This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2010 proceedings
This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2010

α-Optimal User Association and Cell Load Balancing in Wireless Networks

Hongseok Kim[†], Gustavo de Veciana[†], Xiangying Yang[‡] and Muthaiah Venkatachalam[‡]
[†] Department of Electrical and Computer Engineering, The University of Texas at Austin
[‡] Wireless Standardization and Technology Group, Intel Corporation

***Abstract*—In this paper we develop a framework for user association in infrastructure-based wireless networks, specifically focused on flow-level cell load balancing under spatially inhomogeneous traffic distributions. Our work encompasses several different user association policies: rate-optimal, throughput-optimal, delay-optimal, and load-equalizing, which we collectively denote α-optimal user association. We prove that the optimal load vector ρ^* that minimizes a generalized system performance function is the fixed point of a certain mapping. Based on this mapping we propose and analyze an iterative distributed user association policy that adapts to spatial traffic loads and converges to a globally optimal allocation.**

I. Introduction

Fourth generation wireless cellular standards such as IEEE802.16m WiMAX2 and LTE-Advanced are designed to support broadband data services (in addition to voice) so as to meet growing demands for connectivity, e.g., file transfers and web browsing, on mobile platforms [1], [2]. One of the important problems in multi-cell data networks is properly associating mobile terminals (MTs) with serving base stations (BSs); this is usually referred to as the *user association problem*. In selecting the serving BS, two metrics - instantaneous achievable rate at the physical layer and cell load - should be considered. Since the achievable rate is computed from the received signal-to-interference-plus-noise ratio (SINR), the simplest (and thus widely accepted) rule is to choose the BS that gives the strongest downlink pilot signal. However, this rule is naive in the sense that it does not consider either inter-cell interference or cell load balancing.

There have been many efforts in the literature towards developing user association rules considering interference avoidance and/or cell load balancing [3]–[13]. To avoid interference when frequency is universally reused and inter-cell interference is severe, *centralized* approaches have been considered [5], [8], [10], [11]. The basic idea is to schedule users across cells so that they do not severely interfere with each other. This is called inter-cell coordinated scheduling. Earlier work on load balancing also mostly assumed a centralized controller that governs the BSs and the MTs with access to all the necessary information [3], [6], [7], [9], [13]. However, centralized approaches, for either interference avoidance and/or load balancing, may require excessive computational complexity and message overhead, which increase exponentially in the size of the network. Such centralized functionality is usually implemented in a server deep in the core network, which only allows slow adaptation at relatively long time scales. To avoid relying on a centralized controller, current systems are usually based on *fractional frequency reuse* or *interference randomization* [1], [2]. Distributed cell load balancing is also being considered as a basic requirement in upcoming standards. For example, IEEE 802.16m WiMAX2 recently included parameters such as cell load and cell type in the system information broadcast [1], [14].

In this paper we investigate *distributed* user association policies. We will not consider interference avoidance that requires inter-cell coordinated scheduling. So, our approach is reasonable when fractional frequency reuse or interference randomization are being used so that inter-cell interference can be roughly considered as static noise. We focus on developing a theory and algorithms for user association that adapt to *spatially inhomogeneous* traffic. 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 sometimes referred to as *flow-level dynamics* [5], [15].

Interestingly, even though user association in a dynamic setting can be viewed as a routing problem among queues, it is still not well understood; most work to date, is ad hoc in nature, and does not address dynamic systems [3], [4], [7], [10], [13], [16], [17]. The work in [5], [6], [8] explores flow-level dynamics for load balancing, but assumes a centralized controller. In particular none of these efforts fully explore the role of load balancing under spatially inhomogeneous traffic distributions.

One of the main challenges in developing a distributed user association policy is achieving global performance optimum without relying on a centralized controller, and doing so to track changes in traffic distributions; for example, day and night have quite different spatial traffic distributions as may traffic on an hourly (or faster timescale) basis. Our proposed mechanism, denoted *α-optimal user association*, effectively overcomes these challenges.

Contributions. We highlight the contributions of this paper as follows. First, we provide a theoretical framework for user association, specifically focused on load balancing under spatially inhomogeneous traffic distributions in an infrastructure-based wireless network. We formulate the user association problem as a convex optimization problem. Then we show a fixed point optimality condition characterizing the spatial partitions (cell coverage areas) associated with minimizing a

This work was supported in part by a gift from the Intel Research Council and NSF Award CNS-0721532.

978-1-4244-5837-0/10/$26.00 ©2010 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2010 proceedings
This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2010

general system-level performance function. The optimal spatial partition is shown to be unique up to a set of traffic measure zero – this will be explained in the sequel. The optimality condition reveals many interesting facts, e.g. : cell loads are not interchangeable, and *balancing* loads to minimize delay does not imply *equalizing* loads at the BSs; Voronoi cells need not be delay optimal even if the traffic loads are spatially homogeneous; and cell coverage areas need not be contiguous, i.e., can be fragmented.

Second, we present a distributed algorithm and prove its convergence to a global optimum irrespective of the initial condition. Our algorithm could in principle track slowly varying traffic loads. It is also very simple and easily implementable; one need only implement a simple greedy behavior by MTs to achieve a global optimum. The proposed algorithm supports a family of load balancing objectives as α ranges from 0 to ∞: rate-optimal ($\alpha = 0$), throughput-optimal ($\alpha \geq 1$), delay-optimal ($\alpha = 2$), and equalizing BS loads ($\alpha = \infty$). Our work is general and applicable to various scenarios. For example, our model for achievable rate at the physical layer can capture shadowing. We do not assume the Tx power of BSs are the same, so our work is also applicable to heterogeneous BS deployments such as macro, micro, pico and even femto cells. Finally, our user association rule can easily address handoffs [14].

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, WiMAX2 or 3GPP-LTE. For simplicity, we focus on downlink communications but our method is also applicable to the uplink. We assume that other cell interference is static, and can be considered as noise [4], [10], [13]. We consider a region $\mathcal{L} \in \mathbb{R}^2$ 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. 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.

Definition 1 (traffic load measure): We define the traffic load *measure*, $m(\cdot)$, of a Borel set $\mathcal{G}$ as $m(\mathcal{G}) = \int_{\mathcal{G}} \gamma(x)dx$.

Assumption 2.1 (capacity function): We assume the physical capacity each BS $i \in \mathcal{B}$ can deliver to location x, $c_i(x)$, is a Borel measurable function and for any $\eta > 0$ and $i, j \in \mathcal{B}$, the set

$$\mathcal{D}_{ij}(\eta) = \{x \in \mathcal{L} | c_i(x)/c_j(x) = \eta\} \tag{1}$$

has traffic load measure zero, i.e., $m(D_{ij}(\eta)) = 0$. Also to avoid unnecessary technicalities we assume $c_i(x) > 0$ for all $i \in \mathcal{B}$ and $x \in \mathcal{L}$.

As will be seen in the sequel this implies that cell 'boundaries' have zero traffic load measure. Note this model allows for a fairly general but *deterministic* capacity function.

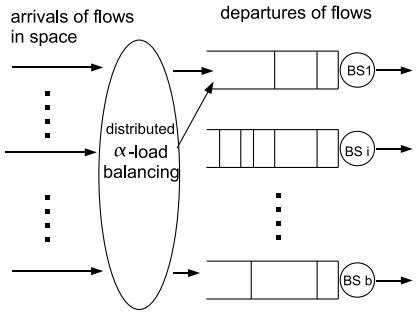

Fig. 1. Flow-level queueing model for user association problem.

Remark 2.1: When $c_i(x)$ is discrete valued, $\mathcal{D}_{ij}(\eta)$ may not have traffic load measure zero, so *non-trivial* tie breaking rules are necessary.

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

$$c_i(x) = \log_2(1 + SINR_i(x)) \tag{2}$$

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 is randomized and/or fractional frequency reuse is used to mitigate interference, the sum of total interference power seen by the MT can be simply treated as 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 [4], [10], [13]. The $SINR_i(x)$ is then given by

$$SINR_i(x) = \frac{P_i g_i(x)}{\sigma^2 + I(x)}, \tag{3}$$

where P_i denotes the transmission power 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. Also σ^2 is noise power and $I(x)$ is the *average* interference seen by the MT at location x. It should be noted that $c_i(x)$ is *location-dependent* but not necessarily determined by the distance from the BS i. For example, $c_i(x)$ can be very small in a shadowed area where $g_i(x)$ is very small. Hence, $c_i(x)$ can capture shadowing as well.

The *system-load density* $\varrho_i(x)$ is then defined by $\varrho_i(x) := \frac{\gamma(x)}{c_i(x)}$, which denotes the fraction of time required to deliver traffic load $\gamma(x)$ from BS i to location x. We assume that $\min_i \varrho_i(x)$ 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 user association policy considering the physical capacity and cell load so as to minimize the system cost function given below. 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. We will see that for our system model and Assumption 2.1 the optimal routing policy is deterministic, i.e., $p_i^*(x) \in \{0, 1\}$,

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2010 proceedings
This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2010

which also uniquely determines spatial cell coverage areas $\{\mathcal{L}_i\}$.

Definition 2 (Feasibility): The set $\mathcal{F}$ of *feasible* BS loads $\rho = (\rho_1, \cdots, \rho_b)$, is given by

$$\mathcal{F} = \Big\{ \rho \quad | \quad \rho_i = \int_{\mathcal{L}} \varrho_i(x) p_i(x) dx, \tag{4}$$

$$0 \le \rho_i \le 1 - \epsilon, \tag{5}$$

$$\sum_i p_i(x) = 1, \tag{6}$$

$$0 \le p_i(x) \le 1, \forall i \in \mathcal{B} \text{ and } \forall x \in \mathcal{L} \Big\}, \tag{7}$$

where ϵ is an arbitrarily small positive constant.

Lemma 1: The feasible set $\mathcal{F}$ is convex.

Proof: Consider two load vectors $\rho^1 \in \mathcal{F}$ and $\rho^2 \in \mathcal{F}$, $\rho^1 \neq \rho^2$. Then, there exist associated $p^1(x) = (p_1^1(x), \cdots, p_b^1(x))$ and $p^2(x) = (p_1^2(x), \cdots, p_b^2(x))$ such that $\rho_i^1 = \int \varrho_i(x) p_i^1(x) dx$ and $\rho_i^2 = \int \varrho_i(x) p_i^2(x) dx$ for all $i \in \mathcal{B}$. Now we make ρ as a convex combination of ρ_1 and ρ_2, i.e., for $\theta \in [0,1]$, $\rho_i = \theta\rho_i^1 + (1-\theta)\rho_i^2 = \int \varrho_i(x)[\theta p_i^1(x) + (1-\theta)p_i^2(x)]dx$ for all $i \in \mathcal{B}$. Let $p(x)$ be the routing probability associated with ρ. Then, $p_i(x) = \theta p_i^1(x) + (1-\theta)p_i^2(x)$, and it satisfies (4) to (7). Hence ρ is feasible, and so $\mathcal{F}$ is a convex set. ■

We formulate our problem as a convex optimization as follows.

Problem 1:

$$\min_{\rho} \left\{ \phi_\alpha(\rho) = \sum_i \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} \,\Big|\, \rho \in \mathcal{F} \right\} \tag{8}$$

where $\alpha \ge 0$ is a parameter specifying the desired degree of load balancing. When $\alpha = 1$ the objective function is defined as $\sum_i \log(\frac{1}{1-\rho_i})$. Problem 1 is said to be feasible if $\mathcal{F}$ is non-empty.

C. Motivation for the objective function

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. 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. 1 which tracks the arrival and departure processes of users (or flows, file requests), see e.g., [18]–[20].

Let $\mathbf{N_i} = (N_i(t), t \ge 0)$ denote a random process representing the number of ongoing file transfers served by BS i at time t. Then, if the system is stationary, the stationary distribution π_i of N_i is identical to that of an $M/GI/1$ multi-class processor sharing system [21], and given by $\pi_i(n_i) = (1-\rho_i)\rho_i^{n_i}$. Multi-class reflects the fact that users see different service rates and file sizes based on their locations. We consider infinitely many classes because we address this problem in a continuous space $\mathcal{L}$. The average number of flows at BS i is then simply given by $E[N_i] = \frac{\rho_i}{1-\rho_i}$ and total number of flows in $\mathcal{L}$ is $E[N] = \sum_i E[N_i] = \sum_i \frac{\rho_i}{1-\rho_i}$. From Little's formula, minimizing the average number of flows is equivalent to minimizing the average delay experienced by a *typical* flow. Minimizing $\sum_i \frac{\rho_i}{1-\rho_i}$ is equivalent to (8) when $\alpha = 2$ because $\sum_i \left(\frac{\rho_i}{1-\rho_i} + 1\right) = \sum_i \frac{1}{1-\rho_i}$, which does not change the optimization problem.

D. α-optimal user association

Before discussing the optimal user association and how to achieve it, we first discuss the implications of this framework. The solution to Problem 1 gives a unified approach that allows the mobile terminals to select the BS considering signal strength (a user point of view) and the degree of load balancing (the network point of view). Throughout this paper we will see that if Problem 1 is feasible, the optimal decision made by the mobile terminal located at x is to join BS $i(x)$ given by

$$i(x) = \underset{j \in \mathcal{B}}{\operatorname{argmax}}\, c_j(x)(1-\rho_j^*)^\alpha, \;\; \forall x \in \mathcal{L} \tag{9}$$

where $\rho^* = (\rho_1^*, \cdots, \rho_b^*)$ denotes an optimal load vector, i.e., solution to Problem 1.

Remark 2.2 (Tie-breaking): A location $x \in \mathcal{L}$ is called a cell boundary if a tie of argmax operation in (9) happens at x. Based on Assumption 2.1, cell boundaries have traffic load measure zero; nevertheless, for completeness *if a tie happens, we shall hereafter assume that the MT at such a location chooses the lower indexed BS.*

From (9) the mobile terminal chooses a BS that provides the highest physical capacity weighted by a power of *BS's idle time*. By a BS's idle time we refer to the fraction of time it is inactive, i.e., $1 - \rho_i$. Depending on the value of α we categorize α-optimal user association policies into four cases.

1) Rate-optimal policy: When $\alpha = 0$, the decision is purely based on user's perspective, i.e., based on the physical capacity only (or SINR), and oblivious of network traffic condition. In this case one can show that α-optimal user association maximizes the *arithmetic* mean of the BSs' idle times.

2) Throughput-optimal policy: As α increases, the BS selection criteria gradually shifts from user's perspective to network perspective, and $\alpha = 1$ is a critical point. This is because $\phi_\alpha(\rho)$ goes to infinity with loads close to 1 only if $\alpha \ge 1$ and ensures a stable behavior. When $\alpha = 1$, it can be shown that the *geometric* mean of the BSs' idle time is maximized.

3) Delay-optimal policy: When $\alpha = 2$, average file transfer delay is minimized as we have seen. In addition, one can show that the *harmonic* mean of the BSs' idle time is maximized.

4) Equalizing-load policy: As α further increases, the rule is such that more emphasis is placed on the traffic loads rather than the physical capacity. One can show that as $\alpha \to \infty$, α-optimal user association minimizes the maximum utilization, i.e., min-max utilization, and furthermore it equalizes the utilization of all the BSs.

III. Distributed Iteration Achieving Optimality

In this section we propose a distributed adaptive user association algorithm that achieves the global optimum of Problem 1 in an iterative manner. The algorithm is simple; BSs periodically share their time average loads with MTs, and

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2010 proceedings
This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2010

MTs use this information to make decisions over these periods. We will show that if spatial loads are temporally stationary, the load vector eventually converges to the unique solution of Problem 1, which in turn determines spatial coverage areas associated with each BS. However to show convergence we shall assume the following simplifying assumption.

Assumption 3.1 (Separation of time scales): We shall assume the flow arrival and departure process is very fast relative to the period on which BSs advertise their loads. In particular, once the BSs advertise their load vector, prior to the next update the BSs are able to measure the new steady state loads associated with MT decisions under the advertised vector.

A. Distributed-decision algorithm

The algorithm involves two parts.

Mobile terminal: At the start of the k-th period MTs receive $\rho^{(k)}$, e.g., through broadcast control messages from BSs.[1] Then, a new flow request for a MT located at x simply selects the BS $i(x)$ using the deterministic rule given by (9) where ρ^* is replaced by $\rho^{(k)}$. Let $\mathcal{L}_i^{(k)}$ denote the coverage area of BS i at k-th period. Then, a new spatial partition $\mathcal{P}^{(k)} = \{\mathcal{L}_1^{(k)}, \cdots, \mathcal{L}_b^{(k)}\}$ is determined by $\rho^{(k)}$ and given by

$$\mathcal{L}_i^{(k)} = \left\{ x \in \mathcal{L} | i = \operatorname*{argmax}_j c_j(x) \left[1 - \rho_j^{(k)}\right]^\alpha \right\}, \ \forall i \in \mathcal{B}. \quad (10)$$

Base station: During the k-th period BSs measure their average utilizations. Due to Assumption 3.1, the measured utilization converges to $T_i(\rho^{(k)})$ given by

$$T_i(\rho^{(k)}) = \min \left[\int_{\mathcal{L}_i^{(k)}} \varrho_i(x) dx, 1 - \epsilon \right], \quad \forall i \in \mathcal{B}. \quad (11)$$

Note that the measured utilization, i.e., average busy fractional time of the BS i cannot exceed 1. To avoid unnecessary technicalities we introduce an arbitrarily small positive constant ϵ. It can be shown that $T(\rho) = \{T_i(\rho)\}$ is a continuous mapping defined on $[0, 1-\epsilon]^b$ to itself.

After $T(\rho^{(k)})$ is measured, BSs compute and advertise their next broadcast message $\rho^{(k+1)}$ given by

$$\rho^{(k+1)} = \beta^{(k)} \rho^{(k)} + (1 - \beta^{(k)}) T(\rho^{(k)}) := S(\rho^{(k)}) \quad (12)$$

where $\beta^{(k)} \in [0, 1)$ is an exponential-averaging parameter. It should be noted that $T(\rho^{(k)})$ corresponds to the average loads seen during the k-th period while $\rho^{(k)}$ is an exponential average of $T(\rho^{(\ell)})$ across periods, i.e., $\ell = 0, \cdots, k-1$ with some initial loads $\rho^{(0)} \in \mathcal{F}$.

B. Fixed point achieves optimality

Note that if $\rho^{(k)}$ converges it must converge to a fixed point of (12), i.e., a solution to

$$\rho^* = T(\rho^*). \quad (13)$$

Due to the space limitations the proof that (12) converges is provided in [22]. Below we will show that $T(\cdot)$ has a unique fixed point ρ^* corresponding to the optimal load vector associated with Problem 1.

[1]IEEE 802.16m facilitates this type of message structure [1], [14].

Theorem 1: Suppose that Problem 1 is feasible. Then, T has a unique fixed point which is the optimal solution to Problem 1. In addition, under Assumption 2.1 this fixed point determines a unique optimal spatial partition $\mathcal{P}^*$ up to a set of traffic measure zero.

Proof: Since T is a continuous mapping defined on compact set $[0, 1-\epsilon]^b$ to itself, by Brouwer's fixed point theorem, a solution of $T(\rho^*) = \rho^*$ must exist. Now we prove that ρ^* is the optimal solution of Problem 1. Since $\phi_\alpha(\rho)$ is a convex function over a convex set, if ρ^* satisfies the following condition

$$\langle \nabla \phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \quad (14)$$

for all $\rho \in \mathcal{F}$ where $\Delta\rho^* = \rho - \rho^*$, then ρ^* is the optimal solution of Problem 1.

Let $p(x)$ and $p^*(x)$ be the associated routing probabilities for ρ and ρ^*, respectively. From (10), (11) and (13) we have

$$p_i^*(x) = \mathbf{1}\left\{ i = \operatorname*{argmax}_j c_j(x) \left(1 - \rho_j^*\right)^\alpha \right\}, \quad (15)$$

and then the inner production is computed such as

$$\begin{aligned} &\langle \nabla \phi_\alpha(\rho^*), \Delta\rho^* \rangle \\ &= \sum_i \frac{1}{(1-\rho_i^*)^\alpha} (\rho_i - \rho_i^*) \\ &= \sum_i \frac{\int_{\mathcal{L}} \varrho_i(x) \left(p_i(x) - p_i^*(x)\right) dx}{(1-\rho_i^*)^\alpha} \\ &= \int_{\mathcal{L}} \gamma(x) \left[\sum_i \frac{p_i(x) - p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha} \right] dx. \end{aligned} \quad (16)$$

Note that

$$\sum_i \frac{p_i(x)}{c_i(x)(1-\rho_i^*)^\alpha} \geq \sum_i \frac{p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha}$$

holds because $p_i^*(x)$ in (15) is an indicator for the maximizer of $c_j(x)(1-\rho_j^*)^\alpha$, for all $j \in \mathcal{B}$. Hence, $\langle \nabla \phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0$. When $\alpha > 0$ Problem 1 is strictly convex, and ρ^* should be unique, and so is the fixed point. When $\alpha = 0$, the optimal policy selects the BS that gives the highest $c_i(x)$ without considering load. Hence $T(\rho)$ is independent of the load vector ρ and a constant function, which ensures that ρ^* is unique.

In addition we can show that ρ^* has a corresponding spatial partition $\mathcal{P}^* = \{\mathcal{L}_i^*, i \in \mathcal{B}\}$ which is unique up to a set of traffic measure zero. Suppose that there are two such partitions $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ associated with ρ^*, and there exists a set $\mathcal{M} \subset \mathcal{L}$ with non-zero traffic measure where $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ differ, i.e., user associations are different. In particular, without loss of generality on $\mathcal{M}$, under $\mathcal{P}_1^*$, users at those locations associated with BS 1, while under $\mathcal{P}_2^*$ they associate with BS 2. It follows that on $\mathcal{M}$ there must be a tie, yet by Assumption 2.1 such sets have traffic measure zero. This is then a contradiction. It follows that the induced partition $\mathcal{P}^*$ is unique except on sets which have zero traffic measure. ■

Remark 3.1: Utilizations can also be indirectly estimated by measuring the average number of flows in the system. For example in an $M/GI/1$ processor sharing queue, the average number of flows is given by $E[N_i] = \frac{\rho_i}{1-\rho_i}$, which in turn

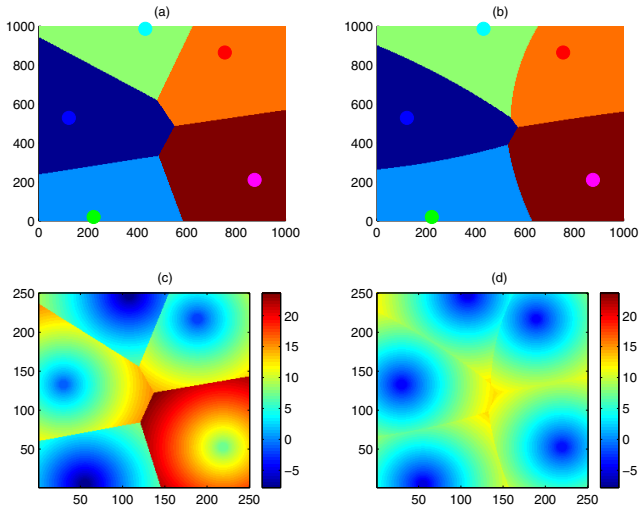

Fig. 2. Voronoi cells vs Delay-optimal cells (a–b), and spatial distribution of conditional average delay (dB scale) in each case.

yields $\rho_i = \frac{E[N_i]}{E[N_i]+1}$. Replacing $\rho_i = \frac{E[N_i]}{E[N_i]+1}$ into (9) when $\alpha = 1$ gives

$$i(x) = \underset{j}{\operatorname{argmax}} \frac{c_j(x)}{E[N_j]+1}. \tag{17}$$

This rule is a special case of α-optimal user association proposed as a heuristic in [4], [10] and [13].

Example 1 (Spatial delay smoothing): This example shows the BS coverage areas and geographical distribution of average file transfer delays. Five BSs are randomly placed in 1000m $\times$ 1000m region. As an example of inhomogeneous traffic loads, a linearly increasing load in the diagonal direction is considered. The Tx power of all the BSs was normalized to 1. We assume hereafter that the Tx power is 1, unless otherwise specified, throughout the paper. In addition, $c_i(x)$ is computed using pathloss exponent 3. Fig. 2 (a) shows the partition when $\alpha = 0$ (Voronoi cells), and Fig. 2 (b) shows the partition when $\alpha = 2$ (delay-optimal cells). Fig 2 (c) and (d) show the conditional average file transfer delays (dB scale) at x, which 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. For simplicity, we set $1/\mu(x) = 1$ and show the average 1-bit transmit time. The benefit of delay-optimal load balancing is clearly shown in Fig. 2. A slight modification of the cell coverages significantly improves the delay performance, specifically, of the congested cell at the right lower corner.

IV. Conclusion and Future Work

In this paper we proposed a theoretical (and also practical) framework for user association problem in wireless networks. We specifically focused on distributed load balancing under spatially inhomogeneous traffic distributions and showed the optimality condition of cell coverage areas that minimizes generalized system performance function. Interestingly, the optimal user association policy, i.e., routing of flows to appropriate BSs is deterministic even though probabilistic routing is allowed. This deterministic property enables us to develop a simple distributed-decision algorithm at the MTs, which is easily implementable and compliant with upcoming standards, e.g., WiMAX2. Our distributed algorithm converges to the global optimum and also is robust to changes of traffic distributions. Our work will be extended to the case where the system cannot be stabilized due to excessive traffic loads. Under such heavy traffic regimes, we will propose optimal admission control policies considering tradeoffs between two QoS metrics: average delay vs. maintaining a minimum level of connectivity to users independent of their location.

References

[1] "IEEE P802.16m-2007 draft standards for local and metropolitan area networks part 16: Air interface for fixed broadcast wireless access systems," *IEEE Standard 802.16m*, 2007.
[2] 3GPP Long Term Evolution, "Physical layer aspects of UTRA high speed downlink packet access," *Technical Report TR25.814*, 2006.
[3] S Das, H Viswanathan, and G Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. IEEE INFOCOM*, 2003, pp. 786–96.
[4] 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.
[5] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *European Wireless Conference*, 2005.
[6] A. Zemlianov and G. de Veciana, "Load balancing in wireless systems supporting end nodes with dual mode capabilities," in *Proc. Conference on Information Sciences and Systems (CISS)*, Mar. 2005.
[7] K. Navaie and H. Yanikomeroglu, "Downlink joint base-station assignment and packet scheduling for cellular CDMA/TDMA networks," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, 2006, pp. 4339–44.
[8] S. Liu and J. Virtamo, "Inter-cell coordination with inhomogeneous traffic distribution," in *Next Generation Internet Design and Engineering Conference*, 2006.
[9] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, 2006.
[10] K. Son, S. Chong, and G. de Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Comm.*, to appear.
[11] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic aware system-level coordination of wireless systems," *submitted to IEEE/ACM Trans. Networking*, June 2009.
[12] B. Rengarajan and G. de Veciana, "Load balancing is not optimal in wireless systems with dynamic intereference," *International Workshop on Network Control and Optimization, Submitted*, 2009.
[13] P. Hande, S. Patil, and H. Myung, "Distributed load-balancing in a multi-carrier system," in *Proc. IEEE Wireless Comm. and Net. Conf.*, 2009.
[14] H. Kim, X. Yang, M. Venkatachalam, Y.-S. Chen, K. Chou, I.-K. Fu, and P. Cheng, "Handover and load balancing rules for 16m," in *IEEE C802.16m-09/0136r1*, Jan. 2009, pp. 1024–37.
[15] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *Proc. IEEE INFOCOM*, vol. 1, pp. 321–331, Apr. 2003.
[16] G. Bianchi and I. Tinnirello, "Improving load balancing mechanisms in wireless packet networks," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, 2002, pp. 891–95.
[17] H. Velayos, V. Aleo, and G. Karlsson, "Load balancing in overlapping wireless LAN cells," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Jun. 2004, pp. 3833–36.
[18] 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.
[19] T. Bonald and J. W. Roberts, "Congestion at flow level and the impact of user behaviour," *Computer Networks*, pp. 521–536, 2003.
[20] H. Kim and G. de Veciana, "Losing opportunism: evaluating service integration in an opportunistic wireless system," in *Proc. IEEE INFOCOM*, 2007.
[21] J. Walrand, *An introduction to queueing networks*, Prentice Hall, 1998.
[22] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "α-optimal user association and cell load balancing in wireless networks," *Technical Report, UT-Austin, available at http://users.ece.utexas.edu/ hkim4/index_files/alpha_optimal2009.pdf*, 2009.