Spectrum Breathing and Cell Load Balancing for Self Organizing Wireless Networks

Hongseok Kim,[†] Hyea Youn Kim,[†] Yunhee Cho,[‡] and Seung-Hwan Lee[‡]
[†]Department of Electronic Engineering, Sogang University, Seoul, Korea
[‡]Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
Email: {hongseok, hyeayoun}@sogang.ac.kr, {yhcho, lsh}@etri.re.kr

Abstract—**In this paper, we develop a self organizing mechanism for spectrum breathing and user association in cellular networks employing fractional frequency reuse. In doing this we specifically focus on flow-level cell load balancing under spatially inhomogeneous traffic distributions. Our work adaptively changes the spectrum bandwidth of each base station (BS) so that spectrums of the BSs breathe in and out to balance the loads of the BSs. Spectrum breathing is further combined with α-optimal user association for better load balancing. Our problem is challenging in the sense that the problem is not necessarily a convex optimization. To tackle the difficulty we decouple spectrum breathing and user association by introducing the concept of reference load, which is independent of the spectrum bandwidth. We propose an iterative algorithm that always converges to a fixed point, which is possibly an optimal solution. We verify that the proposed algorithm achieves the optimality by performing exhaustive search. Our extensive simulations demonstrate that the proposed algorithm significantly improves the system performs: decreasing the delay more than 10 times or increasing the admittable traffic load by more than 125%.**

I. Introduction

In the upcoming wireless cellular systems such as LTE-Advanced or Beyond 4G systems, cells are densely deployed to meet the growing demands for connectivity and thus intercell interference becomes more critical [1], [2]. Hence, intensive research is going on to mitigate the intercell interference. From information theoretic perspective, it is well known that allowing all cells to operate in the same spectrum bandwidth (called *full frequency reuse*) maximizes the total system capacity. However, full frequency reuse may excessively degrade the performance of cell edge users unless intercell interference is properly managed. To this end many intercell interference coordination (ICIC) techniques were proposed, see [3] for a series of research. However, ICIC techniques assuming full frequency reuse often require complex coordination, scheduling and/or message passing among base stations (BSs) and mobile terminals (MTs) [4], [5].

Hence, many of practical solutions are still based on *fractional frequency reuse* (FFR) [3], [6]–[15]. FFR basically lets adjacent cells (or sectors[1]) use a *disjoint* set of frequency bands, and its several variants exist. For example, the inner band for the cell center region employ full frequency reuse while the outter band for cell edge region employ partial frequency reuse. Transmit power of frequency bands may depend on whether the band is inner or outter [6], [7], [10]. The frequency partitioning ratio between inner and outter bands were adjusted to maximize the cell capacity [11]. In [12], the performances of various FFR techniques were compared by extensive simulations.

However, none of prior work so far, e.g., [3], [6]–[12], [14] directly considered the imbalance of cell loads in partitioning the frequency bands, and thus they may not perform well under spatially inhomogeneous traffic distribution. For better cell load balancing, [15] recently proposed an idea of adjusting the frequency bandwidth when each cell sees different load. However, they only focused on the allocation patterns of frequency bands for minimizing interference; they neglected how to determine the *optimal* frequency bandwidth partitioning in maximizing the system performance.

In this paper, we investigate the way of determining optimal frequency bandwidth partitioning and how they are related with the cell loads. We call it *spectrum breathing* because more load in one cell adaptively increases the frequency bandwidth (or spectrum bandwidth hereafter) of that cell by borrowing the spectrum from other cells. Our approach is reasonable when intercell interference can be roughly considered as static noise, which is indeed the case with fractional frequency reuse. We focus on developing a theory and algorithms of spectrum breathing that adapt to *spatially inhomogeneous* traffic distribution. Specifically, we consider stochastic traffic loads where new file transfers, or equivalently *flows*, are initiated at random and leave the system after being served; this is typically referred to as *flow-level dynamics* [16]–[19]. Unlike the previous work of FFR [7], [8], [10]–[12], [15] where static user population is assumed, our work is unique in capturing the impact of *dynamic* user population on the long term average performance, which is closer to user-perceived performance, e.g., file download time. Note that user experience is based on the *time average* performance rather than the *instantaneous* capacity. Analyzing such systems is, however, more challenging because of the dynamic nature of the system, and its analysis is hardly tractable in general.

$\mathcal{W}_i$

Contributions: We highlight the contributions of this paper as follows.

First, we provide a theoretical framework for cell load

This research was supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2012-11-921-04-001 performed by ETRI).

[1] For simplicity we will use the terminology *cells* even for *sectors* unless otherwise specified.

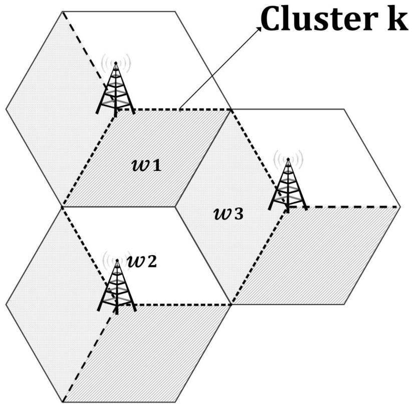

Fig. 1. Spectrum partitioning among BSs in the same interfering group.

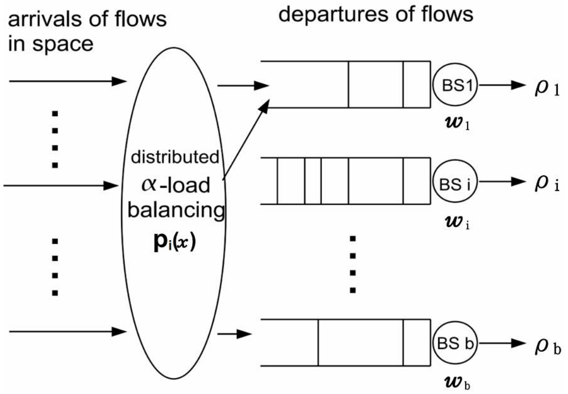

Fig. 2. Flow-level queueing model for user association problem.

balancing with spectrum breathing and user association where spectrum bandwidths adaptively change in a *self-organizing* manner according to spatially inhomogeneous traffic distribution in wireless networks with FFR. We formulate spectrum breathing with user association as an optimization problem. Using flow-level dynamics our model captures the spatial variability (inhomogeneous traffic distribution) as well as the temporal variability (dynamic user population) to better understand the real systems.

Second, we propose algorithms to solve the nonconvex optimization problem. Since the objective function is not generally convex, achieving optimality is challenging. By exploiting the *conditional* convexity, our algorithm converges to a fixed point that is possibly an optimal point. By performing exhaustive search we verify that the proposed algorithm achieves optimality though not guaranteed. The proposed algorithm is simple and thus applicable to conventional FFR systems with minimal modification.

Third, our simulation results show that spectrum breathing can significantly improve the performance; network traffic can be admitted by more than 125% in achieving the same delay performance, or the average delay per file transfer is reduced by more than 10 times compared to cell load balancing with delay-optimal user association only. The performance improvement for the cell edge users is also substantial. Considering that FFR was originally intended to improve the cell edge users' performance, the proposed technique is a promising solution. Finally, it turns out that spectrum breathing *without* the help of delay-optimal user association also surprisingly performs well, and thus no iteration is required at all. This makes spectrum breathing much more practical.

II. System Model

A. Assumptions

We consider an infrastructure based wireless communication system with multiple base stations. Target systems could be, but are not limited to, 3GPP-LTE. For simplicity, we focus on downlink spectrum breathing and user association but our method is also applicable to the uplink and could perhaps be combined to address more complex scenarios, as long as inter-cell interference can be reasonably assumed to be static. We assume that other cell interference is static, and can be considered as noise [5], [20], [21]. We consider a region $\mathcal{L} \subset \mathbb{R}^2$ which is served by a set of base stations $\mathcal{B}$. Let $x \in \mathcal{L}$ denote a location and $i \in \mathcal{B}$ be a BS index. Let $k \subset \mathcal{B}$ be a *cluster*, a set of BSs who partition the system spectrum w_{sys} so that for any cluster k, $\sum_{i \in k} w_i = w_{sys}$ where w_i is the spectrum used by the BS i, see Fig. 1. We assume that file transfer requests follow an inhomogeneous Poisson point process with arrival rate per unit area $\lambda(x)$ and file sizes which are independently distributed with mean $1/\mu(x)$ at location $x \in \mathcal{L}$, so the traffic load density is defined by $\gamma(x) := \frac{\lambda(x)}{\mu(x)}$; we assume $\gamma(x) < \infty$ for $x \in \mathcal{L}$. This captures spatial traffic variability. For example, a hot spot can be characterized by a high arrival rate and/or possibly large file sizes.

For simplicity, we use Shannon capacity to model the transmission rate to a user, i.e.,

$$c_i(x, w_i) = w_i \log_2(1 + SINR_i(x)) \tag{1}$$

where $SINR_i(x)$ is the received signal to interference plus noise ratio at location x for the signal from BS i. Since we assumed that interference from other clusters is randomized and fractional frequency reuse is used to mitigate interference, the sum of total interference power seen by the MT can be simply treated as *location dependent*, but *static* interference, i.e., another Gaussian-like noise [1], [2]. This static inter-cell interference model has also been adopted in previous load-balancing work [5], [20]–[23]. The $SINR_i(x)$ is thus given by

$$SINR_i(x) = \frac{P_i g_i(x)}{\sigma^2 + I_i(x)}, \tag{2}$$

where P_i denotes the transmission power spectral density (W/Hz) of BS i, $g_i(x)$ denotes the total channel gain from the BS i to the MT at location x, including path loss, shadowing, and other factors if any. Note, however, that fast fading is not considered here because the time scale for measuring $g_i(x)$ is assumed to be much larger. In addition, σ^2 denotes noise power spectral density (W/Hz) and $I_i(x)$ denotes the *average* interference power spectral density seen by the MT at location x. It should be noted that $c_i(x, w_i)$ is *location-dependent* but not necessarily determined by the distance from the BS i. For example, $c_i(x, w_i)$ can be very small in a shadowed area where $g_i(x)$ is very small. Hence, $c_i(x, w_i)$ can capture shadowing as well.

The *system-load density* $\varrho_i(x, w_i)$ is then defined by

$$\varrho_i(x, w_i) := \frac{\gamma(x)}{c_i(x, w_i)}, \tag{3}$$

which denotes the fraction of time required to deliver traffic load $\gamma(x)$ from BS i to location x. We assume that $\min_i \varrho_i(x, w_i)$ is finite, i.e., at least one BS has physical capacity to location $x \in \mathcal{L}$ that is not arbitrarily close to zero.

B. Problem formulation

Our problem is to find an optimal spectrum partition $w = (w_1, \cdots, w_b)$ combined with user association considering the physical capacity and cell load so as to minimize the system cost function given in (9). In doing this we introduce a routing function $p_i(x)$, which specifies the *probability* that a flow at location x is associated with BS i.

Definition 1 (Feasibility): The set $\mathcal{F}(w)$ of *feasible* BS loads $\rho(w) = (\rho_1(w_1), \cdots, \rho_b(w_b))$ along with the spectrum bandwidth $w = (w_1, \cdots, w_b)$ is given by

$$\begin{aligned} \mathcal{F}(w) = \Big\{ \rho(w) \quad | \quad & \rho_i(w_i) = \int_{\mathcal{L}} \varrho_i(x, w_i) p_i(x) dx, && (4) \\ & 0 \leq \rho_i(w_i) \leq 1 - \epsilon, && (5) \\ & \sum_{i \in \mathcal{B}} p_i(x) = 1, && (6) \\ & 0 \leq p_i(x) \leq 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} && (7) \\ & \sum_{i \in k} w_i = w_{sys}, \forall k \subset \mathcal{B} \Big\}, && (8) \end{aligned}$$

where ϵ is an arbitrarily small positive constant. Hence, the feasible BS loads $\rho(w)$ has the associated routing probability vector $p(x) = (p_1(x), \cdots, p_b(x))$ for $\forall x \in \mathcal{L}$. Note that the set $\mathcal{F}(w)$ is a function of w, i.e., how to allocate the spectrum bandwidth in accordance with the spatial traffic distribution.

Lemma 1: The feasible set $\mathcal{F}(w)$ is convex for a given w.

Proof: The proof is omitted due to space limitation. ■

We formulate our problem as an optimization as follows. **Problem 1**:

$$\min_{\rho(w), w} \quad \phi_\alpha(\rho(w)) = \sum_{i \in \mathcal{B}} \frac{(1 - \rho_i(w_i))^{1-\alpha}}{\alpha - 1} \tag{9}$$

$$s.t. \quad \rho(w) \in \mathcal{F}(w) \tag{10}$$

where $\alpha \geq 0$ is a parameter specifying the desired degree of load balancing. The condition of $\sum_{i \in k} w_i = w_{sys}$ in (8) states that BSs in the same cluster should have a disjoint spectrum bandwidth. When $\alpha = 1$, the objective function is defined as $\sum_i \log(\frac{1}{1-\rho_i})$. Optimizing $\phi_\alpha(\rho)$ for the case $\alpha = 2$ corresponds to minimizing the overall average flow delay in the system if MTs that are associated with a BS are served by a temporally fair scheduler. Please see [22] for the details about the objective function. We consider a dynamic system where new flows (or file transfer requests) arrive randomly (Poisson) into the system and leave after being served. The dynamics of this system are captured by a *flow-level queuing model* as shown in Fig. 2 which tracks the arrival and departure processes of users (or flows, file requests), see e.g., [18], [24], [25].

III. Decoupling-based Approach

A. Inter-dependency of ρ, w and $p_i(x)$

Note that w determines the spectrum partitioning and should depend on the load, i.e., we aim to allocate more spectrum to the cells with higher load and vice versa so that the spectrum bandwidth adaptively follows the spatial traffic load variation. In our problem formulation, the objective function $\phi_\alpha(\rho)$ is a function of $\rho(w) \in \mathcal{F}(w)$, so ϕ_α is essentially a function of w. It is straightforward that ϕ_α is a convex function of ρ, but ρ is a function of both w and user association $p(x)$ as seen from (1), (3) and (4). Furthermore, user association $p(x)$ should depend on both w and ρ as seen from Fig. 2. Hence, it is not clear how our objective function is related with both ρ and w. In addition, the feasible set $\mathcal{F}$ is a function of w. Indeed the complex inter-dependency of ρ, w, $p(x)$ and $\mathcal{F}(w)$ makes our optimization problem very challenging to tackle.

B. Reference-load and decoupling the problem

To make our problem analytically tractable, we decouple spectrum partitioning w and user association $p(x)$ so that their impacts on ρ is separable. First, instead of using $c_i(x, w_i)$ in (1) we use the spectral efficiency,

$$\tilde{c}_i(x) \quad = \quad \log_2(1 + SINR_i(x)) \tag{11}$$

which is independent of w_i. Note that here we assume that the transmit power of the BS is constant over unit bandwidth, so SINR remains constant irrespective of the use of w_i. In turn, this implies that if BS i uses large spectrum, then, the BS i uses more power proportional to the spectrum bandwidth, which makes sense in practice. Now the service rate at location x provided by BS i having spectrum bandwidth w_i is given by $c_i(x, w_i) = w_i \tilde{c}_i(x)$. Second, we define *reference-load density* as

$$\tilde{\rho}_i(x) \quad = \quad \frac{\gamma(x)}{\tilde{c}_i(x)}. \tag{12}$$

Then, we define *reference-load* of BS i as

$$\tilde{\rho}_i \quad = \quad \int_{\mathcal{L}} \tilde{\varrho}_i(x) p_i(x) dx. \tag{13}$$

The virtue of reference-load $\tilde{\rho} = (\tilde{\rho}_1, \cdots, \tilde{\rho}_b)$ is such that it is independent of the spectrum partitioning w but only depends on the user association probability $p(x)$. Hence, it *decouples* the dependency between w and $\tilde{\rho}$ and it simplifies our algorithm development. It is straightforward that $\rho_i(x) = \frac{\tilde{\rho}_i(x)}{w_i}$, and the system-load ρ_i and the reference-load $\tilde{\rho}_i$ is related by

$$\rho_i \quad = \quad \frac{\tilde{\rho}_i}{w_i}. \tag{14}$$

C. Reformulation of the problem

Now we reformulate our problem as a function of spectrum bandwidth w and the reference-load $\tilde{\rho}$. Let $v = (w, \tilde{\rho})$. For simplicity, hereafter we focus on the case when $\alpha = 2$ that minimizes the flow-level delay. Then, our objective function becomes

Problem 2:

$$\min_{v\in\mathcal{V}} \quad \phi(v) = \sum_{i\in\mathcal{B}} \frac{w_i}{w_i - \tilde{\rho}_i} \tag{15}$$

$$s.t \quad v = (w, \tilde{\rho}),$$

$$\sum_{i\in k} w_i = w_{sys}, \forall k \subset \mathcal{B} \tag{16}$$

$$\tilde{\rho}_i = \int_{\mathcal{L}} \tilde{\varrho}_i(x) p_i(x) dx, \tag{17}$$

$$0 \le \tilde{\rho}_i \le w_i - \epsilon, \tag{18}$$

$$\sum_{i\in\mathcal{B}} p_i(x) = 1, \tag{19}$$

$$0 \le p_i(x) \le 1, \forall i \in \mathcal{B} \text{ and } \forall x \in \mathcal{L} \tag{20}$$

where ϵ is an arbitrarily small constant.

Lemma 2: The feasible set $\mathcal{V}$ is a convex set.

Proof: The proof is omitted due to a space limitation. ■

Even though $\mathcal{V}$ is a convex set, the convexity of $\phi(v)$ over $\mathcal{V}$ is not guaranteed, which makes our problem still challenging. If $\phi(v)$ were a quasi-convex function, then the optimal solution could be found by several well known methods. However, unfortunately, $\phi(v)$ does not satisfy the quasi-convexity property for all points of $v \in \mathcal{V}$. Based on our numerical experiments, less than 2% of randomly selected points in $\mathcal{V}$ do not satisfy the quasi-convexity property. To this end we propose an iterative algorithm that converges to a fixed point, which is possibly an optimum point.

IV. Two-stage iterative algorithm exploiting conditional convexity

In solving Problem 2 we exploit the conditional convexity of the objective function.

Lemma 3: $\phi(w, \tilde{\rho})$ is a convex function of w given $\tilde{\rho}$, and a convex function of $\tilde{\rho}$ given w.

Proof: The proof is omitted due to space limitation. ■

From Lemma 3 we design our algorithm in two-stage as follows.

A. Distributed α-optimal user association

First let's consider the case when w is fixed. Then, the problem boils down to the case of distributed α-optimal user association for cell load balancing that converges to the optimal point given w. Refer to [22] for details. When $\alpha = 2$, the delay-optimal user association is then given as follows.

1) MT side: At the start of the m-th period MTs receive $\rho^{(m)}$ through broadcast control messages from BSs. Then, a new flow request for a MT located at x simply selects the BS $i(x)$ using the deterministic rule given by

$$i(x) = \operatorname*{argmax}_{j\in\mathcal{B}} c_j(x, w_j) \left(1 - \rho_j^{(m)}\right)^{\alpha}, \ \forall x \in \mathcal{L}. \tag{21}$$

2) BS side: This defines a new spatial partition $\mathcal{P}^{(m)} = \{\mathcal{L}_1^{(m)}, \cdots, \mathcal{L}_b^{(m)}\}$ where $\mathcal{L}_i^{(m)}$ denotes the coverage area of BS i at m-th period. Specifically, $\mathcal{L}_i^{(m)}$ depends on the broadcast load $\rho^{(m)}$ as follows:

$$\mathcal{L}_i^{(m)} = \left\{ x \in \mathcal{L} | i = \operatorname*{argmax}_{j\in\mathcal{B}} c_j(x, w_j) \left[1 - \rho_j^{(m)}\right]^{\alpha} \right\}, \ \forall i \in \mathcal{B}.$$

Then, mathematically, the new utilization of BS i is computed by (4). In practice the BS i can compute $\rho_i^{(m+1)}$ by simply averaging the busy fractional time of BS i, and broadcasts it again. This iteration converges to the conditionally optimal system load $\rho^*(w) = (\rho_1^*(w), \cdots, \rho_b^*(w))$ given w, and we have the conditionally optimal reference-load $\tilde{\rho}^*(w)$. Note from (21) that the deterministic user association is optimal even though we initially introduce the probabilistic user association $p_i(x)$.

B. Spectrum breathing

Given $\tilde{\rho}$ the objective function is convex with respect to w. In addition, given $\tilde{\rho}$, the spectrum partitioning is performed cluster by cluster. Hence from now on we focus on the BSs in cluster k rather than all the BSs in $\mathcal{B}$. We first find the optimal spectrum partitioning of cluster k using Lagrange multiplier λ. The Lagrangian is given by

$$L(w, \lambda) = \sum_{i\in k} \frac{w_i}{w_i - \tilde{\rho}_i} + \lambda(\sum_{i\in k} w_i - w_{sys}). \tag{22}$$

Then, from KKT condition, we have

$$\frac{\partial L}{\partial w_i} = \frac{-\tilde{\rho}_i}{(w_i - \tilde{\rho}_i)^2} + \lambda = 0. \tag{23}$$

Let the optimal Lagrange multiplier be denoted by λ^*, then the optimal spectrum w_i^* is given by

$$w_i^* = \tilde{\rho}_i + \sqrt{\frac{\tilde{\rho}_i}{\lambda^*}} \tag{24}$$

and λ^* is determined from the constraint $\sum_i w_i^* = w_{sys}$, and given by

$$\lambda^* = \left(\frac{\sum_{i\in k} \sqrt{\tilde{\rho}_i}}{w_{sys} - \sum_{i\in k} \tilde{\rho}_i} \right)^2. \tag{25}$$

Thus, the conditionally optimal spectrum partition $w^*(\tilde{\rho})$ is

$$w_i^*(\tilde{\rho}) = \tilde{\rho}_i + \frac{\sqrt{\tilde{\rho}_i}}{\sum_{i\in k} \sqrt{\tilde{\rho}_i}} \left(w_{sys} - \sum_{i\in k} \tilde{\rho}_i \right) \tag{26}$$

Roughly speaking (26) implies that the spectrum w_i becomes larger if the reference-load $\tilde{\rho}_i$ is high, and vice versa. This intuitively makes sense because large $\tilde{\rho}_i$ implies that BS i is congested, and it would be better to allocate more spectrum to BS i.

C. Algorithm design

Once the spectrum partition w is updated to $w^*(\tilde{\rho})$, new ρ_i is obtained by applying again user association given in Section IV-A. After the user association is done, the spectrum breathing is again performed. The iteration between spectrum breathing and user association proceeds until converging to a fixed point.

Proposition 4.1: The proposed algorithm converges a fixed point of $(w^*(\tilde{\rho}^*), \tilde{\rho}^*(w^*)) \in \mathcal{V}$.

Proof: Since $\phi(w, \tilde{\rho})$ is a conditionally convex function for both w and $\tilde{\rho}$, fixing one variable and minimizing ϕ with respect to the other variable always decreases ϕ at each process

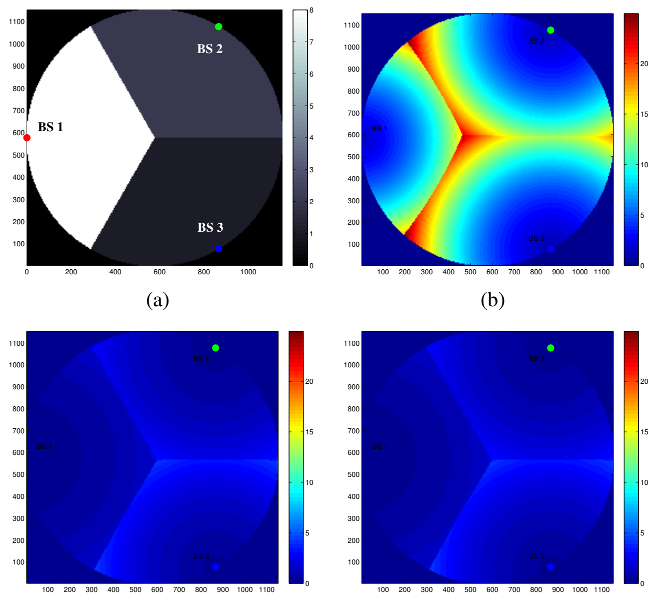

Fig. 3. Cell coverage and conditional delay over space. (a) Simulation setup (b) Case 2: delay-optimal user association (DOUA) only (c) Case 3: spectrum breathing with DOUA (d) Case 4: spectrum breathing with ROUA.

of user association and spectrum breathing. Since $\phi(w, \tilde{\rho})$ is bounded above by zero, ϕ should converge, and so does $(w, \tilde{\rho})$. ■

We summarize the proposed algorithm as follows.

Algorithm: Spectrum Breathing with User Association

```
Initialize $w = (\frac{w_{sys}}{|k|}, \cdots, \frac{w_{sys}}{|k|})$ where $|k|$ is the number
  of BSs in the cluster $k \subset \mathcal{B}$.
repeat
    User Association:
    Given $w$,
    repeat
    MTs choose BSs according to (21).
    BSs update $\rho$ and broadcast it.
    until $\rho$ converges to some point $\rho^*(w) \in \mathcal{F}(w)$
    Spectrum Breathing:
    Given $\tilde{\rho}$,
    update $w = w^*(\tilde{\rho})$ according to (26)
until $(w, \tilde{\rho})$ converges to a fixed point given by
  $(w^*(\tilde{\rho}^*), \tilde{\rho}^*(w^*)) \in \mathcal{V}$.
```

D. Necessary condition of optimality

Proposition 4.2: If the algorithm converges to an optimal point, the following fixed point condition should be satisfied

$$w_i^* = \rho_i^* w_i^* + \frac{\sqrt{\rho_i^* w_i^*}}{\sum_{i \in k} \sqrt{\rho_i^* w_i^*}} \left(w_{sys} - \sum_{i \in k} \rho_i^* w_i^* \right) \quad (27)$$

for $\forall i \in k$ where $w^* = w^*(\tilde{\rho}^*)$ and $\tilde{\rho}^* = \tilde{\rho}^*(w^*)$.

Proof: Applying (14) to (26) completes the proof. ■

V. Performance Evaluation

For numerical experiments, we consider the case of three BSs that are separated by 1,000 m each as shown in Fig. 3(a). The maximum power of the BSs is 40 dBm, w_{sys} is 10 MHz, and σ^2 is -174dBm/Hz. To see whether the proposed algorithm works well even under highly inhomogeneous traffic distribution, we consider the load ratio of BSs is 8:2:1 as shown in Fig. 3(a). We compare four different combinations depending on the use of spectrum breathing and user association as follows:

- Case 1: Rate-optimal user association (ROUA) Only. This is the conventional case when $\alpha = 0$, i.e., users are associated with the closest BS (no load balancing), and also no spectrum breathing is applied.
- Case 2: Delay-optimal user association (DOUA) [22] is applied for load balancing, but spectrum breathing is not applied.
- Case 3: Spectrum breathing with DOUA is applied.
- Case 4: Spectrum breathing with ROUA is applied.[2]

Fig. 3 (b)-(d) show the cell coverage and conditional average delay for Case 2–4, respectively. Note that we do not include Case 1 here because the system becomes unstable due to high traffic load. The conditional average delay experienced by the flow at location x when served by BS i is given by

$$E[D_i|X = x] = \frac{1}{\lambda(x)} \frac{\varrho_i(x)}{\rho_i} \frac{\rho_i}{1 - \rho_i} = \frac{1}{\mu(x) c_i(x)(1 - \rho_i)}$$

in the case of an $M/GI/1$ multi-class processor sharing system model [22]. Comparing Fig. 3(b) and Fig. 3(c) clearly shows that the delay performance at cell edge is significantly improved by applying spectrum breathing. We also see that the cell coverages are changed. In Fig. 3(b) the cell coverage of BS 1 shrinks to balance cell loads, but still cell edge users experience very large delay. By contrast, spectrum breathing successfully balances the cell loads; initially, each BS has 3.33 MHz of bandwidth, but after the iteration of spectrum breathing and user association, the bandwidths converge to 6.964 MHz, 1.962 MHz, 1.704 MHz, respectively. As a result, the congested BS 1 borrows spectrum from its neighboring BSs and uses twice than before, which contributes to significantly reducing the average delay. We also plot the cumulative distribution function (CDF) of the conditional flow delay in Fig. 4 for Case 2–4. As can be seen, the improvement of delay performance becomes more significant for cell edge users. Considering that FFR was originally intended to improve the cell edge users performance, the proposed technique is a promising solution.

A. Delay vs. load

To see the performance under various traffic loads, we compare Case 1–4 in Fig. 5 by increasing the traffic loads. As expected, the delay of Case 1 blows up very quickly even for the medium load. Applying DOUA (Case 2) is beneficial, but the performance is improved further in Case 3. When spectrum breathing is incorporated with delay-optimal user association, the average delay performance is significantly improved compared to Case 1 and Case 2; in achieving the same delay performance, the allowable traffic load is increased by

[2] Reference capacity $\tilde{c}_i(x)$ is used in ROUA instead of $c_i(x, w_i)$.

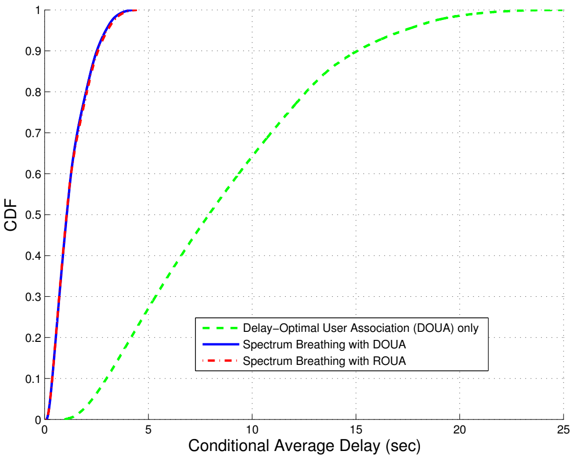

Fig. 4. CDF of conditional average delay for Case 2–4.

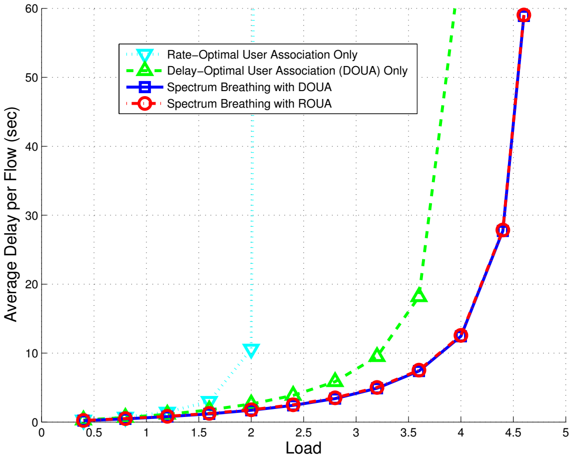

Fig. 5. The comparison of average delay vs load for Case 1–4.

125% and 20% compared to Case 1 and Case 2, respectively. Hence, without increasing the infrastructural facilities, more traffic can be accommodated. More interestingly, Case 3 and Case 4 show almost the same performances, i.e., spectrum breathing alone is almost as good as spectrum breathing with DOUA. This result is remarkable because we do not need to wait until delay-optimal user association converges, which makes the proposed algorithm more practical.

Finally, recall that the proposed algorithm always converges to a fixed point, but the result may not be optimal. Hence, we perform an exhaustive search for the two BSs case, i.e., try all the possible combinations of spectrum partitioning and user association. Our proposed algorithm indeed achieves the optimality, so the optimality gap is zero, but due to the space limitation the results are not included here.

VI. Conclusion

In this paper we investigated a technique that adaptively change the spectrum partitioning ratio in FFR system in accordance with the spatially inhomogeneous traffic distribution. We formulated an optimization problem using flow-level dynamics, but the problem turned out to be non-convex optimization. Hence, we proposed an algorithm of spectrum breathing combined with delay-optimal user association that always converges to a fixed point, which is possibly the optimal solution of the original problem. By extensive simulation, we demonstrated that spectrum breathing with user association substantially improved the flow-level delay performance: 10 times reduction in delay, or more than 125% increase of the affordable traffic. Furthermore, spectrum breathing without delay-optimal user association achieved almost the same performance, which makes the proposed solution much more practical.

References

[1] 3GPP Long Term Evolution, "LTE release 10 & beyond (LTE-Advanced)," .
[2] "IEEE Std 802.16m part 16: Air interface for broadcast wireless access systems: Advanced air interface," *IEEE Standard 802.16m*, 2010.
[3] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Comm. Mag.*, pp. 74–81, Apr. 2009.
[4] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multi-cell OFDMA network," *IEEE Trans. on Veh. Technol.*, vol. 58, no. 6, pp. 12835–2848, Jul. 2009.
[5] K. Son, S. Chong, and G. de Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Comm.*, vol. 8, no. 5, pp. 3566–76, July 2009.
[6] Ericsson, "Downlink inter-cell interference coordination/avoidance evaluation of frequency reuse," *Tech. Rep. TSG-RAN WG1 R1-061374, 3GPP*, May 2006.
[7] X. Zhang, C. He, L. Jiang, and J. Xu, "Inter-cell interference coordination based on softer frequency reuse in OFDMA cellular systems," in *Neural Networks and Signal Processing*, June 2008.
[8] B. Krasniqi and C. F. Mecklenbrauker, "Efficiency of partial frequency reuse in power used depending on user's selection for cellular networks," *Personal Indoor and Mobile Radio Comm.*, pp. 268–272, Sept. 2011.
[9] R. Kwan and C. Leung, "A survey of scheduling and interferencemitigation in LTE," *Journal of Electrical and Computer Engineering - Special issue on LTE/LTE-Advanced*, Jan. 2010.
[10] L. Shu, X. Wen, Z. Liu, W. Zheng, and Y. Sun, "Queue analysis of soft frequency reuse scheme in LTE-Advanced," in *International Conference on Computer Modeling and Simulation*, Jan. 2010.
[11] B. Krasniqi, M. Wrulich, and C. F. Mecklenbrauker, "Network-load dependent partial frequency reuse for LTE," *International Symposium on Comm. and Information Technologies*, pp. 672–676, Sept. 2009.
[12] X. Fan, S. Chen, and X. Zhang, "An inter-cell interference coordination techniquebased on users' ratio and multi-level frequency allocations," in *Wireless Communications, Networking and Mobile Computing*, Sept. 2007.
[13] A Simonsson, "Frequency reuse and intercell interference co-ordination in E-UTRA," in *Proc. IEEE Veh. Technol. Conf.*, Apr. 2007.
[14] X. Mao, A. Maaref, and K. H. Teo, "Adaptive soft frequency reuse for inter-cell interference coordination in SC-FDMA based 3GPP LTE uplinks," in *Proc. IEEE Glob. Telecom. Conf.*, Nov. 2008.
[15] H. Zhang, N. Prasad, and S. Rangarajan, "Method and systems for dynamic and configuration based fractional frequency reuse for uneven load distributions," in *US Patent Application Publication, US 2001/0070911 A1*, Mar. 2011.
[16] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *European Wireless Conference*, 2005.
[17] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *Proc. IEEE INFOCOM*, vol. 1, pp. 321–331, Apr. 2003.
[18] H. Kim and G. de Veciana, "Losing opportunism: evaluating service integration in an opportunistic wireless system," in *Proc. IEEE INFOCOM*, 2007.
[19] H. Kim and G. de Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy," *IEEE/ACM Trans. Networking.*, vol. 18, June 2010.
[20] A. Sang, M. Madihian, X. Wang, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," in *ACM Mobicom*, Sep. 2004.
[21] P. Hande, S. Patil, and H. Myung, "Distributed load-balancing in a multi-carrier system," in *Proc. IEEE Wireless Comm. and Net. Conf.*, 2009.
[22] H Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Networking.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.
[23] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: A practical interference management in heterogeneous wireless access networks," *IEEE Jour. Select. Areas in Comm.*, vol. 29, no. 6, pp. 1260–1271, June 2011.
[24] S. B. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *ACM SIGCOMM*, 2001.
[25] T. Bonald and J. W. Roberts, "Congestion at flow level and the impact of user behavior," *Computer Networks*, pp. 521–536, 2003.