

Received June 2, 2020, accepted June 27, 2020, date of publication July 6, 2020, date of current version July 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007193

# Joint Resource Allocation, User Association, and Power Control for 5G LTE-Based Heterogeneous Networks

JAIN-SHING LIU<sup>1</sup>, (Member, IEEE), CHUN-HUNG RICHARD LIN<sup>2</sup>,  
AND YU-CHEN HU<sup>3</sup>, (Senior Member, IEEE)

<sup>1</sup>Department of Computer Science and Information Engineering, Providence University, Taichung City 43301, Taiwan

<sup>2</sup>Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan

<sup>3</sup>Department of Computer Science and Information Management, Providence University, Taichung City 43301, Taiwan

Corresponding author: Chun-Hung Richard Lin (lin@cse.nsysu.edu.tw)


This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-126-003-MY2.

**ABSTRACT** The aim of 5G wireless networks to provide Mbps and Gbps data rates to end users is expected to be fulfilled by the advanced technologies such as multi-input multi-output (MIMO), carrier aggregation (CA), inter/intra-cell communication, and adaptive modulation and coding techniques, which would be all realized in the Long Term Evolution-Advanced (LTE-A) heterogeneous network constituted by macrocells (MCs) and small cells (SCs) adopting these 5G advanced techniques. Given the potential of significantly increasing the network performance, the resource allocation (RA) problem involved becomes harder than ever especially when MIMO and CA are included in the RA problem involving multiple types of resources to be concurrently determined for the global optimization. Facing this challenge, we develop a framework to jointly optimize energy efficiency (EE), spectrum efficiency (SE), and queue length for downlink transmissions with an overall and comprehensive consideration of dynamically allocating resource blocks (RBs), component carriers (CCs), modulation and coding schemes (MCSs), and deciding user association (UA) with a power control (PC) mechanism on discrete power levels (PLs) in the heterogeneous LTE-based MIMO wireless networks. Specially, for the complex joint RA, UA, and PC problem, we conduct a mixed integer programming model to accommodate the stochastic optimization problem involved with the drift-plus-penalty (DPP) approach for Lyapunov opportunistic optimization. In particular, although it involves a nondeterministic polynomial time (NP) problem, we can still show a reduced problem to be solved easily through linear relaxation when its coefficient matrix is totally unimodular (TUM), and to be solved efficiently as well even when the TUM property is not guaranteed. Based on the reduction, we further develop a distributed or semi-distributed algorithm operated on two levels to approach the optimal results with lower complexity if the UA requirement can be relaxed. Finally, apart from exhibiting its performance on the weighting parameters, the numerical experiments also show our approach to make a good tradeoff among SE, EE, and queue length, and outperform the greedy-based state-of-the-art algorithms.

**INDEX TERMS** LTE-A heterogeneous wireless networks, MIMO, carrier aggregation, multi-resource allocation, user association, power control.

## I. INTRODUCTION

The key techniques for 5G such as multiple-input multiple-output (MIMO), carrier aggregation (CA), inter/intra-cell communication, and adaptive modulation and coding techniques would be all realized in the Long Term Evolution-Advanced (LTE-A) heterogeneous network constituted by

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

macro-cells (MCs) and small-cells (SCs) to fulfill the rapid demand of smart phones and mobile internet services. Given that, the long term evolution-advanced (LTE-A) standard continuously evolves to support the very high data rates required by the international mobile telecommunications-advanced (IMT-Advanced) systems [1]. In future 5G networks, dense deployment of small cells (SCs) such as picocells and femtocells has been envisioned to improve the overall network capacity, spectrum efficiency (SE), and

energy efficiency (EE). In particular, the 5G cellular wireless systems with a multi-tier architecture consisting of MCs and different types of SCs are expected to serve user equipments (UEs) with varying quality-of-service (QoS) requirements. Resource allocation (RA) in such 5G systems will be extremely complex due to many different types of resources to be concurrently allocated in the irregular and pseudo-random network topology, and the existing management approaches may not be sufficient.

Specifically, for a LTE-A heterogeneous network wherein several SCs are deployed within a MC service area, how to allocate the radio resources involved while deciding the UE-to-cell association (UA) and the power levels for all transmitters would be a major issue especially when MC and its SCs share the radio resources from the same service provider. To resolve this issue, SCs can use their own frequency bands different from those of MC [2]. However, the easy solution of dedicated radio usage would sacrifice SE. Alternatively, SCs can totally or partially use the same frequency bands of MC as noted in [3] adopting a co-channel deployment or a partial co-channel deployment to manage the radio resources. However, it could cause cross-tier interference, and hence certain partial co-channel deployments were considered in [4], [5] to alleviate this difficulty. Apart from the above, the heterogeneous network is also complicated by the 5G advanced techniques such as carrier aggregation (CA) and multi-input multi-output (MIMO) which are anticipated as the vital breakthrough that is necessary for 5G [6]. Specifically, they will continue to serve as the key techniques of 5G because CA can aggregate several component carriers (CCs) to support high data rate transmission, and MIMO can enhance signal to noise ratio (SNR) through transmit diversity (TD) or increase data rate through spatial multiplexing (SM).

In the heterogeneous and complex environment, RA would be tremendously difficult due to many unique constraints from these techniques to be satisfied simultaneously and the existing RA schemes can hardly be sufficient. For instance, the previous works [7], [8] addressed the problem of downlink radio resource allocation with CA by employing load balancing mechanisms that assign CCs to UEs first and then schedule RBs of CCs for them at every transmission time interval (TTI) to optimize the radio resource usage. Given their contributions, these works, however, consider only heuristic or greedy approaches, which provide no performance guarantees and no benefits brought by MIMO. In addition, even though more sophisticated greedy algorithms such as those in [9], [10] have been proposed to guarantee performance lower bounds, they still involve no MIMO. In fact, most of the RA researches for CA-enabled LTE networks [11]–[13] assumed CC selection, resource block (RB) allocation, and MCS assignment as completely separate problems, which may lead to the degradation of network performance. By jointly considering these problems in a framework, the related works [9], [10], [14]–[16] exhibited more complete solutions among the others. However, they

still lack to consider MIMO [9], [10], [14], [15], heterogeneous network with multiple cells [9], [15], [16], or power control [9], [10], [15], [16].

Taking these issues into account, in this work we study a joint RA, UA, and PC problem for downlink transmissions in 5G LTE-A heterogeneous wireless networks. In particular, we consider discrete power control to reflect the fact that 3GPP LTE cellular networks only support discrete power levels (PLs) in the downlink via a user-specific data-to-pilot-power offset parameter [1], which would be still useful in the 5G framework. Further, if the discrete power allocation (PC) and the UE-to-cell association (UA) are referred to as a kind of resource allocation, respectively, then RA, UA, and PC in this work could be collectively denoted by UE/RB/CC/MCS/cell/PL allocation problem or multi-resource allocation (MRA) problem for short. Our objective is then to jointly optimize EE, SE and queue length in the long-term, under the constraints resulted from the MRA problem, in addition to those from the LTE. For example, we would take into account the constraint that a UE can be only exclusively served by a single cell, and the constraint that the same RB can not be allocated to two different cells if there is interference between the two cells, which would avoid complicating the power control.

Given the various constraints to be addressed concurrently, the joint stochastic optimization problem is expressed here as a mixed integer programming problem via a transformation, whose solution typically requires prohibitive time complexity. As far as we know, the long-term metrics of EE, SE, and queue length have not been jointly investigated in the LTE-based heterogeneous wireless networks subject to all the specific constraints including those just being exemplified. In particular, although SE, EE, or their relationship had been investigated in, e.g., [17]–[19], these related works typically presume static channel state and infinite queue length. Based on these assumptions, the channel state information (CSI) and the queue state information (QSI) could be ignored, and the arrival traffic may not be transmitted in time and the corresponding queue length would accumulate unboundedly when the data is given in burst without the awareness of CSI. Unlike the above, in this work we try to maximize EE and SE while guaranteeing the network stability in the heterogeneous networks, wherein RA, UA and discrete PC are jointly considered to be a complex MRA problem. Apart from the MRA that is a combinatorial problem to be solved, our work involves also the stochastic nature caused by varying channel and traffic state. For the combined difficulty, in the high layer, we formulate the joint optimization problem as a stochastic optimization problem and resolve it through the drift plus penalty (DPP) approach in the framework of Lyapunov optimization to accommodate time-varying channel conditions and traffic arrivals without prior knowledge of them. In the lower layer, we show that even it is NP in general, the MRA problem can still be easily solved through linear relaxation when its coefficient matrix is totally unimodular (TUM). Then, inspired by the TUM

property, we reduce this problem and develop a distributed or semi-distributed algorithm operated on two levels to approach the optimal result with lower computational complexity if the UA requirement can be relaxed. Finally, to know its performance, we conduct numerical experiments, showing that our approach can make a good tradeoff among EE, SE and queue length, and outperform the greedy-based start-of-the-art algorithms, while showing the performance metrics varied by the weighting parameters. Specially, it is also exhibited that the two-level MRA algorithm proposed would obtain a solution allowing the system to achieve more than half of the optimal spectrum efficiency (SE) and throughput, and approach the optimal energy efficiency (EE) while maintaining a larger queue length, which represents a significant performance gain against the computation cost decreased from nondeterministic polynomial (NP) to polynomial (P). As a summary, the characteristics of this work can be outlined as follows:

- Unlike existing studies on LTE-based wireless networks where the performance metrics such as SE, EE, and delay, the system techniques such as MIMO, CA, and PC, or both, are partially considered, in this work a synthetic framework is proposed to jointly consider EE, SE, and queue length in the 5G LTE-based heterogeneous networks equipped with MIMO, CA, and PC. To also account for the time-varying channel and traffic, the joint design is formulated as a stochastic multi-objective optimization (MOO) problem subject to the constraints on the network stability, the constraints from the multi-cell environment, and the unique RA rules for LTE.
- A Lyapunov DPP technique is adopted to transform the MOO problem to a mixed integer linear programming problem. Further, a linear programming model is found to easily solve the high-dimensional MRA problem involved when the coefficient matrix is totally unimodular (TUM) in a reduced model. Based on this model, a distributed two-level MRA algorithm is proposed for more computationally efficient solutions to the NP allocation problem, in addition to the joint optimization algorithm developed for the whole stochastic MOO problem.
- Apart from the Lyapunov approach to guarantee the network stability even without prior knowledge of the system state on channel and traffic, this work is also aided by the weighted sum method that introduces different weights to the objective function to make an optimal tradeoff among SE, EE, and queue length.

The remainder of this paper is organized as follows. First, in Sec. II we introduce the system model and formulate our problem in terms of transmission modes for LTE, energy efficiency model, and scheduling constraints. Then, in Sec. III we adopt the weighted sum method to formulate the MOO problem, and use the Lyapunov DPP approach to optimize EE, SE, and queue length subject to the various constraints involved. Following that, a distributed two-level RA algorithm is proposed in Sec. IV to resolve the NP problem if the

UA requirement can be relaxed. The performance analysis and numerical experiments are given in Secs. V and VI, respectively, and finally conclusions are drawn in Sec. VII.

## II. SYSTEM MODEL AND PROBLEM FORMULATION

In this work, we consider 5G LTE-based networks with a multi-tier and multi-cell heterogeneous architecture, as exemplified in Fig. 1, which consists of  $|\mathcal{S}|$  base stations (BSs) (including a micro base station (MBS) and  $|\mathcal{S}| - 1$  small base stations (SBSs)), and  $|\mathcal{U}|$  UEs located in their service areas. A number of  $|\mathcal{C}|$  CCs obtained with CA are deployed in the environment, and without loss of generality, each CC has the same number of  $|\mathcal{B}|$  RBs to be allocated. Then,  $|\mathcal{L}|$  MCSs are dedicated to each RB for transmission. Similarly,  $|\mathcal{P}|$  discrete power levels (PLs) denoted by  $P = \{\sigma_1, \sigma_2, \dots, \sigma_{|\mathcal{P}|}\} \times P_{max}$  are designed for each BS, where  $0 < \sigma_1 < \sigma_2, \dots, < \sigma_{|\mathcal{P}|} = 1$  and  $P_{max}$  denotes the maximum power.

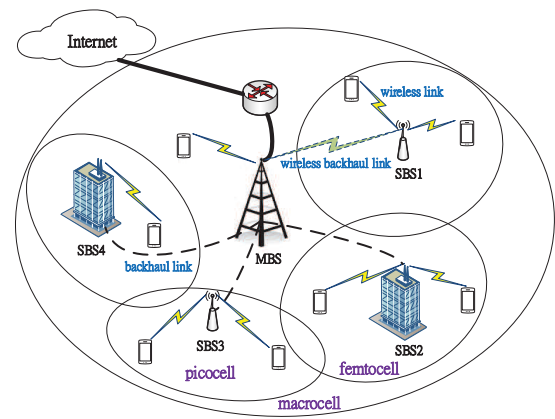


FIGURE 1. An example of the heterogeneous network.

### A. TRANSMISSION MODE

As specified in LTE [20], there are different multi-input multi-output transmission modes (TMs) to be used in the system, which can be distinguished in terms of the antenna mapping and the type of CSI feedback adopted. In this work, we consider transmit diversity (TD) and spatial multiplexing (SM) as the two major types of TMs usually thought for the future 5G network development. As specified, TD sends the same data via different antennas, and each antenna stream uses different coding, which could enhance the signal-to-noise ratio (SNR) and reduce the block error rate (BLER). Specially, in the case of two antennas, TD could be done based on space-frequency block coding (SFBC), and in the case of four antenna ports, TD can be realized through a combination of SFBC and frequency-switched transmit diversity (FSTD) [21], [22]. Different from the above, SM supports spatial multiplexing of two to four layers that are multiplexed to two to four antennas respectively. In the specification, the open loop spatial multiplexing does not rely on pre-coding matrix indicator (PMI) being reported by UE and selects PMI based on a predefined method, while the closed loop

spatial multiplexing selects PMI based on the CSI feedback of UE [21]. In this work, the latter is assumed for the MRA problem to be introduced. Specifically, with TD, RBs are allocated from a single transport block (TB) per carrier component (CC). In contrast, with SM, RBs are allocated from two TBs per CC. Taking both into account, we define a transmission mode table (Table 1) for our MRA problem wherein each index specifies a specific TM mode used and the modulation and coding scheme (MCS) per TB adopted. As shown therein, 29 MCSs defined in the 3GPP LTE standard [23] are considered in this work.

TABLE 1. The index table for transmission modes.

Index	TM	MCS for TB 1	MCS for TB 2
1	TD	0	×
2	TD	1	×
⋮	⋮	⋮	⋮
29	TD	28	×
30	SM	0	0
31	SM	0	1
⋮	⋮	⋮	⋮
869	SM	28	27
870	SM	28	28

For such a stochastic transmission system wherein traffic arrivals and channel conditions are both time-varying, we propose an online algorithm that can dynamically resolve the stochastic optimization problem with the Lyapunov DPP approach to achieve the system stability while maximizing the network utilization. To this end, we first introduce the channel and power model, the spectrum and energy efficiency, and the queueing dynamic. Then, we formulate the constraints specific to the system, and present a complete programming model for the optimization problem addressing the complex RA, UA, and PC issues to be involved.

**B. CHANNEL AND POWER MODEL**

For high-rate networks with reduced degree of mobility, it is vital for a resource allocation (RA) algorithm is conducted to accommodate a slow fading network wherein channel conditions would remain unchanged during an allocation period (Ch. 6 of [24]). Accordingly, for downlink transmissions in the networks, the signal to interference and noise ratio (SINR) from BS *s* to UE *u* using RB *b* of CC *c* at PL *p* in time *t* can be represented by

$$SNR_{s,u}^{c,b,p}(t) \triangleq \frac{P_{s,u}^p(t) |h_{s,u}^{c,b}(t)|^2 d_{s,u}^{-\rho}(t)}{N_{s,u}^{c,b}(t) + I_{s,u}^{c,b}(t)} \quad (1)$$

where the channel gain from BS *s* to UE *u* using RB *b* of CC *c* is denoted by  $h_{s,u}^{c,b}$ , the distance from *s* to *u* by  $d_{s,u}$ , and the path-loss factor by  $\rho$ . In addition, when *s* transmits to *u* on RB *b* of CC *c*, the noise on *u* is denoted by  $N_{s,u}^{c,b}$ . The channel is supposed to be Rayleigh fading and its gain to be exponentially distributed, and further, an empirical downlink SINR to channel quality indicator (CQI) mapping for LTE is adopted

to estimate the CQIs returned to BSs [25]. If MBS is in charge of RA and these CQIs are collected, MBS would decide each MCS index  $\ell(u, c, b, s, p)$  for the downlink transmission from BS *s* to UE *u* using RB *b* of CC *c* at PL *p*. Then, it can transmit the decision to all SBSs it associates. Here, based on the 3GPP specification [26], the data rate would be represented by  $r_l$  through a mapping table for each RB based on MCS *l* in the index  $\ell$ . Let  $\Omega_{u,c,b,s,p}$  be the index of the highest-rate MCS on *u, c, b, s,* and *p*. As  $\Omega_{u,c,b,s,p}$  is the highest MCS to be obtained, the achieved transmission rate  $v_{u,c,b,\ell,s,p}$  on  $\ell$  would be  $\sum_l r_l$  where  $r_l$  is the data rate corresponding to MCS *l* in TB  $i \in \{1, 2\}$  in the entry of index  $\ell$ , and it would be 0 if *l* is not available in the TM, or exceeds  $\Omega_{u,c,b,s,p}$  which would result in unacceptable error rate. Providing a RA matrix **X** which accommodates RA, UA, and PC through a high-dimensional (6-dimensional) representation instead of defining different variables for different metrics brought by RA, UA, and PC, respectively, to unnecessarily complicate its representation, the total data rate can be simply obtained by  $R_{tot}(t) = \sum_{\forall u,c,b,\ell,s,p} (x_{u,c,b,\ell,s,p}(t) \times v_{u,c,b,\ell,s,p}(t))$ . Similarly, the total power consumption can be obtained by  $P_{tot}(t) = \sum_{\forall u,c,b,\ell,s,p} (x_{u,c,b,\ell,s,p}(t) (P_{s,u}^p(t) + P_{s,u}^c))$ , where  $P_{s,u}^p(t)$  denotes the transmit power from *s* to *u* at power level *p*, and  $P_{s,u}^c$  denotes the constant circuit power for *s* and *u*.

**C. SPECTRUM/ENERGY EFFICIENCY AND QUEUEING DYNAMIC**

Unlike the previous works focusing on SE, EE, or both at the moment of observation [27], [28], in the stochastic system with channel conditions and traffic arrivals to be time-varying, we pay our attention to the limits of the time-average expectations of these metrics. Specifically, the long-term time-averaged expected transmission data rate and energy consumption can be represented by

$$\bar{R}_{tot} = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R_{tot}(\tau)\} \quad (2)$$

$$\bar{P}_{tot} = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{P_{tot}(\tau)\} \quad (3)$$

Here, the data rate  $R_{tot}$  is normalized by the channel bandwidth. The spectrum efficiency (SE) is defined as the long-term average data rate on all the transmissions in (2), i.e.,

$$\eta_{SE} = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R_{tot}(\tau)\} = \bar{R}_{tot} \quad (4)$$

Following that, the energy efficiency is defined as the ratio of the long-term aggregated data rate (i.e.,  $\eta_{SE}$ ) to the long-term total energy consumption. That is,

$$\begin{aligned} \eta_{EE} &= \frac{\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R_{tot}(\tau)\}}{\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{P_{tot}(\tau)\}} \\ &= \frac{\eta_{SE}}{\sum_{\forall u,c,b,\ell,s,p} (x_{u,c,b,\ell,s,p} (\bar{P}_{s,u}^p + P_{s,u}^c))} \end{aligned} \quad (5)$$

where  $\bar{P}_{s,u}^p = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{\tau=0}^{T-1} P_{s,u}^p(\tau)$  denotes the average transmit power from  $s$  to  $u$  at power level  $p$ .

Apart from SE and EE, the queueing delay (or queue length) is also jointly optimized in this work. For this, the system is considered to be time-slotted, and the downlink traffics to UEs at time  $t$  are aggregately represented by  $\mathbf{A}(t) \triangleq (A_1(t), \dots, A_{|\mathcal{U}|}(t))$ , which are independently and identically distributed over  $t$  with  $\mathbb{E}\{\mathbf{A}(t)\} = \lambda \triangleq (\lambda_1, \dots, \lambda_{|\mathcal{U}|})$ . In addition, to ensure the system stability, it is assumed that  $A_u(t)$  will not exceed a peak or maximum value  $A_u^{max}$ , i.e.,  $0 \leq A_u(t) \leq A_u^{max}, \forall u \in \mathcal{U}$  and  $\forall t \geq 0$ . The assumption is based on the fact that the statistics of  $\mathbf{A}(t)$  are usually unknown and the capacity region involved is hard to estimate in a practical system. In the real situation, a flow control is generally required to limit the admitted traffic  $R_u(t)$  to be lower than the arrival  $A_u(t)$  for the system stability. Thus, an admission control algorithm is required here to determine  $R_u(t)$  from  $A_u(t)$ , and a BS is conducted to make RA decisions to provide link rates  $\mu_u(t)$  to serve the admitted traffics. With the admitted traffic  $R_u(t)$  and the service rate  $\mu_u(t)$ , the data queue dynamic for UE  $u \in \mathcal{U}$  can then be formulated as

$$Q_u(t+1) = \max\{Q_u(t) - \mu_u(t), 0\} + R_u(t) \quad (6)$$

In the above, the queue is considered stable if it has a bounded time-averaged backlog and finite average queueing delay [29]. According to Little's law [30], the average delay would be proportional to the average queue length with a specific arrival rate. Thus, it could use the queue length and further the queue stability to describe the delay. Accordingly, the strong stability for the average data queue length on  $u$  would be considered for the system, defined as

$$\bar{Q} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{u \in \mathcal{U}} \mathbb{E}\{Q_u(t)\} < \infty \quad (7)$$

### D. MULTIPLE RESOURCE ALLOCATION AND ITS CONSTRAINTS

For the MRA problem, we use  $u \in \mathcal{U}$  to denote a UE,  $c \in \mathcal{C}$  a CC,  $b \in \mathcal{B}$  an RB,  $\ell \in \mathcal{L}$  an MCS index,  $s \in \mathcal{S}$  a cell, and  $p \in \mathcal{P}$  a PL, as already appeared in the index of  $x$ . At each TTI  $t$ , the system aims to allocate UEs/CCs/RBs/MCSs/cells/PLs simultaneously or decide  $x_{u,c,b,\ell,s,p}(t)$  to maximize the network utility. Because the channel condition is time-varying, each UE is conducted to use the reference signals from MBS or SBSs for estimating the channel condition, and accordingly, to transmit its CQIs to MBS or SBSs. Then, a CQI for RB can be mapped to the highest-rate MCS for a UE using the RB [26], and hence the channel conditions on all UEs and RBs can be perceived by the system through the CQIs reported [9]. All cells can eventually share the channel state information (CSI) for the optimization with each other via a backhaul network.

Unlike the previous works [9], [10], [15] paying no attention to MIMO, our system accommodates 2 different MIMO transmission modes (TMs), i.e., transmit diversity (TD) and

spatial multiplexing (SM). As noted before, the TM indices for both selected MIMO TM per CC and MCS per TB are summarized in Table 1, wherein the first 29 MCS indices are given for TD since only a single TB per CC is considered while  $29 \times 29 = 841$  indices starting at 30 are given for SM as two TBs per CC will be used in this TM.

Providing that, for scheduling the multiple resources in the network with CA and MIMO, we have the following constraints which ignore the time index  $t$  for brevity. First, as the basic unit for transmission, RB can at most be assigned to a single UE  $u$  on a certain MCS  $\ell$ , and represented by

$$\sum_{\forall p \in \mathcal{P}} x_{u,c,b,\ell,s,p} \leq y_{u,c,\ell,s}^1, \quad \forall u \in \mathcal{U}, \quad \forall c \in \mathcal{C}, \quad \forall b \in \mathcal{B}, \quad \forall \ell \in \mathcal{L}, \quad \forall s \in \mathcal{S} \quad (8)$$

In the left hand side of (8), the summation on all  $p$ , i.e.,  $\sum_{\forall p \in \mathcal{P}}$ , is used to ensure that each RB can be transmitted at only one power level  $p$ . In the right hand side,  $y_{u,c,\ell,s}^1$  is defined to be an auxiliary binary variable representing a CC allocation, where  $y_{u,c,\ell,s}^1 = 1$  denotes that CC  $c$  is assigned to UE  $u$  in cell  $s$  with TM index  $\ell$ . Complying with LTE, all RBs allocated to UE  $u$  in CC  $c$  should have identical  $\ell$ , i.e.,

$$\sum_{\forall \ell \in \mathcal{L}} y_{u,c,\ell,s}^1 \leq 1, \quad \forall u \in \mathcal{U}, \quad \forall c \in \mathcal{C}, \quad \forall s \in \mathcal{S} \quad (9)$$

Next, a monopoly principle specific to the multi-cell environment is considered that a UE  $u$  can only be served by a single cell  $s$ , and given that, it can not be served by the other cells  $s' \in \mathcal{S} \setminus s$ . To realize the above in a linear form, we have the following two constraints. The first is

$$\sum_{\forall \ell \in \mathcal{L}, \forall p \in \mathcal{P}} x_{u,c,b,\ell,s,p} \leq y_{u,s}^2, \quad \forall u \in \mathcal{U}, \quad \forall c \in \mathcal{C}, \quad \forall b \in \mathcal{B}, \quad \forall s \in \mathcal{S} \quad (10)$$

where  $\sum_{\forall \ell \in \mathcal{L}}$  on  $x$  is used to denote that each RB can be transmitted with only one TM index in addition to the fact that only one power level is adopted for the transmission as already noted previously by  $\sum_{\forall p \in \mathcal{P}}$  in (8). In addition,  $y_{u,s}^2$  is an auxiliary binary variable used to represent an RB allocated to UE  $u$  in cell  $s$  if its value is 1, and 0 otherwise. Given that, the monopoly principle that a UE  $u$  can only be served by a single cell  $s$ , is further enforced by the second constraint:

$$\sum_{\forall s \in \mathcal{S}} y_{u,s}^2 \leq 1, \quad \forall u \in \mathcal{U} \quad (11)$$

In addition, to reduce the inter-cell interference, another monopoly principle about the multi-cell would be also specified that if an RB  $b$  of CC  $c$  is already allocated to a cell  $s$ , it cannot be assigned to its neighboring cells  $s' \in \mathcal{N}_s$  which would cause significant interferences to  $s$ . That is, a specific RB can be either allocated in a cell  $s$  or its neighboring cells  $s'$ , but not both. This involves a logical either-or constraint which can be transformed to regular linear constraints with the aid of an auxiliary binary variable,  $y_{c,b,s}^3$ , to denote whether RB  $b$  of CC  $c$  can be allocated to cell  $s$  or not.

Accordingly, there are two conditions to be specified. The first is for RB  $b$  of CC  $c$  allocated to cell  $s$ , denoted by

$$\sum_{\forall u \in \mathcal{U}, \forall \ell \in \mathcal{L}, \forall p \in \mathcal{P}} x_{u,c,b,\ell,s,p} \leq 1 - y_{c,b,s}^3, \quad \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S} \quad (12)$$

The second is for RB  $b$  of CC  $c$  allocated to its neighboring cells  $s' \in N_s$ , which can be similarly denoted by

$$\sum_{\forall u \in \mathcal{U}, \forall \ell \in \mathcal{L}, \forall p \in \mathcal{P}} x_{u,c,b,\ell,s',p} \leq y_{c,b,s}^3, \quad \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S}, \forall s' \in N_s \quad (13)$$

Further, it is worth noting that in a LTE-based network, the number of CC allocated to cell  $s$  would be limited to a certain number, say  $f_s$ . For example, a UE of LTE 8/9 can only use 1 CC while a LTE UE is allowed to use 2 CCs. Specifically, the cardinality constraint can be realized by the following inequalities. The first is

$$\sum_{\forall u \in \mathcal{U}, \forall b \in \mathcal{B}, \forall \ell \in \mathcal{L}, \forall p \in \mathcal{P}} x_{u,c,b,\ell,s,p} \leq y_{c,s}^4, \quad \forall c \in \mathcal{C}, \forall s \in \mathcal{S} \quad (14)$$

where  $y_{c,s}^4$  is a binary variable whose value 1 denotes CC  $c$  being allocated to cell  $s$ , and 0 otherwise. With the aid of auxiliary variable  $y_{c,s}^4$ , the cardinality of  $f_s$  is further enforced by the second constraint:

$$\sum_{\forall c \in \mathcal{C}} y_{c,s}^4 \leq f_s, \quad \forall s \in \mathcal{S} \quad (15)$$

Apart from cell in the above, UE could have its own cardinality constraint on CC as well; that is, each UE  $u$  can be allocated at most  $d_u$  CCs for its transmission. Similarly, it can be realized by linear inequalities, and the first is

$$\sum_{\forall b \in \mathcal{B}, \forall \ell \in \mathcal{L}, \forall s \in \mathcal{S}, \forall p \in \mathcal{P}} x_{u,c,b,\ell,s,p} \leq y_{u,c}^5, \quad \forall u \in \mathcal{U}, \forall c \in \mathcal{C} \quad (16)$$

where  $y_{u,c}^5$  is a binary variable whose value 1 denotes CC  $c$  being allocated to UE  $u$ , and 0 otherwise. Then, with  $y_{u,c}^5$ , the cardinality constraint on  $d_u$  can be finalized by

$$\sum_{\forall c \in \mathcal{C}} y_{u,c}^5 \leq d_u, \quad \forall u \in \mathcal{U} \quad (17)$$

### E. SERVICE RATE, THROUGHPUT, AND POWER CONSUMPTION

Providing the MRA satisfying the above constraints, the service rate  $\mu_u$  (adopted in (6)) can be obtained by

$$\mu_u(t) = \sum_{\forall c,b,\ell,s,p} x_{u,c,b,\ell,s,p}(t) \times v_{u,c,b,\ell,s,p}(t), \quad \forall u \in \mathcal{U} \quad (18)$$

Given that, the total data rate introduced previously can be also represented by  $R_{tot}(t) = \sum_u \mu_u(t)$ . Clearly, there are two parameters directly impacting the data queue dynamic (6). The first is the data rate  $\mu_u(t)$  just given. The second is the throughput  $r_u(t) \triangleq \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R_u(\tau)\}$  that denotes

the admitted and transmitted data rate for  $u$ . In this work, the time-average throughput  $r_u$  in the long term serves as one of the performance metrics, which should be higher than the requirement  $R_u^{req}$  from  $u$ . Taking these into account, we have

$$r_u \triangleq \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R_u(\tau)\} \geq R_u^{req} \quad (19)$$

Clearly, in the long term, the time-average throughput  $r_u$  will not exceed the time-average arrival rate  $\lambda_u$ , i.e.,

$$0 \leq r_u \leq \lambda_u \quad (20)$$

Similarly, in the short term, the throughput of  $u$  at time  $t$ , i.e.,  $R_u(t)$ , can not exceed its arrival rate  $A_u(t)$ , i.e.,

$$0 \leq R_u(t) \leq A_u(t) \leq A_u^{max} \quad (21)$$

In parallel with the above, the power consumption of  $u$  at time  $t$  can be obtained by  $P_s(t) = \sum_u P_{s,u}^p(t) + P_{s,u}^c(t)$ , and in the long term, it can not exceed the maximum  $P_s^{max}$  as well. That is,

$$p_s \triangleq \lim_{t \rightarrow \infty} \frac{1}{t} P_s(t) \leq P_s^{max} \quad (22)$$

### F. PROBLEM FORMULATION

As mentioned before, our work is to maximize SE and EE simultaneously constrained by the queue stability, involving the three key performance metrics (SE, EE, and queue length) to be optimized at the same time. Here, as the average queue length links the stability and the delay, the queueing delay can be managed by investigating the queue stability. Apart from the queue length or delay to be enforced by using constraints, SE and EE representing our research targets would serve as the objectives in the stochastic MOO problem. However, due to the different units between SE and EE, it is more conveniently considered to maximize the normalized data rate  $\bar{R}_{tot}/R_{max}$  and to minimize the normalized total power consumption  $\sum_{\forall u,c,b,\ell,s,p} (x_{u,c,b,\ell,s,p}(\bar{P}_{s,u}^p + P_{s,u}^c))/P_{max}$ , where  $R_{max}$  denotes the maximum data rate and  $P_{max}$  the maximum power in the system. Now, given the throughput constraints in (19)-(21), the power consumption constraint in (22), and the scheduling constraints in (8)-(17) in addition to the queue stability constraint, we can formulate the MOO problem with the following programming model:

$$\begin{aligned} & \text{Maximize} \quad \frac{\bar{R}_{tot}}{R_{max}} \text{ and} \\ & \text{Maximize} \quad - \frac{\sum_{\forall u,c,b,\ell,s,p} x_{u,c,b,\ell,s,p}(\bar{P}_{s,u}^p + P_{s,u}^c)}{P_{max}} \\ & \text{subject to} \quad \text{C1: } \bar{Q} < \infty \\ & \quad \text{C2: } r_u \geq \mathcal{R}_u^{req} \quad \forall u \\ & \quad \text{C3: } p_s \leq P_s^{max} \quad \forall s \\ & \quad \text{C4: } 0 \leq r_u \leq \lambda_u, \quad \forall u \\ & \quad \text{C5: } 0 \leq R_u(t) \leq A_u(t) \leq A_u^{max}, \quad \forall u, \forall t \\ & \quad \text{C6: } (8) - (17), \quad \forall t \end{aligned} \quad (23)$$

As shown readily, it is a highly challenging stochastic optimization problem involving a large amount of stochastic information on channel conditions and traffic arrivals to be considered, and a high-dimensional variable representation on 6 different types of resources to be determined. This requires an online control and scheduling algorithm to obtain the solutions within a reasonable time limit. In addition, for the MOO problem, it is essential to jointly optimize the multiple types of resources cooperatively representing RA, UA, and PC, which is always a complicated mixed integer programming problem that is NP in general. In addition, BS also needs to concurrently maximize the data rate and minimize the power consumption while keeping the average queue length to be stable, which requires BS to maintain a good balance among SE, EE, and queue length.

### III. STOCHASTIC OPTIMIZATION BASED ON LYAPUNOV DPP TECHNIQUE

To resolve the MOO in (23), we adopt the Lyapunov DPP technique to design an online algorithm with admission control to resolve the complex joint RA, UA, and PC problem. Specifically, it involves the following key components.

#### A. VIRTUAL QUEUES

In (23), C2 represents the stability constraint to ensure the arrivals to be eventually served by the system. To address this constraint, we define virtual queues  $\mathbf{H}(t) = \{H_1(t), H_2(t), \dots, H_{|\mathcal{U}|}(t)\}$  [29], and after the initial state  $H_u(0) = 0$ , conduct the queue to be updated by

$$H_u(t + 1) = \max\{H_u(t) - R_u(t), 0\} + \mathcal{R}_u^{req} \quad (24)$$

In addition, we transform the average power constraint C3 to a queue stability problem by introducing virtual queues  $\mathbf{Z}(t) = \{Z_1(t), Z_2(t), \dots, Z_s(t)\}$ , and after the virtual  $Z_s$  initialized with 0, update it by

$$Z_s(t + 1) = \max\{Z_s(t) - \mathcal{P}_s^{max}, 0\} + P_s(t) \quad (25)$$

By the transformation, if the virtual power queues  $Z_s(t)$ ,  $\forall s \in \mathcal{S}$  are all stable, the power constraint C3 will be satisfied [29].

#### B. LYAPUNOV DPP AND PROBLEM TRANSFORMATION

Next, according to the Lyapunov DPP, we can define  $\Theta(t) \triangleq \{Q(t), H(t), Z(t)\}$  as a concatenated vector of all  $Q_u(t)$ ,  $H_u(t)$  and  $Z_s(t)$  queues, and then introduce a quadratic Lyapunov function to represent a scalar metric of queue congestion as follows:

$$L(\Theta(t)) \triangleq \frac{1}{2} \sum_{u \in \mathcal{U}} Q_u(t)^2 + \frac{1}{2} \sum_{u \in \mathcal{U}} H_u(t)^2 + \frac{1}{2} \sum_{s \in \mathcal{S}} Z_s(t)^2 \quad (26)$$

It can be seen that a small value of the Lyapunov function would represent small lengths of the data and virtual queues. Thus, by pushing the Lyapunov function towards a lower congestion state, the queue stability can be ensured. To be clearer, a one-slot conditional Lyapunov drift is defined by

$$\Delta(\Theta(t)) \triangleq \mathbb{E}[L(\Theta(t + 1)) - L(\Theta(t)) | \Theta(t)] \quad (27)$$

Clearly, the drift function denotes the expected change of the Lyapunov function between two contiguous slots conditioned on the current state  $\Theta(t)$ . By using the drift  $\Delta(\Theta(t))$ , we can force the Lyapunov function in a lower congestion state, and make queues keep stable to control the queueing delay [29]. Given that, the queue stability constraint C1, the average throughput constraint C2, and the average power constraint C3 can be transformed to minimize the drift. Specifically, to accommodate the different units used in the metrics, we let  $\tilde{R}_{tot}(t) = R_{tot}(t)/R_{max}$  and  $\tilde{P}_{tot}(t) = \frac{\sum_{u,c,b,\ell,s,p} x_{u,c,b,\ell,s,p} (P_{s,u}^{p,t} + P_{s,u}^c)}{P_{max}}$ . Similarly, we divide  $R_u(t)$ ,  $\mathcal{R}_u^{req}$ ,  $r_u$ ,  $\lambda_u$ ,  $A_u(t)$ ,  $A_u^{max}$ , and  $\bar{Q}$  by  $R_{max}$  as  $\tilde{R}_u(t)$ ,  $\tilde{\mathcal{R}}_u^{req}$ ,  $\tilde{r}_u$ ,  $\tilde{\lambda}_u$ ,  $\tilde{A}_u(t)$ ,  $\tilde{A}_u^{max}$  and  $\tilde{Q}$ , respectively, and divide  $P_s(t)$  and  $\mathcal{P}_s^{max}$  by  $P_{max}$  as  $\tilde{P}_s(t)$  and  $\tilde{\mathcal{P}}_s^{max}$  as well, for the normalization. Providing the above, the MOO problem is transformed via the weighted sum method to

$$\begin{aligned} & \text{Minimize } \hat{V} \Delta(\Theta(t)) - W \mathbb{E}\{\tilde{R}_{tot}(t) | \Theta(t)\} \\ & \quad + (1 - W) \mathbb{E}\{\tilde{P}_{tot}(t) | \Theta(t)\} \\ & \text{subject to C4: } 0 \leq \tilde{r}_u \leq \tilde{\lambda}_u, \quad \forall u \\ & \quad \text{C5: } 0 \leq \tilde{R}_u(t) \leq \tilde{A}_u(t) \leq \tilde{A}_u^{max}, \quad \forall u, \forall t \\ & \quad \text{C6: (8) - (17),} \quad \forall t \end{aligned} \quad (28)$$

where  $\hat{V}$  denotes the weight on the queue length, and  $W$  the weight on the system performances including the spectrum efficiency and the power consumption which cooperatively exhibit the energy efficiency. To further accommodate the quantitative metric difference between the queue length and the system performances, we let  $\hat{V}/\omega = V$  with  $\omega$  to absorb the difference. Here, the weighted sum method is adopted as it is extensively used for MOO problems to provide not only multiple solution points by varying the weights consistently, but also a single solution point reflecting the preferences incorporated in the selection of a single set of weights [31].

However, due to the drift term involved, directly solving (28) is still challenging even given the weighted sum method. For this, our DPP-based dynamic control algorithm is conducted to make decisions on allocating UEs/RBs/CCs/MCSs/cells/PLs to minimize an upper bound of the following drift-plus-penalty at each time slot  $t$ . Specifically, such an upper bound on the the Lyapunov drift  $\Delta(\tilde{\Theta}(t))$  resulted from the normalized metrics  $\tilde{Q}_u(t)$ ,  $\tilde{H}_u(t)$ , and  $\tilde{Z}_s(t)$  can be shown by the following theorem.

*Theorem 1:* For all  $t$  and  $\Theta(t)$ , the drift-plus-penalty with any joint RA, UA, and PC strategy satisfies the inequality:

$$\begin{aligned} \Delta(\tilde{\Theta}(t)) & \leq \Gamma + \sum_u \tilde{Q}_u(t) \mathbb{E}\{\tilde{R}_u(t) | \tilde{\Theta}(t)\} \\ & \quad - \sum_u \tilde{Q}_u(t) \mathbb{E}\{\tilde{\mu}_u(t) | \tilde{\Theta}(t)\} \\ & \quad - \sum_u \tilde{H}_u(t) \mathbb{E}\{\tilde{R}_u(t) | \tilde{\Theta}(t)\} \\ & \quad + \sum_s \tilde{Z}_s(t) \mathbb{E}\{\tilde{P}_s(t) | \tilde{\Theta}(t)\} \\ & \quad + \sum_u \tilde{\mathcal{R}}_u^{req} \tilde{H}_u(t) - \sum_u \tilde{\mathcal{P}}_u^{req} \tilde{Z}_u(t) \end{aligned} \quad (29)$$

where  $\Gamma = \sum_u (\tilde{A}_u^{max})^2 + \sum_s (\tilde{P}_s^{max})^2 + \frac{1}{2} \sum_u (\tilde{\mu}_u^{max})^2 + \frac{1}{2} \sum_u (\tilde{R}_u^{req})^2$ , and in this expression,  $\tilde{\mu}_u^{max}$  is the maximum of  $\tilde{\mu}_u$  that can be obtained on  $u$ ,  $\tilde{A}_u^{max}$  is  $A_u^{max}/R_{max}$ , and  $\tilde{R}_u^{req}$  is the maximum request allowed for  $u$ .

*Proof.* See Appendix A.

Given that, we can multiply both sides of (29) by  $V$  and add  $-W\mathbb{E}\{\tilde{R}_{tot}(t)|\Theta(t)\} + (1 - W)\mathbb{E}\{\tilde{P}_{tot}(t)|\Theta(t)\}$  at both sides to obtain an upper bound of the objective in (28) as

$$\begin{aligned} & V\Delta(\tilde{\Theta}(t)) - W\mathbb{E}\{\tilde{R}_{tot}(t)|\Theta(t)\} + (1 - W)\mathbb{E}\{\tilde{P}_{tot}(t)|\Theta(t)\} \\ & \leq V\Gamma + V \sum_u \tilde{Q}_u(t)\mathbb{E}\{\tilde{R}_u(t)|\tilde{\Theta}(t)\} \\ & \quad - V \sum_u \tilde{Q}_u(t)\mathbb{E}\{\tilde{\mu}_u(t)|\tilde{\Theta}(t)\} \\ & \quad - V \sum_u \tilde{H}_u(t)\mathbb{E}\{\tilde{R}_u(t)|\tilde{\Theta}(t)\} \\ & \quad + V \sum_s \tilde{Z}_s(t)\mathbb{E}\{\tilde{P}_s(t)|\tilde{\Theta}(t)\} \\ & \quad + V \sum_u \tilde{\mathcal{R}}_u^{req}\tilde{H}_u(t) - V \sum_s \tilde{\mathcal{P}}_s^{req}\tilde{Z}_s(t) \\ & \quad - W\mathbb{E}\{\tilde{R}_{tot}(t)|\Theta(t)\} + (1 - W)\mathbb{E}\{\tilde{P}_{tot}(t)|\Theta(t)\} \quad (30) \end{aligned}$$

wherein  $\tilde{Q}_u(t)$  denotes the data queue in (6) obtained with the normalized  $\tilde{R}_u(t)$  and  $\tilde{\mu}_u(t)$ ,  $\tilde{H}_u(t)$  obtained with  $\tilde{R}_u(t)$  and  $\tilde{\mathcal{R}}_u^{req}$ , and  $\tilde{Z}_s(t)$  obtained with  $\tilde{P}_s(t)$  and  $\tilde{\mathcal{P}}_s^{max}$ . For a more concise representation, the constants ( $V$ ,  $\Gamma$ ,  $\tilde{\mathcal{R}}_u^{req}$ ,  $\tilde{\mathcal{P}}_s^{req}$ , and  $P_{s,u}^c$ ) and the involved terms can be neglected. In addition,  $\tilde{R}_{tot}$  and  $\tilde{P}_{tot}$  can be shown in terms of  $\tilde{\mu}_u(t)$  and  $\tilde{P}_s(t)$ , respectively, while ignoring any constant to be involved. Further, as  $\tilde{R}_u(t) \leq \tilde{A}_u(t)$  with the assumption of  $\tilde{A}_u(t) \leq \tilde{A}_u^{max}(t)$  taken at every slot  $t$  in C5 implies  $\tilde{r}_u \triangleq \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\tilde{R}_u(\tau)\} \leq \tilde{\lambda}_u \triangleq \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\tilde{A}_u(\tau)\}$  in C4, the latter (C4) as well as the other expectation operations could be removed when considering the optimization at each slot  $t$  by employing the concept of opportunistically minimizing the expectation. Finally, by optimizing the right hand side of (30), we can transform problem (28) to

$$\begin{aligned} & \text{Maximize} \sum_u V(\tilde{Q}_u(t) - \tilde{H}_u(t))\tilde{R}_u(t) \\ & \quad + \sum_u (V\tilde{Q}_u(t) + W)\tilde{\mu}_u(t) \\ & \quad - \sum_s (V\tilde{Z}_s(t) + (1 - W))\tilde{P}_s(t) \\ & \text{subject to } 0 \leq \tilde{R}_u(t) \leq \tilde{A}_u(t), \quad \forall u, \forall t \\ & \quad \text{scheduling constraints (8) - (17),} \quad \forall t \quad (31) \end{aligned}$$

### C. PROBLEM DECOMPOSITION

In the sequel, we aim to find a solution to the MOO problem by decoupling the programming model (31) to an admission control sub-problem and a transmission control sub-problem, which can be solved independently and simultaneously.

#### 1) TRAFFIC ADMISSION CONTROL

From the objective of (31), we can see that the first term,  $\sum_u V(\tilde{Q}_u(t) - \tilde{H}_u(t))\tilde{R}_u(t)$ , can be used to admit the traffic

out of  $\tilde{A}_u(t)$  arrivals to transmit subject to the constraints involving  $\tilde{R}_u(t)$ . Specifically, by decoupling this term from the joint problem, a traffic admission control sub-problem can be formulated as

$$\begin{aligned} & \text{Minimize}_{R_u(t)} \sum_{u \in \mathcal{U}} (V\tilde{Q}_u(t) - V\tilde{H}_u(t))\tilde{R}_u(t) \\ & \text{subject to } 0 \leq \tilde{R}_u(t) \leq \tilde{A}_u(t), \quad \forall u, \forall t \quad (32) \end{aligned}$$

For the linear problem obtained, we have a simple threshold-based admission control strategy as

$$\tilde{R}_u(t) = \begin{cases} \tilde{A}_u(t), & \tilde{H}_u(t) > \tilde{Q}_u(t) \\ 0, & \text{otherwise} \end{cases} \quad (33)$$

This strategy clearly shows that the newly arrivals can be admitted to transmit only when the the virtual queue length  $\tilde{H}_u(t)$  is larger than the actual data queue length  $\tilde{Q}_u(t)$ . Otherwise, the arrivals will not be admitted so that the data queue can be stable. In fact, the threshold-based admission control would reduce the value of  $\tilde{H}_u(t)$  to push  $\tilde{r}_u(t)$  towards  $\tilde{R}_u(t)$ , thus stabilizing all the data queues involved.

#### 2) MULTI-RESOURCE ALLOCATIONS FOR TRANSMISSION CONTROL

As the MRA problem for the transmission control is a core of the framework, our main challenge is to concurrently determine UEs, CCs, RBs, MCSs, cells, and PLs at each time slot  $t$  to make an optimal tradeoff decision among SE, EE, and queue length. To this end, the MRA problem is formulated by decoupling the joint optimization problem to concurrently consider the terms with respect to the transmission data rate  $\tilde{\mu}_u(t)$  and the energy consumption  $\tilde{P}_s(t)$  in the right hand side of (30) subject to the scheduling constraints. Specifically, with their signs reversed for changing minimization to maximization, the transmission control sub-problem can be represented by

$$\begin{aligned} & \text{Maximize}_{x_{u,c,b,\ell,s,p}(t)} \sum_u (V\tilde{Q}_u(t) + W)\tilde{\mu}_u(t) \\ & \quad - \sum_s (V\tilde{Z}_s(t) + (1 - W))\tilde{P}_s(t) \\ & \text{subject to scheduling constraints (8) - (17),} \quad \forall t \quad (34) \end{aligned}$$

As shown in Sec. II-D, the variables  $x_{u,c,b,\ell,s,p}$  in the scheduling constraints are binary, and thus the transmission control sub-problem involving only these variables is a binary integer programming (BIP) problem that is NP hard in general. Moreover, as  $x$  is not only binary but also high-dimensional, finding an optimal solution to this problem would be time-consuming even given optimization tools. To address this challenge, in addition to solving the BIP with an IP solver, we develop also a distributed or semi-distributed algorithm wherein the network nodes, or SBSs, can perform the allocation independently or by the minimal assistance of the central controller, or MBS, to be a more efficient solution for practical implementations because of reduced computational complexity.



**IV. DISTRIBUTED TWO-LEVEL MULTI-RESOURCE ALLOCATION ALGORITHM**

To realize a distributed or semi-distributed scheduling algorithm to resolve the MRA problem in (34) for the downlink transmissions, we first reduce the programming model to a single-cell problem with only one power level. Inspired by the reduced model, we then propose a distributed two-level MRA algorithm for more computationally efficient solutions.

**A. REDUCED MODEL**

As the first step for the reduced model, the binary variable is reduced to  $x_{u,c,b,\ell}$  that involves no cells  $s$  and PLs  $p$ . Given that, the constraints to avoid allocating RBs to neighboring cells are no longer necessary, and can be reduced to

$$\sum_u \sum_\ell x_{u,c,b,\ell} \leq 1, \quad \forall c \in \mathcal{C}, b \in \mathcal{B} \quad (35)$$

In addition, the CC cardinality constraint on each cell is also unnecessary and can be ignored while the constraint on the number of CC that each UE can have could be rewritten as

$$\sum_c \sum_\ell y_{u,c,\ell} \leq d_u, \quad \forall u \in \mathcal{U} \quad (36)$$

where  $y_{u,c,\ell}$  is a new auxiliary variable adopted here, and its value 1 represents that CC  $c$  is assigned to UE  $u$  with TM index  $\ell$ , and 0 otherwise. Further, because this scenario considers only one power level in a single cell, constraints (8) and (9) can be simplified as

$$\begin{cases} x_{u,c,\ell,b} \leq y_{u,c,\ell}, & \forall c \in \mathcal{C}, b \in \mathcal{B}, u \in \mathcal{U}, \ell \in \mathcal{L} \\ \sum_\ell y_{u,c,\ell} \leq 1, & \forall c \in \mathcal{C}, u \in \mathcal{U} \end{cases} \quad (37)$$

Now, with a linear object function subject to the constraints (35), (36), and (37), we can reduce our model to a single-cell problem without multiple discrete power levels as that in [9] using certain equivalent *max* operations in their constraints. There is no doubt that an IP problem like the above is NP-hard in general. However, for the reduced problem, the coefficient matrix of constraints could be a totally unimodular (TUM) matrix whose determinate for every square submatrix equals -1, 0, or 1. To show this possibility, we first let  $\mathbf{u} \triangleq |\mathcal{U}|$ ,  $\mathbf{c} \triangleq |\mathcal{C}|$ ,  $\ell \triangleq |\mathcal{L}|$ , and  $\mathbf{b} \triangleq |\mathcal{B}|$  to more concisely represent these quantities, and then order the variables  $x$  and  $y$  to construct the following vectors

$$\mathbf{x} = [\bar{x}_1, \dots, \bar{x}_m, \dots, \bar{x}_{\mathbf{u}\mathbf{c}\mathbf{b}}]^T \quad (38)$$

$$\mathbf{y} = [\bar{y}_1, \dots, \bar{y}_n, \dots, \bar{y}_{\mathbf{u}\mathbf{c}\ell}]^T \quad (39)$$

where  $\bar{x}_m = x_{u,c,\ell,b}$  and  $\bar{y}_n = y_{u,c,\ell}$  with their indices calculated by

$$\begin{aligned} m &= (u - 1)\mathbf{c}\mathbf{b} + (c - 1)\ell\mathbf{b} + (\ell - 1)\mathbf{b} + b \\ n &= (u - 1)\mathbf{c}\ell + (c - 1)\ell + \ell \end{aligned}$$

These variable vectors are then concatenated and transposed to a new vector, say  $\mathbf{v}$ , as follows:

$$\mathbf{v} = [\mathbf{x}^T \ \mathbf{y}^T]^T \quad (40)$$

whose length is  $\mathbf{u}\mathbf{c}\mathbf{b} + \mathbf{u}\mathbf{c}\ell$ . Given that, the constraints (35), (36), and (37) can be more concisely represented. Specifically, let  $\text{blkd}(K, k)$  be a block diagonal matrix wherein a block  $K$  is repeated  $k$  times in the diagonal line and the other elements are all 0. Given that, the identify matrix  $\mathbf{1}_b$  is denoted by  $\text{blkd}(1, \mathbf{b})$ , which leads to  $\beta_1 = [\mathbf{1}_b \cdots \mathbf{1}_b]_{\mathbf{b} \times \mathbf{b}\ell}$ . Then, (35) can be reformulated by

$$A_1 \mathbf{x} \leq \mathbf{1} \quad (41)$$

where  $A_1 = [\text{blkd}(\beta_1, \mathbf{c}) \cdots \text{blkd}(\beta_1, \mathbf{c})]_{\mathbf{c}\mathbf{b} \times \mathbf{c}\mathbf{b}\mathbf{u}\ell}$ . Further, let  $\mathbf{d} = [d_1 \cdots d_u]$  and  $A_2 = \text{blkd}(\mathbf{1}_{\mathbf{c}\ell}^T, \mathbf{u})$ . Then, (36) can be rewritten as

$$A_2 \mathbf{y} \leq \mathbf{d} \quad (42)$$

Similarly, by defining  $\beta_3 = -\text{blkd}(\mathbf{1}_b, \mathbf{u}\mathbf{c}\ell)$  and then  $A_3 = [\mathbf{1}_{\mathbf{u}\mathbf{c}\mathbf{b}}, \beta_3]_{\mathbf{u}\mathbf{c}\mathbf{b} \times (\mathbf{u}\mathbf{c}\mathbf{b} + \mathbf{u}\mathbf{c}\ell)}$  in addition to  $A_4 = \text{blkd}(\mathbf{1}_\ell^T, \mathbf{u}\mathbf{c})$ , we can transform (37) to

$$\begin{cases} A_3 \mathbf{v} \leq \mathbf{0} \\ A_4 \mathbf{y} \leq \mathbf{1} \end{cases} \quad (43)$$

Finally, by integrating all the constraints into a canonical form, we have

$$A \mathbf{v} \leq \mathbf{c} \quad (44)$$

where

$$A = \begin{bmatrix} A_1^T & 0 & \mathbf{1}_{\mathbf{u}\mathbf{c}\mathbf{b}} & 0 \\ 0 & A_2^T & \beta_3^T & A_4^T \end{bmatrix}^T \quad (45)$$

and

$$\mathbf{c} = [\mathbf{1}_{\mathbf{c}\mathbf{b}}^T \ \mathbf{d}^T \ \mathbf{0}_{\mathbf{u}\mathbf{c}\mathbf{b}}^T \ \mathbf{1}_{\mathbf{u}\mathbf{c}}^T]^T \quad (46)$$

As noted before, the coefficient matrix of constraints,  $A$ , could be totally unimodular (TUM). As a simple example, given  $\mathbf{u} = 2$ ,  $\mathbf{c} = 1$ ,  $\ell = 2$  and  $\mathbf{b} = 1$ , the coefficient matrix  $A$  will be

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad (47)$$

Thanks to the small size, its TUM can be verified by exhaustively checking every square submatrix of  $A$  to have determinant equal to  $\pm 1$  or 0. As Schrijver revealed [32], if  $A$  of an integer linear problem (ILP) is TUM, the ILP can be relaxed to a linear programming problem (LP) by removing the integrality constraints. Then, the LP relaxation of an ILP could be solved through any standard LP technique. This is further verified in our numerical experiments for the two-level approach wherein the optimal results for the single-cell problem could be obtained by using a LP solver in

usual cases. However, the TUM property is not guaranteed. For example, given  $\mathbf{u} = 2$ ,  $\mathbf{c} = 1$ ,  $\ell = 2$  and  $\mathbf{b} = 2$ , the coefficient matrix  $A$  becomes

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad (48)$$

In the above, the determinant of its square submatrix could be 2 in addition to  $\pm 1$ , or 0. Although it is not always TUM, the non-TUM integer programming problems only contribute a small part of the experiments in Sec. VI, and have a coefficient matrix like (48) being 2-regular<sup>1</sup> which could be still solved efficiently as that shown in [34].

### B. DISTRIBUTED TWO-LEVEL MRA ALGORITHM

Inspired by the reduced problem or model, we next propose a distributed or semi-distributed two-level MRA algorithm to resolve the MRA problem for more computationally efficient solutions. Specifically, in view of the start-of-the-art RA algorithms such as those implemented in [9] adopting a two-step assignment in a single cell, we would first allocate CCs in an optimal sense, and then allocate the other resources in the multi-cell scenario.

#### 1) THE FIRST LEVEL RA (ON CC)

For the CC allocation, we sort  $\mathbf{c}$  CCs in decreasing order on the data rate perceived by cell  $s$  according to the 3GPP table-based representation shown in Sec. II-B, for each  $s \in \mathcal{S}$ , resulting in a sorted list  $\mathbf{C}_s = \{C_{s_1}, \dots, C_{s_c}\}$ . Then, we allocate these CCs by solving the following CC allocation optimization problem:

$$\begin{aligned} & \text{Maximize} \quad \sum_{\forall s \in \mathcal{S}} \sum_{i=1}^c W_i C_{s_i} x_{s_i}^s \\ & \text{subject to D1:} \quad \sum_{c \in \mathcal{C}} x_c^s \geq 1 \quad \forall s \\ & \quad \quad \quad \text{D2:} \quad \sum_{c \in \mathcal{C}} x_c^s \leq f_s \quad \forall s \\ & \quad \quad \quad \text{D3:} \quad \sum_{s' \in N_s \cup s} x_c^{s'} \leq 1 \quad \forall s, \forall c \end{aligned} \quad (49)$$

wherein  $W_i = \frac{1}{i}$  represents the weight of CC  $C_{s_i}$  that is the  $i$ th element in the list  $\mathbf{C}_s$  in decreasing order. This weight

<sup>1</sup>As shown by Lemma 2.7 in [33], if for each non-singular square submatrix  $R$  of a matrix  $A$ ,  $\det(R) \in \{\pm 1, \pm k\}$ , then  $A$  is  $k$ -regular. In this example,  $k$  is 2, and according to Theorem 3.8 in [33], 2-regular matrices are the rational matrices that ensure half-integral polyhedra.

implies that the higher SNR or data rate the cell  $s$  perceived, the higher preference in the objective function. In addition,  $x_c^s$  is a binary variable deciding whether cell  $s$  is allocated with CC  $c$  or not, and  $x_{s_i}^s$  denotes the variable corresponding to CC  $C_{s_i}$ . Given that, D1 denotes the constraint to enforce each cell to be allocated at least one CC for its UEs. D2 ensures that each cell can have at most  $f_s$  CCs. D3 represents a constraint that as long as a CC is allocated to cell  $s$ , it can not be assigned to its neighboring cells  $s' \in N_s$  to prevent excess interference.

Although the above CC allocation seems to be less complex and may be solved more easily when compared with the joint optimization problem in question, it is still an IP problem that is NP in general if no special structures are imposed. To resolve this problem, we conduct a greedy algorithm performed by MBS to reduce the complexity of solving the RA problem on CC while giving suboptimal solutions to be satisfactory enough. To this end, the data rates of cell  $s$  on different CCs  $c$  based on the CQIs reported by UEs are recorded in a table  $\mathbf{L}$  of MBS, in addition to a table  $\mathbf{C}$  initialized to be empty to record the allocation results. As shown in Algorithm 1, for each CC  $c \in \mathcal{C}$ , it finds the cell  $s^*$  that has the highest data rate obtainable, and if this rate is non-negative and the number of CC allocated to this cell does not exceed its limit  $f_{s^*}$ , then it permits the allocation and records this with  $\mathbf{C}(c, s^*) = 1$ . While allocating, it forbids the neighboring cells  $\hat{s} \in N(s^*)$  to allocate the same CC by setting  $\mathbf{L}(c, \hat{s}) = -1$  so that these neighboring cells have no  $v^*$  greater than 0 to enter the procedure starting at line 6.

---

#### Algorithm 1 The Greedy CC Allocation Algorithm

---

- 1: (Given)  $\mathbf{L}(c, s)$ ,  $\forall c \in \mathcal{C}$ , and  $\forall s \in \mathcal{S}$ ;
  - 2: (Initialization)  $\mathbf{C}(c, s) = 0$ ,  $\forall c \in \mathcal{C}$ , and  $\forall s \in \mathcal{S}$ ;
  - 3: **for**  $c \in \mathcal{C}$  **do**
  - 4:    $v^* = \max_s \mathbf{L}(c, s)$  and  $s^* = \arg \max_s \mathbf{L}(c, s)$
  - 5:   **while**  $v^* > 0$  **do**
  - 6:     **if**  $\sum_c \mathbf{C}(c, s^*) < f_{s^*}$  **then**
  - 7:        $\mathbf{C}(c, s^*) = 1$
  - 8:        $\mathbf{L}(c, \hat{s}) = -1$ ,  $\forall \hat{s} \in N(s^*)$
  - 9:     **else**
  - 10:        $\mathbf{L}(c, s^*) = -1$
  - 11:     **end if**
  - 12:   **end while**
  - 13: **end for**
- 

#### 2) THE SECOND LEVEL RA (ON THE OTHER RESOURCES)

As shown in the literature such as [9], given a number of CCs, the research works usually focus on the RA problem to allocate the radio resources (RBs and CCs) to UEs in a single cell. However, without the viewpoint of multiple cells, they inevitably ignore the problem that if multiple neighboring cells are allocated with the same CCs, they could use the same RBs of these CCs to cause inter-cell interference. In this work, the first level RA is already conducted to avoid the inter-cell interferences by allocating different CCs to the neighboring cells. Then, supposing that UE associating with the nearest

BS is given in advance at a larger time scale, our programming model could be degraded for the second level RA to consider only one cell but still take into account multiple PLs and the other resources. This is different from the reduced model in Sec. IV-A or that shown in [9] addressing no power control. Specifically, without the notion of  $s$ , the scheduling constraints (8) and (9) could be reformulated for the second level RA as

$$\sum_{\forall p \in \mathcal{P}} x'_{u,c,b,\ell,p} \leq y'_{u,c,\ell}, \quad \forall u \in \mathcal{U}, \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall \ell \in \mathcal{L} \quad (50)$$

$$\sum_{\forall \ell \in \mathcal{L}} y'_{u,c,\ell} \leq 1, \quad \forall u \in \mathcal{U}, \forall c \in \mathcal{C} \quad (51)$$

where  $x'_{u,c,b,\ell,p}$  corresponds to the binary variable  $x_{u,c,b,\ell,s,p}$ , and  $y'_{u,c,\ell}$  to  $y_{u,c,\ell,s}$ . In addition, similar to that for the reduced model in Sec. IV-A, the constraints specific to the multi-cell environment should be also deleted or modified to fit the single-cell scenario, as summarized as follows:

- First, the monopoly constraint on UE to be served by a single cell, represented by (10) and (11), are no longer needed and thus deleted here.
- Second, the monopoly constraint to enforce that a specific RB is either allocated in a cell  $s$  or its neighboring cells  $s' \in N_s$  would be modified to consider only the allocation of an RB of a CC for a given cell. Thus, the constraints (12) and (13) can be reduced to

$$\sum_{\forall u \in \mathcal{U}, \forall \ell \in \mathcal{L}, \forall p \in \mathcal{P}} x'_{u,c,b,\ell,p} \leq 1, \quad \forall c \in \mathcal{C}, \forall b \in \mathcal{B} \quad (52)$$

- Third, the cardinality constraint  $f_s$  on  $s$ , i.e., the number of CC allocated to cell  $s$  not to exceed the limit  $f_s$ , has been done by the first level allocation on CC, and can be eliminated. That is, (14) and (15) can be removed here.
- Fourth, the cardinality constraint that the number of CC allocated to UE  $u$  can not exceed  $d_u$  is still valid despite the number of cells. However, in the case of given  $s$ , the notion on cells should be eliminated, and thus (16) and (17) would be changed to

$$\sum_{\forall \ell \in \mathcal{L}, \forall p \in \mathcal{P}} x'_{u,c,b,\ell,p} \leq y'_{u,c}, \quad \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall u \in \mathcal{U} \quad (53)$$

$$\sum_{\forall c \in \mathcal{C}} y'_{u,c} \leq d_u, \quad \forall u \in \mathcal{U} \quad (54)$$

Here, we would rather use  $y'_{u,c}$  to correspond to  $y_{u,c}^5$  given in Sec. II-D to preserve a similar representation than adopt a new auxiliary variable like  $y_{u,c,l}$  in Sec. IV-A which is tailored for the notational simplicity in the reduced model considering no power control.

### 3) HARDNESS RESULT

As shown readily in the above, providing that CCs are given by MBS, the binary variables  $x'$  in the second level RA as well as  $\tilde{r}_u$  and  $\tilde{R}_u$  shown in Sec. III-C can be independently decided by SBSs for their UEs at each time slot  $t$ , leading

to the Lyapunov DPP-based dynamic control in Sec. III-B realized in a distributive manner. In addition, as indicated in Sec. IV-A, the coefficient matrix involved could be TUM or possibly 2-regular, leading to the corresponding RA problem solved efficiently as shown in our numerical experiments.

The above is worth noting because an MRA problem is NP-hard in general if no special structures are imposed. In our case, even the reduced model considered in Sec. IV-A is NP-hard, similar to those already proved in the literature. For example, by mapped to the well-known 3-SAT problem, a radio resource scheduling problem in [16] was proved to be NP-hard subject to the constraints: (i) each RB can be assigned up to one UE, and (ii) only one MIMO mode can be selected for all assigned RBs for a UE. Here, when considered with only one CC and two MCSs, the reduced model can be regarded as a special case of the scheduling problem, in which the MCS selection is equivalent to the MIMO mode selection in the scheduling problem, and hence it could be reduced to a NP-hard problem through the same way of 3-SAT mapping. In contrast to the NP-hardness, if the coefficient matrix in this model is TUM, the LP resulted would be of order  $(\mathbf{u} + \mathbf{cb}) \times (\mathbf{uclb} + \mathbf{ucl})$  maximizing the objective over all  $\mathbf{v}$  in  $\mathbb{R}^{(\mathbf{uclb} + \mathbf{ucl})}$  such that  $\mathbf{A}\mathbf{v} \leq \mathbf{c}$ . According to [35], the worst-case complexity of solving the LP problem would be  $O(\sqrt{\mathbf{u} + \mathbf{cb} + \mathbf{uclb} + \mathbf{ucl}} \ln \frac{1}{\epsilon})$ , where  $\epsilon$  determines the accuracy of the solutions obtained with the barrier method in [35]. When  $\mathbf{u}$  approaches infinity, it could be simply represented by  $O(\mathbf{u}^{1/2})$ , showing a significant improvement on the time complexity for the NP-hard problem. On the other hand, the state-of-the-art greedy algorithm in [9] that is based on submodular set functions has the time complexity  $O(\mathbf{u}^2 d_u \mathbf{cb}\ell)$ , or simply  $O(\mathbf{u}^2)$  if  $\mathbf{u}$  approaches infinity, which is clearly higher than the former.

Finally, the overall optimization algorithm for solving the stochastic MOO problem (28) is given in **Algorithm 2** as a summary for easy reference.

## V. PERFORMANCE BOUNDS

Thanks to the Lyapunov DPP approach, the data and virtual queues involved will be mean rate stable, which can be proved as that shown in [29]. In addition to the stability on queues, the weighted sum function defined by

$$\begin{aligned} f(x(t)) &= W \frac{R_{tot}(t)}{R_{max}} - (1 - W) \frac{P_{tot}(t)}{P_{max}} \\ &= W \tilde{R}_{tot}(t) - (1 - W) \tilde{P}_{tot}(t) \end{aligned} \quad (55)$$

is also considered that can capture the performance of EE and SE for a specific range of  $W$  [36]. Specifically, with  $x(t)$  to denote  $x_{u,c,b,\ell,s,p}(t)$  for the simplicity on its representation, we have the following theorem regarding this metric:

*Theorem 2:* If problem (28) is feasible, problem (31) is optimally solved, and  $\mathbb{E}\{L(\tilde{\Theta}(0))\} < \infty$ , then the weighted sum function  $f(x(t))$  has the performance bounds as

$$f_{opt} - \Gamma V < \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{\tau=0}^{T-1} \mathbb{E}\{f(x(\tau))\} < f_{opt} \quad (56)$$

**Algorithm 2** The Optimization Algorithm for Solving the Stochastic MOO Problem (28)

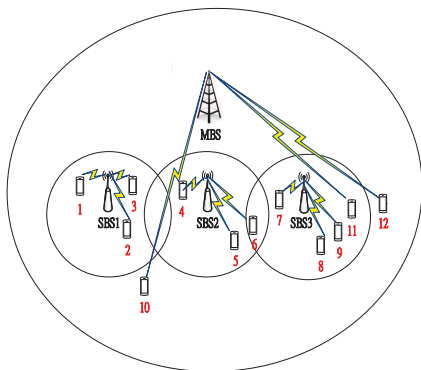
- 1: (Given) weighting parameters  $V, W, T, \tilde{R}_u^{req}, \forall u, \tilde{P}_s^{req}, \forall s$ ;
- 2: (Initialization)  $t = 0, \tilde{Q}_u(t) = 0, \tilde{H}_u(t) = 0, \forall u$ , and  $\tilde{Z}_s(t) = 0, \forall s$ ;
- 3: **while**  $t < T$  **do**
- 4:   Observe  $\tilde{Q}_u(t), \tilde{H}_u(t), \tilde{Z}_s(t), \tilde{R}_u(t), \tilde{\mu}_u(t)$ , and  $\tilde{P}_s(t)$ ;
- 5:   (Traffic admission Control) Determine admitted traffic  $\tilde{R}_u(t), \forall u$  with (33);
- 6:   (MRA for transmission control) Determine  $x_{u,c,b,\ell,s,p}$  by solving (34) with either an IP solver or the distributed two-level MRA algorithm proposed;
- 7:    $t = t + 1$ ;
- 8:   (Queue Updates) Update the data queues  $\tilde{Q}_u(t), \forall u$ , through (6) and virtual queues  $\tilde{H}_u(t), \forall u$ , through (24), and  $\tilde{Z}_s(t), \forall s$ , through (25), with the normalized metrics involved;
- 9: **end while**

where  $f_{opt}$  denotes the maximum value of  $\overline{E}\{f(x(t))\}$  obtained by any solution satisfying C1-C6 in this problem.

*Proof.* Please refer to Appendix B.

**VI. NUMERICAL EXPERIMENTS**

To numerically evaluate our proposal, we conduct a simulation topology consisting of 1 micro base station (MBS) and 3 small base stations (SBSs). As shown in Fig. 2, each base station (MBS or SBS) initially serves 3 user equipments (UEs) that are located within its transmission range. The other parameters for the experiments are summarized in Fig. 3 for reference. As shown therein, the numbers of resources for the multiple resource allocation (MRA) problem involved would be significantly high enough for an optimization tool to obtain a solution to the high-dimensional combinatorial problem within a reasonable period of time. Given that, each UE is simulated to estimate the channel quality on each resource block (RB) of each component carrier (CC) by using reference signals transmitted from base



**FIGURE 2.** Simulation topology of the multi-tier multi-cell network.

Parameter	Setting
Number of UEs ( $u$ )	12
Number of cells ( $s$ )	4
Number of CCs ( $c$ )	5
Number of MCS index ( $\ell$ )	870 (refer to Table 1)
Number of RBs per CC ( $b$ )	10
Maximum number of CCs for BS ( $f_s$ )	2
Maximum number of CCs for UE ( $d_u$ )	2
Transmission powers of MBS ( $\mathcal{P}$ )	$\{0.3P^t, 0.5P^t, P^t\}$
Transmission powers of SBS ( $\mathcal{P}$ )	$\{0.05P^t, 0.1P^t, 0.2P^t\}$
Noise powers ( $N_{s,u}^{c,b}$ )	-110 dbm
SNR-CQI index mapping	refer to [37]
CQI-MCS index mapping	refer to [37]
MCS index mapping to modulation and TBS index tables	refer to TS 36.213 [1]

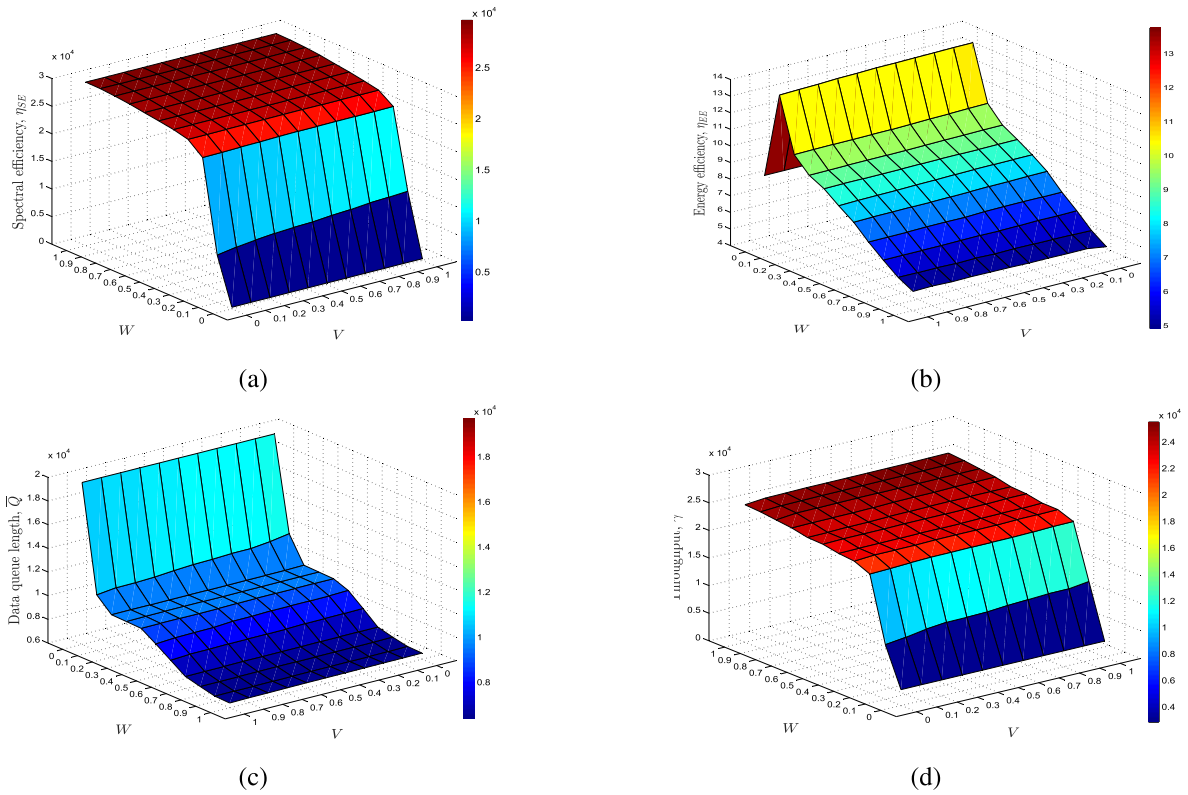
**FIGURE 3.** Parameter setting in the experiments, where  $P^t = 29$  dbm.

stations (BSs), measuring their signal to noise ratios (SNRs) and using a mapping table such as that in [37] to obtain the channel quality indicators (CQIs) to be reported to BSs. Here, SNR of each RB perceived by UE is assumed to be a random variable uniformly distributed in the range between  $-5$  and  $22.38$  according to the SNR-CQI index mapping in [37], so that the allocation results would involve all possible mapping values in the simulation. Given that, the CQIs collected, the MCS index mapping, and TBS index tables specified in [38] are resulted and used by BS to estimate the achievable data rates for UEs required by the different algorithms for comparison in the experiments.

**A. PERFORMANCE TRADEOFF ON SE, EE, AND QUEUE LENGTH**

In the first set of experiments, we focus on the critical factors to impact the optimization algorithm. To show this, the time-average spectrum efficiency  $\eta_{SE}$ , energy efficiency  $\eta_{EE}$ , data queue length  $\bar{Q}$ , and throughput  $\gamma$  are represented by the mean values of the corresponding metrics in their normal scales obtained from all UEs involved.

As shown in Figs. 4(a) and 4(b), the performance trends of spectrum efficiency (SE) and energy efficiency (EE) are shown by varying the system parameters (or weights),  $W$  and  $V$ , given  $\omega = 0.002$ . Specifically, in Fig. 4(a), it can be seen that, given the same  $V$ , the SE value would continuously increase with  $W$  due to the higher transmitted power adopted to enhance the transmission rate. Unlike the performance trend on SE, it can be seen in Fig. 4(b) that EE would rise up first and then turn to decrease as  $W$  increases. The trend on EE can be so observed because when  $W$  is high, the transmit power consumption is negligible as compared to the circuit power consumption, and in this condition SE would grow faster than the total power consumption dominated by the circuit. Thus, as shown in these sub-figures (Figs. 4(a) and 4(b)), when  $W$  is low, EE increases as SE increases. However, after



**FIGURE 4.** Impacts of varying  $V$  and  $W$  upon (a) spectral efficiency  $\eta_{SE}$ , (b) energy efficiency  $\eta_{EE}$ , (c) data queue length  $\bar{Q}$ , and (d) throughput  $\gamma$ .

the peak value of EE, the transmit power would dominate the total power consumption instead of the circuit power, and the increment of the total power consumption would be larger than that of SE afterward, leading to EE gradually decreased as shown in this sub-figure (Fig. 4(b)).

From another viewpoint, it can be also seen that, for a given  $W$ , increasing  $V$  could decrease SE or EE at a milder degree than the above. Specifically, when  $W < 0.9$  along with the other parameters in the experiments, EE is exhibited to decrease as  $V$  increases while SE is shown to increase with the growth of  $V$ . However, when  $W \geq 0.9$ , the trend is slightly reversed with some fluctuations. This trend could be observed because, when  $W$  is small, increasing  $V$  would require the wireless links to increase their data rates so as to decrease the average queue backlog, which leads to a better SE. At the same time, as the emphasis on the data rate is given to all links in the network, the total transmit power would thus increase, which eventually degrades EE. As just indicated, this trend is slightly changed when  $W$  is large. In this case, increasing  $V$  provides a stronger enforcement on reducing the queue length to a lower congestion state as indicated. However, due to a large weight  $W$  on the data rate, lowering the energy consumption would contribute more to EE when a larger data rate is resulted in this case.

Apart from the above, the results on the average queue length ( $\bar{Q}$ ) are summarized in Fig. 4(c). As exhibited therein,

although with slight fluctuations, this metric has the trend to descend with  $W$  and  $V$ . With respect to  $V$ , it has been shown in the above that as  $V$  increases, the system would emphasize more on decreasing the queue length, and therefore the queue backlog declines. With respect to  $W$ , it can be seen from the objective function that a larger  $W$  represents a stronger emphasis on SE. When applied, the transmission data rate is enhanced and thus the average queue length is decreased.

Finally, we present the performance trend on the system throughput  $\gamma$  defined as the sum of UE throughputs, i.e.,  $\gamma = \sum_{u \in \mathcal{U}} r_u$ . By comparing Fig. 4(a) with Fig. 4(d), we can see that the SE and the throughput have the same trend because the admission control is based on the average queue length, and a higher transmission data rate (SE) would reduce the average queue length, which leads to more traffic to be admitted for entering the queues of UEs (i.e., increasing the throughput). However, it can be also seen that when  $W < 0.2$ , the SE ( $\eta_{SE}$ ) is less than the throughput ( $\gamma$ ). This could be observed because in the experiments the traffic arrival rate is conducted to saturate the system; that is, its value is much higher than the throughput requirement,  $A_u(t) \gg R_u^{req}, \forall u, t$ . When  $W < 0.2$  in the experiments, the service rate could not fulfill the requirement, and queue will build up after admitting the arrivals. Given that, the transmission data rate would take a long time to resolve the queue backlog, leading to a low SE and a high queue length as shown in Fig. 4(a) and Fig. 4(c),

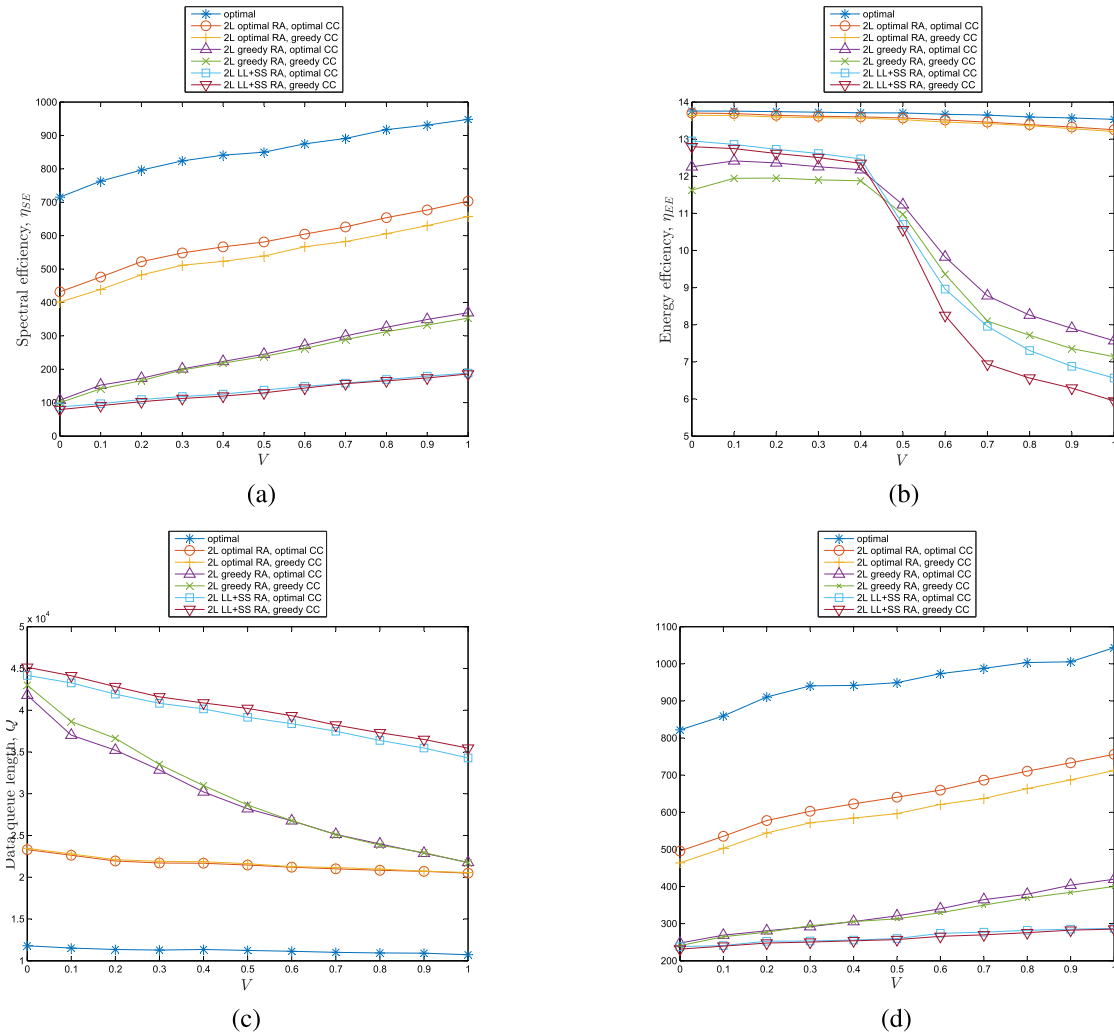


FIGURE 5. Performance comparison on (a) spectral efficiency  $\eta_{SE}$ , (b) energy efficiency  $\eta_{EE}$ , (c) data queue length  $\bar{Q}$ , and (d) throughput  $\gamma$ , by varying  $V$ .

respectively, in the  $W < 0.2$  range. In this range, the system throughput in the long term would be larger than the SE while satisfying the minimum requirement  $R_u^{req}, \forall u$ . In contrast, when  $W$  increases to be larger than 0.2, the service rate and the SE approach the system capacity that is much higher than  $R_u^{req}, \forall u$ . In this situation ( $W \geq 0.2$ ), the queue length could decrease as shown in Fig. 4(c). Therefore, to satisfy the throughput requirement  $R_u^{req}, \forall u$  that is less than the capacity, the throughput  $\gamma$  would be less than the SE according to the admission control, which can be observed when comparing Fig 4(a) with Fig. 4(d), as well.

### B. PERFORMANCE COMPARISON

In the second set of experiments, we let  $W$  be 0.1 and vary  $V$  to exemplify the performance differences between the optimization algorithm (Algorithm 2) using an IP solver, and that using the distributed two-level MRA algorithm to resolve the MRA problem. For the distributed two-level algorithm in the latter, we use the second level RA algorithm proposed in

Sec. IV-B2 to obtain the optimal solutions in the reduced programming model, or adopt the greedy algorithm as well as the LL+RS algorithm in [9] to obtain the heuristic solutions in the second level. Similarly, we use either an optimal IP solver or the greedy CC allocation algorithm shown in Algorithm 1 to resolve the first level RA problem on CC (49). Specifically, by first indicating the RA algorithm used in the second level, and then that used in the first level, we denote the variant methods based on the distributed two-level MRA algorithm by 2L “optimal/greedy/LL+SS RA” with “optimal/greedy CC”, resulting in  $3 \times 2 = 6$  different two-level (2L) method-names as shown in the legends of Fig. 5.

In Fig. 5(a), providing  $W = 0.1$ , the SE is shown to increase as  $V$  increases, complying with the results in the first set of experiments with  $W < 0.9$ . Clearly, all the two-level methods have the same trend. However, the 2L optimal RA-based methods would outperform the 2L greedy-based methods (including those using the greedy RA algorithm and the LL+SS RA algorithm in the second level) in spite of the

CC allocation obtained either optimally or greedily in the first level. In addition, it can be also seen that in the greedy-based methods, the methods using the greedy RA algorithm would outperform those using the LL+RS RA algorithm, in spite of the CC allocation as well. Nevertheless, how to allocate CC still has its own impact on the performance, exemplified by the fact that the greedy CC allocation algorithm would improve the computational complexity at the cost of slightly decreasing SE when compared with the optimal.

Similarly, from Fig. 5(b), we can see that by means of the second level RA algorithm (in Sec. IV-B2) and the greedy CC allocation algorithm (i.e., **Algorithm 1**) in the first level, the 2L optimal RA with greedy CC method can approach the optimal EE. This suggests that using the distributed two-level algorithm proposed in Sec. IV-B to replace an IP solver to resolve the MRA problem is a good way to trade the EE performance off against the time complexity. In addition, as already noted in Sec. VI-A, the trend is the same for all the methods for comparison that EE would decrease as  $V$  increases if  $W < 0.9$ . However, when compared with the 2L optimal RA-based methods, the 2L greedy-based methods have worse performances and would drop even more significantly, especially when  $V$  increases larger than 0.4.

From the viewpoint of queue length, the tradeoff would be seen more clearly. Specifically, although the 2L optimal RA-based methods can approach the joint optimization with an IP solver in terms of EE, the queue length would be the cost, as shown by its value not so close to the optimal exhibited in Fig. 5(c). In addition, among the 2L greedy-based methods, the methods with the greedy RA algorithm would be better than those with the LL+RS RA algorithm in terms of queue length. The performance difference also provides a valuable reference to choose the greedy-based methods in addition to that based on SE and EE. Further, from Fig. 5(d), we can see that the performance trend on the throughput is the same as that on SE. It is expected because the traffic allowed for transmission reflects the data rate resulted in the dynamic system through the Lyapunov-based admission control. More specifically, the trend complies with that shown in Sec. VI-A, which can be also observed here by comparing Fig. 5(a) with Fig. 5(d) to show that  $\gamma$  would be larger than SE when  $W < 0.2$  (in this case  $W = 0.1$ ), and these performance metrics would increase as  $V$  grows in this case. Finally, as shown in all the sub-figures of Fig. 5, the greedy CC allocation algorithm might decrease the performance metrics only at a slight degree when compared with the optimal CC allocation. Apart from the CC allocation, the whole two-level MRA algorithm would lead the system to achieve more than half of the optimal SE and  $\gamma$ , and approach the optimal EE while maintaining a larger queue length. These confirm our design aim to obtain an effective algorithm to reduce the complexity for the RA optimization while obtaining sub-optimal solutions to be satisfactory enough.

Next recall that, in Sec. II-B, the channel condition is assumed to remain unchanged during an allocation period which leads to a CQI for RB to be mapped to the highest-rate

MCS for a UE using the RB [26], and the channel conditions on all UEs and RBs can be perceived by the system through the CQIs reported [9]. However, if certain issues affecting the condition arise, e.g., the channel suddenly varies fast during the period, or the BSs adopt a large-scale CSI (including path-loss and shadowing) scheme for SINR to reduce the CQI overhead [40], the MCSs obtained could be overestimated to result in an unacceptable bit error rate for the transmission. For this, in addition to the full CSI assumption adopted before (represented here by  $P_e = 0$ ), we assume the over-estimation error to be happened with a probability  $P_e = 0.1, 0.3, \text{ or } 0.5$ , which diminishes the data rate of a MRA to 0 due to the unacceptable bit error rate resulted, while fixing  $W = 0.1$  and  $V = 0.5$  to exemplify the performance trend. Otherwise, the data rate is considered to be correctly represented by  $v_{u,c,b,\ell,s,p}(t)$ .

The results are now summarized in Fig. 6 for reference. Specifically, it is shown in Figs. 6(a) and 6(b) that SE and EE would degrade on their metrics, respectively, and if normalized with respect to the error-free results (i.e., SE and EE with  $P_e = 0$ ), their values would be around the four levels of  $P_e$  (i.e., 0, 0.1, 0.3, and 0.5) for each method with little fluctuations, as expected. In Fig. 6(d), the throughput is similarly shown to degrade, but if normalized with respect to its error-free result, it would be lower than the  $P_e$  levels, respectively, except the first 0. That is to say, even with a less data rate due to  $P_e$ , the throughput or admitted traffic does not decrease as much, and the excess traffic can be absorbed by the increased queue length as shown in Fig. 6(c). This reveals a unique merit brought by the DPP-based dynamic control with the data and virtual queues to achieve the system stability while maximizing the network utility.

Finally, we show that the distributed two-level RA algorithm (represented here by 2L RA-based and greedy-based methods) would be computationally efficient in terms of the LP optimal ratio and the objective improvement. First, by *LP optimal ratio*, we mean the number of optimal results obtained by a linear programming (LP) problem solver divided by the total number of results in the experiments. As shown in Table 2, this ratio is around 96% despite the CC allocation algorithm, exhibiting the fact that most of the experiment instances would have their coefficient matrices to be TUM and could be solved easily without an integer programming (IP) problem solver for the MRA problem. Second, by *objective improvement*, we mean the improvement degree on the objective function values obtained by

**TABLE 2.** Performance comparison between the methods based on integer programming and those based on linear programming.

	Optimal CC	Greedy CC
LP optimal ratio	95.57%	96.12%
objective improvement	-0.0651%	-0.0668%
$\eta_{SE}$ improvement	-2.735%	-2.696%
$\eta_{EE}$ improvement	0.348%	0.202%
$Q$ improvement	-0.530%	-0.179%
$\gamma$ improvement	-1.812%	-2.002%

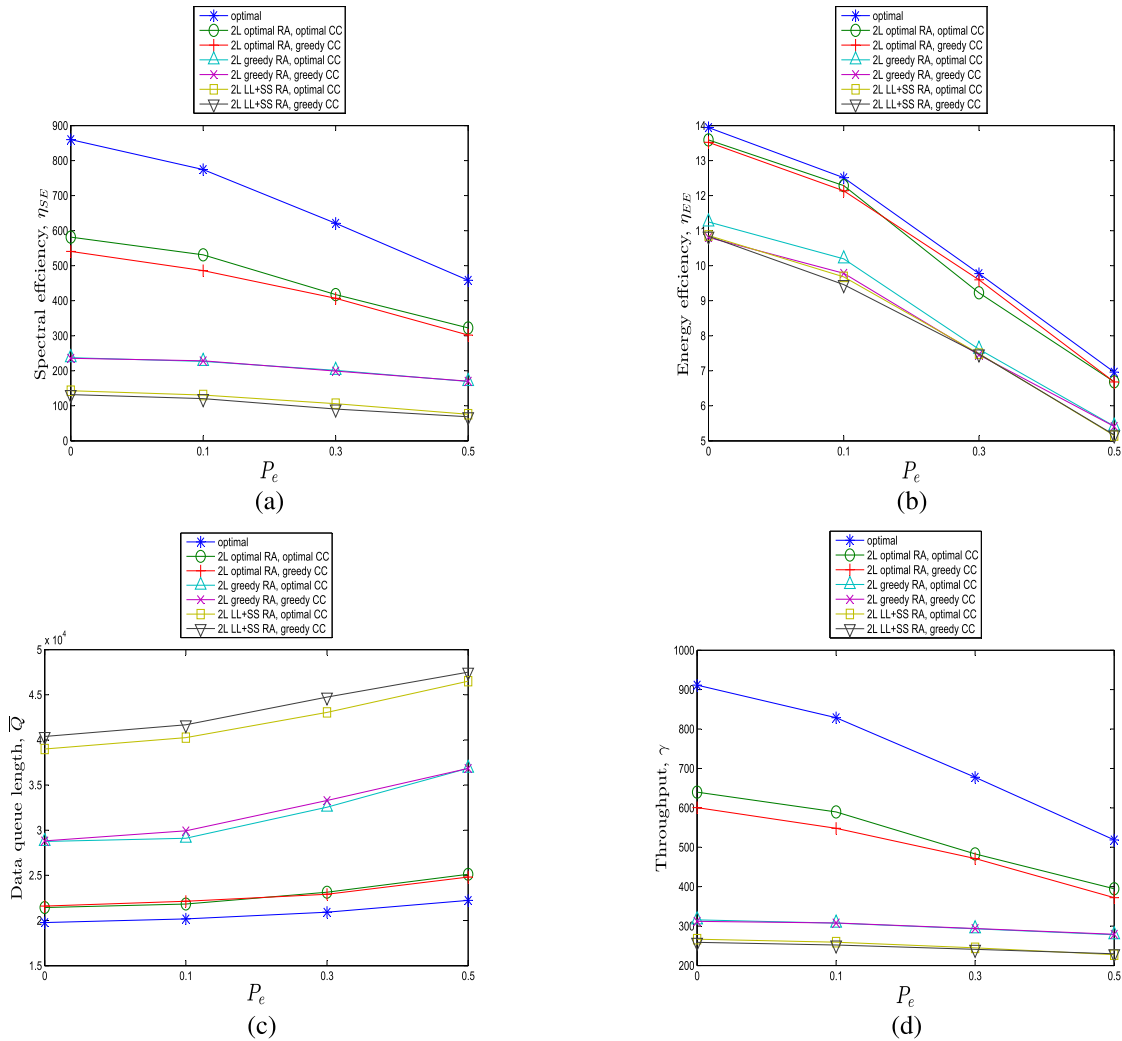


FIGURE 6. Performance comparison on (a)  $\eta_{SE}$ , (b)  $\eta_{EE}$ , (c)  $\bar{Q}$ , and (d)  $\gamma$ , by varying  $P_e$  while fixing  $W = 0.1$  and  $V = 0.5$ .

a linear programming problem solver compared with those by an integer programming problem solver. Here, the small negative value around  $-0.066\%$  exemplifies the performance trend that LP would only slightly degrade the objective function while significantly improve the computational complexity from nondeterministic polynomial (NP) to polynomial (P). Taking a closer look to see the components in the objective function, i.e.,  $SE$ ,  $EE$ , and queue length ( $\bar{Q}$ ), we can find also that as the objective is a weighted sum of these metrics, it is not necessary that all of them would be degraded as the objective function itself. Specifically,  $SE$  and  $\bar{Q}$  are degraded while  $EE$  is improved, and  $\gamma$  is also degraded, reflecting the same trend on  $SE$ .

As shown in the above, most of the problems encountered have their coefficient matrices to be TUM and can be solved efficiently. However, a MRA problem may still have its coefficient matrix without the property of TUM or 2-regular. In this case, developing a (re)formulation that leads to an integer polyhedron for a subset of constraints can be

valuable. This in fact motivates the study to find the classes of constraints so that specific cutting plans can be found to yield an exact representation or a tighter approximation of the convex hull of the feasible integer points. It further invokes the decomposition-based approaches that decompose the original IP formulation into several subproblems, one or more of which can be solved efficiently through an exact or approximated representation developed for the integer solutions [39]. In our work, such a development for the MRA problem would be very complex if it is not impossible, which requires our future study.

## VII. CONCLUSION

In this work, we have addressed a joint stochastic optimization problem on energy efficiency (EE), spectrum efficiency (SE), and queue length for downlink transmissions in the 5G LTE-based heterogeneous wireless networks with the advanced techniques for 5G such as multi-input multi-output (MIMO) and carrier aggregation (CA). Specifically, for the



multiple objectives to be optimized concurrently, we proposed a Lyapunov optimization framework on the downlink transmissions with an overall and comprehensive consideration for the joint problem concurrently involving resource allocation (RA), user association (UA), and power control (PC) in the complex networks. In particular, for the multiple resource allocation (MRA) problem involved, which is NP-hard and served as a key issue in this work, we had shown a reduced problem to be solved easily via linear relaxation when its coefficient matrix is totally unimodular (TUM), and accordingly, developed a distributed or semi-distributed algorithm with low computational complexity to resolve this NP-hard problem. Finally, in the numerical experiments, we have demonstrated that our framework can make a good tradeoff among EE, SE, and queue length, provide the design insights on the tradeoff decision through the system weights designed, and produce the performance metrics outperforming the greedy-based state-of-the-art counterparts.

**APPENDIX A**

**PROOF OF THEOREM 1**

First, by squaring both sides of the data queue dynamic  $\tilde{Q}_u(t+1) = \max\{\tilde{Q}_u(t) - \tilde{\mu}_u(t), 0\} + \tilde{R}_u(t)$ , we have

$$\tilde{Q}_u(t+1)^2 \leq \tilde{Q}_u(t)^2 + \tilde{R}_u(t)^2 + \tilde{\mu}_u(t)^2 + 2\tilde{Q}_u(t)(\tilde{R}_u(t) - \tilde{\mu}_u(t)) \quad (57)$$

as any  $A \geq 0, b \geq 0, Q \geq 0, (\max\{Q-b, 0\}+A)^2 \leq Q^2+A^2+b^2+2Q(A-b)$  would hold. Then, by summing over all  $u \in \mathcal{U}$  and taking the fact  $\tilde{R}_u(t) \leq \tilde{A}_u^{max}(t)$  and  $\tilde{\mu}_u(t) \leq \tilde{\mu}_u^{max}(t)$  into account, we have

$$\sum_{u \in \mathcal{U}} (\tilde{Q}_u(t+1)^2 - \tilde{Q}_u(t)^2) \leq \sum_{u \in \mathcal{U}} (\tilde{A}_u^{max})^2 + \sum_{u \in \mathcal{U}} (\tilde{\mu}_u^{max})^2 + 2 \sum_{u \in \mathcal{U}} \tilde{Q}_u(t)(\tilde{R}_u(t) - \tilde{\mu}_u(t)) \quad (58)$$

Similarly, for the virtual queue dynamics  $\tilde{H}_u$  and  $\tilde{Z}_s$ , we have

$$\begin{aligned} & \sum_{u \in \mathcal{U}} (\tilde{R}_u(t+1)^2 - \tilde{R}_u(t)^2) \\ & \leq \sum_{u \in \mathcal{U}} (\tilde{R}_u^{req})^2 + \sum_{u \in \mathcal{U}} (\tilde{A}_u^{max})^2 + 2 \sum_{u \in \mathcal{U}} \tilde{R}_u(t)(\tilde{R}_u^{req} - \tilde{R}_u(t)) \end{aligned} \quad (59)$$

and

$$\begin{aligned} & \sum_{s \in \mathcal{S}} (\tilde{Z}_s(t+1)^2 - \tilde{Z}_s(t)^2) \\ & \leq 2 \sum_{s \in \mathcal{S}} (\tilde{P}_s^{max})^2 + 2 \sum_{s \in \mathcal{S}} Z_u(t)(\tilde{P}_s(t) - \tilde{P}_s^{max}) \end{aligned} \quad (60)$$

Next, by combining these bounds and taking the expectation with respect to  $\tilde{\Theta}(t)$  on the both sides of the result, we derive

the one-slot conditional Lyapunov drift as

$$\begin{aligned} \Delta(\tilde{\Theta}(t)) & \leq \Gamma + \mathbb{E} \left\{ \sum_u \tilde{Q}(t)(\tilde{R}_u(t) - \tilde{\mu}_u(t)) | \tilde{\Theta}(t) \right\} \\ & \quad + \mathbb{E} \left\{ \sum_{u \in \mathcal{U}} \tilde{H}_u(t) (\tilde{R}_u^{req} - \tilde{R}_u(t)) | \tilde{\Theta}(t) \right\} \\ & \quad + \mathbb{E} \left\{ \sum_{u \in \mathcal{U}} \tilde{Z}_u(t) (\tilde{P}_u(t) - \tilde{P}_u^{req}) | \tilde{\Theta}(t) \right\} \end{aligned} \quad (61)$$

In the above,  $\Gamma = \sum_{u \in \mathcal{U}} (\tilde{A}_u^{max})^2 + \sum_{s \in \mathcal{S}} (\tilde{P}_s^{max})^2 + \frac{1}{2} \sum_{u \in \mathcal{U}} (\tilde{\mu}_u^{max})^2 + \frac{1}{2} \sum_{u \in \mathcal{U}} (\tilde{R}_u^{req})^2$  is obtained by combining the constant terms in the right hand sides of (58), (59) and (60), where  $\tilde{\mu}_u^{max}$  denotes the maximum of the transmission rate  $\tilde{\mu}_u$  that can be obtained on  $u$ ,  $\tilde{A}_u^{max}$  is  $A_u^{max}/R_{max}$ , and  $\tilde{R}_u^{req}$  represents the maximum request allowed for  $u$ . Finally, by removing the expectation operations on the constant terms, we can obtain the inequality shown in (29).

**APPENDIX B**

**PROOF OF THEOREM 2**

Assume  $\mathbb{E}\{R_u(t)\} \leq \eta_1, \mathbb{E}\{P_s(t)\} \leq \eta_2$ , and  $\mathbb{E}\{A_u(t)\} \leq \eta_3$ , where  $\eta_1, \eta_2$ , and  $\eta_3$  are finite positive constants. According to Theorem 4.5 in [29], if problem (28) is feasible and the boundedness assumption is given, then for any  $\delta > 0$  there is one policy that can satisfy

$$\mathbb{E}\{f(x^*(t)) | \Theta(t)\} = \mathbb{E}\{f(x^*(t))\} \leq f_{opt} - \delta \quad (62)$$

$$\mathbb{E}\{\tilde{R}_u^*(t) | \Theta(t)\} = \mathbb{E}\{\tilde{R}_u^*(t)\} \geq \tilde{R}_u^{req} - \delta \quad (63)$$

$$\mathbb{E}\{\tilde{P}_s^*(t) | \Theta(t)\} = \mathbb{E}\{\tilde{P}_s^*(t)\} \leq \tilde{P}_s^{max} + \delta \quad (64)$$

$$\mathbb{E}\{\tilde{R}_u^*(t) - \tilde{\mu}_u(t) | \Theta(t)\} = \mathbb{E}\{\tilde{R}_u^*(t) - \tilde{\mu}_u(t)\} \leq \delta \quad (65)$$

As the optimization is turned to minimize the R.H.S of (30), the optimal solution to this problem must satisfy

$$\begin{aligned} & V \Delta(\tilde{\Theta}(t)) - \mathbb{E}\{f(x(t)) | \tilde{\Theta}(t)\} \\ & \leq V \Gamma + V \sum_u \tilde{Q}_u(t) \mathbb{E}\{\tilde{R}_u(t) | \tilde{\Theta}(t)\} \\ & \quad - V \sum_u \tilde{Q}_u(t) \mathbb{E}\{\tilde{\mu}_u(t) | \tilde{\Theta}(t)\} \\ & \quad - V \sum_u \tilde{H}_u(t) \mathbb{E}\{\tilde{R}_u(t) | \tilde{\Theta}(t)\} \\ & \quad + V \sum_s \tilde{Z}_s(t) \mathbb{E}\{\tilde{P}_s(t) | \tilde{\Theta}(t)\} \\ & \quad + V \sum_u \tilde{R}_u^{req} \tilde{H}_u(t) - V \sum_s \tilde{P}_s^{req} \tilde{Z}_s(t) \\ & \quad - \mathbb{E}\{f(x^*(t)) | \tilde{\Theta}(t)\} \end{aligned} \quad (66)$$

Taking (62)-(65) into (66), we have

$$\begin{aligned} & V \Delta(\tilde{\Theta}(t)) - \mathbb{E}\{f(x(t)) | \tilde{\Theta}(t)\} \\ & \leq V \Gamma + \delta V \sum_u \tilde{Q}_u(t) + \delta V \sum_u \tilde{H}_u(t) \\ & \quad + \delta V \sum_s \tilde{Z}_s(t) - f_{opt} + \delta \end{aligned} \quad (67)$$

When  $\delta \rightarrow 0$ , (67) reduces to

$$V\Delta(\tilde{\Theta}(t)) - \mathbb{E}\{f(x(t))|\tilde{\Theta}(t)\} \leq V\Gamma - f_{opt} \quad (68)$$

Given that, we can take expectations of both sides of (68) and then adopt the law of iterated expectations, resulting in

$$\begin{aligned} V\mathbb{E}\{L(\tilde{\Theta}(t+1))\} - V\mathbb{E}\{L(\tilde{\Theta}(t))\} \\ \leq \mathbb{E}\{f(x(t))\} + V\Gamma - f_{opt} \end{aligned} \quad (69)$$

The results for all  $t \in \{0, 1, \dots, T-1\}$  can be further summed up, and dealt with the law of telescoping sums to yield

$$\begin{aligned} V\mathbb{E}\{L(\tilde{\Theta}(T))\} - V\mathbb{E}\{L(\tilde{\Theta}(0))\} \\ \leq \sum_{t=0}^{T-1} \mathbb{E}\{f(x(t))\} + TV\Gamma - Tf_{opt} \end{aligned} \quad (70)$$

Next, by rearranging the above while neglecting non-negative terms when appropriate, we can obtain, for all  $T > 0$ ,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{f(x(t))\} \geq f_{opt} - V\Gamma - \frac{V\mathbb{E}\{L(\tilde{\Theta}(0))\}}{T} \quad (71)$$

As  $T \rightarrow \infty$ , the lower bound is obtained while the upper bound is clearly represented by  $f_{opt}$  as the the maximum value of  $\overline{E}\{f(x(t))\}$  as defined, which completes the proof.

## REFERENCES

- [1] *Evolved Universal Terrestrial Radio Access (E-Utra): Physical Layer Procedures*, document 3GPP TS 36.213 Version 8.4.0 Release 8, 2009.
- [2] V. Chandrasekhar and J. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.
- [3] Y. Sun, R. P. Jover, and X. Wang, "Uplink interference mitigation for OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 614–625, Feb. 2012.
- [4] M. Andrews, V. Capdevielle, A. Feki, and P. Gupta, "Autonomous spectrum sharing for mixed LTE femto and macro cells deployments," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM)*, San Diego, CA, USA, Mar. 2010, pp. 1–5.
- [5] W. S. Jeon, J. Kim, and D. G. Jeong, "Downlink radio resource partitioning with fractional frequency reuse in femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 308–321, Jan. 2014.
- [6] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
- [7] Y. Wang, K. Pedersen, T. Sorensen, and P. Mogensen, "Carrier load balancing and packet scheduling for multi-carrier systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1780–1789, May 2010.
- [8] Y. Wang, K. I. Pedersen, T. B. Sorensen, and P. E. Mogensen, "Utility maximization in LTE-advanced systems with carrier aggregation," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, May 2011, pp. 1–5.
- [9] H.-S. Liao, P.-Y. Chen, and W.-T. Chen, "An efficient downlink radio resource allocation with carrier aggregation in LTE-advanced networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2229–2239, Oct. 2014.
- [10] J.-S. Liu, "Joint downlink resource allocation in LTE-advanced heterogeneous networks," *Comput. Netw.*, vol. 146, pp. 85–103, Dec. 2018.
- [11] S. Rostami, K. Arshad, and P. Rapajic, "Resource allocation algorithms for OFDM based wireless systems," in *Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Aug. 2015, pp. 1105–1110.
- [12] H. Lee, S. Vahid, and K. Moessner, "A survey of radio resource management for spectrum aggregation in LTE-advanced," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 745–760, 2nd Quart., 2014.
- [13] S. Rostami, K. Arshad, and P. Rapajic, "A joint resource allocation and link adaptation algorithm with carrier aggregation for 5G LTE-advanced network," in *Proc. 22nd Int. Conf. Telecommun. (ICT)*, Apr. 2015, pp. 102–106.
- [14] J.-S. Liu, C.-H. Lin, and H.-C. Huang, "Joint congestion control and resource allocation for energy-efficient transmission in 5G heterogeneous networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, p. 227, Sep. 2019.
- [15] S. Rostami, K. Arshad, and P. Rapajic, "Optimum radio resource management in carrier aggregation based LTE-advanced systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 580–589, Jan. 2018.
- [16] S.-B. Lee, I. Pefkianakis, S. Choudhury, S. Xu, and S. Lu, "Exploiting spatial, frequency, and multiuser diversity in 3GPP LTE cellular networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 11, pp. 1652–1665, Nov. 2012.
- [17] L. Gallo, F. Negro, I. Ghauri, and D. T. M. Slock, "Weighted sum rate maximization in the underlay cognitive MISO interference channel," in *Proc. IEEE 22nd Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2011, pp. 661–665.
- [18] K. Yang, S. Martin, C. Xing, J. Wu, and R. Fan, "Energy-efficient power control for Device-to-Device communications," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3208–3220, Dec. 2016.
- [19] Z. Zhou, M. Dong, K. Ota, J. Wu, and T. Sato, "Energy efficiency and spectral efficiency tradeoff in device-to-device (D2D) communications," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 485–488, Oct. 2014.
- [20] *LTE; Evolved Universal Terrestrial Radio Access (E-Utra); Physical Layer Procedures*, document 3GPP TS 36.213 version 10.1.0 Release 10, Apr. 2010.
- [21] E. Dahlman, J. Skold, and S. Parkvall, *4G: LTE/LTE-Advanced for Mobile Broadband*. New York, NY, USA: Academic, 2011.
- [22] A. Lozano and N. Jindal, "Transmit diversity vs. spatial multiplexing in modern MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 186–197, Jan. 2010.
- [23] *LTE; Evolved Universal Terrestrial Radio Access (E-Utra); Physical Channels And Modulation*, document 3GPP TS 36.211 version 14.2.0 Release 14, Mar. 2017.
- [24] D. T. Ngo and T. Le-Ngoc, *Architectures of Small-Cell Networks and Interference Management*, 1st ed. New York, NY, USA: Springer, 2014.
- [25] Y. Ji, F. Chen, and L. Liu, "MCS selection for performance improvement in downlink TD-LTE system," in *Proc. 2nd Int. Conf. Bus. Comput. Global Informatization*, Oct. 2012, pp. 687–690.
- [26] *Terrestrial Radio Access (E-Utra); Physical Channels and Modulation*, document TS 36.211, 3GPP, Evolved Universal, Jun. 2012.
- [27] C. Xiong, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy- and spectral-efficiency tradeoff in downlink OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3874–3886, Nov. 2011.
- [28] C. He, B. Sheng, P. Zhu, X. You, and G. Y. Li, "Energy- and spectral-efficiency tradeoff for distributed antenna systems with proportional fairness," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 894–902, May 2013.
- [29] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan and Claypool, 2010.
- [30] L. Kleinrock, *Queueing Systems: Theory*, vol. 1. New York, NY, USA: Wiley, 1975.
- [31] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: New insights," *Structural Multidisciplinary Optim.*, vol. 41, no. 6, pp. 853–862, Jun. 2010.
- [32] A. Schrijver, *Theory of Linear and Integer Programming*. Chichester, U.K.: Wiley, 1986.
- [33] B. Kotnyek, "A generalization of totally unimodular and network matrices." Ph.D. dissertation, London School Econ. Political Sci., London, U.K., 2002.
- [34] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2028–2035.
- [35] R. Robere, *Interior Point Methods and Linear Programming*. Toronto, ON, Canada: Univ. of Toronto, Dec. 2012.
- [36] O. Amin, E. Bedeer, M. H. Ahmed, and O. A. Dobre, "Energy efficiency–spectral efficiency tradeoff: A multiobjective optimization approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 1975–1981, Apr. 2016.
- [37] A. Tiwari, S. Suman, and P. Viswanathan, "Long term evolution (LTE) PROTOCOL verification of MAC scheduling algorithms in NetSim," Tetcos, Bengaluru, India, Tetcos White Paper, 2014.
- [38] *Terrestrial Radio Access (E-Utra); Physical Layer Procedures*, document TS 36.213, 3GPP, Evolved Universal, Jun. 2012.

- [39] Y. Pochet and L. A. Wolsey, *Production Planning by Mixed Integer Programming*. New York, NY, USA: Springer-Verlag, 2006.
- [40] W. Feng, Y. Wang, N. Ge, J. Lu, and J. Zhang, "Virtual MIMO in multi-cell distributed antenna systems: Coordinated transmissions with large-scale CSIT," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2067–2081, Oct. 2013, doi: 10.1109/JSAC.2013.1311009.



**JAIN-SHING LIU** (Member, IEEE) was born in Taipei, Taiwan, in 1970. He received the Ph.D. degree from the Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan.

He is currently a Professor with the Department of Computer Science and Information Engineering, Providence University, Taichung City, Taiwan. His research interests include design and performance analysis of wireless communication protocols, wireless local area networks, wireless sensor networks, personal communication networks, and wireless rechargeable networks. He is a member of the IEICE. Apart from other academic services being done, he more recently serves as a (technical) Program Committee Member of several international conferences such as the Second European Conference on Information and Communication Technology (ECICT), the Second International Conference on Advances in Computer Technology, Information Science, and Communications (CTISC), the Second International Conference on Advances in Computer Vision, Image, and Virtualization (CVIV), the Fourth International Conference on Communication and Future Internet (ICCFI), and the Third International Conference on Information Management and Processing (ICIMP). He was a recipient of the Best Paper Award from the Tenth Mobile Computing Workshop, in 2004, and was included in *Marquis Who's Who in Science and Engineering* (Tenth Anniversary Edition).



**CHUN-HUNG RICHARD LIN** was born in Kaohsiung, Taiwan. He received the B.S. and M.S. degrees from the Department of Computer Science and Information Engineering, National Taiwan University, in 1987 and 1989, respectively, and the Ph.D. degree from the Department of Computer Science, University of California at Los Angeles (UCLA), Los Angeles, in 1996. He joined National Chung Cheng University, Taiwan, in 1996. Since August 2000, he has been with the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung. His research interests include the design and control of mobile communication networks, the Internet of Things, edge computing and device AI, the mobile Internet, distributed computing, and embedded operating system design and implementation. He is a Senior Member of the ACM. He received the Investigative Research Award from the Pan Wen Yuan Foundation, Taiwan, in 2000, and the Junior Professor Research Award from National Sun Yat-sen University, in 2001.



**YU-CHEN HU** (Senior Member, IEEE) received the Ph.D. degree in computer science and information engineering from the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, in 1999. He is currently a Professor with the Department of Computer Science and Information Management, Providence University, Taichung City, Taiwan. His research interests include image and signal processing, data compression, information security, data engineering, and computer networks. He is also a member of the Phi Tau Phi Society, China.

...