System Modeling and Performance Evaluation of Rate Allocation Schemes for Packet Data Services in Wideband CDMA Systems

Yu-Kwong Kwok, *Senior Member*, *IEEE*, and Vincent K.N. Lau, *Senior Member*, *IEEE*

Abstract—To fully exploit the potential of a wideband CDMA-based mobile Internet computing system, an efficient algorithm is needed for judiciously performing rate allocation, so as to orchestrate and allocate bandwidth for voice services and high data rate applications. However, in existing standards (e.g., cdma2000), only a first-come-first-served equal sharing allocation algorithm is used, potentially leading to a low bandwidth utilization and inadequate support of high data rate multimedia mobile applications (e.g., video/audio files swapping, multimedia messaging services, etc.). In this paper, we first analytically model the rate allocation problem that captures realistic system constraints such as downlink power limits and control, uplink interference effects, physical channel adaptation, and soft handoff. We then suggest six efficient rate allocation schemes that are designed based on different philosophies: rate optimal, fairness-based, and user-oriented. Simulations are performed to evaluate the effectiveness of the rate allocation schemes using realistic system parameters in our model.

Index Terms—Rate allocation algorithms, wideband CDMA, high data rate services, optimal, wireless data networks.

1 Introduction

The 21st century is, beyond doubt, the age of ubiquitous mobile computing with applications surrounding the Internet and business databases operations [1], [3], [7], [14], [15], [33], [35], [39], [44], as we have witnessed tremendous advancements in multiple synergistic developments in processor technologies, wireless communications, and protocol designs. Indeed, today's technologies can almost realize the scenario in which users, carrying powerful hand-held devices supported by a high-speed wireless network, perform computations anytime and anywhere. We still need to complete a couple of key steps before we can really enjoy such convenience of computing anywhere in a tetherless manner. In particular, as we can expect that a majority portion of the future mobile Internet computing systems will be supported by wideband CDMA (code division multiple access) networks [4], [9], [18], [46] (also known as "third generation" cellular networks, or 3G), there is one prominent key problem that we must tackle in order to fully deliver the potential of such a mobile computing system. Specifically, the challenge is how we can judiciously utilize the precious bandwidth of the 3G CDMA systems. In the existing standards (e.g., cdma2000), resources in the packet-based channel are allocated in a first-come-first-served manner [40], which is undeniably inefficient in the utilization of bandwidth. In order to support a large number of concurrent users demanding high data rates for their multimedia applications, an optimized rate allocation scheme is needed.

• *The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: {ykwok, knlau}@eee.hku.hk.*

Manuscript received 19 Sept. 2001; revised 31 July 2002; accepted 30 Oct. 2002.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number 117654.

The question of how to efficiently allocate the packet channels in a wideband CDMA system is still largely unanswered because of several technical difficulties. First, the spreading process in the physical layer limits the permissible data rates in limited wireless spectrum. Nevertheless, from an information theoretic point of view, the efficiency of utilizing the allocated spectrum could be increased in order to support packet data services and, in fact, this is the major motivation behind the wideband CDMA systems. Second, the law of large number does not hold for the relatively small number of packet data users. Thus, the intrinsic advantage of perfect statistical multiplexing in CDMA systems does not apply to high-speed packet data users. In other words, packet data transmissions from data users have to be *coordinated* carefully and, to achieve this goal, we need to devise an intelligent rate allocation scheme that works under realistic constraints such as considering both the downlink (from the base-station to the mobile device) and uplink (from the mobile device to the base-station), channel adaptation, as well as the soft handoff effects.

Liu et al. [31] performed the pioneering research of systematically deriving the channel code assignment and power allocation rules in a multicode DS-CDMA system. They suggested the Maximum Capacity Power Allocation (MCPA) rule and provided a number of useful insights in the overall system design and analysis. However, the code assignment part was incomplete in that only a set up capacity rules are given without specifying how to *order* the requests for applying the rules and the user Quality of Service (QoS) is ignored. Choi and Shin [5] extended Liu et al.'s model to handle two different classes of traffic. Again, using Liu et al.'s model, Kim and Sung [20] derived estimated capacity expressions for a multicode DS-CDMA system. Still, these research efforts did not address the rate allocation mechanism with QoS issues considered. Lu and Brodersen [32] proposed an integrated approach for power control and

scheduling with error correction coding support. However, they only addressed the downlink. Kim and Honig [19] analyzed the resource allocation problem for multiple classes of DS-CDMA traffic. To keep the analytical models tractable, only Gaussian noise was considered and fading effects were ignored. Joshi et al. [16] investigated the performance of various scheduling heuristics and obtained a number of useful insights. For example, they found that scheduling approaches that exploit request sizes can perform better. However, only downlink was considered. Kumar et al. [21], [22], [40] recently explored the problem of rate allocation in wideband CDMA systems and they provided a number of useful design guidelines based on practical system parameters. However, the allocation algorithm they used, called BALI, is only a simplistic equal sharing approach.

In Section 2, we first analytically model the rate allocation problem framework that captures downlink power limits, uplink interference conditions, and soft handoff. Furthermore, our model also incorporates a channel adaptive physical layer, which is widely envisioned to be a crucial component in modern wireless communication networks (e.g., in the 3G systems). Our model is generic enough so that practical system parameters, such as those of WCDMA or cdma2000, can be readily plugged in it to derive realistic allocation constraints. With the model, in Section 3, we suggest six rate allocation heuristics that can be classified into three categories: rate or utility optimal approaches, approaches that maintain fairness among users, and user-oriented heuristics. We discuss the features and rationales of these six schemes in detail. In Section 4, we present the results of our detailed simulation study, in which we generate performance data in terms of capacity, coverage, admission probability, outage probability, average delay, and average throughput. Concluding remarks are provided in the last section.

2 System Modeling

In this section, we first describe the general features of the wideband CDMA systems considered in our study. We then derive the generic power and interference constraints for rate allocation. Finally, we describe the use of a channel adaptive physical layer to enhance the bandwidth efficiency of the system.

2.1 Overview

In third generation wideband CDMA cellular networks, upon being admitted into the system, each user is assigned a *fundamental channel* (FCH) (e.g., as in the cdma2000 standard [9]), which can be used for power control, low bit rate services (e.g., voice or short message service) and, most importantly, making high data rate transmission requests. If a user's request is successfully granted, the user will get allocated a certain number of *supplemental channels* (SCH's) on which the high bandwidth data can be transmitted. The more SCHs a user gets, the higher the data rate he/she can transmit: Suppose the user is allocated m SCHs, then the data rate is $m \times R_{FCH}$, where R_{FCH} is the data rate of a FCH. Such channel aggregation can be implemented by code aggregation schemes [12], [31] and other related methods [40]. Thus, the rates that can be allocated are discrete values (e.g., 14.4 kbps, 28.8 kbps, 57.6 kbps, up to 1.843 Mbps).

We assume that the channel is synchronous and is divided into frames (e.g., in cdma2000, the frame duration is 20 msec). In each frame, there are a request minislot and an announcement minislot, and the remaining time in the frame is for data payload transmission. The user can submit a high data rate transmission request to the base-station in the request minislot and the base-station, as governed by the rate allocation policy, announces the allocation results in the announcement minislot in a prespecified later frame. Usually, there is a system-wide parameter called "burst duration," which determines how often a user can submit a request and, in turn, how much time the base-station has for computing the allocation results. Thus, the system is synchronized at the burst duration level also. We assume that a high data rate user always has enough amount of data to send in order to completely utilize the burst duration even at the highest possible data rate. This can be easily implemented by imposing a policy in configuring the mobile devices such that the devices will always accumulate enough data before making a request (the amount of data to be accumulated is indeed not very large: 5 kbytes if the highest rate is 2 Mbps and the duration is 20 msec).

2.2 Power and Interference Constraints

A rate allocation policy has to work under two system-wide constraints: 1) the downlink link power budget (in dB), and 2) the uplink interference limit. Now, consider that there are n high data rate transmission requests. In cell k, suppose the power required to support a single SCH of request j is τ_{jk}, which can be measured at the mobile device and its value depends on the current channel condition of the user. If the current downlink power level is Γ_k and the power budget is Γ_{max}, then we have the following constraint:

$$\Gamma_k + \sum_{j=1}^{n} m_j \tau_{jk} \leq \Gamma_{max}, \tag{1}$$

where m_j denotes the number of SCHs to be allocated to user j.

For the uplink, we need to consider two different cases: a cell k in soft handoff with the mobile device j (e.g., the host cell) and a cell k' not in soft handoff with the device j (e.g., as depicted in Fig. 1, the data user is in soft handoff with cells 0, 1, and 2, which constitute the *active set*, but not with cell 3). In the former case, the extra interference λ_{jk} introduced by device j is given by:

$$\lambda_{jk} = m_j \mu_{jk}, \tag{2}$$

where μ_{jk} denotes the uplink received power at the base-station k for one SCH, which depends on the channel condition (controls the processing gain required; detailed later), current total received power Λ_k, and the received bit energy to interference ratio. Here, for simplicity, we use m_j again, meaning that the user request is for both uplink and downlink, and is symmetric. That is, the same number of SCH's will be allocated to the user for both the uplink and downlink. Of course, this is not necessarily true in practice and such an assumption is not needed for our model to be useful. However, the physical implication of this assumption deserves a bit more elaboration here. While previously, people may consider that high data rate services are asymmetric (i.e., downlink is much more demanding than the uplink, just like the Web surfing situation), we believe that this assumption may not be appropriate for wireless Internet devices because a quickly emerging hot application, called Multimedia Messaging Services (an extention of

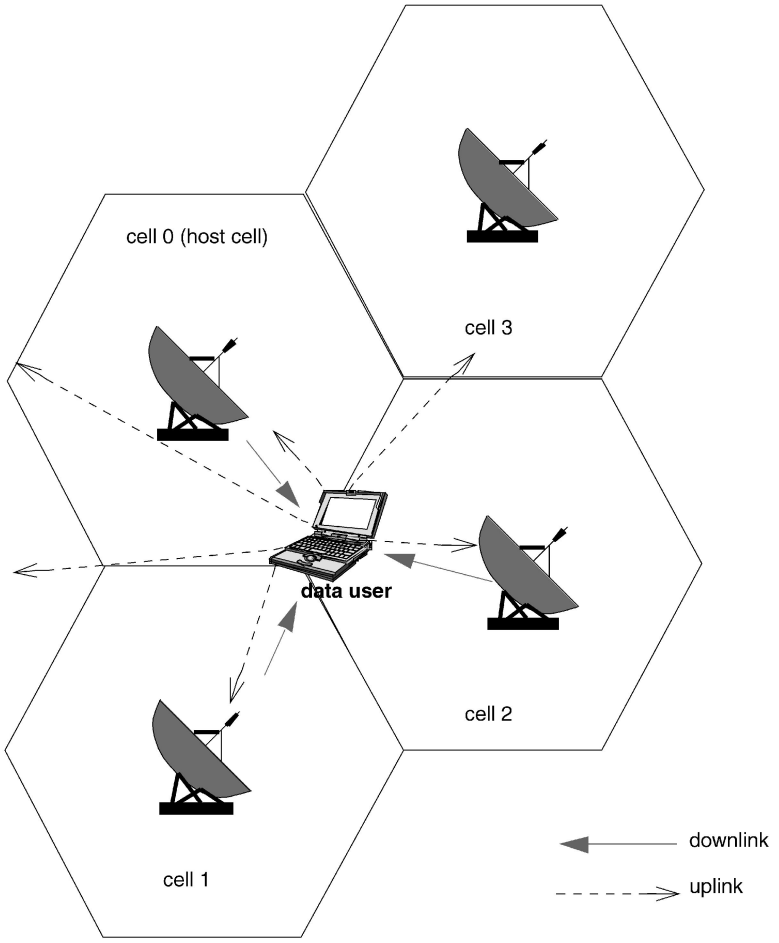


Fig. 1. The scenario in which the soft handoff active set of a data user consists of cells 0, 1, and 2, but not cell 3.

the very successful Short Message Service) [38], requires the user mobile device to be capable of transmitting high bandwidth data (e.g., a video clip) to the peer devices via the base-station (hence, high data rate is needed on the uplink as well).

Note that it is infeasible to obtain measurements of the received bit energy to interference ratio and, thus, we need to make use of the uplink pilot strength p_{jk}^{UL}, and the transmit power ratio of the data channel and the pilot channel, denoted by ϕ_j, of request j. We have:

$$\mu_{jk} = \phi_j p_{jk}^{UL} \Lambda_k. \tag{3}$$

Thus, for a cell k in soft handoff with user j, the extra interference introduced is:

$$\lambda_{jk} = m_j \phi_j p_{jk}^{UL} \Lambda_k. \tag{4}$$

For a cell k' not in soft handoff, the major difference is that it cannot obtain the uplink pilot strength p_{jk}^{UL} also. However, it is practical to accurately estimate its value using the downlink pilot strength using the relative path loss (given by the ratio of the downlink pilot strength of the cell k' and a soft handoff cell k) as follows:

$$p_{jk'}^{UL} = p_{jk}^{UL} \frac{p_{jk'}^{DL}}{p_{jk}^{DL}}. \tag{5}$$

Thus, the extra interference $\mu_{jk'}$ is given by:

$$\lambda_{jk'} = m_j \phi_j p_{jk}^{UL} \frac{p_{jk'}^{DL}}{p_{jk}^{DL}} \Lambda_{k'}. \tag{6}$$

In summary, we have the following constraint for the uplink:

$$\Lambda_k + \sum_{j=1}^{n} \lambda_{jk} \leq \Lambda_{max}, \tag{7}$$

where λ_{jk} is computed by using (4) if k is in soft handoff with user j, and by using (6) for other cells.

2.3 Channel Adaptation

The key quantities, τ_{jk} (power required for a downlink SCH) and μ_{jk} (interference of an uplink SCH), are critically affected by the channel condition. Indeed, by using the channel state information (CSI), which is the signal-to-interference ratio (SIR) as measured by the mobile device and reported to the base-station (i.e., the feedback CSI as indicated in Fig. 2), τ_{jk} and μ_{jk} can be carefully selected to maximize the achievable data rate for a certain specified bit error rate (BER) (in other words, the same data rate can be achieved by using possibly smaller values of τ_{jk} and μ_{jk}). This channel adaptation process can be implemented by incorporating a variable throughput adaptive coding and modulation physical layer in the system (both at the base-station and at the mobile devices) [24], [29], [30]. For

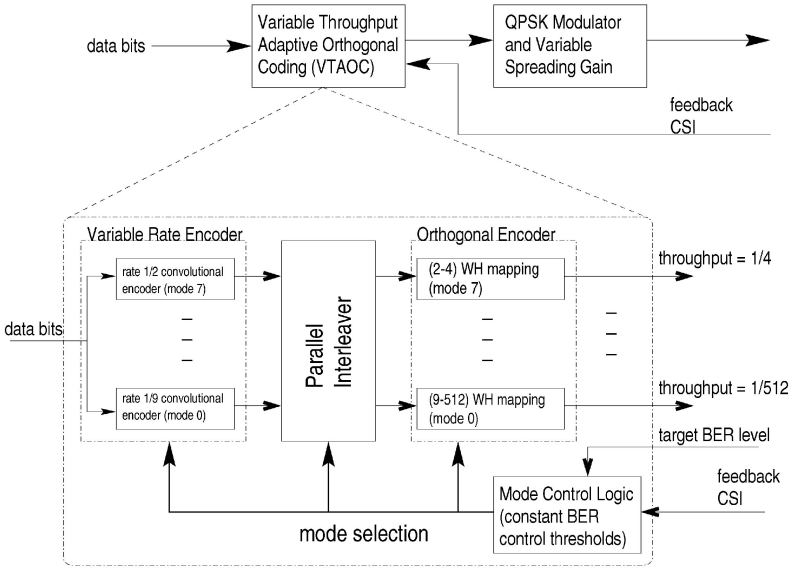


Fig. 2. Block diagram of the variable throughput adaptive physical layer.

example, the VTAOC scheme [27], [28], as shown in Fig. 2, can be used.

In our performance study, we use an 8-mode (symbol-by-symbol) VTAOC module (note that using the VTAOC scheme [28] is just for illustration only and certainly other similar schemes, such as those proposed in [10], [49], can be used in our framework). The available instantaneous physical layer throughput, which is defined as the number of information bits carried per modulation symbol, ranges from $1/2^2$ to $1/2^9$ depending on which path of convolutional encoding and Walsh-Haadamard mapping the data bit stream is to traverse, as governed by the CSI. Specifically, $\text{transmission mode-}q$ is chosen from the current information bit if the CSI falls within the *adaptation thresholds*, (ζ_{q-1}, ζ_q). The operation and the performance of the VTAOC scheme is determined by the set of adaptation thresholds $\{\zeta_0, \zeta_1, \ldots, \zeta_M\}$. In this paper, it is assumed that the VTAOC scheme is operated in the constant BER mode [28], in which the adaptation thresholds are set optimally to maintain a target transmission error level over a range of CSI values as illustrated in Fig. 3a. When the channel condition is good, a higher transmission mode could be used and the system enjoys a higher physical layer throughput at a given power level. On the other hand, when the channel condition is bad, a lower mode is used to maintain the target error level at the expense of a lower transmission throughput. Essentially, with reference to the BER curves (against CSI) and the adaptation thresholds (as shown in Fig. 3b), the values of τ_{jk} and μ_{jk} can be proportionately scaled down or up to allocate a certain number of SCH's to a user.

3 Rate Allocation Schemes

In our performance evaluation study, we consider six different approaches: throughput optimal approach, near optimal utility-based approach, channel adaptive fair heuristic, proportional fair heuristic, smallest backlog first, and longest delay first. We describe them in detail in the following sections.

3.1 Optimal and Near Optimal Approaches

We consider a throughput optimal approach in which we use the following objective function of maximizing the aggregate rate:

$$J_T(\vec{m}) = \sum_{j=1}^{n} m_j. \tag{8}$$

Thus, using $J_T(\vec{m})$, together with the power and interference constraints (i.e., (1) and (7)), constitute a linear programming problem. Using a linear program solver (in our study, we use AMPL/CPLEX [8], [2]), optimal rate vector can be computed such that the system utilization (or bandwidth efficiency) is maximized. However, this approach requires exponential time in the worst case (time complexity of $O(n^3 n!)$ [41]) and, thus, may not be practicable. Furthermore, as we will illustrate in Section 4, this approach may not necessarily give the best quality of service (QoS) to the users.

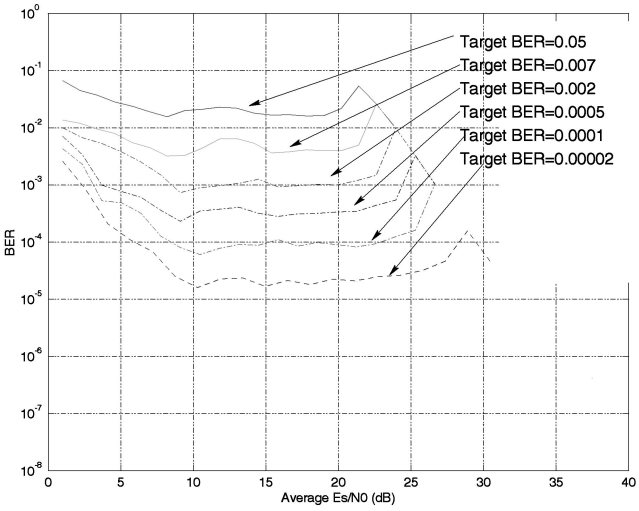
Because maximizing the total throughput may not be always desirable from the user's cost point of view, we also consider an approach that optimizes the aggregate utility:

$$J_U(\vec{m}) = \sum_{j=1}^{n} w_j \log(m_j + 1). \tag{9}$$

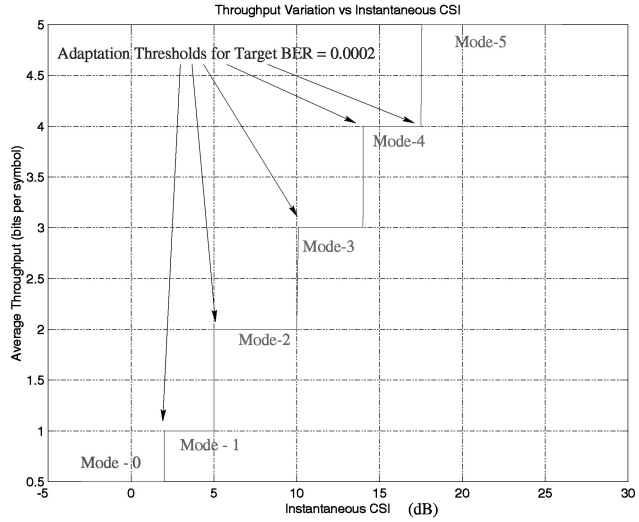
Here, utility is captured by the weighted sum of the logarithmic functions of the allocated number of SCH's. Note that the term 1 inside the logarithmic function is to ensure that a user j receives zero utility only when its allocated value of m_j is zero. The logarithmic function is used because it is a concave function and, thus, can model the diminishing return of the rate allocations.

Unfortunately, because the objective function $J_U(\vec{m})$ is nonlinear,[1] we cannot obtain optimal solutions efficiently using a linear program solver. Thus, we resort to a near optimal approach of using a genetic algorithm. Specifically, we encode each m_j as a 4-bit number (i.e., the maximum value

1. In a preliminary study [25], we also considered another nonlinear objective function.



(a)



(b)

Fig. 3. BER and throughput of the VTAOC scheme. (a) Instantaneous BER and the adaptation range. (b) BER and throughput of the VTAOC scheme.

of m_j is only 16, but this is legitimate in practice because, for a reasonable number of background load in a cell, the number of SCHs a user can obtain is less than 10; this will be evident in Section 4) and then concatenate all the m_js as a single binary string (again, in practice, the number of high data rate requests is less than 10 and, thus, the string will be around 40 bits long). Using such encoding, a population of randomly generated strings, each of which representing a different allocation, are used (in our study, the population size, denoted by N_p is 2^{20}, using about 5 Mbytes of memory). As a binary encoding is used, we can apply the two genetic operators, crossover and mutation, with their usual definitions: For crossover, two strings are randomly selected from the population, a random position (the crossover point) within the bit string is chosen, and then the two strings exchange their parts at the crossover point; for mutation, a string is randomly selected and a randomly selected bit is flipped. The invocations of crossover and mutation are governed by the crossover and mutation rates, respectively. With reference to the genetic algorithm literature [6], [23], [26], we set crossover probability as 0.1 and mutation probability as 0.02. The genetic algorithm is iterative in nature and each iteration is called a generation (in our study, the number of generations is 1,000). In each generation, the crossover and mutation operators are applied N_p times. At the end of a generation, the *fitness* value of each string is computed. We define fitness as the value of $J_U(\vec{m})$ and an invalid string (i.e., one that violates the power and interference constraints) is assigned a fitness of zero. Using the fitness values, a selection process can be applied in that after eliminating all the zero fitness strings from the population, new strings are generated by duplicating the remaining strings in a proportionate manner: The fittest string gets the largest share of duplication, and so on. A new generation can begin using the new population.

The time complexity of the genetic algorithm is $O(nN_p)$ because each evaluation of the fitness requires $O(n)$ time. With carefully crafted implementation [6], the genetic algorithm can be quite efficient in practice. The merit of a genetic approach is that it can generate near optimal solutions [23], [26].

3.2 Approaches that Maintain Fairness

The fairness concept has its roots in the design of the packet service disciplines at a router in which different sessions of packets are contending for a single outgoing link [48]. Assuming that the packets are infinitesimally divisible (i.e., the fluid model), a fair policy can be implemented by using the generalized processor sharing (GPS) approach [42]: In a round-robin manner, each packet session gets an infinitesimally small share of transmission time on the outgoing link. The shares may be weighted by the sessions' requested rates, which are determined when they are admitted in the system. However, because, in practice, packets are not infinitesimally divisible and are of variable sizes, the GPS approach cannot be used in actual implementation. Thus, a variety of different approaches are devised to approximate the behavior of the GPS approach [48]. These approaches differ in the time complexity and the analytical performance in terms of delay [48].

For wireless networks, there are also a number of recent attempts in designing fair allocation policies (for a brief survey, see [47]). The major difference between the rate allocation problem for wireline networks and that for wireless networks is that, in a wireless network, the channel quality is time-varying and location dependent. Thus, it is possible that the channel quality of a user can be so poor that he/she cannot successfully transmit data. Thus, the wireline policies cannot be directly applied in a wireless setting. The general approach of tackling the problem is that a wireline policy is used as a reference system in which the channel is assumed to be error free. The allocation for each user is then computed in this ideal system. In the real system, for each user we try to allocate the computed rate share to the user as far as possible subject to the constraint that the user's channel condition can support such an allocation. If a user's channel condition happens to be poor for one or more rounds of allocations, the user will then be a *lagging* because he/she gets a smaller amount of rate share as compared to the one computed in the reference system. On the other hand, the "surplus" rate allocations (because the users with poor channel quality cannot transmit) are shared proportionately by the users with better channel

conditions. Thus, these users get more rate shares as compared to those computed in the reference system, and are called *leading* users. The existing algorithms for fair wireless rate allocation are mostly based on this general principle [47]. However, because these algorithms are designed using a very simplified channel model—a 2-state model (either good or bad), they are not suitable for use in our model, in which the channel is accurately simulated taking into account fast fading and shadowing effects.

In our study, we devise a channel adaptive fair rate allocation policy using a priority metric Q_i that is simple to implement:

$$Q_i = \gamma_i e^{-\beta \Delta_i}, \tag{10}$$

where γ_i is the SIR (i.e., the channel state) of user i, Δ_i is the leading amount (in terms of actual number of information bits transmitted), and β is a scaling factor for balancing the effects of γ_i and Δ_i. We use the STFQ (start-time fair queueing) [48], which is an efficient variant of the GPS approach, as the reference allocation algorithm. Thus, the allocation policy works by using the following "filling" procedure:

1. Sort the requests in descending order of Q_i.
2. Allocate one SCH to the first request and check the constraints. If the constraints are satisfied, repeat this step with the next request; otherwise, undo the allocation and stop.
3. Update the Q_i of all requests. Go back to Step (1).

The time complexity of the above channel adaptive fair policy is $O(\mathcal{M} n \log n)$, where $\mathcal{M}$ is the largest possible value of m_j. As mentioned before, in practice, the value of $\mathcal{M}$ is quite small (less than 10) and, thus, the channel adaptive fair policy is quite fast.

We also consider another fair allocation heuristic called proportional fair algorithm [13], [34], [36]. The notion of proportional fairness, introduced by Kelly [17], is formally defined as follows: For a rate allocation vector $\vec{x}$, the allocation is proportional fair if:

$$\sum_{i=1}^{n} \frac{(x_i^* - x_i)}{x_i} \leq 0, \tag{11}$$

where $\vec{x^*}$ is any arbitrary allocation vector. In wireless networks, it has been shown [13] that proportional fair allocation can be approximately achieved by using the following metric S_i:

$$S_i = \frac{r_i}{R_i(T)}, \tag{12}$$

where r_i is the current maximum achievable rate (depending on the channel condition) and $R_i(T)$ is the average rate achieved in the past time window T. Thus, we can use a similar procedure as in the channel adaptive fair approach described above by replacing Q_i by S_i.

3.3 User-Oriented Heuristics

The schemes described in the previous sections are based on a more global view of the system: maximizing system throughput, maximizing aggregate utility, and maintaining system-wide user fairness. We also consider two heuristics that are more user-oriented. The first one is called Longest Queue First (LQF), which gives a higher priority to a user that has a larger backlog of packets to be sent. This heuristic is useful for mobile devices with only a limited amount of buffer space. Indeed, variants of LQF are found to be highly effective in a recent study [16]. The second one is called Largest Delay First (LDF), which gives a higher priority to a user that has its head of line packets delayed by the longest amount of time. However, it is important to note that unlike the four schemes described earlier, these user-oriented heuristics do not make use of the channel adaptation mechanism in that users are not selected based on their relative channel conditions. Thus, despite the fact that a channel adaptive physical layer (e.g., the VTAOC scheme) is still incorporated in the model, there is no synergy between the rate allocation policy and the physical layer.

4 Performance Results

Under our system model, the six rate allocation schemes described in Section 3 are evaluated by simulations. We use a 7-cell environment: a center hexagonal cell with six surrounding cells. We focus on the center cell and model the surrounding cells as background load. However, when a user is situated within the soft handoff region (which is assumed to be at a distance of 0.9 $\times$ cell radius from the base-station), three neighboring cells (e.g., as shown in Fig. 1) will participate in soft handoff. In the simulation of the physical layer, we employ a path loss exponent of 4 (i.e., signals are attenuated as d^{-4}, where d is the distance of propagation). We model the Rayleigh fast fading and log-normal shadowing (with variance of 8 dB) environments [37], [43], [45]. Power control is incorporated also. The other physical layer parameters are listed in Table 1.

We use a simple model for the mobility of the users. Each user (voice or data) selects a random starting position in the center cell, which is uniformly distributed over the cell. The direction of motion is also randomly selected. The motion is rectilinear and a random inward direction is selected again if the user hits the boundary of the center cell. This is done so to maintain the same number of users in the system throughput the whole simulation period. The speed of motion is assumed to be constant (at 35 miles/hr). A voice source is assumed to be continuously toggling between talkspurt and silence states. The duration of a talkspurt and a silence period are assumed to be exponentially distributed with mean 1.0 and 1.35 seconds, respectively (chosen according to the empirical study in [11]). We assume a talkspurt and a silence period start only at a frame boundary. For high data rate source, we model it as generating large data files (e.g., video clip files for multimedia messaging services). The arrival time of the file data generated by a mobile device is assumed to be exponentially distributed with mean equal to 1 second. The data size is also assumed to be exponentially distributed with mean equal to 10 kbytes. Again, we assume that the data packets are generated at a frame boundary. Each test case (with certain combination of number of voice users and number of data users) is run for 1,000 seconds of simulation time and is repeated 10 times with different random number seeds to obtain the average results.

We consider two aspects of performance: system wide and user QoS. For the system-wide aspect, we measure the capacity, coverage, data request admission probability, and voice outage probability. For the user QoS aspect, we measure the average delay and average throughput. We describe these performance metrics in more detail below.

TABLE 1
Physical Layer Parameters

Parameter	Value
Path loss exponent	4
log-normal shadowing variance	8 dB
Channel bandwidth	5 MHz
Cell radius	800 m
Chip rate	7.7328 Mcps
Modulation	QPSK modulation with quadrature spreading
FCH rate	14.4 kbps
SCH rate	[14.4 kbps–1.8432 Mbps]
processing gain (FCH)	512
processing gain (SCH)	[512-4]
BER (FCH)	10^{-2}
BER (SCH)	10^{-4}
E_s/I_0 (FCH)	7dB
E_s/I_0 (SCH)	13dB
Frame duration	20 msec
Burst duration	5 frames
Average adjacent cell load	50%, 75%

- We quantify the capacity of the system as the total number of SCH's available. This is an important performance parameter for the downlink. Obviously, as we have analyzed in Section 2, the capacity is limited by the power and interference constraints. Because of the power level required and interference generated by a user critically depends on his/her distance from the base-station (e.g., path loss), the capacity of the system highly depends on the geographical distribution of the users.
- Coverage is a performance metric closely related to capacity, but it indicates performance from another angle. Specifically, coverage can be defined as the fraction of cell area that a data user can be served. Again, this is also an important performance parameter for the downlink. Obviously, the coverage area depends on how many SCHs are to be allocated to the data user. For example, suppose that a data user is to be allocated only one SCH, then he/she may successfully get allocated even if he/she is situated at the boundary of the cell. On the other hand, if the user wants to get a large number of SCH's, then that may be successful only if the user is situated near to the base-station.
- Data request admission probability is defined as the fraction of cases where a data request successfully gets allocated a positive number of SCH's (note that m_j may end up to be zero in the throughput optimal approach). This metric is an important performance parameter for the uplink. It should be noted that this metric also reflects the level of QoS the system can offer to a user in the sense that how often a user's data "call" can successfully be made.
- Voice outage probability is an important parameter in that admitting the requests of data users should not disturb the QoS of the existing voice users (in other words, voice service is treated as the default service that should be always available). Again, this is an important performance parameter for the uplink. The outage probability is computed by calculating the fraction of frames that a voice packet cannot be successfully transmitted because the received E_s/I_0 falls below the threshold (i.e., 7 dB). Due to the isochronous nature of voice, such a packet is useless (i.e., need not be retransmitted) and the quality of the voice service has been disrupted.
- Each data packet is time stamped when it is generated. Thus, the delay experienced by a packet can be calculated by subtracting this time stamp from the receipt time of the packet at the base-station. Average delay is obtained by taking the mean over all delay values within the whole duration of a test case. Average throughput is computed in a similar fashion.

Fig. 4 shows the capacity results for adjacent cell load of 50 percent and 75 percent, which represent moderate background load and heavy background load, respectively. These results are obtained using five data users and 25 voice users. In the figure, we use the following abbreviations for the six rate allocation schemes: ROPT (throughput optimal using a linear program solver), UTIL (utility optimal using a genetic algorithm), CAF (channel adaptive fair heuristic), PF (proportional fair heuristic), LQF (longest queue first heuristic), and LDF (largest delay first heuristic). As can be seen, the capacity of the system drops significantly when the data users are midway between the base-station and the cell boundary. This is because the power level requirements increase significantly as the user gets farther away from the base-station. At the cell boundary, the situation is even worse, not only because of the distance effect, but also soft handoff load. Indeed, at the cell boundary, we find that there is at most one high data rate user that can be served (should he/she be selected) because the capacity is just around 10 SCHs. The performance ranking of the six schemes is as follows: ROPT, UTIL, CAF, PF, LQF, and LDF. In fact, the performance of LQF and LDF are very similar. ROPT clearly excels in the capacity aspect because ROPT inherently optimizes the power and interference "usage" to allocate rates to users. UTIL optimizes the logarithm of rate and, thus, is slightly inferior to ROPT. CAF and PF

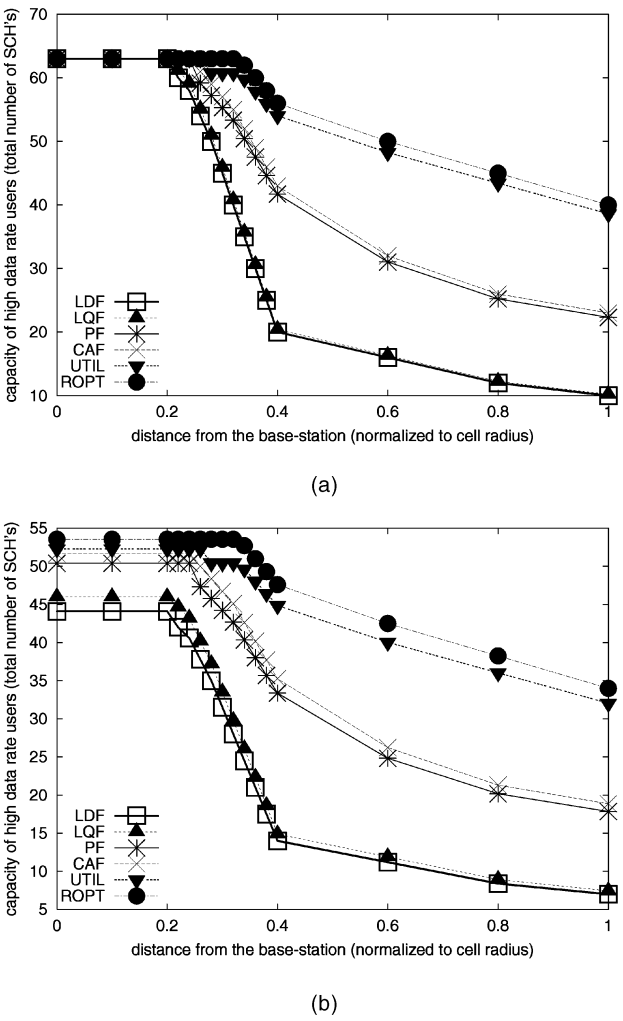


Fig. 4. Number of SCHs available on the downlink with varying distances from the base-station. (a) Adjacent cell load = 50 percent. (b) Adjacent cell load = 75 percent.

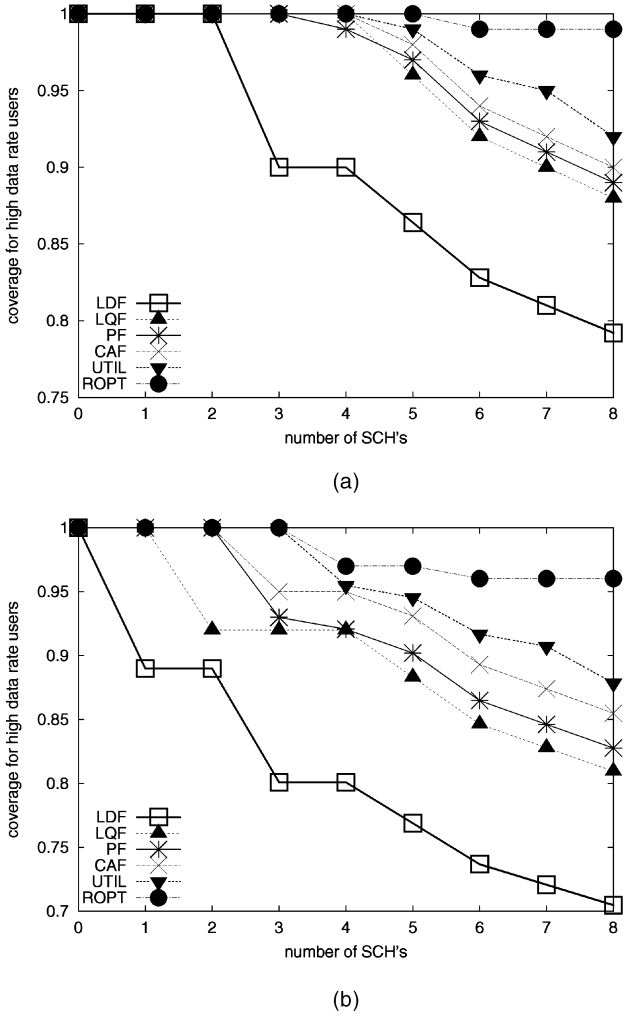


Fig. 5. Downlink link coverage area as a function of the number of SCH's allocated. (a) Adjacent cell load = 50 percent. (b) Adjacent cell load = 75 percent.

performs similarly. With a heavy background load (i.e., 75 percent adjacent cell load), the performance of the system becomes worse remarkably. This is because the higher load in the surrounding cells introduces a much higher interference to the users at midcell area as well as the cell boundary.

Fig. 5 shows the coverage results (again with five data users and 25 voice users, and for adjacent cell load of 50 percent and 75 percent). We can see that the coverage of ROPT is very much higher than all the other schemes, including even UTIL. This is because, except ROPT, all schemes do not pay particular attention to maximizing the number of SCH's that can be allocated with a high efficiency of utilizing the power budget. UTIL is a distant second and the other four schemes are much worse. A scrutiny of the simulation trace reveals that all schemes, except ROPT, need to cater for one or more user-oriented constraints: utility (diminishing return), fairness, and delay. Thus, the power budget is almost always not utilized at the maximum possible extent. When the background load is higher (i.e., 75 percent of adjacent cell load), the performance of LQF and LDF are remarkably worse. This can be explicated by the observation that these schemes make allocation "mistakes" more frequently (i.e., allocating SCHs to a user with a poor channel condition or is far away from the base-station) under a tight power budget.

Fig. 6 shows the uplink admission probability results (five data users and 25 voice users; adjacent cell load: 50 percent and 75 percent). We can see that similar to the situation of the downlink capacity, the admission probability drops significantly when the data users are midway between the base-station and the cell boundary. It is interesting to note that, except for ROPT and UTIL, the admission probabilities of all schemes are not close to one even when the data users are very near the base-station. This is because in CAF, PF, LQF, and LDF, a data user with a very good position and (hence) very good channel condition does not necessarily get allocated. For example, if it happens that a severely lagging user encounters a good channel state, an even better user (in terms of channel state) may not get allocated any SCH because the latter may be leading by a large margin. The results for a heavy background load environment show similar trends.

Fig. 7 shows the uplink outage probabilities of the voice users. These results are obtained by using five data users and varying number of voice users (from 15 to 33). We can

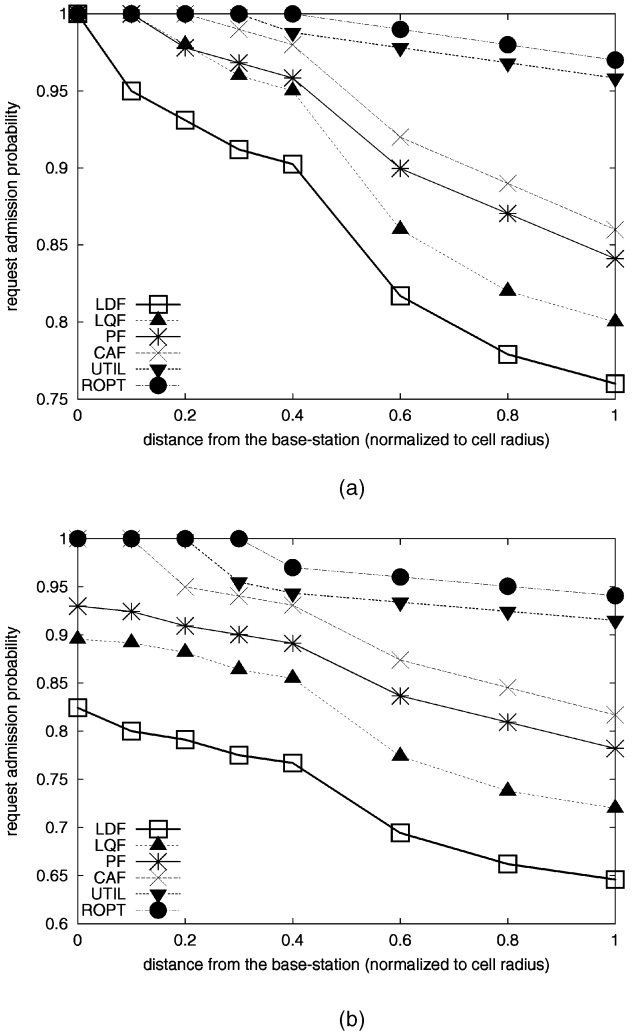


Fig. 6. Data request admission probabilities on the uplink. (a) Adjacent cell load = 50 percent. (b) Adjacent cell load = 75 percent.

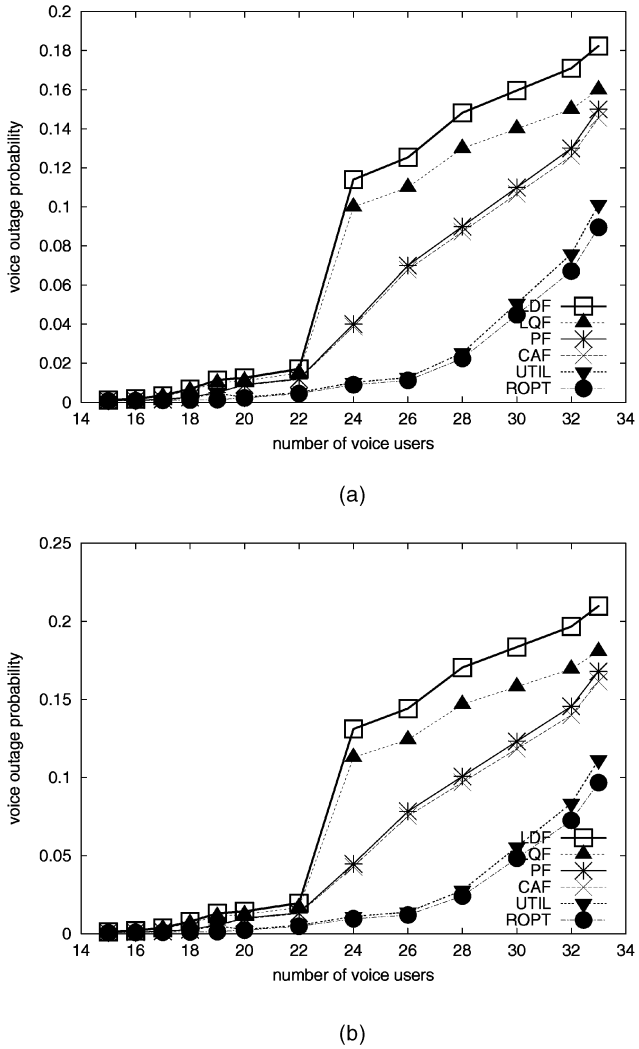


Fig. 7. Voice outage probabilities on the uplink. (a) Adjacent cell load = 50 percent. (b) Voice outage probabilities on the uplink.

see that even for ROPT, the voice user's QoS becomes unacceptable (cannot get service in 5 percent of the time) when there are about 30 voice users in the system. For other rate allocation schemes, the situation is even worse. The reason for this phenomenon is that the interference introduced by the data users and, more importantly, by the peer voice users, are quite high and the system capacity is reached most of the time. Thus, a stable operating point of our simulated system should be around 20 voice users with four to six data users.

Now, let us consider the user QoS aspect of the performance, in terms of average data delay and average data throughput. We use 20 voice users and vary the number of data users from three to seven. Moreover, we use a heavy background load of 75 percent. These results are shown in Fig. 8. As expected, the average throughput of ROPT is much more higher than the other schemes. UTIL is obviously worse than ROPT but not by a large margin. The performance of CAF and PF are very similar. Finally, the performance of LQF and LDF are also very close. These results by and large concur with the observations we have seen from the system-wide results (i.e., coverage, capacity, etc.). However, it is interesting to note that for the average delay, the ranking of the schemes is quite different: CAF and PF are among the best, followed by LDF, UTIL, ROPT and, finally, LQF. This phenomenon was intriguing and, thus, we looked at the simulation traces very carefully. We find that the ROPT and UTIL schemes, only focus on rate as they are so designed, often times generate a rate allocation vector $\vec{m}$ with many zeros: Effectively, some data users do not get allocation for several burst durations. A large delay thus results for such users. On the other hand, while CAF and PF do not explicitly cater for the delay metric, the action of trying to balance the service shares among the users has the effect of controlling the delay also. Of course, for many cases, the absolute values of the delays are not as low as those best cases achieved by LDF. However, the LDF scheme usually cannot allocate a high rate for the data users and, thus, the gain in choosing largest delay users is frequently offset by the loss in rate. Thus, LDF does not perform well overall.

5 Concluding Remarks

In this paper, we have provided an analytical model of the rate allocation problem for wideband CDMA systems.

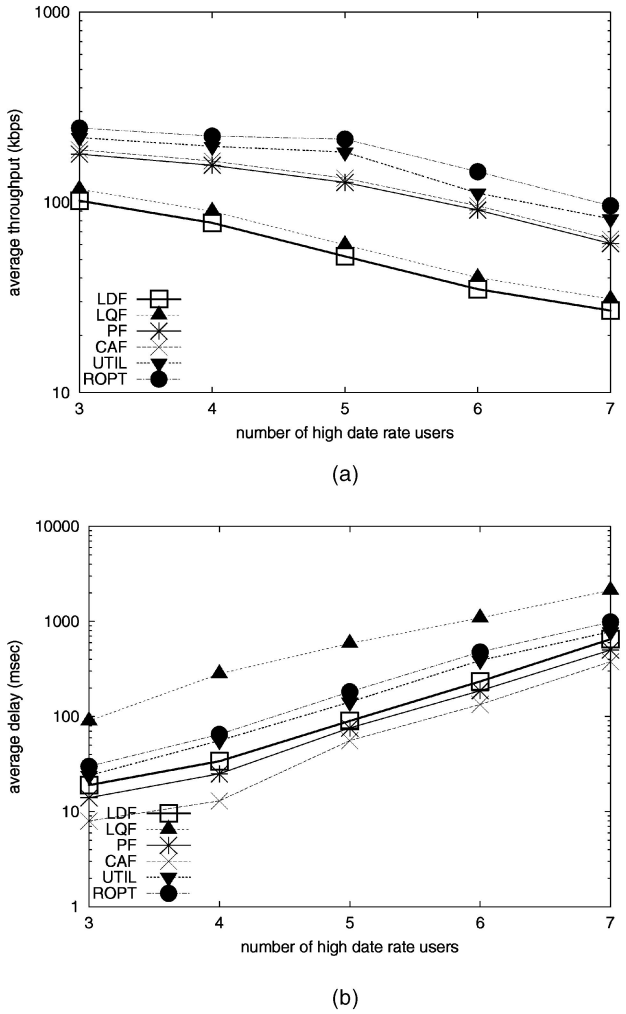


Fig. 8. Average delay and throughput of data requests with 20 voice users and adjacent cell load of 75 percent. (a) Average throughput. (b) Average delay.

Using the model, we suggest six different rate allocation schemes. These schemes are designed based on different philosophies: rate or utility optimal, fairness-based, and user-oriented. The design rationales and distinctive features of the six algorithms are also discussed in detail. In our simulation study, the rate and utility optimal approaches are found to exhibit the best performance in maximizing the overall system throughput. However, they may not be always able to provide a good quality of service to the users, in particular, in terms of delay. Furthermore, the time complexity of these approaches can also be prohibitively high for practical use. The fairness-based approaches seem to be a good compromise and are good candidates for practical implementations in a real system.

We would like to point out one interesting avenue of further research. We believe that it would be extremely useful if we could devise insightful utility functions for the mobile devices and the base-station (may be different) and then solve the optimization problem involving the system constraints we have presented in this paper. A distributed algorithm may then result from the solution of the optimization problem such that the mobile devices and the base-station can independently optimize their respective utility functions, and yet achieve a globally optimal aggregate utility. A successful instance of this technique has been demonstrated in [17]. However, the problem we mentioned here would be much more difficult because of the many degrees of freedom involved (imperfect power control, multiple access interference, heterogeneous SIR required by different applications, channel adaptive coding and modulation, etc.).

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful and constructive comments on this paper. This research was jointly supported by a HKU URC research grant under contract number 10203413, and by a grant from the Hong Kong Research Grants Council under contract number HKU 7024/00E. Preliminary results of this research were presented at the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM) 2001, Rome, Italy, July 2001.

REFERENCES

[1] G. Alonso, R. Gunthor, M. Kamath, D. Agrawal, A.E. Abbadi, and C. Mohan, "Exotica/FMDC: A Workflow Management System for Mobile and Disconnected Clients," *Distributed and Parallel Databases,* vol. 4, pp. 229-247, 1996.

[2] AMPL Web Site, http://www.ampl.com/, 2002.

[3] D. Barbara, "Mobile Computing and Databases—A Survey," *IEEE Trans. Knowledge and Data Eng.,* vol. 11, no. 1, pp. 108-117, Jan./Feb. 1999.

[4] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A Bandwidth-Efficient High-Speed Wireless Data Services for Nomadic Users," *IEEE Comm. Magazine,* vol. 38, no. 7, pp. 70-77, July 2000.

[5] S. Choi and K.G. Shin, "An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic," *IEEE/ACM Trans. Networking,* vol. 7, no. 5, pp. 616-628, Oct. 1999.

[6] *The Handbook of Genetic Algorithms.* New York: Van N. Reinhold, L.D. Davis, ed., 1991.

[7] R. Dube, C.D. Rais, and S.K. Tripathi, "Improving NFS Performance over Wireless Links," *IEEE Trans. Computers,* vol. 46, no. 3, pp. 290-298, Mar. 1997.

[8] R. Fourer, D.M. Gay, and B.W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming.* Wadsworth Publishing Company, 1992.

[9] V.K. Garg, *IS-95 CDMA and cdma2000.* Prentice-Hall, 2000.

[10] A.J. Goldsmith and S.-G. Chua, "Variable-Rate Variable-Power MQAM for Fading Channels," *IEEE Trans. Comm.,* vol. 45, no. 10, pp. 1218-1230, Oct. 1997.

[11] J. Gruber, L. Strawczynski, "Subjective Effects of Variable Delay and Speech Clipping in Dynamically Managed Voice Systems," *IEEE Trans. Comm.,* vol. 33, no. 8, pp. 801-808, Aug. 1985.

[12] C.-L.I and and R.D. Gitlin, "Multicode CDMA Wireless Personal Comm. Networks," *Proc. Int'l Conf. Comm.,* vol. 2, pp. 1060-1064, June 1995.

[13] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR: A High Efficiency High Data Rate Personal Communication Wireless System," *Proc. IEEE Vehicular Technology Conf,* vol. 3, pp. 1854-1858, May 2000.

[14] J. Jing, A. Helal, and A. Elmagarmid, "Client-Server Computing in Mobile Environments," *ACM Computing Surveys,* vol. 31, no. 2, pp. 117-157, June 1999.

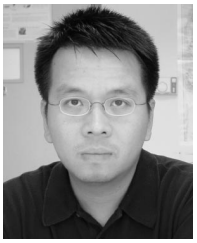
[15] A.D. Joseph, J.A. Tauber, and M.F. Kaashoek, "Mobile Computing with the Rover Toolkit," *IEEE Trans. Computers,* vol. 46, no. 3, pp. 337-352, Mar. 1997.

[16] N. Joshi, S.R. Kadaba, S. Patel, and G.S. Sundaram, "Downlink Scheduling in CDMA Data Networks," *Proc. MOBICOM,* pp. 179-190, Aug. 2000.

[17] F. Kelly, "Charging and Rate Control for Elastic Traffic," *European Trans. Telecomm.,* vol. 8, pp. 33-37, 1997.

[18] K.I. Kim, *Handbook of CDMA System Design, Engineering, and Optimization.* Prentice-Hall, 2000.

[19] J.B. Kim and M.L. Honig, "Resource Allocation for Multiple Classes of DS-CDMA Traffic," *IEEE Trans. Vehicular Technology,* vol. 49, no. 2, pp. 506-519, Mar. 2000.
[20] D.K. Kim and D.K. Sung, "Capacity Estimation for a Multicode CDMA System with SIR-Based Power Control," *IEEE Trans. Vehicular Technology,* vol. 50, no. 3, pp. 701-709, May 2001.
[21] D.N. Knisely, S. Kumar, S. Laha, and S. Nanda, "Evolution of Wireless Data Services: IS-95 to cdma2000," *IEEE Comm. Magazine,* vol. 36, no. 10, pp. 140-149, Oct. 1998.
[22] S. Kumar and S. Nanda, "High Data-Rate Packet Comminication for Cellular Networks Using CDMA: Algorithms and Performance," *IEEE J. Selected Areas in Comm.,* vol. 17, no. 3, pp. 472-492, Mar. 1999.
[23] Y.-K. Kwok and I. Ahmad, "Efficient Scheduling of Arbitrary Task Graphs to Multiprocessors Using A Parallel Genetic Algorithm," *J. Parallel and Distributed Computing,* vol. 47, no. 1, pp. 58-77, Nov. 1997.
[24] Y.-K. Kwok and V.K.N. Lau, "A Quantitative Comparison of Multiple Access Control Protocols for Wireless ATM," *IEEE Trans. Vehicular Technology,* vol. 50, no. 3, pp. 796-815, May 2001.
[25] Y.-K. Kwok and V.K.N. Lau, "Design and Analysis of a New Approach to Multiple Burst Admission Control for cdma2000," *Proc. ACM SIGMOBILE Seventh Ann. Int'l Conf. Mobile Computing and Networking,* pp. 310-321, July 2001.
[26] Y.-K. Kwok, A.A. Maciejewski, H.J. Siegel, A. Ghafoor, and I. Ahmad, "Evaluation of a Semi-Static Approach to Mapping Dynamic Iterative Tasks onto Heterogeneous Computing Systems," *Proc. Fourth Int'l Symp. Parallel Architectures, Algorithms, and Networks,* pp. 204-209, June 1999.
[27] V.K.N. Lau, "Channel Capacity and Error Exponents of Variable Rate Adaptive Channel Coding for Rayleigh Fading Channels," *IEEE Trans. Comm.,* vol. 47, no. 9, pp. 1345-1356, Sept. 1999.
[28] V.K.N. Lau, "Performance Analysis of Variable Rate: Symbol-By-Symbol Adaptive Bit Interleaved Coded Modulation for Rayleigh Fading Channels," *IEEE Trans. Vehicular Technology,* vol. 51, no. 3, pp. 537-550, May 2002.
[29] V.K.N. Lau and Y.-K. Kwok, "Synergy between Adaptive Channel Coding and Media Access Control for Wireless ATM," *Proc. Vehicular Technology Conf.,* vol. 3, pp. 1735-1739, Sept. 1999.
[30] V.K.N. Lau and Y.-K. Kwok, "CHARISMA: A Novel Channel-Adaptive TDMA-Based Multiple Access Control Protocol for Integrated Wireless Voice and Data Services," *Proc. Second IEEE Wireless Comm. and Networking Conf.,* vol. 2, pp. 507-511, Sept. 2000.
[31] Z. Liu, M.J. Karol, M. El Zarki, and K.Y. Eng, "Channel Access and Interference Issues in Multicode DS-CDMA Wireless Packet (ATM) Networks," *Wireless Networks,* vol. 2, pp. 173-193, 1996.
[32] Y. Lu and R.W. Brodersen, "Integrating Power Control, Error Correction Coding, and Scheduling for a CDMA Downlink System," *IEEE J. Selected Areas in Comm.,* vol. 17, no. 5, pp. 978-989, June 1999.
[33] Q. Lu and M. Satyanarayanan, "Resource Conservation in a Mobile Transaction System," *IEEE Trans. Computers,* vol. 46, no. 3, pp. 299-311, Mar. 1997.
[34] P. Marbach, "Priority Services and Max-Min Fairness," *Proc. INFOCOM,* vol. 1, pp. 266-275, June 2002.
[35] A. Massari, S. Weissman, and P.K. Chrysanthis, "Supporting Mobile Database Access Through Query by Icons," *Parallel and Distributed Databases,* vol. 4, pp. 249-269, 1996.
[36] L. Massoulie and J. Roberts, "Bandwidth Sharing: Objectives and Algorithms," *IEEE/ACM Trans. Networking,* vol. 10, no. 3, pp. 320-328, June 2002.
[37] A. Mawira, "Models for the Spatial Correlation Functions of the (log)-Normal Component of the Variability of VHF/UHF Field Strength in Urban Environment," *Proc. Third IEEE Int'l Conf. Personal, Indoor, and Mobile Radio Comm.,* pp. 436-440, Oct. 1992.
[38] MobileMMS.com, http://www.mobilemms.com/, 2002.
[39] K. Nakano, S. Olariu, and J.L. Schwing, "Broadcast-Efficient Protocols for Mobile Radio Networks," *IEEE Trans. Parallel and Distributed Systems,* vol. 10, no. 12, pp. 1276-1289, Dec. 1999.
[40] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation Techniques in Wireless Packet Data Services," *IEEE Comm. Magazine,* vol. 38, no. 1, pp. 54-64, Jan. 2000.
[41] S.G. Nash and A. Sofer, *Linear and Nonlinear Programming.* McGraw Hill, 1996.
[42] A.K. Parekh and R.G. Gallagar, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Network: The Single-Node Case," *IEEE/ACM Trans. Networking,* vol. 1, no. 3, pp. 344-357, June 1993.
[43] J.D. Parsons, *The Mobile Radio Propagation Channel,* second ed. Wiley, 2000.
[44] A.P. Sista, O. Wolfson, and Y. Huang, "Minimization of Communication Cost Through Caching in Mobile Environments," *IEEE Trans. Parallel and Distributed Systems,* vol. 9, no. 4, pp. 378-390, Apr. 1998.
[45] G.L. Stuber, *Principles of Mobile Communications,* second ed., Kluwer Academic Publishers, 2001.
[46] A.J. Viterbi, *CDMA: Principles of Spread Spectrum Communications,* Addison-Wesley, 1995.
[47] L. Wang, Y.-K. Kwok, W.C. Lau, and V.K.N. Lau, "Channel Capacity Fair Queueing in Wireless Networks: Issues and A New Algorithm," *Proc. Int'l Conf. Comm.,* vol. 5, pp. 3116-3120, Apr. 2002.
[48] H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks," *Proc. IEEE,* vol. 83, no. 10, pp. 1374-1396, Oct. 1995.
[49] M. Zorzi, R.R. Rao, "The Role of Error Corrections in the Design of Protocols for Packet Switched Services," *Proc. 35th Ann. Allerton Conf. Comm., Control, and Computing,* pp. 749-758, Sept. 1997.



Yu-Kwong Kwok received the BSc degree in computer engineering from the University of Hong Kong in 1991, the MPhil and PhD degrees in computer science from the Hong Kong University of Science and Technology (HKUST) in 1994 and 1997, respectively. He is an associate professor in the Department of Electrical and Electronic Engineering at the University of Hong Kong. Before joining the University of Hong Kong, he was a visiting scholar in the parallel processing laboratory at the School of Electrical and Computer Engineering at Purdue University. His research interests include wireless networking, mobile computing, network protocols, and distributed computing algorithms. He is a member of the ACM, the IEEE Computer Society, and the IEEE Communications Society, and a senior member of the IEEE.

Vincent K.N. Lau received the BEng (Distinction 1st Hons) in electrical engineering in 1992 from the University of Hong Kong, and the PhD degree in 1997 from the University of Cambridge. He joined the HK Telecom after graduation for three years as a system engineer, responsible for transmission systems design. He was awarded the Sir Edward Youde Memorial Fellowship and the Croucher Foundation in 1995. He joined the Lucent Technologies-Bell labs in the US as a member of technical staff and was engaged in the algorithm design, standardization, and prototype development of CDMA2000 systems. He joined the University of Hong Kong in 1999 as assistant professor and was appointed the codirector of information engineering programme as well as the codirector of 3G Technology center. In July 2001, he left the university and returned to the Wireless Advanced Technology Lab of Lucent Technologies. His research interests include digital transceiver design, adaptive modulation and channel coding, CDMA power control, soft handoff and CREST factor control algorithms, jointly adaptive multiple access protocols, as well as short-range wireless ad-hoc networking. He is currently working on BLAST-MIMO systems, iterative decoding, and UMTS call processing protocol stack design. He is a senior member of the IEEE.