Liu *et al. EURASIP Journal on Wireless Communications and Networking* (2019) 2019:227
https://doi.org/10.1186/s13638-019-1532-z

EURASIP Journal on Wireless Communications and Networking

RESEARCH

Open Access

Joint congestion control and resource allocation for energy-efficient transmission in 5G heterogeneous networks

Jain-Shing Liu[1], Chun-Hung Lin[2*] and Heng-Chih Huang[2]

Abstract

The deployment of small cells with carrier aggregation (CA) is a significant feature of fifth generation (5G) mobile communication systems which could be characterized by the multi-dimensional heterogeneity on their diversified requirements upon different resources. Taking the heterogeneity into account, we consider here a joint optimization problem wherein multiple kinds of resources are concurrently allocated to optimize the system throughput utility while enhancing the network energy efficiency (EE) and maintaining the system stability. Especially, for the high-dimensional non-deterministic polynomial (NP)-hard allocation problem embedded, we conduct a mathematical programming model involving nonlinear integer constraints to seek the long-term stable utility on throughput and introduce an iterative optimal modulation and coding scheme-based (optimal MCS-based) heuristic algorithm as an effective solver. In addition, as data traffic and channel condition will be time-varying in the real world, an admission control based on the Lyapunov technique that requires no prior knowledge on channel information is proposed to reduce the system overhead. Finally, not only the performance bound is derived in theory, but also the numerical experiments are conduced to reveal its characteristics with respect to the system parameter V and the EE requirement.

Keywords: Heterogeneous wireless networks, Joint optimization, Energy efficiency, Carrier aggregation

1 Introduction

For the next generation of mobile internet connectivity, 5G networks aim to offer increased data rate, shortened latency, improved energy efficiency, reduced cost, and other desired features. To this end, the communication society has proposed many techniques from different aspects such as dense heterogeneous networks, cloud-based radio access networks, energy-aware communications, and wireless energy harvesting [1]. Among these, the dense heterogeneous networks (HetNets) based on Long Term Evolution-Advanced (LTE-A) including carrier aggregation (CA) as its key feature would be particularly useful since the aggregation can achieve wider bandwidth and better energy efficiency (EE) [2]. Specially, with the aid of 4G framework, small-cells (SCs) that represent pico-cells, femto-cells, etc., can be more easily deployed to improve the 5G capacity by offloading the traffic from a macro cell (MC) to SCs [3].

Providing these benefits, designing HetNets, however, is a challenging work. One of the hardest challenges is caused by its resource and interference management because both MCs and SCs in a 5G network would tend to utilize the radio resources from the same service provider. To reduce the overhead emerged, the cells would be arranged under the so-called co-channel deployment, i.e., by spatially reusing the available spectrum or, specifically, by using a different set of channels or resource blocks (RBs) for macro base stations (MBSs) and small base stations (SBSs), as noted, e.g., in [4]. However, it is only a step toward the solution to the CA-capable LTE system that allows several component carriers (CCs) to be aggregated. That is, given CA, this system is still complicated by its requirement to modify the radio resource management (RRM) entity, including CC selection, RB allocation, modulation and coding scheme (MCS) assignment, and power allocation. For this complexity, many researches

*Correspondence: lin@cse.nsysu.edu.tw
[2]Department of Computer Science and Engineering, National Sun Yat-Sen University, 804 Kaohsiung, Taiwan
Full list of author information is available at the end of the article

had been done to develop the approaches on RRM that can properly allocate RBs, CCs [5–7], and even MCSs [8] to increase the performance. Now, as the standard evolves, more attentions are paid to the heterogeneous networks wherein the multiple types of resources would be allocated between MCs and SCs that are connected by backhaul links in a multi-tier sense [9, 10]. In such networks, high-capacity fiber backhaul (e.g., IEEE 802.3av 10G-EPON) will play a major role that consistently provides data rates 100 times higher than cellular networks to help in reaching the envisioned 10 Gbps peak data rates required by 5G [10]. Here, we focus on the multi-cell multi-tier networks equipped with high-capacity backhaul and introduce a solution based on discrete power control[1] , reflecting the fact that 3GPP LTE cellular networks only support discrete power levels in the downlink via a user-specific data-to-pilot-power offset parameter [13].

Given that, a joint congestion control and downlink resource allocation problem is particularly considered with the objective to maximize a long-term throughput utility subject to a system-wide EE requirement. The major challenge of this optimization problem is brought by the various constraints that are specific to the LTE-A system with CA. For this, a high-dimensional allocation problem involved is first formulated as a programming model whose constraints involve integer variables coupled with a nonlinear form, and optimally solving such a model at each transmission time interval (TTI) is impractical. In addition, for the data traffic and channel condition involved would be both time-varying in the system, an admission control is usually required to stabilize the data queue for each user equipment (UE). Thus, to address the combinatorial problem with queueing stability, an iterative optimal MCS-based heuristic algorithm inspired by the iterative linear programming-based heuristic [14, 15] is proposed to resolve the NP-hard allocation problem involved in the low layer. Then, as LTE would be a stochastic system with time-varying traffic and channel as noted, we further address its queueing stability problem at the high layer for the system. This is challenging because unlike deterministic optimization, stochastic optimization is usually hard to solve, and even harder than most well-known combinatorial optimization problems [16]. Given that, the Lyapunov-based optimization is considered to be a very useful technique to enable constrained optimization of time averages in general stochastic systems [17]. Accordingly, a Lyapunov optimization framework is developed to address the high layer problem focusing on the time-varying data traffic and channel condition without a priori knowledge of arrivals. By combining the solutions from the two layers, we are able to approach the optimal tradeoff with a control parameter V and satisfy the long-term EE requirement simultaneously. More specifically, the characteristics of this work can be summarized as follows:

- For the high-dimensional resource allocation optimization problem in the 5G LTE-A multi-tier multi-cell heterogeneous wireless networks that is a NP hard combinatorial problem, we first transform the corresponding nonlinear integer programming model into a linear counterpart that can be solved by conventional techniques.
- Then, an iterative optimal MCS-based heuristic algorithm or IOMHA for short inspired by the iterative linear programming-based heuristic is developed to approach the optima within a time limit. Given that, a two-layer method is proposed for the stochastic programming problem so that the data queue of each UE can be stabilized in the high layer based on the resources efficiently allocated in the low layer.
- Using the Lyapunov optimization framework, we realize a formulation to strike a balance between average throughput and average delay while guaranteeing the required EE performance and accommodating both traffic variations in the long term and channel fading in the short term, in the heterogeneous networks.
- We show that with the EE constraint enforced, the proposed algorithm has its performance advantage especially on EE through our simulation study. In the study, by gradually improving its result, our IOMHA is also shown to resolve the complex allocation problem effectively, trading the optimality of the NP-hard optimization problem off against a lower and controllable complexity to approach the optimal solution iteratively, in contrast to the other algorithms shown in, e.g., [5, 8], which would be done only once for obtaining suboptimal solutions to their allocation problems in LTE without a chance for further improvements.

The remainder of this paper is organized as follows. First, the related works are summarized in Section 2. Then, the scheduling constraints and queueing dynamics of the joint optimization problem are formulated in Section 3. The online control method based on Lyapunov drift-plus-penalty technique for this problem is proposed in Section 4, and the iterative optimal-MCS-based heuristic algorithm involved is introduced in Section 5. Given that, the performance bounds and evaluations of this work are presented in Sections 6 and 7, respectively. Finally, conclusions are drawn in Section 8.

[1] It is so considered according to the note shown in [11, 12] that discrete power control can offer two main benefits over continuous power control: (i) the transmitter design is simplified and, more importantly, (ii) the overhead of information exchange among network nodes is significantly reduced.

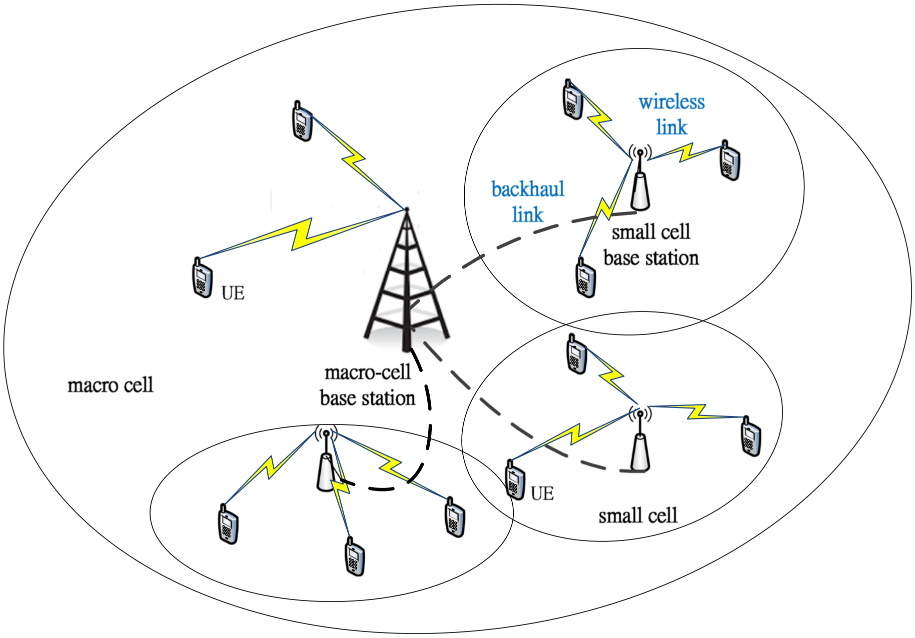

Fig. 1 An example of the heterogeneous network

2 Related works

For 5G networking, there are many networks continuously explored with variant aims on different performance metrics. Among these, energy efficiency (EE) plays a vital role in 5G as the future networks should effectively reduce the overall carbon footprint for the world to be sustainable. With respect to this issue, the authors in [18] had studied energy efficiency of resource allocation in orthogonal frequency division multiple access (OFDMA) downlink networks where the circuit power consumption and the minimum data rate required were both considered. More recently, the authors in [9] investigated energy-efficient power allocation and wireless backhaul bandwidth allocation in OFDMA heterogeneous small cell networks. Specifically, they proposed a near optimal iteration resource algorithm to solve the power and bandwidth allocation problem and suggested also a suboptimal low-complexity algorithm to this end.

Apart from the above, downlink radio resource allocation methods in LTE system with CA are particularly noted here for their potential on EE even without direct objectives for this aim. As surveyed in [19], a two-step allocation method was considered in [5–7] that first uses a load balancing scheme to assign CCs to UEs, and then schedules RBs of these CCs to reduce the computational complexity for the NP-hard RB/CC allocation problem. In addition, different joint allocation approaches had also been done with various efforts to reduce the time complexity. For example, the work in [20] divided the optimization problem into a number of subproblems for each CC to optimize its RB allocation independently. Then, after RB assignment, an iterative resource adjustment algorithm was performed to meet the CA capability requirement for UEs. Despite their differences, these approaches mainly focus on RB/CC allocation and pay no attention to the other constraints specific to LTE/LTE-A.

In addition, if categorized by using the number of cells, the authors in [21] have recently proposed for a single-cell scenario a downlink scheduling algorithm aiming to maximize the weighted sum of throughput constrained by the allocation rules of LTE. Similarly, the authors in [8] have addressed a downlink resource scheduling problem that takes also into account the MCS constraint for LTE, with a greedy-based algorithm to maximize the system throughput. With the notable performance gains obtained, these algorithms, however, consider no queueing dynamic resulted from the dynamic traffics that should be involved also. Next, as another category, for a multi-cell scenario, the authors in [22] proposed a resource allocation algorithm that accounts for MCS, RB, and transmit power, with inter-cell interference coordination, but ignores MCS constraint, CA, and queueing dynamic. In addition, the previous work in [23] considered a dynamic resource allocation algorithm for downlink transmission in a multi-cell network. However, it considered no discrete power allocation in the downlink and ignored the EE performance that is one of the most important factors impacting the system. Here, for the 5G multi-tier multi-cell networks based on LTE-A with discrete power levels, we first transform the nonlinear integer scheduling constraints to be involved into their linear counterparts as the previous. Then, an optimal-MCS-based heuristic algorithm inspired by the iterative linear programming-based heuristic is proposed to approach the optima within a time limit. Finally, a drift-plus-penalty approach for joint admission control and resource allocation with the requirement on EE and queueing stability is constructed that iteratively resolves the stochastic optimization problem involved for the long-term optimal throughput utility.

Table 1 A list of important symbols used in the problem formulation

$\mathcal{S}$	Set of base stations (BSs), with index $s \in \left\{1, ..., \mathbf{s} \triangleq \lvert\mathcal{S}\rvert\right\}$
$\mathcal{P}$	Set of power levels (PLs), with index $p \in \left\{1, ..., \mathbf{p} \triangleq \lvert\mathcal{P}\rvert\right\}$
$\mathcal{U}$	Set of user equipments (UEs), with index $u \in \left\{1, ..., \mathbf{u} \triangleq \lvert\mathcal{U}\rvert\right\}$
$\mathcal{C}$	Set of component carriers (CCs), with index $c \in \left\{1, ..., \mathbf{c} \triangleq \lvert\mathcal{C}\rvert\right\}$
$\mathcal{B}$	Set of resource blocks (RBs) per CC, with index $b \in \left\{1, ..., \mathbf{b} \triangleq \lvert\mathcal{B}\rvert\right\}$
$\mathcal{L}$	Set of modulation and coding schemes (MCSs) per RB, with index $l \in \left\{1, ..., \mathbf{l} \triangleq \lvert\mathcal{L}\rvert\right\}$
$\mathcal{N}_s$	Set of neighboring cells of cell s, which may interfere with s when given the same RBs
f_s	Maximum number of CCs able to be used by cell s
k_u	Maximum number of CCs able to be assigned to UE u
W	Balance weight
$\Psi_{u,c,b,s,p}$	Index of the highest-rate MCS used by UE u on RB b of CC c at PL p of cell s
r_l	Achieved transmission rate of an RB on MCS l
$\underline{e}_{u,c,b,l,s,p}$	Binary variable showing if RB b of CC c on MCS l at PL p of cell s is assigned to UE u
$v(\underline{e})$	Achieved transmission rate with the allocation indicated by $\underline{e}$
$\underline{\hat{e}}, \underline{\hat{e}}_1, ..., \underline{\hat{e}}_{10}$ and $\underline{\tilde{e}}_4, \underline{\tilde{e}}_5$	Binary variables similar to $\underline{e}$ with certain elements to be varied while fixing the others, for the different scenarios shown in Section 3.2 and Eqs. (5)–(12), (14), (16), and (18)–(21)
$\underline{\hat{y}}_1, ..., \underline{\hat{y}}_5$	Auxiliary variables for the different scenarios shown in Eqs. (13), (15), (17), and (18)–(21)
$P_{\text{tot}}(t), \overline{P}_{\text{tot}}, P_{\max}$	Total power consumption at time t, its limit of the time-average expectation, and the maximum transmit power
$R_{\text{tot}}(t), \overline{R}_{\text{tot}}$	Total data rate at time t, and its limit of the time-average expectation
$Q(t), \overline{Q}$	Data queue length at time t, and its limit of the time-average expectation
$Z_u(t), H_u(t)$	Virtual queue lengths at time t corresponding to (27-C5) and (27-C6), respectively, for UE u
$A_u(t), A_u^{\max}, \lambda_u$	Traffic arrival of UE u at time t, its maximum value allowed, and arrival rate of UE u
$R_u(t), \mu_u(t)$	Admitted traffic of UE u, and link (service) rate of UE u, at time t
$\bar{r}_u$	Time-average throughput of UE u
$\eta_{EE}, \eta_{EE}^{\text{req}}$	Energy efficiency and its requirement

3 Methods

3.1 System model and problem formulation

In the sequel, we consider a multi-tier multi-cell heterogeneous network as exemplified in Fig. 1, consisting of $\mathbf{s}$ base stations (including a MBS and $\mathbf{s} - 1$ SBSs) and $\mathbf{u}$ UEs located in the service area of these cells. In addition, the network is equipped with a number of $\mathbf{c}$ CCs. Each CC has $\mathbf{b}$ RBs, and each RB can use one of $\mathbf{l}$ MCSs for transmission. Further, there are $\mathbf{p}$ discrete power levels (PLs), and MBS/SBSs can choose among $P = \{\sigma_1 P_{\max}, \sigma_2 P_{\max}, ..., \sigma_{\mathbf{p}=|P|} P_{\max}\}$ to transmit, where $0 < \sigma_1 < \sigma_2, ..., < \sigma_{\mathbf{p}=|P|} = 1$ and $P_{\max}$ denotes the maximum power as that in [12]. In summary, there are $\mathcal{U}, \mathcal{C}, \mathcal{B}, \mathcal{L}, \mathcal{S}$, and $\mathcal{P}$ to represent the set of UEs, the set of CCs, the set of RBs per CC, the set of MCSs per RB, the set of base stations (BSs), and the set of power levels (PLs) with u, c, b, l, s, and p as their indices, and $\mathbf{u} = |\mathcal{U}|, \mathbf{c} = |\mathcal{C}|, \mathbf{b} = |\mathcal{B}|, \mathbf{l} = |\mathcal{L}|, \mathbf{s} = |\mathcal{S}|$, and $\mathbf{p} = |\mathcal{P}|$ as their cardinalities, respectively. Given that, we focus on downlink transmission in the 5G heterogeneous network based on LTE-A and consider a stochastic communication system whose traffic load is changed from time to time, requiring an online admission algorithm for its stability. Further, its channel condition is also time-varying. For this condition, a UE would inspect reference signals currently transmitted from MBS or SBSs to estimate the channel quality of each RB [24]. After that, it will send a feedback report with the channel quality indicator (CQI) whose value would then be mapped to the highest-rate MCS adoptable by the UE for receiving the corresponding RB from MBS/SBSs [25]. Then, with the information from UEs and SBSs, MBS is responsible for admission control, resource scheduling, and link adaption. For easy reference, the important symbols for the formulation are summarized in Table 1 in advance.

3.2 Multi-resource allocation

To show the multiple types of resources involved more concisely, we denote by $\underline{e}$ a binary variable or an element of the feasible set Ξ representing all possible allocations, where $\underline{e}_{u,c,b,l,s,p} \triangleq (u_{\underline{e}} = u, c_{\underline{e}} = c, b_{\underline{e}} = b, l_{\underline{e}} = l, s_{\underline{e}} = s, p_{\underline{e}} = p)$ with value of 1 exhibits that RB b of CC c on MCS l at PL p of cell s is assigned to UE u, and 0 otherwise. Further, let $\Psi_{u,c,b,s,p}$ be the index of the highest-rate MCS allowed among the possible transmissions, $\underline{\hat{e}}_{u,c,b,s,p} = (u, c, b, \hat{l}, s, p), \forall \hat{l} \in \mathcal{L}$. Given that, the achieved transmission rate with the allocation, $v(\underline{e})$, is the data rate of an RB on MCS l, r_l, for $l \leq \Psi_{u,c,b,s,p}$, and 0 otherwise.

3.3 Channel, power, and energy efficiency model

Accordingly, the allocation (or scheduling) algorithm is conducted to accommodate a slow fading network wherein channel condition would remain unchanged during the resource allocation period (Ch. 6 of [26]), which complies with the high-rate network with reduced degree of mobility. In this situation, the signal-to-noise ratio (SNR) from BS s to UE u using RB b of CC c at PL p in time t can be represented by

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

where $h_{s,u}^{c,b}$ is the channel gain from transmitter (MBS or SBS) s to receiver (UE) u using RB b of CC c, and $d_{s,u}$ is the distance from s to u. The channel is considered to be Rayleigh fading which yields the channel gain following the exponential distribution. In addition, ρ is the path-loss factor and $N_{s,u}^{c,b}$ is the noise experienced by u when s transmits to u on RB b of CC c. Providing that, an empirical downlink SNR to CQI mapping for LTE such as that in [27, 28] could be used to estimate the CQIs to be returned to BSs. Then, according to the CQIs collected, MBS would decide each MCS index l for the downlink transmission from BS $s_{\underline{e}} = s$ to UE $u_{\underline{e}} = u$ using RB $b_{\underline{e}} = b$ of CC $c_{\underline{e}} = c$ at PL $p_{\underline{e}} = p$, in terms of $\underline{e}$, and transmit the decisions to all SBSs it associates via the backhaul network. Consequently, as 3GPP specifies the transmit data rate of each MCS index l using table representation [24], the data rate $v(\underline{e})$ would be obtained through a function or table mapping, r_l, for each RB on MCS l. Given that and the feasible allocation set Ξ, the total data rate can be given by $R_{\text{tot}}(t) = \sum_{\underline{e}\in\Xi}\left(\underline{e}(t) \times v\left(\underline{e}(t)\right)\right)$. Similarly, the total power consumption can be obtained by $P_{\text{tot}}(t) = \sum_{\underline{e}\in\Xi}\left(\underline{e}(t)\left(P_{s,u}^{p}(t) + P_{s,u}^{c}\right)\right)$, where $P_{s,u}^{p}(t)$ is the transmit power from $s_{\underline{e}} = s$ to $u_{\underline{e}} = u$ at power level $p_{\underline{e}} = p$, and $P_{s,u}^{c}$ is the constant circuit power for this transmission.

Specifically, in the stochastic system, we are interested in the limits of the time-average expectations of the above metrics. That is,

$$\overline{R}_{\text{tot}} = \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{R_{\text{tot}}(\tau)\} \tag{2}$$

$$\overline{P}_{\text{tot}} = \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{P_{\text{tot}}(\tau)\} \tag{3}$$

In terms of the long-term metrics, the energy efficiency is considered as the ratio of the long-term aggregated rate to the long-term total energy consumption as

$$\eta_{EE} = \frac{\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{R_{\text{tot}}(\tau)\}}{W\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{P_{\text{tot}}(\tau)\}} = \frac{\overline{R}_{\text{tot}}}{W\overline{P}_{\text{tot}}} \tag{4}$$

where W is used to accommodate the quantitative difference between the two metrics in the ratio.

3.4 Scheduling constraints

For the heterogeneous network with CA, we have the following scheduling constraints. First, as the basic unit for the transmission, each RB can be assigned to a single UE u at most with a certain MCS l. To show this, we let $\underline{\hat{e}}_1 \triangleq (\hat{u}, c, b, \hat{l}, s, p)$ be the binary allocation variables with different $\hat{u} \in \mathcal{U}$ and $\hat{l} \in \mathcal{L}$ while fixing $c_{\underline{\hat{e}}_1} = c, b_{\underline{\hat{e}}_1} = b, s_{\underline{\hat{e}}_1} = s$, and $p_{\underline{\hat{e}}_1} = p$. Given that, this constraint can be simply shown by

$$\sum_{\forall\underline{\hat{e}}_1=(\hat{u},c,b,\hat{l},s,p)} \mathbb{1}\left\{\underline{\hat{e}}_1\right\} \leq \mathbb{1}, \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \tag{5}$$

where $\mathbb{1}\{x\}$ denotes an indicator function whose value is 1 if x is true, and 0 otherwise. In addition, the notations are given without the time index t for brevity. Further, according to LTE-A, it is required that if a UE u is assigned a CC c by a BS s serving it, then all RBs of c allocated to u should use the same MCS l to transmit. More specifically, the MCS constraint based on LTE-A is considered as

$$\sum_{\forall l\in\mathcal{L}} \mathbb{1}\left\{\sum_{\forall\underline{\hat{e}}_2=(u,c,\hat{b},\hat{l},s,p)} \mathbb{1}\left\{\underline{\hat{e}}_2 | l_{\underline{\hat{e}}_2} = l\right\}\right\} \leq 1, \quad \forall u \in \mathcal{U}, \forall c \in \mathcal{C}, \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \tag{6}$$

As noted in Section 3.2, a UE can only use a MCS less than or equal to $\Psi_{u,c,b,s,p}$. If not, it could lead to an unacceptable bit error rate on transmission, and the transmission should be discarded. Thus, we have the following constraint

$$\sum_{\forall\underline{\hat{e}}_3=(u,c,b,\hat{l},s,p)} \mathbb{1}\left\{\underline{\hat{e}}_3 | l_{\underline{\hat{e}}_3} > \Psi_{u,c,b,s,p}\right\} = 0, \quad \forall u \in \mathcal{U}, \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \tag{7}$$

Apart from the above without special notion on the number of cells involved, here, we take into account the constraints specific to the multi-cell environment as well as follows. First, to reduce overheads on backhaul, it is commonly considered that a UE is only served by a single BS s, which implies a monopoly constraint as

$$\sum_{\forall\underline{\hat{e}}_4=(u,\hat{c},\hat{b},\hat{l},s,\hat{p})} \mathbb{1}\left\{\underline{\hat{e}}_4\right\} \times \sum_{\forall\underline{\tilde{e}}_4=(u,\tilde{c},\tilde{b},\tilde{l},\tilde{s},\tilde{p})} \mathbb{1}\left\{\underline{\tilde{e}}_4 | \tilde{s} \in \mathcal{S}\backslash s\right\} = 0, \quad \forall u \in \mathcal{U}, \forall s \in \mathcal{S} \tag{8}$$

Second, even given the spatial reuse principle, it should be still considered that an RB b of CC c already allocated to a BS s can not assigned to its neighboring BSs $s' \in \mathcal{N}_s$ to avoid the leading cause of inter-cell interference. Consequently, it also implies a monopoly constraint as

$$\sum_{\forall \underline{\hat{e}}_5=(\hat{u},c,b,\hat{l},s,\hat{p})} \mathbb{1}\left\{\underline{\hat{e}}_5\right\} \times \sum_{\forall \underline{\tilde{e}}_5=(\tilde{u},c,b,\tilde{l},\tilde{s},\tilde{p})} \mathbb{1}\left\{\underline{\tilde{e}}_5 | \tilde{s} \in \mathcal{N}_s\right\} = 0, \\ \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S} \quad (9)$$

Moreover, there are two cardinality constraints to be involved. First, each UE u has its own limitation on the number of CC allocated by a BS s, denoted by k_u. For example, it can be enforced that a UE of LTE 8/9 can only use 1 CC while a LTE-A UE would use 2 CCs. In general, such a constraint can be written by

$$\sum_{\forall c \in C} \mathbb{1}\left\{ \sum_{\forall \underline{\hat{e}}_6=(u,\hat{c},\hat{b},\hat{l},\hat{s},\hat{p})} \mathbb{1}\left\{\underline{\hat{e}}_6 | c_{\underline{\hat{e}}_6} = c\right\} \right\} \leq k_u, \forall u \in \mathcal{U} \quad (10)$$

Similarly, a cardinality constraint for each BS s to equip with at most f_s CCs for communication can be represented by

$$\sum_{\forall c \in C} \mathbb{1}\left\{ \sum_{\forall \underline{\hat{e}}_7=(\hat{u},\hat{c},\hat{b},\hat{l},s,\hat{p})} \mathbb{1}\left\{\underline{\hat{e}}_7 | c_{\underline{\hat{e}}_7} = c\right\} \right\} \leq f_s, \forall s \in \mathcal{S} \quad (11)$$

3.5 Linear transformation of scheduling constraints

As shown in above, the indicator functions with complex conditions could be nonlinear on the binary integer variables involved. For those especially involving logical operations, we refer to the work in [29] showing that two either-or constraints $f(x_1, x_2, ..., x_n) \leq 0$ and $g(x_1, x_2, ..., x_n) \leq 0$ can be transformed to $f(x_1, x_2, ..., x_n) \leq My$ and $g(x_1, x_2, ..., x_n) \leq M(1-y)$ with a large number M and auxiliary binary variable y such that $f(x_1, x_2, ..., x_n) \leq M$ and $g(x_1, x_2, ..., x_n) \leq M$. Here, given a certain l, the condition in the outer indicator function in (6) implies a logic operation to choose among the multiple binary variables $\underline{\hat{e}}_8 \triangleq (u, c, \hat{b}, l, s, p)$ which satisfy the condition $l_{\underline{\hat{e}}_8} = l$ shown in the inner indicator function can be transformed to $\sum \underline{\hat{e}}_8 \leq \mathbf{b}\underline{y}_1$, where $\underline{\hat{y}}_1 \triangleq (u, c, \hat{l}, s, p)$ is defined to play the role of the auxiliary variable y, and $\mathbf{b} = |\mathcal{B}|$ defined before plays the role of M. Given that, the constraint that all RBs should be assigned only the same MCS in this context can be done by $\sum \underline{\hat{y}}_1 \leq 1$ on the auxiliary variables. Therefore, (6) can be transformed to the linear counterparts as

$$\sum_{\underline{\hat{e}}_8=(u,c,\hat{b},l,s,p)} \underline{\hat{e}}_8 \leq \mathbf{b}\underline{\hat{y}}_1, \\ \forall u \in \mathcal{U}, \forall c \in \mathcal{C}, \forall l \in \mathcal{L}, \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \quad (12)$$

$$\sum_{\underline{\hat{y}}_1=(u,c,\hat{l},s,p)} \underline{\hat{y}}_1 \leq 1, \forall u \in \mathcal{U}, \forall c \in \mathcal{C}, \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \quad (13)$$

In addition, the inner indicator functions in (10) and (11) could be also regarded as the logic operations to choose among the binary variables that can satisfy the conditions specified, and can apply a transformation like the above. Specifically, with the aid of the auxiliary variables $\underline{\hat{y}}_2$ in addition to the binary variables $\underline{\hat{e}}_9$, both shown below, (10) can be represented by

$$\sum_{\underline{\hat{e}}_9=(u,c,\hat{b},\hat{l},\hat{s},\hat{p})} \underline{\hat{e}}_9 \leq \mathbf{b}\underline{\hat{y}}_2, \quad \forall u \in \mathcal{U}, \forall c \in \mathcal{C} \quad (14)$$

$$\sum_{\underline{\hat{y}}_2=(u,\hat{c})} \underline{\hat{y}}_2 \leq k_u, \qquad \forall u \in \mathcal{U} \quad (15)$$

Similarly, with the auxiliary variables $\underline{\hat{y}}_3$ and the binary variables $\underline{\hat{e}}_{10}$ shown below, (11) can be transformed to

$$\sum_{\underline{\hat{e}}_{10}=(\hat{u},c,\hat{b},\hat{l},s,\hat{p})} \underline{\hat{e}}_{10} \leq \mathbf{b}\underline{\hat{y}}_3, \quad \forall c \in \mathcal{C}, \forall s \in \mathcal{S} \quad (16)$$

$$\sum_{\underline{\hat{y}}_3=(s,\hat{c})} \underline{\hat{y}}_3 \leq f_s, \qquad \forall s \in \mathcal{S} \quad (17)$$

Apart from these, the monopoly constraints shown in (8) could be rewritten with linear forms as well. To this end, let $\sum_{\underline{\hat{e}}_4=(u,\hat{c},\hat{b},\hat{l},s,\hat{p})} \underline{\hat{e}}_4$ be the first metric for transforming the logical either-or constraints in [29] (here, $\underline{\hat{e}}_4$ is directly drawn because $\mathbb{1}\left\{\underline{\hat{e}}_4\right\} = \underline{\hat{e}}_4$) and $\sum_{\underline{\tilde{e}}_4=(u,\tilde{c},\tilde{b},\tilde{l},\tilde{s}\in\mathcal{S}\backslash s,\tilde{p})} \underline{\tilde{e}}_4$ be the second metric. Then, by introducing also the large number, $M = \mathbf{ucblsp}$, and the auxiliary binary variables $\underline{\hat{y}}_4$, we can transform (8) into its linear counterparts as

$$\sum_{\underline{\hat{e}}_4=(u,\hat{c},\hat{b},\hat{l},s,\hat{p})} \underline{\hat{e}}_4 \leq M\underline{\hat{y}}_4, \quad \forall u \in \mathcal{U}, \forall s \in \mathcal{S} \quad (18)$$

$$\sum_{\underline{\tilde{e}}_4=(u,\tilde{c},\tilde{b},\tilde{l},\tilde{s}\in\mathcal{S}\backslash s,\tilde{p})} \underline{\tilde{e}}_4 \leq M\left(1-\underline{\hat{y}}_4\right), \quad \forall u \in \mathcal{U}, \forall s \in \mathcal{S} \quad (19)$$

Similarly, by introducing the auxiliary binary variables $\underline{\hat{y}}_5$ and M into (9), we have the linear counterparts as

$$\sum_{\forall \underline{\hat{e}}_5=(\hat{u},c,b,\hat{l},s,\hat{p})} \underline{\hat{e}}_5 \leq M\underline{\hat{y}}_5, \quad \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S} \quad (20)$$

$$\sum_{\forall \underline{\tilde{e}}_5=(\tilde{u},c,b,\tilde{l},\tilde{s}\in\mathcal{N}_s,\tilde{p})} \underline{\tilde{e}}_5 \leq M(1-\underline{\hat{y}}_5), \forall c \in \mathcal{C}, \forall b \in \mathcal{B}, \forall s \in \mathcal{S} \quad (21)$$

3.6 Stochastic system and queue dynamic

Now, even though the scheduling constraints can be linearly transformed, the design of 5G heterogenous networks in dynamic is still challenged by stochastic channel condition and time-varying data traffic. Specifically, the random channel gains are considered to be exponentially distributed, and the downlink traffics to UEs in time t are represented by an vector $\mathbf{A}(t) \overset{\triangle}{=} (A_1(t), ..., A_{\mathbf{u}}(t))$, according to an independently and identically distributed (i.i.d.) distribution over t whose expectations would be $\mathbb{E}\{A(t)\} = \lambda \overset{\triangle}{=} (\lambda_1, ..., \lambda_{\mathbf{u}})$. In addition, it is assumed that a maximum A_u^{max} exists that any non-negative traffic arrival $A_u(t)$ will not exceed. Given that, however, the statistics of $\mathbf{A}(t)$ are still unknown and its capacity region is also hard to estimate for a real system. Thus, without flow control, the data queues can not be stabilized in general. For this issue, an admission control method is proposed here to determine $R_u(t)$ out of $A_u(t)$, followed by an allocation algorithm introduced next to provide link rates $\mu_u(t)$ for serving the admitted traffic. To realize this mechanism, the data queueing dynamic for UE $u \in \mathcal{U}$ is formulated first by

$$Q_u(t+1) = \max\{Q_u(t) - \mu_u(t), 0\} + R_u(t) \tag{22}$$

Then, the average data queue length on each u would be conducted to be strongly stable as

$$\overline{Q} \overset{\triangle}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{u\in\mathcal{U}} \mathbb{E}\{Q_u(t)\} < \infty \tag{23}$$

Note that in (22), the service rate μ_u defined for a UE u can be obtained by

$$\mu_u = \sum_{\underline{\hat{e}}_u = (u,\hat{c},\hat{b},\hat{l},\hat{s},\hat{p})} \left(\underline{\hat{e}}_u \times v(\underline{\hat{e}}_u)\right), \quad \forall u \in \mathcal{U} \tag{24}$$

Similarly, we ignore the time index t in above for brevity. As a result, $R_{tot} = \sum_{u\in\mathcal{U}} \mu_u$. Moreover, we can see that not only the resource scheduling to provide service, but also the throughput $r_u(t) \overset{\triangle}{=} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R_u(\tau)\}$ to represent its performance contributes the queueing dynamic (22). Given that, the time-average throughput $\overline{r}_u$, which represents the admitted and transmitted data for u in the long term, is considered as the key metric in the time-varying system for optimization:

$$\overline{r}_u \overset{\triangle}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_u(t)\} \tag{25}$$

3.7 Problem formulation

Taking all the above into account, we can now formulate the joint congestion control and resource allocation with EE-delay tradeoff problem (JCREEP) for the heterogeneous wireless network by the following stochastic programming model:

$$\begin{array}{lll}
\textbf{Maximize} & \sum_{u\in\mathcal{U}} \phi(\overline{r}_u) & \\
\textbf{subject to} & \text{C1:}\overline{Q} < \infty & \\
 & \text{C2:}0 \le r_u \le \lambda_u, & \forall u \\
 & \text{C3:}0 \le R_u(t) \le A_u(t) \le A_u^{\max}, & \forall u, \forall t \\
 & \text{C4:}(5), (7), (12) - (21), & \forall t \\
 & \text{C5:}\eta_{EE} \ge \eta_{EE}^{\text{req}} &
\end{array} \tag{26}$$

In above, (26-C1) denotes the strong stability of data queues in the long term. (26-C2) and (26-C3) exhibit the constraints enforcing that the average and instantaneous throughput to be feasible. (26-C4) shows the resource scheduling constraints in linear forms most done in Section 3.5. Note that, even with the linear forms, the constraints (5), (7), and (12)-(21) are still involving the specific binary integer variables $\underline{\hat{e}}, \underline{\hat{y}}$, or both, and deciding these binary variables concurrently for the optimization is a combinatorial problem that is NP-hard if no special structures are imposed. Finally, (26-C5) ensures that the EE performance will achieve the requirement η_{EE}^{req} predefined. It is worth noting here that, by using EE in (4) as one of the constraints rather than the objective function, we can maximize the system utility and guarantee EE of the whole system simultaneously, which may not be achieved by simply optimizing the EE metric as the program objective as that in the related works [30, 31].

4 Optimization for the stochastic system

With the aid of Lyapunov drift-plus-penalty technique and the iterative heuristic algorithm to be introduced, we would next develop an online control framework to resolve (26) composed of the resource allocation problem and the traffic admission control problem in the stochastic system.

4.1 Equivalent transformation

As shown in (26), JCREEP involves a function $\phi(\overline{x})$ with a time-average parameter, say $\overline{x}$, rather than a time-average function $\overline{\phi(x)}$ with a pure parameter, say x. To use the Lyapunov drift-plus-penalty technique in the optimization as shown in [17], we would reformat JCREEP to involve the latter by first introducing an infinite sequence of random vectors in $\mathbb{R}$ as $\gamma = (\gamma_1(t), ..., \gamma_{\mathbf{u}}(t))$. Then, we define a time-average metric $\overline{\gamma}_u \overset{\triangle}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{\gamma_u(t)\}$ and a time-average function $\overline{\phi(\gamma_u)} \overset{\triangle}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{\phi(\overline{\gamma}_u(t))\}$. With these, JCREEP can be transformed to an equivalent problem, say eJCREEP, as follows:

$$\begin{array}{lll} \textbf{Maximize} & \sum_{u\in\mathcal{U}} \overline{\phi(\gamma_u)} & \\ \textbf{subject to} & \text{C1:}\overline{Q} < \infty & \\ & \text{C2:}0 \le r_u \le \lambda_u, & \forall u \\ & \text{C3:}0 \le R_u(t) \le A_u(t) \le A_u^{\max}, & \forall u, \forall t \\ & \text{C4:(5), (7), (12)} - \text{(21)}, & \forall t \\ & \text{C5:}\eta_{EE} \ge \eta_{EE}^{\text{req}} & \\ & \text{C6:}\overline{\gamma}_u \le \overline{r}_u, & \forall u \\ & \text{C7:}0 \le \gamma_u(t) \le A_u^{\max}, & \forall u, \forall t \end{array} \tag{27}$$

4.2 Virtual queues

In eJCREEP, (27-C6) denotes the constraints to ensure the system stability representing the fact that the arrivals would be eventually served. To conform these constraints, we define a virtual queue H_u for each $u \in \mathcal{U}$. Specifically, given an initial value $H_u(0) = 0$, such a queue will be updated by

$$H_u(t+1) = \max\{H_u(t) - R_u(t), 0\} + \gamma_u(t) \tag{28}$$

In addition, for the EE performance requirement in (27-C5), we define a virtual queue Z which evolves as

$$Z(t+1) = \max\{Z(t) - R_{tot}(t), 0\} + W\eta_{EE}P_{tot}(t) \tag{29}$$

In terms of queueing dynamic similar to (22), the variables $\gamma_u(t)$ and $W\eta_{EE}P_{tot}(t)$ can be regarded as the arrivals of the virtual queues in (28) and (29), while $R_u(t)$ and $R_{tot}(t)$ as the service rates of these virtual queues, respectively.

4.3 Online control based on Lyapunov drift-plus-penalty

Given $H_u(t), Z(t)$, and $Q_u(t)$ for the online control method, we define $\Theta(t) \triangleq \{Q_u(t), H_u(t), Z(t) : u \in \mathcal{U}\}$, a vector concatenating all the data and virtual queues involved. Further, for realizing a scalar metric to reflect the queue congestion, we define a quadratic Lyapunov function corresponding to the system as

$$L(\Theta(t)) \triangleq \frac{1}{2}\left\{\sum_{u\in\mathcal{U}} Q_u(t)^2 + \sum_{u\in\mathcal{U}} H_u(t)^2 + Z(t)^2\right\} \tag{30}$$

Here, a small value of $L(\Theta(t))$ implies that the sizes of data queues and virtual queues are all small and that the queues have strong stability. Given that, the queue stability can be ensured by persistently pushing the Lyapunov function toward a lower congestion state. Thus, to stabilize these queues, a one-slot conditional Lyapunov drift can be defined by

$$\Delta(\Theta(t)) \triangleq \mathbb{E}[L(\Theta(t+1)) - L(\Theta(t))|\Theta(t)] \tag{31}$$

Now, apart from satisfying the average constraints and optimizing the system throughput utility, with this drift, our online dynamic control algorithm can observe the data and virtual queues, the current channel conditions, and the traffic states at each slot t so that $R_u(t)$ can be determined and the resources be allocated to support $\gamma_u(t)$, by minimizing a bound on the following Lyapunov conditional drift-plus-penalty expression:

$$\Delta(\Theta(t)) - V\mathbb{E}\left\{\sum_{u\in\mathcal{U}} \overline{\phi(\gamma_u(t))}|\Theta(t)\right\} \tag{32}$$

In above, the system parameter V is a non-negative weight to represent the emphasis on the utility maximization compared with the queue stability and can be flexibly chosen to make a tradeoff between them. More precisely, with the above queueing dynamics, an upper bound for the drift-plus-penalty-based algorithm can be obtained with the following theorem.

Theorem 1 *At slot t, for any observed queue state* $\Theta(t)$*, and* $V \ge 0$*, the Lyapunov drift-plus-penalty algorithm can satisfy the following inequality:*

$$\begin{aligned} &\Delta(\Theta(t)) - V\mathbb{E}\left\{\sum_{u\in\mathcal{U}} \overline{\phi(\gamma_u(t))}|\Theta(t)\right\} \le \\ &\quad \Gamma - V\mathbb{E}\left\{\sum_{u\in\mathcal{U}} \overline{\phi(\gamma_u(t))}|\Theta(t)\right\} + \\ &\quad \mathbb{E}\left\{\sum_{u\in\mathcal{U}} Q_u(t)\left(R_u(t) - \mu_u(t)\right)|\Theta(t)\right\} + \\ &\quad \mathbb{E}\left\{\sum_{u\in\mathcal{U}} H_u(t)\left(\gamma_u(t) - R_u(t)\right)|\Theta(t)\right\} + \\ &\quad \mathbb{E}\left\{Z(t)\left(W\eta EE^{\text{req}}P_{\text{tot}}(t) - R_{\text{tot}}(t)\right)|\Theta(t)\right\} \end{aligned} \tag{33}$$

where $\Gamma = \frac{1}{2}\left(3\sum_{u\in\mathcal{U}}\left(A_u^{\max}\right)^2 + \sum_{u\in\mathcal{U}}\left(\mu_u^{\max}\right)^2 + \left(P_{\text{tot}}^{\max}(t)\right)^2 + \left(R_{\text{tot}}^{\max}(t)\right)^2\right)$*, and* $\mu_u^{\max}$ *denotes the maximum transmission rate that can be obtained on u.*

Proof Please refer to Appendix 1. □

4.4 Solving problem by decomposition

By observing the inequality in Theorem 1, we can decide to minimize the bound given in the right-hand side (R.H.S.) of (33) at every time slot for the optimization. This is more convenient than directly minimizing the drift-plus-penalty function itself because the minimization on R.H.S. could be decoupled to a series of independent subproblems that can be solved independently and simultaneously, as shown as follows.

4.4.1 Auxiliary variables

The first subproblem is to determine the optimal auxiliary variables γ_u, conducted to track the stability constraint shown in (27-C6). Specifically, the optimal γ_u can be resulted from minimizing $-\mathbb{E}\left\{\sum_{u\in\mathcal{U}}\left(V\overline{\phi(\gamma_u(t))} - H_u(t)\gamma_u(t)\right)|\Theta(t)\right\}$ that is obtained by slightly rearranging the relevant terms in the R.H.S of (33). Clearly, for the minimization, a concave nondecreasing system utility $\phi(\cdot)$ for each u should be given at first. Here, the well-known utility function $\log(1+\nu_u\gamma_u)$ is considered as an example wherein ν_u denotes a weight to maintain, e.g., the proportional fairness among UEs. Further, since the variables are independent among UEs, the minimization on $\gamma_u(t)$ can be decoupled from the joint optimization. Finally, by reversing the sign in the objective for minimization, we have an equivalent maximization problem as

$$\begin{aligned}&\underset{\gamma_u(t)}{\textbf{Maximize}}\ \ V\overline{\phi(\gamma_u(t))}-H_u(t)\gamma_u(t)\\&\textbf{subject to}\ \ 0\le\gamma_u(t)\le A_u^{\max},\qquad \forall u\in\mathcal{U}\end{aligned}\tag{34}$$

Obviously, it is a convex optimization problem. To find its optimum, we can first differentiate the objective function $V\overline{\phi(\gamma_u(t))}-H_u(t)\gamma_u(t)$ with respect to $\gamma_u(t)$ and then make the result equal to zero. For the log utility function just exemplified, we can solve the equation resulted to obtain its solution as

$$\gamma_u(t)=\begin{cases}0, & H_u(t)>\nu_u V\\ \frac{V}{H_u(t)}-\frac{1}{\nu_u}, & \frac{V}{A_u^{\max}+\frac{1}{\nu_u}}\le H_u(t)\le\nu_u V\\ A_u^{\max}, & H_u(t)<\frac{V}{A_u^{\max}+\frac{1}{\nu_u}}\end{cases}\tag{35}$$

4.4.2 Admission control

Recall that for the system stability, our algorithm can admit only $R_u(t)$ out of $A_u(t)$ arrivals to transmit. For the traffic admission control subproblem in hand, we can observe the second and third expectations in R.H.S. of (33) to minimize $\mathbb{E}\{\sum_{u\in\mathcal{U}}R_u(t)\ (Q_u(t)-H_u(t))\,|\Theta(t)\}$, which leads to the optimal traffic admission control at each TTI, as follows:

$$\begin{aligned}&\underset{R_u(t)}{\textbf{Minimize}}\ \ \textstyle\sum_{u\in\mathcal{U}}R_u(t)\ (Q_u(t)-H_u(t))\\&\textbf{subject to}\ \ 0\le R_u(t)\le A_u(t),\quad \forall u\in\mathcal{U}\end{aligned}\tag{36}$$

This is clearly a linear problem, and a simple threshold-based admission control strategy for this problem can be derived as

$$R_u(t)=\begin{cases}A_u(t), & H_u(t)>Q_u(t)\\ 0, & \text{otherwise}\end{cases}\tag{37}$$

As the threshold would imply, only when the virtual queue $H_u(t)$ is accumulated larger than the data queue $Q_u(t)$, the new arrival $A_u(t)$ can then be admitted; otherwise, they will be denied to ensure the data traffic stability. That is, with the simple threshold, the admission control will be conducted to reduce $H_u(t)$ to push $\gamma_u(t)$ toward $R_u(t)$ and increase the throughput $R_u(t)$ to improve the system utility simultaneously.

4.4.3 Resource allocation for energy efficient transmission

As the kernel issue of eJCREEP, how to concurrently determine the multiple kinds of resources at each TTI for EE transmission is a NP-hard combinatorial problem, in general, without special structures imposed. Here, with the aid of the drift-plus-penalty technique developed, such a high-dimensional allocation subproblem can be decomposed as minimizing $-\mathbb{E}\left\{\sum_{u\in\mathcal{U}}Q_u(t)\mu_u(t)+Z(t)\left(R_{\text{tot}}(t)-W\eta_{EE}^{\text{req}}P_{tot}(t)\right)|\Theta(t)\right\}$ without knowing the channel states in advance. Similarly, by negating the objective, we have an equivalent maximization problem as

$$\begin{aligned}&\underset{\underline{e}(t)}{\textbf{Maximize}}\ \ \nu(t)=\sum_{u\in\mathcal{U}}(\alpha_u(t)\mu_u(t)-\beta(t)P_u(t))\\&\textbf{subject to}\ \ (5),(7),(12)-(21)\end{aligned}\tag{38}$$

where $\alpha_u(t)=Q_u(t)+Z(t),\beta(t)=W\eta_{EE}^{\text{req}}Z(t)$, and $P_u(t)=\sum_{s\in\mathcal{S}}P_{s,u}^p+P_{s,u}^c$. As shown in Sections 3.4 and 3.5, the scheduling constraints are composed by the binary integer variables involved, and the combinatorial problem would be NP-hard, despite the optimization tools. Thus, instead of directly using an integer programming tool to solve this problem which would be still time-consuming when the inputs are not small enough, we design in the sequel a more computationally efficient algorithm based on the iterative linear programming-based heuristic (ILPH) to obtain a suboptimal solution that can be done within a time limit required.

5 Iterative optimal MCS-based heuristic algorithm

As shown in [15], iterative linear programming-based heuristic (ILPH) is a useful approach to resolve 0-1 integer programs, which is done by solving a series of small subproblems obtained from linear programming relaxations. Specifically, at each iteration, ILPH will conduct an LP-relaxation of the current problem P to generate one constraint. Then, a reduced problem induced from an optimal solution of the LP-relaxation is solved to obtain a feasible solution for the initial problem. After that, if the stopping criterion is satisfied, then the solutions found are returned. Otherwise, a pseudo-cut is added to P and the process is repeated.

In our work, the binary variable $\underline{e}$ for resource allocation is highly dimensional so that even solving a corresponding LP-relaxation problem could be time-consuming unless the input size is trivially small. Thus, a MCS-based reallocation approach is conducted here to reduce the overhead. For doing so, we define $J^0(\underline{e})=\{j\in(u,c,b,l,s,p):\underline{e}_j=0\},J^1(\underline{e})=\{j\in(u,c,b,l,s,p):\underline{e}_j=1\}$, and $J(\underline{e})=J^0(\underline{e})\cup J^1(\underline{e})$ similar to that in [15]. Then, an iterative optimal MCS-based heuristic algorithm (IOMHA) is introduced to restrict the search process to visiting the optimal solutions already generated from the time-limited optimization on P by adding a pseudo-cut at each iteration. As tabulated in Algorithm 1 with details, IOMHA

Algorithm 1 The Iterative Optimal-MCS-based Heuristic Algorithm (IOMHA)

```
1:  For the problem instance $P$ in (38), set $t = T_B$, initial $I = I_B$, and $\Omega = \emptyset$;
2:  while $t > 0$ and $I \geq I_B$ do
3:      Solve $P$ with the time limit $T_l$ to obtain a feasible (or an optimal) solution $\underline{e}^*$ with utility $\nu^*$;
4:      Set $J^1(\underline{e}^*) = \emptyset, J^0(\underline{e}^*) = \emptyset, \nu^o = \nu^*$, and $\Omega = \Omega \cup \nu^0$;
5:      for all $\underline{e}$ with different $u, c, s$ and $p$ do
6:          $l' \leftarrow$ value of $\hat{l}$ with the maximum value of $\sum_{\forall \underline{\hat{e}}=(u,c,\hat{b},\hat{l},s,p)} \mathbb{1}\left\{h(\underline{\hat{e}})|l_{\underline{\hat{e}}} = \hat{l}\right\}$
7:              among all $\hat{l} \in [\min_{\forall b \in \mathcal{B}} \Psi_{u,c,b,s,p}, \max_{\forall b \in \mathcal{B}} \Psi_{u,c,b,s,p}]$ if $\underline{e} == 1$;
8:          for all $b$ do
9:              $J^1(\underline{e}^*) = J^1(\underline{e}^*) \cup \underline{e}$ if $l_{\underline{e}} = l'$, and $J^0(\underline{e}^*) = J^0(\underline{e}^*) \cup \underline{e}$ if $l_{\underline{e}} \neq l'$;
10:     Generate the set of constraints $\{\hat{f}\underline{e} = C\}$, where $\hat{f}_j = 1, \forall j \in J$ while $C_j = 1$ if $j \in J^1$ and 0 if $j \in J^0$;
11:     Solve the problem $Q = (P|\{\hat{f}\underline{e} = C\})$ with the time limit $T_l$ to obtain a feasible (or an optimal) solution $\underline{\hat{e}}$ giving
    $\hat{\nu}$;
12:     Generate the cut $\{f\underline{e} \leq |J^1(\underline{e}^*)| - 1\}$;
13:     Update the Problem $P$ by adding the above constraint: $P = (P|\{f\underline{e} \leq |J^1(\underline{e}^*)| - 1\})$;
14:     Set $t = t - 2T_l, I = \frac{\hat{\nu}}{\nu^o}$, and $\Omega = \Omega \cup \hat{\nu}$;
15: Return $\underline{e} = \arg\max_{\underline{\hat{e}}}\{\hat{\nu} \in \Omega\}$;
```

first solves the maximization problem instance P in (38) to find a feasible solution $\underline{e}^*$ with utility ν^*. If the solution is not optimal, it might be improved by boosting the MCS of remaining RBs to find the largest MCS usable by all considered RBs [8]. However, instead of using the primitive method, IOMHA further attempts to make the utility contributed by the UE larger by releasing more RBs of the considered CC to render its remaining RBs able to employ an even higher-rate MCS. To this end, consider the utility $h(\underline{e}) = ((Q_u + Z)\nu(\underline{e}) - W\eta_{EE}^{\text{req}} ZP_u)$ without the time index t for brevity. Given that, if a UE u served by a BS s has some RB(s) of CC c^* at PL p re-allocated to UE u^*, we search the MCS l' that makes the largest the total UE utility contributed by all remaining RBs of c^* assigned to u among all maximum MCSs employable by these RBs (lines 5–7). Then, we reassign MCS l' to UE u on the transmission of CC c^* from BS s and release the allocations without any utility contribution, producing $J^1(\underline{e}^*)$ and $J^0(\underline{e}^*)$ (lines 8–9). The reassignment further forms a new set of constraints $\{\hat{f}\underline{e} = C\}$, where $\hat{f}_j = 1, \forall j \in J$ while $C_j = 1$ if $j \in J^1$ and 0 if $j \in J^0$ (line 10), and we solve the corresponding problem $Q = (P|\{\hat{f}x = C\})$ with the time limit T_l to obtain a feasible (or an optimal) solution $\underline{\hat{e}}$ giving utility $\hat{\nu}$ (line 11). If the improvement $I = \frac{\hat{\nu}}{\nu^o}$ does not exceed a given low bound I_B, the process would stop. Otherwise, based on Propositions 1 and 2 in [15], a pseudo-cut $\{f\underline{e} \leq |J^1(\underline{e}^*)| - 1\}$, where $f_j = 2\underline{e}_j^* - 1$ if $\underline{e}_j^* \in J(\underline{e}^*)$ and 0 otherwise, will be added when the remaining time $t = t - 2T_l$ allows, and the problem will be updated as $P = (P|\{f\underline{e} \leq |J^1(\underline{e}^*)| - 1\})$ that would be solved to seek further improvements (lines 12–14). Finally, the allocation result $\underline{e}$ corresponding to the best utility found during the searching process will be returned (line 15).

6 Performance bounds

As shown in above, IOMHA is an approximation algorithm to resolve the high-dimensional allocation subproblem involved. However, if the optimal solutions can be given, the overall algorithm for eJCREEP can operate under the performance bounds on, e.g., data queue lengths, as shown in the following theorem.

Theorem 2 *Given arbitrary traffic arrival rates and an energy efficiency requirement, the algorithm solving eJCREEP with a fixed control parameter $V \geq 0$ can guarantee the bounds on data queue lengths as*

$$Q_u(t) \leq Q_u^{\max} = \nu_u V + 2A_u^{\max} \tag{39}$$

Proof Please refer to Appendix 2. □

Apart from the above, the other performance bounds for the Lyapunov drift-plus-penalty framework can be also found in a similar way. For example, a drift-plus-penalty approach had been shown, e.g., in [32], to achieve $O(\epsilon)$ approximation with a convergence time of $O\left(1/\epsilon^2\right)$ with $\epsilon = 1/V$.

7 Results and discussion

7.1 Environment setting

In this section, we numerically evaluate our optimization algorithm through a simulation topology as shown in Fig. 2, wherein 1 MBS and 3 SBSs are deployed, and

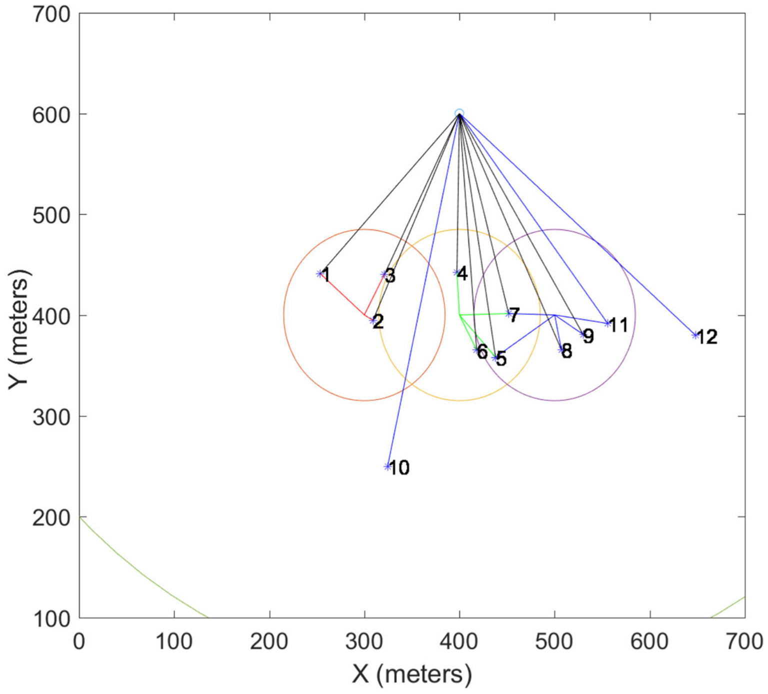

Fig. 2 Initial topology for the experiments

each of them initially serves 3 UEs located within its transmission range for their downlink transmissions before the resource allocation. In addition to **s** = 4 and **u** = 3 just indicated, there are also **c**=5, **b**=10, **l**=29, and **p**=3, other resources contributing to the overall complexity that is significant enough to evaluate the high-dimensional allocation problem involved. Further, each UE in cell is conducted to dynamically change its position according to the random waypoint (RWP) model [33], and the channel condition is assumed to be varied time to time on each RB as that in [34]. Given the time-varying environment, MBS is conducted to perform the proposed algorithm with $T_l = 1000, W = 1, \nu_u = 1$, along with the other key parameters summarized in Table 2. Based on the above setting, the performance results are resulted and summarized in the sequel.

7.2 Result analysis

To be specific, the performance metrics include time-average utility, throughput, data queue length, and energy efficiency (EE) denoted by $\overline{\phi}, \overline{\gamma}, \overline{Q}$, and $\overline{\eta_{EE}}$, respectively; each of them is represented by its mean value obtained from all UEs carrying out 100 times of the algorithm per experiment. Given these metrics, our algorithm is then conducted by varying V and η_{EE}^{req} to focus on the performance tradeoffs among throughput, data queue length (or delay), and energy efficiency (EE) in the experiments exemplifying the performance trends. To this end, $A_u(t), \forall u \in \mathcal{U}$ at each slot t is randomly

Table 2 Parameters in the experiments, with $P^t = 29$ dbm denoting the maximum transmit power $P_{\max}$

Parameter	Setting
Macro cell (MC) transmission radius	550 m
Macro cell (SC) path-loss factor (ρ)	4
Small cell (SC) transmission radius	85 m
Small cell path loss factor (ρ)	3
Mobility model	Random waypoint
UE speed (m/s) intervals	[0.2 2.2] for SC, [2 20] for MC
Number of user equipments (**u**)	12
Number of cells or BSs (**s**)	4
Number of carrier components (**c**)	5
Number of resource blocks per CC (**b**)	10
Maximum number of CCs for BS (f_s)	2
Maximum number of CCs for UE (k_u)	2
Available MCS (l)	29 MCSs in TS 36.213 [13]
Transmit powers of MBS	$\{0.3P^t, 0.5P^t, 1P^t\}$
Transmit powers of SBS	$\{0.05P^t, 0.1P^t, 0.2P^t\}$
Noise powers $\left(N_{s,u}^{c,b}\right)$	− 110 dbm
SNR-CQI index mapping	Refer to [27]
CQI-MCS index mapping	Refer to [27]
MCS index mapping to modulation and TBS index tables	Refer to TS 36.213 [13]

generated by the Poisson distribution with the mean value obtained by the maximum $TBS = 680$ multiplying with a given constant $C_1 = 14$ which represents a possible varying traffic arrival at time t under the maximum allowable rate $A_u^{\max} = TBS * C_2$, where $C_2 = 20$. Following that, the time-varying Rayleigh channel conditions are simulated by using the random channel gains obtained by the exponential distribution with the mean value of 1. Consequently, a wide range of V sampled at $\left[10^1, 10^3, 10^6, 10^7, 10^8, 10^9 10^{11} 10^{15}\right]$, and that of η_{EE}^{req} at $[1, 2, 4, 8]$ are combinatorially examined to know their impacts on the algorithm in general.

The experiment results are summarized in Fig. 3. Specifically, from Fig. 3a and b, we can see that as V increases, the utility and throughput improve significantly and converge to their maximum levels for larger V. This is expected because the achieved utility would increase to optimum at the speed of $O(1/V)$ when V increases, which implies a control emphasizing more on the throughput. However, as shown by the curves remaining nearly the same for large V, we can see also that the improvement will diminish with an excessive increment of V which may then aggravate the congestion as the data queue length would rise as V increases. In addition, it can be noted that as V increases, the system would more emphasize on the throughput utility as noted before, which could increase γ_u (with (35)) and then H_u (with (28)), leading to more arrivals to be admitted (with (37)) and eventually an increased data queue length (with (22)). Specifically, Fig. 3c exhibits that the increasing data queue length due to the increment of V would increase the average delay, and thus, the tradeoff between throughput and delay emerges, which well confirms Theorem 2.

On the other hand, the performance differences on EE obtained by different EE requirements versus different V are exhibited in Fig. 3d. To show its implication, we note that in the simulations, the EE value obtained from different V without any EE requirement is 3.58 on average,

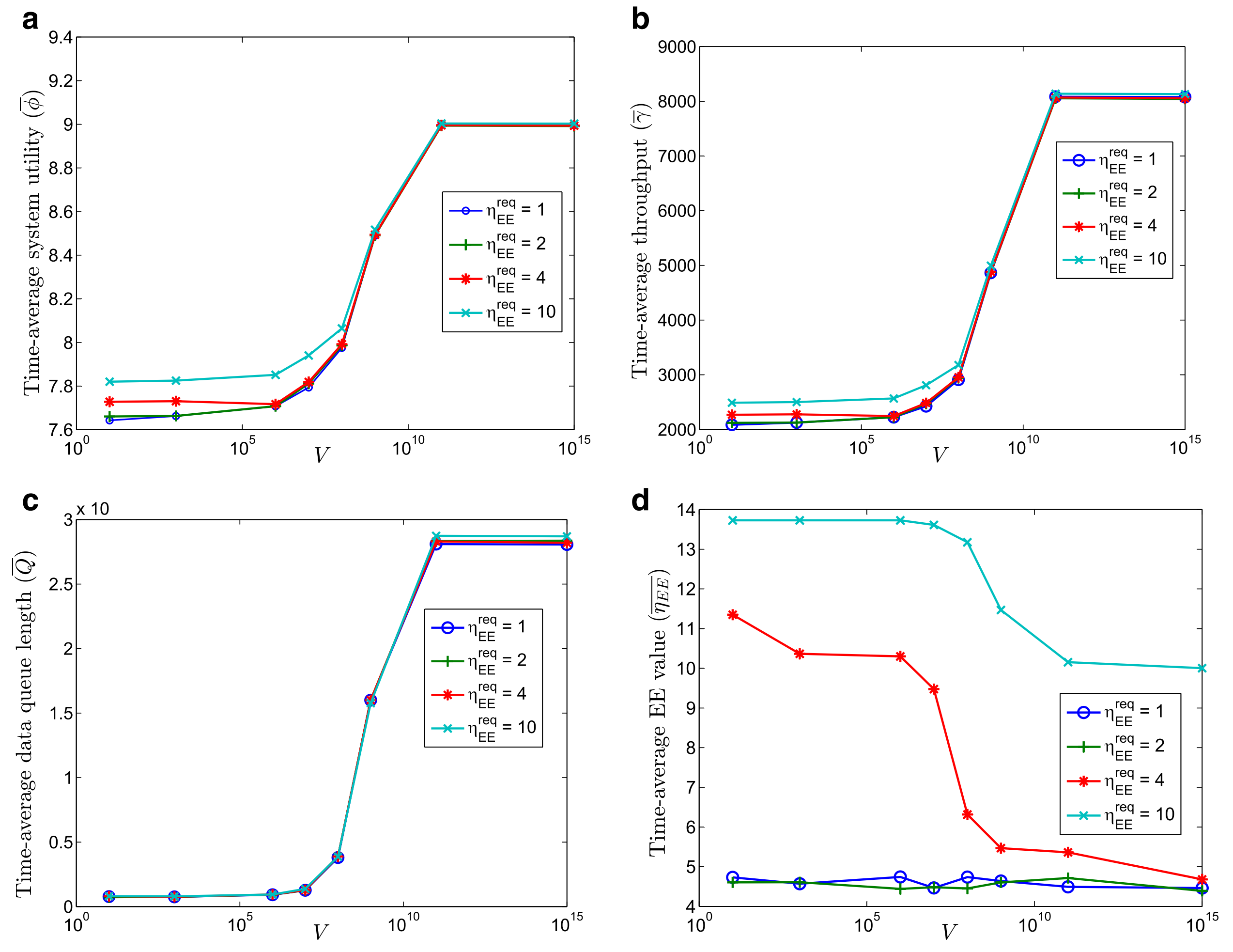

Fig. 3 Impacts of varying V upon the time-average **a** utility $\overline{\phi}$, **b** throughput $\overline{\gamma}$, **c** data queue length $\overline{Q}$, and **d** energy efficiency $\overline{\eta_{EE}}$

denoted here by *EE threshold*. Clearly, when $\eta_{EE}^{\rm req} = 1$ and 2 that are smaller than the threshold, the EE values actually obtained shown in this figure as well as the throughput-delay tradeoffs shown in the above are very similar, despite V. On the other hand, when $\eta_{EE}^{\rm req}$ increases to 4 and 10 that are larger than the threshold, the average throughput would increase especially when V is smaller (see Fig. 3b). This phenomenon can be explained by the aid of Fig. 4 that is obtained with $V = 10^1$. As shown therein (Fig. 4), to guarantee $\eta_{EE}^{\rm req}$, the network would decrease the transmit power level, and thus encourage the transmissions of small cells by allocating more RBs to SBSs that achieve a higher spectrum reuse gain, followed by the increment of the EE obtained and the average throughput. When V is smaller (such as $V = 10^1$ as exemplified), the EE performance gain obtained by a higher EE requirement $\left(\eta_{EE}^{\rm req}\right)$ is more significant. On the other hand, as V increases, the system would more emphasize on the throughput utility and pay less attention to EE, and hence, the EE gain would decrease and become less significant (see Fig. 3d). These results confirm that our algorithm actually represents a controllable method that can approach the optimal throughput while satisfying the EE requirement by simply manipulating the parameter V to achieve the performance tradeoff required by the system.

7.3 Performance comparison

As our IOMHA is conducted to concurrently allocate multiple types of resources in the multi-tier multi-cell networks, which can hardly correspond to an existing method in the related works that did not consider the resources: UEs, RBs, CCs, MCSs, cells, and PLs, and the EE constraint at the same time. However, to explore its performance benefits in eJCREEP, we extend the greedy algorithm in [8] (called Greedy), and the LL+RS algorithm introduced therein for comparison, to involve the multiple cells and discrete power levels in this question, resulting in more comparable methods for our work, when compared with the other algorithms possessing certain properties such as continuous power allocation that is hard to changed for the sake of comparison. As introduced in [8], in the first step, the LL+SS algorithm based on [5] will perform CC assignment with the concept of Least Load (LL) by which each UE is assigned the CCs with the least number of UEs. In the second step, it assigns RBs of each CC to UE by its packet scheduling function while resolving the MCS constraint in the scheduling. Given that, LL+SS as well as Greedy still did not consider allocating CCs to multiple cells and utilizing discrete PLs. To address this issue, we first allocate CCs to different cells with the objective of maximizing the sum of SNR values of CCs perceived by cells while complying with our multi-cell constraint that a UE can be only served by a single BS s and that each BS s can equip with at most f_s CCs, as the first level of the extension. After allocating CCs to each cell, Greedy and LL+SS can then be run to play the role of IOMHA in eJCREEP with discrete PLs, respectively, to solve the allocation problem in Section 4.4.3, as the second level of the extension.

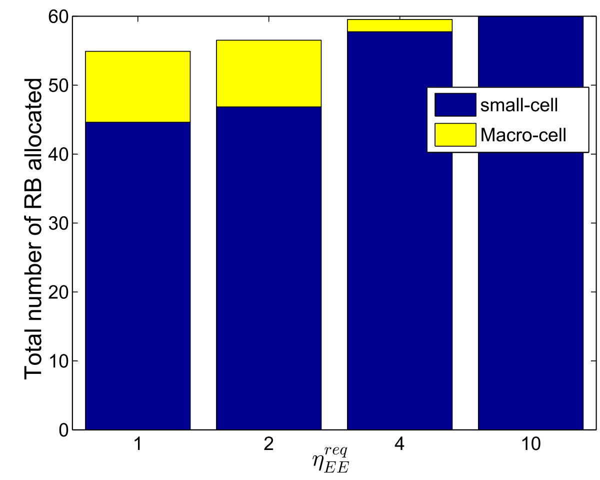

Fig. 4 RB allocation results with $V = 10^1$

In addition, for a more general condition to be encountered, we do not restrict ourselves to consider only the SNR values based on the distances and channel models in the previous set of experiments. Instead, we assume that SNR of each RB perceived by UE would be a random variable uniformally distributed in the range between -5 and 22.38 according to the SNR-CQI index mapping in [27], exemplifying an allocation that can involve all possible mapping values and their results in the simulation. In this case, we solve the allocation problem (38) with an optimization tool for the optimum without limiting its solving time while approaching the optimal result by using IOMHA with a reasonable time constraint represented by $T_B = 1000$ and $T_l = 500$ and obtaining the suboptimal solutions based on Greedy and LL+SS, respectively, to see their performance differences varied with different V. Specifically, in view of the results revealed in the previous experiment set, we use $V = \left\{10^1, 10^6, 10^{15}\right\}$ to exemplify a possible low/midle/high system parameter causing the performance tradeoff in the same spectrum of V from 10^1 to 10^{15} considered in Section 7.2, while fixing $\eta_{EE}^{\rm req} = 10$ and remaining the other parameters.

The comparison results are now summarized in Fig. 5. As shown in Fig. 5a, while complying with the performance trend shown in Section 7.2, IOMHA exhibits its throughput to approach the optimal value which is significantly higher than that to be achieved by Greedy and LL+SS despite V. This confirms the benefit of the joint optimization that can concurrently decide the CC

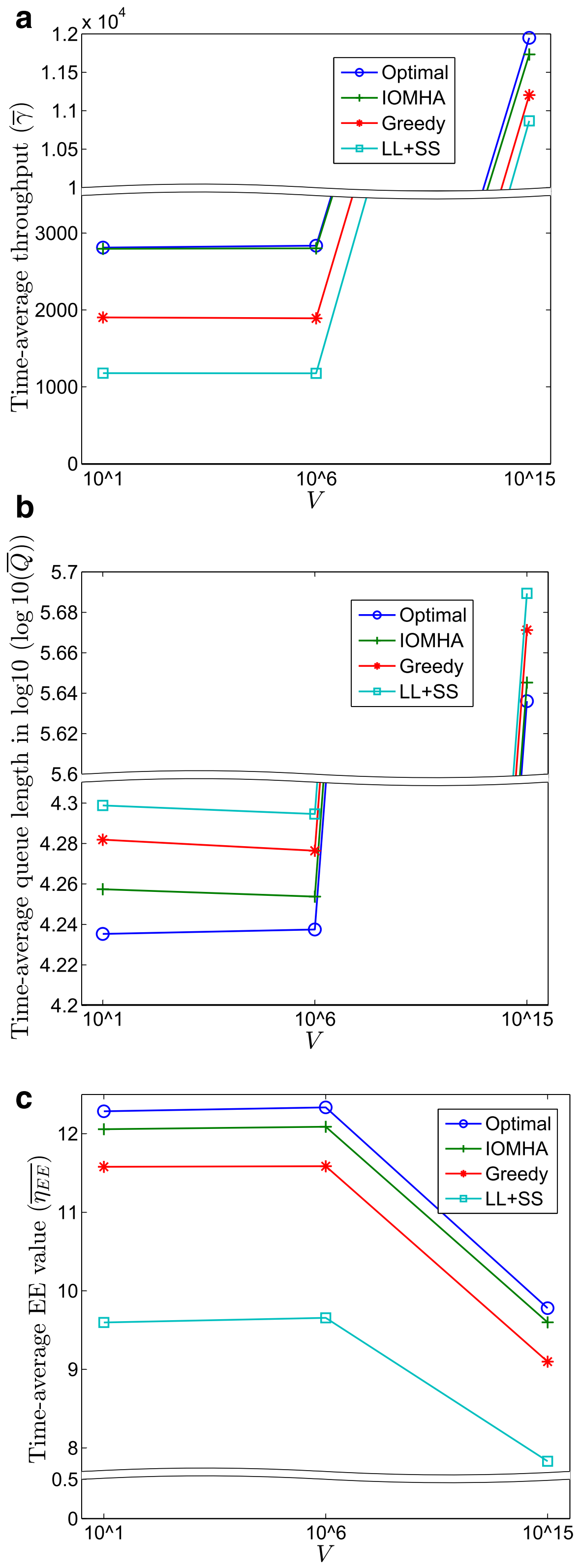

Fig. 5 Performance comparison on the time-average **a** throughput $\overline{\gamma}$, **b** data queue length $\overline{Q}$, and **c** energy efficiency $\overline{\eta_{EE}}$ that are optimally obtained and resulted from IOMHA, Greedy, and LL+SS, respectively

allocation to cells and the allocation of the RBs involved to UEs while complying with the MCS constraint and the other constraints. In contrast to the joint approach, the related works [5, 8] usually schedule RBs with or without the MCS constraint, based on the assumption of pre-allocated CCs. Here, without the joint optimization gain, Greedy is worse than IOMHA as a result, but it still outperforms LL+SS which is consistent with the observation shown in [8].

In Fig. 5b, the data queue length is shown in log10 magnitude to focus on the performance differences brought by the different methods in this metric. If applying a normal scale, the larger queue lengthes resulted from a high V (10^{15}) would be the focus of the figure, making the results from a lower V (10^1 or 10^6) insignificant even though the relative differences among them are all large enough despite V. In this representation, it is clearly shown that IOMHA yields a lower queue length than Greedy and LL+SS throughout the three V parameters, which also denotes a lower delay to be obtained by our method.

Finally, in Fig. 5c, the decreasing trend for the EE performance is exhibited to be the same as that observed from Fig. 3d. While all the methods under comparison have the same trend as expected, the EE performance resulted from IOMHA in eJCREEP is only slightly lower than the optimum, and Greedy has the result lower than ours but still outperforms LL+SS significantly. Taking all the tree metrics (throughput, queue length or delay, and EE) into account, it could be noted that using IOMHA with a proper time constraint to resolve the resource allocation problem and gradually improve the result would be a good method to trade the optimality for the eJCREEP that is NP optimization problem off against a lower and controllable complexity. That is, using IOMHA in eJCREEP would be better than simply adopting on-the-fly methods such as Greedy and LL+SS in this problem that can be done only once for a suboptimal solution to the complex allocation problem involved without a chance for further improvements.

8 Conclusions

In this work, we have addressed an optimization problem on the throughput utility while satisfying the EE requirement with time-varying channel condition and data traffic realized by the carrier aggregation technique in 5G heterogeneous wireless networks. For obtaining a practical solution, the high-dimensional NP-hard allocation problem involved was first formulated with a programming model involving nonlinear integer constraints, and then reformatted to be an equivalent problem involving only linear integer constraints. However, finding an optimal solution for the mixed integer programming model without special structures imposed would be still NP-hard and time-consuming, even with the linear form. For this challenge, an iterative optimal MCS-based heuristic algorithm (IOMHA) was proposed to approach the optimum within a limited period of time demanded by the user, in the low layer. Given that, a Lyapunov optimization framework was developed to resolve the problem in the high layer that can admit time-varying traffics without a priori knowledge of arrivals. Then, with the solutions from the two layers, we completed an approach that can make an optimal tradeoff with a system control parameter V and satisfy the long-term EE requirement simultaneously. Finally, the proposed framework was verified to reveal the performance tradeoffs among throughput, delay, and energy efficiency, showing that it can serve as an efficient way to address such a complex optimization problem, exhibiting the performance trends on the tradeoffs for the future works. In particular, as a resource allocation problem for nowadays stochastic networks becomes more challenging to meet fast convergence and tolerable delay requirement, a machine learning approach involving batch training could be developed as our future work while preserving the stochastic network optimization context guarantees queue stability with our Lyapunov drift-plus-penalty framework that can take the advantage of the iterative optimal MCS-based heuristic algorithm proposed to flexibly adjust its convergence time required by the system.

Appendix 1

Proof of Theorem 1:

By leveraging the fact that $A \geq 0, b \geq 0, Q \geq 0, (\max\{Q - b, 0\} + A)^2 \leq Q^2 + A^2 + b^2 + 2Q(A - b)$, we can square both sides of (22), (28), and (29), and sum the squares for (22) and (28) over all u, leading to

$$\sum_{u\in\mathcal{U}} \left(Q_u(t+1)^2 - Q_u(t)^2\right) \leq \sum_{u\in\mathcal{U}} (A_u)^2 + \sum_{u\in\mathcal{U}} (\mu_u)^2 + 2\sum_{u\in\mathcal{U}} Q_u(t)(R_u(t) - \mu_u(t)) \qquad (40)$$

$$\sum_{u\in\mathcal{U}} \left(H_u(t+1)^2 - H_u(t)^2\right) \leq 2\sum_{u\in\mathcal{U}} (A_u)^2 + 2\sum_{u\in\mathcal{U}} H_u(t)(\gamma_u(t) - R_u(t)) \qquad (41)$$

$$Z(t+1)^2 - Z(t)^2 \leq (P_{\text{tot}}(t))^2 + (R_{\text{tot}}(t))^2 + 2\left(W\eta_{EE}^{\text{req}} P_{\text{tot}}(t) - R_{\text{tot}}(t)\right) \qquad (42)$$

Let $A_u^{\max}$ and $\mu_u^{\max}$ be the upper bounds of $A_u(t)$ and $\mu_u(t), \forall t$, respectively. Further let $R_{tot}^{\max}(t)$ be $\sum_{\forall \underline{e}\in\Xi} \nu(\underline{e}(t))$, and $P_{\text{tot}}^{\max}(t)$ be $W\eta_{EE}^{\text{req}}\left(\sum_{\forall \underline{e}\in\Xi} \mathbf{I}(\underline{e})\left(P_{s,u}^{p}(t) + P_{s,u}^{c}\right)\right)$, where $\mathbf{I}(x) = 1, \forall x$. In addition, consider $R_u(t) \leq A_u^{\max}$ and $\gamma_u(t) \leq A_u^{\max}$. After taking these definitions and considerations into (40), (41), and (42), we can then combine the resulted equations and take the expectation with respect to $\Theta(t)$ on both sides of the combination, which eventually leads to the one-slot conditional Lyapunov drift as follows:

$$\Delta(\Theta(t)) \leq \Gamma + \mathbb{E}\left\{\sum_{u\in\mathcal{U}} Q_u(t)\left(R_u(t) - \mu_u(t)\right)|\Theta(t)\right\} + \mathbb{E}\left\{\sum_{u\in\mathcal{U}} H_u(t)\left(\gamma_u(t) - R_u(t)\right)|\Theta(t)\right\} + \mathbb{E}\left\{Z(t)\left(W\eta_{EE}^{\text{req}} P_{\text{tot}}(t) - R_{\text{tot}}(t)\right)|\Theta(t)\right\} \qquad (43)$$

where $\Gamma = \frac{1}{2}\left(3\sum_{u\in\mathcal{U}}\left(A_u^{\max}\right)^2 + \sum_{u\in\mathcal{U}}\left(\mu_u^{\max}\right)^2 + \left(P_{\text{tot}}^{\max}(t)\right)^2 + (R_{\text{tot}}^{\max}(t))^2\right)$. Finally, (33) is obtained by adding $-V\mathbb{E}\left\{\sum_{u\in\mathcal{U}} \overline{\phi(\gamma_u(t))}|\Theta(t)\right\}$ on both sides of (43).

Appendix 2

Proof of Theorem 2

For the performance bound, we would first show that $H_u^{\max} \triangleq \nu_u V + A_u^{\max}$, will be the upper bound of $H_u(t)$. It is done by induction, showing that if this bound holds at time slot t, it will be true at time $t+1$ also. More specifically, because $\gamma_u(t)$ can not exceed $A_u^{\max}$, the algorithm can increase $H_u(t)$ with an amount of at most $A_u^{\max}$ at slot t based on (37), and thus, if $H_u(t) \leq \nu_u V, H_u(t{+}1)$ will not exceed $\nu_u V + A_u^{\max}$. Otherwise, if $H_u(t) > \nu_u V, \gamma_u(t)$ will be 0 according to (35). In this case, $H_u(t)$ will not increase in $t+1$, and hence, $H_u(t+1) \leq H_u(t)$ which is bounded above by $H_u^{\max}$.

Next, we proceed to prove $Q_u(t)$ to be bounded with respect to $H_u^{\max}$ shown above, which can be also done by induction. First, this bound is assumed to be true at t. Given the induction hypothesis and the relationship $R_u(t) \leq A_u(t) \leq A_u^{\max}, Q_u$ will increase by at most $A_u^{\max}$ in one slot. Recall that $H_u^{\max} \triangleq \nu_u V + A_u^{\max}$ is the upper bound of $H_u(t)$. Then, if $Q_u(t) \leq H_u^{\max}$, then $Q_u(t+1)$ will not exceed $H_u^{\max} + A_u^{\max} = \left(\nu_u V + A_u^{\max}\right) + A_u^{\max} = \nu_u V + 2A_u^{\max}$, according to the data queueing dynamic (22) which increases $Q_u(t)$ by at most $R_u(t)$ while $R_u(t)$ can increase by at most $A_u^{\max}$ based on (37). Otherwise, if $Q_u(t) > H_u^{\max}$, then $R_u(t)$ will be 0 according to (37) as well. Finally, both cases confirm $Q_u^{\max} \triangleq H_u^{\max} + A_u^{\max} = \nu_u V + 2A_u^{\max}$ to be the bound shown in (39), and the proof is done.

Abbreviations

3GPP: 3rd Generation Partnership Project; 5G: Fifth generation; BS: Base station; CA: Carrier aggregation; CC: Component carrier; CQI: Channel quality indicator; EE: Energy efficiency; eJCREEP: Equivalent joint congestion control and resource allocation with EE-delay tradeoff problem; HetNet: Heterogeneous network; IOMHA: Iterative optimal-MCS-based heuristic algorithm; RWP: Random waypoint; JCREEP: Joint congestion control and resource allocation with EE-delay tradeoff problem; LTE: Long-Term Evolution; LTE-A: Long-Term Evolution-Advanced; MBS: Macro base station; MC: Macro cell; MCS: Modulation and coding scheme; NP: Non-deterministic polynomial time; OFDMA: Orthogonal frequency division multiple access; PL: Power level; RB:

Resource block; RRM: Radio resource management; SBS: Small base station; SC: Small-cell; SNR: Signalto-noise ratio; TTI: Transmission time interval; UE: User equipment

Authors' contributions

All authors contribute to the concept, the design, and developments of the algorithm and the simulation results in this manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Ministry of Science and Technology, Republic of China, under grant MOST 105-2221-E-126-001-MY3.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science and Information Engineering, Providence University, 43301 Taichung, Taiwan. [2]Department of Computer Science and Engineering, National Sun Yat-Sen University, 804 Kaohsiung, Taiwan.

Received: 16 February 2019 Accepted: 7 August 2019
Published online: 11 September 2019

References

1. E. Hossain, M. Hasan, 5g cellular: key enabling technologies and research challenges. IEEE Instrum. Meas. Mag. **18**, 11–21 (2015)
2. M. Deghani, K. Arshad, Lte-advanced radio access enhancements: A survey. Wirel. Pers. Commun. **80**(3) (2014)
3. FemtoForum, Femtocells Natural Solution for Offload. Tech. Rep. (2010). https://www.slideshare.net/wandalex/femtocells-a-natural-solution-for-offload
4. F. D. Ganni, A. Pratap, R. Misra, in *Proceedings of the 7th ACM International Workshop on Mobility, Interference, and MiddleWare Management in HetNets (MobiMWareHN'17)*. Distributed algorithm for resource allocation in downlink heterogeneous small cell networks (USA Article, New York, 2017), pp. 5–6
5. Y. Wang, K. I. Pedersen, T. B. Sorensen, P. E. Mogensen, Carrier load balancing and packet scheduling for multi-carrier systems. IEEE Trans. Wirel. Commun. **9**(5), 1780–1789 (2010)
6. Y. Wang, K. I. Pedersen, T. B. Sorensen, P. E. Mogensen, in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*. Utility Maximization in LTE-Advanced Systems with Carrier Aggregation, (2011), pp. 1–5. https://doi.org/10.1109/VETECS.2011.5956494
7. L. Zhang, K. Zheng, W. Wang, L. Huang, Performance analysis on carrier scheduling schemes in the long-term evolution-advanced system with carrier aggregation. IET Commun. **5**(5), 612–619 (2011)
8. H. S. Liao, P. Y. Chen, W. T. Chen, An efficient downlink radio resource allocation with carrier aggregation in lte-advanced networks. IEEE Trans. Mob. Comput. **13**(10), 2229–2239 (2014)
9. H. Zhang, H. Liu, J. Cheng, V. C. M. Leung, Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks. IEEE Trans. Commun. **66**(4), 1705–1716 (2018)
10. H. Beyranvand, M. Levesque, M. Maier, J. A. Salehi, C. Verikoukis, D. Tipper, Toward 5g: Fiwi enhanced lte-a hetnets with reliable low-latency fiber backhaul sharing and wifi offloading. IEEE/ACM Trans. Netw. **25**(2), 690–707 (2017)
11. H. Zhang, L. Venturino, N. Prasad, P. Li, S. Rangarajan, X. Wang, Weighted sum-rate maximization in multi-cell networks via coordinated scheduling and discrete power control. IEEE J. Sel. Areas Commun. **29**(6), 1214–1224 (2011)
12. J. Zheng, Y. Cai, Y. Liu, Y. Xu, B. D. and, and x. (sherman) shen, "optimal power allocation and user scheduling in multicell networks: Base station cooperation using a game-theoretic approach," in. IEEE Trans. Wirel. Commun. **13**(12), 6928–6942 (2014)
13. G. T. version 8.4.0 Release 8, Evolved universal terrestrial radio access (e-utra): Physical layer procedures. https://www.3gpp.org/ftp/Specs/archive/36_series/36.213/
14. A. L. Soyster, B. Lev, W. Slivka, Zero-one programming with many variables and few constraints. Eur. J. Oper. Res. **2**, 195–201 (1978)
15. S. Hanafi, C. Wilbaut, Improved convergent heuristics for the 0-1 multidimensional knapsack problem. Ann. Oper. Res. **183**(1), 125–142 (2011)
16. M. Dyer, L. Stougie, Computational complexity of stochastic programming problems. Math. Program. **106**(3), 423–432 (2006)
17. M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. (Morgan and Claypool Publishers, San Rafael, 2010)
18. D. W. K. Ng, E. S. Lo, R. Schober, Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity. IEEE Trans. Wirel. Commun. **11**(10), 3618–3631 (2012)
19. H. Lee, S. Vahid, K. Moessner, A survey of radio resource management for spectrum aggregation in LTE-advanced. IEEE Commun. Surv. Tutor. **16**(2), 745–760 (2014)
20. F. Wu, Y. Mao, S. Leng, X. Huang, in *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, ed. by N. S. W. Sydney. A Carrier Aggregation Based Resource Allocation Scheme for Pervasive Wireless Networks, (2011), pp. 196–201. https://doi.org/10.1109/DASC.2011.54
21. H. Mahdavi-Doost, N. Prasad, S. Rangarajan, in *2016 8th International Conference on Communication Systems and Networks (COMSNETS)*. Energy efficient downlink scheduling in LTE-Advanced networks, (Bangalore, 2016), pp. 1–8. https://doi.org/10.1109/COMSNETS.2016.7439928
22. D. Lopez-Perez, X. Chu, A. V. Vasilakos, H. Claussen, On distributed and coordinated resource allocation for interference mitigation in self-organizing LTE networks. IEEE/ACM Trans. Netw. **21**(4), 1145–1158 (2013)
23. J.-s. Liu, Joint downlink resource allocation in LTE-advanced heterogeneous networks. Comput. Netw. **146**, 85–103 (2018)
24. E. Universal, 3GPP TS 36.211 3GPP TSG RAN enolved universal terrisrial radio access (E-UTRA) physical channels and modulation, version 11.0.0, release 11, 2012, 3GPP. 3rd Generation Partnership Project(3GPP) (2012)
25. 3rd Generation Partnership Project, 3GPP TS 36.213 v10.6.0, Evolved Universal Terrestrial Radio Access (E-UTRA). Physical layer procedure (2012)
26. D. T. Ngo, T. Le-Ngoc, *Architectures of Small-cell Networks and Interference Management (1st Ed.)* (Springer, New York, 2014)
27. S. S. A. Tiwari, *LONG TERM EVOLUTION (LTE) PROTOCOL Verification of MAC Scheduling Algorithms in NetSim*. (Tetcos White Paper, Tetcos, 2014)
28. M. T. Kawser, N. I. B. Hamid, M. N. Hasan, M. S. Alam, M. M. Rahman, Downlink snr to cqi mapping for different multiple antenna techniques in lte. Int. J. Inf. Electron. Eng. **2**(5), 756–760 (2012)
29. W. L. Winston, M. Venkataramanan, *Introduction To Mathematical Programming*. (Duxbury Resource Center, Belmont CA, 2002)
30. Y. Li, M. Sheng, Y. Shi, X. Ma, W. Jiao, Energy efficiency and delay tradeoff for time-varying and interference-free wireless networks. IEEE Trans. Wirel. Commun. **13**(11), 5921–5931 (2014)
31. M. Sheng, Y. Li, X. Wang, J. Li, Y. Shi, Energy efficiency and delay tradeoff in device-to-device communications underlaying cellular networks. IEEE J. Sel. Areas Commun. **34**, 92–106 (2016)
32. M. J. Neely, A simple convergence time analysis of drift-plus-penalty for stochastic optimization and convex program. arXiv:1412.0791v1 [math.OC], 1–10 (2014)
33. D. B. Johnson, D. A. Maltz, in *Mobile Computing. The Kluwer International Series in Engineering and Computer Science*, ed. by T. Imielinski, H. F. Korth. Dynamic Source Routing in Ad Hoc Wireless Networks, vol. 353 (Springer, Boston, 1996)
34. H. Liao, P. Chen, W. Chen, An efficient downlink radio resource allocation with carrier aggregation in LTE-advanced networks. IEEE Trans. Mob. Comput. **13**(10), 2229–2239 (2014)

Publisher's Note