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

Hongseok Kim, *Member, IEEE*, Gustavo de Veciana, *Fellow, IEEE*, Xiangying Yang, *Member, IEEE*, and Muthaiah Venkatachalam, *Associate Member, IEEE*

***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. We then address admission control policies for the case where the system is overloaded. For an appropriate system-level cost function, the optimal admission control policy blocks all flows at cells edges. However, providing a minimum level of connectivity to all spatial locations might be desirable. To this end, a location-dependent random blocking and user association policy are proposed.**

***Index Terms*—Delay-optimal, flow-level dynamics, load balancing, throughput-optimal, user association, wireless network.**

I. Introduction

FOURTH-GENERATION wireless cellular standards such as IEEE 802.16m 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 multicell 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 considers neither intercell interference nor cell load balancing.

There have been many efforts in the literature toward developing user association rules considering interference avoidance and/or cell load balancing [3]–[13]. To avoid interference when frequency is universally reused and intercell 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 intercell 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]. 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 timescales. 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 intercell coordinated, i.e., centralized scheduling. Therefore, our approach is reasonable when fractional frequency reuse or interference randomization are being used so that intercell 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], and [8] explores flow-level dynamics for load balancing, but assumes a centralized controller. In particular, none of these efforts fully explores the role of load balancing under spatially *inhomogeneous* traffic distributions in a distributed way.

Recently, [18] includes an analysis of the stability/capacity of systems with (and without) server interaction. In the case

Manuscript received February 16, 2010; revised January 05, 2011; accepted May 14, 2011; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor A. Proutiere. Date of publication June 13, 2011; date of current version February 15, 2012. This work was supported in part by the Intel Research Council and the NSF under Award CNS-0721532.

H. Kim is with the Department of Electronic Engineering, Sogang University, Seoul 121-742, Korea (e-mail: hongseok@ieee.org).

G. de Veciana is with the Wireless Networking and Communications Group (WNCG), The University of Texas at Austin, Austin, TX 78712 USA (e-mail: gustavo@ece.utexas.edu).

X. Yang and M. Venkatachalam are with the Wireless Standard Group, Intel Corporation, Hillsboro, OR 97124 USA (e-mail: xiangying.yang@intel.com; muthaiah.venkatachalam@intel.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2011.2157937

of no server interaction (e.g., static intercell interference), they characterize the system stability/capacity region based on static server assignments. In particular given *a priori* measurements for the spatial traffic loads, they propose an optimization that would in principle result in a static assignment achieving the largest proportional increase of the load (i.e., capacity). By contrast, the distributed α-optimal user association policy developed in this paper is aimed to adapt to changing loads. It does so without requiring direct knowledge of the spatial traffic loads and has the aim of not only achieving stability, when possible, but also of minimizing mean flow delay when flow scheduling is temporally fair. Based on our approach, we also devise a framework for admission control to be used when the system is overloaded or cannot be stabilized. In the case of server interaction, [18] explicitly characterizes the stability region for a two-server system and provides a lower bound on the stability region for a multiple-server system. However, no practical user association rule was given. The characteristics of performance-optimal user association polices in systems with dynamic interaction are studied in [11]. The insights developed therein suggest an intriguing practical adaptive heuristic policy that performs quite well [19].

One of the main challenges in developing a distributed user association policy is achieving a 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 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: 1) Cell loads are not interchangeable, and *balancing* loads to minimize delay does not imply *equalizing* loads at the BSs. 2) Voronoi cells need not be delay optimal even if the traffic loads are spatially homogeneous. 3) 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 be easily extended to cover handover by applying proper triggers and target ranking that govern network mobility management control [14].

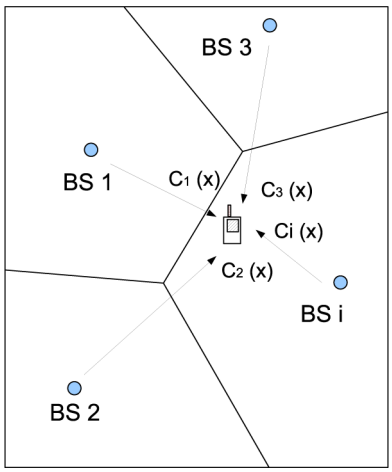


Fig. 1. User association problem considering the capacity and the traffic loads.

Third, we further extend our α-optimal user association to admission control policies, which have not been discussed so far in the literature. Specifically, we consider possible admission control policies when the system cannot be stabilized or is subject to excessively high loads. The work in [20] and [21] suggests that admission control is indeed required for best effort traffic in these circumstances. The optimal policy that minimizes our generalized system-level performance function plus blocking cost results in blocking flows around the boundaries of BS coverage areas. In practice, this may not be desirable, so we propose a policy that admits flows at the cell edge with a fixed probability, giving a minimum level of "connectivity" to all users.

Organization: The paper is organized as follows. In Section II, we describe our system model and assumptions. Section III is devoted to the distributed algorithm and the fixed-point optimality condition of user association under inhomogeneous traffic distribution. We prove the convergence of our algorithm in Section IV. We consider admission control in Section V and conclude the paper in Section VII.

II. System Model

A. Assumptions

We consider an infrastructure-based wireless communication system with multiple base stations; see Fig. 1. Target systems could be, but are not limited to, WiMAX2 or 3GPP-LTE. For simplicity, we focus on downlink user association, but our method is also applicable to the uplink user association and could perhaps be combined to address more complex scenarios, as long as intercell interference can be reasonably assumed to be static. We assume that other cell interference is static and can be considered as noise [4], [10], [13]. We consider a region $\mathcal{L} \subset \mathbb{R}^2$ that 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 that 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. Similar analysis on the spatial flow-level dynamics was done in [15] and [22], but for the homogeneous Poisson point process with uniform arrival rate, i.e., $\lambda(x) = \lambda$.

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.

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

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

$$c_i(x) = \log_2(1 + \mathrm{SINR}_i(x)) \tag{2}$$

where $\mathrm{SINR}_i(x)$ 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 *location-dependent*, but *static* interference, i.e., another Gaussian-like noise [1], [2]. This static intercell interference model has also been adopted in previous load-balancing work [4], [10], [13]. We discuss in detail about our assumption on the static intercell interference at the end of this section. The $\mathrm{SINR}_i(x)$ is thus given by

$$\mathrm{SINR}_i(x) = \frac{P_i g_i(x)}{\sigma^2 + I_i(x)} \tag{3}$$

where P_i denotes the transmission power of BS i, and $g_i(x)$ 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 timescale for measuring $g_i(x)$ is assumed to be much larger. In addition, σ^2 denotes noise power and $I_i(x)$ denotes 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

$$\varrho_i(x) := \frac{\gamma(x)}{c_i(x)}$$

TABLE I
NOTATION SUMMARY

α	degree of load balancing
x	location in continuous space $\mathcal{L}$
$i \in \mathcal{B}$	BS index
b	$:=\|\mathcal{B}\|$, the number of the BSs
$\lambda(x)$	flow arrival rate per unit area
$1/\mu(x)$	average file size at x
$\gamma(x)$	$:= \frac{\lambda(x)}{\mu(x)}$, inhomogeneous traffic load density
$c_i(x)$	the physical capacity at x from BS i
$\varrho_i(x)$	$:=\frac{\gamma(x)}{c_i(x)}$ system-load density (fractional time)
$p_i(x)$	the routing probability to BS i at x
$p(x)$	$:=(p_1(x), \cdots, p_b(x))$
$q_i(x)$	the binary-valued routing probability to i at x
ρ_i	$:=\int_{\mathcal{L}} \varrho_i(x)p_i(x)dx$ or $\int_{\mathcal{L}_i} \varrho_i(x)dx$
ρ	$:= (\rho_1, \cdots, \rho_b)$
$\mathcal{L}_i$	the coverage of BS i
$\mathcal{P}$	$:=\{\mathcal{L}_1, \cdots, \mathcal{L}_b\}$, the spatial partition
$\mathcal{F}$	a set of feasible ρ
$\partial\mathcal{F}^o$	a set of $T(\rho)$, $\rho \in \mathcal{F}$
$T(\rho)$	a mapping function in $\mathbb{R}^b$
β	exponential averaging parameter
$S(\rho)$	$:= \beta\rho + (1-\beta)T(\rho)$
(k)	iteration index in the superscript

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. Our notation is summarized in Table I.

This paper focuses on scenarios where users see (a roughly) static interference from neighboring cells. In what follows, we discuss when this is likely to be the case. Dynamic interference is most problematic in systems with a frequency reuse of 1, i.e., when all cells (or sectors) operate on the same frequency. In such a scenario, the interference will vary significantly depending on the activity in neighboring cells, possibly coverage to QoS, e.g., voice/video users at the cell edge.

For this reason, upcoming practical wireless systems such as WiMAX2 and LTE-Advanced include different strategies to mitigate the impact of intercell interference by ensuring adjacent cells (or sectors) operate in different frequencies. For example, when frequency reuse 3 is used, each cell (or each sector) can have either one of three different frequencies, say F1, F2, or F3. However, in order to maximize the spectrum efficiency, a frequency reuse 1 and 3 can be used together [23], i.e., frequency reuse 1 for the cell center and frequency reuse 3 for the cell edge. This approach has been generalized in IEEE 802.16 m, where reuse-3 is achieved in the same carrier frequency via partial bandwidth usage; this is usually referred to as fractional frequency reuse (FFR). A TDM-based enhanced Inter-Cell-Interference-Cancellation (eICIC) approach has also been developed to ensure that mobile terminals see no interference from the closest interfering cells [24]. In addition to FFR/eICIC, the cell radius is carefully chosen, taking into consideration power budgets and path loss. This typically results in interfering cells (or sectors) that are sufficiently separated so that interference from the cells operating on the same frequency becomes negligible.

When the above mechanisms are in place, even if the intercell interference depends on the activity of neighboring cells/sectors, the variation (and the interference itself) can be reasonably neglected and modeled perhaps as an (averaged) static value. We note, however, that studying the one with frequency reuse 1 where dynamic intercell interference plays a role is still of interest as it can lead to a larger stability region and better performance, e.g., delay for best-effort flows. For this reason, there has been quite a bit of work in this direction, see e.g., [5], [11], [18], [19], and [25].

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

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

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

$$\sum_{i \in \mathcal{B}} p_i(x) = 1, \tag{6}$$

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

where ϵ is an arbitrarily small positive constant. Hence, the feasible BS loads ρ has the associated routing probability vector $p(x) = (p_1(x), \ldots, p_b(x))\ \forall x \in \mathcal{L}$.

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 $p^1(x) = (p_1^1(x), \ldots, p_b^1(x))$ and $p^2(x) = (p_1^2(x), \ldots, p_b^2(x))$ 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)–(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:

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

where $\alpha \ge 0$ is a parameter specifying the desired degree of load balancing. When $\alpha = 1$, the objective function is defined as $\sum_i \log(\frac{1}{1-\rho_i})$. Our objective function has a similar form with the α-fair utility function [26]. However, we have a notion of cost instead of utility, so we need to minimize it. In addition, instead of the functional similarity, the implication of α is not exactly matched. Problem 1 is said to be feasible if $\mathcal{F}$ is nonempty. Otherwise, we shall require admission control, which will be discussed in Section V.

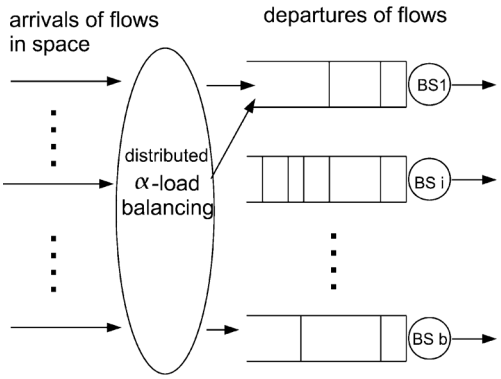


Fig. 2. Flow-level queueing model for user association problem.

C. Motivation for the Objective Function

Now, we describe the motivation for our objective function (8). 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. 2, which tracks the arrival and departure processes of users (or flows, file requests); see e.g., [27]–[29], [34].

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 $M/GI/1$ multiclass processor sharing system [30], and given by $\pi_i(n_i) = (1-\rho_i)\rho_i^{n_i}$. Multiclass 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 $\sum_i (\frac{\rho_i}{1-\rho_i} + 1) = \sum_i \frac{1}{1-\rho_i}$, 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

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

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

Remark 2.2 (Deterministic User Association is Optimal): It should be noted that the optimal user association rule (9) is indeed *deterministic* under Assumption 2.1 even though we formulate the problem with *probabilistic* user association by introducing $p_i(x)$. This will be made clear in the proof of Theorem 1.

Remark 2.3 (Tie-Breaking): A location $x \in \mathcal{L}$ is called a cell boundary if there is a tie in the argmax operation in (9) 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 a *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 a 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 the user's perspective to the 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 $\alpha \geq 1$ 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.

Remark 2.4: Hence, it should be noted that load balancing does not necessarily imply equalizing the loads of all BSs; different values of α have different implications.

Remark 2.5: There are some (but not exact) analogies between α-optimal user association and α-fair utility. When $\alpha = 0$, load balancing is ignored, which is similar to the case when fairness is ignored. When α goes to infinity, the user association becomes the min-max utilization policy, which is similar to the case of max-min fairness. However, for other values of α, the exact similarity is not present.

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 MTs use this information to make decisions over these periods. Since this algorithm is totally distributed, i.e., does not require any centralized computation, we do not have an algorithmic complexity issue here. 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 Timescales): We shall assume the flow arrival and departure processes are 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 kth period, MTs receive $\rho^{(k)}$, e.g., through broadcast control messages from BSs.[1] Then, a new flow request for an MT located at x simply selects the BS $i(x)$ using the deterministic rule given by

$$i(x) = \underset{j\in\mathcal{B}}{\mathrm{argmax}}\, c_j(x)\left(1-\rho_j^{(k)}\right)^{\alpha} \qquad \forall x \in \mathcal{L}. \tag{10}$$

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

$$\mathcal{L}_i^{(k)} = \left\{ x \in \mathcal{L} | i = \underset{j\in\mathcal{B}}{\mathrm{argmax}}\, c_j(x)\left[1-\rho_j^{(k)}\right]^{\alpha} \right\} \qquad \forall i \in \mathcal{B}. \tag{11}$$

Base Station: During the kth period, BSs *measure* their average utilizations. Due to Assumption 3.1, the measured utilization of BS i converges to some value, denoted by $T_i(\rho^{(k)})$

$$T_i\left(\rho^{(k)}\right) = \min\left[\int_{\mathcal{L}_i^{(k)}} \varrho_i(x)dx, 1-\epsilon\right] \qquad \forall i \in \mathcal{B}. \tag{12}$$

Note that BS i simply measures its utilization, yet this is mathematically captured by (12). In addition, the measured utilization, of course, cannot exceed 1. Hence, to avoid unnecessary technicalities, we introduce an arbitrarily small positive constant ϵ. It can be shown that $T(\rho) = (T_1(\rho), \ldots, T_b(\rho))$ is a continuous mapping defined on $[0, 1-\epsilon]^b$ to itself. Note that mapping $T(\rho)$ is the *mathematical model* capturing the user association dynamics, i.e., when BSs broadcast ρ, and the associated user association policy is followed, the BSs will eventually see a new load vector $T(\rho)$.

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

$$\begin{aligned}\rho^{(k+1)} &= \beta^{(k)}\rho^{(k)} + \left(1-\beta^{(k)}\right)T\left(\rho^{(k)}\right)\\ &:= S\left(\rho^{(k)}\right)\end{aligned} \tag{13}$$

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 kth period while $\rho^{(k)}$ is an exponential average of $T(\rho^{(\ell)})$ across periods, i.e., $\ell = 0, \ldots, k-1$ with some initial loads $\rho^{(0)} \in \mathcal{F}$.

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

B. Fixed Point Achieves Optimality

Note that if $\rho^{(k)}$ converges, it must converge to a fixed point of (13), i.e., a solution to

$$\rho^* = T(\rho^*). \tag{14}$$

The proof that (13) converges to ρ^* is provided in Section IV. Below, we will show that $T(\cdot)$ has a unique fixed point ρ^* corresponding to the optimal load vector associated with Problem 1.

Theorem 1: Suppose that Problem 1 is feasible. Then, T has a unique fixed point that 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. Since $\phi_\alpha(\rho)$ is a convex function over a convex set $\mathcal{F}$, if ρ^* satisfies the following condition:

$$\langle \nabla \phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{15}$$

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 (11), (12), and (14), the fixed point ρ^* generates the *deterministic* cell coverage, and thus the association rule is also *deterministic*, i.e.,

$$p_i^*(x) = \mathbf{1}\left\{ i = \operatorname*{argmax}_{j\in\mathcal{B}} c_j(x)(1-\rho_j^*)^\alpha \right\} \tag{16}$$

and then the inner product can be computed such as

$$\begin{aligned}\langle \nabla \phi_\alpha(\rho^*), \Delta\rho^* \rangle &= \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^*)^\alpha}(\rho_i - \rho_i^*) \\ &= \sum_{i\in\mathcal{B}} \frac{\int_{\mathcal{L}} \varrho_i(x)(p_i(x) - p_i^*(x))dx}{(1-\rho_i^*)^\alpha} \\ &= \int_{\mathcal{L}} \gamma(x) \left[\sum_{i\in\mathcal{B}} \frac{p_i(x) - p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha} \right] dx. \end{aligned} \tag{17}$$

Note that

$$\sum_{i\in\mathcal{B}} \frac{p_i(x)}{c_i(x)(1-\rho_i^*)^\alpha} \geq \sum_{i\in\mathcal{B}} \frac{p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha}$$

holds because $p_i^*(x)$ in (16) 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, one 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 nonzero 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 associate 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 that have zero traffic measure. ■

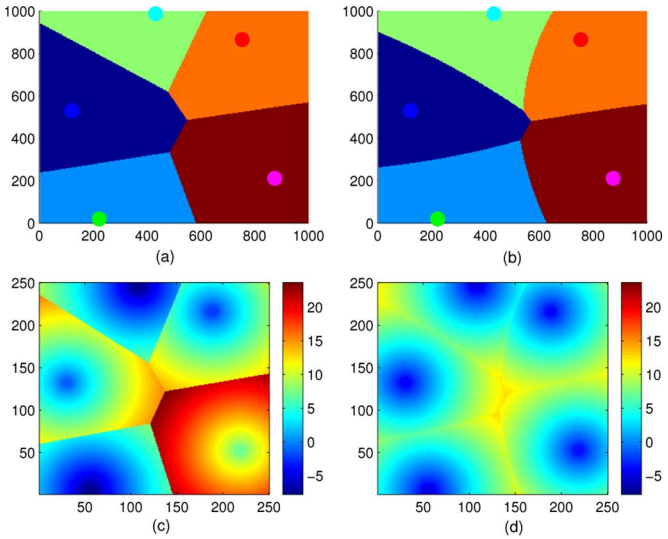


Fig. 3. (a), (b) Voronoi cells versus delay-optimal cells and (c), (d) spatial distribution of conditional average delay (dB scale) in each case.

C. Examples

We provide some examples to exhibit the properties of α-optimal user association.

Example 1: Rather than computing the utilization from busy fractional time, utilizations can 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 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

$$i(x) = \operatorname*{argmax}_{j\in\mathcal{B}} \frac{c_j(x)}{E[N_j]+1}. \tag{18}$$

This rule was proposed as a heuristic in [4], [10], and [13], which turns out to be a special case of our α-optimal user association.

Example 2 (Spatial Delay Smoothing): This example shows the BS coverage areas and geographical distribution of average file transfer delays. Five BSs are randomly placed in a 1000×1000 m^2 region. As an example of inhomogeneous traffic loads, a linearly increasing load along the diagonal direction from the bottom left to the top right 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 a path-loss exponent 3. Fig. 3(a) shows the partition when $\alpha = 0$ (Voronoi cells), and Fig. 3(b) shows the partition when $\alpha = 2$ (delay-optimal cells). Fig. 3(c) and (d) shows the conditional average file transfer delays (dB scale) at x, which is given by

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

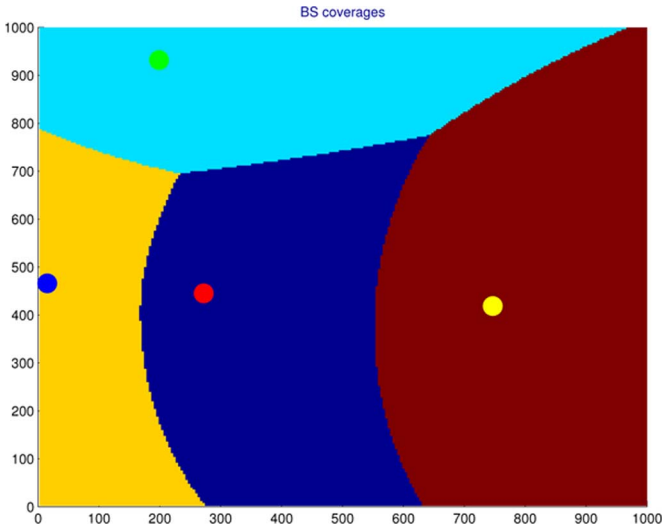


Fig. 4. Delay-optimal cells obtained for spatially homogeneous traffic loads. Voronoi cells are not delay-optimal even if the traffic loads are spatially homogeneous.

in the case of an $M/GI/1$ multiclass processor sharing system model. For simplicity, we set $1/\mu(x) = 1$ and show the average 1-b transmit time. The benefit of delay-optimal load balancing is clearly exhibited in Fig. 3. A slight modification of the cell coverages significantly improves the delay performance, specifically, for the congested cell at the lower right corner.

Example 3 (Voronoi Cells Versus Delay-Optimal Cells): One might think that Voronoi cells are delay-optimal for homogeneous traffic loads. However, this is not necessarily true. Consider a case where the traffic loads are homogeneous, i.e., $\lambda(x) = \lambda$ and $1/\mu(x) = 1/\mu$. Then, from (16), the delay-optimal cell boundary ℓ_{ij} for two adjacent cells $\mathcal{L}_i$ and $\mathcal{L}_j$ is given by

$$\ell_{ij} = \{x | c_i(x)(1-\rho_i^*)^2 = c_j(x)(1-\rho_j^*)^2\}. \qquad (19)$$

Since $c_i(x) = c_j(x)$ at the Voronoi cell boundaries, (19) is satisfied when $\rho_i^* = \rho_j^*$. However, two adjacent Voronoi cells do not necessarily have the same loads, i.e., $\rho_i^* = \rho_j^*$. In fact, Voronoi cells are delay-optimal *only if* in addition Voronoi cells have the same loads, which can be achieved when all the BSs are deployed symmetrically, e.g., hexagonal structure. Fig. 4 shows an example of delay-optimal cells that are far from Voronoi cells even though the traffic loads are homogeneous.

Remark 3.1: In general, α-optimal user association gives the following cell boundary:

$$\ell_{ij} = \{x | c_i(x)(1-\rho_i^*)^\alpha = c_j(x)(1-\rho_j^*)^\alpha\}. \qquad (20)$$

We note that [18] also characterizes the cell boundary associated with the capacity achieving static association policy for a given traffic load. In this scenario, they show the ratio of $c_i(x)$ and $c_j(x)$ needs to be constant at the cell boundary. By contrast, (20) explicitly characterizes the load-dependent ratio. These cell boundaries need not be the same, as they correspond to different objective functions.

Example 4 (Fragmented Cells): One might think that coverage areas associated with BSs should be contiguous. However, optimal BS coverage areas may be fragmented.

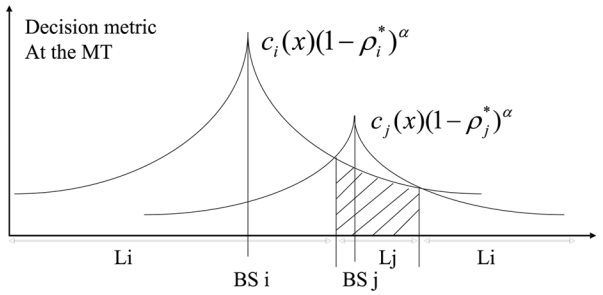


Fig. 5. Illustration of fragmented cell coverage areas.

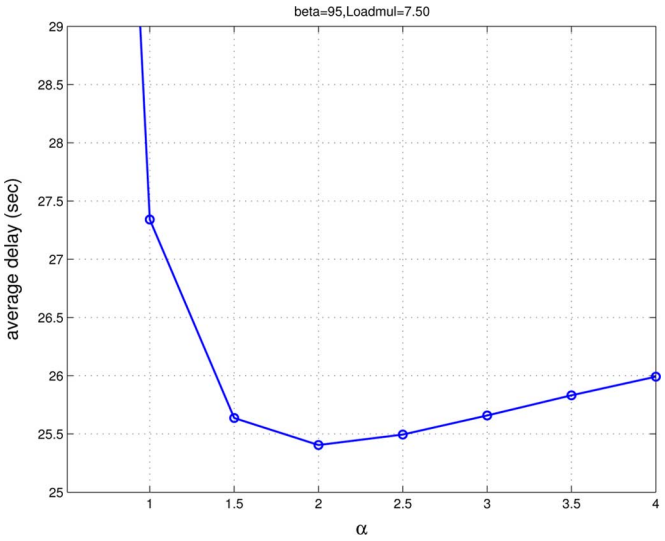


Fig. 6. Average delay obtained for different values of α. When $\alpha = 2$, minimum average delay is achieved.

Fragmented cells can exist because $(1-\rho_i^*)^\alpha$ and $(1-\rho_j^*)^\alpha$ in (19) play a role in determining the boundary. Fig. 5 illustrates $c_i(x)(1-\rho_i^*)^\alpha$ and $c_j(x)(1-\rho_j^*)^\alpha$ in 1-D and shows how noncontiguous coverage areas may arise depending on ρ_i^* and ρ_j^* even if two BSs have the *same* Tx power.

Example 5 (Delay for Various α): Fig. 6 shows the average delay performance for different α. Four BSs are randomly placed on 1000×1000 m^2. For illustrative purposes, the traffic loads are chosen to grow linearly along the diagonal direction. We exclude the results when $\alpha < 1$ because they result in excessive delays. It can be clearly seen that $\alpha = 2$ minimizes the average delay.

IV. Convergence of Distributed Iteration

In this section, we prove that the distributed α-optimal user association algorithm converges to the global optimum load vector ρ^*. When $\alpha = 0$, $T(\rho)$ is constant and (13) with $\beta = 0$ converges in one iteration, so hereafter we focus on the case when $\alpha > 0$.

A. Proof of Convergence

As seen earlier, the proposed algorithm can be interpreted as iteratively applying the mapping S in (13) to an initial load $\rho^{(0)}$. We shall prove the convergence of the loads by first considering the characteristics of the T mapping. If T were a contraction mapping, then iterating T would guarantee convergence to the

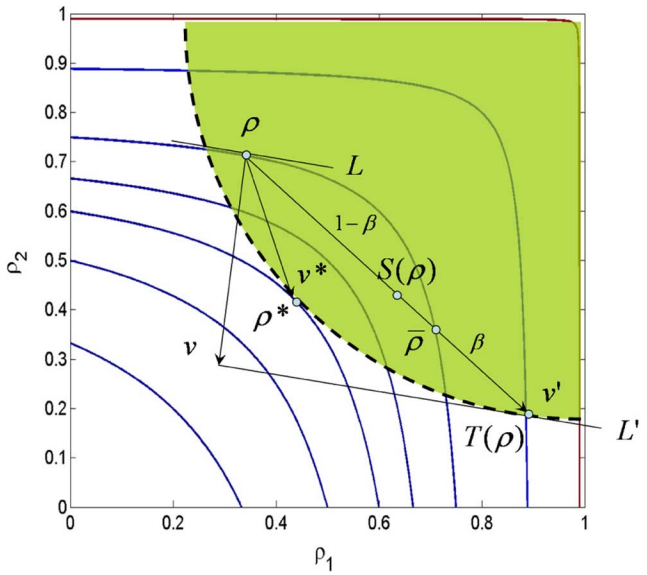


Fig. 7. Convergence property of S mapping. The shaded region represents the convex set $\mathcal{F}$, and the dashed line represents the set $\{T(\rho)|\rho \in \mathcal{F}\}$. In addition, $v = -\nabla\phi_\alpha(\rho)$, and L is the tangent line of the level set at ρ.

unique fixed point associated with the global optimum. However, T is not necessarily a contraction mapping, in particular when the system is highly loaded. Therefore, the proposed algorithm is a damped version of T, i.e., $S(\rho) = \beta\rho + (1-\beta)T(\rho)$. We first show the following two lemmas associated with the T mapping, and then prove the convergence of the S mapping.

Lemma 2: If $\rho \in \mathcal{F}$, then $T(\rho)$ is on the boundary of $\mathcal{F}$ that faces the origin; see, e.g., Fig. 7.

The proof is included in that of Lemma 3.

In the case of two BSs, the fixed point of T can be visualized as shown in Fig. 7. The dashed line denotes a set $\partial\mathcal{F}^o = \{T(\rho)|\rho \in \mathcal{F}\}$, i.e., the boundary of $\mathcal{F}$ facing the origin. Since the level sets of $\phi_\alpha(\rho)$ are concave functions (solid lines), ρ^* is the point where the level set touches a convex set $\mathcal{F}$. Note that the shape of $\mathcal{F}$, i.e., the shaded region in the figure, and ρ^* depend on the spatial traffic distribution.

Remark 4.1: From (11) and (12), $T(\rho)$ is associated with deterministic BS coverage areas, and the routing probability that specifies $T(\rho)$ is binary, i.e., either 1 or 0. Hence, in describing the routing probability associated with $T(\rho)$, we will use the notation $q_i(x) \in \{0, 1\}$ instead of $p_i(x)$.

Next, we show two *key* properties of T mapping. The first is that $T(\rho) - \rho$ is a descent direction of $\phi_\alpha(\rho)$. The second is that $T(\rho) - \rho$ is a vector that minimizes the inner product with $\nabla\phi_\alpha(\rho)$. This is formally stated in the following lemma.

Lemma 3 (Descent Direction): For $\rho \in \mathcal{F}$ and $\rho \neq \rho^*$, $T(\rho)$ gives a descent direction at ρ, i.e.,

$$\langle \nabla\phi_\alpha(\rho), T(\rho) - \rho\rangle < 0.$$

In addition, $T(\rho)$ is the feasible load vector that minimizes the inner product with the gradient at ρ, i.e.,

$$T(\rho) = \underset{\hat{\rho}\in\mathcal{F}}{\operatorname{argmin}} \langle \nabla\phi_\alpha(\rho), \hat{\rho} - \rho\rangle. \tag{21}$$

Proof: Let $p(x)$ and $q(x)$ be the routing probability associated with ρ and $T(\rho)$, respectively. From (12), T_i is associated with deterministic cell coverage area $\mathcal{L}_i$, and thus its routing probability $q_i(x)$ is given by binary, i.e.,

$$q_i(x) = \mathbf{1}\left\{i = \underset{j\in\mathcal{B}}{\operatorname{argmax}}\, c_j(x)(1-\rho_j)^\alpha\right\} \qquad \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \tag{22}$$

with ties broken in favor of lowest index BS. Let $\Delta\rho = T(\rho) - \rho$. Then, $\langle \nabla\phi_\alpha(\rho), \Delta\rho\rangle$ can be computed as follows:

$$\begin{aligned}\langle \nabla\phi_\alpha(\rho), \Delta\rho\rangle &= \sum_{i\in\mathcal{B}} \frac{T_i(\rho) - \rho_i}{(1-\rho_i)^\alpha} \\ &= \sum_{i\in\mathcal{B}} \frac{\int_{\mathcal{L}} \varrho_i(x)\,(q_i(x) - p_i(x))\,dx}{(1-\rho_i)^\alpha} \\ &= \int_{\mathcal{L}} \gamma(x) \left(\sum_{i\in\mathcal{B}} \frac{q_i(x) - p_i(x)}{c_i(x)(1-\rho_i)^\alpha}\right) dx.\end{aligned}$$

By definition, $q_i(x)$ satisfies

$$\sum_{i\in\mathcal{B}} \frac{q_i(x) - p_i(x)}{c_i(x)(1-\rho_i)^\alpha} \leq 0. \tag{23}$$

Since $\rho \neq \rho^*$, it must be that $p(x) \neq q(x)$ on a set that has nonzero traffic load measure. Then, multiplying (23) by $\gamma(x)$ and integrating over $\mathcal{L}$ gives $\langle \nabla\phi_\alpha(\rho), \Delta\rho\rangle < 0$.

Furthermore, we have the following property:

$$q_i(x) = \underset{\hat{p}_i(x)}{\operatorname{argmin}} \sum_{i\in\mathcal{B}} \frac{\hat{p}_i(x) - p_i(x)}{c_i(x)(1-\rho_j)^\alpha} \tag{24}$$

because (23) holds for arbitrary $p_i(x) \neq q_i(x)$. Then, multiplying (24) with $\gamma(x)$ and integrating (24) over $\mathcal{L}$ proves (21). Finally, (21) implies that $T(\rho)$ is on the boundary of $\mathcal{F}$, and Lemma 2 is proved. ■

Fig. 7 exhibits $T(\rho)$. Suppose that v is the opposite direction of $\nabla\phi_\alpha(\rho)$ and L is the tangent line of the level set at ρ. Then, the feasible vector that maximizes the inner product with v can be found by drawing a line L' that is parallel to L and tangent to the boundary $\partial\mathcal{F}^o$; the tangent point is then $T(\rho)$. In Fig. 7, we see the case of $\phi_\alpha(T(\rho)) > \phi_\alpha(\rho)$, which implies that $T(\rho)$ gives a descent direction, but it does not necessarily result in a monotonic decreasing sequence $\phi_\alpha(\rho^{(k)})$. Indeed, T mapping can overshoot along the descent direction, in particular when the loads are high. Introducing the weighting parameter β in (13) alleviates such overshooting. Fig. 7 shows $\phi_\alpha(S(\rho)) < \phi_\alpha(\rho)$ if $S(\rho)$ is selected between ρ and $\bar{\rho}$, where $\bar{\rho}$ is the intersection of $T(\rho) - \rho$ and the level set at ρ. Based on this, we prove the convergence of S iteration in Lemma 4 and Theorem 2.

Lemma 4: For $\rho \in \mathcal{F}$ and $\rho \neq \rho^*$, there exists $\beta \in [0, 1)$ such that $\phi_\alpha(S(\rho)) < \phi_\alpha(\rho)$.

Proof: Since $S(\rho) - \rho = \beta\rho + (1-\beta)T(\rho) - \rho = (1-\beta)(T(\rho) - \rho)$, $S(\rho) - \rho$ is also a descent direction. Since the level sets of $\phi_\alpha(\rho)$ are strictly concave functions when $\alpha > 0$ and $S(\rho)$ gives a descent direction at ρ, there exists a $\beta \in [0, 1)$ that makes $\phi_\alpha(S(\rho)) < \phi_\alpha(\rho)$. ■

Theorem 2 (Convergence): Suppose that Problem 1 is feasible. If $\rho^{(0)} \in \mathcal{F}$ and $\beta^{(k)}$ is chosen so that $\phi_\alpha(S(\rho^{(k)})) < \phi_\alpha(\rho^{(k)})$, then $\rho^{(k+1)} = S(\rho^{(k)})$ converges to ρ^*.

Proof: $\phi_\alpha(\rho^{(k)})$ is a monotonically decreasing sequence in k and also lower-bounded by 0, so $\phi_\alpha(\rho^{(k)})$ converges. Suppose

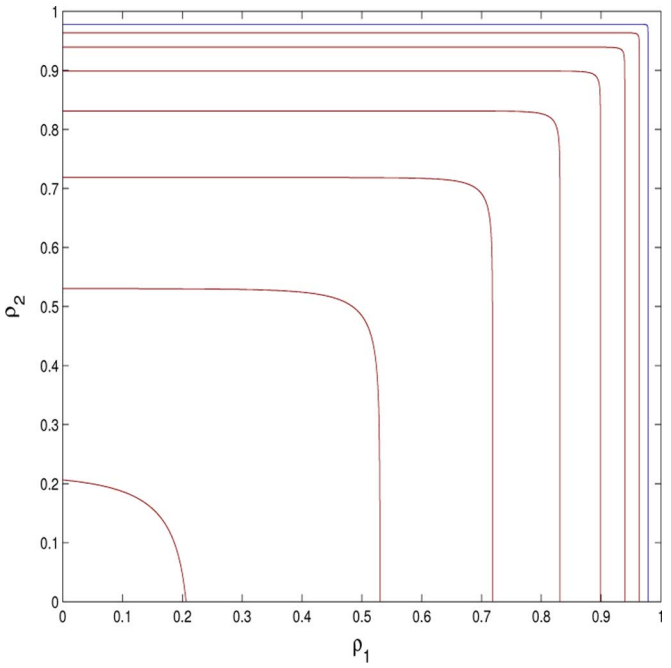


Fig. 8. Level sets of $\phi_\alpha(\rho)$ when $\alpha = 10$ (two-BSs case).

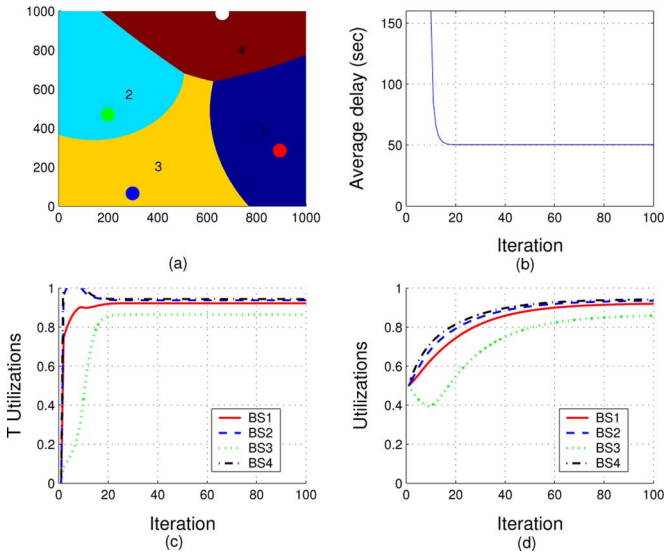


Fig. 9. Example of convergence: (a) delay-optimal partition; (b) average delay; (c) $T(\rho^{(k)})$; (d) $\rho^{(k)}$.

that $\phi_\alpha(\rho^{(k)})$ converges to something other than $\phi_\alpha(\rho^*)$. Then, S produces a descent direction again, and by Lemma 4, $\phi_\alpha(\rho^{(k)})$ can further decrease in the next iteration. This contradicts the convergence assumption, and $\rho^{(k)}$ should converge to ρ^*. ■

Remark 4.2: A fixed β close to 1 generally works well for the convergence. However, the magnitude of β guaranteeing convergence depends on the network load. When the system is not congested, even $\beta = 0$ can guarantee convergence as $T(\rho)$ may be a contraction mapping. However, when the system is congested, β needs to be close to 1, e.g., 0.95–0.99. The convergence speed also depends on β. When the loads are low, β can be small and exhibits fast convergence. In practice, β is a design parameter that should be selected to balance speed of convergence versus stable system behavior.

Remark 4.3: Note that MT decides its serving BS in a greedy way so as to maximize its own decision metric (11). Nevertheless, this algorithm converges to the global optimum. This is an interesting property because greedy behaviors of terminals degrade overall system performance in many cases.

Remark 4.4: As stated earlier without proof, the optimal solution equalizes ρ_i for all $i \in \mathcal{B}$ when $\alpha = \infty$. This can be easily proven when the level sets of $\phi_\alpha(\rho)$ are plotted. Fig. 8 shows the level sets when $\alpha = 10$. In fact, the level sets become more and more sharp as α grows, and thus the optimal utilization where the level set touches $\mathcal{F}$ occurs when ρ_i are all equal.

B. Convergence Independent of Initial Condition

So far, we assumed $\rho^{(0)} \in \mathcal{F}$, and then $\rho^{(k)}$ remains in the feasible set $\mathcal{F}$ during the iteration. One can, however, show that the iteration converges to the optimal point as long as $\rho^{(0)} \in [0, 1-\epsilon]^b$. This property is important in real implementation because it makes the algorithm *robust* to changes in the traffic spatial distribution. As an example, suppose that at time $t = t_0$, the stationary file arrival process with $\lambda^1(x)$ changes to another stationary process with $\lambda^2(x)$. However, the optimal solution for $\lambda^1(x)$ may not be in the feasible set associated with $\lambda_2(x)$. Nevertheless, our algorithm would converge to new optimal point. A proof of convergence for $\rho^{(0)} \in [0, 1-\epsilon]^b$ is given in the Appendix.

Example 6: Fig. 9 shows the convergence of two different utilizations: $T(\rho^{(k)})$ and $\rho^{(k)}$ when $\beta = 0.95$. The iteration can start at any $\rho^{(0)} \in [0, 1-\epsilon]^b$, so we simply pick up $\rho^{(0)} = (0.5, 0.5, 0.5, 0.5)$, which gives Voronoi cells at the first iteration. In this example, traffic loads are chosen so that Voronoi cells cannot stabilize the system. Hence, the delays would be infinite for the first few iterations; see Fig. 9(b) and (c). Nevertheless, our algorithm converges quickly to the optimal point.

Remark 4.5 (Throughput-Optimality): The proposed algorithm is throughput-optimal when $\alpha \geq 1$. This is because $\phi_\alpha(\rho)$ goes to infinity when ρ_i approaches 1. Then, if the system can be stabilized, there exists a partition $\mathcal{P}$ and corresponding ρ such that $\phi_\alpha(\rho) < \infty$. Then, $\phi_\alpha(\rho^*) \leq \phi_\alpha(\rho)$, i.e., $\phi_\alpha(\rho^*)$ is also finite. Since the algorithm converges to ρ^* for any $\rho^{(0)} \in [0, 1-\epsilon]^b$, it stabilizes the system if the system can be stabilized.

V. Admission Control

So far, we have assumed that Problem 1 is feasible, i.e., the system can be stabilized and Problem 1 has a solution. However, when the traffic loads are too high, the system may not be stabilizable or may perform very poorly so admission control is required. In this section, we consider admission policies for such regimes. Our objective is to minimize a system cost function that includes a cost associated with blocking flows. We assume that the blocking cost is proportional to the volume of the blocked traffic. Since flow blocking determines users' satisfaction and unsatisfied users may switch operators, such admission control policies would reflect operators' business concerns.

A. Optimality Condition

We assume the flows that are blocked are routed to a sink, or *null* BS. Let $\mathcal{B}_0$ denote a set of all BSs including the null BS, and redefine ρ as $\rho = (\rho_0, \rho_1, \ldots, \rho_b)$. It should be noted that ρ_0 is

not a utilization; it is defined as $\rho_0 := \int_{\mathcal{L}} \gamma(x)p_0(x)\,dx$, where $p_0(x)$ is the flow blocking probability at location x. Hence, ρ_0 can be greater than 1. The total blocking cost is given by $\xi\rho_0$, where ξ is a parameter that captures blocking cost per bit. We define a feasible set $\mathcal{F}_0$ including ρ_0 as

$$\mathcal{F}_0 = \Bigg\{\rho \quad | \quad \rho_0 = \int_{\mathcal{L}} \gamma(x)p_0(x)\,dx, \\ \rho_i = \int_{\mathcal{L}} \varrho_i(x)p_i(x)\,dx \qquad \forall i \in \mathcal{B}, \\ 0 \le \rho_i \le 1-\epsilon \qquad \forall i \in \mathcal{B}, \\ \sum_{i\in\mathcal{B}_0} p_i(x) = 1 \qquad \forall x \in \mathcal{L}, \\ 0 \le p_i(x) \le 1 \qquad \forall i \in \mathcal{B}_0 \text{ and } \forall x \in \mathcal{L}\Bigg\}.$$

It can be shown that $\mathcal{F}_0 \in \mathbb{R}^{b+1}$ is a convex set. Our objective function is then given by Problem 2.

Problem 2:

$$\min_{\rho}\left\{\phi_\alpha^\xi(\rho) = \sum_{i\in\mathcal{B}} \frac{(1-\rho_i)^{(1-\alpha)}}{\alpha-1} + \xi\rho_0 \,\middle|\, \rho \in \mathcal{F}_0\right\}. \quad (25)$$

Note that Problem 2 is a simple convex generalization of Problem 1. If the system can be stabilized, Problem 2 is equivalent to Problem 1 as ξ goes to infinity. An optimality condition for this problem based on which we can develop adaptive admission control and user association policy is proposed next.

We propose an iterative algorithm for Problem 2 whose behavior is similar to those given for Problem 1, i.e., (10)–(13). At the start of the kth period, MTs choose their serving BSs using (10), and BSs measure the utilization, which converges to (12) under the Assumption 3.1. BSs update their broadcast load vectors using (13). The difference is that the user association policy now involves admission control, i.e., though MT uses the same association rule, a BS may block an MT based on a threshold. As a consequence, the coverage area (11) is revised as follows:

$$\mathcal{L}_i^{(k)} = \left\{x \in \mathcal{L} | i = \operatorname*{argmax}_{j\in\mathcal{B}_0} \nu_j(x)\right\} \qquad \forall i \in \mathcal{B}_0 \quad (26)$$

where

$$\nu_j(x) = \begin{cases} \frac{1}{\xi}, & \text{if } j = 0 \\ c_j(x)\left(1-\rho_j^{(k)}\right)^\alpha, & \text{if } j \in \mathcal{B}. \end{cases} \quad (27)$$

Note that $\mathcal{L}_0^{(k)}$ denotes the area where flows are blocked. Intuitively, (26) and (27) say a BS blocks flows that do not see good performance as compared to threshold $1/\xi$. The threshold is the inverse of blocking cost per bit. Thus, if the blocking cost is high, the BS is less likely to block flows, and vice versa. The meaning of a threshold $1/\xi$ depends on α: When $\alpha = 0$, $1/\xi$ is simply the minimum achievable rate; when $\alpha = 1$, $1/\xi$ corresponds to the expected throughput, and thus the expected throughput of a MT at x, i.e., $\max_j c_j(x)(1-\rho_j^{(k)})$ needs to exceed $1/\xi$ in order to be admitted; when $\alpha = 2$, ξ corresponds to a maximum marginal 1 bit transmit time. Note that if the admission control is enforced, flows around the cell edge are first to be blocked if shadow fading is not considered. It is also reported in [5] that admission control under a heavily congested system blocks the flows around the cell edge. This, of course, is because users at the cell edge consume most of the system resources, i.e., time.

Following our previous approach, the optimal user association and admission control policy are related with the fixed point of a certain mapping. To derive it, in addition to (12), we define

$$T_0(\rho^{(k)}) := \int_{\mathcal{L}_0^{(k)}} \gamma(x)\,dx \quad (28)$$

and T is redefined on $[0,M]\times[0,1-\epsilon]^b$ to itself where $M < \infty$. It can be shown that $\rho^{(k)}$ converges to ρ^*, i.e., a fixed point of T, using the same technique as for Theorem 1.

Theorem 3: T has a unique fixed point that is the optimal solution to Problem 2. In addition, under Assumption 2.1, this fixed point determines a unique optimal spatial partition up to a set of traffic measure zero.

Proof: Since T is a continuous mapping defined on compact set $[0,M]\times[0,1-\epsilon]^b$ 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 2. Since $\phi_\alpha^\xi(\rho)$ is a convex function over a convex set $\mathcal{F}_0$, if ρ^* satisfies the following condition:

$$\left\langle \nabla\phi_\alpha^\xi(\rho^*), \Delta\rho^* \right\rangle \ge 0 \quad (29)$$

for all $\rho \in \mathcal{F}_0$ where $\Delta\rho^* = \rho - \rho^*$, then ρ^* is the optimal solution of Problem 2.

Let $p(x)$ and $p^*(x)$ be the associated routing probabilities for ρ and ρ^*, respectively. From (12), (14), (26), and (28), the fixed point ρ^* generates the *deterministic* cell coverage, and thus the association rule is also *deterministic*, i.e.,

$$p_i^*(x) = \mathbf{1}\left\{i = \operatorname*{argmax}_{j\in\mathcal{B}_0} \nu_j^*(x)\right\} \quad (30)$$

where $\nu^*(x)$ is given by (27) with ρ^*. Then, the inner product is computed such as $\langle \nabla\phi_\alpha^\xi(\rho^*), \Delta\rho^*\rangle$

$$\begin{aligned} &= \sum_{i\in\mathcal{B}_0} \frac{\partial\phi_\alpha^\xi}{\partial\rho_i}(\rho_i - \rho_i^*) \\ &= \sum_{i\in\mathcal{B}} \frac{\int_{\mathcal{L}} \varrho_i(x)\,(p_i(x) - p_i^*(x))\,dx}{(1-\rho_i^*)^\alpha} \\ &\quad + \xi\int_{\mathcal{L}} \gamma(x)(p_0(x) - p_0^*(x))dx \qquad (31) \\ &= \int_{\mathcal{L}} \gamma(x)\left[\sum_{i\in\mathcal{B}} \frac{p_i(x) - p_i^*(x)}{c_i(x)(1-\rho_i^*)^\alpha} + \xi(p_0(x) - p_0^*(x))\right]dx \\ &= \int_{\mathcal{L}} \gamma(x)\left[\sum_{i\in\mathcal{B}_0} \frac{p_i(x) - p_i^*(x)}{\nu_i(x)}\right]dx \\ &\overset{(a)}{\ge} 0 \qquad (32) \end{aligned}$$

where (a) follows from that $p_i^*(x)$ is the maximizer of $\nu_i^*(x)$ for $i \in \mathcal{B}_0$ (or the minimizer of their inverses). The uniqueness of the spatial partition can be proven similarly as that of Problem 1. ■

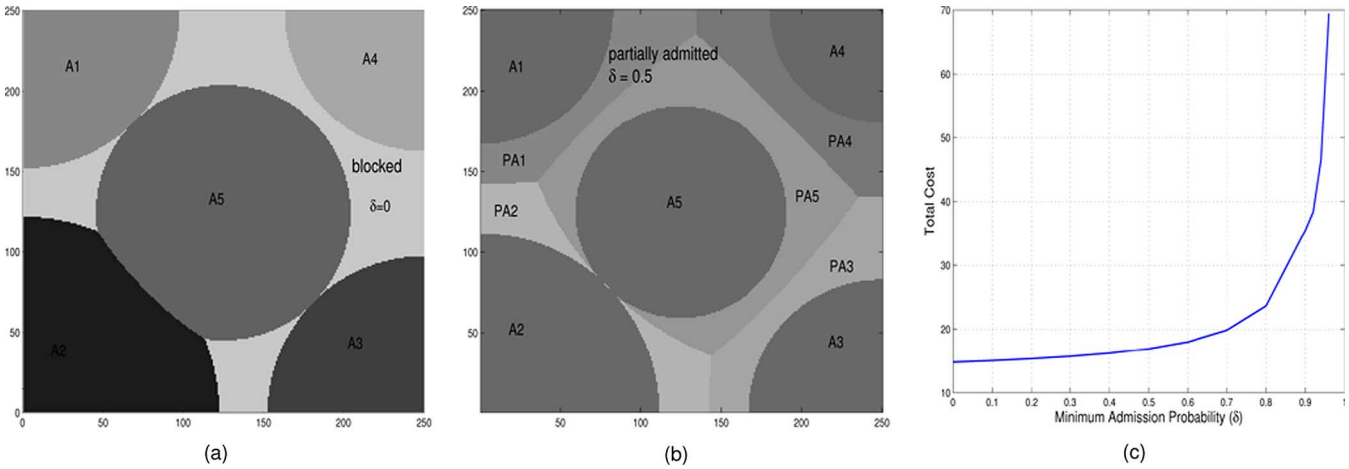


Fig. 10. Cell coverage areas under admission control. (a) $\delta = 0$. (b) $\delta = 0.5$. Each cell has completely admitted area A i and partially admitted area PA i. (c) Tradeoff between δ and performance.

B. Adding Flow Acceptance Probability Constraint

Fig. 10 shows an example of admission control policy when the traffic load is heavy. For illustrative purposes, BS1–BS4 are placed at the four corners, and BS5 at the center. Fig. 10(a) shows the coverage areas of five BSs, i.e., A1–A5. Traffic load increases along the diagonal direction and is such that the system is not stabilizable. As can be seen, the flows around the cell edge are blocked (bright gray areas). Note that the complete call blocking around the cell edge raises issues of fairness, and thus the service provider might want to admit the flows at cell edges even if they see poor performance. In addition, allowing some level of minimal connectivity might be beneficial from a higher-layer QoS perspective, i.e., the tradeoff between delay and service outage probability. However, providing such minimum connectivity will compromise overall delay performance and lead to additional blocking for customers closer to the BSs. To capture the tradeoff between the fairness and delay performance, we add the following constraint:

$$0 \le p_0(x) \le 1 - \delta \tag{33}$$

in $\mathcal{F}_0$, where δ specifies the minimum probability of flow acceptance.

Theorem 4: An optimal user association policy of Problem 2 with additional constraint (33) is still (9), but with probability $1 - \delta$, the flow is blocked if $\max_{i \in \mathcal{B}} c_i(x)(1 - \rho_i^*)^\alpha < 1/\xi$.

Proof: Since $\mathcal{F}^0$ is still a convex set with additional constraint of (33), it is sufficent to show that (32) is satisfied when ρ^* and its associated $p^*(x)$ are given as follows:

if $\max_{i \in \mathcal{B}} \nu_i^*(x) \ge \frac{1}{\xi}$

$$p_i^*(x) = \mathbf{1}\left\{ i = \underset{j \in \mathcal{B}}{\operatorname{argmax}}\, \nu_j^*(x) \right\} \qquad \forall i \in \mathcal{B}_0 \tag{34}$$

otherwise

$$p_0^*(x) = 1 - \delta, \tag{35}$$

$$p_i^*(x) = \delta \times \mathbf{1}\left\{ i = \underset{j \in \mathcal{B}}{\operatorname{argmax}}\, \nu_j^*(x) \right\} \qquad \forall i \in \mathcal{B}. \tag{36}$$

The proof is essentially the same as the proof of Theorem 3, and condition (a) in (32) is satisfied when $p_i^*(x)$ is chosen as stated in (34)–(36) because $p_i^*(x)$ gives most of its weight on at most two $\nu_i^*(x)$ under the constraint of (33). ■

Fig. 10(b) shows the areas where the flows are partially admitted with a fixed probability (denoted by PA i, $i = 1, \ldots, 5$) and completely admitted (denoted by A i). In this example, $\delta = 0.5$ is used. Comparing Fig. 10(a) and (b) shows that areas where flows are admitted with probability 1 shrink as δ increases from 0 to 0.5. In addition, increasing δ also degrades overall system performance. Fig. 10(c) shows its tradeoff: As δ grows, $\phi_\alpha^\xi(\rho)$ increases. Hence, δ should be carefully chosen considering the tradeoff between minimum probability of flow acceptance and performance degradation.

Remark 5.1: The addition of a minimum probability of flow acceptance δ irrespective of location means once more that the network may not be stabilizable. That is, even if flows are blocked with probability $1 - \delta$ everywhere, there may not exist a user association policy that stabilizes the remaining load. We envisage the service provider having sufficient knowledge of the traffic loads on its network to balance the selection of the two parameters ξ and δ: balancing blocking (and stability) versus flow-level performance.

VI. Discussion

A. α-Optimal Versus State-Dependent User Associations

Our proposed user association is based on estimating and adapting to a base station's long-term utilizations ρ. One might instead consider using a state-dependent policy, i.e., using the *current* number of flows. For example, one can consider the following state-dependent policy:

$$i(x,t) = \underset{j \in \mathcal{B}}{\operatorname{argmax}} \frac{c_j(x)}{n_j(t) + 1} \tag{37}$$

where $n_j(t)$ denotes the current number of flows served by the BS j at time t. The denominator is augmented by one to include the MT that would be joining BS j. Under this "greedy" policy, an MT associates with the BS that currently affords it the best *instantaneous* effective service assuming temporally fair scheduling.

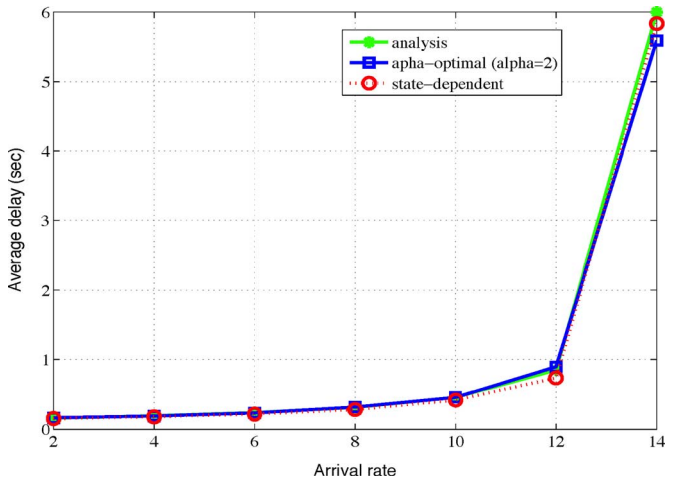


Fig. 11. Event-driven simulation for ρ-dependent and the state-dependent policies.

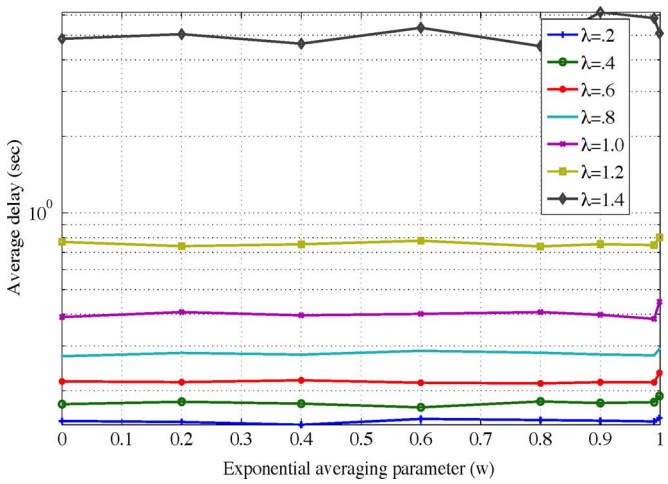


Fig. 12. Insensitivity of the averaging window size in estimating the load.

The performance for such a state-dependent policy is analytically intractable, so we shall determine the average flow-level delays via event-driven simulations. For illustrative purposes, we consider a linear network where two BSs are placed on a line segmented apart by 500 m, and $\lambda(x)$ is linearly increasing in x. In Fig. 11, the line with $*$ is for the optimal average delay obtained from analysis, and the lines with $\square$ and $\bigcirc$ demonstrate the delay-optimal user association and state-dependent user association given by (37), respectively. Interestingly, for this scenario, the state-dependent policy exhibits roughly the same performance as our proposed policy.

To further investigate state-dependent policies, let us compare (37) to

$$i(x) = \underset{j \in \mathcal{B}}{\operatorname{argmax}} \frac{c_j(x)}{E[N_j] + 1}. \tag{38}$$

Recall that (38) is α-optimal user association for $\alpha = 1$. To relate (37) to (38), consider estimating $E[N_j]$ exponential averaging, i.e., the estimated average number of flows up to time t is given by

$$\tilde{n}_j(t) = w\tilde{n}_j(t-1) + (1-w)n_j(t)$$

where $0 \leq w < 1$. Note that when w is zero, (38) reduces to (37). Thus, by varying w from 0 to 1, we can see the impact of the exponential averaging parameter (window) on delay performance. Interestingly, service capacity is shared based on processor-sharing service discipline, and the delay performance is not significantly affected by the averaging parameter; see Fig. 12. This "insensitivity" is beneficial in practice since, in principle, the BSs need not broadcast the load estimation too frequently. Note that implementing the state-dependent policy requires the current state to be broadcast whenever the state changes (i.e., the arrival or departure of flow happens); see the time index t of $i(x,t)$ in (37) instead of $i(x)$ in (38).

Note that although the state-dependent policy in (37) exhibits good delay performance, it may not stabilize the system under all traffic scenarios. Indeed, [31] proposed a state-dependent user association called *MinDrift server assignment*, which is throughput-optimal for a work-conserving discipline. However, it requires knowledge of the mean file size $1/\mu(x)$ generated by flows at location x, which in practice may not be available—e.g., flows generated at a coffee house might have larger means than those on the road. By contrast, our α-optimal user association does not require the knowledge of $1/\mu(x)$, but throughput-optimal when $\alpha \geq 1$; these quantities are implicitly estimated by the measured long-term utilizations of the base stations.

B. Applicability to OFDMA Systems

Since emerging broadband standards are based on OFDMA, e.g., WiMAX2 and LTE-Advanced, we briefly discuss the applicability of our work to such systems. The base station model discussed in this paper captures a system with a single resource that is time-shared (equally) among active users. OFDMA systems differ in that users can be scheduled (opportunistically) across various frequency bands in a manner that need not be temporally fair, e.g., to compensate for heterogenous user locations. We envisage two ways in which our model can be applied: The first is strict application, whereas the second is perhaps a more realistic relaxation of the model. First, consider a system with B base stations, each of which has C transport blocks, sets of frequencies. Transport blocks correspond to the minimum unit of spectrum that can be allocated to a user. Suppose a mobile terminal is associated with a specific transport block at a given base station, thus users see a $B \times C$ server system.[2] As long as users associated with a block are scheduled in a temporally fair way, our system model directly applies to this setting. Furthermore, to minimize the intercell interference, the available bands to users may depend on the location, e.g., cell center or cell edge may have different frequency band. Indeed, this is the combination of cell load balancing and interference avoidance and is already proposed for 4G OFDMA systems; see [10] and [13].

In our second scenario, all the frequency resources available at a base station are shared by all the users according to a chosen scheduler, e.g., proportionally fair scheduling

[2]The case where multiple transport blocks can be associated with the mobile terminal becomes a more complex problem because the number of blocks becomes another decision variable.

(for WiMAX2, see [32, Appendix F]; for LTE-Advanced, see [33, Section A.3]). The base station's utilization is simply the fraction of transport blocks utilized per frame. The overall cost in Problem 1 is interpreted as a natural proxy for congestion to be minimized. The base station's capacity to serve a given user is determined based on an averaged channel quality indication across transport blocks or otherwise estimated. Under these relaxed assumptions, the proposed user association policy is a natural fit for use in most OFDMA settings where intercell interference has been proactively mitigated.

VII. Conclusion

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 or LTE-Advanced. Our distributed algorithm converges to the global optimum and also is robust to changes of traffic distributions. Finally, our work was extended to the case where the system cannot be stabilized due to excessive traffic loads. Under such heavy traffic regimes, we proposed optimal admission control policies considering tradeoffs between two QoS metrics: average delay versus maintaining a minimum level of connectivity to users independent of their location.

Appendix

We will prove the convergence of our iterative algorithm irrespective of the feasibility of the initial load vector using an *affine-invariant* property of the T mapping given as follows.

Definition 3 (Affine Set of ρ): For $\rho \in [0, 1-\epsilon]^b$, we define an affine set $\mathcal{A}(\rho) = \{\tilde{\rho} \in [0, 1-\epsilon]^b | \tilde{\rho} = \theta\rho + (1-\theta)e, \theta \geq 0\}$ where $e = (1, \ldots, 1)$. Hence, $\mathcal{A}$ is a set of points on the line connecting ρ and e; see Fig. 13.

Lemma 5 (Affine Invariance of T): For $\rho \in [0, 1-\epsilon]^b$ and $\tilde{\rho} \in \mathcal{A}(\rho)$, we have $T(\rho) = T(\tilde{\rho})$, which implies that all the points in the affine set yield the same partition by T. In fact, T is not a one-to-one mapping, but many-to-one, and thus noninvertible.

Proof: From (11), we see that scaling of $(1-\rho)$ does not change the decision rule because the decision metrics for all BSs are scaled in the same way. Hence, $T(\rho) = T(\tilde{\rho})$. ■

Lemma 6: For $\rho \in [0, 1-\epsilon]^b \cap \mathcal{F}^c$, there exist $\tilde{\rho} \in \mathcal{A}(\rho)$ and $\beta, \tilde{\beta} \in [0, 1)$ such that $\phi_\alpha(S(\tilde{\rho})) < \phi_\alpha(\tilde{\rho})$, where $S(\tilde{\rho}) = \tilde{\beta}\tilde{\rho} + (1-\tilde{\beta})T(\tilde{\rho})$; see Fig. 13.

Proof: We consider a mirror image of ρ, denoted by $\tilde{\rho}$, such that $\tilde{\rho} \in \mathcal{F} \cap \mathcal{A}(\rho)$. Then, $T(\rho) = T(\tilde{\rho})$ by Lemma 5. Note that $\tilde{\rho}$ is not unique. Let $\tilde{L}$ denote a line connecting $\tilde{\rho}$ and $T(\tilde{\rho})$. Similarly, let L denote a line connecting ρ and $T(\rho)$. We pick up $S(\rho)$ with some β on line L and determine $S(\tilde{\rho})$ as an intersection of $\tilde{L}$ and $\mathcal{A}(S(\rho))$. From Lemma 4, if β is sufficiently close to 1, which in turn implies $\tilde{\beta}$ is also sufficiently close to 1, we have $\phi_\alpha(S(\tilde{\rho})) < \phi_\alpha(\tilde{\rho})$. ■

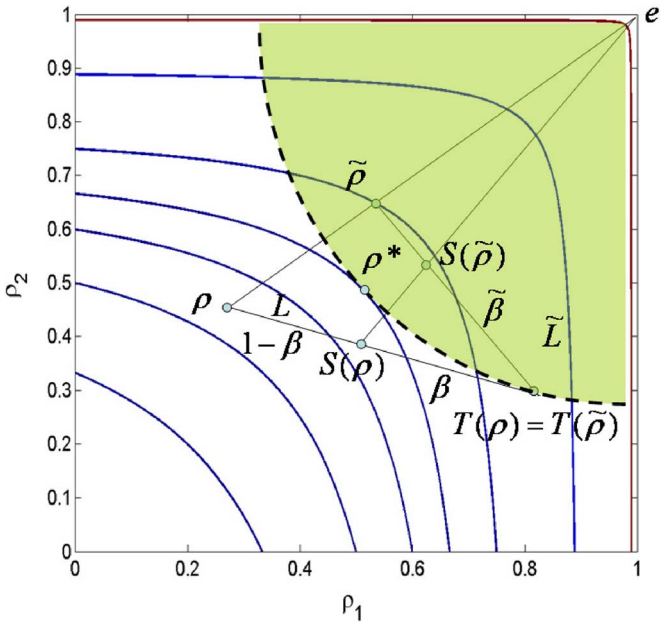


Fig. 13. Convergence property starting from arbitrary ρ.

Theorem 5: If Problem 1 is feasible, $S(\rho^{(k)})$ converges to ρ^* for any $\rho^{(0)} \in [0, 1-\epsilon]^b$.

Proof: If β is sufficiently close to 1, there exists a mirror sequence $\tilde{\rho}^{(k)}$ that converges to ρ^* by Lemma 6 and Theorem 2. Since $T(\rho^{(k)}) = T(\tilde{\rho}^{(k)})$, $T(\rho^{(k)})$ also converge to ρ^*. Since T is a continuous mapping, $\rho^{(k)}$ also converges to ρ^*. ■

References

[1] *IEEE Std 802.16m Part 16: Air Interface for Broadcast Wireless Access Systems: Advanced Air Interface*, IEEE Standard 802.16m, 2010.

[2] 3GPP Long Term Evolution, "LTE release 10 & beyond (LTE-Advanced)," 2011.

[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 *Proc. ACM MobiCom*, Sep. 2004, pp. 302–314.

[5] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," presented at the Eur. Wireless Conf., 2005.

[6] A. Zemlianov and G. de Veciana, "Load balancing in wireless systems supporting end nodes with dual mode capabilities," presented at the 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 ICC*, 2006, pp. 4339–4344.

[8] S. Liu and J. Virtamo, "Inter-cell coordination with inhomogeneous traffic distribution," in *Proc. Next Generation Internet Design Eng. Conf.*, 2006, pp. 64–71.

[9] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, 2006, pp. 1–12.

[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. 5, pp. 3566–3576, Jul. 2009.

[11] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic aware system-level coordination of wireless systems," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 721–734, Jun. 2011.

[12] B. Rengarajan and G. de Veciana, "Practical adaptive user association policies for wireless systems with dynamic interference," *IEEE/ACM Trans. Netw.*, Mar. 2010, submitted for publication.

[13] P. Hande, S. Patil, and H. Myung, "Distributed load-balancing in a multi-carrier system," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2009, pp. 1–6.

[14] H. Kim, X. Yang, M. Venkatachalam, Y.-S. Chen, K. Chou, I.-K. Fu, and P. Cheng, "Handover and load balancing rules for 16 m," IEEE C802.16 m-09/0136r1, Jan. 2009, pp. 1024–1037.
[15] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM*, Apr. 2003, vol. 1, pp. 321–331.
[16] G. Bianchi and I. Tinnirello, "Improving load balancing mechanisms in wireless packet networks," in *Proc. IEEE ICC*, 2002, pp. 891–895.
[17] H. Velayos, V. Aleo, and G. Karlsson, "Load balancing in overlapping wireless LAN cells," in *Proc. IEEE ICC*, Jun. 2004, pp. 3833–3836.
[18] S. Borst, N. Hegde, and A. Proutiere, "Interacting queues with server selection and coordinated scheduling—Application to cellular data networks," *Ann. Oper. Res.*, vol. 170, pp. 59–78, 2009.
[19] B. Rengarajan and G. de Veciana, "Practical distributed user association policies for wireless systems with dynamic interference," *IEEE Trans. Netw.*, Mar. 2010, submitted for publication.
[20] N. Benameur, S. Ben Fredj, F. Delcoigne, S. Oueslait-Boulahia, and J. W. Roberts, "Integrated admission control for streaming and elastic traffic," in *Proc. QofIS*, 2001, pp. 69–81.
[21] N. Benameur, S. Ben Fredj, S. Oueslait-Boulahia, and J. W. Roberts, "Quality of service and flow level admission control in the internet," *Comput. Netw.*, vol. 40, pp. 57–71, 2002.
[22] T. Bonald and A. Proutiere, "Wireless downlink channels: User performance and cell dimensioning," in *Proc. ACM MobiCom*, 2003, pp. 339–352.
[23] K. Etemad and M.-Y. Lai, *WiMAX Technology and Network Evolution*. Piscataway, NJ: IEEE Press, 2000.
[24] Qualcomm, "Introduction of enhanced icic, draft1_r2 106897," 3GPP TSG-RAN WG2 Meeting 72, 2010.
[25] T. Bonald, S. Borst, N. Hegde, and A. Proutier, "Wireless data performance in multi-cell scenarios," in *Proc. ACM SIGMETRICS*, 2004, pp. 378–380.
[26] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.
[27] S. B. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," in *Proc. ACM SIGCOMM*, 2001, pp. 111–122.
[28] T. Bonald and J. W. Roberts, "Congestion at flow level and the impact of user behavior," *Comput. Netw.*, vol. 42, pp. 521–536, 2003.
[29] H. Kim and G. de Veciana, "Losing opportunism: Evaluating service integration in an opportunistic wireless system," in *Proc. IEEE INFOCOM*, 2007, pp. 982–990.
[30] J. Walrand, *An Introduction to Queueing Networks*. Upper Saddle River, NJ: Prentice-Hall, 1998.
[31] A. L. Stolyar, "Optimal routing in output-queued flexible server systems," *Probab. Eng. Inf. Sci.*, vol. 19, no. 2, pp. 141–189, 2005.
[32] WiMAX2, "IEEE 802.16m Evaluation Methodology Document (EMD)," 2009.
[33] LTE-Advanced, "3rd generation partnership project; Technical specification group radio access network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)," 2010.
[34] H. Kim and G. de Veciana, "Leveraging dynamic space capacity in wireless systems to conserve mobile terminals' energy," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 802–815, Jun. 2010.

Hongseok Kim (S'06–M'10) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1998 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2009.

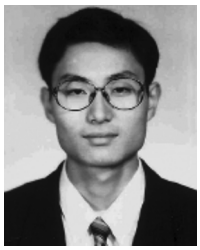
From 2000 to 2005, he was a Member of Technical Staff with Korea Telecom Network Laboratory, Daejeon, Korea. He participated in FSAN and ITU-T SG16 FS-VDSL standardization from 2000 to 2003. He was a Post-Doctoral Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ, from 2009 to 2010. He was a Member of Technical Staff with Bell Laboratories, Alcatel-Lucent, Murray Hill, NJ. He is currently with the Department of Electrical Engineering, Sogang University, Seoul, Korea. His research interests are network resource allocation and optimization including cross-layer design of wireless communication systems, MIMO and OFDMA, network economics, and Smart Grid.

Dr. Kim is the recipient of the Korea Government Overseas Scholarship in 2005–2008.

Gustavo de Veciana (S'88–M'94–SM'01–F'09) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1987, 1990, and 1993, respectively.

He is currently a Professor with the Department of Electrical and Computer Engineering at the University of Texas at Austin. He served as the Associate Director and then Director of the university's Wireless Networking and Communications Group (WNCG) from 2004 to 2008. His research focuses on the design, analysis, and control of telecommunication networks. Current interests include measurement, modeling and performance evaluation, wireless and sensor networks, and architectures and algorithms to design reliable computing and network systems.

Dr. de Veciana has served as an Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING and as Co-Chair of ACM CoNEXT 2008. He is the recipient of a General Motors Foundation Centennial Fellowship in Electrical Engineering and a National Science Foundation (NSF) CAREER Award in 1996. He is the co-recipient of the IEEE William McCalla Best ICCAD Paper Award 2000 and the Best Paper in the ACM *Transactions on Design Automation of Electronic Systems* from 2002 to 2004.



Xiangying Yang (S'01–M'05) received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1998, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin in 2000 and 2005, respectively.

He is now a Research Scientist/System Engineer with the Mobility Group Wireless Standard and Technology Organization, Intel Corporation, Hillsboro, OR. He is one of the major contributors to the IEEE 802.16m Work Group on the topics of QoS, mobility management, security, and enhanced features such as multi-carrier/relay support. His research interests include wireless broadband radio access networks, end-to-end performance and mobility management, wireless security protocols, Web caching, and peer-to-peer applications.

Muthaiah Venkatachalam (A'10) received the B.Tech. degree (with honors) in electronics and information engineering from the Indian Institute of Technology, Kharagpur, India, in 1999, and the M.S. degree in telecommunications and information systems engineering from the University of Texas at Austin in 2000.

He is the Director of System Architecture with the Mobile Wireless Group, Intel, Hillsboro, OR. He is responsible for technology development and industry standardization of 4G-based MAC, network and service aspects, as well as interworking between technologies. He has contributed significantly to IEEE 802.16 and WiMAX, which has helped shape the 4G wireless technology landscape in the industry. He has led the Intel team to complete the MAC-layer definition, design, and specification for the next-generation mobile WiMAX. He has chaired the technology development in the industry for various aspects such as femto cells and self-optimizing networks, location-based services, device power management, etc. Previously at Intel, he was the Lead Software/System Architect for the IXP2300 network processor and has driven Intel's efforts on developing network processor-based IP and ATM traffic management solutions; processing architectures for Intel's IXP23xx network processor family; and system architectures for broadband access, wireless access, and metropolitan optical networking systems. As the network processor architect at Intel, he has helped secure several design wins for Intel IXPs. He has several publications with 15 issued patents and numerous patents pending. He has served as an editor for *Computer Networks*.

Mr. Venkatachalam has also served as an organizing committee member for the 5th–8th Workshops on Media and Streaming Processors.