This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

Cross-Layer Rate Control and Resource Allocation in Spectrum-Sharing OFDMA Small Cell Networks with Delay Constraints

Yashuang Guo, Qinghai Yang, Jiayi Liu and Kyung Sup Kwak

***Abstract*—In this paper, we present a dynamic resource management scheme for delay-aware applications in two-tier small cell networks (SCNs). We propose the scheme of joint rate control at the transport layer and resource allocation at the physical layer to manage the cross-tier interference. The joint rate control and resource allocation scheme is designed to maximize the time-averaged sum capacity of small cell users in the SCN subject to each small cell user's delay constraint and an interference constraint imposed by the macrocell. By using Lyapunov optimization technique, we develop a delay-guaranteed capacity optimal algorithm (DCOA) to obtain the optimal rate control and resource allocation decisions. We show that without prior knowledge of the data arrivals and channel statistics, DCOA achieves a capacity of SCN that can arbitrarily approach the optimal capacity achieved by the algorithm with the complete knowledge of data arrivals and channel statistics. Simulations results confirm the theoretical analysis on the performance of DCOA and also show the adaptiveness of DCOA.**

***Index Terms*—Small cells networks, cross-tier interference, delay constraints, cross-layer.**

I. Introduction

Small cell networks (SCNs), which are composed of low-power supplied nodes (such as micro-, pico- and femtocells) have been introduced as a novel networking paradigm based on the idea of deploying short-range, low-power, and low-cost base stations underlaying the macro-cellular network to handle the unprecedent traffic demand in next-generation wireless networks [1–4]. Orthogonal frequency division multiple access (OFDMA) based SCNs have been considered in major wireless communication standards, e.g., LTE/LTE- Advanced [5].

Due to spectrum scarcity and implementation difficulty, spectrum-sharing, rather than spectrum splitting, between SCNs and macrocells is preferable from the operators perspective [6]. However, cross-tier interference imposed from SCNs to macrocell could be severe in spectrum-sharing based two-tier networks [7]. Meanwhile, many real-life applications (such as video streaming and online gaming) are delay-sensitive, which have stringent requirements on delay. In fact, the primary objective of deploying SCNs is to improve wireless networks' capacity while satisfying the satisfactory QoS (e.g., delay) performance for small cell users (SUEs) [8]. Therefore, resource management schemes that could satisfy SUEs' delay requirements with consideration of cross-tier interference is needed in SCNs [9].

Power allocation has been widely used for maximizing the capacity of SCNs whilst alleviating the cross-tier interference in two-tier networks. Power control was utilized in [11] to ensure the adequate signal-to-interference-plus-noise ratio (SINR) for SUEs. A Lagrangian dual decomposition based power allocation scheme was proposed with cross-tier interference mitigation in [12], on the other hand, channel allocation was applied to suppress the cross-tier interference. A hybrid frequency assignment scheme was proposed for SCN deployed within the coverage of a macrocell in [13]. Subchannel allocation in SCN was performed via a correlated equilibrium game-theoretic approach for minimizing the interference to primary macro base station (MBS) in [14]. Moreover, joint power and subchannel allocation algorithm was proposed for maximizing the total capacity of densely deployed SCN in [15].

As a common feature, most of the existing works [11–15] only focused on the physical layer performance (e.g. capacity), while disregarding the bursty data arrivals and the delay requirement of SUEs. The authors in [16] considered the cross-tier interference-aware resource management with bursty data arrivals. However, since their work focused on maximizing the SCNs' capacity, the delay performance was ignored. Recently, Li et al [17] proposed a delay-aware resource allocation algorithm based on Markov decision process for minimizing the sum of average delay of all users. Since the authors focused on the sum of average delay of all users, how to perform resource allocation whilst providing explicit delay guarantee for an individual user was ignored. A cross-layer scheduling algorithm was developed in [18] for maximizing the time-average throughput of single cell OFDMA networks subject to user's delay constraint. However, without consideration of the cross-tier interference, their works cannot be directly applied to spectrum-sharing SCNs. Therefore, it is worth studying how to perform the resource management in spectrum-sharing SCNs while taking into account both SUE's delay and cross-tier interference constraints.

In addition, interference mitigation in spectrum underlay

This research was supported in part by NSF China (61471287), MSIP, Korea, under ITRC Program (IITP-2015-H8501-15-1019), and 111 Project (B08038).

Y. Guo, Q. Yang, and J. Liu are are with State Key Laboratory of ISN, School of the Telecomm. Engineering, and also with the Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, No.2 Taibainan-lu, Xi'an, 710071, Shaanxi, China.

K. S. Kwak is with the Department of Information and Communication Engineering, Inha University, #100 Inha-Ro, Nam-gu, Incheon, 22212, Korea.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

systems is also a crucial issue considered in cognitive radio (CR) networks [19]. Interference temperature limit is introduced in CR networks to constrain the interference from a secondary network to a primary network, which has priority for utilizing the same spectrum [20]. Interference suppression based on resource allocation strategies has also been studied in CR networks. In literature [21], a dual decomposition method based subchannel selection and power allocation, subject to interference temperature limit, was studied in CR networks. Joint subchannel and power allocation for maximizing the system capacity considering the interference temperature limit on each subchannel of active primary users for multi-cell CR networks was investigated in [22, 23]. However, the interference temperature cannot be directly applied in SCNs [26], because of the absence of cognitive capabilities for SUEs. To solve this problem, the interference temperature limit can be delivered to SCN by backhaul from the MBS [24–26].

In this paper, we focus on the delay-guaranteed resource management for two-tier SCNs, in which a central macrocell is overlaid with spectrum-sharing small cells. The main contributions of this work are summarized below:

- We introduce a cross-tier interference temperature limit to protect MUEs from severe cross-tier interference. We utilize rate control (RC) at the transport-layer and interference-aware resource allocation (RA) (e.g., joint power allocation and subchannel assignment) at the physical layer to manage the cross-tier interference.
- We employ stochastic optimization model to maximize the long-term time-averaged capacity of SCNs subject to each SUE's delay constraint, minimum data rate constraint, and the interference temperature limit constraint.
- We develop a delay-guaranteed capacity optimal algorithm (DCOA) to obtain the optimal RC and RA decisions without prior knowledge of the data arrivals and channel statistics. Particularly, both of the the RC and RA in DCOA have closed-form solutions.

The remainder of this paper is organized as follows. Section II provides an overview of the system model followed by the problem formulation in Section III. We present the DCOA in Section IV. The performance of the proposed DCOA is analyzed in Section V. Simulation results are presented in Section VI, and Section VII concludes our paper.

II. System Model

As shown in Fig. 1, we consider the downlink of a two-tier OFDMA-based SCN, where K co-channel small-cell base stations (SBSs) are overlaid on the coverage of a macrocell. Let B and U denote the numbers of active MUEs in the macrocell and SUEs in each small cell, respectively. All small cells are assumed to be closed access, i.e., SCNs only provide services to the pre-registered SUEs [39]. Specifically, this paper considers the sparse deployment scenario of SCN as in [25–27], e.g., the scenario of suburban area, where the co-tier interference between neighboring SBSs is negligible compared with the cross-tier interference.

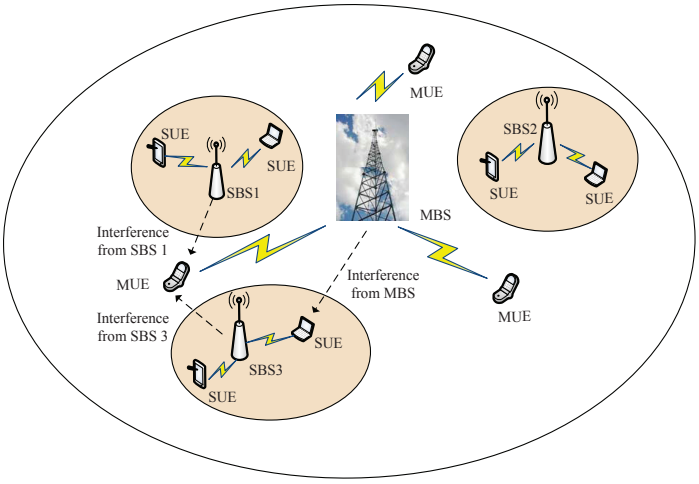


Fig. 1. System Model.

A. Traffic Model and Rate Control

The SCN operates in a time-slotted manner with time slot index $t \in \{0, 1, 2, \cdots\}$. There is a busty source data that arrives randomly at each SBS every time slot destined to an SUE. The SBS maintains a buffer for each SUE to temporally store the arrived data before the data is transmitted to its corresponding SUE. Let $A_{k,u}(t)$ denote the amount of arrived data for SUE u in small cell k at time slot t. We assume that $A_{k,u}(t)$ is independently and identically distributed (i.i.d)[1] over time slots with a long-term time-averaged arrival rate $\lambda_{k,u}$, namely $\mathbb{E}\{A_{k,u}(t)\} = \lambda_{k,u}$. In the heavy-load case, where the downlink system cannot support all SUEs' arrived data under continuous poor channel conditions, there is a need to adjust the admission rate, i.e., the amount of data from transport layer admitted to the data queue (or buffer) at SBS to avoid blocking [16, 37]. Denote $R_{k,u}(t)$ as the admission rate of the data queue for SUE u in small cell k. Then we have the following constraint on the RC decision $R_{k,u}(t)$

$$R_{k,u}(t) \leq A_{k,u}(t) \leq A_{max}, \tag{1}$$

where A_{max} is a constant upper bound on SUE's data arrival. Evidently, the maximum amount of data admitted by an SUE cannot exceed the amount of arrived data at each time slot. RC decisions on how much data to be admitted to the SCN are taken by each SBS according to a certain policy which would be specified in Section IV.

The RC constraint in (1) indicates that part of the arrived data will be rejected if the heavy load case occurs in practical SCN. We assume that there are no transport layer buffers and thus the rejected data will be perceived as invalid and be dropped immediately as performed in [16, 28, 37]. Actually, RC is reasonable in practice since real-life applications such as video

[1]Here, the i.i.d assumptions imposed on SUE's traffic and channel condition are only to guarantee the existence of a rate control policy in **Lemma 2**, but they are not crucial for the performance of the algorithm [29] (Page 72). The proposed algorithm can also be effective for arbitrary (possible non-i.i.d) channels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

TABLE I
SUMMARY OF KEY NOTATIONS

Notation	Meaning
K, N	Number of small cells, and number of subchannels
U	Number of SUEs in each small cell
k, u, n	Indices of small cell, SUE, and subchannel
B	Number of MUEs in the macrocell
F	Bandwidth of the SCN
$h^S_{k,u,n}(t)$	Channel gain on subchannel n from SBS k to SUE u at time slot t
$h^M_{k,u,n}(t)$	Channel gain on subchannel n from MBS to SUE u in small cell k at time slot t
$h^S_{k,b,n}(t)$	Channel gain on subchannel n from SBS k to MUE b at time slot t
$p^S_{k,u,n}(t)$	Transmit power allocated on subchannel n to SUE u in small cell k at time slot t
$p^M_{b,n}(t)$	Transmit power of MBS on subchannel n to the assigned MUE b
$w_{k,u,n}(t)$	Binary decision variable on whether to allocate subchannel n to SUE k in small cell k at time slot t
$I^{th}_n(t)$	Maximum tolerable interference level on subchannel n suffered by MUEs at time slot t
$\gamma^S_{k,u,n}(t)$	Received SINR of SUE u on subchannel n in small cell k at time slot t
σ^2	Power spectral density of noise
$C_{k,u,n}(t)$	Transmission rate on subchannel n of SUE u in small cell k at time slot t
$Q_{k,u}(t)$	Data queue length for SUE u in small cell k at time slot t
$A_{k,u}(t)$	Bits of data that arrived at SBS k for SUE u at time slot t
$R_{k,u}(t)$	Admission rate of SUE u in small cell k
$C_{k,u}(t)$	Transmission rate of SUE u in small cell k at time slot t
$X_{k,u}(t), Y_{k,u}(t), Z_{k,u}(t)$	Queue length of virtual queues X, Y, and Z for SUE u in small cell k at time slot t
$\nu_{k,u}(t)$	Virtual admission rate of SUE u in small cell k at time slot t
$O_{k,u}$	Average rate requirement of SUE u in small cell k
$D_{k,u}$	Delay threshold of SUE u in small cell k
q_{max}	Buffer size of each SUE in each SBS
V	Lyapunov control parameter
$\mathbf{G}(t)$	System state at time slot t

streaming can tolerate some packets loss, but has strict delay constraint so that occasional packet loss is allowed [18].

B. Downlink Physical Layer Model and Resource Allocation

The downlink RA (e.g., power allocation and subchannel assignment) takes place in each SBS with the assistance of MBS. Specifically, at the beginning of each time slot t, each SBS observes the current queue state information (QSI) at the transmission queues. Meanwhile, we assume that the channel state information (CSI) is reported to each SBS via the feedback channel without any delay and error.

Exclusive subchannel allocation: The OFDMA system has a bandwidth of F, which is divided equally into N subchannels. The subchannel set is denoted as $\mathcal{N} = \{1, \cdots, N\}$. Each subchannel $n \in \mathcal{N}$ can be allocated to at most one SUE in each small cell at each time slot t to avoid the intra-cell-interference. Let $w_{k,u,n}(t) \in \{0, 1\}$ be the subchannel assignment index for the SCN, where $w_{k,u,n}(t) = 1$ denotes the n-th subcarrier is assigned to SUE u in small cell k at time slot t. Otherwise, $w_{k,u,n}(t) = 0$. Thus, we have

$$\sum_{u=1}^{U} w_{k,u,n}(t) \leq 1, \forall k, n, t. \tag{2}$$

SBS's power constraint: Let $p^S_{k,u,n}(t)$ be the transmit power allocated on subchannel n to SUE u in small cell k at time slot t, and P_{max} be the peak transmit power of SBS. Each SBS's transmit power is also limited

$$\sum_{n=1}^{N}\sum_{u=1}^{U} w_{k,u,n}(t) p^S_{k,u,n}(t) \leq P_{max}. \tag{3}$$

Cross-tier interference constraint: We impose an interference temperature limit to constrain the cross-tier interference suffered by the MUE. Let $I^{th}_n(t)$ denote the maximum tolerable interference level on subchannel n for the assigned MUE b at time slot t, we have

$$\sum_{k=1}^{K}\sum_{u=1}^{U} w_{k,u,n}(t) p^S_{k,u,n}(t) h^S_{k,b,n}(t) \leq I^{th}_n(t), \forall n, t, \tag{4}$$

where $h^S_{k,b,n}(t)$ is the interference channel gain on subchannel n from SBS k to MUE b served by the MBS at time slot t. The interference constraint means that SBSs are only permitted to transmit signals on the same subchannel with MBS as long as the total interference is kept under a tolerable level.

Let $h^S_{k,u,n}(t)$ and $h^M_{k,u,n}(t)$ denote the channel gain on subchannel n from SBS k to SUE u in small cell k at time slot t and the interference channel gain from MBS to SUE u in small cell k at time slot t, respectively. Let $p^M_{b,n}(t)$ denote the transmit power of MBS on subchannel n to MUE b at time slot t. Then, the received SINR of SUE u in small cell k occupying the subchannel n is given by:

$$\gamma^S_{k,u,n}(t) = \frac{p^S_{k,u,n}(t)|h^S_{k,u,n}(t)|^2}{p^M_{b,n}(t)h^M_{k,u,n}(t) + \sigma^2}, \tag{5}$$

where $p^M_{b,n}(t)h^M_{k,u,n}(t)$ is the interference caused by the macrocell on subchannel n, and σ^2 is the power of additive white Gaussian noise (AWGN) per subchannel. In the paper, $h_{k,u,n}(t)$ is assumed to be i.i.d over time slots, and takes values in a finite state space. Furthermore, $h_{k,u,n}(t)$ keeps constant during one time slot, but potentially changes from one time slot to another.

Based on Shannon's capacity, the transmission rate on subchannel n of SUE u in small cell k at time slot t is given by

$$C_{k,u,n}(t) = w_{k,u,n}(t)\log_2(1 + \gamma^S_{k,u,n}(t)). \tag{6}$$

C. Queueing Model and System Dynamics

Let $Q_{k,u}(t)$ be the data backlog of the queue at SBS k for SUE u at time slot t. Given the RC and RA decisions, the data queues for SUEs evolve over time as follows

$$Q_{k,u}(t+1) = [Q_{k,u}(t) - C_{k,u}(t)]^+ + R_{k,u}(t), \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

where $C_{k,u}(t)$ and $R_{k,u}(t)$ are the service rate and the input rate of queue $Q_{k,u}$ at time slot t, respectively, and $[x]^+ = \max(x, 0)$. Let $R(t)$ and $C(t)$ be the input rate and service rate at time slot t for queue $Q(t)$, respectively. For discrete time process $Q(t)$ evolves as following

$$Q(t+1) = [Q(t) - C(t)]^+ + R(t), \tag{8}$$

$Q(t)$ is defined as *strongly stable* [29] if :

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{Q(t)\} < \infty. \tag{9}$$

A multi-queue network is strongly stable if all the individual queues are strongly stable. According to the *strong stability Theorem* in [29], for finite variables $R(t)$ and $C(t)$, strong stability implies the rate stability of $Q(t)$. The definition of rate stability can be found in [29] and omitted here. According to the *Rate Stability Theorem* in [29], The discrete queue $Q(t)$ is rate stable if and only if the time-average service rate c satisfies

$$c \geq r, \tag{10}$$

where $c = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} C(t)$ and $r = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} R(t)$. In this paper, we assume that $R_{k,u}(t)$ and $C_{k,u}(t)$ are both finite $\forall k, u$. The assumption is reasonable since all physical quantities such as the admission rates, the transmission rates, and the transmit power are all bounded in real SCN. Thus, if a queue is strongly stable in the SCNs, the time-averaged service rate c satisfies $c \geq r$.

D. Rate and Delay Constraints

Since many delay-aware applications, such as video and online gaming, typically require an upper bound on delay and an lower bound on rate [38], we impose both a time-averaged rate and a time-averaged delay constraints for each user (or the associated application).

1) Rate Constraint: The rate constraint is expressed as

$$r_{k,u} \geq O_{k,u} \tag{11}$$

where $O_{k,u}$ is the rate requirement of SUE u in small cell k [19, 20].

2) Delay Constraint: The queuing delay is defined as the time length that a packet waits in a queue until it can be transmitted. The delay constraint is expressed as

$$\rho_{k,u} \leq D_{k,u}, \tag{12}$$

where $D_{k,u}$ is the upper bound of the time-averaged delay of SUE u in small cell k.

III. Problem Formulation

The objective of this paper is to maximize the capacity of the SCN while satisfying each SUE's delay and rate constraints as well as the cross-tier interference constraint. Meanwhile, the stability of the network as well as the resource allocation constraints of the small cells must also be satisfied. Hence, the problem is formulated by maximizing the admission rates of all SUEs:

$$\max_{\mathbf{R},\mathbf{W},\mathbf{P}} \quad \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U} \mathbb{E}\left\{R_{k,u}(t)\right\} \tag{13}$$

$$\text{s.t.} \quad w_{k,u,n}(t) \in \{0,1\}, \forall k, u, n, t, \tag{C1}$$

$$\sum_{u=1}^{U} w_{k,u,n}(t) \leq 1, \forall k, n, t, \tag{C2}$$

$$0 \leq \sum_{n=1}^{N}\sum_{u=1}^{U} w_{k,u,n}(t) p^S_{k,u,n}(t) \leq P_{max}, \forall k, t, \tag{C3}$$

$$\sum_{k=1}^{K}\sum_{u=1}^{U} w_{k,u,n}(t) p^S_{k,u,n}(t) h^M_{k,b,n}(t) \leq I^{th}_n(t), \forall n, t, \tag{C4}$$

$$0 \leq R_{k,u}(t) \leq A_{k,u}(t), \forall k, u, t, \tag{C5}$$

$$\text{Queues } Q_{k,u}(t) \text{ are strongly stable}, \forall k, u, \tag{C6}$$

$$r_{k,u} \geq O_{k,u} \forall k, u, \tag{C7}$$

$$\rho_{k,u} \leq D_{k,u}, \forall k, u, \tag{C8}$$

where $\mathbf{R}(t) = \{R_{k,u}(t)\}$, $\mathbf{W}(t) = \{w_{k,u,n}(t)\}$, and $\mathbf{P}(t) = \{p^S_{k,u,n}(t)\}$ are the admission rate, subcarrier assignment and power allocation matrices of the network, respectively. C1 and C2 are the subchannel allocation constraints employed to ensure that a subchannel can at most be occupied by one user at each time slot. C3 is SBS's peak transmit power constraint. C4 is the cross-tier interference constraint to prevent the MUE from intolerable interferences from SBSs. C5 is the RC constraint to guarantee the amount of admission data at each time slot is no greater than the amount of arrived data. C6 is the network stability constraint. C7 and C8 are SUE's rate and delay constraints.[2]

Theoretically, we can find the optimal solution to the problem (13) if we have the statistical knowledge of CSI and data arrivals in advance by methods such as dynamic programming [46]. But these methods are computationally complex and suffer from the curse of dimensionality. Moreover, it will be highly costly to get channel statistics in practical scenarios. Thus, in the paper, we resort to Lyapunov optimization technique, since the algorithms developed from Lyapunov optimization technique do not need the prior knowledge and have low computational complexity.

IV. Online Algorithm

In this section, we shall design the delay-guaranteed capacity optimal algorithm (DCOA) in detail. Before we introduce the design, it is worth noticing that the original problem (13) has a long-term average limitations on queuing delay and rate. In order to model the average delay and rate constraints, we introduce the concept of virtual queue [29, 30]. The virtual queue $Y(t)$ associated with the average rate constraint evolves

[2]Throughout this paper, we assume that the minimum rate vector, i.e., $\mathbf{O} = \{O_{k,u}\}$, is inside of the capacity region of the SCN.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

as follows

$$Y_{k,u}(t+1) = [Y_{k,u}(t) - \nu_{k,u}(t)]^+ + O_{k,u}. \quad (14)$$

And the virtual queue $Z(t)$ associated with the delay constraint evolves as follows

$$Z_{k,u}(t+1) = [Z_{k,u}(t) - D_{k,u}\nu_{k,u}(t)]^+ + Q_{k,u}(t). \quad (15)$$

In addition, we consider that there is a finite buffer size, denoted by q_{max} for each SUE at each SBS. To ensure the worst case SUE data queue length, virtual queue $X_{k,u}(t)$ is introduced to assist in developing our algorithm, which would guarantee that the actual queue $Q_{k,u}(t)$ is bounded deterministically in the worst case

$$X_{k,u}(t+1) = [X_{k,u}(t) - R_{k,u}(t)]^+ + \nu_{k,u}(t), \quad (16)$$

where $\nu_{k,u}(t)$ is the virtual admission rate of queue $Q_{k,u}(t)$, which is upper bounded by $A_{k,u}(t)$. Note that all of the virtual queues $X(t)$, $Y(t)$, and $Z(t)$ do not stand for any real queues or data. They are only generated by the proposed algorithm.

For data queue $Q_{k,u}(t)$, define $c_{k,u} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} C_{k,u}(t)$ and $r_{k,u} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} R_{k,u}(t)$. According to the *Rate Stability Theorem* in (10), if $Q_{k,u}(t)$ is strongly stable, then the time-averaged service rate $c_{k,u} \geq r_{k,u}$. Similarly, for virtual queue $X_{k,u}(t)$, define $\phi_{k,u} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \nu_{k,u}(t)$. Then if virtual queue $Y_{k,u}(t)$ is strongly stable, the time-averaged virtual admission rate $\nu_{k,u}(t)$ satisfies

$$\phi_{k,u} \geq O_{k,u}. \quad (17)$$

And if $Z_{k,u}(t)$ is strongly stable, we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Q_{k,u}(t) \leq D_{k,u}\phi_{k,u}. \quad (18)$$

Moreover, if $X_{k,u}(t)$ is stable, we have

$$r_{k,u} \geq \phi_{k,u}, \quad (19)$$

$$\rho_{k,u} = \frac{1}{r_{k,u}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Q_{k,u}(t) \leq D_{k,u}. \quad (20)$$

By *Little's Theorem*, time-averaged delay $\rho_{k,u}$ is approximated [45] by[3]

$$\rho_{k,u} = \frac{\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{Q_{k,u}(t)\}}{\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_{k,u}(t)\}}.$$

Thus, (20) ensures that the average delay of queue $Q_{k,u}(t)$ is less than or equal to the threshold $D_{k,u}$ with probability one, which means C8 is satisfied.

[3]Note that we consider a heavy-loaded SCN. Propagation delays are assumed to be negligible compared to queueing delays and thus are omitted [45].

A. Problem Transformation

Remark 1: From the above analysis and according to the *(Rate stability Theorem)*, if the data queues and the three virtual queues (X, Y, and Z) are stable for all SUEs, we know that the network is stable (i.e., date queues at all SBSs are stable) and the delay constraint and rate constraints are both satisfied. Therefore, we can transform the original problem in (13) into a problem of maximizing the capacity of the SCN subject to the queue stability constraints along with C1, C2, C3, C4 and C5. The transformed problem is formulated as follows

$$\max_{\mathbf{R},\mathbf{W},\mathbf{P}} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=1}^{K} \sum_{u=1}^{U} \mathbb{E}\{R_{k,u}(t)\} \quad (21)$$

$$\text{s.t.} \quad \text{C1, C2, C3, C4 , and C5,}$$
$$\text{Queues } Q_{k,u}(t), X_{k,u}(t), Y_{k,u}(t) \text{ and } Z_{k,u}(t)$$
$$\text{are stable, } \forall k, u.$$

Let $\mathbf{Q} = \{Q_{k,u}(t)\}, \mathbf{X} = \{X_{k,u}(t)\}, \mathbf{Y} = \{Y_{k,u}(t)\}$, and $\mathbf{Z} = \{Z_{k,u}(t)\}$ denote the queue backlogs of Q, X, Y and Z, respectively. Let $\mathbf{G}(t) = [\mathbf{Q}(t), \mathbf{X}(t), \mathbf{Y}(t), \mathbf{Z}(t)]$ denote the concatenated queue backlog of the SCN. Define the following Lyapunov function[4]

$$\begin{aligned} L(\mathbf{G}(t)) =& \frac{1}{2} \sum_{k=1}^{K} \sum_{u=1}^{U} \frac{1}{q_{max}} X_{k,u}(t) Q_{k,u}^2(t) \\ &+ \frac{1}{2} \sum_{k=1}^{K} \sum_{u=1}^{U} \frac{q_{max} - A_{max}}{q_{max}} X_{k,u}^2(t) \\ &+ \frac{1}{2} \sum_{k=1}^{K} \sum_{u=1}^{U} Y_{k,u}^2(t) + \frac{1}{2} \sum_{k=1}^{K} \sum_{u=1}^{U} Z_{k,u}^2(t). \end{aligned} \quad (22)$$

Without loss of generality, we assume that all queues are empty when $t = 0$ such that $L(\mathbf{G}(0)) = 0$. Define the one-slot conditional Lyapunov drift $\Delta(\mathbf{G}(t))$ as follows

$$\Delta(\mathbf{G}(t)) = \mathbb{E}\{L(\mathbf{G}(t+1)) - L(\mathbf{G}(t)) | \mathbf{G}(t)\}. \quad (23)$$

Subtracting from (23) the conditional expectation of $\nu(t) = \sum_{k=1}^{K} \sum_{u=1}^{U} \nu_{k,u}(t)$, we obtain the following drift-minus-reward term:

$$\Delta(\mathbf{G}(t)) - V\mathbb{E}\{\nu(t)|\mathbf{G}(t)\}, \quad (24)$$

where V is a nonnegative tunable parameter. It will be later shown in **Theorem 3** that when V is sufficiently large, DCOA approaches the optimal capacity. According to the design principle of Lyapunov optimization [29, 37], the RC and RA decisions should be chosen to minimize an upper bound of (24) at each time slot t.

[4]Please note that we employ the Lyapunov function in (22) instead of the traditional quadratic Lyapunov function [30] $\frac{1}{2}\left[Q^2(t) + X^2(t) + Y^2(t) + Z^2(t)\right]$ since the Lyapunov function in (22) is not only important in guaranteeing the finite buffer size constraint q_{max}, but also important in guaranteeing the whole network stability (including both the data queue and the virtual queues).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

Theorem 1: (Upper Bound of the Drift-Minus-Reward Term) Suppose $h_{k,u,n}(t)$ is i.i.d over time slots. Under any control algorithms, the *drift-minus-reward term* [29, 37] is upper bounded by [5]:

$$\begin{aligned}
&\Delta(\mathbf{G}(t)) - V\mathbb{E}\{\nu(t)|\mathbf{G}(t)\} \\
&\le B + \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\frac{R_{k,u}^2(t)+C_{k,u}^2(t)}{2q_{max}}X_{k,u}(t)|\mathbf{G}(t)\right\} \\
&+ \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{R_{k,u}(t)\frac{X_{k,u}(t)}{q_{max}}[Q_{k,u}(t)-(q_{max}-A_{max})]|\mathbf{G}(t)\right\} \\
&+ \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\nu_{k,u}(t)\left(\frac{q_{max}-A_{max}}{q_{max}}X_{k,u}(t)-Y_{k,u}(t)\right)|\mathbf{G}(t)\right\} \\
&- \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\nu_{k,u}(t)\left(D_{k,u}Z_{k,u}(t)+V\right)|\mathbf{G}(t)\right\} \\
&- \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{C_{k,u}(t)\frac{X_{k,u}(t)Q_{k,u}(t)}{q_{max}}|\mathbf{G}(t)\right\} \\
&+ \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{Y_{k,u}(t)O_{k,u}+Z_{k,u}(t)Q_{k,u}(t)|\mathbf{G}(t)\right\}, \qquad (25)
\end{aligned}$$

where B is a positive constant, which satisfies the following inequality for all t:

$$\begin{aligned}
&B \ge B(t) \\
&= \sum_{k=1}^{K}\sum_{u=1}^{U}\left\{\frac{q_{max}A_{max}}{2}+\frac{q_{max}-A_{max}}{2q_{max}}\left[R_{k,u}^2(t)+\nu_{k,u}^2(t)\right]\right\} \\
&+ \sum_{k=1}^{K}\sum_{u=1}^{U}\left\{\frac{D_{k,u}^2\nu_{k,u}^2}{2}+\frac{1}{2}q_{max}^2+\frac{1}{2}\left[\nu_{k,u}^2(t)+O_{k,u}^2\right]\right\}.
\end{aligned}$$

Proof: Please refer to Appendix A for the proof. ■

By **Theorem 1**, we have transformed the problem in (21) into minimizing the Right-Hand Side (R.H.S) of (25) at each time slot t subject to the RC constraint C5 and the RA constraints C1, C1, C3, and C4. Thus, the original stochastic optimization problem in (13) is transformed into a series of successive instantaneous static optimization problems.

B. Algorithm Design

Algorithm 1 describes the pseudo-code of DCOA. At each slot t, the algorithm performs the following four control operations: (1) RC in each SBS, which decides the admission rate for each SUE; (2) Virtual Rate Control in each SBS, which decides the virtual admission rate for each SUE; (3) RA in each SBS, which decides the subcarrier assignment and power allocation; and (4) Queues updating for $\{Q_{k,u}(t)\}, \{X_{k,u}(t)\}, \{Y_{k,u}(t)\}$, and $\{Z_{k,u}(t)\}$.

[5]Please note than the second term on the R.H.S of (25) is not evaluated in Algorithm 1 since ignoring this term in Algorithm 1 can significantly reduce the difficulty of solving the RA subproblem and maintaining the network stability from the perspective of math. But, this term can be later covered (in the proof from (54) to (55)) if the constraints in (47) are satisfied.

Algorithm 1 Delay-Guaranteed Capacity Optimal Algorithm DCOA

At each time slot t, observe the queue states $\mathbf{Q}(t), \mathbf{X}(t), \mathbf{Y}(t), \mathbf{Z}(t)$, and the channel condition $\mathbf{H}(t) = \{h_{k,u,n}^S(t)\}$

Step 1: RC
Compute $R_{k,u}(t)$ according to (27)

Step 2: Virtual Rate Control
Compute $\nu_{k,u}(t)$ according to (29)

Step 3: RA
Compute $w_{k,u,n}(t)$ and $p_{k,u,n}^S(t)$ according to **Algorithm 2**

Step 4: Update the queues

1) Rate Control: Observe that the third term on the R.H.S of (25) only involves with the RC decision $R_{k,u}(t)$. Since there are no coupled constraints between the $K \times U$ RC decisions, we can decompose the minimization of this term into $K \times U$ subproblems as follows:

$$\begin{aligned}
\min \quad & R_{k,u}(t)\frac{X_{k,u}(t)}{q_{max}}(Q_{k,u}(t)-q_{max}+A_{max}) \\
\text{s.t.} \quad & \text{C5}.
\end{aligned} \qquad (26)$$

The corresponding solution to (26) is

$$R_{k,u}(t) = \begin{cases} 0, & \text{if} \quad Q_{k,u}(t)-q_{max}+A_{max} > 0, \\ A_{k,u}(t), & \text{otherwise}. \end{cases} \qquad (27)$$

Here, we can have an intuitive explanation on the RC rules. They work like valves. When an actual queue exceeds $q_{max} - A_{max}$, the corresponding valve would turn off and no data would be admitted.

As to virtual variable $\nu_{k,u}(t)$, there is also its respective virtual rate control algorithm in each SBS so as to update the virtual queues $X_{k,u}(t), Y_{k,u}(t)$ and $Z_{k,u}(t)$. Observe that the fourth and fifth terms on the R.H.S of (25) only involve with the virtual rate control decision $\nu_{k,u}(t)$. We can decompose the minimization of the two terms into $K \times U$ subproblems as follows

$$\begin{aligned}
&\min_{0\le\nu_{k,u}(t)\le A_{k,u}(t)} \nu_{k,u}(t) \\
&\times \left(\frac{q_{max}-A_{max}}{q_{max}}X_{k,u}(t)-Y_{k,u}(t)-D_{k,u}Z_{k,u}(t)-V\right).
\end{aligned} \qquad (28)$$

The solution to (28) is

$$\nu_{k,u}(t) = \begin{cases} 0, & \text{if } \frac{q_{max}-A_{max}}{q_{max}}X_{k,u}(t) \\ & -Y_{k,u}(t)-D_{k,u}Z_{k,u}-V > 0, \\ A_{k,u}(t), & \text{otherwise}. \end{cases} \qquad (29)$$

2) Resource Allocation: Observe that the sixth term on the R.H.S of (25) only involves with the RA decisions $p_{k,u,n}(t)$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

$w_{k,u,n}(t)$. We reformulate the sixth term as follows:

$$\max_{\{w_{k,u,n}(t)\},\{p^S_{k,u,n}(t)\}} \sum_{k=1}^{K}\sum_{u=1}^{U}\sum_{n=1}^{N} \frac{Q_{k,u}(t)X_{k,u}(t)}{q_{max}} C_{k,u,n}(t)$$
$$\text{s.t.} \quad \text{C1, C2, C3, and C4.} \tag{30}$$

The optimization problem in (30) is a non-convex mixed integer programming problem because of the integer constraint for subchannel allocation in C1. The optimal solution of (30) under the constraints of C1, C2, C3, and C4 can be obtained by a Brute-force method, which has a high computational complexity. To make the problem tractable, we relax $w_{k,u,n}(t)$ to be a continuous real variable in the range [0,1], where $w_{k,u,n}(t)$ can be considered as a time-sharing factor for subchannel n. The time-sharing relaxation was widely used to transform non-convex combinatorial optimization problems into convex optimization problems in multichannel OFDMA systems [6, 26, 31–33].

Denote the actual power allocated on subchannel n to SUE u in small cell k at time slot t as $\widetilde{p}^S_{k,u,n}(t) = w_{k,u,n}(t)p^S_{k,u,n}(t)$. Now, the problem in (30) can be converted into:

$$\max_{\{\widetilde{p}^S_{k,u,n}(t)\},\{\widetilde{w}_{k,u,n}(t)\}} \sum_{k=1}^{K}\sum_{u=1}^{U}\sum_{n=1}^{N} \frac{Q_{k,u}(t)X_{k,u}(t)}{q_{max}} \widetilde{w}_{k,u,n}(t) \log_2\left(1+\frac{\widetilde{p}^S_{k,u,n}(t)|h^S_{k,u,n}(t)|^2}{\widetilde{w}_{k,u,n}(t)*\left(p^M_{b,n}(t)h^M_{k,u,n}(t)+\sigma^2\right)}\right) \tag{31}$$
$$\text{s.t.} \quad \widetilde{w}_{k,u,n}(t) \in [0,1], \forall k,u,n,t, \tag{C1}$$
$$\sum_{u=1}^{U} \widetilde{w}_{k,u,n}(t) \le 1, \forall k,n,t, \tag{C2}$$
$$\sum_{n=1}^{N}\sum_{u=1}^{U} \widetilde{p}^S_{k,u,n}(t) \le P_{max}, \forall k,t, \tag{C3}$$
$$\sum_{k=1}^{K}\sum_{u=1}^{U} \widetilde{p}^S_{k,u,n}(t)h^S_{k,b,n}(t) \le I^{th}_n(t). \tag{C4}$$

As the inequality constraints in (31) are convex, and the objective function is jointly concave with respective to $\widetilde{p}^S_{k,u,n}(t)$ and $\widetilde{w}_{k,u,n}(t)$, the optimization problem in (31) is concave. Being a concave optimization problem, the transformed optimization problem in (31) has a unique optimal solution, that is, the local solution is the optimal solution, which can be obtained in polynomial time.

In this subsection, the subchannel assignment and power allocation optimization in (31) is solved by using Lagrangian dual decomposition method. The Lagrangian function is given by

$$\begin{aligned} &L\left(\{\widetilde{w}_{k,u,n}(t)\},\{\widetilde{p}^S_{k,u,n}(t)\},\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\zeta}\right) \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{u=1}^{U} \frac{Q_{k,u}(t)X_{k,u}(t)}{q_{max}} \widetilde{w}_{k,u,n}(t) \\ &\times \log_2\left(1+\frac{\widetilde{p}^S_{k,u,n}(t)|h^S_{k,u,n}(t)|^2}{\widetilde{w}_{k,u,n}(t)*\left(p^M_{b,n}(t)h^M_{k,u,n}(t)+\sigma^2\right)}\right) \\ &- \sum_{k=1}^{K}\lambda_k\left(\sum_{u=1}^{U}\sum_{n=1}^{N}\widetilde{p}^S_{k,u,n}(t) - P_{max}\right) \\ &- \sum_{n=1}^{N}\eta_n\left(\sum_{k=1}^{K}\sum_{u=1}^{U}\widetilde{p}^S_{k,u,n}(t)h^S_{k,b,n}(t) - I^{th}_n(t)\right) \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K}\zeta_{k,n}\left(1-\sum_{u=1}^{U}\widetilde{w}_{k,u,n}(t)\right), \end{aligned} \tag{32}$$

where $\boldsymbol{\lambda} = \{\lambda_k\}$, $\boldsymbol{\eta} = \{\eta_n\}$, and $\boldsymbol{\zeta} = \{\zeta_{k,n}\}$ are the Lagrange multipliers (also called dual variables) matrix (or vector) for the constraints C3, C4, and C2 in (31), respectively. Thus, the Lagrangian dual function is defined as:

$$g(\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\zeta}) = \max_{\{\widetilde{w}_{k,u,n}(t)\},\{\widetilde{p}^S_{k,u,n}(t)\}} L\left(\{\widetilde{w}_{k,u,n}(t)\},\{\widetilde{p}^S_{k,u,n}(t)\},\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\zeta}\right). \tag{33}$$

The dual problem can be expressed as

$$\min_{\boldsymbol{\lambda},\boldsymbol{\eta}} \quad g(\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\zeta}) \tag{34}$$
$$\text{s.t.} \quad \boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\zeta} \succeq 0.$$

We decompose the Lagrangian dual function in (33) into a master problem together with $K \times N$ subproblems. The dual problem can be solved iteratively with each SBS solving the corresponding local subproblem in each iteration using local information. Accordingly, the Lagrangian function in (33) is rewritten as

$$\begin{aligned} &L\left(\{\widetilde{w}_{k,u,n}(t)\},\{\widetilde{p}^S_{k,u,n}(t)\},\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\zeta}\right) \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N} L_{k,n}\left(\{\widetilde{w}_{k,u,n}(t)\},\{\widetilde{p}_{k,u,n}(t)\},\lambda_k,\eta_n,\zeta_{k,n}\right) \\ &+ \sum_{k=1}^{K}\lambda_k P_{max} + \sum_{n=1}^{N}\eta_n I^{th}_n(t), \end{aligned} \tag{35}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

where

$$\begin{aligned}
&L_{k,n}(t)\left(\{\widetilde{w}_{k,u,n}(t)\},\{\widetilde{p}^{S}_{k,u,n}(t)\},\lambda_k,\eta_n,\zeta_{k,n}\right)\\
&=\sum_{u=1}^{U}\frac{Q_{k,u}(t)X_{k,u}(t)}{q_{max}}\widetilde{w}_{k,u,n}(t)\\
&\log_2\left(1+\frac{\widetilde{p}^{S}_{k,u,n}(t)|h^{S}_{k,u,n}(t)|^2}{\widetilde{w}_{k,u,n}(t)*\left(p^{M}_{b,n}(t)h^{M}_{k,u,n}(t)+\sigma^2\right)}\right)\\
&-\sum_{u=1}^{U}\lambda_k\widetilde{p}^{S}_{k,u,n}(t)-\sum_{u=1}^{U}\eta_n\widetilde{p}^{S}_{k,u,n}(t)h^{S}_{k,b,n}(t)\\
&+\zeta_{k,n}\left(1-\sum_{u=1}^{U}\widetilde{w}_{k,u,n}(t)\right).
\end{aligned}\tag{36}$$

Then, taking the partial derivative of $L_{k,u,n}(t)$ with respect to $\widetilde{p}^{S}_{k,u,n}(t)$ yields

$$\begin{aligned}
&\frac{\partial L_{k,n}(t)-\tau_{k,u,n}\widetilde{p}_{k,u,n}(t)}{\partial\widetilde{p}_{k,u,n}(t)}=\frac{\widetilde{w}_{k,u,n}(t)Q_{k,u}(t)X_{k,u}(t)}{q_{max}\ln 2}\\
&\times\left(\frac{|h^{S}_{k,u,n}(t)|^2}{\widetilde{w}_{k,u,n}(t)I_{k,u,n}(t)+\widetilde{P}^{S}_{k,u,n}(t)|h^{S}_{k,u,n}(t)|^2}\right)\\
&-\lambda_k-\eta_n h^{S}_{k,b,n}(t)-\tau_{k,u,n},
\end{aligned}\tag{37}$$

where $I_{k,u,n}(t)=p^{M}_{b,n}(t)h^{M}_{k,u,n}(t)+\sigma^2$, and $\tau_{k,u,n}$ is the Lagrange multiple associated with the implicit constraint that $\widetilde{p}_{k,u,n}(t)\geq 0$ in each subproblem.

According to the Karush-Kuhn-Tucker (KKT) conditions [44], the optimal power allocation of the subproblems, denoted by $\widetilde{p}^{S*}_{k,u,n}(t)$, must satisfy the following constraints:

$$\begin{aligned}
&\frac{\partial L_{k,n}(t)-\tau_{k,u,n}\widetilde{p}_{k,u,n}(t)}{\partial\widetilde{p}_{k,u,n}(t)}=0,\\
&\tau_{k,u,n}\widetilde{p}^{S}_{k,u,n}(t)=0,\\
&\widetilde{p}_{k,u,n}(t)\geq 0,\\
&\tau_{k,u,n}\geq 0.
\end{aligned}$$

Thus, by making the partial derivation in (37) (equal) to be zero, $\widetilde{p}^{S*}_{k,u,n}(t)$ is expressed as following:

$$\begin{aligned}
&\widetilde{p}^{S*}_{k,u,n}(t)=\\
&\left[\frac{Q_{k,u}(t)X_{k,u}(t)}{q_{max}\ln 2}\left(\frac{1}{\lambda_k+\eta_n h^{S}_{k,b,n}(t)}\right)-\frac{I_{k,u,n}(t)}{|h^{S}_{k,u,n}(t)|^2}\right]^{+}\\
&\times\widetilde{w}^{S}_{k,u,n}(t),\forall k,u,n,t.
\end{aligned}\tag{38}$$

Remark 2: *(Structure of the Power allocation and Subcarrier Assignment)* The power allocation in (38) is a function of CSI and QSI. It has the form of *multilevel water-filling*, where the power allocation is adaptive to both QSI and CSI. In addition, from (38) we can know that the water level is influenced by the production of $Q_{k,u}(t)X_{k,u}(t)$, $h^{S}_{k,b,n}(t)$ as well as $\frac{I_{k,u,n}(t)}{|h^{S}_{k,u,n}(t)|^2}$. That is, more cross-tier interference channel gain $h^{S}_{k,b,n}(t)$ results in lower water level and reduces the interference suffered by MUE. SUE with higher $\frac{|h^{S}_{k,u,n}(t)|^2}{I_{k,u,n}(t)}$ will be allocated more power on channel n. Moreover, SUE with higher value of the production of $Q_{k,u}(t)X_{k,u}(t)$ will be allocated more power to balance SUEs' delay requirements.

As to the Lagrange multiplier $\{\lambda_k\},\{\eta_n\}$, we use subgradient method to update it as shown in (39) and (40), respectively.

$$\lambda_k^{i+1}=[\lambda_k^{i}-\theta_1^{i}(P_{max}-\sum_{u=1}^{U}\sum_{n=1}^{N}\mathbf{p}^{S*}_{k,u,n}(t))]^{+},\forall k,t,\tag{39}$$

$$\eta_n^{i+1}=[\eta_n^{i}-\theta_2^{i}(I_n^{th}-\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbf{p}^{S*}_{k,u,n}(t))h^{S}_{k,b,n}(t)]^{+},\forall n,t,\tag{40}$$

where index i stands for the iteration number, and $\theta_1^{i},\theta_2^{i}$ are the step sizes of iteration i. I_{max} is the maximum number of iterations. When the subgradient method converges, RA is finished.

Then we will make use of the results of power allocation for subcarrier assignment. Observing that (36) can be decomposed into U independent subproblems. Each subproblem is formulated as following:

$$\begin{aligned}
&L_{k,n}(\mathbf{P})=\sum_{u=1}^{U}L_{k,u,n}(\mathbf{P}),\\
&L_{k,u,n}(\mathbf{P})=\frac{Q_{k,u}(t)X_{k,u}(t)}{q_{max}}\log_2\left(1+\frac{\widetilde{p}^{S*}_{k,u,n}(t)|h^{S}_{k,u,n}(t)|^2}{p^{M}_{b,n}(t)h^{M}_{k,u,n}(t)+\sigma^2}\right)\\
&-\lambda_{k,u}\widetilde{p}^{S*}_{k,u,n}(t)-\eta_n\widetilde{p}^{S*}_{k,u,n}(t)h^{S}_{k,b,n}(t)-\zeta_{k,n}\widetilde{w}_{k,u,n}(t).
\end{aligned}\tag{41}$$

Substituting (38) into (41), the objective of subcarrier assignment is to maximize $L_{k,n}(\mathbf{P})$ for all SUEs in small cell k. For any subcarrier n, it will be assigned to the SUE who has the biggest $L_{k,u,n}(\mathbf{P})$. Let n_u^{*} be the result of subcarrier n's assignment, which is given by:

$$n_u^{*}=\arg\max_{u}L_{k,u,n},\text{and},L_{k,n_u^{*},n}>0.\tag{42}$$

Thus, the optimal subchannel assignment decision, $w_{k,u,n}(t)$, is expressed as following:

$$w^{*}_{k,u,n}(t)=\begin{cases}1, & \text{if}\quad u=n_u^{*},\\ 0, & \text{otherwise.}\end{cases}\tag{43}$$

For sufficiently large N, a joint power allocation and subcarrier assignment strategy that assigns every subchannel to the user with the largest $L_{k,u,n}$ will result in negligible performance loss relative to the optimum OFDMA solution. This relaxation technique was also used in [6, 26, 33].

Although the above equations (38)–(43) give a solution for the joint power allocation and subchannel allocation problem, it still remains to design an algorithm to indicate the execution structure and the executing entity for the equations. Therefore, we propose **Algorithm 2**, which gives the procedures of the implementation.

Denote the iteration number at step 9 In **Algorithm 2** by j. In **Algorithm 2**, each SBS update $\boldsymbol{\lambda}$ with convergence factor ε_1 means that each SBS update $\widetilde{p}^{S}_{k,u,n}(t)$ until $|\lambda_k^{j+1}-\lambda_k^{j}|\leq\varepsilon_1$. In addition, $h^{S}_{k,b,n}(t)$ required in (38), (40) and (41) for the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

Algorithm 2 Distributed Lagrange Duality Optimization

1: Initialize I_{max} and Lagrange variables vectors $\boldsymbol{\lambda}, \boldsymbol{\eta}$, set iteration number $i = 1$, and convergence factor ε_1, ε_2
2: **Repeat**
3: **for** $k = 1$ to K **do**
4: **for** $n = 1$ to N **do**
5: **for** $u = 1$ to U **do**
6: each SBS computes $\widetilde{p}^{S}_{k,u,n}(t)$ according to (38)
7: each SBS computes calculate $L_{k,u,n}(t)$ according to (41)
8: each SBS computes $w_{k,u,n}(t)$ according to (43)
9: each SBS updates $\boldsymbol{\lambda}$ with convergence factor ε_1 according to (39)
10: **end for**
11: **end for**
12: **end for**
13: MBS updates $\boldsymbol{\eta}$ according to (40), and broadcasts those values to all SBSs via backhaul, $i = i + 1$
14: **until** convergence $\left(\sum_{n=1}^{N} |\eta_n^i - \eta_n^{i-1}| \leq \varepsilon_2\right)$ or $i = I_{max}$

downlink can be estimated at SBS k if the symmetry between the uplink and the downlink exists [25, 26]. Or, it can be assumed that there is direct wired connection between a SBS and the MBS for the SBS to coordinate with the central MBS [7, 10], according to a candidate scheme proposed for 3GPP HeNB mobility enhancement [23].

3) Computation Complexity of DCOA: The DCOA algorithm can be implemented at each SBS, but requires MBS' coordination for updating $\boldsymbol{\eta}$. The computation complexity of DCOA at each time slot is $O(KU) + O(I_{max}\frac{NKU}{\varepsilon_1^2})$, where $O(KU)$ is the computation complexity of the RC and virtual rate control algorithms, and $O(\frac{NU}{\varepsilon_1^2})$ is the computation complexity of the RA algorithm that is implemented in each SBS. The DCOA algorithm with a computation complexity of polynomial complexity, $O(KU)+O(I_{max}\frac{NKU}{\varepsilon_1^2})$, facilitates the practical implementation.

V. Algorithm Performance

Before the analysis it is necessary to introduce some auxiliary variables. We denote Λ as the capacity region of the SCN consisting of all admissible, i.e., consists of all feasible admission rates stabilizable by some RC and RA algorithms without considering QoS requirements (i.e., delay constraints and minimum data rate constraints). Let $\boldsymbol{r}^* = \{r^*_{k,u}\}$ be the solution to the following problem

$$\max_{\boldsymbol{r}, \boldsymbol{r} \in \Lambda} \sum_{k=1}^{K}\sum_{u=1}^{U} r_{k,u} \tag{44}$$

We let $\boldsymbol{r}^*_\epsilon = \{r^*_{k,u,\epsilon}\}$ denote the solution to

$$\max_{\boldsymbol{r}, \boldsymbol{r}+\epsilon \in \Lambda} \sum_{k=1}^{K}\sum_{u=1}^{U} r_{k,u,\epsilon}$$
$$\text{s.t.} \quad r_{k,u,\epsilon} \geq O_{k,u},$$

where ϵ is a positive number that can be chosen arbitrarily small. For simplicity of analysis, we assume that $O_{k,u}$ is in the interior of Λ, and without loss of generality, we assume that there exists ϵ such that $r^*_{k,u,\epsilon} \geq O_{k,u} + \epsilon$. According to [35, 36], it is true that

$$\lim_{\epsilon \to 0} \sum_{k=1}^{K}\sum_{u=1}^{U} r^*_{k,u,\epsilon} = \sum_{k=1}^{K}\sum_{u=1}^{U} r^*_{k,u}, \tag{45}$$

The performance of the DCOA are listed in **Theorem 2** and **Theorem 3**.

Theorem 2: Employing the proposed DCOA, all the actual data queues have deterministic worst-case bounds:

$$Q_{k,u}(t) \leq q_{max}, \forall k, u, t. \tag{46}$$

Proof: Please refer to Appendix B for the proof. ■

Theorem 3:

(a) Given $\epsilon > 0$, if

$$q_{max} > \frac{C^2_{max} + A^2_{max}}{2\epsilon} + A_{max} \quad \text{and} \quad D_{k,u} > \frac{q_{max}}{\phi^*_{k,u,\epsilon}}, \forall k, u, \tag{47}$$

where C_{max} is the maximum downlink transmission rate in the SCN. DCOA can achieve a downlink capacity

$$\lim_{T\to\infty} \inf \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U} \mathbb{E}\{r_{k,u}\} \geq \sum_{k=1}^{K}\sum_{u=1}^{U} r^*_{k,u,\epsilon} - \frac{B}{V}. \tag{48}$$

(b) DCOA ensures that the virtual queues X, Y, Z have an upper bound:

$$\lim_{T\to\infty} \sup \frac{1}{T}\sum_{t=0}^{T-1} \left\{ \sum_{k=1}^{K}\sum_{u=1}^{U} [X_{k,u}(t) + Y_{k,u}(t) + Z_{k,u}(t)] \right\}$$
$$\leq \frac{B'}{\delta}, \tag{49}$$

where $B' = B + VB_1$, and B_1 is a finite positive constant.

Proof: Please refer to Appendix C for the proof. ■

Remark 3 (Network Stability): The inequalities (46) in **Theorem 2** and (49) in **Theorem 3(b)** indicate that DCOA stabilizes the actual queues and the virtual queues. As immediate results, DCOA stabilizes the SCN and satisfies each SUE's delay and rate requirements. In addition, **Theorem 2** indicates that all the data queues are deterministically bounded by q_{max}, which guarantees the finite buffer size at each SBS.

Remark 4 (Optimal Capacity Performance): Eq. (48) gives the lower-bound of the downlink capacity that the proposed DCOA can achieve. Since B is a constant independent of V, the DCOA would achieve a time-averaged capacity arbitrarily close to $\sum_{k=1}^{K}\sum_{u=1}^{U} r^*_{k,u,\epsilon}$ for some $\epsilon \geq 0$ by choosing a sufficiently large $V \geq 0$. In addition, as shown in (48), when ϵ tends to zeros, DCOA would achieve a downlink capacity arbitrarily to $\sum_{k=1}^{K}\sum_{u=1}^{U} r^*_{k,u}$ with a tradeoff in queue length bound q_{max} and delay constraint $D_{k,u}$, both of which are lower-bounded by the reciprocal terms of ϵ as shown in (47).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

In DCOA, the control parameter V, which is typically chosen to be large, does not affect the actual queue backlog upper bound. However, a larger V increases the upper bound of the virtual queue backlogs as shown in (49) and results in a slow convergence time of the virtual queues. Thus, by establishing virtual queues, V's influence on queue length [16, 30] is shifted from actual queues to virtual queues.

VI. Simulation Results

We simulate an SCN where an macrocell is underlaid with $K = 2$ uniformly and randomly distributed small cells. In the simulations, SUEs users are uniformly distributed in the coverage area of their serving SBS; The bandwidth $F = 10$MHz, $P_{max} = 0.1$W, $I_n^{th}(t) = 2 \times 10^{-10}, \forall n, t$, $N = B = 10$, $U = 2$ and $\sigma^2 = (F/N)N_0$, where $N_0 = -174$dBm/Hz is the AWGN power spectral density. The coverage radius of the macrocell is 500 m, and that of a small cell is 10 m. The path loss model P_d between a SUE and a SBS/MBS is modeled as $P_d = 15.3 + 37.6 \log 10(d)$ [16], where d is the distance from a SUE to a SBS/MBS. Channel gains are modeled considering both poss loss and shadow fading. Specifically, we define $h(t) = 10^{\frac{-P_d+\Psi_B}{10}}$ as the channel gain between the SUE and the SBS/MBS [43], where Ψ_B is normally distributed random variable with mean zero and and deviation $\delta_{\Psi_B} = 10$dB. Without loss of generality, MBS transmits on each channel at its peak transmit power $p_{b,n}^M(t) = 1$W $\forall n, t,$, the bursty data arrival is Poisson distributed with average arrival rate $\boldsymbol{\lambda}$ being $[10, 10; 10, 10]$, the rate requirements $\mathbf{O}$ are $[2, 2; 2, 2]$, the delay requirements $\mathbf{D}$ are $\{50, 50; 50, 50\}$, and maximum buffer size q_{max} is 500 in the simulations. The simulation is carried out for $T = 2000$ consecutive time slots.

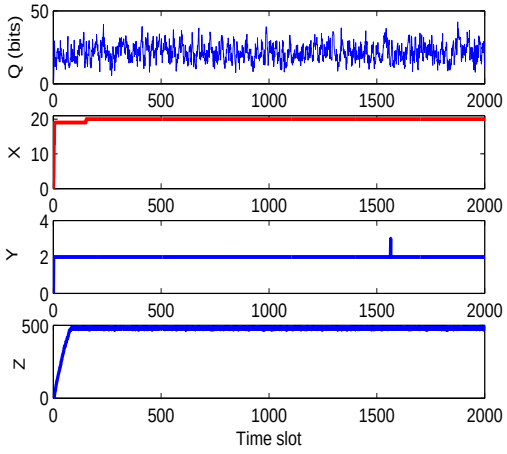


Fig. 2. Queue Stability.

First, we demonstrate the queue stability in Fig. 2 with $V = 10$. Because all SUEs' data queues Q and virtual queues X, Y, Z enjoy similar trends, we take queues of SUE $u = 1$ in small cell $k = 1$ as an example. Fig. 2 shows the dynamics of SUE's data queue $Q_{1,1}$, virtual queue $X_{1,1}, Y_{1,1}, Z_{1,1}$. We observe that the actual data queues are strictly lower than the buffer size $q_{max} = 500$, which verifies **Theorem 2**. In addition, from Fig. 2, we also observe that all of the virtual queues are bounded, which verifies **Theorem 3(b)** and means that the long-term time-averaged constraints of delay and rate for SUEs are both satisfied. Fig. 2 shows the proposed DCOA is effective for maximizing the capacity of spectrum-sharing SCNs while satisfying SUEs' delay and rate constraints.

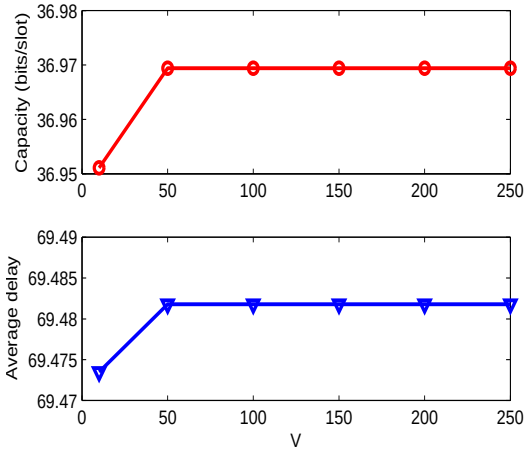


Fig. 3. Network performance of DCOA versus the value of V.

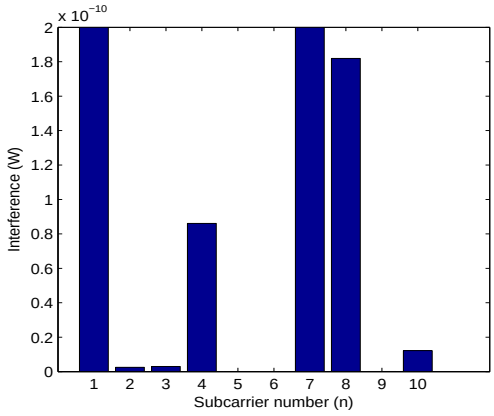
Fig. 3 shows the downlink capacity of the SCN and the average delay of the SCN, denoted by $\rho = \frac{1}{KU} \sum_{k=1}^{K} \sum_{u=1}^{U} \rho_{k,u}$, of DCOA by varying the control parameter V. With an increasing V, we observe that the downlink capacity of the SCN is slightly increased while the average end-to-end delay is slightly decreased.

However, although a large value of V increases the upper bound of the virtual queue backlogs (as stated in **Theorem 3** (b)), it has a negative effect on the convergence rates of virtual queues. For example, specifically, for $V = 10000$, it takes more than 1000 time slots for the virtual queues to converge. In comparison, in general, it only takes about less than 100 time slots for the virtual queues to converge when $V = 50$. Thus, $V = 50$ is sufficiently large for maximizing the capacity of the SCN since a larger V does not lead to a noticeable increase in the capacity while resulting in a smaller convergence rate of virtual queues.

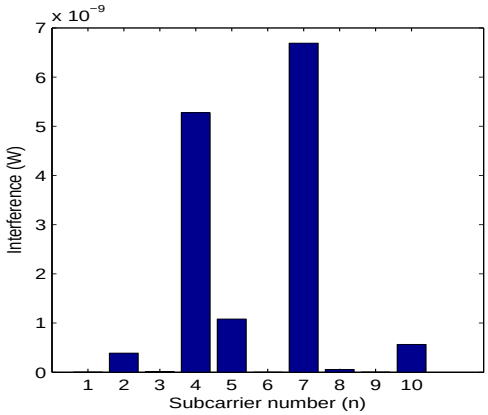
As an example, Fig. 4 (a) and Fig. 4 (b) show the accumulated interference, $I_n(t) = \sum_{k=1}^{K} \sum_{u=1}^{U} w_{k,u,n}(t) p_{k,u,n}^S(t) h_{k,b,n}^S(t)$, at time slot $t = 1200$ for the $N = 10$ subcarriers, with and without the interference constraints, respectively, with $V = 50$, $\mathbf{O} = [2, 2; 2, 2]$, and $\mathbf{D} = \{50, 50, 50, 50\}$. From Figs. 4 (a) and 4(b), we can observe that the imposed inter-cell interference constraints are necessary to ensure that MUEs are not severely affected by SUEs' utilization of the same channel.

Fig. 5 shows the performance of DCOA by varying the maximum buffer size q_{max} with $V = 50$, $\mathbf{O} = [2, 2; 2, 2]$, and $\mathbf{D} = \{50, 50, 50, 50\}$. With an increasing q_{max}, we observe

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology



(a) Interference with the interference constraints



(b) Interference without the interference constraints

Fig. 4. Interference.

that both the downlink capacity and the average delay are increased. Intuitively, this is because a larger value of q_{max} means that more data is admitted into the data queue at SBS for transmission, and thus the average delay is increased.

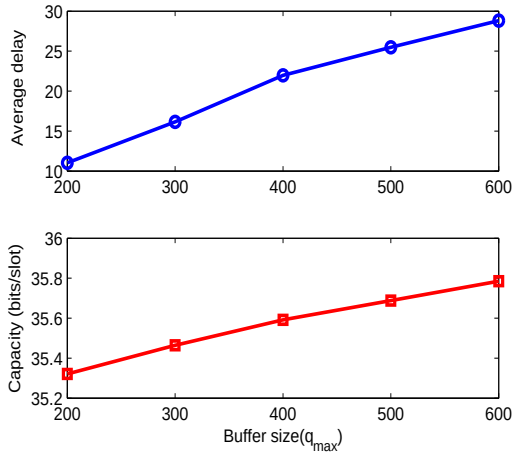


Fig. 5. Network performance of DCOA versus buffer size q_{max}.

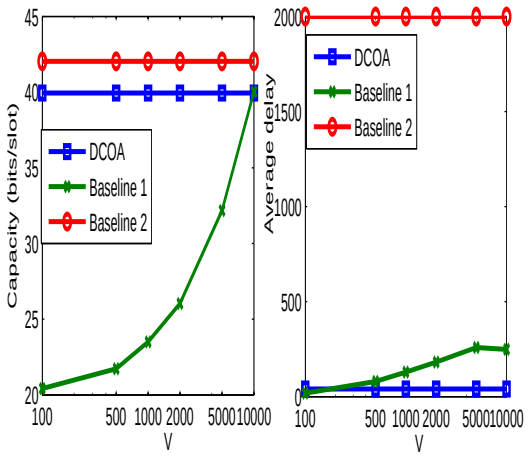


Fig. 6. Network performance of DCOA compared with two baselines.

Finally, we compare the performance of DCOA with two baselines in Fig. 6. Baseline 1 is the JACRA algorithm in [16], which maximizes the capacity of SCNs only subject to the data queue stability without consideration of SUE's delay constraint. [16] shows that baseline 1 (JACRA) can achieve the near-to-optimal capacity of the SCN subject to the network stability when the control parameter V is sufficiently enough. Baseline 2 is the capacity optimal algorithm for maximizing the (instantaneous) capacity of SCNs at each time slot without consideration of the bursty data arrivals at SUEs in [26].

From Fig. 6 (b), we observe that the DCOA algorithm always satisfies SUEs' delay constraints whereas it is evident that neither of the two baselines can provide guaranteed delay performance for SUEs. Moreover, from Fig. 6 (a) we observe that DCOA not only satisfy SUEs' delay requirements, it can also achieve almost the same (about %99) downlink capacity performance as baseline 1.

In addition, from Figs. 6 (a) and 6 (b), we observe that although baseline 2 achieves the best performance in downlink capacity of the SCN, it achieves the worst performance among the three algorithms in terms of the average delay. This is because by simply assuming infinite buffers at SBSs, baseline 2 is not aware of QSI, hence, leading to the worst delay performance. In fact, the achieved downlink capacity of baseline 2 may not be effective, because in our simulation, we observe that the data queues at SBSs are not even stable by using baseline 2. In addition, we also observe that the better downlink capacity of baseline 1 (JACRA) over the proposed DCOA is achieved at the expense of the larger delay by increasing the control parameter V. In comparison, DCOA not only provides guaranteed delay performance, but also achieves almost the same downlink capacity performance as baseline 1. Actually, the much better delay performance of the DCOA algorithm over the JACRA algorithm in baseline 1 is owing to the establishment of the virtual queues X, Y and Z. With the virtual queues, "the burden of V" is shifted from actual queues to virtual queues and that SUEs' delay constraints are satisfied

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

with finite buffer sizes for all actual data queues at SBSs (as stated in **Remark** 4).

VII. Conclusions and Future Research

In this paper, we have investigated the delay-guaranteed resource allocation for spectrum-sharing SCNs with cross-tier interference. We have introduced an interference temperature limit to protect MUEs from severe cross-tier interference. Stochastic optimization model is employed to maximize the long-term time-averaged capacity of SUEs subject to each SUE's delay constraint, minimum data rate constraint, and an interference temperature limit constraint. We have developed a delay-guaranteed rate control and resource allocation algorithm (DCOA) to obtain the optimal RC and RA decisions without prior statistical information. The DCOA can provide guaranteed delay for SUE in SCN with bursty data arrivals.

Our work is suitable for delay-aware applications such as video and voice, which typically have a time-averaged delay requirement. In our future work, we want to (1) investigate the effect of imperfect network state information such as CSI and/or QSI on DCOA; (2) extend the work to more complex type of wireless network technology, e.g., wireless network virtualization (WNV) [47], in which a low-complexity and fully distributed RA algorithm may be desirable; and (3) modify the proposed DCOA by introducing less variables (or virtual queues) based on a more efficient Lyapunov function.

Appendix A
Proof of Theorem 1

Lemma 1: For any nonnegative real numbers x, y and z, there holds $[\max(x-y,0)+z]^2 \leq x^2+y^2+z^2-2x(y-z)$ [29].

From **Lemma 1**, we have

$$\begin{aligned}
&\frac{1}{2q_{max}}\left[X_{k,u}(t+1)Q_{k,u}^2(t+1)-X_{k,u}(t)Q_{k,u}^2(t)\right]\\
&\leq \frac{1}{2q_{max}}\left[\left(X_{k,u}(t)+\nu_{k,u}(t)\right)Q_{k,u}^2(t+1)-X_{k,u}(t)Q_{k,u}^2(t)\right]\\
&\leq \frac{1}{2q_{max}}\left\{\nu_{k,u}(t)Q_{k,u}^2(t+1)\right\}\\
&+\frac{1}{2q_{max}}X_{k,u}(t)\\
&\times\left[R_{k,u}^2(t)+C_{k,u}^2-2Q_{k,u}(t)\left(C_{k,u}(t)-R_{k,u}(t)\right)\right]\\
&\leq \frac{A_{max}q_{max}}{2}+\frac{R_{k,u}^2(t)+C_{k,u}^2(t)}{2q_{max}}X_{k,u}(t)\\
&-\frac{C_{k,u}(t)-R_{k,u}(t)}{q_{max}}X_{k,u}(t)Q_{k,u}(t). \qquad (50)
\end{aligned}$$

By squaring both sides of the queue dynamics (7), (16), (14), and (15), and by employing (50), we obtain

$$\begin{aligned}
&\Delta(\mathbf{G}(t)) \triangleq \mathbb{E}\{L\left(\mathbf{G}(t+1)\right)-L\left(\mathbf{G}(t)\right)|\mathbf{G}(t)\}\\
&\leq B+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left(\frac{R_{k,u}^2(t)+C_{k,u}^2(t)}{2q_{max}}X_{k,u}(t)|\mathbf{G}(t)\right)\\
&-\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\frac{X_{k,u}(t)Q_{k,u}(t)}{q_{max}}C_{k,u}(t)|\mathbf{G}(t)\right\}\\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{R_{k,u}(t)\frac{X_{k,u}(t)}{q_{max}}[Q_{k,u}(t)-(q_{max}-A_{max})]|\mathbf{G}(t)\right\}\\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\nu_{k,u}(t)\left(\frac{q_{max}-A_{max}}{q_{max}}X_{k,u}(t)-Y_{k,u}(t)\right)|\mathbf{G}(t)\right\}\\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\nu_{k,u}(t)\left(-D_{k,u}Z_{k,u}(t)-V\right)|\mathbf{G}(t)\right\}\\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{Y_{k,u}(t)O_{k,u}+Z_{k,u}(t)Q_{k,u}(t)|\mathbf{G}(t)\right\}. \qquad (51)
\end{aligned}$$

By adding and subtracting the term $V\mathbb{E}\{\nu(t)|\mathbf{G}(t)\}$ to the R.H.S. of (50), we can prove (25).

Appendix B
Proof of Theorem 2

We use mathematical induction to prove (46). When $t=0$, we have $Q_{k,u}(t)=0, \forall k,u$. Now we suppose that at time slot t, we have $Q_{k,u}(t) \leq q_{max}, \forall k,u$. There are two cases as follows.

- Case 1: If $Q_{k,u}(t) \leq q_{max}-A_{max}$, then according to the data queue dynamics in (8), we have

$$\begin{aligned}
Q_{k,u}(t+1) &\leq Q_{k,u}(t)+R_{k,u}(t)\\
&\leq q_{max}-A_{max}+R_{k,u}(t) \leq q_{max}. \qquad (52)
\end{aligned}$$

- Case 2: if $Q_{k,u}(t) > q_{max}-A_{max}$, then according to the RC in (27), we have $R_{k,u}(t)=0$. Thus $Q_{k,u}(t+1)=\left(Q_{k,u}(t)-C_{k,u}(t)\right)^+ \leq Q_{k,u}(t) < q_{max}$.
  Therefore, $Q_{k,u}(t) \leq q_{max}, \forall k,u,t$.

Appendix C
Proof of Theorem 3

Lemma 2: For any feasible rate vector $\boldsymbol{\pi} \in \Lambda$ with $\pi_{k,u} \geq O_{k,u}, \forall O_{k,u} \in \Lambda$, there exists a stationary randomized scheduling and resource allocation algorithm that chooses scheduling and resource allocation strategy independent of queue backlogs and yields:

$$\mathbb{E}\{C_{k,u}(t)\}=\mathbb{E}\{R_{k,u}(t)\}=\mathbb{E}\{\nu_{k,u}(t)\}=\pi_{k,u}, \forall k,u. \quad (53)$$

Notice that, the stationary randomized scheduling and resource allocation policy algorithm makes decisions only depending on channel condition and independent of queue backlogs. Furthermore it may not fulfill the delay constraints. Similar proof can be found in [30, 37] and the proof of **Lemma 2** is omitted here.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

Note that $\pi_{k,u}$ can take values as $r^*_{k,u,\epsilon}$ or $r^*_{k,u,2\epsilon}$, where we recall $r^*_{k,u} \in \Lambda$ and $r^*_{k,u,\epsilon} \geq O_{k,u} + \epsilon, \forall, k, u$. Thus, We can control the admission rate of r ranging from $r^*_{k,u}$ to $r^*_{k,u,\epsilon}$ or to $r^*_{k,u,2\epsilon}$ arbitrarily and resulting in that both $r^*_{k,u,\epsilon}$ and $r^*_{k,u,2\epsilon}$ are within Λ.

Because DCOA minimizes the R.H.S of (25) over C1,C2, C3, C4 and C5, we have

$$\begin{aligned}
&\Delta\left(\mathbf{G}(t)\right) - V\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\{\nu_{k,u}(t)|\mathbf{G}(t)\} \\
&\leq B + \sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\frac{R^2_{k,u}(t)+C^2_{k,u}(t)}{2q_{max}}X_k(t)|\mathbf{G}(t)\right\}. \\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{R_{k,u}(t)\frac{X_{k,u}(t)}{q_{max}}[Q_{k,u}(t)-(q_{max}-A_{max})]|\mathbf{G}(t)\right\} \\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\nu_{k,u}(t)\left(\frac{q_{max}-A_{max}}{q_{max}}X_{k,u}(t)-Y_{k,u}(t)\right)|\mathbf{G}(t)\right\} \\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{\nu_{k,u}(t)\left(-D_{k,u}Z_{k,u}(t)-V\right)|\mathbf{G}(t)\right\} \\
&-\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{C_{k,u}(t)\frac{X_{k,u}(t)Q_{k,u}(t)}{q_{max}}|\mathbf{G}(t)\right\} \\
&+\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\left\{Y_{k,u}(t)O_{k,u}+Z_{k,u}(t)Q_{k,u}(t)|\mathbf{G}(t)\right\}. \qquad (54)
\end{aligned}$$

Note that the third term, the fourth and fifth terms, and the sixth term of the R.H.S of (54) are minimized by the RC (26), the virtual rate control (28), the RA (30) policy, respectively, over a set of feasible algorithms including the stationary randomized algorithm introduced in **Lemma 2**. We can substitute into the third and the sixth terms of (54) a stationary randomized algorithm with admission rate vector $r^*_{k,u,2\epsilon}$, and into the fourth and fifth terms of (54) a stationary randomized algorithm with admitted admission rate vector $r^*_{k,u,\epsilon}$. Thus, since the proposed DCOA minimize the R.H.S of (54) over all possible policies including the policy in **Lemma 2**, we can get

$$\begin{aligned}
&\Delta\left(\mathbf{G}(t)\right) - V\mathbb{E}\left\{\sum_{k=1}^{K}\sum_{u=1}^{U}\nu_{k,u}(t)|\mathbf{G}(t)\right\} \\
&\leq B - V\sum_{k=1}^{K}\sum_{u=1}^{U}r^*_{k,u,\epsilon} - \epsilon\sum_{k=1}^{K}\sum_{u=1}^{U}Y_{k,u} \\
&-\sum_{k=1}^{K}\sum_{u=1}^{U}Z_{k,u}(t)(D_{k,u}r^*_{k,u,\epsilon}-q_{max}) \\
&-\sum_{k=1}^{K}\sum_{u=1}^{U}\frac{X_{k,u}(t)}{q_{max}}\left[\epsilon(q_{max}-A_{max})-\frac{R^2_{k,u}(t)+C^2_{k,u}(t)}{2}\right]. \qquad (55)
\end{aligned}$$

When (47) holds, we can find that $\epsilon_1 > 0$ such that $\epsilon_1 \leq D_{k,u}r^*_{k,u,\epsilon} - q_{max}$ and $\epsilon_1 \leq \frac{2\epsilon(q_{max}-A_{max})-R^2_{k,u}(t)-C^2_{k,u}(t)}{2q_{max}}$. Thus, we have

$$\begin{aligned}
&\Delta\left(\mathbf{G}(t)\right) - V\mathbb{E}\{\nu(t)|\mathbf{G}(t)\} \\
&\leq B - \delta\sum_{k=1}^{K}\sum_{u=1}^{U}(X_{k,u}(t)+Y_{k,u}(t)+Z_{k,u}(t)) \\
&-V\sum_{k=1}^{K}\sum_{u=1}^{U}r^*_{k,u,\epsilon}. \qquad (56)
\end{aligned}$$

By taking iterated expectation and using telescoping sums over $t \in \{0, 1, ..., T-1\}$ in the above inequality, we get

$$\begin{aligned}
&\mathbb{E}\left\{L(\mathbf{G}(T)\right\} - \mathbb{E}\left\{L(\mathbf{G}(0)\right\} - \sum_{t=0}^{T-1}V\mathbb{E}\{\nu(t)\} \\
&\leq TB - \delta\sum_{t=0}^{T-1}\mathbb{E}\left\{\sum_{k=1}^{K}\sum_{u=1}^{U}(X_{k,u}(t)+Y_{k,u}(t)+Z_{k,u}(t))\right\} \\
&-V\sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U}r^*_{k,u}, \qquad (57)
\end{aligned}$$

where $\sigma \triangleq \min\{\epsilon, \epsilon_1\}$. Considering the fact that $\mathbb{E}\left\{L(\mathbf{G}(T)\right\}$ is nonnegative and taking limsup of T on both sides of (57), rearranging the terms, and ignoring unimportant terms, we get

$$\begin{aligned}
&\delta\sum_{t=0}^{T-1}\mathbb{E}\left\{\sum_{k=1}^{K}\sum_{u=1}^{U}(X_{k,u}(t)+Y_{k,u}(t)+Z_{k,u}(t))\right\} \\
&\leq \mathbb{E}\left\{L(\mathbf{G}(0)\right\} + \sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U}V\mathbb{E}\{\nu_{k,u}(t)\} + TB \\
&-V\sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U}r^*_{k,u,\epsilon}. \qquad (58)
\end{aligned}$$

Dividing (58) by $T\delta$, we have

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{\sum_{k=1}^{K}\sum_{u=1}^{U}(X_{k,u}(t)+Y_{k,u}(t)+Z_{k,u}(t)\right\} \\
&\leq \frac{\mathbb{E}\left\{L(\mathbf{G}(0)\right\}}{\delta T} + \frac{TB}{\delta T} + \frac{V\sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U}\mathbb{E}\{\nu_{k,u}(t)\}}{\delta T} \\
&-\frac{V\sum_{t=0}^{T-1}\sum_{k=1}^{K}\sum_{u=1}^{U}r^*_{k,u,\epsilon}}{\delta T}. \qquad (59)
\end{aligned}$$

Since $\lim_{T\to\infty}\frac{1}{T}\sup\sum_{t=0}^{T-1}\sum_{k=0}^{K}\sum_{u=1}^{U}\mathbb{E}\{\nu_{k,u}(t)\}$ is bounded above (say, by a constant B_1 with $B_1 \leq KUA_{max}$, taking a limit as $T \to \infty$, we have

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{\sum_{k=1}^{K}\sum_{u=1}^{U}(X_{k,u}(t)+Y_{k,u}(t)+Z_{k,u}(t)\right\} \\
&\leq \frac{B}{\delta} + \frac{V}{\delta}(B_1 - r^*_\epsilon), \\
&\leq \frac{B}{\delta} + \frac{VB_1}{\sigma}, \qquad (60)
\end{aligned}$$

which prove (b).

Similarly, we can prove (a).

REFERENCES

[1] Ericsson, "5G Radio Access, Research, and Vision," white paper, 2013.

[2] V. Chandrasekhar and J. G. Andrews, "Femtocell networks: A survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59-67, Sept. 2008.

[3] J. G. Andrews, H. Claussen, and M. Dohler, "Femtocells: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497-508, April 2012.

[4] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: Technical and business perspectives," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 104-112, April 2013.

[5] D. L. Perez, A. Valcarce, and G. D. L. Roche, "OFDMA femtocells: A roadmap on interference avoidance," *IEEE Communications Magazine*, vol. 47, no. 9, pp. 41-48, Sept. 2009.

[6] J. H. Yun and K. G. Shin, "Adaptive interference management of OFDMA femtocells for co-channel deployment," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1225-1241, June 2011.

[7] K. Son, S. Lee, and Y. Yi, "Refim: A practical interference management in heterogeneous wireless access networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1260-1272, June 2011.

[8] X. Chen, J. Wu, and Y. Cai, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 627-640, April 2015.

[9] M. Yavuz, F. Meshkati, and S. Nanda, "Interference management and performance analysis of UMTS/HSPA+ Femtocells," *IEEE Communications Magazine*, vol. 47, no. 9, pp. 102-109, Sept. 2009.

[10] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A stackelberg game approach," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 538-549, April 2012.

[11] D. C. Oh, H. C. Lee, and Y. H. Lee, "Power control and beamforming for femtocells in the presence of channel uncertainty," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 6, pp. 2545-2554, July 2011.

[12] J. Zhang, Z. Zhang, and K. Wu, "Optimal distributed subchannel, rate and power allocation algorithm in OFDM-based two-tier femtocell networks," in *Proceedings of IEEE VTC*, pp. 1-5, May 2010.

[13] I. Guvenc, M. R. Jeong, and F. Watanabe, "A hybrid frequency assignment for femtocells and coverage area analysis for co-channel operation," *IEEE Communications Letters*, vol. 12, no. 12, pp. 880-882, Dec. 2008.

[14] J. W. Huang and V. Krishnamurthy, "Cognitive base stations in LTE/3GPP femtocells: A correlated equilibrium game-theoretic approach," *IEEE Transactions on Wireless Communications*, vol. 59, no. 12, pp. 3485-3493, Dec. 2011.

[15] J. Kim and D. H. Cho, "A joint power and subchannel allocation scheme maximizing system capacity in indoor dense mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 9, pp. 4340-4353, Nov. 2010.

[16] X. Xiang, C. Lin, and X. Chen, "Toward optimal admission control and resource allocation for LTE-A femtocell uplink," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3247-3261, July, 2015.

[17] J. Li, Y. Li, and A. Cheng, "Delay aware cell association and user scheduling in heterogeneous overlay networks," in *Proceedings of IEEE PIMRC*, pp. 106-110, Sept. 2013.

[18] X. Zhu, B. Yang, and C. Chen, "Cross-layer scheduling for OFDMA-based cognitive radio systems with delay and security constraints," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5919-5934, Dec. 2015.

[19] C. Jiang, Y. Chen, and Y. Gao, "Joint spectrum sensing and access evolutionary game in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2470-2483, May 2013.

[20] C. Jiang, Y. Chen, and R. K. J. Liu, "Renewal-theoretical dynamic spectrum access in cognitive radio networks with unknown primary behavior," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 3, pp. 406-416, March 2013.

[21] D. T. Ngo and T. L. Ngoc, "Distributed resource allocation for cognitive radio networks with spectrum-sharing constraints," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 7, pp. 3436-3449, Sept. 2011.

[22] K. W. Choi, E. Hossain, and D. I. Kim, "Downlink subchannel and power allocation in multi-cell OFDMA cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 7, pp. 2259-2271, July 2011.

[23] Y. Ma, D. I. Kim, and Z. Wu, "Optimization of OFDMA-based cellular cognitive radio networks," *IEEE Transactions Communications*, vol. 58, no. 8, pp. 2265-2276, Aug. 2010.

[24] *Way Forward Proposal for (H)eNB to HeNB Mobility*, 3GPP Std. R3-101 849, 2010. femtocells for co-channel deployment," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1225-1241, June 2011.

[25] W. Jing, Z. Lu, and H. Zhang, "Energy-saving resource allocation scheme with QoS provisioning in OFDMA femtocell networks," in *Proceedings of IEEE ICC*, pp. 912-917, June 2014.

[26] H. Zhang, C. Jiang, and N. C. Beaulieu, "Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services," *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2366-2377, July 2014.

[27] T. Nakamura, S. Nagata, and A. Benjebbour, "Trends in small cell enhancements in LTE advanced," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 98-105, Feb. 2013.

[28] R. Urgaonkar and M. J. Neely, "Opportunistic scheduling with reliability guarantees in cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 766-777, June 2009.

[29] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[30] M. J. Neely, "Energy optimal control for time varying wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 2915-2934, July 2006.

[31] C. Y. Wong, R. Cheng, and K. Lataief, "Multiuser OFDM with adaptive subcarrier, bit, power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747-1758, Oct. 1999.

[32] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310- 1322, July 2006.

[33] M. Tao, Y. C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2190-2201, June 2008.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[35] M. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2003.

[36] A. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401-457.

[37] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource alloca-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2016.2606644, IEEE Transactions on Vehicular Technology

tion and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1-144, April 2006.
[38] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, New Jersey 07632, Prentice Hall, 1992.
[39] P. Xia, V. Chandrasekhar, and J. Andrews, "Open vs. closed access femtocells in the uplink," *IEEE Transactions Wireless Communications*, vol. 9, no. 12, pp. 3798-3809, Dec. 2010.
[40] A. E. Essaili, L. Zhou, and D. Schroeder, "QoE-driven live and on-demand LTE uplink video transmission," in *proceedings of IEEE Multimedia Signal Processing (MMSP)*, pp. 1-6, Oct. 2011.
[41] J. P. M. Gea, R. A. Pardo, and H. Wehbe, "Optimization framework for uplink video transmission in HetNets," in *proceedings of ACM Mobile Video Delivery (MoVidD)*, pp. 1-6, March 2014.
[42] D. Wang, L. Toni, and P. C. Cosman, "Uplink resource management for multiuser OFDM video transmission systems: Analysis and algorithm design," *IEEE Transactions on Communications*, vol. 61, no. 5, pp. 2060-2073, May 2013.
[43] 3GPP, "FDD base station (BS) classification (release 11)," TR 25.951 V11.0.0, Technical Report, Sep. 2012.
[44] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[45] Y. Song, C. Zhang, and Y. Fang, "Revenue maximization in time-varying multi-hop wireless networks: A dynamic pricing approach," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 7, pp. 1237-1245, Aug. 2012.
[46] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Massachusetts: Athena Scientific, 2007.
[47] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials,*, vol. 17, no. 1, pp. 358-380, Firstquarter 2015.

Yashuang Guo received the B.S. degree in electronic and information engineering from Dalian University, China, in 2011. She is currently pursuing a Ph.D. degree in Communication and Information Systems at Xidian University. Her research interests include cross-layer design, QoS provisioning and applications of stochastic optimization in wireless networks.

Qinghai Yang received his B.S. degree in Communication Engineering from Shandong University of Technology, China in 1998, M.S. degree in Information and Communication Systems from Xidian University, China in 2001, and Ph. D. in Communication Engineering from Inha University, Korea in 2007 with university-president award. From 2007 to 2008, he was a research fellow at UWB-ITRC, Korea. Since 2008, he is with Xidian University, China . His current research interest lies in the fields of autonomic communication, content delivery networks and LTE-A techniques.

Jiayi Liu is currently working as a Lecture in Xidian University. She received her PhD degree in November 2013 in Computer Ssiences from Telecom-Bretagne, Rennes, France. She obtained a Master of Science in Computer Science from Rennes1 University in 2009, and a Bachelors of Science in Electronic Engineering from Xidian University in Xi'an China, in 2007. Her research interests include video dilivery systems, content distribution in mobile network, resource allocation and optimization problems.

Kyung Sup Kwak(M'81) received the B.S. degree from the Inha University, Inchon, Korea in 1977, and the M.S. degree from the Uni-versity of Southern California in 1981 and the Ph.D. degree from the University of California at San Diego in 1988, under the Inha University Fellowship and the Korea Electric Association Abroad Scholarship Grants, respectively. He worked for Hughes Network System, San Diego, USA, and IBM Network Research Center, Research Triangle Park, USA from 1988 to 1990 and has been with Inha University from 1990 as the Inha Fellow Professor and now as the Inha Hanlim Professor and is the director UWB Wireless Communications Research Center, Korea. His research interests include multiple access communication systems, mobile communication systems, UWB radio systems and ad-hoc networks, high-performance wireless Internet. Mr. Kwak is members of IEEE, IEICE, KICS and KIEE.