

$$0 \leq \bar{m}^f_{k,g} \leq N_{tot}, \quad \bar{m}^f_{k,g} \in \mathbb{Z}, \tag{16b}$$

$$\sum_{k\in\hat{\mathcal{K}}^f_g} \bar{m}^f_{k,g} \leq K_{\max} N_{tot}, \tag{16c}$$

$$\sum_{g\in\hat{\mathcal{C}}^f_k} \bar{m}^f_{k,g} \leq N_{tot}, \tag{16d}$$

where (16a), (16b), (16c) and (16d) are derived from the constraints in (11a), (3), (4) and (5), respectively.

Problem P2 is integer programming, and directly solving this problem is still with exponential complexity. To reduce the computational complexity, we first relax the integer variables to the real ones. In practice, when the number of available RBs in each BS is large, the performance loss caused by the relaxation is negligible. To deal with the non-differential objective function caused by the minimization function, we use a standard approach to max-min problems by introducing an auxiliary variable x to (16) and considering extra constraints [24]. Then, the relaxed problem can be equivalently converted into a linear programming problem as follows,

$$\mathsf{P3}: \max_{x,\{\bar{m}^f_{k,g}\}} \quad \omega\frac{1}{K^f}\sum_{k\in\mathcal{K}^f}\hat{R}^f_k + (1-\omega)x \tag{17}$$

$$\text{s.t. } x \leq \hat{R}^f_k, \tag{17a}$$

$$0 \leq \bar{m}^f_{k,g} \leq N_{tot}, \tag{17b}$$

$$(16a),\ (16c),\ (16d).$$

Problem P3 can be solved with an efficient optimization algorithm, such as the interior-point algorithm with polynomial complexity $\mathcal{O}(n^{3.5})$, where n is the number of optimization variables [27]. In (17), the number of elements in $\{\bar{m}^f_{k,g}\}_{k\in\mathcal{K}^f, g\in\hat{\mathcal{C}}^f_k}$ is equal to the total number of connections from all users, i.e., $\sum_{k\in\mathcal{K}^f}|\hat{\mathcal{C}}^f_k|$. By adding the introduced auxiliary variable x, the total number of optimization variables is $\sum_{k\in\mathcal{K}^f}|\hat{\mathcal{C}}^f_k|+1$. Therefore, the computational complexity of P3 is obtained as

$$\mathcal{O}_{\mathsf{P3}} = \mathcal{O}\left(\left(\sum_{k\in\mathcal{K}^f}|\hat{\mathcal{C}}^f_k|+1\right)^{3.5}\right). \tag{18}$$

Next, the optimal solution of the relaxed problem P3 is slightly adjusted to the integers $\{\bar{m}^{f*}_{k,g}\}$ with constraints of the minimum rate requirements and total available resources to obtain the optimal solution of problem P2.

Finally, the CP informs each BS of the average number of RBs reserved for each connected user in the next frame, i.e., $\bar{m}^{f*}_{k,g}$. Then, BS_g allocates $\bar{m}^{f*}_{k,g}$ RBs to UE_k in each time slot. Although the instantaneous rate $R^{f,t}_k$ fluctuates around the optimal average rate $R^{f*}_k = \bar{m}^{f*}_{k,g}\hat{r}^f_{k,g}$ due to the impacts of the instantaneous channel gains, the average rate R^f_k approaches to R^{f*}_k. Therefore, the optimized resource reservation scheme can provide the maximal weighted rate.

However, the instantaneous rate $R^{f,t}_k$ may not satisfy the minimum rate requirement in (1). In the following, we adjust the number of reserved RBs slightly to meet the requirement according to the estimated instantaneous channel gains at the beginning of each time slot in the fth frame. From (10), we know that the instantaneous rate is contributed by multiple connected BSs. Therefore, to satisfy the requirement in (1), the instantaneous rate provided by BS_g should meet

$$R^{f,t}_{k,g} \geq R^{f,\text{th}}_{k,g}, \tag{19}$$

where $R^{f,\text{th}}_{k,g} = R^{\text{th}}_k\hat{R}^f_{k,g}/(\sum_{g'\in\hat{\mathcal{C}}^f_k}\hat{R}^f_{k,g'})$ is the minimum rate requirement of UE_k supported by BS_g in the fth frame.

When (19) is unsatisfied, BS_g should allocate more than $\bar{m}^{f*}_{k,g}$ RBs to UE_k by borrowing RBs from other users. We know that for UE_k, except the required RBs to meet (19), the average number of remaining RBs can be estimated as $\bar{m}^{f*}_{k,g} - R^{f,\text{th}}_{k,g}/\hat{r}^f_{k,g}$. Considering that users with more remaining RBs are easier to meet the QoS requirements, the user whose reserved RBs are not enough can borrow RBs from the ones with abundant remaining RBs.

B. PROACTIVE HANDOVER

1) HANDOVER WITH DUAL CONNECTIVITY

From (13), we can see that solving the resource reservation problem P3 depends on the predicted sets of connected BSs of all users in the next frame, i.e., $\hat{\mathcal{C}}^f_k, \forall k$. A simple and direct method of selecting the next connected BSs for each user is according to the predicted average channel gains from all BSs in the next frame, i.e.,

$$\hat{\mathcal{C}}^f_k = \arg\max_2 \hat{\boldsymbol{\alpha}}^f_k, \tag{20}$$

where $\arg\max_n$ denotes the index set of n BSs with maximal average channel gains, $\hat{\boldsymbol{\alpha}}^f_k = [\hat{\alpha}^f_{k,1}, \cdots, \hat{\alpha}^f_{k,G}]$ is the predicted average channel gain vector in the fth frame. If $\hat{\mathcal{C}}^f_k \neq \mathcal{C}^{f-1}_k$, the user needs to execute the handover by connecting to the new BSs, otherwise, the user still maintains the previous connection. However, owing to the fluctuation of shadowing and the mobility of users, connecting to the next BSs according to (20) may make mobile users suffer from frequent handovers, which will lead to the ping-pong effect and affect the QoS.

To avoid frequent handovers, traditional reactive handover schemes divide the handover process into two separate

phases, i.e., *handover trigger* and *handover execution* [28]. The handover trigger is determined by the user based on the measured average channel gains in the current frame. Only when the trigger condition has been fulfilled for a duration of Time-to-Trigger (TTT), the handover is executed at the BS. A large TTT can help decrease frequent handovers, which however may cause delayed handovers, increasing the handover failures. To improve the QoS of mobile users by avoiding both frequent and delayed handovers, it is necessary to propose a proactive handover scheme by predicting the next connected BSs of each user when the handover is triggered. By considering the future average channel gains from the neighbor BSs in the duration of TTT (i.e., T_{p} frames), the expected set of connected BSs is obtained. In this way, the proactive handover can be executed in advance (immediately after the handover is triggered) without the need of waiting for a duration of TTT.

2) PROACTIVE HANDOVER WITH LOW COMPLEXITY

From (18), we can see that the computational complexity of solving problem **P3** depends on the total number of connections, i.e., $\sum_{k\in\mathcal{K}} |\hat{\mathcal{C}}_k^f|$. To further reduce the computational complexity, we first consider the minimal numbers of connections for different users while maintaining their QoS requirements. For cell-center users, say the user in the first location as shown in Figure 2, there is no need to adopt dual connectivity. While for cell-edge users, say the user in the second location, it is necessary to establish dual connectivity by connecting to two neighbor BSs with larger average channel gains for guaranteeing the QoS. Then, by establishing dual connectivity only for cell-edge users rather than all users, the complexity of solving problem **P3** can be reduced. Moreover, both the signaling overhead and the energy consumption can be reduced by avoiding the establishment of the costly dual connectivity for the cell-center users.

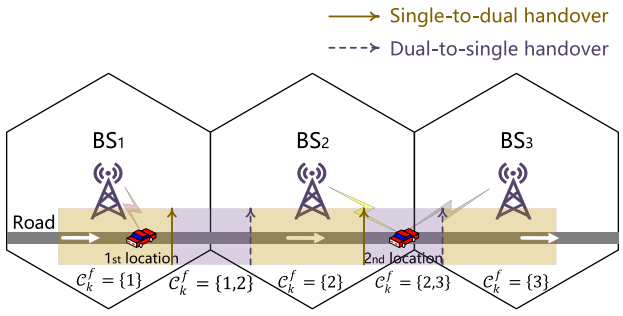


FIGURE 2. An illustration of the connection process of a user when moving along the road.

As illustrated in Figure 2, after partitioning all the users into cell-center and cell-edge users, there exist two types of handovers in the network, i.e., *single-to-dual* handover for cell-center users and *dual-to-single* handover for cell-edge users. The proactive handover can be executed by predicting the next connected one or two BSs, i.e., $\hat{\mathcal{C}}_k^f$. Yet, it is hard to predict $\hat{\mathcal{C}}_k^f$ since its number of elements varies for different handovers. In fact, after the handover is triggered at the end of the $(f-1)$th frame, given the predicted number of connections $|\hat{\mathcal{C}}_k^f|$ of UE_k, the user can obtain the next connected one or two BSs according to the largest measured average channel gains in the $(f-1)$th frame as

$$\hat{\mathcal{C}}_k^f = \arg\max_{|\hat{\mathcal{C}}_k^f|} \boldsymbol{\alpha}_k^{f-1}. \tag{21}$$

Consequently, predicting $|\hat{\mathcal{C}}_k^f|$ is equivalent to predicting $\hat{\mathcal{C}}_k^f$, but the former is more simple and effective than the latter. If $|\hat{\mathcal{C}}_k^f| \neq |\mathcal{C}_k^{f-1}|$, the handover is executed and UE_k connects to the one or two BSs in $\hat{\mathcal{C}}_k^f$, otherwise, the previous connection is maintained. To avoid frequent handovers, we can consider the future average channel gains in the subsequent T_{p} frames to obtain the expected number of connections. Besides, since the handover trigger condition has been met in the $(f-1)$th frame, using the measured average channel gains in the $(f-1)$th frame is sufficient to determine $\hat{\mathcal{C}}_k^f$.

When moving along the road, each user may change its connection (i.e., the set of connected BSs) during the handover process. To ensure the QoS of each cell-edge user, it should provide a sufficient duration to establish dual connectivity. Then, the trigger conditions for single-to-dual and dual-to-single handovers should be set with different offset values and TTT durations (i.e. T_{p} frames), which can be expressed as

$$10\log_{10}\left(\frac{\alpha_{k,g'}^{f-1}}{\alpha_{k,g}^{f-1}}\right) > \begin{cases} \alpha_{\mathrm{sd}}^{\mathrm{th}}(\mathrm{dB}), & \forall\, \mathcal{C}_k^{f-1}=\{g\}, \\ \alpha_{\mathrm{ds}}^{\mathrm{th}}(\mathrm{dB}), & \forall\, \mathcal{C}_k^{f-1}=\{g,g'\}, \end{cases} \tag{22}$$

where $\alpha_{\mathrm{sd}}^{\mathrm{th}} < 0$ and $\alpha_{\mathrm{ds}}^{\mathrm{th}} > 0$ are respectively the offset values for triggering the single-to-dual and dual-to-single handovers, and T_{p} is set as T_{sd} frames or T_{ds} frames to determine the expected number of connections in the next frame respectively for the single-to-dual or dual-to-single handover. In general, $T_{\mathrm{sd}} \neq T_{\mathrm{ds}}$.

Next, we illustrate the connection process of a user when traversing across the second cell, as shown in Figure 2. When the user is to enter the second cell, the user stays connected to both BS_1 (i.e., the previous serving BS) and BS_2. At the end of each frame, the user judges whether the dual-to-single handover trigger condition is satisfied according to (22). If the trigger condition is satisfied at the end of the $(f-1)$th frame, the user reports the average channel gains in an observation window to BS_2, and BS_2 predicts the number of connections in the fth frame. If the predicted number of connections is different from that in the $(f-1)$th frame, the handover is executed by disconnecting to BS_1, otherwise, the user remains connected to both BSs. To avoid frequent reports from the user and frequent predictions at the BS, the predicted set of connected BSs will sustain unchanged for a duration of T_{ds} frames. After T_{ds} frames, the user restarts to judge whether the trigger condition is satisfied.

When the user has entered the second cell, the user only stays connected to BS_2. If the single-to-dual handover trigger

condition is satisfied, similar procedures (i.e., measurement report from the user and prediction at the BS) to the aforementioned dual-to-single handover are carried out. For the same purpose, the predicted set of connected BSs will sustain unchanged for a duration of T_{sd} frames. Finally, when the user is about to leave the second cell, the handover is completed by connecting to both BS_2 and BS_3.

C. INFORMATION PREDICTION WITH NEURAL NETWORKS

To implement the proposed proactive handover and resource reservation scheme, it is necessary to predict the number of connections for each user in the fth frame $|\hat{\mathcal{C}}_k^f|$ after the handover is triggered, which is helpful to obtain the next set of connected BSs $\hat{\mathcal{C}}_k^f$ according to (21), and to predict the average channel gain vector in the fth frame $\hat{\boldsymbol{\alpha}}_k^f$. For notational simplicity, we omit the subscript k in all the parameters throughout this subsection.

1) PREDICTING NUMBER OF CONNECTIONS AT EACH BS

When the handover is triggered at a user, the connected BS predicts the number of connections for the user in the fth frame based on its average channel gain vectors in an observation window, i.e., $[\boldsymbol{\alpha}^{f-T_{\mathrm{o}}}, \cdots, \boldsymbol{\alpha}^{f-1}]$, where T_{o} is the number of frames in the observation window.

Predicting the number of connections is a classification problem with two classes, corresponding to single connectivity and dual connectivity. When making the prediction at BS_g, each sample for training and testing the supervised learning model consists of a $T_{\mathrm{o}}G$-dimensional input vector $\boldsymbol{x}_g$ and a two-dimensional expected output vector (i.e., label) $\boldsymbol{y}_g$. Specifically, the input vector $\boldsymbol{x}_g$ is formed by a time series of average channel gain vectors in the observation window, i.e., $\boldsymbol{x}_g = [\boldsymbol{\alpha}^{f-T_{\mathrm{o}}}, \cdots, \boldsymbol{\alpha}^{f-1}]$. The expected output vector $\boldsymbol{y}_g = \boldsymbol{c}^f$ is a one-hot vector, where the index of the nonzero element indicates the expected number of connections in the next frame, i.e., $\boldsymbol{c}^f = [1, 0]$ or $[0, 1]$ respectively represents that the user should be with single connectivity or dual connectivity in the fth frame.

To obtain the expected output vector $\boldsymbol{y}_g$, we introduce a simple criterion to decide the appropriate number of connections in the next frame. The criterion is based on the average channel gains in the subsequent T_{p} frames after the handover trigger condition is satisfied, which can be obtained at BS_g by periodically reporting the average channel gains from users in the training phase. We take the single-to-dual handover as an example, while the criterion can be also applied to the dual-to-single handover. When the handover trigger condition is satisfied at the end of the $(f-1)$th frame, BS_g determines the appropriate number of connections in the fth frame according to the average channel gains from the fth frame to the $(f+T_{\mathrm{sd}}-1)$th frame. Considering that the average channel gains fluctuate due to shadowing, we relax the handover execution condition as that the trigger condition is satisfied during more than $\kappa_{\mathrm{sd}}\%$ of the total number of frames. The corresponding parameters of determining $\boldsymbol{y}_g$ for the dual-to-single handover are T_{ds} and $\kappa_{\mathrm{ds}}\%$.

We adopt a widely-used neural network, MLP [29], as an illustration to make the prediction (called MLP_1). The employed MLP_1 consists of input layer, hidden layers, and output layer, as shown in Figure 3. We select ReLU function (i.e., $y = \max(x, 0)$) as the activation function for hidden layers, which can achieve similar performance to other activation functions, such as tanh and sigmoid functions according to our simulation. We select the softmax function (i.e., $y_i = e^{x_i}/\sum_{j=1}^{n} e^{x_j}, i = 1, \cdots, n$ with n classes) as the activation function for the output layer, which is typical for the classification problem. With the generated training samples, each BS trains MLP_1 to minimize the cross-entropy between the output vectors of the neural network and the expected output vectors for all the training samples. After training, each BS can predict the number of connections in the next frame with its own trained model when the handover is triggered.

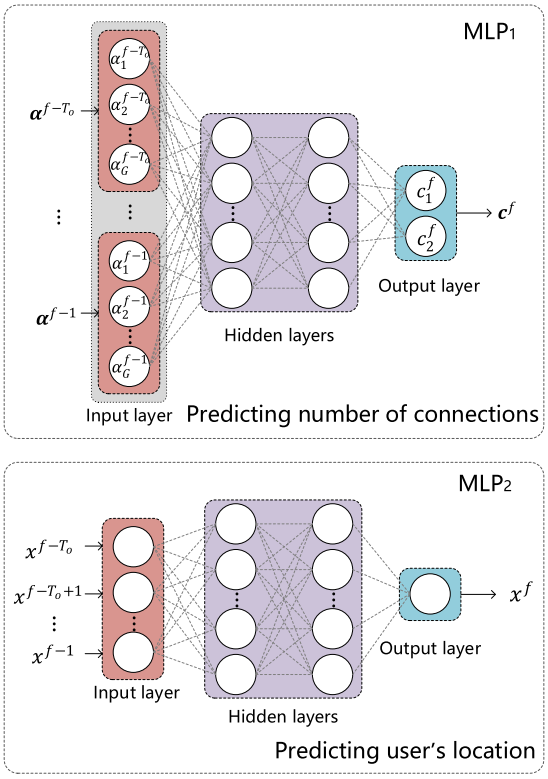


FIGURE 3. Structures of two MLPs for predicting the number of connections (MLP_1) and predicting the user's location (MLP_2).

2) PREDICTING USER's LOCATION AT CP

When directly predicting the average channel gain vector in the next frame for each user, the prediction accuracy is low due to shadowing. Instead, we first predict the user's location, and then translate it to average channel gains from all BSs with the help of a radio map stored at the CP.

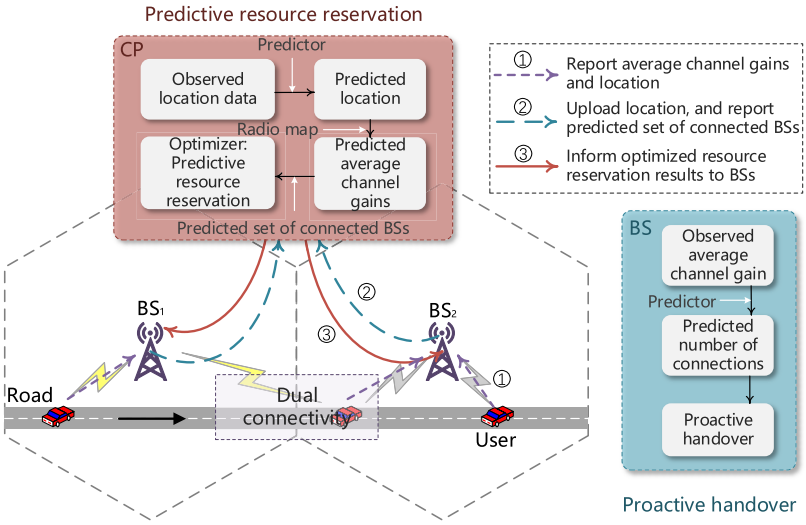


FIGURE 4. An illustration of the complete procedure of the proactive handover and resource reservation scheme.

The CP can predict the user's location in the next frame with the uploaded location data from the user in an observation window, which is measured at the user's mobile equipment via global positioning system. Location prediction is a regression problem. Each sample generated at the CP for training and testing the learning model consists of a T_o-dimensional input vector $\boldsymbol{x}_{cp}$ and a one-dimensional output value y_{cp}. Specifically, the input vector $\boldsymbol{x}_{cp}$ is formed by a time series of locations in the observation window, i.e., $\boldsymbol{x}_{cp} = [x^{f-T_o}, \cdots, x^{f-1}]$, where x^f is the location in the fth frame. The expected output value $y_{cp} = x^f$ is the location in the next frame.

Again, we employ a MLP to make the prediction (called MLP_2). MLP_2 is with similar structure to MLP_1 as shown in Figure 3. We also use ReLU as the activation function in the hidden layers, but we do not adopt any non-linear activation function in the output layer, since the linear output yields better performance according to our simulation. With the training samples at the CP, MLP_2 is trained to minimize the mean square error between the output values of the neural network and the expected output values for all the training samples. At the end of each frame in the prediction phase, the CP can predict the location of each user in the next frame, which is used to obtain the average channel gains from all BSs with the help of the pre-stored radio map.

D. COMPLETE PROCEDURE OF PROACTIVE MOBILITY MANAGEMENT

The proposed proactive mobility management scheme is composed of two key parts, as illustrated in Figure 4, where the proactive handover and the predictive resource reservation are performed at each BS and the CP, respectively. We next provide the procedures of these two parts as follows:

- *Proactive Handover:* When the handover trigger condition is satisfied at a user at the end of a frame, say the $(f-1)$th frame, the user reports the recorded average channel gain vectors in the observation window to the connected BS [1] (with larger average channel gain for the cell-edge user). With the reported average channel gain vectors, the BS predicts the number of connections in the next frame (i.e., the fth frame) for this user, and then determines whether to execute the handover by changing the number of connections. The predicted number of connections can also help the BS to determine the next set of connected BSs for this user.
- *Predictive Resource Reservation:* At the end of each frame, each user reports the measured location in this frame to a connected BS. Then, the location is uploaded to the CP from the BS together with the predicted number of connections (hence the predicted set of connected BSs) if the handover is to be executed in the next frame. The CP predicts each user's location in the next frame, which is then translated into the predicted average channel gain vector with the help of the pre-stored radio map. Finally, the CP optimizes the average number of RBs reserved for each user in every time slot of the next frame with the predicted sets of connected BSs and the predicted average channel gain vectors of all users, and then inform the optimized results to all BSs. At the beginning of each time slot in the next frame, the resource allocation decision is obtained at each BS according to the optimized average numbers of RBs and the estimated instantaneous channel gains of these users.

[1] To avoid occupying the storage space, each user only records the average channel gain vectors in the observation window, and discards the recorded data before this window.

In Figure 4, the uplink signaling overhead can be calculated from Steps ① and ②, while the downlink signaling overhead can be calculated from Step ③.

- *Step ①:* The user always reports the measured location (including latitude and longitude, i.e., two real numbers) in a frame to the connected BS. If the handover is triggered at the user, the user also uploads the average channel gains in the observation window (i.e., T_oG real numbers) to the connected BS.
- *Step ②:* The BS forwards the reported location (i.e., two real numbers) to the CP. If the handover is to be executed according to the predicted number of connections, the BS also uploads the predicted set of connected BSs (including one or two real numbers, which depends on single or dual connectivity of the user) to the CP.
- *Step ③:* The CP informs the optimized resource reservation decision for each user in the next frame to the one or two connected BSs (including one or two real numbers).

The signaling overhead depends on the states, i.e., with or without triggering handover in Step ①, with or without executing handover in Step ②, and single or dual connectivity in Steps ② and ③. We can see that the maximal signaling overhead in the uplink is $(T_oG + 6)$ real numbers, and the maximal overhead in the downlink is two real numbers.

Remark 1: For the proposed scheme, the predictive resource reservation is applicable to homogeneous cellular networks with arbitrary cell size and arbitrary carrier frequency, while the proactive handover between single connectivity and dual connectivity is developed for homogeneous networks under a low-carrier frequency (say sub-6 GHz). When the system operates in a higher frequency band, say the mmWave band, the data transmission is highly susceptible to blockages, which easily incurs the connection intermittency. To extend this scheme to mmWave homogeneous networks, even the cell-center users should keep connected to both BSs to cope with blockages. By predicting the next connected two BSs, the proactive handover can be performed from a dual-connectivity state to another dual-connectivity state.

IV. SIMULATION RESULTS

In this section, we first show the performance of predicting the number of connections and the user's location. Then, we evaluate the handover performance of the reactive and proactive handover schemes with dual connectivity. Finally, we compare the system performance achieved by the proposed proactive handover and resource reservation scheme with three baselines.

A. SIMULATION SETUPS

Consider a homogeneous cellular network consisting of $G = 4$ macro BSs, where each BS with radius of 250 m is located at the center of a hexagonal cell as shown in Figure 5. The vehicles travel along two straight roads in one direction, and the two roads are with minimum distance of 60 m and 120 m from the BSs, respectively. Users move along the two roads

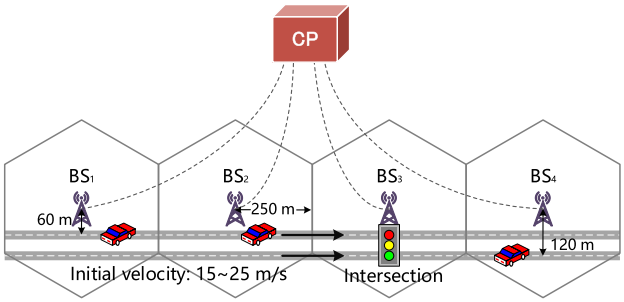


FIGURE 5. Considered network and road topology.

with random initial velocities of $15 \sim 25$ m/s. The acceleration of each user in each frame is a Gaussian random variable with zero mean and standard deviation of 1 m/s^2. There is an intersection with a traffic light at each road, and each user stops $0 \sim 10$ s randomly at the intersection. To avoid the boundary effect, a user re-enters the road from the other side when arriving at the end of a road.

For the dual connectivity-aided handover, the offset value of the trigger condition for single-to-dual or dual-to-single handover is set as $\alpha_{\rm sd}^{\rm th} = -6$ dB or $\alpha_{\rm ds}^{\rm th} = 6$ dB. The expected number of connections, i.e., the label of a sample for training or testing MLP$_1$, is obtained based on the future average channel gains in the duration of $T_{\rm sd} = 10$ frames (i.e., 2 s) or $T_{\rm ds} = 20$ frames (i.e., 4 s), and the fraction of frames that satisfy the handover trigger condition is set as $\kappa_{\rm sd}\% = 80\%$ or $\kappa_{\rm ds}\% = 90\%$. We set different parameters for these two types of handovers, so as to reduce the impact of the shadowing on cell-edge users, ensure each cell-edge user to establish dual connectivity, and avoid frequent handovers.

The other simulation parameters are listed in Table 1 [22], [30]. These setups are used in the sequel unless otherwise specified.

TABLE 1. Simulation parameters.

Simulation parameters	Values
Transmit power of each BS, $P_{\max}$	46 dBm
Noise power, σ^2	-95 dBm
Number of transmit antennas, $N_{\rm tx}$	8
Maximum number of served users at one RB, $K_{\max}$	4
System bandwidth, W	18 MHz
Bandwidth of one RB, W_0	180 KHz
Total number of RBs in the system, $N_{\rm tot}$	100
Pathloss at 1 meter	35.3 dB
Pathloss exponent, β	3.76
Standard derivation of log-normal shadowing, σ_X	6 dB
Correlation distance of shadowing	50 m
Small-scale fading channels among time slots	Rayleigh
Total number of served users on both roads, $K^f = K$	80
Minimum rate requirement of each user, $R_k^{\rm th} = R^{\rm th}$	2 Mbps
Velocity of each mobile user	$15 \sim 25$ m/s
Duration of each time slot, $\Delta_{\rm s}$	1 ms
Duration of each frame, $\Delta_{\rm f}$	200 ms
Length of observation window, $T_{\rm o}$	5 frames

B. PREDICTION PERFORMANCE

In this subsection, we show the prediction results of the two MLPs, including the prediction performance of the number of connections, the user's location, as well as the corresponding average channel gain from the nearest BS. The fine-tuned hyper-parameters are listed in Table 2.

TABLE 2. Fine-tuned hyper-parameters of MLPs.

Hyper-parameters	Values	
	MLP_1	MLP_2
Number of input nodes	20	5
Number of hidden layers	1	
Number of hidden nodes	50	20
Number of output nodes	2	1
Learning algorithm	Adam [31]	
Initial learning rate	10^{-3}	10^{-4}
Mini-batch size	16	32

1) PREDICTING NUMBER OF CONNECTIONS

We first generate the training and test samples for MLP_1. We synthesize the trajectory and obtain the average channel gains at the corresponding locations for each user according to the simulation setups. Specifically, we first sample the trajectory of a user with duration of a frame to obtain the locations of the user. Then, with the considered network topology, pathloss and shadowing for the four cells and two roads, we can simulate the average channel gains of the user from the four BSs in each location. With the periodically reported average channel gains from users in the training phase, each BS can generate its own samples according to the criterion in Section III-C1. The training set and test set are independently generated at each BS with $2,000$ and 500 samples, respectively.

The prediction accuracy at BS_g is computed as

$$A_g = \frac{\sum_{n=1}^{N_{\text{te}}} \mathbf{1}(\hat{c}_g^{(n)} = c_g^{(n)})}{N_{\text{te}}} \times 100\%, \tag{23}$$

where $c_g^{(n)}$ and $\hat{c}_g^{(n)}$ are the real and predicted number of connections of the nth sample at BS_g, N_{te} is the number of samples in the test set, and $\mathbf{1}(\cdot)$ is the indicator function. We show the accuracy of predicting the number of connections of each user at BS_2 and BS_3, since the two types of handovers, i.e., single-to-dual and dual-to-single handovers, are included in both cells when each user traverses across these two cells. With the well-trained MLP_1, we obtain $A_2 = 98.2\%$ and $A_3 = 99.8\%$ in the test sets of BS_2 and BS_3, respectively. We can see that the accuracy of predicting the number of connections of each user in the next frame at each BS is very high.

2) PREDICTING USER's LOCATION

We generate the training and test samples for MLP_2 from the synthetic trajectories by sliding the observation window with the step of a frame. The training set and test set are independently generated at the CP with $10,000$ and $2,500$ samples, respectively.

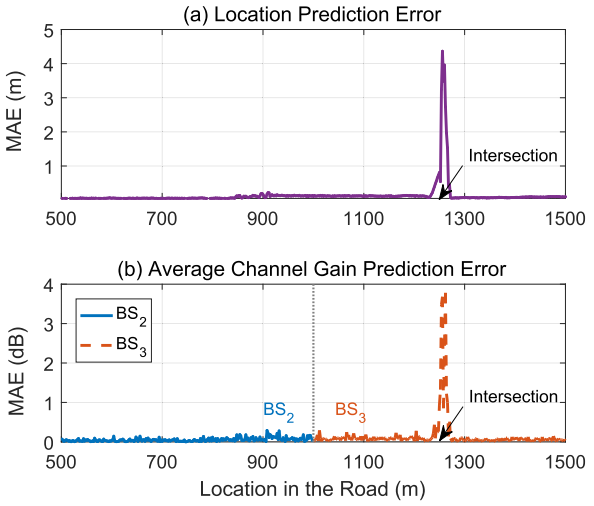


FIGURE 6. MAEs of predicting the user's location and the average channel gains from two BSs.

In Figure 6, we show the mean absolute errors (MAEs) of predicting the user's location and the corresponding average channel gain from the nearest BS, e.g., BS_2 or BS_3, versus the locations in the roads, where the average channel gain is obtained by translating the user's location with the help of the pre-stored radio map at the CP. We can see from Figure 6(a) that the MAE of predicting the user's location is about 0 m when the location is far away from the intersection (at the location of $1,250$ m), while it can reach four meters near the intersection. This is because compared with the random acceleration, the random stopping time is hard to predict. It is shown in Figure 6(b) that the MAE of predicting the average channel gain from the nearest BS is with similar trends to the MAE of predicting the user's location. The maximal MAE of predicting the average channel gain approaches to 4 dB near the intersection, while the prediction MAEs at other locations are nearly 0 dB. This result indicates that only the users near the intersection exhibit large prediction errors.

C. HANDOVER AND SYSTEM PERFORMANCE

In this subsection, we evaluate the handover and system performance of the proposed proactive mobility management scheme.

1) BASELINES

We compare the performance from the following aspects:

- *Reactive vs Proactive Handover:* After the handover is triggered, the reactive handover scheme needs to wait for a duration of TTT to determine whether to execute the handover. While the proactive handover scheme can execute the handover in advance immediately after the handover is triggered by predicting the number of connections in the next frame.
- *Single vs Dual Connectivity:* With single connectivity, each user switches the connected BS between the current serving BS and the neighbor BS. The offset value is

set as 3 dB, and the duration of TTT is set as 200 ms. With the aid of dual connectivity, each user changes its number of connections by establishing or canceling the dual connectivity. Correspondingly, the offset value is set as $\alpha_{\text{sd}}^{\text{th}}$ or $\alpha_{\text{ds}}^{\text{th}}$, and the duration of TTT is set as T_{sd} or T_{ds} frames for the single-to dual or dual-to-single handover.

- *Non-Predictive vs Predictive Resource Reservation:* The resource reservation is optimized based on the measured average channel gains in the current frame for the non-predictive method, and based on the predicted ones in the next frame for the predictive method.

In particular, we compare the proposed method (with legend "Proposed") with three baselines in Table 3.

TABLE 3. Proposed scheme and baselines.

Schemes	Handover		Reservation
Reactive (Single+Non-pre)	Reactive	Single	Non-predictive
Reactive (Dual+Non-pre)		Dual	
Proactive (Dual+Non-pre)	Proactive		
Proposed			Predictive

2) HANDOVER PERFORMANCE

We first evaluate the numbers of handovers of three handover schemes with dual connectivity considering different velocities of mobile users, as shown in Table 4. The proactive handover by predicting the average channel gains in the next frame $\hat{\boldsymbol{\alpha}}_k^f$ and the reactive handover are adopted for comparing with the proposed proactive handover by predicting the number of connections in the next frame $|\hat{\mathcal{C}}_k^f|$. We can see that both the proposed proactive handover and the reactive handover can decrease the number of handovers, hence effectively avoid frequent handovers.

In addition to the number of handovers, we are concerned with the impact of the delay of executing the handover on the service quality of a mobile user. To this end, we show the cumulative distribution functions (CDFs) of the maximal average SINR received from the connected BSs during the TTT for the proposed proactive and reactive single-to-dual handover schemes in Figure 7. The average SINR received from the connected BS in each frame is computed according to (12). We can see that the received maximal average SINR of the proactive handover is evidently higher than that of the reactive handover under two considered mobility

TABLE 4. Comparison of numbers of handovers.

Handover Schemes	Number of handovers		
	10∼15 m/s	15∼20 m/s	20∼25 m/s
Proactive (Predicting $\hat{\boldsymbol{\alpha}}_k^f$)	28	24	22
Proactive (Predicting $\lvert\hat{\mathcal{C}}_k^f\rvert$)	6	6	6
Reactive	6	6	6

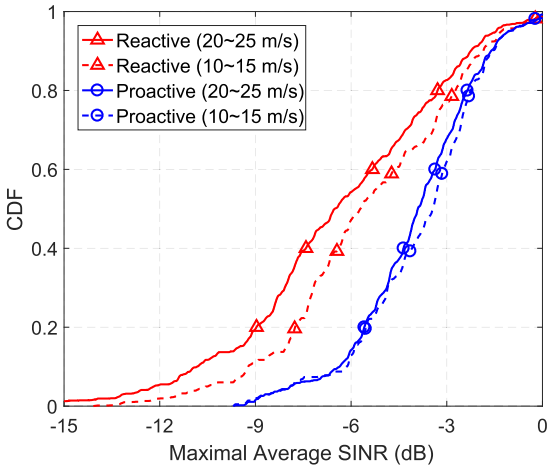


FIGURE 7. CDFs of maximal average SINR in the duration of TTT for the proactive and reactive single-to-dual handovers.

scenarios with different velocities. This is because by using the proactive handover, each user can execute the handover by connecting to two BSs immediately after the handover is triggered, which does not suffer from the delay of TTT. We can also see that for the reactive handover, the maximal average SINR received from the serving BS may drop to less than -12 dB, corresponding to the average rate of 15.89 Kbps when served with one RB. It means that once the minimum rate requirement of the user is high, say 2 Mbps, the user's QoS requirement cannot be satisfied even when allocating all the RBs (i.e., 100 RBs) to this user, leading to the service interruption inevitably. In contrast, the proactive handover can significantly improve the service quality of each mobile user by avoiding the delayed handovers, hence effectively reduce service interruptions.

3) SYSTEM PERFORMANCE

Performance metrics: We consider the following metrics to evaluate the system performance. Users requesting the real-time service may suffer from service outages. Considering the QoS requirement of each user, the outage probability in the fth frame can be estimated as

$$\eta^f = \frac{1}{N_{\text{s}} K^f} \sum_{t=1}^{N_{\text{s}}} \sum_{k \in \mathcal{K}^f} \mathbf{1}(R_k^{f,t} < R_k^{\text{th}}). \tag{24}$$

To evaluate the system performance from either the network perspective or the user fairness perspective, we consider the network rate and the minimal rate with satisfied QoS requirements of users in the fth frame, which are respectively defined as

$$R_{\text{net}}^f = \frac{1}{N_{\text{s}} K^f} \sum_{t=1}^{N_{\text{s}}} \sum_{k \in \mathcal{K}^f} \mathbf{1}(R_k^{f,t} \geq R_k^{\text{th}}) R_k^{f,t}, \tag{25}$$

$$R_{\text{min}}^f = \frac{1}{N_{\text{s}}} \sum_{t=1}^{N_{\text{s}}} \min_{k \in \mathcal{K}^f} \mathbf{1}(R_k^{f,t} \geq R_k^{\text{th}}) R_k^{f,t}. \tag{26}$$

With (24), (25) and (26), the other performance metrics for evaluation are summarized in Table 5, where T_{e} is the number

of frames in the evaluation period to show the performance in long term, and is set as $T_e = 500$ frames. Subscripts "p" and "r" denote the performance of the proposed method and the single-connectivity reactive method with non-predictive resource reservation, respectively. By comparing these two methods, we can obtain the performance gain provided by the joint design of proactive handover, dual connectivity, and predictive resource reservation. All the results are averaged over 100 Monte-Carlo trials. In each trial, K users with the same minimum rate requirement of R^{th} initiate the service at random locations with random initial velocities.

Performance under $\omega = 1$: We first evaluate the performance for maximizing the network rate in each frame under $\omega = 1$. We show the impacts of the minimum rate requirement R^{th} and the total number of served users K on the average network rate $\bar{R}_{\text{net}}$ and the outage probability $\bar{\eta}$ in Figure 8, on the gain of outage probability $G_{\bar{\eta}}$ in Table 6, and on the relative gain of network rate G^f_{net} in Figure 9, respectively.

TABLE 5. Performance metrics.

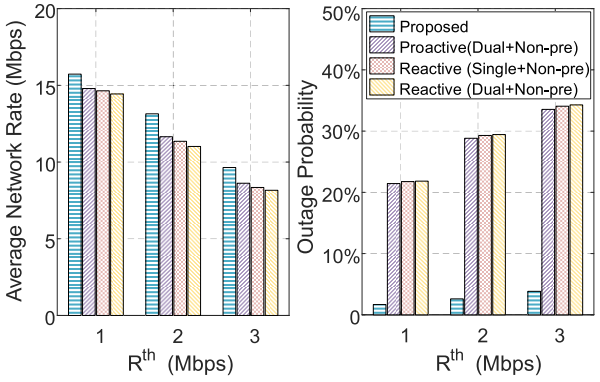
Outage probability	$\bar{\eta} = \frac{1}{T_e}\sum_{f=1}^{T_e} \eta^f$
Average network rate	$\bar{R}_{\text{net}} = \frac{1}{T_e}\sum_{f=1}^{T_e} R^f_{\text{net}}$
Average minimal rate	$\bar{R}_{\text{min}} = \frac{1}{T_e}\sum_{f=1}^{T_e} R^f_{\text{min}}$
Gain of outage probability	$G_{\bar{\eta}} = \bar{\eta}_{\text{r}} - \bar{\eta}_{\text{p}}$
Relative gain of network rate	$G^f_{\text{net}} = \frac{R^f_{\text{net,p}} - R^f_{\text{net,r}}}{R^f_{\text{net,r}}}$
Relative gain of minimal rate	$G^f_{\text{min}} = \frac{R^f_{\text{min,p}} - R^f_{\text{min,r}}}{R^f_{\text{min,r}}}$

From Figure 8, we can see that the proposed scheme outperforms the baselines obviously for arbitrary R^{th} and K. Furthermore, both the average network rates and the outage probabilities of these baselines are very close. It indicates that the performance improvement on the average network rate and the outage probability from proactive handover with dual connectivity is marginal, and optimizing the predictive resource reservation is crucial to boost the network rate meanwhile meet the QoS requirement.

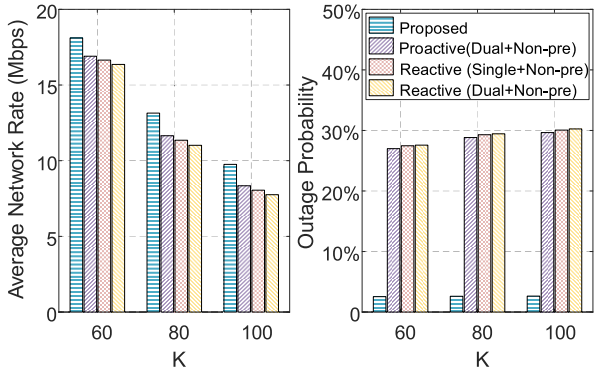
From the gain of outage probability $G_{\bar{\eta}}$ in Table 6 and the complementary cumulative distribution function (CCDF) of the relative gain of network rate G^f_{net} in Figure 9, we can see that with the increasing of R^{th} or K, both $G_{\bar{\eta}}$ and G^f_{net} increase. This is because when increasing R^{th} or K, the system needs to consume more resources to satisfy higher QoS requirement or requirements from more users. With the predictive resource reservation, the proposed method can improve the resource usage efficiency, hence provide larger performance gains in terms of both the outage probability and the network rate.

When $K = 100$ and $R^{\text{th}} = 2$ Mbps, i.e., the traffic load is heavy, the proposed method can reduce 27.4% of outage probability, and improve more than 17% of network rate with the probability of 50%. Compared to the baselines, the proposed method with $\omega = 1$ can reduce service outages significantly meanwhile guarantee higher network rate.

Performance under $\omega = 0$: We next evaluate the performance for maximizing the minimal rate in each frame under



(a) Impact of minimum rate requirement R^{th} ($K = 80$).



(b) Impact of total number of served users K ($R^{\text{th}} = 2$ Mbps).

FIGURE 8. Comparison of average network rates and outage probabilities among different schemes, $\omega = 1$.

TABLE 6. Gain of outage probability, $\omega = 1$.

$G_{\bar{\eta}}$	$K = 60$	$K = 80$	$K = 100$
$R^{\text{th}} = 1$ Mbps	18.2%	20.1%	21.8%
$R^{\text{th}} = 2$ Mbps	24.9%	26.7%	27.4%
$R^{\text{th}} = 3$ Mbps	28.4%	30.2%	32.6%

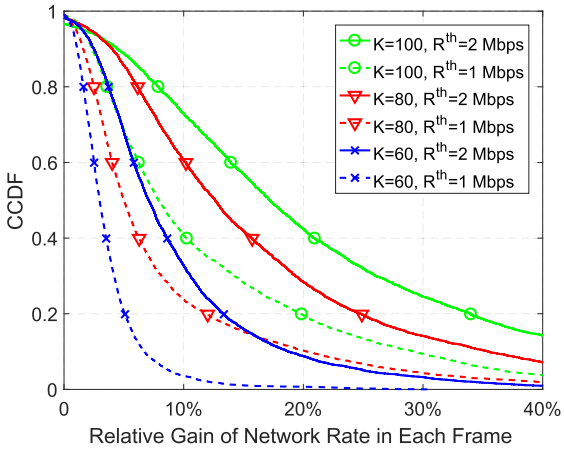
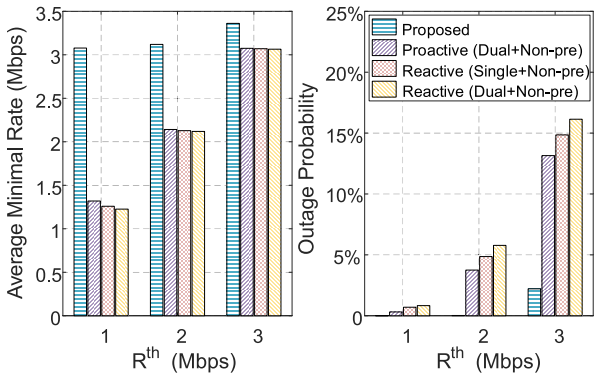
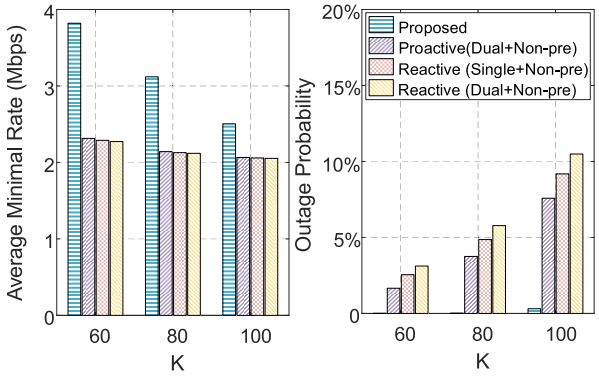


FIGURE 9. CCDF of relative gain of network rate in each frame, $\omega = 1$.

$\omega = 0$. We show the impacts of R^{th} and K on the average minimal rate $\bar{R}_{\text{min}}$ and $\bar{\eta}$ in Figure 10, and on the relative gain of minimal rate G^f_{min} in Figure 11, respectively.



(a) Impact of minimum rate requirement R^{th} ($K = 80$).



(b) Impact of total number of served users K ($R^{\text{th}} = 2$ Mbps).

FIGURE 10. Comparison of average minimal rates and outage probabilities among different schemes, $\omega = 0$.

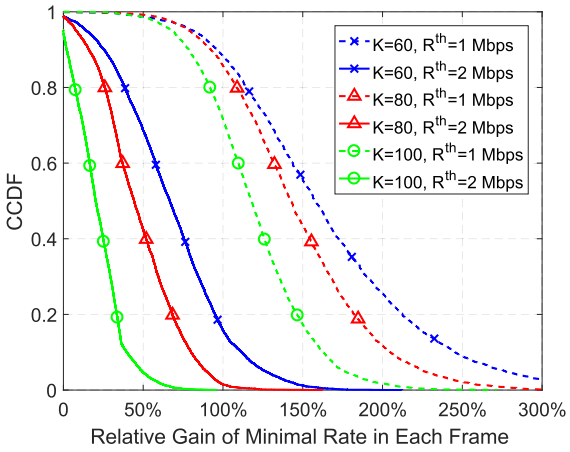


FIGURE 11. CCDF of relative gain of minimal rate in each frame, $\omega = 0$.

As shown in Figure 10, the proposed scheme outperforms the baselines obviously for arbitrary R^{th} and K. By comparing Figures 10(a) and 10(b), we can see that the proposed scheme shows remarkable gain in terms of $\bar{R}_{\min}$ but slight gain in terms of $\bar{\eta}$. This is because the baselines can also achieve low outage probabilities. However, the proposed method can enhance the minimal rate through the predictive resource reservation, which can improve the user fairness efficiently.

From the CCDF of the relative gain of minimal rate $G^f_{\min}$ in Figure 11, we can see that with the decrease of R^{th} or K, $G^f_{\min}$ increases. It means that when R^{th} or K decreases, the system can exploit more residual resources to improve the user fairness. When $K = 60$ and $R^{\text{th}} = 1$ Mbps, i.e., the traffic load is light, the proposed method can avoid outage, and improve more than 160% of minimal rate with the probability of 50%. Compared with the baselines, the proposed method with $\omega = 0$ can improve the minimal rate significantly meanwhile guarantee less service outages.

V. CONCLUSION

In this article, we proposed a dual connectivity-aided proactive handover and resource reservation scheme for mobile users requesting realtime service. The predictive resource reservation was optimized to improve the network throughput or user fairness meanwhile satisfy the minimum data rate requirement of each user. The proactive handover was proposed to avoid frequent and delayed handovers, and reduce the computational complexity for optimizing the resource reservation. Simulation results demonstrated that the predictive resource reservation is crucial to boost data rate while meet the QoS requirement of each mobile user. Compared to the existing reactive schemes, the proposed scheme can either reduce service outages significantly meanwhile support higher network throughput when the traffic load is heavy, or improve user fairness significantly meanwhile incur less service outages when the traffic load is light.

REFERENCES

[1] H. Zhang and L. Dai, "Mobility prediction: A survey on state-of-the-art schemes and future applications," *IEEE Access*, vol. 7, pp. 802–822, 2019.
[2] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann, "Zero-zero mobility: Intra-frequency handovers with zero interruption and zero failures," *IEEE Netw.*, vol. 32, no. 2, pp. 48–54, Mar. 2018.
[3] M. Tayyab, X. Gelabert, and R. Jantti, "A survey on handover management: From LTE to NR," *IEEE Access*, vol. 7, pp. 118907–118930, 2019.
[4] H.-S. Park, Y. Lee, T.-J. Kim, B.-C. Kim, and J.-Y. Lee, "Handover mechanism in NR for ultra-reliable low-latency communications," *IEEE Netw.*, vol. 32, no. 2, pp. 41–47, Mar. 2018.
[5] C. Rosa, K. Pedersen, H. Wang, P.-H. Michaelsen, S. Barbera, E. Malkamaki, T. Henttonen, and B. Sebire, "Dual connectivity for LTE small cell evolution: Functionality and performance aspects," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 137–143, Jun. 2016.
[6] A. Al-Dulaimi, S. Mumtaz, S. Al-Rubaye, S. Zhang, and C.-L. I, "A framework of network connectivity management in multi-clouds infrastructure," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 104–110, Jun. 2019.
[7] M. Centenaro, D. Laselva, J. Steiner, K. Pedersen, and P. Mogensen, "Resource-efficient dual connectivity for ultra-reliable low-latency communication," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.
[8] C. Wang, Z. Zhao, Q. Sun, and H. Zhang, "Deep learning-based intelligent dual connectivity for mobility management in dense network," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–15.
[9] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017.
[10] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using Sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, Sep. 2020.
[11] C. Lee, H. Cho, S. Song, and J.-M. Chung, "Prediction-based conditional handover for 5G mm-wave networks: A deep-learning approach," *IEEE Veh. Technol. Mag.*, vol. 15, no. 1, pp. 54–62, Mar. 2020.

[12] L. Sun, J. Hou, and T. Shu, "Optimal handover policy for mmWave cellular networks: A multi-armed bandit approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.
[13] Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover control in wireless systems via asynchronous multiuser deep reinforcement learning," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4296–4307, Dec. 2018.
[14] C.-Y. Lin, K.-C. Chen, D. Wickramasuriya, S.-Y. Lien, and R. D. Gitlin, "Anticipatory mobility management by big data analytics for ultra-low latency mobile networking," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.
[15] C.-L. I, Q. Sun, Z. Liu, S. Zhang, and S. Han, "The big-data-driven intelligent wireless network: Architecture, use cases, solutions, and future trends," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 20–29, Dec. 2017.
[16] W.-S. Soh and H. S. Kim, "A predictive bandwidth reservation scheme using mobile positioning and road topology information," *IEEE/ACM Trans. Netw.*, vol. 14, no. 5, pp. 1078–1091, Oct. 2006.
[17] A. Nadembega, A. Hafid, and T. Taleb, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2561–2576, Jun. 2015.
[18] C. She and C. Yang, "Energy efficient resource allocation for hybrid services with future channel gains," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 165–179, Mar. 2020.
[19] K. Guo, T. Liu, C. Yang, and Z. Xiong, "Interference coordination and resource allocation planning with predicted average channel gains for HetNets," *IEEE Access*, vol. 6, pp. 60137–60151, 2018.
[20] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013–2026, Jun. 2014.
[21] J. Guo, C. Yang, and C.-L. I, "Exploiting future radio resources with End-to-End prediction by deep learning," *IEEE Access*, vol. 6, pp. 75729–75747, 2018.
[22] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, document TR 36.814 V9.2.0, 3GPP, Mar. 2017.
[23] Z. Zhou, C. Zhang, J. Wang, B. Gu, S. Mumtaz, J. Rodriguez, and X. Zhao, "Energy-efficient resource allocation for energy harvesting-based cognitive machine-to-machine communications," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 3, pp. 595–607, Sep. 2019.
[24] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.
[25] A. Schrijver, *Theory of Linear and Integer Programming*. Hoboken, NJ, USA: Wiley, 1998.
[26] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan, "Distributed energy efficient fair user association in massive MIMO enabled HetNets," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1770–1773, Oct. 2015.
[27] N. Karmarkar, "A new polynomial-time algorithm for linear programming," *Combinatorica*, vol. 4, no. 4, pp. 373–395, Dec. 1984.
[28] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol Specification*, document TS 36.331 V15.8.0, 3GPP, Dec. 2019.
[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[30] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document TS 36.211 V16.3.0, 3GPP, Sep. 2020.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

KAIQIANG QI (Student Member, IEEE) received the B.S. degree in electronics engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), China, in 2015, where he is currently pursuing the Ph.D. degree in information and communication engineering with the School of Electronics and Information Engineering. His recent research interests include the areas of wireless edge caching based on mobile big data and machine learning, and learning-based mobility management.

TINGTING LIU (Member, IEEE) received the B.S. and Ph.D. degrees in signal and information processing from Beihang University (formerly Beijing University of Aeronautics and Astronautics), China, in 2004 and 2011, respectively. From 2008 to 2010, she was a Visiting Student with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. She is currently an Associate Professor with the School of Electronics and Information Engineering, Beihang University. Her research interests include predictive resource management, interference management for future wireless networks. Her Ph.D. thesis received the 2012 Excellent Beijing Doctoral Thesis Award and the 2013 National Excellent Doctoral Thesis Nomination Award.

CHENYANG YANG (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Beihang University [formerly Beijing University of Aeronautics and Astronautics, (BUAA)], China, in 1997. She has been a Full Professor with the School of Electronics and Information Engineering, BUAA, since 1999. She has authored or coauthored over 200 articles and filed over 80 patents in the fields of energy efficient transmission, CoMP, interference management, cognitive radio, and relay. Her recent research interests include wireless caching, wireless big data, and URLLC. She received the 1st Teaching and Research Award Program for the Outstanding Young Teachers of Higher Education Institutions, Ministry of Education of China. She was the Chair of the Beijing Chapter of IEEE Communications Society from 2008 to 2012 and the MDC Chair of APB of the IEEE Communications Society from 2011 to 2013. She has served as a TPC Member, the TPC Co-Chair, or the Track Co-Chair for IEEE conferences. She has served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATION and a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS.

SHIQIANG SUO received the B.S. degree in automation from Tsinghua University, China, in 1999, and the M.A. degree in communication and information system from the China Academy of Telecommunication Technology, China, in 2002. He is currently pursuing the Ph.D. degree in information and communication engineering with Beihang University (formerly Beijing University of Aeronautics and Astronautics). He is currently a General Manager with the New Technology Department, DaTang Mobile Communication Equipment Company, Ltd., Beijing, China. He has authored or coauthored over 450 patents in the fields of 3G, 4G, and 5G mobile communication technologies. His recent research interests include 6G and future new technologies.

YUANFANG HUANG received the B.S. degree in automation from Tsinghua University, China, in 1997, and the M.A. degree in communication and information system from the China Academy of Telecommunication Technology, China, in 2003. She is currently a Senior Engineer with DaTang Mobile Communication Equipment Company, Ltd., Beijing, China. Her main research interests include 5G wireless communication and RAN intelligent control.