Title
Cross Layer Dynamic Resource Allocation with Targeted Throughput for WCDMA Data

Journal
IEEE Transactions on Wireless Communications, 7(12)

ISSN
1536-1276

Authors
Zhang, Peifang
Jordan, Scott

Publication Date
2008-12-01

DOI
10.1109/T-WC.2008.070312

Peer reviewed

Cross Layer Dynamic Resource Allocation with Targeted Throughput for WCDMA Data

Peifang Zhang and Scott Jordan, *Member, IEEE*

***Abstract*—We consider resource allocation for elastic wireless applications that measure utility by target connection average throughput and achieved throughput. We construct a framework for connection access control and rate scheduling that supports this class of applications by maximizing long term average utility. We present a decomposition of the problem into connection access control and rate scheduling layers. The connection access control layer considers current commitments and decides whether to admit new sessions. For admitted users, the layer sets a target throughput to be achieved of the lifetime of the session. The rate scheduling layer adjusts the instantaneous rates of each connection on the basis of how well it is achieving each performance goal and on the relative strength of each connection's current channel. We illustrate how commonly used connection access control and rate scheduling techniques can be applied to design these two layers using an exchange of information regarding future target throughputs and achieved throughputs. Through numerical analysis, we show how wireless channel time diversity and multi-user diversity can be exploited to construct a rate scheduling algorithm that is superior to proportional fairness. Finally, this utility-based framework is compared to a method that does not use utility.**

***Index Terms*—Connection access control; rate scheduling.**

I. Introduction

In recent years, wide area wideband wireless networks have grown tremendously with a fusion of multimedia applications, which imposes a large challenge on network throughput and throughput management. We consider resource allocation for elastic wireless applications that do not require a constant bit rate (CBR) but do require more predictable throughput than that provided by best effort (BE) service. While a great deal of research has addressed resource allocation for CBR traffic and for BE traffic, relatively little has addressed this intermediate class of elastic traffic. We propose that such elastic applications should be able to communicate a performance goal to the network, and the network should strive to achieve this goal. In the vast majority of the literature, the performance goals of variable bit rate traffic focused on loss, delay, or short-term throughput. In contrast, we consider here performance goals stated in terms of the *average throughput* to be achieved over the *lifetime of the connection*.

Manuscript received March 21, 2007; revised March 19, 2008 and June 24, 2008; accepted September 13, 2008. The associate editor coordinating the review of this paper and approving it for publication was K. Leung. Portions of this paper appeared in the IEEE 2006 Wireless Communications and Networking Conference. This material is based upon work supported by the National Science Foundation under grant CNS-0137103.

N. Chen is with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92697, USA (e-mail: pfzhang@gmail.com).

S. Jordan is with the Department of Computer Science, University of California, Irvine, CA 92697, USA (e-mail: sjordan@uci.edu).

Digital Object Identifier 10.1109/T-WC.2008.070312

Our goal is to construct a framework for connection access control, resource allocation, and rate scheduling for this intermediate class of traffic. In our framework, applications express their satisfaction with the throughput achieved over their connection lifetime using a utility function that depends on both the performance goal and on the achieved throughput. Applications are presumed to be more satisfied with higher throughput, and given a level of achieved throughput, by a performance goal closer to the achieved result. The reasoning here is that some elastic applications can use the information of a performance goal to improve the quality of the application, *if* the performance goal is achieved.

In our framework, new sessions present this utility function to the network. The network considers its current commitments and decides whether to admit or block the new session. If admitted, the network sets a performance goal in terms of a target throughput to be achieved of the lifetime of the session. In addition, during each session, the network adjusts the instantaneous rates of each connection on the basis of how well it is achieving each performance goal and on the relative strength of each connection's current channel. Our goal is to maximize the long term average utility per unit time. Such maximization involves a complex balance between the number of connections admitted, the power dedicated to maximizing achieved throughput, and the power dedicated to ensuring that target throughputs are achieved. If too many connections are admitted, then utility falls when targets are not met. Similarly, if the network focuses too much on connections with relatively good channels, this may increase aggregate throughput but may decrease utility if targets are not met. In order to manage the interaction between these different effects, we decompose the problem into two layers. A connection access control and resource allocation (CAC-RA) layer decides whether to admit new sessions and sets target throughput for those admitted. A rate scheduling (RS) layer sets instantaneous rates for each live connection. The two layers communicate with each other; the CAC-RA layer occasionally presents to the RS layer a set of future target throughputs for each connection on the basis of whether long-term targets are being met, and the RS layer periodically responds with information on achieved throughputs.

There is a great body of literature on connection access control, resource allocation, and rate scheduling for wireless networks. However, none of this literature integrates CAC and RS to support a class of traffic that evaluates performance on the basis of both target throughput and achieved average throughput. Several utility-based power control schemes integrate RS and CAC, see e.g. [1]–[7]. In addition, there are many cross-layer approaches that integrate physical layer and networking layers aspects of power and rate assignment, but do not address CAC; see [8] for a survey. Finally, there are similar cross-layer approaches for wired networks, see e.g. [9]. However, none of this literature considers the resulting achieved throughput over a connection lifetime. On the flip side, there are papers in the RS literature that consider connection lifetime performance measures such as achieved throughput. Target throughputs are often expressed in terms of fairness, and are often implemented through schedulers that explicitly consider the cumulative throughput achieved by each connection. Common examples include proportional fair scheduling and max-min policies (see e.g. [10]–[12]). In addition, a number of papers attempt to combine efficiency and fairness, and in doing so consider achieved throughput, see e.g. [1], [13]–[16]. However, this wireless RS literature typically focuses exclusively on the decision of rate assignment; arrival and departure of connections is taken as a given, usually because a separate CAC algorithm does not partner with the RS module. In contrast, here we consider acceptance or blocking of new connections as strictly under our control, since we jointly consider CAC and RS. Finally, there is a vast literature on wireless connection access control literature, which often addresses connection-level QoS. This literature typically focuses exclusively on the admit or reject decision [17]; rate scheduling is taken as a given, usually because a separate RS algorithm does not partner with the CAC module. This is particular true of most measurement-based access control (MBAC) approaches, which typically consider source rates as uncontrollable except through flow control. In contrast, here we consider source rates as strictly under our control, since we jointly consider CAC and RS.

The novelty of this paper, therefore, is to integrate CAC and RS to support a class of traffic that evaluates performance on the basis of both target throughput and achieved average throughput. We view our contributions as follows. First, we construct a framework for connection access control, resource allocation, and rate scheduling in which applications express their satisfaction with the throughput achieved over their connection lifetime using a utility function that depends on both the performance goal and on the achieved throughput. Second, we present a decomposition of the problem into a CAC-RA layer that decides whether to admit new sessions and sets target throughputs for those admitted and a RS layer that sets instantaneous rates for each live connection. Third, we illustrate how commonly used CAC and RS techniques can be applied to design these two layers using an exchange of information regarding future target throughputs and achieved throughputs. Fourth, the RS module is shown to be superior to the widely applied proportional fairness scheduler, in terms of both efficiency and fairness. However, as the first attempt to integrate CAC and RS to support a class of traffic that evaluates performance on the basis of both target throughput and achieved average throughput, we do not view the architecture presented as ready for implementation; in particular, we expect that use of more nuanced CAC and RS techniques presented in the literature could replace those used here in order to improve the overall performance.

The rest of the paper is organized as follows. The system model is presented and the main problem is formulated in section II. In section III, we decompose the problem into a set of two problems that can be implemented in two communicating layers. In sections IV and V, connection access control and rate scheduling algorithms are derived. Finally in section VI numerical results are presented to illustrate how the two layers together support a class of targeted throughput traffic, how the interaction between them works, how the RS module outperforms proportional fairness, and how the utility-based CAC outperforms a similar approach that does not consider utility.

II. System Description and Problem Formulation

We consider a single cell downlink WCDMA network where a base station transmits to a number of nomadic users. Time t is indexed by frames of equal length. Users, representing connections, are indexed by i in the order of arrival. Upon arrival, a user is either admitted into the system or blocked from access by the network. Blocked users leave the system and admitted users enter the system immediately. Denote user i's arrival time by s_i and its departure time by o_i. At time t, denote user i's time in the system, called user i's *age*, by $a_i(t) = t - s_i$. During a user's connection, we assume there is always traffic destined to each user.

Our goal is to construct a framework for connection access control, resource allocation, and rate scheduling in which applications express their satisfaction with the throughput achieved over their connection lifetime using a utility function that depends on both the performance goal and on the achieved throughput. In the following, we first introduce a novel utility function, which serves as a measure of users' satisfaction with received QoS, expressed in terms of time-average throughput over the connection duration. We then consider the mapping between rate and power, and constraints on both resources. Based on these metrics and constraints, we pose a system optimization problem to maximize time-average aggregate utility subject to transmission power and rate constraints. Finally, for purposes of illustration, we assume specific traffic and channel models.

We start with metrics describing each user. Denote user i's target throughput by X_i and its achieved throughput over its connection lifetime by $x_{f,i}$. For convenience, denote the ratio of achieved throughput to target throughput by $g_{f,i} = x_{f,i}/X_i$, called user i's achievement ratio. User i expresses its satisfaction with the throughput achieved over its connection lifetime using a utility function $U_i(X_i, g_{f,i})$ that depends on both the performance goal and on the achieved throughput, or equivalently on the target X_i and the achievement ratio $g_{f,i}$. We note that this form of utility function, with its dependence upon the combination of target throughput and achieved throughput, is novel. While many other papers have represented a user's satisfaction using a utility function, typically it only expresses satisfaction with short-term or instantaneous throughput. In addition, we believe that using a utility function that allows the network to determine the target throughput may be superior to approaches in which the user dictates a fixed target throughput upon arrival, since in this manner the network may adjust targets in response to operating conditions. Such a soft throughput guarantee may help avoid the infeasibility and efficiency degradation issues associated with hard throughput guarantees. Alternatively, the utility function could be chosen by the network rather than the user, on the basis of the type of application or to balance short-term and long-term goals.

We propose that the the utility function $U_i(X_i, g_{f,i})$ should have the following properties: (1) $U_i(X_i, g_{f,i})$ is increasing both in X_i and in $g_{f,i}$; (2) $U_i(X_i, g_{f,i})$ is concave in X_i; (3) $U_i(X_i, g_{f,i})$ is convex in $g_{f,i}$ if $g_{f,i} < 1$, and concave in $g_{f,i}$ otherwise; (4) For fixed $X_i g_{f,i} = C$, $U_i(X_i, g_{f,i})$ is maximum at $X_i = C$ (i.e. $g_{f,i} = 1$); and (5) $\frac{\partial U_i(X_i, g_{f,i})}{\partial X_i}|_{X_i=0}$ is finite. When utility has been used in the literature to represent satisfaction with instantaneous throughput, it is common to assume that utility is increasing with throughput, see e.g. [4], [18]; in property 1, we assume that utility is similarly increasing with both target throughput and achieved throughput (or equivalently with both target throughput and the achievement ratio). When there is no target, it is common to assume that utility is a concave function of instantaneous throughput; here, in property 2 we assume that utility is a concave function of the target throughput. However, once a target is fixed, in property 3 we assume that utility is a mixed convex and concave function of the achievement ratio. This form, known as a *sigmoid-shape* utility, is common when a target is present, see e.g. [1]; it reflects decreasing absolute increments in utility as the achieved throughput differs on either side from the target. Property 4 assumes that, *given a level of achieved throughput*, applications are more satisfied the closer the performance goal is to the achieved result. As a consequence, an application may be more satisfied by a moderate achieved throughput that is close to a moderate performance goal than by a high achieved throughput that falls significantly short of a higher performance goal. Finally, a target throughput of 0, $X_i = 0$, will indicate that the user is blocked from the system; property 5 is required to ensure that the blocking probability of new connection requests is positive.

Denote the instantaneous rate assigned to user i at time t by $r_i(t)$. Then user i's average throughput, from entry into the network until time t, denoted by $x_i(t)$, is the sum of its assigned rates divided by its age: $x_i(t) = [\sum_{\tau=s_i}^{t} r_i(\tau)]/a_i(t), \forall s_i \le t \le o_i$. User i's achieved throughput over its connection lifetime is thus given by $x_{f,i} = x_i(o_i)$. We now turn to the mapping between rate and power, and constraints on both resources. Denote the downlink transmission power allocated to user i by the base station at time t by $P_i(t)$. We assume that there is a constraint on the total power allocated:

$$\sum_{i \in A(t)} P_i(t) \le \mathcal{P}, \tag{1}$$

where $A(t)$ is the set of active users. We also assume that there is an upper bound on the instantaneous rate that can be assigned to any individual user,

$$0 \le r_i(t) \le 1. \tag{2}$$

This upper bound on instantaneous rate implies that there exists a corresponding upper bound on a user's throughput target which we denote by X_{max}. At this stage, we assume a general mapping from the instantaneous rate of each user to the required downlink transmission power allocated to user i, which is expected to depend on each user's channel. The power constraint (1) and the rate constraint (2) are thus expected to be interrelated.

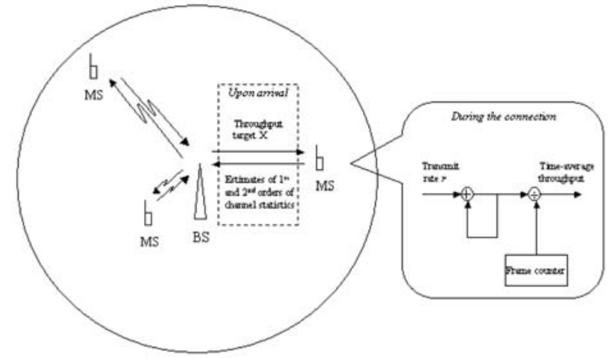


Fig. 1. Illustration of system model.

Based on these metrics and constraints, we now turn to posing a system optimization problem to maximize time-average aggregate utility subject to transmission power and rate constraints. We formulate this as a centralized nonlinear optimization problem, called **Problem Centralized:**

$$\max_{\{X_i, \{r_i(t), \forall t\}, \forall i\}} \lim_{T \to \infty} \frac{\sum_{i=1}^{N(T)} U_i(X_i, g_{f,i})}{T}, \text{s. t.}(1) \ \& \ (2) \tag{3}$$

where $N(T)$ denotes the number of users who depart during (0,T].

Recall that our design goal is to construct a framework for connection access control, resource allocation, and rate scheduling in which applications express their satisfaction with the throughput achieved over their connection lifetime using a utility function that depends on both the performance goal and on the achieved throughput. Let us examine whether the optimization problem achieves this goal. User utilities do depend on both the target throughput and on the achievement ratio. The decision space is a joint CAC and RS problem: to jointly decide which users gain entry ($X_i > 0$), choose target throughputs X_i for admitted users, and set rate allocations $r_i(t)$ at all times t for admitted users. The optimization metric is the total long term average utility per unit time, based on the target throughputs and achieved throughputs of all departed users. The feasible region of target throughputs and instantaneous rates that can be assigned are constrained by power and rate constraints. The model is illustrated in Fig. 1.

There are two principal difficulties with this formulation, however. First, the decisions are off-line, in that they can potentially depend not only on current and past system states but also upon the future, in particular upon future channel conditions, arrivals and departures. Second, the CAC and RS decisions are jointly made, which is likely to be far too complicated. In the next section, we formulate an on-line problem and decompose the CAC and RS algorithms.

III. A TWO LAYER CAC-RA AND RS DECOMPOSITION

In this section, we decompose *Problem Centralized* into a set of two problems that can be implemented in two communicating layers. The motivation for the definition of each layer comes from the recognition that CAC operates on the connection level time scale defined by the connection request interarrival time, whereas wireless RS operates on the packet level time scale defined by frame duration and slow fading. Our CAC-RA layer will be tasked with the decision of whether to admit new sessions and with setting target throughputs for those admitted. Our RS layer will be tasked with rate and power allocation of each frame for each active connection. In order to accomplish this decomposition, we need to create a pair of on-line problems with individual decision variables, optimization objectives, and constraints.

The interface between the two layers will be defined by the information exchanged between them. We propose that CAC-RA should have sole knowledge of utility, that RS should have sole knowledge of channel attenuation, and that the two layers should exchange information about target and achieved throughput. Denote the future target throughput for user i by y_i. Recalling that the cumulative throughput achieved by user i is given by $x_i(t)$, denote the corresponding achievement ratio for user i at time t as $g_i(t) = x_i(t)/X_i$. The general idea is that the CAC-RA layer should occasionally instruct the RS layer to attempt to achieve $y_i, i \in A(t)$, and that the RS layer should periodically report back what throughputs $g_i(t)$ it has been able to achieve. The CAC-RA layer will thus be able to adjust future targets both on the basis of the number of active users in the system and on the difference between the long term target throughputs and the actual achieved throughputs. The RS layer will be able to adjust power and rate allocations both on the basis of the future target throughput and on the relative strength of each user's channel.

We turn to separating the optimization objective into a pair for each sub-problem, and on transforming *Problem Centralized* from an off-line version that depends on future information to a dynamic on-line problem that relies only on past and current information. We start with the objective function for the CAC-RA layer. A common approach to MBAC is to consider only those users that have arrived so far and to ignore predictions of future arrivals. We adopt that approach here. In that spirit, we restrict the optimization metric at time t to those users who have arrived by time t, so that the total utility becomes $\sum_{i\in A(t)} U_i(X_i, g_{f,i})$. From this point on, we make two traffic assumptions: that users arrive according to a Poisson process with a rate of λ users/frame, and that user i's connection duration is Exponentially distributed with mean $1/\mu_i$ sec, independent of other users connection durations.

This CAC-RA optimization metric still relies on future actions through the final achievement ratios $g_{f,i}$. To remove this dependence, we construct a prediction of user i's final achievement ratio $g_{f,i}$, denoted by $\hat{g}_{f,i}$. By assuming that the CAC-RA layer will maintain the same target future throughput for user i from time t until the end of user i's connection, the prediction is:

$$\hat{g}_{f,i} = g_i(t) \cdot \frac{a_i(t)}{a_i(t) + 1/\mu_i} + \frac{y_i}{X_i} \cdot \frac{1/\mu_i}{a_i(t) + 1/\mu_i}, \quad (4)$$

If connection durations are assumed to be other than Exponentially distributed, then the weights of the two terms would be changed correspondingly. In addition, more complex prediction methods could be used, if one wanted to incorporate knowledge about the distribution of future active users.

With the reliance on future actions removed, we can now pose an optimization objective for the CAC-RA layer. Upon arrival of user j, the layer must decide whether to admit or block the arrival; if admitted (denoted by $\mathbf{I}_{X_j>0}$), it must set a long term target throughput X_j for the arrival, and it must choose a set of future target throughputs y_i for all active users:

$$\max_{\{X_j, y_i, \forall i\}} \left[\sum_{i\in A(t)} U_i(X_i, \hat{g}_{f,i}) + \mathbf{I}_{X_j>0} \cdot U_j(X_j, y_j/X_j) \right].$$

The first term in the objective is a summation over current users, with their final achievement ratios replaced by the predictions. The second term considers admission and target throughput assignment for a new arrival. At user departures, the objective is similar but without the presence of the X_j decision variable or the second term, and with the departed user excluded from $A(t)$.

We now turn to the objective function for the RS layer. Rather than passing the future target throughputs y_i to the RS layer, the CAC-RA layer encodes these into the predicted achievement ratios $\hat{g}_{f,i}$, which the RS layer interprets as target achievement ratios. On the frame scale, the RS layer must determine individual rates, $r_i(t)$. Denote the wireless channel of user i at time t by $h_i(t)$. Rate scheduling will be done on the basis of the current channel condition, these target achievement ratios, and the current achievement ratios. The overall system objective in (3) is to maximize the average utility $(\sum_{i=1}^{N(T)} U_i(X_i, g_{f,i}))/T$. Utility information, however, is restricted to the CAC-RA layer, so the RS optimization metric should be expressed in terms of throughput. Due to the sigmoid shape of the utility function, it is reasonable to set the goal of the RS layer as achieving the future target throughputs, y_i, e.g. $g_{f,i} = 1\ \forall i$. However, the CAC-RA layer may err on how many users it admits, and posing an equality goal for the RS layer is likely to be too rigid.

A common approach in the RS literature is to attempt to maximize the smallest rate. Here, since the goal is not to equalize rates but to maximize utility based on target throughputs, we could consider maximizing the smallest achievement ratio $g_{f,i}$. However, such a formulation would ignore utility and rob the CAC-RA layer of its ability to put different weights on different users. Therefore, we set the RS objective as maximizing the smallest ratio $g_i(s)/\hat{g}_{f,i}$ at some future time s. The CAC-RA layer thus can influence the behavior of the RS layer by appropriate choice of y_i and thus of $\hat{g}_{f,i}$. It remains to determine over what time frame the RS layer should attempt to accomplish this objective, and how. We consider these issues in section V below. The resulting optimization metric for the RS layer is $\max_{\{r_i(t),\forall i\},\forall t} \min_{i\in A(t)} g_i(s)/\hat{g}_{f,i}$.

Each layer now has an optimization metric and a set of decision variables that depend only on information and variables that are at the disposal of that layer. We turn next to the constraint for each layer's optimization problem. The constraints in *Problem Centralized* depend on information and variables from both layers. We need to decompose these constraints so that each layer has constraints relevant to its own optimization metric and decision variables. We start with the simpler RS layer. The rate constraint (2) can be applied directly to this layer without modification. Up to this point, we have assumed a general mapping from the instantaneous rate of each user to the required downlink transmission power allocated to user i. From this point on, we need to incorporate a specific mapping. For simplicity, we assume that the location of a new user is assumed to be fixed during the connection, and to be uniformly distributed within the cell. We assume the wireless channel $h_i(t)$ of user i at time t consists of distance-based pathloss and slow fading, $h_i(t) = (d_i/d_0)^n \psi_i(t)$, where d_0 is a far-field reference distance, d_i is user i's distance to the base station, n is the pathloss exponent, and $\psi_i(t)$ is a log-normal random variable with zero mean and user-specific variance $\sigma^2_{\psi_{dB,i}}$. We assume there is a linear relationship between power and rate, $P_i(t) = h_i(t)r_i(t)$, so that the power constraint (1) can be applied to the RS layer as $\sum_{i\in A(t)} h_i(t)r_i(t) \leq \mathcal{P}$. Of course, different mappings between rate and power and different channel models can be incorporated into the general framework, with corresponding changes to the form of this constraint.

We now turn to constraints for the CAC-RA layer. Power and rate constraints at the RS layer, in conjunction with the RS algorithm used, should result in constraints on the long term target throughputs X_i and the future target throughputs y_i. Denote the first two moments of channel attenuation for user i by m_{h_i} and $\sigma^2_{h_i}$, which we assume are known to the CAC-RA layer. Since CAC-RA operates at a longer time-scale than RS, a straight-forward approach would be to average the RS power constraint over time. However, any rate scheduler that includes an opportunistic component will base the assigned rate in part on the user's channel, which results in $E[h_i r_i] < m_{h_i} y_i$. Denote a specific rate scheduling policy by Π. We propose a CAC-RA constraint of the form:

$$\sum_{i=1}^{M} m_{h_i} y_i \leq \phi_P(\Pi)\mathcal{P}, \quad (5)$$

where $\phi_P(\Pi) \geq 1$, which we call the *power efficiency factor*, is a measure of how much efficiency is gained by opportunistic components within the rate scheduler; more opportunistic rate schedulers result in higher values of $\phi_P(\Pi)$. The rate limit in the RS layer certainly places a corresponding limit $X_i \leq 1$ on long term target throughputs and $y_i \leq 1$ on future target throughputs. However, any use of opportunistic scheduling results in users receiving less than the maximum rate for some portion of time, regardless of their target throughputs. We therefore similarly replace the rate constraint by $0 \leq y_i, X_i \leq \phi_R(\Pi)$, where $0 < \phi_R(\Pi) < 1$, which we call the *rate fluctuation factor*, depends on the rate scheduling policy. Both $\phi_P(\Pi)$ and $\phi_R(\Pi)$ must be set with knowledge of the rate scheduler. This is discussed further below.

With the decision variables, optimization metrics, and constraints separated by layer, we can now state the final framework. The functionality of each layer and the communication between layers is illustrated in Fig. 2, where i, j and k are used to denote the indices of current, arriving and departing users. The CAC-RA layer operates on a connection time-scale. Upon arrivals or departure of users, it attempts to maximize the total utility of all active users by deciding on acceptance or blocking of new users, assignment of long term target throughput X_j to arrivals, and assignment of future target throughputs y_i to active users. CAC-RA encodes long term target throughputs and future target throughputs into a set of target achievement ratios $\hat{g}_{f,i}$, and passes these to the RS layer. RS operates on a frame time-scale. Once each frame, the scheduler attempts to meet the target achievement ratios of all current users in the system. The rate scheduler must balance efficiency (achieving high throughput ratios) with fairness (achieving similar throughput ratios). Below, we suggest that the rate scheduler can achieve efficiency by adapting the transmission rate to the time-varying fading channel $h_i(t)$, and can achieve fairness using a max-min target objective.

The information flow is as follows. When new users arrive, the system roughly estimates the first two moments of their channel. Upon any admittance to the network or any departure, CAC-RA informs RS of the target achievement ratios for each active user. Fast power control (at a lower layer) provides frame-level channel information to RS. In turn, RS periodically informs CAC-RA of the current achievement ratios for each active user. This separation of functionality by layer implements a negative feedback loop. If CAC-RA sets a target throughput X_i or y_i unrealistically high, RS will achieve a low achievement ratio $g_i(t)$ in the period before the next arrival or departure. This low $g_i(t)$ will be fed back to CAC-RA, which will increase the priority of this user and lower the target achievement ratios of other users, resulting in increasing an increase in future values of $g_i(t)$ for the disaffected user. These effects will be demonstrated in numerical analysis below. In the following two sections, algorithms for each layer are developed. Afterward, we return to analysis of the combined two layer architecture.

IV. Connection Access Control and Resource Allocation

In this section, we consider the CAC-RA problem, and formulate an on-line algorithm. From the previous section, the CAC-RA problem can be stated as **Problem CAC-RA:**

$$\max_{y_i, i\in A(t), X_j} \sum_{i\in A(t)} U_i(X_i, \hat{g}_{f,i}) + \mathbf{I}_{X_j>0} \cdot U_j(X_j, y_j/X_j),$$

$$\text{s. t.} \sum_{i\in A(t)} m_{h_i} y_i \leq \phi_P(\Pi)\mathcal{P}, 0 \leq y_i, X_j \leq \phi_R, \ \forall i \in A(t).$$

The problem must be solved upon user arrivals and departures. The optimization metric is to maximize the expected utility to be gained by current users, plus a new arrival if applicable. The decisions to be made are the future target throughputs to assign to each active user, whether to admit arrivals, and the long term target throughput to assign to an arrival. Since the utility function is a sigmoid-shape, the optimization problem is not a concave program, and hence simple gradient-based searches are not guaranteed to converge to the global optimum. We therefore look for a heuristic algorithm. We will consider assignment of future target throughputs and treatment of new arrivals separately.

CAC-RA

$$\xrightarrow{m_{h_j},\sigma^2_{h_j}} \quad \max \sum_{i\in A(t)} U_i(X_i,\hat{g}_{f,i}) + I_{X_j>0} U_j\left(X_j,\frac{y_j}{X_j}\right)$$

$$\xleftarrow{m_{h_k},\ \sigma^2_{h_k}} \quad \text{s.t.} \sum_{i\in A(t)} m_{h_i} y_i \le \phi_P(\Pi)\mathcal{P}$$

$$0 \le y_i, X_j \le \phi_R$$

RS

$$\xrightarrow[m_{h_j},\sigma^2_{h_j}]{\{X_i,\hat{g}_{f,i}\}} \quad \max \min_i \left\{\frac{g_i(s)}{\hat{g}_{f,i}}\right\}$$

$$\xleftarrow{\{g_i\}} \quad \text{s.t.} \sum_i h_i(t) r_i(t) \le \mathcal{P}$$

$$0 \le r_i(t) \le 1$$

Fast PC

$$\xleftarrow{\{h_i(t)\}}$$

Fig. 2. Decomposition into layers.

With respect to assignment of future target throughputs, we apply a water-filling strategy, which requires treating the convex and concave portions of the utility function separately. We begin by focusing on a pair of users. We start with a tentative assignment of resources, and consider assignment of a rate increment. For an arrival, denote $y_j = X_j$. To focus on the decision variables, denote $V_i(y_i) = U_i(X_i, \hat{g}_{f,i})$. Recall that by assumption $U_i(X_i, \hat{g}_{f,i})$ is convex in $\hat{g}_{f,i}$ when $\hat{g}_{f,i} < 1$ and concave in $\hat{g}_{f,i}$ when $\hat{g}_{f,i} > 1$. We first consider the case in which both users are in the concave portion of the utility curve, and then consider the case in which both users are in the convex portion of the utility curve. For the concave case, denote the normalized marginal utility per unit y_i by $G_i = \frac{1}{m_{h_i}}\frac{dV_i(y_i)}{dy_i}$.

Proposition 1: *For any two users, if both have $\hat{g}_{f,i} \ge 1$, or if one user has $\hat{g}_{f,i} \ge 1$ and the other is an arrival, then it's optimal to assign an infinitesimal rate increment to the user with the higher G_i.*

Proof: We consider the following optimization over two users:

$$\max_{y_k,y_n} V_{tot} = V_k(y_k) + V_n(y_n), \text{s. t. } m_{h_k}y_k + m_{h_n}y_n \le x \quad (6)$$

The Lagrangian is formed as $\mathrm{L} = V_k(y_k) + V_n(y_n) + \alpha(x - m_{h_k}y_k - m_{h_n}y_n)$. Taking the derivative of L over y_i (where $i = k$ or n), we have $\partial\mathrm{L}/\partial y_i = dV_i(y_i)/dy_i - m_{h_i}\alpha$, where α, the shadow cost associated with the power constraint, is positive only when the power constraint is binding. So since both $V_k(y_k)$ and $V_n(y_n)$ are locally concave, L is concave in (y_k, y_n). From standard nonlinear programming, the optimal resource allocation policy is to assign infinitesimal incremental resources to the user with higher $\frac{1}{m_{h_i}}\frac{dV_i(y_i)}{dy_i}$. □

G_i, which we call the *gradient priority*, is the marginal utility that can be earned per unit future throughput target, normalized by the user's mean channel. It can be expressed separately for current and new users:

$$G_i = \begin{cases} \frac{1/\mu_i}{a_i(t)+1/\mu_i}\frac{1}{m_{h_i}X_i}\frac{dU_i(X_i,\hat{g}_{f,i})}{d\hat{g}_{f,i}}, \\ \quad \text{if current user } \& \ \hat{g}_{f,i} > 1, \\ \frac{1}{m_{h_j}}\frac{dU_j(X_j,y_j/X_j)}{dX_j}, \\ \quad \text{if new user.} \end{cases} \quad (7)$$

In the next proposition, we consider two users who are both in the convex portion of the utility curve. Again assume that a tentative resource assignment has been made and consider an incremental power assignment of δ. In this case, the increment need not be infinitesimal, so we can consider a chord of the utility function. We define call this chord the *arch gradient priority*, and define it as:

$$G_i^\delta = \frac{U_i(X_i, \hat{g}_{f,i} + \frac{1/\mu_i}{a_i+1/\mu_i}\frac{\delta}{m_{h_i}X_i}) - U_i(X_i,\hat{g}_{f,i})}{\delta}. \quad (8)$$

Proposition 2: *For any two current users, if both satisfy $\hat{g}_{f,i} + \frac{1/\mu_i}{a_i+1/\mu_i}\frac{\delta}{m_{h_i}X_i} < 1$, then it is optimal to assign the incremental power δ to the user with the higher G_i^δ.*

Proof: Consider the problem in (6) with $\hat{g}_{f,k} < 1, \hat{g}_{f,n} < 1$, and $G_k^\delta > G_n^\delta$. Suppose it's optimal to assign ϵ $(0 \le \epsilon \le \delta)$ to user k and $\delta - \epsilon$ to user n, then

$$\begin{aligned} V_{tot} &= U_k(X_k, \hat{g}_{f,k} + \frac{1/\mu_k}{a_k + 1/\mu_k}\frac{\epsilon}{m_{h_k}X_k}) \\ &\quad + U_n(X_n, \hat{g}_{f,n} + \frac{1/\mu_n}{a_n + 1/\mu_n}\frac{\delta-\epsilon}{m_{h_n}X_n}) \\ &= G_k^\epsilon\epsilon + G_n^{\delta-\epsilon}(\delta-\epsilon) + U_k(X_k,\hat{g}_{f,k}) + U_n(X_n,\hat{g}_{f,n}). \end{aligned} \quad (9)$$

Since $U(X, g)$ is convex at $g < 1$, $G_k^\delta > G_k^\epsilon$ and $G_n^\delta > G_k^{\delta-\epsilon}$ for any $\epsilon < \delta$. Therefore, the maximal V_{tot} in (9) is achieved by $\epsilon = \delta$. □

The case in which two current users are in the convex portion of the utility curve, but an incremental power assignment of δ would push one into the concave portion of the curve, can be easily handled by defining the utility chord from the current assignment to the point where $\hat{g}_{f,i} = 1$, i.e. replace G_i^δ by $G_i^{\delta_1}$ where $0 < \delta_1 < \delta$ is such that $\hat{g}_{f,i} + \frac{1/\mu_i}{a_i+1/\mu_i}\frac{\delta_1}{m_{h_i}X_i} = 1$. The remaining case is that in which one user is in the concave portion of the utility curve and other other user is in the convex portion. The optimal solution to this mixture would involve a global search, not an iterative algorithm, and the resulting complexity would be unacceptably high. Below, we therefore simply directly compare the gradient priority for the user in the concave portion to the arch gradient priority for the user in the convex portion, recognizing that this may lead to suboptimal results.

With insight from these cases, we can now formulate an iterative algorithm to determine the resource allocation for an arbitrary number of users. The algorithm is based on iterative incremental assignment of power, based on the gradient priority users in the concave portion of the utility curve and on the arch gradient priority for users in the convex portion. In the resulting algorithm, a high priority is associated with a good channel (low $m_{h,i}$), a low target throughput, a young user (small a_i), a long expected holding time (large $1/\mu_i$), and/or $\hat{g}_{f,i}$ close to 1 (large $\frac{\Delta U_i(X,g)}{\Delta g}$).

With respect to treatment of arrivals, we look to the measurement-based access control literature for guidance. Although a wide variety of more complex policies have been proposed, a common basic approach is to admit a user if and only if there are currently sufficient resources to satisfy current performance commitments [17]. In our framework, we interpret this philosophy as admitting an arrival if and only if the above algorithm results in allocation of a strictly positive target throughput for the arrival, i.e. $y_j = X_j > 0$. Of course, more complex strategies could be used, if one wanted to incorporate knowledge about the distribution of future active users or if one wanted to directly incorporate current achievement ratios into the decision.

The resulting CAC-RA algorithm is shown in Table I. The effectiveness of this algorithm will be evaluated in section VI.

V. Rate Scheduler

In this section, we consider the rate scheduling (RS) problem, and formulate an on-line algorithm. From section III, the RS problem can be stated as **Problem RS:**

$$\max_{\Pi} \min_{i \in A(t)} \left\{ \frac{g_i(s)}{\hat{g}_{f,i}} \right\},$$
$$\text{s. t.} \sum_{i \in A(t)} h_i(t) r_i(t) \leq \mathcal{P}, 0 \leq r_i(t) \leq 1, \forall i \in A(t).$$

The general goal of the RS layer is to achieve the future target throughputs y_i. These targets have been encoded into the desired achievement ratios $\hat{g}_{f,i}$ that are handed to the RS layer from the CAC-RA layer upon each user arrival or departure. Following much of the rate scheduling literature, we have cast this goal as a max-min problem. However, the goal is not to equalize rates but to maximize utility based on target throughputs, so we set the RS objective as maximizing the smallest ratio of $g_i(s)/\hat{g}_{f,i}$ at some future time s. It remains to determine over what time frame to attempt to accomplish this objective, and how.

The rate scheduler must balance efficiency (high achievement ratios) with fairness (balanced achievement ratios). In the rate scheduling literature, two types of algorithms are common. Some algorithms focus on fairness, often directly using a max-min formulation; conceptually, these algorithms consider a small future time s, and attempt a short-term equalization. In contrast, other algorithms focus on efficiency using opportunistic methods, by adapting the transmission rate to the time-varying fading channel $h_i(t)$; conceptually, these algorithms are using a larger future time s, and attempt longer term maximization. We apply both of these ideas here with a novel tradeoff between opportunistic and max-min components.

For the efficiency component, we propose an opportunistic policy, similar to that used in [19]. The idea is to allocate higher transmit rate to users who are currently experiencing less fading than average. This approach takes advantage of channel variations among multiple users at a particular time (multi-user diversity), and channel fluctuations in time for each user (temporal diversity), using estimates of channel statistics. The resulting algorithm is shown in Table I as Scheduler OS. For the fairness component, we propose a greedy algorithm to solve *Problem RS*, in which $g_i(s)$ is replaced by the current $g_i(t)$. This approach grants absolute priority to users who are not meeting their target achievement ratios, and particularly helps users who have suffered from severe fading for a period of time. The resulting algorithm is shown in Table I as Scheduler MM.

Our proposal is to create a tunable tradeoff between these two policies. We propose to allocate a portion of the power budget to each. Specifically, given an *efficiency index* $0 \leq \theta \leq 1$, the opportunistic policy is given a power budget $\theta\mathcal{P}$, and the max-min policy is given the remaining power. The tunable scheduler is shown in Table I as Scheduler Tunable. At $\theta = 0$ therefore, it reduces to the max-min scheduler, while at $\theta = 1$ it reduces to the opportunistic scheduler. Heavier use of the opportunistic scheduler will increase efficiency but reduce fairness. It will be seen that the tunable scheduler outperforms the well-known proportional fairness throughout a large range of θ. Of course, more complex hybrid policies could be created which may outperform this simple hybrid.

VI. Numerical Analysis

The goal of this paper is to integrate CAC and RS to support a class of traffic that evaluates performance on the basis of both target throughput and achieved average throughput. In previous sections, we presented a framework for connection access control, resource allocation, and rate scheduling in which applications express their satisfaction with the throughput achieved over their connection lifetime using a utility function that depends on both the performance goal and on the achieved throughput. Numerical results are presented in this section to illustrate how the CAC-RA and RS layers together support a class of targeted throughput traffic, how the interaction between CAC-RA and RS works, and the strengths and limitations of the simple MBAC and RS algorithms that we adapted from the literature into our framework. We start by studying the performance of the rate scheduler in isolation. After this, we proceed to the performance of the combined CAC-RA and RS architecture. The following parameters are used throughout this section: far-field reference distance $d_0 = 0.1$, distance from base station d_i is uniformly distributed in (0.1,1), and the path-loss exponent $n = 3.71$. We define the *load* as the ratio of normalized power, $\frac{\lambda}{\mu}\frac{E[m_h]}{2}$, to the power supply $\mathcal{P}$.

A. Performance of the Rate Scheduler

In this subsection, the RS algorithm is isolated and evaluated. Since both opportunistic and max-min rate schedulers have been studied extensively in the literature, the principal focus here is whether such standard algorithms can be adapted to address achievement ratios, rather than simpler performance measures. In addition, we show how the RS algorithm can be used to balance resource efficiency and fairness, and how it can outperform proportional fairness with respect to both metrics.

The following parameters are used throughout this subsection: 100 simultaneous users with target throughputs X_i uniformly distributed in (0,1], connection duration of 3000 frames, shadowing variance $\sigma_{\psi_{dB}} = 6$, and $load$ = 2. We judge the performance of the RS hybrid algorithm by the set

TABLE I
CAC AND RS ALGORITHMS

CAC-RA:

1). Event initialization:
Upon an arrival or departure, update the active user set $A(t)$. Initialize $P_{res} = \phi_P(\Pi)\mathcal{P}$, $X_j = 0$, $y_i = 0$ and $\hat{g}_{f,i} = g_i$, $\forall i \in A(t)$.

2). Priority measure calculation:
$\forall i s.t. \hat{g}_{f,i} < 1$, calculate G_i^{δ} according to (8). $\forall i s.t. \hat{g}_{f,i} > 1$, calculate G_i according to (7).
If the event is an arrival, then calculate G_j according to (7).

3). Resource allocation:
3a. Pick the user with the highest priority measure.
If $\hat{g}_{f,i} < 1$, then assign a rate increment Δ_i ($y_i \leftarrow y_i + \Delta_i$), such that either the updated $\hat{g}_{f,i} = 1$ or $y_i = \phi_R$ or $P_{res} = 0$;
If $\hat{g}_{f,i} > 1$ or the user is an arrival, then assign a rate increment Δ_i, such that either $\Delta_i = \bar{\Delta}$ (a fixed step), or the resulting $y_i = \phi_R$.
3b. Update $\hat{g}_{f,i}$ according to (4), G_i according to (7), reduce P_{res} by $m_{h,i}\Delta_i$, and remove any users with $y_i = \phi_R$ from $A(t)$.
3d. If either $P_{res} = 0$ or $A(t) = \emptyset$, go to step 4; else go to 3a.

4). Admission:
If the event is an arrival, and $X_j > 0$, then admit the new user. Wait for a new arrival or departure, and return to step 1.

Scheduler OS:

1). Initialize $B(t) = A(t)$, and sort $B(t)$ by increasing normalized shadowing, $e_i = \psi_{dB,i}(t)/\sigma_{\psi_{dB,i}}$.
2). In the resulting order, assign the full rate, $r_i(t) = 1$, to as many users as possible as long as (1) is satisfied, and remove these full-rate users from $B(t)$.
3). Assign any residual power to the first remaining user in $B(t)$. Assign $r_i(t) = 0$ to all other users in $B(t)$.

Scheduler MM:

1). Initialize $B(t) = A(t)$, and sort $B(t)$ by increasing $\frac{g_i(t)}{\hat{g}_{f,i}}$.
2). In the resulting order, assign rates subject to (1) and (2).

Scheduler Tunable:

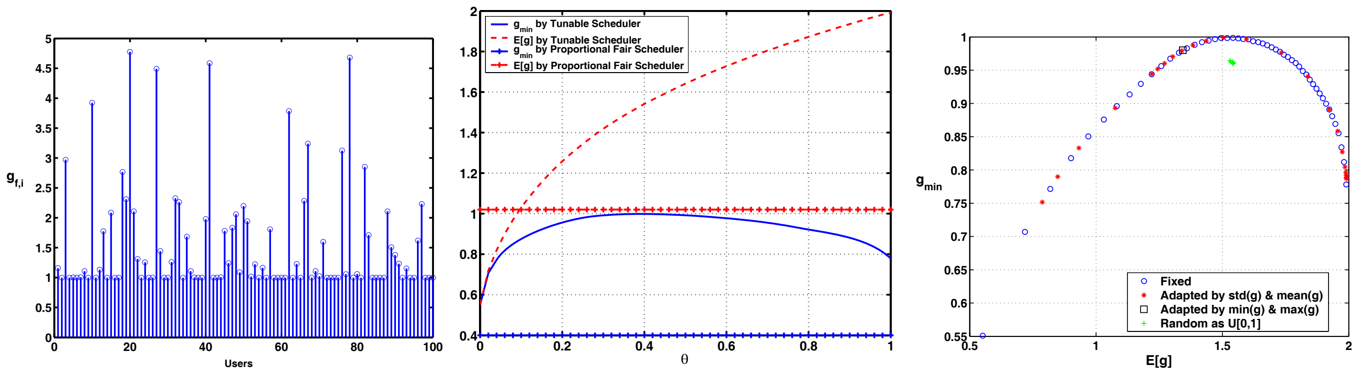
1). Initialize $B(t) = A(t)$. Run Scheduler OS, with $\mathcal{P}$ replaced by $\theta\mathcal{P}$ in (1). Update $\{g_i(t)\}$, and pass leftover power to the next step.
3). Run Scheduler MM, with $\mathcal{P}$ replaced by the leftover power in (1). Update $\{g_i(t)\}$.
4). Update $t = t + 1$ and $a_i(t) = a_i(t) + 1$. If an arrival or departure occurs, report $\{g_i(t)\}$ to CAC-RA. Otherwise, return to step 1.

of users' achievement ratios, $g_{f,i}$, and the minimum and mean achievement ratios.

Users' throughput achievement ratios $g_{f,i}$, resulting from the RS with an efficiency index $\theta = 0.4$, are shown in Fig. 3(a). All users meet or exceed their targets using the tunable scheduler, i.e., $g_i \geq 1, \forall i$. A small number of users achieve significantly higher throughput ratios, because they either have relatively low target throughputs or are close to the base station. In Fig. 3(b), the algorithm tunability is illustrated by varying the efficiency index θ. Recall that when $\theta = 0$, the hybrid scheduler reduces to a max-min scheduler. As the efficiency index increases, heavier use is made of the opportunistic scheduler. The efficiency of the scheduler can be judged by the mean throughput ratio, m_g. Efficiency increases monotonically with the efficiency index, from 0.55 to 2.0. Fairness of the scheduler can be judged by the variation of g among users; tightly clustered throughput ratios imply that users achieve throughput closely proportional to their target throughputs. When $\theta = 0$, the throughput achievement ratios for all users are tightly clustered around 0.55. Fairness falls monotonically with the efficiency index.

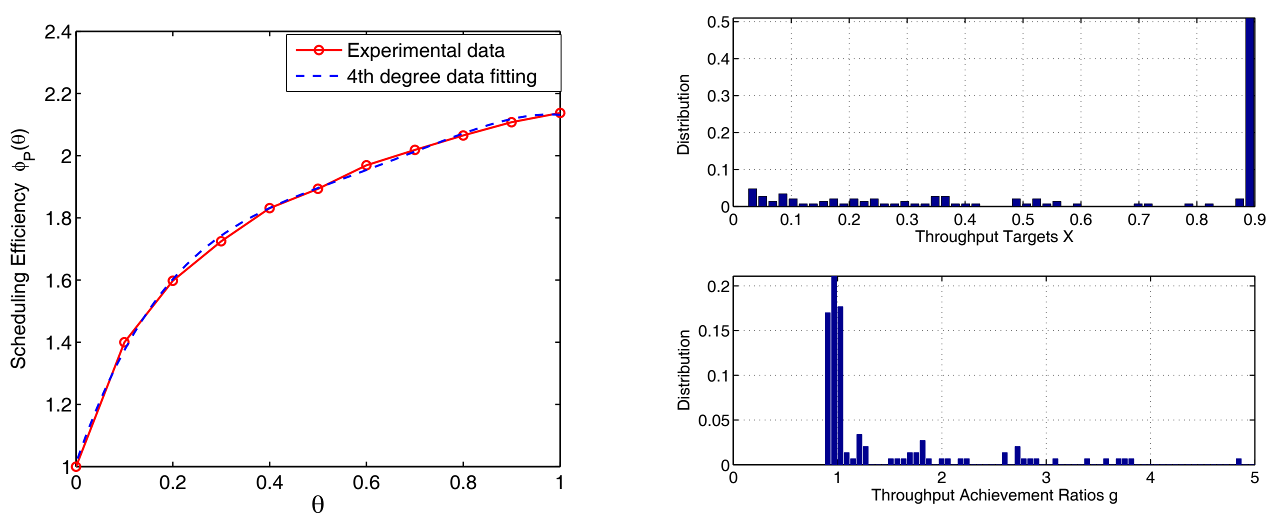
One might expect that the minimum throughput achievement ratio among all users, g_{min}, should be monotonically decreasing with the efficiency index. It is not; indeed the minimum throughput achievement ratio at first *increases* with the efficiency index on $0 < \theta < 0.4$ and then decreases on $0.4 < \theta < 1$. The minimum throughout achievement ratio is a measure that is a mixture of fairness and efficiency. When the efficiency index is low and increasing, an increase in the use of the opportunistic scheduler results in quickly rising average throughput ratios. This increment in the average dominates the corresponding increase in variance, and the minimum throughput ratio increases, i.e. the increment in efficiency affects the worst user more than the decrease in fairness. In contrast, when the efficiency index is high and increasing, an increment in the use of the opportunistic scheduler results in slowly rising average throughput ratios. Now this increment in the average is dominated by the corresponding increase in variance, and the minimum throughput ratio decreases.

A comparison with the proportional fairness rate scheduler is shown in Fig. 3(b); the proposed RS algorithm outperforms proportional fairness for a large region of $\theta > 0.1$ in terms of both mean throughput ratio m_g and minimum throughput ratio g_{min}. We also investigated a tunable RS algorithm that adaptively sets θ on the basis of the first two moments of g or the minimum and maximum of g. As shown in Fig. 3(c), in the plane of two key performance measures m_g and g_{min}, adaptive algorithms achieve the same performance tradeoff as the fixed θ algorithm, while a purely random θ performs worse.



(a) (b) (c)

Fig. 3. (a) Throughput achievement ratio, $g_{f,i}$, for each user; (b) Minimum and mean throughput achievement ratios, g_{min} and $E[g]$, versus efficiency index θ; (c) Fixed θ versus adaptive θ in terms of g_{min} and $E[g]$



(a) (b)

Fig. 4. (a) Estimation of scheduling efficiency in power $\phi_P(\theta)$; (b) Distribution of target throughputs and throughput achievement ratios

Our conclusion is that the proposed tunable RS algorithm, by combining opportunistic and max-min algorithms, addresses achievement ratios well. We expect that more complex opportunistic and max-min algorithms, as well as more complex hybrids (see e.g. [20]), could improve the ability of the RS to achieve targets, and increase its feasible region.

B. Performance of the Proposed CAC-RA and RS Architecture

In this subsection, the hierarchical two layer algorithm is evaluated with an emphasis on cross layer coordination. The goal is to explore whether the decomposition into two layers using an exchange of information about current and target achievement ratios is effective. The following parameters are used throughout this subsection: $\sigma_{\psi_{dB}}$ is uniformly distributed in [5,12], $\lambda = 10^{-2}$, and $\mu_i = 10^{-4}$. For the same of simplicity, all users have the same utility function $U_i(X_i, \hat{g}_{f,i}) = U_1(X_i)U_2(\hat{g}_{f,i})$ where $U_1(X_i) = c_1[(X_i + c_2)^{c_3} - c_2^{c_3}]$ and $U_2(\hat{g}_{f,i}) = c_4\hat{g}_{f,i}^3$ if $\hat{g}_{f,i} < 1$ and $U_2(\hat{g}_{f,i}) = c_5[(\hat{g}_{f,i} + c_6)^{c_7} + c_8]$ if $\hat{g}_{f,i} \geq 1$ and where the constants $c_i > 0\ \forall i$ and $c_3, c_7 < 1$. We judge the performance of the combined CAC-RA and RS on-line algorithms by blocking probability (defined as the ratio of the number of blocked users to the total number of arrivals), average aggregate utility U_{tot}, average target throughput m_X, minimum and average throughput achievement ratios g_{min} and m_g, and the percentage of maximum target throughput. All metrics other than blocking probability apply only to admitted users.

Recall that the CAC-RA algorithm requires estimation of the feasible region of target throughputs that the RS can achieve, that is, estimation of the power efficiency factor, $\phi_P(\Pi)(\theta)$, and the rate fluctuation factor, $\phi_R(\Pi)(\theta)$. As above, θ is chosen to be 0.4. To estimate $\phi_P(\Pi)(\theta)$, a simple empirical approach is applied: two users are associated with randomly generated channel statistics, $\{m_{h_i}, \sigma_{h_i}, \forall i = 1, 2\}$, and varying throughput targets, $\{X_i \in (0, 1), \forall i = 1, 2\}$. These two users are imported to the RS scheduler, where other users with complete information are present under a fixed power supply. In terms of the mean of the resulting achieved throughputs, a scheduler efficiency curve, $\phi_P(\theta)$, is fitted using a fourth degree polynomial, shown in Fig. 4(a). The choice of the rate fluctuation factor $\phi_R(\Pi)(\theta)$ is more ad-hoc; we observe that $g_{min} \approx 1$ when $\theta = 0.4$, and apply a backoff of 90% so that no single user saturates the system, resulting in $\phi_R(\Pi)(0.4) = 0.9$. Certainly, more sophisticated

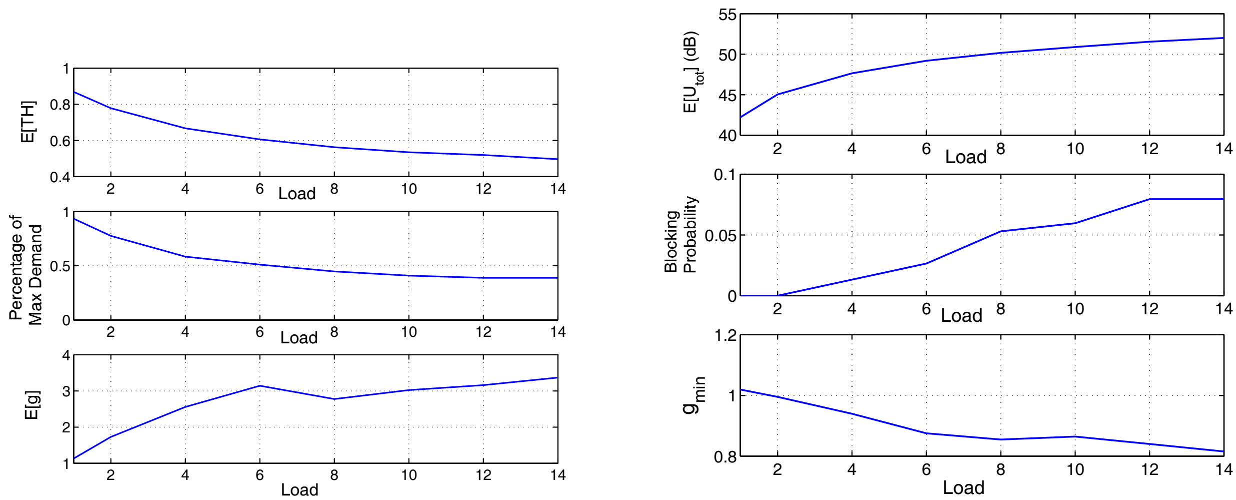


Fig. 5. Comparison between utility assisted optimization and no-utility assisted optimization.

algorithms such as Kalman filters could be developed in the CAC-RA layer to estimate $\phi_P(\theta)$ and $\phi_R(\theta)$.

We first examine the performance of the two layer structure under a fixed load. Given $\theta = 0.4$ and $load = 6$, Fig. 4(b) illustrates the distribution of the target throughputs $\{X_i, \forall i\}$ chosen by the algorithm, and the achieved throughput ratios. We observe that target throughputs have a distribution with heavy ends, especially around $X_{max} = 0.9$, and achievement ratios cluster around 1. The distribution of X_i can be explained as follows. From the global power constraint (5), we should expect an approximately inverse relationship between a user's target throughput and her mean channel condition. Indeed, in the CAC-RA layer, a user's gradient priority is inversely proportional to her mean channel condition. About 50% of users are assigned $X_i = 0.9$, corresponding mostly to users close to the base station.

Next, we examine the effect of variations in load, illustrated in Fig. 5. Recall that *load* is defined as the ratio of normalized power, $\frac{\lambda}{\mu}\frac{E[m_h]}{2}$, to the power supply $\mathcal{P}$. Hence, an increase in load can be associated with an increase in arrival rate, an increase in mean connection duration, or a decrease in the power supply. Here, we decrease the power supply to increase the load. As load increases, the blocking probability increases monotonically. This is an indication that the CAC-RA layer is tightening the admission control, blocking an increasing number of arrivals in an attempt to satisfy the target throughputs of those arrivals that have already been accepted into the system. Users who are admitted are assigned lower targets; the mean target throughput m_X decreases and the proportion who are assigned the maximum target throughput $X = 0.9$ also declines. The decrease in power, however, makes it more difficult to achieve the targets of all users. The distribution of the achieved throughput ratio widens, as evidenced by the simultaneous decrease in the minimum throughput ratio g_{min} and the increase in the average throughput ratio m_g. Nevertheless, the tunable algorithm acts effectively even under very high load, maintaining g_{min} above 0.8. Finally, the aggregate utility per unit power supply increases with load. This is an indication that the system is using the limited power supply more efficiently, allocating it to users that can achieve higher utilities. More complex MBAC algorithms (see e.g. [21]) may improve the performance.

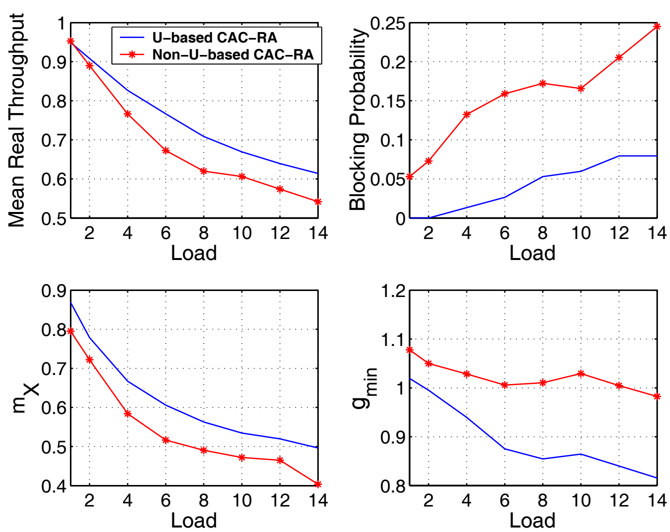


Fig. 6. Comparison between utility assisted optimization and no-utility assisted optimization.

Finally we compare the utility assisted CAC to a non-utility based variant, using the same rate scheduler. In the variant, all users are served in increasing order of $g_{f,i}$, and the remainder of the power is assigned to arrivals, if applicable. Given $\theta = 0.4$ and $load = 6$, as shown in Fig. 6, the variant results in a higher blocking probability, lower throughput demands, lower real throughput, and higher minimum throughput ratio. We conclude that the use of utility in CAC provides useful information that can result in a more intelligent trade off decision between efficiency and fairness.

VII. Conclusion

Our goal was to construct a framework for resource allocation for elastic wireless applications with performance goals stated in terms of the average throughput to be achieved over the connection duration. Our framework allows an application to express its satisfaction as a utility function that depends on both the performance goal and on the achieved throughput. With the goal of maximizing the long term average utility per unit time, we decomposed the problem into a connection access control and resource allocation layer and a rate scheduling layer. The connection access control layer considers its current commitments and decides whether to admit or block new sessions. When admitted, the layer sets a performance goal in terms of a target throughput to be achieved over the lifetime of the session. The rate scheduling layer adjusts the instantaneous rates of each connection on the basis of how well it is achieving each performance goal and on their channel state. We illustrated how commonly used connection access control and rate scheduling techniques can be applied to design these two layers using an exchange of information regarding future target throughputs and achieved throughputs. Through numerical analysis, we have shown that the proposed RS algorithm is tunable between efficiency (total throughput) and fairness (achieved ratios), and outperforms the well-known proportional fairness algorithm for a large range of the tuning parameter θ. We have also shown that the utility assisted CAC algorithm outperforms a variant that does not consider utility in terms of performance measures such as blocking probability and total throughput.

As the first attempt to integrate CAC and rate RS to support a class of traffic that evaluates performance on the basis of both target throughput and achieved average throughput, we do not view the architecture presented as ready for implementation. Although we have shown that standard opportunistic and max-min rate scheduling algorithms can be adapted to address the ratio of achieved to target throughput, we expect that more nuanced RS techniques and hybrids could improve the ability of RS to achieve targets and increase its feasible region. Although we have shown that a simple greedy MBAC algorithm can be adapted to control access on the basis of whether targets are being achieved, we expect that more nuanced MBAC techniques could improve the ability of the CAC-RA layer to balance accepted connections with achieving target throughputs. Finally, additional research to characterize the feasible region of common rate scheduling algorithms would be useful.

References

[1] M. Xiao, N. B. Shroff, and E. K. Chong, "A utility-based power-control scheme in wireless cellular systems," *IEEE/ACM Trans. Networking*, vol. 11, pp. 210–221, 2003.

[2] P. Liu, P. Zhang, S. Jordan, and M. Honig, "Single-cell forward link power allocation using pricing in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 533–543, 2004.

[3] C. Zhou, P. Zhang, M. Honig, and S. Jordan, "Two-cell power allocation for downlink CDMA," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 2256–2266, 2004.

[4] C. Zhou, M. Honig, and S. Jordan, "Utility-based power control for a two-cell CDMA data network," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2764–2776, 2005.

[5] X. Wang, "Wide-band TD-CDMA MAC with minimum-power allocation and rate- and BER-scheduling for wireless multimedia networks," *IEEE/ACM Trans. Networking*, vol. 12, pp. 103–116, 2004.

[6] T. Wong, J. Mark, and K.-C. Chua, "Joint connection level, packet level, and link layer resource allocation for variable bit rate multiclass services in sellular DS-CDMA networks with QoS constraints," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 1536–1545, 2003.

[7] T. Kwon and Y. Choi, "QoS provisioning inWireless/Mobile multimedia networks using an adaptive framework," *Wireless Networks*, vol. 9, pp. 51–59, 2003.

[8] R. Berry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Mag.*, vol. 21, pp. 59–68, 2004.

[9] M. Grossglauser and D. Tse, "A time-scale decomposition approach to measurement-based admission control," *IEEE/ACM Trans. Networking*, vol. 11, pp. 550–563, 2003.

[10] D. Bertsekas and R. Gallager, *Data Networks*. Prentice Hall, 1987.

[11] F. P. Kelly, "Charging and rate control for elastic traffic," *European Trans. Telecommun.*, vol. 8, pp. 33–37, 1997.

[12] J. M. Holtzman, "CDMA forward link waterfilling power control," in *Proc. 51st IEEE Vehicular Technology Conference*, 2000, pp. 1663–67.

[13] L. Song and N. B. Mandayam, "Hierarchical SIR and rate control on the forward link for CDMA data users under delay and error constraints," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1871–1882, 2001.

[14] X. Liu, "Opportunistic scheduling in wireless communication networks," Ph.D. dissertation, Purdue University, 2002.

[15] Z. Han and K. R. Liu, "Joint link quality and power management over wireless networks with fairness constraint and space-time diversity," *IEEE Trans. Veh. Technol.*, vol. 53, pp. 1138–1148, 2004.

[16] L. Tong and P. Ramanathan, "Adaptive power and rate allocation for service curve assurance in DS-CDMA network," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 555–564, 2004.

[17] J. Kim and A. Jamalipour, "Traffic management and QoS provisioning in future wireless IP networks," *IEEE Personal Commun.*, vol. 8, pp. 46–55, 2001.

[18] D. Goodman and N. Mandayam, "Power control for wireless data," *IEEE Personal Commun.*, vol. 7, pp. 48–54, 2000.

[19] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2053–2064, 2001.

[20] K.-K. Leung, C. W. Sung, and V. Bhargava, "An opportunistic power control algorithm with fairness," in *Proc. 2005 International Symposium on Information Theory*, 2005, pp. 690–694.

[21] W. Wang, X. Wang, and A. Nilsson, "Energy-efficient bandwidth allocation in wireless networks: algorithms, analysis, and simulations," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 1103–1114, 2006.