Generalized Proportional Fair Scheduling in Third Generation Wireless Data Networks

Tian Bu Li (Erran) Li Ramachandran Ramjee
Bell Labs, Lucent
tbu,erranlli,ramjee@bell-labs.com

Abstract—**In 3G data networks, network operators would like to balance system throughput while serving users in a fair manner. This is achieved using the notion of proportional fairness. However, so far, proportional fairness has been applied at each base station *independently*. Such an approach can result in non-Pareto optimal bandwidth allocation when considering the network as a whole. Therefore, it is important to consider proportional fairness in a network-wide context with user associations to base stations governed by optimizing a generalized proportional fairness objective. In this paper, we take the first step in formulating and studying this problem rigorously. We show that the general problem is NP-hard and it is also hard to obtain a close-to-optimal solution. We then consider a special case where multi-user diversity only depends on the number of users scheduled together. We propose efficient offline optimal algorithms and heuristic-based greedy online algorithms to solve this problem. Using detailed simulation based on the base station layout of a large service provide in the U.S., we show that our simple online algorithm, which assigns a newly arrived user to a base station that improves the generalized proportional fairness objective the most without changing existing users' association, is very close to the offline optimal solution. The greedy algorithm can achieve significantly better throughput and fairness in heterogeneous user distributions, when compared to an approach that assigns a user to the base station with the best signal strength.**

I. INTRODUCTION

Third Generation (3G) wide-area wireless networks based on the CDMA2000 [1] and UMTS [2] standards are now being increasingly deployed throughout the world. As of December 2004, there were over 146 million CDMA2000 subscribers and over 16 millions UMTS subscribers worldwide [3]. Emerging 3G data standards, EV-DO and HSDPA, promise to deliver broadband mobile internet services with peak rates of 2.4 Mbps and 14.4 Mbps, respectively.

In these third generation wireless data networks, Proportional Fair (PF) scheduling [4] is employed at the base station to schedule downlink flows among different users. Proportional fair is a channel-state based scheduling algorithm that relies on the concept of exploiting user diversity. Consider a model where there are N active users sharing a wireless channel with the channel condition seen by each user varying independently. Better channel conditions translate into higher data rate and vice versa. Each user continuously sends its measured channel condition back to the centralized PF scheduler which resides at the base station. If the channel measurement feedback delay is relatively small compared to the channel rate variation, the scheduler has a good enough estimate of all the users' channel condition when it schedules a packet to be transmitted to the user. Since channel condition varies independently among different users, PF exploits user diversity by selecting the user with the best condition to transmit during different time slots. This approach can increase system throughput substantially compared to a round-robin scheduler. However, such a rate maximizing scheme can be very unfair and users with relatively bad channel conditions can be starved. Hence, the mechanism used in PF is to weight the current rate achievable by a user by the average rate received by a user.

Specifically, at each time slot (every 1.67ms in EV-DO), the decision of the PF scheduler is to schedule the user with the largest $max_i \frac{R_i}{A_i}$, where R_i is the rate achievable by user i and A_i is the average rate of user i. The average rate is computed over a time window as a moving average:

$$\begin{array}{lcll} A_i(t+1) & = & (1-\alpha)A_i(t) + \alpha R_i & \text{if scheduled} \\ A_i(t+1) & = & (1-\alpha)A_i(t) & \text{if not scheduled} \end{array}$$

While PF scheduling achieves high throughput while maintaining proportional fairness among all users at a base station, the association of mobile devices to base stations in todays networks is not based on any fairness considerations. Instead, the mobile device is simply associated to that base station from which it receives the strongest signal. Clearly, such an association algorithm could create load imbalances where some "hotspot" base stations are heavily loaded while neighboring base stations are lightly

loaded. Further, such imbalances will decrease both overall throughput and also decrease fairness among users at neighboring base stations. In fact, such an approach can result in non-Pareto optimal bandwidth allocation when considering the network as a whole.

In this paper, we start with the premise that a fair scheduling algorithm should take a network-wide view and support fairness among all users connected to a network of base stations. To this end, we formulate the generalized proportional fairness (GPF) problem that takes a more macro-view of fairness and includes asssignment of users to base stations as part of an overall strategy to provide fairness across all users attached to the network of base stations. We show that the general problem is NP-hard and it is also hard to obtain a close-to-optimal solution. We then consider a special case, which roughly holds in practice, where multi-user diversity only depends on the number of users scheduled together and all users have equal priority. We propose efficient offline optimal algorithms and heuristic-based greedy online algorithms to solve this problem. Using detailed simulations based on the base station layout of a large service provide in the U.S., we show that our simple online algorithm, which assigns a newly arrived user to a base station that improves the generalized proportional fairness objective the most without changing existing users' association, is very close to the offline optimal solution. The greedy algorithm can achieve significantly better throughput and fairness, in many cases where users are not evenly distributed, when compared to an approach that assigns a user to the base station with the best signal strength. Our results also indicate, as expected, that max-min fairness sacrifices too much overall throughput (less than 60% of what is achieved by our algorithm) for fairness from each user's perspective.

The rest of the paper is structured as follows. In Section II, we present our motivation for generalized proportional fairness and describe our system model. In Section III, we present a rigorous formulation for the generalized proportional fairness problem and show that the general problem is NP-hard and it is also hard to obtain a close-to-optimal solution. In Section IV, we present the details of our offline optimal and online greedy algorithms. In Section VI, we present a detailed evaluation of our algorithms using simulation. In Section VII, we present related work. Finally, in Section VIII, we present our conclusions with a discussion of future work.

II. Motivation and System Model

In this section, we first motivate the need for generalized proportional fairness and then present our system model.

A. Motivation

Consider a single base station in a 3G data network serving its associated users. Assume all users have the same priority. A network operator would like meet users' demands to the greatest extent possible, given resource constraints. However, this does not imply maximizing the total throughput of all users, as such a policy may lead to starvation of users who have poor channel conditions. Therefore, we need to consider fairness constraints in allocating shared resources to multiple users. Two fairness measures are commonly used: max-min fairness and proportional fairness. Informally, an allocation of bandwidth by a BS is max-min fair if there is no way to give more bandwidth to any user without decreasing the allocation of a user with less or equal bandwidth. Max-min fairness achieves ideal fairness from the users' perspective and the system is work-preserving given the constraints on resource availability to users. However, max-min fairness can significantly sacrifice aggregate throughput. For example, suppose there are two users u and v associated with a BS a. Let the average data rate of u and v, denoted by r_{ua} and r_{va}, be 10 and 1 unit respectively. Max-min fairness will allocate both users a bandwidth of $10/11$ (lets ignore multi-user diversity for the moment). User v with a much lower rate will be allocated 10 times ($10/11$ fraction) more slots than u. This clearly sacrifices aggregate system throughput significantly. In order to achieve a better tradeoff between fairness and throughput, Kelly [5] proposed proportional fairness. In our example, proportional fairness is equivalent to time fairness. User u and v will get a bandwidth of 5 and $1/2$ respectively. Proportional fairness achieves much better aggregate throughput than max-min and gives users equal time fairness. For wireless networks with scarce bandwidth, proportional fairness is much more appealing to network operators and thus have been implemented in 3G data networks such as 3G1x-EVDO.

In CDMA-based 3G networks, a mobile device can hear signals from multiple BSs. The user is typically associated to the base station with the strongest signal. Proportional fairness scheduling is then used *independently* at each base station to schedule downlink transmissions to devices associated with each base station. Using this approach, uneven user distribution will result in uneven load distributions at the BSs. In order to handle this load imbalance, techniques that control base station coverage such as

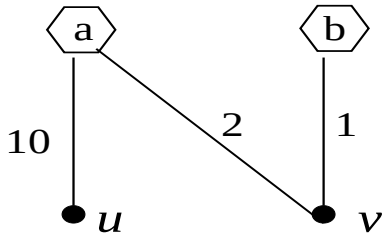


Fig. 1
A WIRELESS NETWORK WITH 2 BSS, a AND b, AND TWO USERS, u AND v.

cell breathing or user association have been proposed [6], [7]. Although Sang et al. [6] consider load balancing and fairness in an integrated framework. The load balancing metric does not consider the fairness objective. In fact, the load balancing techniques may reduce network-wide fairness objective since highly loaded cells will reduce the amount of bandwidth allocated to users at the cell boundary. Their rationale is that this may trigger users to handoff to other less loaded cells. However, without the explicit consideration of fairness, it is unclear whether they will maintain a unified fairness measure. Thus, *we are the first to consider load balancing with the network-wide proportional fairness as the goal*. We argue that proportional fairness should be used when allocating *all shared system resources* (all BSs) to users rather than a single system resource (a single BS). This enables optimal tradeoff between system utilization as a whole and serving users fairly. We achieve this by controlling user associations to base stations.

We now illustrate the benefit of considering user association and proportional fairness jointly using an example 1. In this example, $r_{ua} = 10$, $r_{ub} = 0$, $r_{va} = 2$, $r_{vb} = 1$. If user association is based on strongest signal, then both u and v will associate with BS a. If the system is using proportional fairness scheduling at BS a, then u will get 5 and v will get 1. However, if we consider user association and proportional fairness jointly, users u and v will be associated with BS a and b respectively. This gives u a bandwidth of 10 and v a bandwidth of 1. This certainly is a better allocation than the previous one which considers user association and proportional fairness separately. The allocation of $5, 1$ is not even *Pareto optimal* when compared with the allocation of $10, 1$. A Pareto optimal allocation is one such that, there does not exist another feasible allocation where at least one user gets more bandwidth, and all others get at least the same bandwidth. Now, if we consider max-min fairness allocation in conjunction with user association, this will result in both u and v being associated with BS a, and they will both get a bandwidth of $10/6$.

B. System Model

The latest 3G standard, called the 3G1X Evolution or High Data Rate (HDR), is designed for bursty packet data applications. It provides a peak downlink data rate of 2.4Mbps and an average downlink data rate of 600Kbps within one 1.25MHz CDMA carrier. HDR is commercially available. HDR downlink has much higher peak data rate (2.4Mbps), compared with the uplink peak data rate of 153.6Kbps [8]. Users share the HDR downlink using time multiplexing with time slots of 1.67ms each. At any time instant, data frames are transmitted to one specific user, and the data rate is determined by the user's channel condition. Users monitor the pilot bursts in the downlink channel to estimate the channel conditions in terms of Signal to Noise Ratio(SNR). This SNR is then mapped into a supported data rate, and fed back every time slot to the base station through the data-rate-request channel in the reverse link. The duration of transmission to each user is determined by the downlink scheduling algorithm. HDR uses a scheduling algorithm called Proportional-Fair Scheduling [8]. The scheduler serves the user with the highest ratio of instantaneous downlink channel rate over the average received data rate.

In this paper, we focus on the downlink scheduling of a network of 3G wireless data base stations. When a device powers up or becomes active, we assume the device reports the set of BSs it can hear and the data rate to each of the BSs. The network, possibly the Radio Network Controller (RNC), then determines a user's (device's) association based on this reported information as well as preexisting information of currently associated users at each BS. This can be achieved in a reasonable time since fast fades last 5-50 slots depending on the speed of a mobile, and a device can feed up its rate information each time slot (1.67ms). So it is reasonable to obtain a good estimate of the average rate of a user with a BS in a short period of time (200ms). Alternatively, the network could also first associate a device to the base station with the strongest signal, then instruct the mobile device to report the average rates to all the base stations the device can hear, and then switch the mobile device to the optimal base station.

We determine user association upon arrival or handoff due to mobility. Other work such as [6], [7] have assumed much more frequent user association changes. This would require much more coordination among BSs as to which

packet a BS should transmit at a certain time slot. This would also increase the signaling load to the RNC if RNC determines the association change. Both the work of Sang et al. [6] and Das et al. [7] have assumed that a central controller can determine the scheduling policy of each base station. This would require fine-grained feedbacks from each BS to the central controller. In this paper, we assume that the BS runs an independent proportional fairness scheduler and the network (RNC) only determines the user association to a base station. Sang et al. [6] has also assumed that each mobile determines which BS it wants to be associated with according to its own utility function (which is determined by the throughput it can get from each BS). We remark that, independent user's action may lead to handoff oscillations. To avoid such problems, we have assumed the network (RNC) determines when a user should change its association.

III. Problem Formulation

In this section, we present a rigorous formulation of the network-wide proportional fairness bandwidth allocation problem. We consider a 3G wireless data network consisting of a set of BSs A, and a set of users U. We assume users are static for now and allow for mobile users in later sections. A user u's average data rate if associated with a BS a is denoted as r_{ua}. Let $S_u = \{a | r_{ua} > 0,\ \forall a \in A\}$. Let γ_u be the actual bandwidth allocation to user u by the network. It has been shown in [5] that proportional fairness allocation of network resources is equivalent to the optimization of the following objective function:

$$\max \sum_{u \in U} log(\gamma_u)$$

Typically 3G data users have elastic traffic. The application traffic in most cases uses TCP as the transport protocol. TCP will try to achieve the maximum rate allowed by the system. Therefore, we assume a user will consume all the bandwidth allocated and the queues are backlogged. In 3G data networks, typically there is a limit on how many users can be admitted to the system. For example, in 3G1x-EVDO, there are 60 Walsh codes for orthogonal transmission. This puts an upper bound of 60 active users (per base station per sector-carrier) at any given time. For ease of explanation, we will assume that any user u will be admitted as long as $|S_u| > 0$. We will discuss how to extend our framework to the admission control case later. Note that the limit of 60 users is rarely reached in practice since 60 users sharing an average 600Kbps downlink channel provides very little throughput for each user.

Lets denote x_{ua} the association variable, i.e. $x_{ua} = 1$ if user u is associated with BS a, 0 otherwise. We assume a user can only be associated with one BS at any given time, i.e., only one BS can transmit data through its downlink to the user as in the EV-DO standard [4]. As noted earlier, we pnly consider downlink bandwidth allocation in this paper. Since all users must be admitted, we have the following:

$$\sum_{a \in S_u} x_{ua} = 1$$

Let the bandwidth allocation for users associated with a given BS be proportional fair. Let the number of users associated with BS a be $y_a = \sum_{u \in U} x_{ua}$. Lets denote the set of users associated with a given BS a be Q_a, i.e. $Q_a = \{u | x_{ua} = 1,\ \forall u \in U\}$. In the general case, the multi-user diversity gain can depend on the set of users, not just the number of users. Let $\mathcal{R}$ be the mapping that, given u and a, returns r_{ua}. Users may have different scheduling priorities for service differentiation. We denote the scheduling priority of user u as Ω_u. Let ω be the mapping that given a user u returns ω_u. If $x_{ua} = 1$ for a given user u and BS a, its actual bandwidth γ_u allocation by BS a will be a general function of all the users associated with a. Denote this function by $f_a(Q_a, u, \mathcal{R}, \Omega)$. Of cause, $f_a(Q_a, u, \mathcal{R}, \Omega) = 0$ if $u \notin Q_a$. We now derive the following problem formulation for Generalized Proportional Fairness, GPF1:

$$\max \sum_{a \in A} \sum_{u \in Q_a} log(f_a(Q_a, u, \mathcal{R}, \Omega)) \quad (1)$$

Subject to

$$\sum_{a \in S_u} x_{ua} = 1,\ \forall u \in U \quad (2)$$

$$Q_a = \{u | x_{ua} = 1,\ \forall u \in U\} \quad (3)$$

$$x_{ua} = \{0, 1\} \quad (4)$$

However, as shown in [9], if the relative rate fluctuations are statistically identical, the multi-user diversity gain only depends on the number of users associated with a given BS. This assumption on rate fluctuation is roughly valid when the users have, for example, Rayleigh fading channels and the feasible rate is approximately linear in the SNR (reasonably accurate when the SNR is not too high). With this assumption, according to [9], $\gamma_{ua} = G(y_a) r_{ua} / y_a$ if u is associated with BS a. Thus, we obtain the following restricted version of the GPF1 prob-

lem. We refer to this problem as GPF2.

$$\max \sum_{u\in U}\sum_{a\in S_u} x_{ua} log(r_{ua}\omega_u \frac{G(y_a)}{\Omega_a}) \quad (5)$$

Subject to

$$\sum_{a\in S_u} x_{ua} = 1,\ \forall u \in U \quad (6)$$

$$\Omega_a = \sum_{u:a\in S_u} x_{ua}\omega_u,\ \forall a \in A \quad (7)$$

$$y_a = \sum_{u:a\in S_u} x_{ua},\ \forall a \in A \quad (8)$$

$$x_{ua} = \{0,1\} \quad (9)$$

When all the users have the same priority in GPF2, we have the following special case, GPF3:

$$\max \sum_{u\in U}\sum_{a\in S_u} x_{ua} log(r_{ua} \frac{G(y_a)}{y_a}) \quad (10)$$

Subject to

$$\sum_{a\in S_u} x_{ua} = 1,\ \forall u \in U \quad (11)$$

$$y_a = \sum_{u:a\in S_u} x_{ua},\ \forall a \in A \quad (12)$$

$$x_{ua} = \{0,1\} \quad (13)$$

A. GPF1 and GPF2 are NP-hard and inapproximable

We show that, for GPF1 and GPF2, there does not exist an algorithm that can find the optimal in polynomial time unless $P = NP$, i.e. the problem is NP-hard. Our reduction is via 3-dimensional matching which is known to be NP-complete. The 3-dimensional matching problem is stated as follows.

Definition 1: Given an instance of the following problem: Disjoint sets $B = \{b_1, \ldots, b_n\}$, $C = \{c_1, \ldots, c_n\}$, $D = \{d_1, \ldots, d_n\}$, and a family $F = \{T_1, \ldots, T_m\}$ of triples with $|T_i \cap B| = |T_i \cap C| = |T_i \cap D|$ for $i = 1, \ldots, m$. The question is: does F contain a matching, i.e. a subfamily F' for which $|F| = n$ and $\cup_{T_i \in F'} T_i = B \cup C \cup D$?

Theorem 1: The generalized proportional fairness problem GPF1 is NP-hard.

Proof: Our reduction is along the lines of [10]. Clearly, the answer to the 3D matching decision problem is no if $m < n$. It is easy to check whether F is a matching or not if $m = n$. We simply check whether $\cup_{T_i \in F} T_i = B \cup C \cup D$. So we assume $m > n$ in our reduction.

We call the triples that contain b_j *triples of type* j. Let t_j be the number of triples of type j for $j = 1, \ldots, n$. BS i corresponds to the triple T_i for $i = 1, \ldots, m$. We have $2n$ element users, corresponding to the $2n$ elements of $C \cup D$. There are $t_j - 1$ dummy users of type j for $j = 1, \ldots, n$. Note that, the total number of jobs is $m - n$. For BS i corresponding to a triple of type j, say $T_j = (b_j, c_k, d_l)$, the corresponding element users c_k and d_l have an average rate of R. Their multi-user diversity gain when scheduled together is G where $1 < G(2) < 2$, i.e. each obtain a rate of $GR/2$. Each of the dummy jobs of type j has an average rate of R. However, the multi-user diversity gain $G'(2)$ when an element user and a dummy user of type j is assigned to one of the t_j BSs, is much smaller than G. All other average rates between BSs and users are 0. Except element users corresponding to a T_j, all other users, dummy or element, when associated with a BS, the multi-user diversity gain is given by function $G'()$. Assume $\frac{G(i)}{i} < \frac{G(j)}{j}$ if $j < i$. Similarly the condition holds for G'. We claim that a matching exists iff the GPF1's optimal objective function $F^* \geq F$ where $F = 2nlog(GR/2) + (m-n)logR$.

Suppose there is a matching. For each $T_i = (b_j, c_k, d_l)$ in the matching, associate element user c_k and d_l with BS j. For each j, this leaves $t_j - 1$ idle BSs corresponds to triples of type j not in the matching; associate the $t_j - 1$ dummy users of type j with these $t_j - 1$ BSs. This assignment has an objective function value of $F = 2nlog(GR/2) + (m-n)logR$. Conversely, suppose that there is such an assignment with objective function $F^* \geq F$. Dummy users can only be assigned to the BS of its type. So the $t_j - 1$ dummy users will be assigned to the t_j BSs of type j. If possible, BSs should not be idle. Suppose a user c_k has been assigned to a BS with $L > 1$ users and there is an idle BS user c_k can be assigned, then we can obtain a better solution after moving c_k to the idle BS. This is because $\frac{G(i)}{i} < \frac{G(j)}{j}$ if $j < i$ (for G' as well). Therefore, the set of BSs of each type j gets at least one element user (otherwise, there will be an idle BS). We claim that, if possible assignment exists, no BS should get more than 3 users. This is because reducing the number of users on this BS and assign them to other BS will one user always improve the solution. Therefore, each BS gets assigned at most 2 users in the optimal solution if possible. For those BS with 2 users, it is better to assign the two element users corresponding to BS of type j, this is because of their multi-user diversity gain is better. In addition, no two BSs of the same type, each gets assigned two users. So the best possible solution is an assignment with an objective function value F. In any of these assignment, there exists exactly one BS for each type j, where two element

users corresponding to type j is assigned, and $t_j - 1$ BSs of type j, each get assigned a dummy user. There is a total of $2n$ element users which implies a 3-dimensional matching in the original problem. ■

Similarly we can prove the following theorem.

Theorem 2: The generalized proportional fairness problem GPF2 is NP-hard.

The key idea is to assign low priority to dummy users and high priority to element users; then set the average rate of dummy users with its t_j BSs very high, and set the rate of element users low. This will force element users to be assigned together. Otherwise, they would "steal" the rates of dummy users which results in a worse objective function value.

Corollary 1: There does not exist a polynomial time algorithm that can approximate GPF1 and GPF2 within a factor of $\rho, \forall \rho > 0$, i.e. they are inapproximable.

Proof: Theorem 1 holds even if we assign rate r instead of R for dummy users of type j to its t_j type j BSs. We can pick r such that $F^* = 0$. Suppose we have an approximation algorithm of factor ρ. According to the definition of approximation, the algorithm outputs a solution with objective function value F such that $F^* \leq \rho F$ which means $F \leq 0$. Since $F^* = 0$, this means $F = 0$, thus we have found the optimal which contradicts Theorem 1. Similarly, this holds for problem GPF2. ■

B. Properties of GPF3

We do not yet know whether GPF3 is NP-hard or not. The reduction for GPF1 and GPF2 can not be applied to GPF3 because we can not enforce two element users corresponding to a triple in 3-dimensional matching are assigned to the same BS (GPF3 can associate dummy users and element users together and still achieve optimal). Despite this fact, we do know interesting properties of GPF3 which enable us to design efficient algorithms. We now show these properties.

Proposition 1: If we know y_a in GPF3, then the problem can be solved optimally in polynomial time.

If y_a is fixed, then $p_{ua} = log(r_{ua}\frac{G(y_a)}{y_a})$ is fixed. The problem is then equivalent to finding a maximum weighted bipartite matching with p_{ua} as the weight. The maximum weighted matching problem can be solved optimally in polynomial time.

Proposition 2: If two users u, v are associated with BS a, b respectively in an assignment and $\frac{r_{ua}}{r_{ub}} < \frac{r_{va}}{r_{vb}}$, then we can always swap u and v's association and improve the objective function.

Algorithm **OfflineOPT-KBS**
Input: Network $H = (A, U)$, multi-user diversity gain $G()$, $r_{ua}, \forall a \in, u \in U$
for each $(y_1, \ldots, y_{|A|})$ such that $\sum_{i=1}^{|A|} = n$
  if $y_i > |Q_i|$ where $Q_i = \{u | r_{ui} > 0, \forall u \in U\}$, then next iteration
  if $r_{ui} = 1$ and $y_i > 0$
    $p_{ui} = log(r_{ui}\frac{G(y_i)}{y_i})$
  else $p_{ui} = 0$

  MatchingAlgo($H, \{p_{ui}\}$)
end

Fig. 2
A FORMAL DESCRIPTION OF THE OFFLINE OPTIMAL ALGORITHM

This is easily obtained by comparing the two objective function values before and after u and v are swapped with their associated BSs.

IV. ALGORITHMS

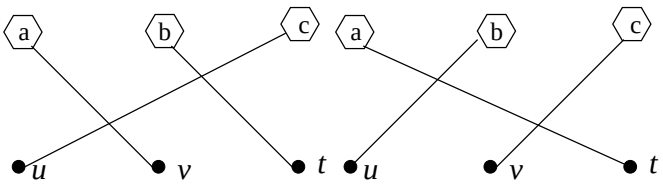
A. Offline Algorithms

We first design an offline algorithm for computing the optimal user assignments to the base stations. If the number of BSs is a constant K, based on Proposition 1, we can obtain a polynomial time algorithm to find the optimal association. The idea is to guess all possible y_a configurations, then solve the the maximum weighted matching problem for each configuration. The algorithm is polynomial because the number of configurations is n^K (let $n = |U|$) and maximum weighted matching has a running time of $O(\sqrt{n^3})$. The total running time is thus $O(n^{K+3/2})$. The OfflineOPT algorithm is presented in Figure 2.

However, when the number of BSs is large, OfflineOPT will be very inefficient. We note that, in our context, users are spatially distributed; users inside a certain region will not be able communicate with BSs further away from the region. Therefore, there is a natural partition. We exploit this spatial property and partition the network into smaller connected components where the number of possible edges (associations) between components is small. We first solve our generalized proportional fairness problem within each component. We then assign users, whose edges cross components, greedily. The formal description of the algorithm is in Figure 3. The algorithm is refered

Algorithm **KComponent**
Input: Network $H = (A, U)$, multi-user diversity gain $G()$, mapping $(\mathcal{R})$ s.t. $R(u, a) = r_{ua}, \forall a \in, u \in U$
Run MinK-Cut Algorithm to obtain $H_1, \ldots, H_K$ connected components
for each $H_i, \forall i = 1, \ldots, K$
  OfflineOPT-KBS(H_i,G,R)
for each user u whose edges cross components
  Greedily assign u to BS a that improves the objective function the most
end

Fig. 3
A FORMAL DESCRIPTION OF THE KCOMPONENT ALGORITHM



(a) Optimal assignment (b) LS assignment

Fig. 4
ALGORITHM LS MAY GET STUCK IN LOCAL OPTIMA

to as KComponent. We use the approximation algorithm in [11] to compute a K cut. We would like to minimize the number of edges across components. Formally, the minimal k-cut problem is defined as follows: a set of edges whose removal leaves k connected components is called a k-cut. The k-cut problem asks for a minimal weight k-cut. We can tune k to trade off optimality with computation time.

In a large network with many users, Algorithm offlineOPT, even when running KComponent, can result in a high computational overhead. In addition, KComponent (also OfflineOPT) only works for GPF3. We thus consider the design of a heuristic algorithm based on efficient local search. We define two operations: Swap and Change. At any given time of the algorithm, if swapping two users can improve the objective function, we then do the swap. If changing a user's association can help, then we carry out the Change operation. To make sure the algorithm runs in polynomial time, we place a lower bound constant δ such that each improvement operation, or L improvement operations, improves the objective function by at least δ. The algorithm is guaranteed to terminate in time $log(\frac{OPT}{\delta})$ iterations. Its running time is $O(log(\frac{OPT}{\delta})n^2)$. We refer to this algorithm as local search (LS).

Note that Algorithm LS can get stuck in local optima. The following is an example. We have three BS a, b, c and three users u, v, t. We have the following rate values, $r_{ua} = 123$,$r_{ub} = 492$, $r_{uc} = 893$, $r_{va} = 415$,$r_{vb} = 217$, $r_{vc} = 659$, $r_{ta} = 526$,$r_{tb} = 756$, $r_{vc} = 367$. It is easy to verify that the optimal assignment 4-a yields an objective function value of 19.5. Algorithm LS outputs the assignment in 4-b with a value of 19.0. We can verify that any single Swap or Change operation will not yield a better assignment. We can replicate this example so that LS search will fail even if we consider Swap operations with $|U| - 1$ users.

B. Online Algorithms

While offline algorithms are useful in computing optimal assignments to compare against, they cannot be used in a dynamic network-setting with mobile users. We therefore consider two greedy online algorithms with very little computational overhead. Our first algorithm assumes that once a user is assigned with a BS, it can not change to associate with a different BS unless due to handoff or connection failure. We greedily associate with the BS such that the objective function improves the most. We refer to this algorithm as Greedy-0.

Given that users' association to base stations can change often in a dynamic setting, we can also potentially change users' association if it improves the system objective function. In the second online algorithm, we assume we can change the association of at most k existing users. Because k is a small number, we can try out all possible cases and pick the one that improves the objective function the most. We refer to this algorithms as Greedy-k

V. EXTENSIONS

We discuss possible extensions to our problem. In practice, a BS can only admit a finite number of users due to the number of available Walsh codes for uplink communication. Let this bound be $C_a, \forall a \in A$. We then have the following constraint:

$$y_a \leq C_a$$

Since some users may be blocked, we need to remove the constraint $\sum_{a \in S_u} x_{ua} = 1$. However, this is not enough. In order to optimize the objective function using this formulation, a system may reject a user u even if there exists

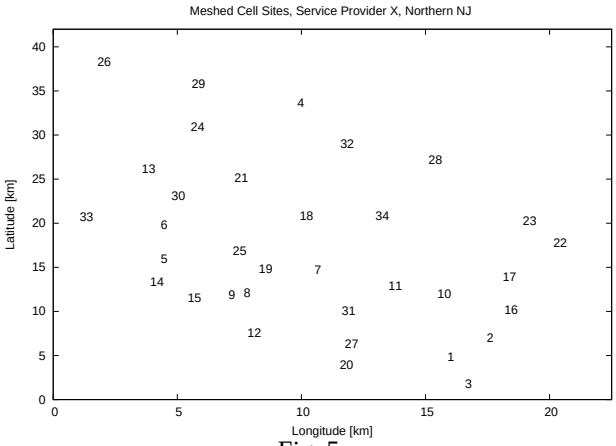


Fig. 5
GEOGRAPHICAL LOCATIONS OF BASE STATIONS

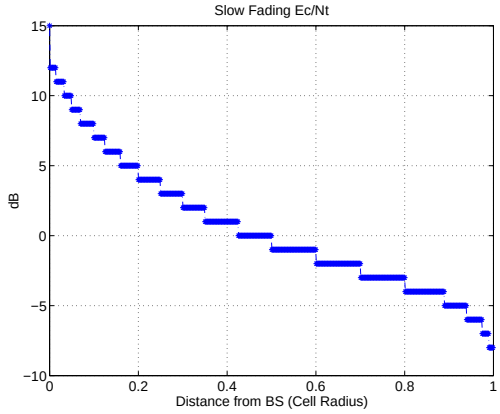


Fig. 6
SLOW FADING

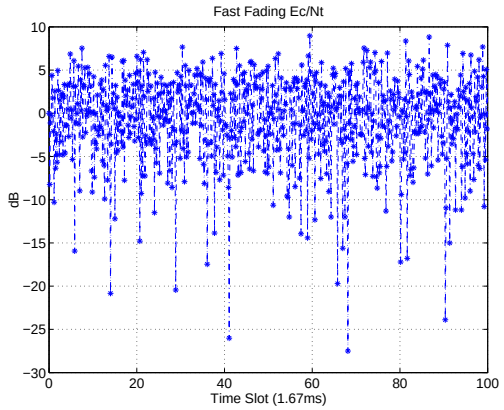


Fig. 7
FAST FADING

a BS a such that $u \in S_a$ and $y_a < C_a$. To make the system work conserving, we define the indicator variable z_a; $z_a = 0$ if $y_a = C_a$, 1 otherwise. We derive the following constraint in order for the system to be work conserving:

$$\sum_{a \in S_u} x_{ua} = 1, \text{if } y_a < C_a, u \in S_a$$

It is easy to see that the GPF1 and GPF2 problem remains to be NP-hard and inapproximable. The OfflineOPT-KBS can be extended to our current setting by simply observing the capacity constraint in the algorithm shown in Figure 2.

VI. EVALUATION

We have demonstrated the general proportional fairness problem is NP hard and presented some online and offline heuristic algorithms. In this section, we evaluate our algorithms using simulations. We first evaluate the quality of the LS algorithm by comparing to the optimal algorithm for a small problem size. We then investigate the advantages of associations based on our GPF3 problem formulation (hereafter, simply referred to as GPF) over Best-Signal and Max-min associations. Finally, we evaluate the performance of two online algorithms, Greedy-0 and Greedy-k. We remark that, except OfflineOPT and KComponent, all our other algorithms (LS, Greedy-0 and Greedy-k) work with all three problem formulations (GPF1,GPF2 and GPF3).

A. Simulation setup

The map of base station layout that we use for the performance evaluation of our algorithms is presented in Figure 5. It is part of a 3G network operated by one of the major service providers in United States. There are a total of 34 base stations within $50 \times 30 km^2$. The longitude and latitude distances are all relative to a reference point picked for good visualization.

In our simulation, we assume all base stations have a uniform radius of 8km and there are a total of 1000 3G HDR mobiles in the network. User requests for radio channel arrives according to a Poisson process with an average rate 0.001/second. We assume that the average radio channel holding time is exponentially distributed with mean of 300 seconds. We divide the map in Figure 5 into 50×30 zones where each zone is one square kilometer. We assume mobiles migrate from one zone to another with an exponentially distributed staying time of mean 60 seconds. We further assume mobiles have the same rates within the same zone and a re-association may be performed when an active mobile moves from one zone to another and its rates change.

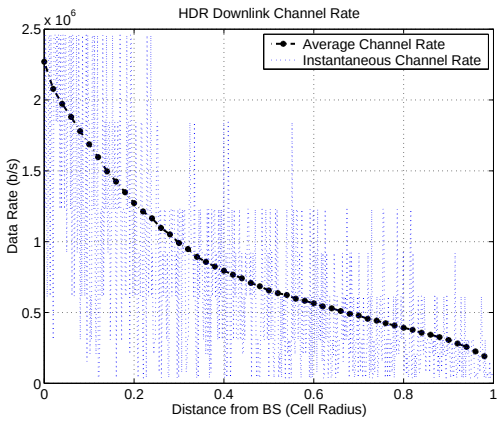


Fig. 8
HDR CHANNEL RATE

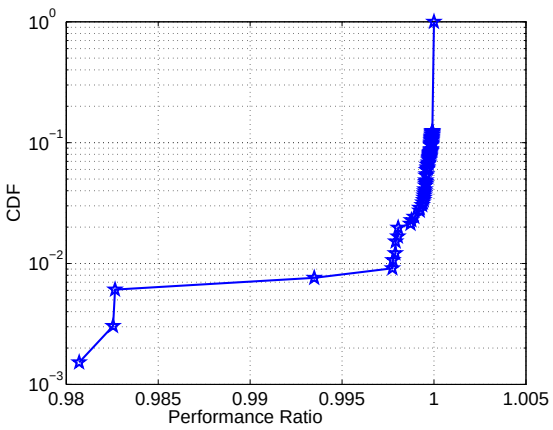


Fig. 9
CDF OF PERFORMANCE RATIOS

The HDR downlink channel is modeled according to the published experimental data in [8], [12]. The HDR downlink channel quality is determined by both slow fading and fast fading. Slow fading is modeled as a function of the client's distance from the HDR base station, as shown in Figure 6. Fast fading is modeled by Jakes' Rayleigh fading [13] as shown in Figure 7. The combined E_c/N_t for both slow and fast fading is then mapped to a table of supported data rate with 1% error [12]. Figure 8 presents a snapshot of HDR downlink instantaneous channel rates, and the average rate over a long time period for clients with different distances from the base station. In our simulations, we use the average rate for determining throughput and base station associations and do not simulate the fast fading or the zero mean shadow fading on a per-mobile basis.

In order to simulate skewed user distributions that are typical in a real 3G network setting, we assign different weights to different zones. When a mobile moves, it choose its next destination zone from one of the neighboring zones with probability proportional to the weights of that zone. In our simulation, we randomly generate a weight for each zone in $(0, 10]$ uniformly.

B. Comparison of local search algorithm with optimal

It has been shown that Algorithm OfflineOPT can find the optimal association in polynomial time. We evaluate the local search (LS) heuristic algorithm for a small network of 6 base stations where the optimal solution can be obtained in reasonable amount of time using OfflineOPT. We picked base stations $5, 8, 9, 14, 15$ and 25 in Figure 5 together with 200 users as input and run the simulation for 10000 seconds. A new association is calculated when a radio channel is requested or terminated. A radio channel request may be triggered by either a mobile user becoming active or an active mobile user moving into a new zone. A radio channel termination may be triggered by the inactive timer timeout or a mobile moving away from a zone. We compute each association with both the optimal algorithm and our LS heuristic algorithm. Figure 9 plots the CDF of the performance ratio, which is defined as the ratio between the value of generalized proportional fairness objective function obtained from the heuristic algorithm and that obtained from the optimal algorithm. We observe that only less than 1% of the performance ratios are less than 0.997. In almost 90% of the cases, the LS heuristic algorithm achieves the same objective function values as the optimal algorithm. Thus, *we can approximate the OfflineOPT algorithm with the much more efficient LS heuristic-based algorithm with high confidence when computing the optimal bounds on large networks.*

C. Comparison of GPF with other association algorithms

In this Section, we compare our GPF association with other association schemes including Best-Signal and Max-Min. We use the LS algorithm to obtain the "optimal" GPF association. The Best-Signal algorithms always assigns a mobile to the BS with the strongest signal regardless of network load. The Max-Min algorithm is the offline algorithm presented in [14]. We do not consider backhaul bottlenecks. Therefore, the max-min fair association computed has an approximation factor of 2 coordinate-wise in terms of the throughput vector (in ascending order) all users get with respect to the optimal max-min fair association.

We evaluate the algorithms using the achieved network aggregate throughput and fairness metrics. The achieved

throughput metric measures the throughput efficiency at the network level whereas the fairness metric focuses on the performance perceived by each individual user.

The fairness metric is measured using fairness index proposed in [15]. The throughput is first normalized by the reference association which we assume an equal bandwidth share. In this case, the fairness index can be computed as following

$$FI_{EQ} = \frac{(\sum t_i)^2}{n \sum t_i^2}$$

The fairness index value is a positive number with max value 1 suggesting an equal throughput among all mobiles. The higher the fairness index is, the closer the allocation is to an equal throughput allocation.

We compare the GPF allocation with the other two by normalizing the achieved throughput using the one obtained from the GPF allocation; the fairness index metric is normalized to users receiving equal share. Figures 10 and 11 plot the CDF of normalized throughput and fairness indexes respectively (thus, GPF has a value of 1 in Figure 10).

We observe from Figure 10 that both Best-Signal and GPF algorithm achieve much higher throughput than Max-Min algorithm. Almost in all cases, the Max-Min algorithm achieves throughput less than 60% of the throughput of GPF. The GPF approach achieves higher throughput in more than 80% cases than Best-Signal. In about 60% cases, the GPF approach achieve more than 10% more throughput than Best-Signal. Figure 11 shows that the bandwidth allocation of our GPF algorithm is always more fair than the allocation of Best-Signal algorithm while the allocation of the Max-Min algorithm is the most fair scheme. However, we observe in Figure 10 the fairness of Max-min is achieved by significantly compromising throughput. Thus, *the GPF association can achieve significantly better throughput and fairness when compared to an approach that assigns a user to the base station with the best signal strength. Also, as expected, max-min fairness sacrifices too much overall throughput (less than 60% of what is achieved by our algorithm) for fairness.*

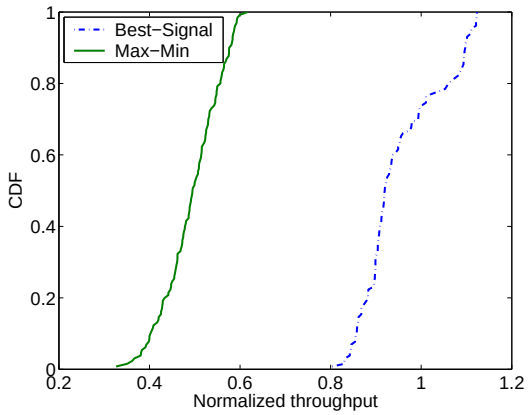


Fig. 10
CDF OF THROUGHPUT NORMALIZED BY THROUGHPUT ACHIEVED FROM GPF ASSOCIATION

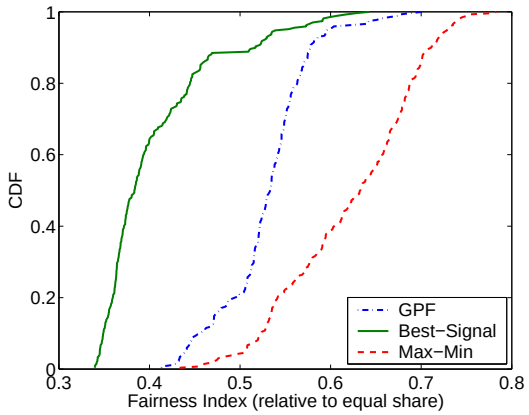


Fig. 11
CDF OF THE FAIRNESS INDEXES RELATIVE TO EQUAL SHARES

D. Comparison of online GPF association algorithms

We have demonstrated that an offline generalized proportional fairness association achieves both better fairness and higher throughput than the currently used Best-Signal approach. In this section, we compare Best-Signal to two online versions of the GPF association,i.e, greedy-0 and greedy-k; they differ in whether a constant number of the existing association can be changed. We would like to see if the gains of using the offline GPF association can be preserved in a practical dynamic environment using the online algorithms. As before, we evaluate these algorithms using both the throughput and fairness metrics.

We normalize the throughput of each mobiles by its GPF allocation, where we approximate the GPF allocation using values obtained from the LS algorithm. Let us denote the throughput achieved by GPF for mobile i being t_i^{GPF} and t_i^X the throughput achieved by association scheme X. We compare the fairness of Best-Signal and on-line GPF association relative to the offline GPF association. The fairness index relative to GPF association for X is calcu-

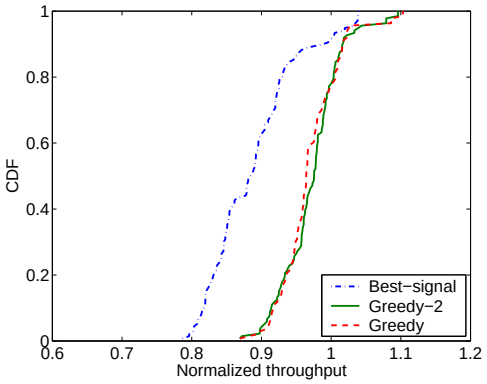


Fig. 12

CDF OF THROUGHPUT NORMALIZED BY THROUGHPUT ACHIEVED FROM GPF ASSOCIATION

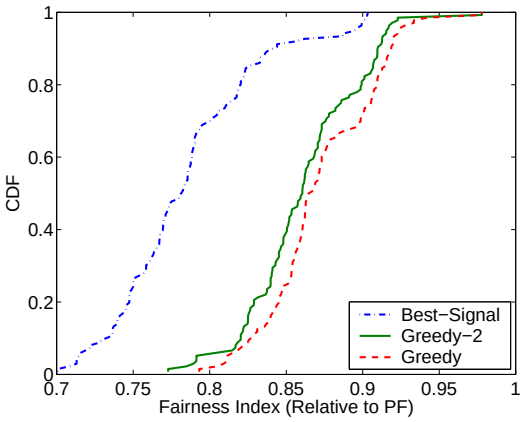


Fig. 13

CDF OF FAIRNESS INDEXES RELATIVE TO GPF ASSOCIATION

lated from

$$FI_{GPF}(X) = \frac{(\sum t_i^X / t_i^{GPF})^2}{n \sum (t_i^X / t_i^{GPF})^2}$$

where n is total number of active mobiles in the network.

We also normalize the throughput from different associations by the throughput achieved by algorithm LS and present the CDF of normalized throughput in Figure 12. We observe that both online algorithms achieve higher throughput than the Best-Signal. There is very little difference between the two online algorithm. *In more than* 70% *cases, the online algorithms achieves throughput greater than* 95% *of that achieved by the offline GPF*.

Figure 13 presents the CDF of the fairness indexes for different association algorithms. *We observe that both online algorithms are more fair than Best-Signal with the Greedy-2 slightly more fair than the Greedy-0*. The fairness indexes for both online algorithms are almost always greater than 0.8.

VII. RELATED WORK

The Proportional fair scheduler was first presented in [4], [16] for the third generation EV-DO wireless data system. Since then, there has been a lot of work on the topic of proportional fair scheduling.

The performance of PF has been analyzed by several The proportional fair algorithm can be shown to maximize the sum of the logarithm of the throughput of the users attached to each base station [17]. The user-level performance of the proportional fair algorithm in a dynamic setting was analyzed in [9]. Extensions to proportional fair algorithm to support quality of service have been proposed in [18], [19], [20], [21]. The authors in [19], [20] show that an exponential rule performs best in gracefully trading-off delay versus throughput.

Several researchers have also examined the fairness aspect of the proportional fair algorithm [22], [23]. It has been shown that, with users experiencing heterogeneous channel quality, the differences in variances of the channel quality can result in unfairness using the proportional fair algorithm [9]. In [23], the author describes a monitoring-based feedback algorithm to correct the unfairness of the proportional fair algorithm under heterogeneous channel conditions. The authors in [22] propose a new definition of fairness that extends the absolute fairness bound measure derived from the Generalized Processor Sharing discipline [24]. They propose an opportunistic proportional fair algorithm that is a weighted combination of the max rate scheduling [19] and proportional fair scheduling. Note that all the above algorithms, like the proportional fair scheduling algorithm, only support fairness among users attached to a single base station. In this paper, we take a network-wide approach for supporting fairness while exploiting user diversity to maximize throughput.

As we have discussed in Section II, user association changes have been used as one of the techniques for load balancing in third generation wireless data networks. The objective of their work [6], [7] is load balancing given certain definition of load at the BS. However, they do not consider fairness in a network-wide view. We remark that a network-wide proportional fairness formulation enables better tradeoff between system utilization and fairness, and is a natural objective of load balancing. As shown in [5], proportional fairness optimizes the aggregate user utilities (total user satisfaction) if each user's utility function

is an increasing, strictly concave and continuously differentiable function of its throughput. Load balancing and fairness have been considered jointly in the wireless LAN context in [14]. However, the fairness metric considered in [14] was max-min fairness. We remark that, in third generation wireless data networks, a better fairness metric is proportional fairness which is currently used in deployed third generation data networks.

VIII. Conclusion and future work

In today's networks, users are associated to base stations with the strongest signal strength and each base station independently executes the proportional fairness scheduling algorithm. This approach can result in non-Pareto optimal bandwidth allocation when considering the network as a whole. Therefore, we formulate and study a generalized proportional fairness problem where user associations to base stations are based on optimizing a generalized proportional fairness objective.

We show that this problem is NP-hard and hard to approximate in general. For the special case, which roughly holds in practice, we obtain efficient offline and online algorithms. Our results show that the throughput and fairness can both be improved in heterogeneous user distributions when compared to an approach that assigns a user to the base station with the best signal strength.

For future research, we would like to incorporate other load balancing techniques such as cell breathing into the consideration of network-wide proportional fairness. In addition, we would like to consider uplink scheduling.

IX. Acknowledgement

We are grateful to the helpful discussions with Sem Borst which was instrumental in arriving at the generalized proportional fairness formulation. We would also like to thank Chandra Chekuri for pointing out the possible connection between 3-dimensional matching and our problem.

References

[1] TIA/EIA/cdma2000, *Mobile Station - Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular Systems*. Washington: Telecommunication Industry Association, 1999.

[2] UMTS, *Release 5*. 3G Partnership Project.

[3] http://www.cdg.org, "Cdma development group."

[4] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and A. Viterbi, "A bandwidth efficient high speed wireless data service for nomadic users," *IEEE Communications Magazine*, July 2000.

[5] F. P. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.

[6] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," in *Proceedings of the 10th annual international conference on Mobile computing and networking*, pp. 302–314, 2004.

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

[8] "1xEV: 1x EVolution IS-856 TIA/EIA Standard - Airlink Overview." QUALCOMM Inc. White Paper, Nov. 2001.

[9] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *In Proceedings of INFOCOM*, (San Francisco, CA), April 2003.

[10] J. K. Lenstra, D. B. Shmoys, and E. Tardos, "Approximation algorithms for scheduling unrelated parallel machines," *Mathematical Programming*, vol. 46, pp. 259–271, 1990.

[11] V. Vazirani, *Approximation Algorithms*. Springer-Verlag New York, Incorporated, Jun 1999.

[12] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A Bandwidth-Efficient High-Speed Wireless Data Service for Nomadic Users," *IEEE Communications Magazine*, vol. 38, pp. 70–77, Jul. 2000.

[13] W. C. Jakes, *Microwave Mobile Communication*. Wiley, 1974.

[14] Y. Bejerano, S.-J. Han, and L. E. Li, "Fairness and load balancing in wireless lans using association control," in *Proceedings of the 10th annual international conference on Mobile computing and networking*, pp. 315–329, 2004.

[15] R. Jain, D. Chiu, and W. Hawe, *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Systems*. DEC TR-301, Littleton, MA: Digital Equipment Corporation, 1984.

[16] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of cdma hdr: a high efficiency-high data rate personal communication wireless system," in *Proceedings of IEEE Vehicular Technology Conference*, May 2000.

[17] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, June 2002.

[18] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, Feb 2001.

[19] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in hdr," in *In Proceedings of ITC-17*, pp. 793–804, September 2001.

[20] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *Analytic Methods in Applied Probability*, vol. 207, pp. 185–202, 2002.

[21] A. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviations and optimality," *Annals of Applied Probability*, vol. 11, no. 1, pp. 1–48, 2001.

[22] K. Norlund, T. Ottosson, and A. Brunstrom, "Fairness measures for best effort traffic in wireless networks," in *In Proceedings of PIMRC*, 2004.

[23] C. Westphal, "Monitoring proportional fairness in cdma2000 high data rate networks," in *In Proceedings of Globecom*, 2004.

[24] S. Keshav, "An engineering approach to computer networking," in *Addisson Wesley*, 1997.