Base Station Association in Wireless Cellular Networks: An Emulation Based Approach

Soohwan Lee, *Student Member, IEEE*, Kyuho Son, *Member, IEEE*, Huazhi Gong, and Yung Yi, *Member, IEEE*

***Abstract*—In order to utilize network resources efficiently and reduce regional congestion, associating mobile stations (MSs) with proper base stations (BSs) is of crucial importance in wireless cellular networks. There have been several load-aware proposals in literature, where most are classified into so-called closed-form approaches. In such approaches, each MS independently and deterministically selects the BS which is expected to provide the highest throughput. The throughput is estimated by a closed-form equation based on the assumption of the Proportional Fair (PF) user scheduler that ensures temporal fairness. However, the closed-form approaches do not perform well when the closed-form equation is not available, e.g., general α-fair user scheduler, where temporal fairness is not guaranteed, or deterministic BS association may make wrong decisions, e.g., under the dynamics of mobility or flow arrivals/departures. In this paper, we propose a novel BS association scheme, called ViSE (Virtual Scheduling based Emulation) to tackle such challenges. It emulates an optimal BS association by running a notion of virtual scheduler, and each MS randomly determines its associated BS with the probability proportional to the throughput virtually allocated by the virtual scheduler. We demonstrate through extensive simulations under various practical scenarios that ViSE outperforms the existing algorithms in terms of user schedulers with diverse fairness and robustness to network dynamics.**

***Index Terms*—Base station association, emulation-based, virtual scheduling, load balancing, cellular networks.**

I. Introduction

PREVALENCE of smart phones is accelerating the increase of mobile data traffic. Many researchers in financial sectors forecast that mobile data traffic will reach about 10.8 exabytes per month by 2016 [1], [2]. To support the high data demand, the standards such as Mobile WiMax (802.16m) and 3GPP LTE have focused on enhancing spatial reuse [3], [4]. Thus, small cells, e.g., femto and pico cells, are expected to rapidly emerge, and future cellular networks will consist of a complex mixture of small and macro BSs. In this trend, an MS is likely to have many candidate serving BSs, and the problem of associating an MS with an appropriate BS is becoming more important [5].

A conventional approach is to connect to the BS providing the highest received signal strength, but this approach does not consider the number of associated MSs in a cell. The user population in a cell has a big impact on the actual MS throughput due to resource sharing among users (i.e., user scheduling). Several BS association schemes adaptive to the offered loads and the number of MSs for better resource utilization and alleviation of regional congestion, have been recently proposed under slightly different models in terms of traffic model, network heterogeneity, bandwidth splitting policy, and centralized/distributed control [6]–[10].

These schemes suggest adaptive association algorithms that can be understood as ones inspired by an optimization framework. In this framework, an objective function that maximizes a long-term aggregate measure, e.g., utility (thus, fairness) or throughput, is considered, and then a per-slot optimal algorithm of deciding BS association and user scheduling is developed. The optimal algorithm is typically impractical due to (i) *spatial hardness* that an MS needs to consider all the BSs as an association candidate (and thus heavy message exchanges), and (ii) *temporal hardness* that MSs may need to change their association at each slot. Frequent association changes in may be undesirable due to large system overheads in the back-haul, e.g., traffic re-routing and service disruptions.

The algorithms proposed in [6]–[10] can be understood as the approximating approaches of the optimal algorithm to trade-off between practicability and performance. The initial work in [6] assumes unsaturated traffic models and proposes the Max-Weight [11] based joint BS association and user scheduling with reduction of spatial hardness using the notion of clustering, i.e., only a group of neighboring BSs are considered for an association change. However, temporal hardness still remains. A natural approach to tackle temporal hardness is to add the "dwell time constraint" that MSs can change BS association only every at least, say T, time slots, called association epoch, where T is a system parameter[1].

More practical BS association algorithms have been proposed in [8]–[10], which we broadly categorize into *closed-form based approaches.* In these approaches, in conjunction with the dwell time constraint, assuming the Proportional Fair (PF) user scheduler [13], at each association epoch, each MS estimates the expected throughput for all neighboring BSs using the closed-form equation driven by the PF scheduler and the measurement at the previous epoch, and then, each MS *independently* selects the BS that provides the maximum estimated throughput as its future BS to be associated.

Manuscript received January 4, 2011; revised July 24, 2011 and January 25, 2012; accepted March 19, 2012. The associate editor coordinating the review of this paper and approving it for publication was B. Liang.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0015042).

S. Lee, H. Gong, and Y. Yi are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea (e-mail: shlee@lanada.kaist.ac.kr, hankgong@gmail.com, yiyung@kaist.edu).

K. Son is with the Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 (e-mail: kyuhoson@usc.edu).

Digital Object Identifier 10.1109/TWC.2012.052412.110022

[1]The dwell time T is set to be 60 ms in the Mobile WiMAX standard of IEEE 802.16m [12], but the actual dwell time constraint can be much larger than 60 ms.

The closed-form approaches perform well only when the closed-form equation reflects the reality well, which often does not hold for the following reasons: First, the key feature that enables a closed-form equation is due to the feature of temporal fairness in the PF, i.e., the per-MS throughput is simply the achievable data rate divided by the number of MSs in the same cell, regarded as the fairness achieved when $\alpha = 1$ in the notion of α-fairness in [14]. It showed that fairness is parameterized by a simple parameter α from the optimization theoretic perspective, including the popular fairness concepts, e.g., sum-throughput maximization ($\alpha = 0$) and max-min fairness ($\alpha = \infty$). Unfortunately, for $\alpha \neq 1$, the closed-form equation is unknown. Second, there may be heavy dynamic scenarios in terms of flow arrivals/departures and mobility. Under such network dynamics, the estimated throughput based on the past, which provides a guideline to select the BS for the future association, may be far from that in the future. The problem of the closed-form approaches lies in the deterministic selection of the future BS, where the wrong decision due to the big difference between the past and the future is sustained for a long time, leading to performance degradation.

In this paper, we propose a significantly different approach, called ViSE (Virtual Scheduling based Emulation). In ViSE, instead of estimating what will happen in the future based on the closed-form equation, each MS *(i) emulates* an optimal algorithm with practical complexity using *virtual scheduling (VS)*, *(ii)* records the virtual BS association histories and the virtual throughput for neighboring BSs over the duration of inter association change epochs, *(iii)* generates a probability distribution from the virtual thorughput for neighboring BSs, and *(iv)* randomly selects the BS with the recorded probability distribution.

There are two key factors that enable ViSE to outperform the closed-form approaches. First, by emulating the optimal algorithm, it is not sensitive to the type of user scheduling algorithms. In particular, ViSE works well for the user scheduler with any general α fairness. Second, by changing association probabilistically, even under network dynamics, we open the possibility of being associated with the BS reflecting the instantaneous network status, which, however, will not be selected if only the average value is used and the association is decided deterministically as in the closed-form approaches. We demonstrate ViSE's performance under various environments, e.g., general α-fair schedulers and the practical dynamics on a real 3G BS deployment topology. Compared to other competitive algorithms, ViSE achieves a performance which is the closest to the optimal algorithm for all α objectives, of course, including the case PF ($\alpha = 1$).

There exist some other related works on BS association. In [7], the authors proposed a deterministic BS association algorithm which jointly considers inter-cell interference management and transmission power control while solving BS association optimization problem with computational intractable complexity. In [15], the authors considered only stochastic arrivals which can be stabilized, and thus delay was a major performance metric for user association[2]. Our focus in this paper is the case of infinite backlog to investigate the maximum allowable aggregate throughput/utility for elastic data traffic.

The remainder of this paper is organized as follows. In Section II, we describe the system model and formulate a BS association problem together with preliminaries. In Section III, we explain ViSE, followed by performance evaluation in Section IV, and we conclude the paper in Section V.

II. Model and Preliminaries

A. Model and notations

We consider orthogonal frequency-division multiple access (OFDMA) based wireless cellular systems, composed of N BSs, K MSs, and total bandwidth B which is equally divided by J sub-carriers. Denote by $\mathcal{K} = \{1, ..., K\}, \mathcal{N} = \{1, ..., N\}$, and $\mathcal{J} = \{1, ..., J\}$, a set of MSs, BSs, and sub-carriers, respectively. We consider only *down-link* transmissions in the time-slotted system indexed by $t = 0, 1, \ldots$.

Each MS should be associated with only one BS $n \in \mathcal{N}$ at each slot. Let $\mathcal{K}_n$ be the set of MSs associated with the BS n. Then, we have that $\mathcal{K} = \cup_{i\in\mathcal{N}}\mathcal{K}_i$, and $\mathcal{K}_n \cap \mathcal{K}_m = \emptyset$, for $n \neq m$. We consider universal frequency reuse, i.e., all BSs can use all the sub-carriers for data transmission. We assume a fixed power allocation in BSs and do not consider dynamic power control schemes[3]. Let $\boldsymbol{I}(t) = (I_{k,j,n}(t) : k \in \mathcal{K}, j \in \mathcal{J}, n \in \mathcal{N})$ be a vector consisting of *BS association* and *user scheduling* indicators (scheduling indicator in short throughout the paper), where $I_{k,j,n}(t) = 1$ if MS k is scheduled on sub-carrier j by BS n, and 0 otherwise. We assume that the infinite amount of traffic destined for each MS is ready.

B. Objective and optimal algorithm

We now formulate an objective that includes the BS association as a key component. Studying this objective provides a guideline to practical BS association algorithms and is useful to understand the existing schemes as well as the scheme in this paper. The objective is described as maximizing the long-term network wide utility whenever possible, i.e., solving the following optimization problem:

$$\max \sum_{k\in\mathcal{K}} U_k(\bar{R}_k), \quad \text{s.t.} \quad (\bar{R}_1, ..., \bar{R}_K) \in \mathbf{\Lambda}, \tag{1}$$

where $\mathbf{\Lambda}$ is the throughput region that is the set due to time-multiplexing of all feasible long-term rate vectors across the user. Mathematically, $\mathbf{\Lambda}$ is the convex hull of the instantaneous rate regions, where an instantaneous rate region is defined for each channel state.

The $\bar{R}_k$ is the long-term throughput achieved by the MS k. We assume the standard conditions of differentiability and strictly increasing concavity of U_k. Of particular interest is the following α-fair utility function [14]: $U_k(x) = x^{1-\alpha}/(1-\alpha)$

[2]The authors propose a general α-optimal user association that can achieve rate-optimal ($\alpha = 0$), delay-optimal ($\alpha = 2$), and load-equalizing ($\alpha \to \infty$).

[3]Our algorithm in this paper does not require any assumption on dynamic power control for inter-cell interference (ICI) management (see e.g., [16]–[21]). We will discuss how our algorithm performs in presence of ICI management schemes in Section IV.

for $\alpha \neq 1$, and $U_k(x) = \log x$ for $\alpha = 1$. The α-fair utility function is known to encompass various popular fairness in literature, including proportional fairness ($\alpha = 1$) and max-min fairness ($\alpha \to \infty$).

Using the stochastic gradient-based technique in, e.g., [22], the optimal slot-by-slot algorithm of BS association and user scheduling can be expressed as the solution of the following per-slot optimization problem **OAS** (Optimal Association and Scheduling):

$$\textbf{OAS}: \quad \max_{\boldsymbol{I}(t)} \sum_{k \in \mathcal{K}} U'_k(\bar{R}_k(t-1)) \cdot r_k(t) \tag{2}$$

$$\text{s.t.} \quad I_{k,j,n}(t) \in \{0,1\}, \ \forall k, j, n, \tag{3}$$

$$\sum_{k \in \mathcal{K}} I_{k,j,n}(t) \leq 1, \ \forall j, n, \tag{4}$$

$$\sum_{n \in \mathcal{N}} \max_{j \in \mathcal{J}} I_{k,j,n}(t) \leq 1, \ \forall k, \tag{5}$$

where $r_k(t) = \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{J}} r_{k,j,n}(t) \cdot I_{k,j,n}(t)$ is the *actual* data rate assigned to user k at slot t, $r_{k,j,n}(t)$ is the *potential* rate to MS k from BS n on sub-carrier j at slot t, and $\bar{R}_k(t) = \frac{1}{t} \sum_{\tau=1}^{t} r_k(\tau)$ is the average long-term throughput for user k until slot t. The scheduling constraint in (4) is imposed since only one user can be scheduled in each BS on each sub-carrier. The association constraint in (5) is due to the fact that each user should be associated with at most one BS.

The **OAS** is hard to implement for the following reasons:

1) *Spatial hardness.* Solving **OAS** requires to examine all the possibilities of associating an MS to any arbitrary BS in the entire network (see $\mathcal{K}$ in (2)). This incurs too much overhead among MSs and BSs at each slot.
2) *High computational complexity.* Even if we assume that the information for solving **OAS** can be collected fast at some centralized coordinator, the algorithmic complexity to solve **OAS** is NP-hard, since the problem can be easily reduced to the maximum weight independent set problem [23], where $U'(\bar{R}(t-1))$ is the weight.
3) *Too frequent association changes.* Assuming that the complexity issues in the above are handled appropriately, **OAS** should allow MSs to change association at each slot. However, too frequent association change generates other system overheads, e.g., traffic re-routing and possible service disruptions.

Existing approaches for practical BS association [6], [8]–[10] can be regarded as approximating distributed heuristics at the cost of performance, incurring some gap from **OAS**. In tackling the issue of *3) too frequent association changes*, an easy approach is to have a *dwell time constraint* that specifies a maximum allowable frequency of association change. This dwell time constraint may have different forms, e.g., deterministic value or probabilistic average. The dwell time constraint enables a time-scale separation between user scheduling and BS association. Thus, when association decision is made, the past-histories are typically exploited, e.g., the channel gain etc., to make efficient association decision for the future. The dwell time constraint is used by the related research [8]–[10]. We will also adopt it in our scheme presented in Section III, and focus on the issues of 1) and 2).

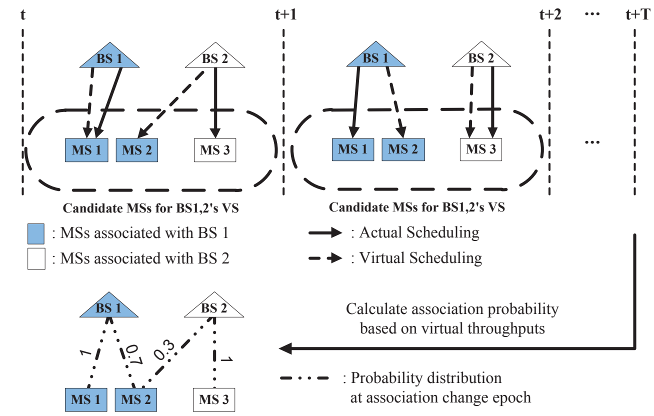


Fig. 1. An example of ViSE operation.

III. Virtual Scheduling based Emulation (ViSE)

A. Overview

The operation of ViSE consists of the following two key steps: *(i)* each MS *emulates* the optimal algorithm with practical complexity by employing a *virtual scheduler (VS)* on BSs and records the virtually achieved throughput from the VS, and *(ii)* at each association change epoch, each MS randomly selects a BS following the probability distribution based on the virtual throughputs. As the name implies, VS does not physically allocate resources to users, and just chooses users and notifies the scheduling results of the scheduled user through down-link control messages. Once a virtually scheduled user receives the result, the user updates per-BS virtual throughput during the past association period. Then, each user calculates the association probability for each BS (which virtually scheduled the user) in proportion to the virtual throughput. Each user then randomly selects its associated BS following the calculated probability for the next association period. The random decision may let an MS stay connected to the already-associated BS. Then, the MS can alter association at any time slot after the dwell time elapses without association change.

To illustrate, consider a simple scenario with two BSs and three MSs, as shown in Fig. 1. Before an association change epoch $t+T$, BS1 and BS2 run VS independently. Each MS memorizes the portion of achieved virtual throughput from each BS at time T. Suppose that the virtual throughputs of MS2 are 0.7 and 0.3 from BS1 and BS2, respectively. Note that over the current association period $[t, t+T]$, MS2 has been served only by BS1. Then, at the association epoch $t+T$, MS2 selects BS1 (resp. BS2) with probability of 0.7 (resp. 0.3).

B. Virtual scheduling

Virtual scheduling is designed to emulate **OAS**, yet with practical complexity, so that ViSE makes right BS association decisions following the direction of **OAS**. In this section, we present the key ideas towards our design goals of VS.

Relaxing OAS for decentralization

The centralized feature and high computational complexity of **OAS** come from the association constraint (5) that requires to carry out the exhaustive search to find optimal solutions.

Consider the following relaxed problem of **OAS** without the association constraint.

$$\textbf{Relaxed-OAS}: \quad \max_{\boldsymbol{I}(t)} \quad \sum_{k\in\mathcal{K}} U_k'(\bar{R}_k(t-1))\cdot r_k(t) \quad (6)$$

$$I_{k,j,n}(t)\in\{0,1\},\ \forall k,j,n, \quad (7)$$

$$\text{subject to} \quad \sum_{k\in\mathcal{K}} I_{k,j,n}(t)\le 1,\ \forall j,n. \quad (8)$$

We first show that the problem **Relaxed-OAS** can be decomposed into multiple sub-problems, enabling decentralization and significant complexity reduction, described in Lemma 3.1.

Lemma 3.1: The problem **Relaxed-OAS** is reduced to $J\times N$ independent sub-problems in which each BS n selects the MS $k^*_{j,n}(t)$ on each sub-carrier j, i.e.,

$$k^*_{j,n}(t)=\arg\max_{k\in\mathcal{K}} U_k'(\bar{R}_k(t-1))\cdot r_{k,j,n}(t). \quad (9)$$

Then, the resulting scheduling indicator vector $\boldsymbol{I}(t)$, i.e., $I_{k,j,n}(t)=1$ when $k=k^*_{j,n}(t)$, and 0 otherwise, is an optimal solution of **Relaxed-OAS**.

Proof: As $U_k'(\bar{R}_k(t-1))$ and $r_{k,j,n}(t)$ are given parameters, it suffices to investigate dependencies among $I_{k,j,n}(t)$. Since the constraint (8) for the given BS n and sub-carrier j does not affect the other BSs and sub-carriers at all, the objective function of **Relaxed-OAS** can be rewritten as follows:

$$\sum_{k\in\mathcal{K}} U_k'(\bar{R}_k(t-1))\cdot\Big[\sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{J}} r_{k,j,n}(t)\cdot I_{k,j,n}(t)\Big]$$
$$=\sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{J}}\Big[\sum_{k\in\mathcal{K}} U_k'(\bar{R}_k(t-1))\cdot r_{k,j,n}(t)\cdot I_{k,j,n}(t)\Big].$$

Accordingly, the relaxed problem can be decomposed and is equivalent to individually solving the $N\times J$ user scheduling sub-problems given by (9) for each BS and sub-carrier. ■

At the cost of performance gap from **OAS**, our first design principle of VS is to solve **Relaxed-OAS** which becomes more practical just by locally running the intra-cell user scheduling in (9) at each BS. It is not easy to quantify the sub-optimality gap, but it does not seem to be significant due to the following reasons: Since the number of edge users are not large in a practical situation where users are distributed randomly, the probability that a user is scheduled by multiple BSs is very low, and we do not actually schedule users for **Relaxed-OAS** in VS, but just to obtain the long-term "virtual" throughput for BS association decisions.

Reducing spatial and temporal overheads

However, the decentralized intra-cell user scheduling in (9) still has heavy signaling overheads in both spatial and temporal senses: Each BS needs to collect information (e.g., instantaneous rate $r_{k,j,n}(t)$ and the gradient of their utility $U_k'(\bar{R}_k(t-1))$) from all the users in the network. Moreover, this feedback should be exchanged at each slot. To develop VS that is practical, the first idea is to restrict the candidate MSs for VS to "local" users, not all users for alleviating spatial overheads. The second idea is to exchange the averaged feedback, not the instantaneous one for reducing temporal overheads. We first describe Virtual Scheduling[4], followed by more details.

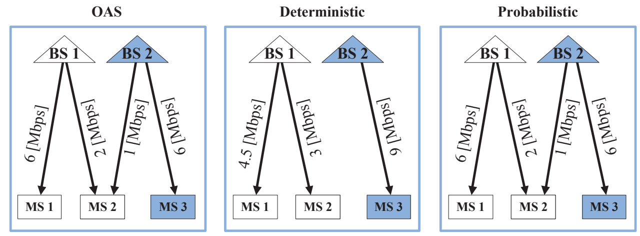


Fig. 2. Load balancing with fine granularity.

Virtual scheduling (VS):
Each BS n selects the MS $k^{(v)}_{j,n}(t)$ on sub-carrier j based on the exponential time-averaged instantaneous rate $\overline{r}_{k,j,n}(t)$ among the reduced set of candidate users $\tilde{\mathcal{K}}_n$ as follows[5]:

$$k^{(v)}_{j,n}(t)=\arg\max_{k\in\tilde{\mathcal{K}}_n} U_k'(\bar{R}^{(v)}_k(t-1))\cdot\overline{r}_{k,j,n}. \quad (10)$$

The total virtual throughput $\bar{R}^{(v)}_k$ is the summation of the virtual throughput provided by each BS, i.e., $\bar{R}^{(v)}_k=\sum_{n\in\mathcal{N}}\bar{R}^{(v)}_{k,n}(t)$, where

$$\bar{R}^{(v)}_{k,n}(t)=(1-\beta)R^{(v)}_{k,n}(t-1)+\beta\sum_{j\in\mathcal{J}}\overline{r}_{k,j,n}(t)I^{(v)}_{k,j,n}(t).$$

Here, the $0<\beta<1$ is a constant and we denote by $I^{(v)}_{k,j,n}(t)$ the indicator that reflects the scheduling result of VS; $I^{(v)}_{k,j,n}(t)=1$ if MS k is selected by the VS of BS n on sub-carrier j, and 0 otherwise.

First, to reduce the spatial feedback, rather than considering all users $\mathcal{K}$ in the network as the candidates, a VS at each BS only considers the set of local users $\tilde{\mathcal{K}}_n$, which consists of users associated with BS n and its neighboring BSs $\mathcal{N}(n)$, where

$$\mathcal{K}\quad\Rightarrow\quad\tilde{\mathcal{K}}_n=\mathcal{K}_n\cup_{m\in\mathcal{N}(n)}\mathcal{K}_m. \quad (11)$$

Since it is unlikely for VS to select the users in other far-field BSs due to their low values of instantaneous rate, this approximation will not degrade the performance significantly. It is also understandable from users' perspective because a handover typically occurs among the neighboring cells.

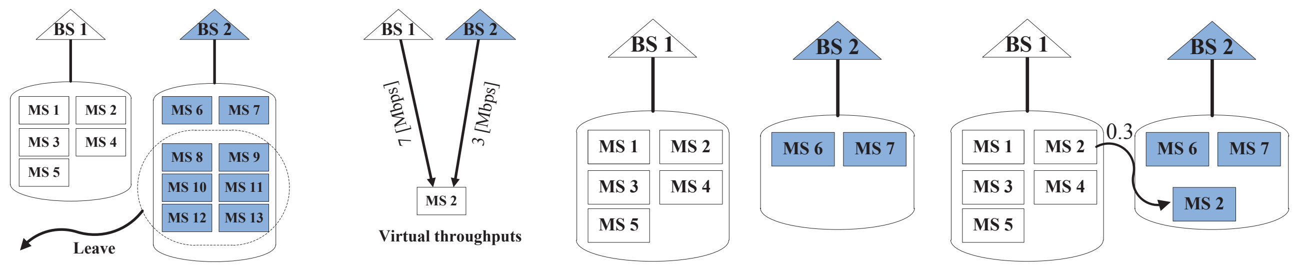
Second, to reduce the temporal feedback, we exchange the following averaged feedback infrequently rather than instantaneous feedback at each slot:

$$r_{k,j,n}(t)\quad\Rightarrow\quad\overline{r}_{k,j,n}. \quad (12)$$

The use of the average, infrequent feedbacks is from the intuition that instantaneous information is not critically necessary, because VS does not actually schedule. Instead, it may suffice that VS follows the macroscopic conditions for obtaining long-term virtual throughputs. There might be

[4]We use the superscript (v) to refer to the values from Virtual Scheduling throughout the paper.

[5]Our VS is originally developed from the fixed set of MSs. However, even if the set of MSs is time-varying (i.e., dynamic MS arrivals and departures), it still works well as will be shown later in subsection IV-D.



(a) Two cell network under flow dynamics (b) Deterministic (c) Probabilistic

Fig. 3. The robustness of probabilistic decision to flow dynamics.

a small loss of precision in calculating the virtual throughput due to slow tracking of instantaneous channel gains. We will present the impact of infrequent feedback in subsection IV-B.

C. BS association decision

Given the result of VS, each MS determines its BS for association. Note that the results by VS are: from which BS, each MS has virtually been scheduled with what rates.

In ViSE, each MS computes the BS association probability distribution that will provide a probability to choose a BS. The probability to be associated with the BS n is set to be proportional to the virtual throughput of BS n, given by:

$$P_{k,n}(t) = \frac{\bar{R}^{(v)}_{k,n}(t)}{\sum_{n\in\mathcal{N}} \bar{R}^{(v)}_{k,n}(t)}, \qquad (13)$$

The intuition is that sticking to the BS providing higher virtual throughput is likely to be more beneficial. According to the association probabilities in (13) for all candidate BSs n, each user randomly determines its associated BS for the next association period. We remark that in the closed-form approaches, BS association is decided deterministically based on the estimated throughput from a closed-form equation. We illustrate why a probabilistic decision is better for improving performance using two examples.

Finer granularity load balancing. Consider a two-cell network with three MSs as depicted in Fig. 2, where the numbers represent the average throughputs received on each link under different channel conditions. It is worthwhile to mention that, in **OAS**, MS2 in the middle receives data both from BS1 and BS2, while MS1 (resp. MS3) receives data from only BS1 (resp. BS2). In a deterministic strategy, MS2 will choose BS1, since from the result of **OAS**, BS1 provides more rates to MS2 than BS2 (2 Mbps vs. 1 Mbps), whereas in the probabilistic strategy, MS2 will be randomly change its association between BS1 and BS2, according to the probability distribution from the result of **OAS**. When the system experiences the similar channel conditions over multiple association periods, the probabilistic strategy guarantees the achieved throughput closer to the result of **OAS** than the deterministic one.

Robustness to dynamics. We now provide an example that the probabilistic decision helps with responding to network dynamics. Fig. 3(a) shows the scenario that over some association period, more MSs are associated with BS2 than BS1, and around an association epoch, 6 MSs move and become further from BS2. In this case, at each association epoch, the baseline throughput (virtual throughput in ViSE or estimated throughput in closed-form approaches) is still unaffected by such mobility significantly, because it is an average over the dwell time $T \gg 1$. In the deterministic strategy (Fig. 3(b)), MS2 will choose BS1, because MS1 believes that BS1 will provide 7 Mbps, which is larger than 3 Mbps by BS2, clearly being a wrong decision in terms of load balancing. The deterministic strategy may also lead to "ping-pong" effects [24] due to unwanted congestion, where in an abrupt change of user arrivals/departures, many users try to change their associations to a same BS at a similar time. However, in the probabilistic strategy (Fig. 3(c)), we have some possibility of choosing BS2 (with prob. of 0.3 in this example), so that we alleviate the ping-pong effect as well as improve the performance due to better load balancing.

We complement the probabilistic algorithm to avoid "unnecessary" association changes. From our simulation results in subsection IV-C, we found that these "unnecessary" association changes do not provide a big performance improvement, yet incurring heavy overheads. As an example, a naive application of our probabilistic association change allows a "center user" located close to its BS to be associated with a neighboring BS, even if its probability is small. Note that the major gain from smart association change is due to the users at the cell edges. Thus, we employ a threshold-based method, where if $P_{k,n} < \delta$, for some threshold δ, then we set $P_{k,n} = 0$. We scale up other probabilities, so that $\sum_n P_{k,n} = 1$ for all k. The threshold δ balances the trade-off between performance and overhead (i.e., the frequency of association changes).

D. User scheduling

Once a BS association is given, a user set $\mathcal{K}$ is equivalent to the disjoint union of $\mathcal{K}_n$, for all $n \in \mathcal{N}$, i.e., $\mathcal{K} = \mathcal{K}_1 \cup \cdots \cup \mathcal{K}_n$, and thus, a constraint of **OAS** that each user should be associated with at most one BS (5), can be removed. Then, **OAS** becomes equivalent to the following per-BS and per-slot optimal scheduling problem:

$$\begin{aligned} &\max_{\boldsymbol{I}(t)} \sum_{k\in\mathcal{K}} U'_k(\bar{R}_k(t-1)) \cdot r_k(t) \\ &= \sum_{n\in\mathcal{N}} \max_{\boldsymbol{I}_n(t)} \sum_{k\in\mathcal{K}(n)} U'_k(\bar{R}_k(t-1)) \cdot r_k(t), \end{aligned}$$

where $\boldsymbol{I}_n(t) \doteq [I_{k,j,n}(t) : k \in \mathcal{K}, j \in \mathcal{J}]$ is the user scheduling indicator vector for a BS n.

Then, the optimal scheduling algorithm for each BS is to schedule a user on each sub-carrier as described in the following:

$$k^*_{j,n}(t) = \arg \max_{k \in \mathcal{K}(n)} U'_k(\bar{R}_k(t-1)) \cdot r_{k,j,n}(t), \quad \forall j, n. \quad (14)$$

E. System requirements for implementation

In order to implement ViSE, signaling message exchange is unavoidable, yet its implementation does not seem to be very challenging, from MSs to BSs, from BSs to MSs, and between BSs, as explained as follows.

Each MS k can measure the per-BS instantaneous rate by hearing down-link pilot signals from multiple BSs. An MS needs to be reported the VS's scheduling results by its associated BS and neighboring BSs. This signaling can be done through down-link control message (e.g., DL-map) at the head of slot. In order to decode DL-maps from multiple BSs, synchronization between BSs may be required. This assumption is reasonable in many current and next generation TDD (Time Division Duplex) systems, such as UTRA-TDD and mobile WiMax (802.16m) [25]. Note that the instantaneous rate for the current associated BS n needs to be reported at each slot for actual scheduling, irrespective of ViSE. In ViSE, as mentioned in subsection III-B, the MS periodically reports the average instantaneous rates from neighboring BSs as well as long-term throughput to its associated BS n for Virtual Scheduling. Then, BS n sends these information to its neighboring BSs $\mathcal{N}(n)$ over a wired backhaul. Using these feedback information, the BS runs VS whose results are informed to the neighboring BSs for calculating the updated weights (the derivatives of the utility), again over a high speed wired backhaul.

We further analyze the signaling overhead of ViSE. Since ViSE uses the average potential rate $\overline{r}_{k,j,n}$ and considers only local users $\tilde{\mathcal{K}}_n$ for virtual scheduler, we can reduce the overhead of a BS n from $O(|\mathcal{N}||\mathcal{K}||\mathcal{J}|)$ to $O(\frac{|\mathcal{N}(n)|\mathcal{K}_n||\mathcal{J}|}{T_f})$ where T_f and $\mathcal{N}(n)$ represent the duration of average information change and the set of neighboring BSs of BS n, respectively.

F. Comparison: ViSE vs. Closed-Form Approach

The work [8]–[10] on efficient BS association by relaxing spatial hardness and high computational complexity, referred to as *closed-form approaches*, is related to our paper. The key idea is to let each MS *locally*, *independently*, and *deterministically* select the BS association, based on the estimation of *"how much rate I will receive if I change my association from the current BS to one of my neighboring BSs."* Then, each MS changes its association to the BS that is estimated to provide the highest rate. The individual estimated rate for the MS k at the BS n is:

$$\frac{\text{Potential estimated rate for MS } k}{\text{Num. of MSs at BS } n}. \quad (15)$$

The estimated rate is regarded as an expectation in the future, using the measured rate over the past. The key underlying assumption of (15) is the PF user scheduler having the property of *temporal fairness*. In *temporal fairness*, the PF scheduler guarantees the equal share of the service time among the MSs in the corresponding cell, allowing a simple throughput equation. We refer the readers to [8]–[10] for more details.

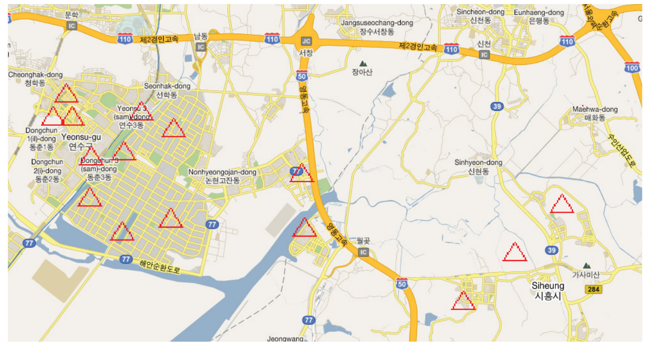


Fig. 4. Real BS deployment topology.

ViSE is designed to overcome the following two limitations of the closed-form approaches. First, the key feature that enables a closed-form equation is the feature of *temporal fairness* in the PF. The PF corresponds to $\alpha = 1$ in the α-fairness [14]. Unfortunately, for $\alpha \neq 1$, the closed-form equation is hard to compute. Second, there may be highly dynamic scenarios in terms of flow arrivals/departures and MSs' mobility. Under such dynamics, the estimated throughput based on the past, which provides a guideline to selecting an associated BS for the future, may be far from the current as well as the future. The problem of the closed-form approaches lies in the *deterministic* selection of the future BS, where the wrong decision, due to the big difference between the past and the future, would be sustained for a long time. We tackle these two problems with *emulation* by Virtual Scheduling and *probabilistic* association changes.

IV. Performance Evaluation

We evaluate the performance of ViSE through extensive simulations under various scenarios, including a real BS deployment topology. First, we verify the superiority of ViSE to other tested algorithms under general α-fairness in subsection IV-A and examine the effect of the infrequent feedbacks and the threshold value for the probabilistic BS association in subsections IV-B and IV-C. Second, we test the robustness of ViSE to flow arrivals/departures and mobility as well as the dynamic power control for interference management in subsections IV-D and IV-E, respectively.

For our simulations, we consider a real BS deployment (15 BSs in 14x9 km^2) in Incheon city, South Korea (see Fig. 4), of a major cellular service provider. We generate users one-by-one and attach them to the closest BS until each BS has 10 users. Our user generation is based on the assumption that the number of BSs per unit area is proportional to the user density. In other words, the average number of users per cell is almost similar because BSs in an urban environment cover a small area and BSs in a rural environment a large area.

For most of our simulations, we consider the fixed maximum transmission power 43dBm for all BSs and allocate the power evenly to all the sub-carriers. In subsection IV-D, we consider dynamic (time-varying & frequency selective) power allocations for interference management by adopting ICI management techniques [16]–[21]. In modeling the propagation environment, a path loss $16.62 + 37.6 \log_{10}(d[m])$ and log-normal shadowing with a standard deviation σ_s=8 dB are

TABLE I

AGGREGATE UTILITY OF DIFFERENT ALGORITHMS BY VARYING α. SEVERAL MEANINGFUL THROUGHPUT METRICS IN [MBPS] ARE ALSO PROVIDED: THE SUM, GEOMETRIC AVERAGE AND MINIMUM OF MS THROUGHPUTS FOR $\alpha = 0.1$, $\alpha = 1$ AND $\alpha = 8$ IN PARENTHESES ( ), RESPECTIVELY; JAIN'S FAIRNESS INDEX IN SQUARE BRACKETS [ ]

α	0.1	1	8
Max-SINR	1.41e8 (722.0) [8.67e-2]	2.06e3 (0.912) [2.45e-1]	-3.13e-36 (0.124) [2.13e-1]
Closed-form	1.41e8 (722.7) [8.67e-2]	2.08e3 (1.041) [2.92e-1]	-1.09e-38 (0.260) [3.22e-1]
ViSE	1.42e8 (722.8) [8.71e-2]	2.08e3 (1.046) [2.93e-1]	-8.97e-39 (0.305) [4.02e-1]
Relaxed-OAS	1.42e8 (723.4) [8.74e-2]	2.09e3 (1.094) [3.57e-1]	-1.13e-39 (0.443) [9.17e-1]

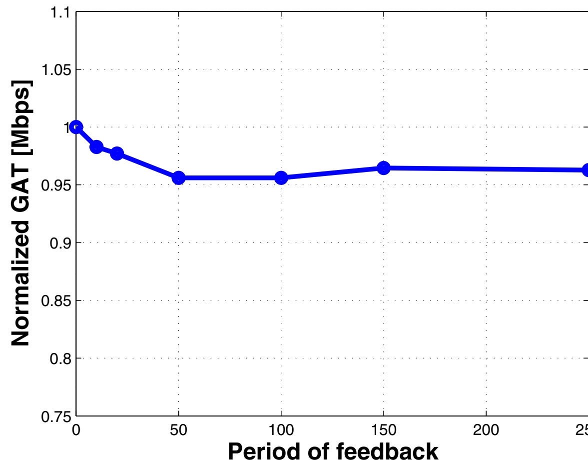


Fig. 5. The effect of infrequent feedback information (α=1).

adopted. The frequency selective fading is also captured by adopting an uncorrelated fading channel model [26]. Once the transmit powers and all the channel gains are determined, the achievable data rates for users are simply calculated by Shanon's formula.

To make a fair comparison between the closed-form approach and ViSE, we basically consider $\alpha = 1$ (i.e., proportional fairness) except the case where general α-fairness is tested in subsection IV-A. This allows us to solely present the advantages that can be achieved from emulation and probabilistic decision. Tested algorithms are *(a)* Max-SINR scheme, *(b)* a closed-form based algorithm [9][6], and *(c)* **Relaxed-OAS**. In order to investigate gap from the optimality, we examine the difference from **Relaxed-OAS**, instead of **OAS**, due to its prohibitive complexity. However, since **Relaxed-OAS** always increase the objective value of **OAS** by ignoring the constraint (5) that a user is scheduled by multiple BSs is very low[7], we believe that it will give a tight upper-bound for all other tested algorithms.

A. Applicability to general fairness

TABLE I shows the aggregate utility for various α ranging from 0.1 to 8. When α goes to zero, the objective is to maximize the sum of MS throughputs, in which case it is enough to simply associate an MS to the closest BS, thus there is little difference in performance across the tested algorithms. However, as α increases (i.e., enforcing more fairness), Max-SINR and the closed-form approaches become far from optimal. Note that ViSE can always achieve performance even closer to that of **Relaxed-OAS** than other algorithms, regardless of α.

Beyond the unitless values of utilities, we also provide real throughput metrics in [Mbps]. For example, the values can be found in parentheses of TABLE I: sum of MS throughputs for $\alpha = 0.1$, geometric average of MS throughputs (GAT) for $\alpha = 1$, and minimum of MS throughputs for $\alpha = 8$. We use these metrics since maximizing this metric is equivalent to our system objective when $\alpha \rightarrow 0$ (maximizing sum throughput), $\alpha = 1$ (proportional fairness) and $\alpha \rightarrow \infty$ (max-min fairness), respectively. Further, we provide Jain's fairness index in square brackets of TABLE I.

B. Impact of infrequent feedback

In order to reduce temporal overheads, in ViSE, the virtual scheduler is designed to utilize the time-averaged potential rate in subsection III-B. We test the impact of time-averaged information, i.e., $\overline{r}_{k,j,n}$ for various feedback periods T_f, which is the same as the averaging period. We take into account $\overline{r}_{k,j,n}$ for this simulation as follows.

$$\overline{r}_{k,j,n} = (1 - \frac{1}{T_f})\overline{r}_{k,j,n}(t-1) + \frac{1}{T_f} r_{k,j,n}(t) \qquad (16)$$

Fig. 5 shows the normalized GAT for different feedback periods. In particular, the case of (period of feedback = 1) is the instantaneous feedback at every slot. As the period of feedback increases, the performance tends to decrease due to information inaccuracy. However, the GAT no longer decreases after 50 slots and the performance degradation is marginal, less than 4.5%. In other words, we can expect that the proposed algorithm achieves at least 95% of the performance, even with infrequent feedback, compared to that with instantaneous feedback.

C. Effect of threshold value

As mentioned in subsection III-C, we employ a threshold to prevent unnecessary association changes. Fig. 6 shows its impact on GAT and the number of association changes, where recall that we use $\alpha = 1$.

As expected, when a threshold value is large, the averaged number of association changes (i.e., handovers) for an MS is small, and the GAT performance also degrades. This is because only a small portion of MSs are considered, and accordingly, the association change becomes too sluggish and even it is impossible to track the optimal algorithm. However,

[6]We tested other closed-form based algorithms as well, but the performance difference was not significant. Thus, we only include the results of [9] for brevity of presentation.

[7]The average probability that the schedulers of multiple BSs simultaneously choose a user is about 0.118.

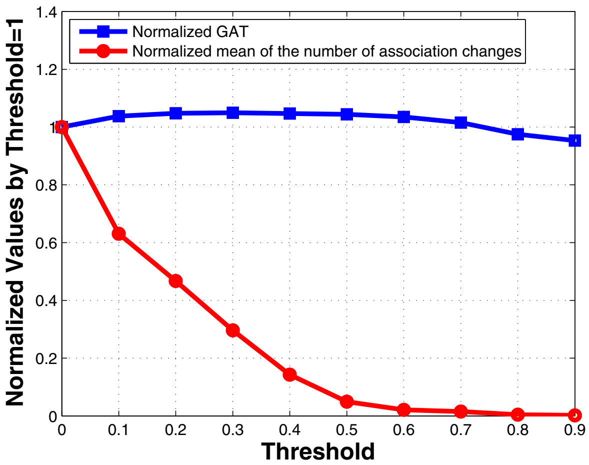
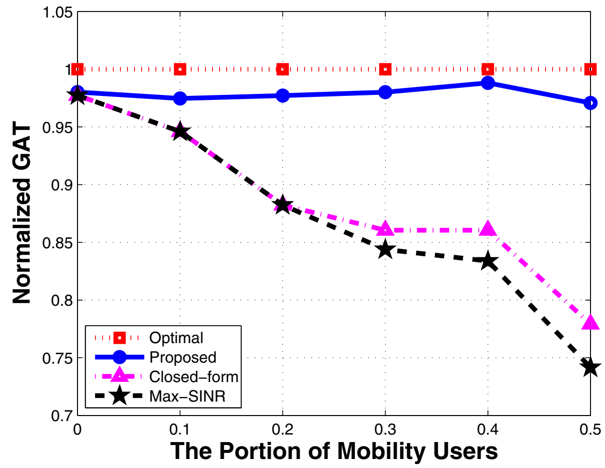
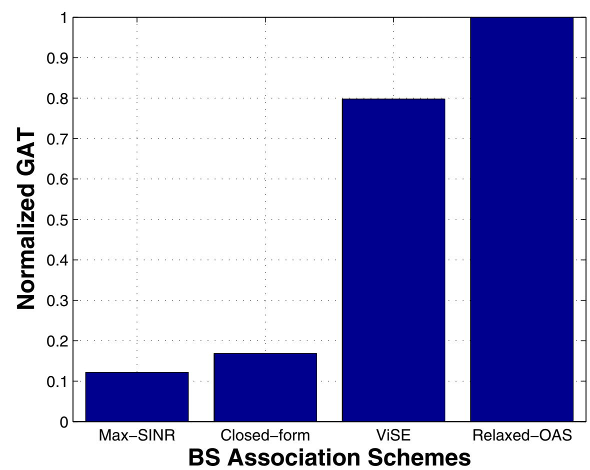


Fig. 6. Impact of threshold on GAT and the number of association changes.



(a) Mobility



(b) Flow arrivlas and departures

Fig. 7. Robustness to network dynamics: (a) mobility and (b) flow arrivals and departures.

we also observe that too small threshold values degrade the GAT performance. This is because an MS can be associated with a BS that provides low throughput, and this "wrong" association change lasts for at least the dwell time, e.g., an MS at the cell center can change its association to a neighboring cell. Therefore, as shown in Fig. 6, it is important to choose a proper threshold value, which is between 0.2 and 0.5 from our simulation results. In Fig. 6, each line represents (i) GAT and (ii) the number of association changes for an MS, and each of these is normalized for the case of threshold=1, respectively. We set the threshold value of 0.3 for our other simulations.

D. Network dynamics: Mobility and flow arrivals/departures

Mobility. We consider a general 7 hexagonal BSs topology with uniformly distributed 70 MSs. In modeling mobility, a random waypoint model with 72 km/h velocity is adopted. We vary the portion of MSs with mobility from 0% to 50%. As shown in Fig. 7(a), ViSE maintains near-optimal performance (about 97%—98% of Relaxed-OAS constantly for the tested degree of mobility), but other algorithms experience performance degradation as mobility increases. This is because the deterministic BS selection should sustain for a long time, even though it made a wrong decision due to the big difference between the past and the future by mobility. Note that for $\alpha = 1$, this performance gap is just due to a better tracking of the optimality and probabilistic selection for BS association change.

Flow dynamics. We also test robustness to flow-level dynamics whose results are shown in Fig. 7(b). To model flow arrival/departure simply, starting from an initial set of MSs, we add an MS in a random location with probability p and remove one of the existing MSs with probability q at each slot. We set the probabilities $[p, q]$=[1.25%, 0.75%]. ViSE achieves the performance of 80% close to **Relaxed-OAS** in terms of normalized GAT, whereas the closed-form approach achieves about 20%, which is as low as the performance of Max-SINR scheme.

E. Robustness to interference management

In Section II-A, we assume that all BSs are set to use a fixed transmit power. However, BSs in the near future are likely to adopt more intelligent power control to increase the efficiency of spectrum sharing, called ICI (Inter-Cell Interference) management. In this subsection, we test the performance of BS association schemes when such ICI management schemes are employed. In literature, there are two types of BS power control algorithms, depending on the time-scale of operation: fast power control [16], [17], [20] and slow power control [18], [19], [21]. In fast power control, it is assumed that the feedback messages for interference management are possible to be exchanged in the order of time-slots, whereas in slow power control, BS powers are slowly tracking the system dynamics. In our simulation, we use [19] and [20] for slow and fast power control, respectively.

As shown in Fig. 8, ViSE generally tracks **Relaxed-OAS** well, especially for the case of slow power control and fixed power. In fast power control, since transmit powers are determined based on which MSs are actually (not virtually)

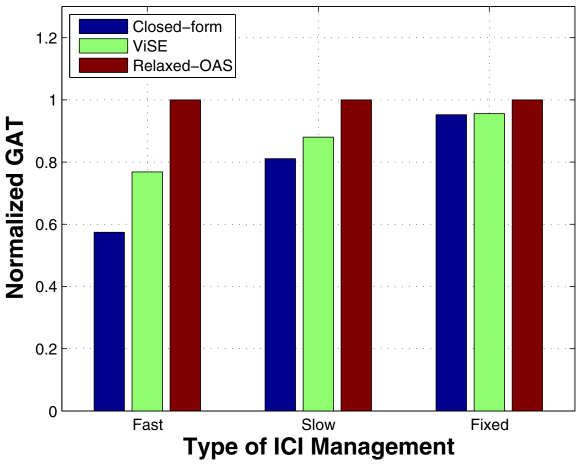


Fig. 8. Impact of ICI management schemes: fast and slow time-scale power control compare to fixed power.

scheduled, the emulation via VS would have a certain inaccuracy, which degrades the normalized GAT performance. However, ViSE still performs much better than the closed form approach.

V. Conclusion

In order to balance the load among BSs towards fairness and efficiency, many algorithms on BS association have been proposed. Existing algorithms typically associate a MS with a BS which provides the maximum expected throughput that is induced by temporal fairness in PF. However, these algorithms are not able to achieve good performance for general α-fairness, and under network dynamics. In this paper, we propose a novel approach based on emulation and probabilistic BS selection. The main idea of our algorithm is to emulate the approximated version of the optimal algorithm and to determine BS association based on statistical probability from the past emulation. Extensive simulations demonstrate that ViSE can adjust well to the user schedulers with general fairness and network dynamics.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments that greatly improved the quality of this paper.

References

[1] Cisco Systems Inc., "Cisco visual networking index: global mobile data traffic forecast update, 2011-2016." Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/\\ ns705/ns827/white_paper_c11-520862.pdf, Feb. 2012.

[2] "Data, data everywhere." Available: http://korea.emc.com/\\collateral/ analyst-reports/ar-the-economist-data-data-everywhere.pdf, Feb. 2010.

[3] 3GPP TR 36.814, "Futher advancement for e-utra - physical layer aspects," Feb. 2009.

[4] IEEE 802.16m-09/0034, "IEEE 802.16m system description document [draft]," July 2009.

[5] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, O. S. Taesang Yoo, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, June 2011.

[6] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. 2003 IEEE Infocom*.

[7] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[8] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," in *Proc. 2004 ACM Mobicom*.

[9] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. 2006 IEEE Infocom*.

[10] K. Son, S. Chong, and G. de Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3566–3576, July 2009.

[11] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Automatic Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[12] IEEE 802.16m-07/002r8, "IEEE 802.16m system requirements," Jan. 2009.

[13] 3GPP2 C.R1002-0 v1.0, "CDMA2000 evaluation methodology," Dec. 2006.

[14] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[15] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "α-optimal user association and cell load balancing in wireless networks," in *Proc. 2010 IEEE Infocom Mini-Conference*.

[16] K. Son, S. Lee, Y. Yi, and S. Chong, "Practical dynamic interference management in multi-carrier multi-cell wireless network: a reference user based approach," in *Proc. 2010 WiOpt*.

[17] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multicell OFDMA networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 6, pp. 2835–2848, July 2009.

[18] K. Son, Y. Yi, and S. Chong, "Adaptive multi-pattern reuse in multi-cell networks," in *Proc. 2009 WiOpt*.

[19] A. L. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination," in *Proc. 2009 IEEE Infocom*.

[20] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: a practical interference management in heterogeneous wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1260–1272, June 2011.

[21] K. Son, Y. Yi, and S. Chong, "Utility-optimal multi-pattern reuse in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 142–153, Jan. 2011.

[22] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, Jan. 2005.

[23] T. H. Corman, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd edition. McGraw-Hill Book Company, 2001.

[24] G. P. Pollini, "Trends in handover design," *IEEE Commun. Mag.*, vol. 34, no. 3, pp. 82–90, Mar. 1996.

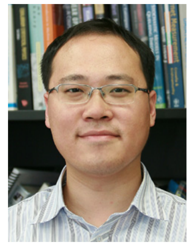
[25] "IEEE p802.16m-2007 draft standards for local and metropolitan area networks part 16: air interface for fixed broadcast wireless access systems," 2007.

[26] K. Fazel and S. Kaiser, *Multi-Carrier and Spread Spectrum Systems*. John Wiley & Sons, 2003.

Soohwan Lee (S'11) received his B.S. degree in the School of Electrical Engineering and Computer Science from Kyungpook National University, South Korea, in 2009, and his M.S. degree in the Department of Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2011. He is currently a Ph.D student in the Department of Electrical Engineering at KAIST. His current research interests include interference management in heterogeneous cellular networks, green wireless networking, and security management in cellular networks.

Kyuho Son (S'03-M'10) received his B.S., M.S. and Ph.D. degrees all in the Department of Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2002, 2004 and 2010, respectively. From 2010 to 2012, he was a post-doctoral research associate in the Department of Electrical Engineering at the University of Southern California, CA. He is currently a senior engineer at T-Mobile, USA. His current research interests lie in the design, analysis and optimization of wireless networks, cognitive radio systems and smart grid. He is a founding member of TSGCC (Technical Subcommittee of Green Communications and Computing) within IEEE Communications Society.

Huazhi Gong graduated in Information and Telecommunication from Xian Jiaotong University, Xian, China, in 1999, and received the Master Degree in Information and Communication from Huazhong University of Sci&Tech, Wuhan, China, in 2003. In 2003-2008, he was with the Department of Information and Communication at Gwangju Institute of Sci&Tech, Gwangju, Korea. Since Sept. 2009, he has worked as post-doctoral fellow in Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests are in the area of wireless network management, and P2P networks.



Yung Yi (S'04-M'06) received his B.S. and the M.S. in the School of Computer Science and Engineering from Seoul National University, South Korea in 1997 and 1999, respectively, and his Ph.D. in the Department of Electrical and Computer Engineering at the University of Texas at Austin in 2006. From 2006 to 2008, he was a post-doctoral research associate in the Department of Electrical Engineering at Princeton University. He is now an associate professor at the Department of Electrical Engineering at KAIST, South Korea. He has been serving as a TPC member at various conferences including ACM Mobihoc, Wicon, WiOpt, IEEE Infocom, ICC, Globecom, ITC, and WASA. His academic service also includes the local arrangement chair of WiOpt 2009 and CFI 2010, the networking area track chair of TENCON 2010, the publication chair of CFI 2011-2012, and the symposium chair of a green computing, networking, and communication area of ICNC 2012. He has served as a guest editor of the special issue on Green Networking and Communication Systems of *IEEE Communications Surveys and Tutorials*, an associate editor of *Elsevier Computer Communications Journal*, and an associate editor of the *Journal of Communications and Networks*. He has also served as the co-chair of the Green Multimedia Communication Interest Group of the IEEE Multimedia Communication Technical Committee. His current research interests include the design and analysis of computer networking and wireless systems, especially congestion control, scheduling, and interference management, with applications in wireless ad hoc networks, broadband access networks, economic aspects of communication networks, and greening of network systems.