Adaptive and Predictive Downlink Resource Management in Next Generation CDMA Networks

Xin Wang, *Member, IEEE,*, Ramachandran Ramjee, *Senior Member, IEEE* and Harish Viswanathan, *Member, IEEE*

***Abstract*— Guard channels have been proposed to minimize handoff call dropping when mobile hosts move from one cell to another. CDMA systems are power- and interference-limited. Therefore, guard capacity in CDMA networks is soft, that is, a given capacity corresponds to variable number of connections. Thus, it is essential to adjust the guard capacity in response to changes in traffic conditions and user mobility. We propose two schemes for managing downlink CDMA radio resources: Guard Capacity Adaptation Based on Dropping (GAD), and Guard Capacity Adaptation Based on Prediction and Dropping (GAPD). In both schemes, the guard capacity of a cell is dynamically adjusted so as to maintain the handoff dropping rate at a target level. In the second scheme, there is an additional, frequent adjustment component where guard capacity is adjusted based on soft handoff prediction. We show through extensive simulations that GAD and GAPD control the handoff dropping rate effectively under varying traffic conditions and system parameters. We also find that GAPD is more robust than GAD to temporal traffic variations and changes in control parameters.**

***Index Terms*— CDMA, Downlink, Handoff, Prediction, Adaptation, Admission Control.**

I. Introduction

The recent trend in personal communication industry is to provide end users with ubiquitous access to the Internet. Mobility and handoff, however, place stringent requirements on network resources. Whenever a mobile host (MH) in an active session moves from one cell to another, network resources need to be allocated at the new base station (BS). New and handoff session requests will compete for connection resources. QoS degradation or forced termination may occur when there is insufficient resources to accommodate the handoff. The trend in cellular networks of reducing the cell size to increase system capacity results in more frequent handoffs, thus making connection-level QoS even more important.

It is widely accepted in the literature that forced termination of an ongoing call (call dropping) is more annoying than the blocking of a new call. Prioritizing handoff calls [1][2] has been considered to reduce handoff failures. Among various handoff prioritization schemes, channel reservation scheme has been a preferred choice because it can reduce handoff failures with minimum overhead. With this scheme, a portion of the link capacity is reserved for handoffs. Under resource constraints, the blocking probability of handoff calls can be kept lower than that of new calls. However, the research literature on channel reservation schemes have focused mainly on time- and frequency-division multiple-access systems. To minimize the call dropping rate, in a frequency-division multiple access (FDMA)/time-division multiple access (TDMA) system where capacity has a hard limit due to the frequency/time allocation, *hard guard channels* such as time-slots and/or frequency channels can be reserved for handoff calls.

There are three important differences in adapting guard channels for reducing handoff dropping in a CDMA system as compared to a FDMA/TDMA system. First, the capacity of the CDMA system is interference or power limited and hence *soft*. In other words, the capacity of a CDMA system is not fixed and is dependent on a number of factors including the location of the mobile users, their speed, their environment path loss characteristics etc. Thus, a given fixed amount of resource cannot be reserved in order to guarantee, for example, a specific limit on handoff dropping probability. Second, due to the dependence on a variety of factors mentioned earlier, the capacity of a CDMA system is also highly variable. Thus, *any solution for improving handoff dropping in CDMA systems must be highly adaptive and cannot rely on assumptions of traffic or mobility patterns.* Third, CDMA systems are not symmetric and different factors affect uplink and downlink resources. This is due to the *soft-handoff* feature of CDMA where a mobile node's transmission in the uplink (also called reverse link, where the information is transmitted from the mobile to the base station) is automatically received by multiple base stations without using any additional radio resources. Thus, soft-handoff in the reverse link, rather than incurring a cost, actually results in a considerable gain in performance[1]. However, in the downlink direction (also called the forward link, where the information is transmitted from the base stations to the mobiles), establishing soft handoff is costly as secondary base stations must now also transmit the same signal as the primary base station. The power required by the handoff connection is no longer available for allocation to other users of the secondary base stations. Thus, it is necessary in CDMA systems to perform admission control separately in the uplink and downlink directions.

To summarize, in CDMA systems, there are no fixed resources that can be used as guard channels. Instead, a certain amount of *soft guard capacity* has to be reserved. Also, given the asymmetry in CDMA uplink and downlink, call admission control has to be performed differently in the two directions. Furthermore, this reserved capacity has to be constantly adapted to variations due to changing traffic patterns, mobility, environment characteristics etc.

Most previous studies on call admission control in CDMA systems have concentrated on capacity management [3][4][5]. None of these studies considered user mobility. Su *et al.* [6] and Ma *et al.* [7] consider mobility through a soft handoff process in which fixed amount of soft channels are reserved as guard capacity to reduce soft handoff failures. Fixed capacity reservation

Manuscript received February 1, 2004; revised December 1, 2004.

Xin Wang is with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260, USA (e-mail: xwang8@cse.buffalo.edu)

Ramachandran Ramjee is with Bell Laboratories, Lucent Technologie, Holmdel, New Jersey 07733, USA (e-mail: ramjee@dnrc.bell-labs.com)

Harish Viswanathan is with Bell Laboratories, Lucent Technologie, Murray Hill, New Jersey 07974, USA (e-mail: harishv@lucent.com)

[1]Mobile nodes transmit at a minimum power to reach a set of base stations, reducing interference and increasing capacity.

is effective only under ideal stationary traffic conditions and cannot effectively handle a variety of traffic characteristics and users' mobility. Some researchers [8][9][10] have proposed dynamic guard bandwidth schemes for reverse link transmission. As mentioned earlier, admission control is essential for the handoff calls in the forward direction.

In order to give priority to the handoff calls, some power can be reserved for handoff calls in advance. In CDMA systems, the transmission power required for a connection is frequently adapted using open-loop and closed-loop [11] power control so that the signal received by a mobile can meet the target signal-to-noise ratio. It is therefore difficult to predict the power required for a handoff call in advance. The adjustment of power allocations for ongoing sessions will also lead to the variations of available capacity of a cell. In addition, CDMA allows the transmission of both voice and different bit rate data. The dynamics in the power requirement for each mobile and the variety of resource requirements of different applications add more complexity to the radio resource management. Until recently, most research about CAC schemes in CDMA networks have been on the reverse link on the basis of interference levels. Park *et al.* [12] studied a CAC scheme on the CDMA forward link, taking into account both the number of codes[2] and interference level. The proposed scheme gives priority to handoff call by reserving fixed amount of codes and interference margin. Little work has been done in the literature to adaptively control the reserved downlink resources so that the transmission quality of a CDMA call during handoff is guaranteed, taking into account the traffic and power dynamics.

In this paper, we present two novel schemes for effectively managing the downlink CDMA radio resources. The two schemes are: Guard Capacity Adaptation Based on Dropping (GAD), and Guard Capacity Adaptation Based on Prediction and Dropping (GAPD). In both schemes, the guard capacity of a cell is dynamically adjusted so as to satisfy a predetermined bound on the handoff dropping probability without over-penalizing new arrivals. The novelties shared by both the proposed mechanisms are as follows:

- There are no assumptions on traffic and mobility patterns. The proposed schemes can handle the power allocation dynamics of CDMA connections, the changing traffic patterns, the diversified resource requirements and traffic loads, and users' mobility.
- Both schemes apply to the mixture of voice and high-speed circuit data applications.

In GAPD scheme, in addition to the relatively slow adjustment of the guard capacity based on the handoff dropping probability in the cell, there is also a frequent adjustment component based on predictions of handoffs from neighboring cells. The intention is to be able to better handle system dynamics and traffic conditions and to be more robust to the choice of system parameters. The novelties of the GAPD scheme are:

- Handoff direction and attempt are predicted in concert with the pilot-strength power measurement for soft handoff detection.
- Aggregation technique is used so that only the total guard capacity predicted needs to be sent to a neighboring cell at the end of each prediction window.
- The use of dual control and aggregation effectively handles the inaccuracy in handoff predictions.

[2]In a CDMA cellular networks, a set of orthogonal codes are assigned to users to spread information bits to the transmission bandwidth.

This paper is organized as follows. In Section II, we review related work. In Section III, we establish admission criteria and define the associated reserve (guard) capacities in the downlink direction. In Section IV and Section V, we describe the GAD scheme and the GAPD scheme respectively. In Section VI, we present our simulation model and in Section VII, we present extensive results evaluating the performance of the GAD and GAPD schemes. Finally, in Section VIII, we present our conclusions.

II. Related Work

A number of attempts have been made to dynamically control the guard channels. The proposed schemes typically take into consideration the active calls in the cell where a new call arrives, as well as in its neighboring cells to which the call is likely to be handed off. One of the challenges for dynamic guard bandwidth management is to predict where the subscribers will move to. Predictions in the literature are generally based on mobility models or GPS monitoring of the mobile locations. Tracking the speed and moving direction of the mobiles is generally costly and not accurate.

Priscoli and Sestini [8] proposed an adaptive scheme to find an optimum balance between the call blocking and dropping probabilities. The proposed algorithm only relies on the parameters of a single cell, such as the Eb/It [3] received by the BS, the number of call drops and call blocks, and the duration of link unavailability at the BS. The authors did not consider neighboring cell load and mobility patterns. While the scheme proposed by Chang *et al.* [9] controlled the reserved capacity according to variations in the soft handoff attempt rate, the capacity adaptation scheme was not presented. Also, bandwidth reservation based on individual soft handoff attempt would lead to significant signaling overhead between cells. Both these schemes were designed to optimize the linear combination of the dropping and blocking probabilities, but not for satisfying the hard constraints on the call dropping probability often required by applications with tight quality requirements.

The distributed call admission (DCA) scheme by Naghshineh and Schwartz [13] targets to keep the connection handoff dropping probability below a specified limit. The admission control algorithm calculates the maximum number of calls that can be admitted to a given cell without violating the QoS of the existing calls in this cell as well as calls in adjacent cells. However, imprecise control decisions can be made due to a number of simplifying approximations in the control algorithms of DCA. The limited results of the original paper [13] and results rebuilt by authors from [10] show that the scheme cannot always guarantee the target call dropping probability.

Instead of controlling the guard bandwidth, the scheme proposed by Wu *et al.* [10] controls the fraction of new calls to be admitted. The information on channel occupancies and new call arrival rates are exchanged periodically up to the third nearest neighboring cells. The major computational complexity of the control algorithm is to obtain the acceptance ratio by solving a nonlinear equation for the average dropping probability on-line. Numerical method was used to obtain coarse-grain solutions.

The shadow cluster mechanism by Levine *et al.* [14] estimates future resource requirements by implementing a *tentative shadow cluster* around an active mobile for every new and handoff call. Simulations show that this mechanism is able to reduce the

[3]The parameter Eb/It represents the ratio of signal bit energy to total interference and thermal noise power spectral density.

percentage of dropped calls in a controlled fashion. However, the scheme requires the precise knowledge of each user's mobility. Therefore, it is most suitable for a strong directional environment such as the highway. Moreover, the proposed scheme could be computationally too expensive to be practical.

Choi *et al.* [15] designed handoff estimation functions to predict a mobile's next cell and estimate its sojourn time probabilistically based on its previously-resided cell and the observed history of handoffs in each cell. The authors assumed that the handoff behavior of a mobile will be probabilistically similar to the mobiles which came from the same previous cell and are now residing in the current cell. The guard bandwidth is adapted based on the estimation of directions and handoff times of ongoing connections in adjacent cells. Each adjacent cell needs to track the active connections. For each new call admission, the scheme requires the checking of the conditions of some potentially overloaded neighboring cells.

Some of the above work deals with channel allocation, or assumes that the connections consume known amounts of resources. Our approach differs significantly since we deal with CDMA downlink resource management, in which capacity is soft (power-constrained). Also, unlike the above work (including CDMA-compatible schemes), we propose schemes in which no assumptions are made about the traffic characteristics and mobility patterns. Accordingly, our schemes are simple to implement, and robust to inaccurate estimations of mobility and to variations of traffic patterns, mobility, cell dimensions, and control parameters.

III. Forward Link Admission Control and Power Allocation

Before describing our resource management algorithms, we first discuss the concepts of power allocation, guard capacity and admission control in CDMA systems. CDMA systems are interference limited and rely on the *processing gain* (the ratio of transmission bandwidth to the information rate) to be able to operate at a low signal-to-interference ratio (SIR). In each channel, the power transmitted by the base station is controlled to keep the SIR at a receiver at a target value. When the maximum limit of the base station output power is reached, the SIR can no longer be maintained at the target level, and calls serviced by the base station are blocked or dropped. As a result, call admission is closely tied to power control. The capacity of the base station is thus not just determined by the information rates, but is dependent on the power available and its distribution across the mobiles. A common approach to admission control in the downlink direction is to admit new calls as long as the output power at the base station is below a certain threshold [3] [16]. A similar power threshold-based admission control policy is used in this paper.

A. Admission Control and Initial Power Allocation

The total power available at a base station is distributed among overhead channels (pilot, paging, and synchronization channels) and traffic channels. The constraint of the total available traffic power on the power allocation to the downlink traffic channels can be expressed as follows. Assume that a cell k has M_k users. With the total traffic power normalized to 1, let the fraction of traffic power (averaged over the time variations because of fast fading) allocated to a user i be denoted as w_{ki}, and the channel activity factor for the user be denoted as v_i. Then, we have the constraint

$$\sum_{i=1}^{M_k} v_i \omega_{ki} \leq 1. \tag{1}$$

A certain fraction of the traffic power can be reserved in order to minimize dropping of handoff calls: we refer to this as the *guard capacity*. We can now express the admission control decisions for new and handoff calls as follows. Let the total traffic power be normalized to 1, and let the currently allocated power, and the guard capacity be represented respectively as Ω_k and Ω_k^g. Also, we denote the initial power requirement for a new call as ω_{new}, the initial power requirement of a handoff call as ω_{hf}. and the activity factors for new and handoff connections as v_{new} and v_{hf} respectively. Then, the admission control criteria are:

- Admit a new connection at cell k iff

$$\Omega_k + v_{\text{new}}\omega_{\text{new}} \leq 1 - \Omega_k^g. \tag{2}$$

- Admit a handoff connection at cell k iff

$$\Omega_k + v_{\text{hf}}\omega_{\text{hf}} \leq 1. \tag{3}$$

Since the base station controls the transmitted power in closed-loop to maintain a targeted SIR, it does not have a-priori knowledge of the power required by a new or handoff call. Therefore, the initial power (ω_{new} and ω_{hf}) must be estimated.

One way to estimate the initial power of a mobile is to use the average of the powers transmitted for existing connections. If M_k mobile connections are admitted in a cell k and the power allocation for a mobile i is ω_{ki}, the estimated initial power for a new mobile is:

$$\omega_{\text{new}} = \frac{R_{\text{new}}}{M_k} \sum_{i=1}^{M_k} \frac{\omega_{ki}}{R_i}, \tag{4}$$

where R_i and R_{new} are the transmission rate of mobile i and the new mobile respectively, and $\frac{1}{M_k}\sum_{i=1}^{M_k} \frac{\omega_{ki}}{R_i}$ represents the average bit energy of all the admitted mobiles.

During soft handoff, a mobile will connect to and receive power from multiple base stations (constituting an active set), and the received signals from the base stations will be combined at the mobile. To facilitate the maximal ratio combining [11] of signals at the mobile, the base stations in an active set will all allocate the same power fraction to the mobile. So the initial power allocation for a handoff call of a mobile i will be equal to the power allocation at its serving base station[4]:

$$\omega_{i,\text{hf}} = \omega_{i,\text{serving}}. \tag{5}$$

[4]During soft handoff, one of the base stations in the active set is selected as serving base station to be in charge of call-related management functions. A serving base station is normally the one that provides the strongest signal to the mobile or the one that has been serving the mobile for the longest time.

B. Theoretical Power Allocation

After a new or handoff call is admitted, with an initial power allocated as above, the allocated power is adjusted by the base-station in closed-loop to maintain a targeted SIR. However, it is theoretically possible to approximately calculate the power allocation required for a certain SIR [11]. We use this calculation to obtain the power allocation in our simulations.

Assume a user i receives signal power from base station k and interference power from the remaining $J-1$ base stations. Suppose that the total power received by user i from the j^{th} base station is S_{ji}. We also assume that a fraction ϕ_k^t of the total power from a base station k is devoted to the traffic channels and a fraction ω_{ki} of the total traffic power is allocated to a mobile i. Then the ratio of signal bit energy E_b to the total interference and thermal noise power spectral density I_t of a user i can be expressed as:

$$\left(\frac{E_b}{I_t}\right)_i = \frac{\omega_{ki}\phi_k^t S_{ki}/R_i}{(h_{ki}S_{ki} + \sum_{j=1,j\neq k}^{J} S_{ji} + N_0 B_w)/B_w} \qquad (6)$$

where R_i is the bit rate of user i, B_w is the spreading bandwidth, and h_{ki} is the self interference coefficient that models the effect of non-orthogonality due to multipath propagation and transmiter and receiver non-linearities. Hence, with a target $(\frac{E_b}{I_t})_i$ for a user i, we can get the relative allocation of power for user i as:

$$\omega_{ki} = \frac{(\frac{E_b}{I_t})_i}{\phi_k^t G_i}(h_{ki} + \sum_{j=1,j\neq k}^{J} S_{ji}/S_{ki} + N_0 B_w/S_{ki}), \qquad (7)$$

where $G_i = B_w/R_i$ is the processing gain of user i.

Equation (6) applies to the case when the mobile receives signal from only a single base station. When the mobile is in soft handoff with a set of base stations $\mathcal{B}$, then the received E_b/I_t for maximal ratio combining is given by [11]

$$\left(\frac{E_b}{I_t}\right)_i = \sum_{k\in\mathcal{B}} \frac{\omega_i\phi_k^t S_{ki}/R_i}{(h_{ki}S_{ki} + \sum_{j=1,j\neq k}^{J} S_{ji} + N_0 B_w)/B_w}, \qquad (8)$$

where ω_i is now the common power fraction transmitted to the mobile by the different base stations in the active set. The required power fraction ω_i can be obtained by inverting the above equation as in (7).

IV. Guard Capacity Adaptation Based on Dropping (GAD)

The guard capacity in a cell is intended to maintain the handoff dropping rate at a sufficiently low level. On the other hand, if the handoff dropping rate is consistently equal to zero, this may indicate that the guard capacity is too large, at the cost of an unnecessarily high new call blocking rate. Clearly, an optimal amount of guard capacity would allow the most efficient use of the air interface capacity. However, a-priori or fixed optimization of the guard capacity over some known parameters is not feasible in a practical implementation. This is because the traffic pattern in a cell is not known in advance, and varies over the lifetime of the cellular network. Also, as discussed in Section III, the transmission power required for a CDMA connection is frequently adjusted to maintain the signal-to-interference ratio.

The basic objective of both our resource management schemes is to dynamically adapt the guard capacity for the efficient use of the traffic capacity of a cell. In this section, we discuss the GAD scheme. In this scheme, the handoff arrival and dropping rates are monitored by a cell. The handoff dropping rate is maintained at a target level by adjusting the guard capacity, based on a constrained integral control law [17]. With the measured handoff dropping rate of a cell k represented as $B_{k,hf}$ and the target dropping rate as $B^*_{k,hf}$, the guard capacity Ω_k^g for a period n is calculated as:

$$\Omega_k^g[n] = \min\{[\Omega_k^g[n-1] + \sigma_k(B_{k,hf} - B^*_{k,hf})/B^*_{k,hf}]^+, \Omega_k^{g,\max}\}. \qquad (9)$$

where the parameter σ_k controls the adaptation speed of the guard capacity, and $\Omega_k^{g,\max}$ is the maximum guard capacity allowed for cell k. Note that $[x]^+$ requires x to be not less than 0.

In an integral controller such as ours, a higher σ_k leads to a faster response, but also leads to larger oscillations and possible instabilities. Also, if $|\sigma_k(B_{k,hf} - B^*_{k,hf})/B^*_{k,hf}|$ is too large, Ω_k^g may be absorbed into an extreme state. Therefore, the value of σ_k should be constrained [5].

V. Guard Capacity Adaptation Based on Prediction and Dropping (GAPD)

The basic concept of the GAPD scheme is to anticipate the soft handoffs to a cell before they occur, in addition to monitoring the handoff dropping rate in the cell, as in Section IV. The guard capacity is adjusted based on both the predicted handoffs as well as the handoff dropping rate. One of the challenges in this approach is the prediction of soft handoff calls to a cell, and signaling of the anticipated handoffs to that cell. This is discussed in Section V-A. The adjustment of the guard capacity is then discussed in Section V-B.

A. Soft Handoff Prediction

We begin with a brief explanation of how a soft handoff is initiated in CDMA systems. During inter-cell handoff, a mobile sends and receives information from both new and old base stations. The pilots of the cells involved in the soft handoff are categorized into an *active set*. A mobile periodically measures the pilot signal strength received from neighboring cells. If the mobile finds a neighboring BS with a pilot signal strength E_c/I_o higher than a predetermined threshold T_{ADD}, the mobile transfers the BS associated with the pilot into the *candidate set* and sends a Pilot Strength Measurement Message to the serving base station, which will send a handoff request to the target base station. If the BS can be added into the active set, the serving base station sends a Handover Direction Message to the mobile. If the pilot signal from either the old BS or the new BS drops below threshold T_{DROP} for an amount of time T_{TDROP}, the corresponding link is released.

Since the measured pilot signal strength is used to initiate soft handoff, we propose using the pilot signal strength to predict soft handoff. We define a new parameter, a soft handoff *prediction threshold* T_{PREDICT} that is set lower than T_{ADD}. When a mobile detects that the pilot signal strength from a neighboring cell exceeds T_{PREDICT}, the mobile predicts the neighboring cell as a handoff target. The mobile signals its serving base station

[5] With the range constraints, care must be taken that Ω_k^g does not get absorbed into the extreme states. Assume that ϵ is the largest error that occurs once the system is in closed-loop operation. The parameter Ω_k^g can be prevented from being absorbed into an extreme state if $\sigma_k < \frac{\Omega_k^{g,\max} B^*_{k,hf}}{\epsilon}$.

indicating that it is approaching the predicted cell, and the serving base station identifies the mobile as candidate for handoff to the neighboring cell in the impending future and signals to the cell to reserve guard capacity. On the other hand, if a mobile detects that the pilot strength from the cell originally predicted as a handoff target drops below T_{DROP} for a time period T_{TDROP} before its reaching T_{ADD}, it signals the serving base station to cancel the handoff prediction. Again, the serving base station identifies the mobile accordingly.

If the pilot strength from a cell predicted as handoff target reaches T_{ADD} and the cell can admit the mobile, the predicted cell is added into the mobile's active set and the mobile initiates soft handoff. Irrespective of whether the mobile is admitted into the target cell, the corresponding guard capacity is no longer needed, and the target base station reduces the guard capacity accordingly.

1) Prediction Aggregation and Signaling: If the serving base station needs to inform a neighboring base station about each handoff prediction, signaling overhead may become excessive. Therefore, we define a *prediction window* with length W_p, over which predictions are aggregated. For a target cell k, at time intervals W_p, a serving cell j calculates a net predicted required power Ω^p_{jk}, which is given by the difference between the total estimated power requirement corresponding to handoff predictions (pilot strength is higher than T_{PREDICT}), and the total estimated power requirement corresponding to withdrawn predictions (pilot strength drops below T_{DROP} before reaching T_{ADD}) during the time interval W_p. Therefore,

$$\Omega^p_{jk} = \sum_{i \in C^p_{jk}} \omega_i - \sum_{i' \in C^q_{jk}} \omega_{i'}. \tag{10}$$

In Equation 10, ω_i is the power of the active session i at the time of prediction, $\omega_{i'}$ is the power of the active session i' at the time the prediction was withdrawn, C^p_{jk} is the set of indices of the active sessions predicted to handoff to cell k, and C^q_{jk} is the set of indices of the active sessions with withdrawn handoff predictions to cell k. If Ω^p_{jk} is non-zero, the serving base station sends a guard capacity update message containing Ω^p_{jk} to the target base station k at the end of the prediction window. In Section V-B, we will describe the algorithms according to which the guard capacity is actually adapted, based on the net estimated power requirement Ω^p_{jk}, initiated handoffs to the target base station, and the handoff dropping rate.

2) Prediction Parameters: We now discuss the trade-offs involved in selecting values for the various prediction parameters. First, the length of the prediction window, W_p, trades off the signaling overhead with the granularity of guard capacity adaptation.

For each handoff prediction, we define a *prediction interval*, as the time interval between the measured pilot signal strength from a neighboring cell reaching T_{PREDICT} (when handoff is anticipated) and its reaching T_{ADD} (when handoff can be performed). When a predicted target cell receives a handoff prediction, it may not have sufficient spare capacity (that is not currently consumed by active mobiles or already booked as guard capacity) to set the required guard capacity. The longer the prediction interval, the more likely it is that the predicted target handoff cell can set aside guard capacity corresponding to the predicted handoff, as other mobiles release capacity, or consume less capacity than predicted.

The length of the prediction interval is dependent on the prediction threshold T_{PREDICT} relative to T_{ADD}, and the mobile's speed and moving direction. Reducing T_{PREDICT} increases the prediction interval, but setting T_{PREDICT} too low will cause more incorrectly predicted handoffs, and may in turn result in excessive guard capacity and a higher new call blocking rate. Ideally, since the prediction interval depends on the mobile's speed and direction, each mobile should have its own handoff prediction threshold T_{PREDICT}. However, the mobility characteristics of a mobile are generally not known a priori.

The guard capacity set aside in a cell is generally shared by all the mobiles that handoff to this cell. Therefore, the resource needed by a fast moving mobile with a short prediction interval can be borrowed from slower moving mobiles with earlier handoff predictions. We will see in our performance studies that due to this guard capacity sharing, the sensitivity of the performance to the prediction threshold is reduced.

In IS-95A, the handoff thresholds T_{ADD} and T_{DROP} are set as constants. However, some locations in the cell only receive weak pilots (requiring a lower threshold) and other locations receive a few strong and dominant pilots (requiring higher handoff thresholds). As a result, IS-95B proposes dynamic thresholds. We take this into account by setting the thresholds T_{PREDICT} relative to T_{ADD}, instead of as absolute values.

B. Guard Capacity Adaptation Based on Prediction and Dropping

Having established the handoff prediction strategy, we now consider the actual adaptation of the guard capacity. This adaptation is carried out at two different time scales: a rapid adaptation in response to handoff predictions, and a longer-term adaptation based on the handoff dropping probability of the cell.

1) Adaptation upon Prediction - Fast Control: At the end of a prediction window, if the total predicted power for a neighboring cell is not zero, the serving base station sends an estimated aggregate power requirement to the neighboring target base station. However, several problems may arise if this power is added to the guard capacity directly. Since the transmission power in a channel is adjusted frequently to maintain the signal quality at the mobile, the power requirement of a mobile at the time of handoff can be different (lower or higher) from the estimated power at the time of handoff prediction. Also, some mobiles that were originally predicted to handoff into a cell may end their calls or change direction before they arrive at the cell, resulting in higher than necessary guard capacity setting, and possibly a higher new call blocking probability B_n. Finally, if a cell always sets aside sufficient capacity for every anticipated handoff, the handoff failure probability B_{hf} is theoretically zero. In practice, the handoff failure probability is only required to be below a desired value, say, 1%.

In order to track the power requirement dynamics and compensate for the prediction errors, and hence maintain the correct trade-off between high capacity utilization (low B_n) and low handoff failure probability, we introduce a *scaling factor* for the predicted power requirement. A cell k adapts its guard capacity by scaling the predicted power by a factor α_k. Each cell has its own prediction scaling factor, which is adjusted at the end of every prediction window based on the moving average handoff dropping probability of the cell, using an integral control law. The scaling factor α_k for a cell k during the m^{th} prediction window is given by:

$$\alpha_k[m] = \min\{\max\{\alpha_k^{\min}, \alpha_k[m-1] + \sigma_k^{\alpha}(B_{k,hf} - B^*_{k,hf})/B^*_{k,hf}\}, \alpha_k^{\max}\} \tag{11}$$

where parameter σ_k^{α} controls the adjustment speed of α_k, and $\alpha_k^{\min}$ and $\alpha_k^{\max}$ are the minimum and maximum values of α_k.

2) Adapting Minimum Guard Capacity - Slow Control: As mentioned in Section V-A.2, even if a handoff can be correctly predicted, if the target cell is highly loaded, the target cell may not be able to release the required amount of guard capacity by the time of the handoff. One solution is to make the prediction interval variable (by making T_{PREDICT} variable), and adjust the prediction interval based on the handoff dropping probability. When the handoff dropping probability increases, the prediction interval could be increased in response, thus allowing the target cell more time to use freed-up resources to increase the guard capacity. However, periodically conveying the new prediction threshold to each mobile would increase the signaling overhead in the air interface. In addition, the prediction interval cannot be controlled by prediction threshold alone, but also depends on each mobile's speed and moving direction.

We consider an alternative solution. The problem arises because the target cell allows the guard capacity to fall too low, in response to dynamics in the handoff predictions and actual handoff attempts. Accordingly, we introduce a certain amount of minimum guard capacity $\Omega_k^{\text{g,min}}$, which remains practically constant on the time-scale of the handoff prediction process, independent of handoff predictions and attempt rate.

However, to make it easier to estimate the right amount of minimum guard capacity, we make the minimum guard capacity dependent on the handoff dropping rate over a longer time-scale. We use a similar control scheme to that used for adapting the guard capacity in the GAD scheme (equation 9). However, we use a longer control window LW_p, with the control loop driven by mismatch between the *long-term* measured handoff dropping probability $B_{k,hf}^{l}$ and the target value $B_{k,hf}^{*}$.

3) Guard Capacity Adaptation: We now summarize the overall guard capacity adaptation under GAPD. The guard capacity adaptation of a cell is driven by several inputs: predicted soft handoffs, the soft handoff attempt rate, and the short-term and long-term mismatch between the measured and target handoff blocking rates. Consider a cell k. At the end of the m^{th} prediction window, the cell adjusts its guard capacity based on handoff predictions, by an amount $\alpha_k[m]\Omega_k^p$. Here Ω_k^p is the total estimated power requirement predicted by all its neighboring cells during that prediction window, that is,

$$\Omega_k^p = \sum_{j \in N_k} \Omega_{jk}^p, \tag{12}$$

where N_k is the set of indices of cell k's neighbors with mobiles predicted to handoff to cell k.

At the same time, cell k also reduces the guard capacity by an amount $\alpha_k[m-1]\Omega_k^f$, corresponding to attempted handoffs to the cell. Here Ω_k^f is the total power requirement of all attempted handoffs to the cell in that prediction window,

$$\Omega_k^f = \sum_{i \in C_k^f} \omega_{i,hf}, \tag{13}$$

where $\omega_{i,hf}$ is the power of the session i at handoff time, and C_k^f is the set of indices of active sessions that attempt handoff to cell k. Note that no matter whether a handoff is admitted or rejected by the cell, the guard capacity reserved for the session is no longer needed after the handoff is performed.

Therefore, the guard capacity for cell k during the m^{th} handoff prediction period can be written as:

$$\begin{aligned} \Omega_k^g[m] \quad &= \quad \Omega_k^g[m-1] + \alpha_k[m]\Omega_k^p - \alpha_k[m-1]\Omega_k^f, \\ &\text{and} \qquad \Omega_k^{\text{g,min}} \le \Omega_k^g[m] \le \Omega_k^{\text{g,max}}. \end{aligned} \tag{14}$$

We should note that the guard capacity adaptation in a cell occurs fairly independently of other cells. Although the prediction threshold and add and drop thresholds are assumed to be the same across all cells, the predicted power scaling factor α_k, and the minimum guard capacity $\Omega_k^{\text{g, min}}$ are adjusted independently in each cell k, based on the short-term and long-term variations of handoff dropping probability in the cell. The length of the prediction window W_p and the long control window LW_p can also be different in each cell, and the requests for guard capacity from neighboring cells need not be synchronized with each other.

VI. Simulation Model

We describe the simulation set-up in this section, and discuss simulation results in Section VII. We simulate the GAD and GAPD schemes, as well as a scheme with fixed guard capacity (FG). We introduce the path loss model used for the simulations in Section VI-A, and state our assumptions and default parameter values in Section VI-B.

A. Path Loss Model

We consider path loss and shadowing in our path model. Since only signal strength measurements and transmit power values averaged over time scales corresponding to the fast fading are considered for handoff decisions and admission control, we do not include fast fading in our simulations. The path loss is modeled using the COST231-Hata model proposed by Mogensen [18]. The signal from the base station to the user is assumed to decay at the rate of 3.5^{th} power of the distance. We assume each base station has the same power P. The signal received by a user from all the base stations except the one that is serving the user is treated as interference. Considering only path loss, the interference power from each interfering base station j to a user i is

$$S_{ji} = \frac{P * c}{d_{ji}^{3.5}}, \tag{15}$$

where d_{ji} is the distance between the base station j and the user i. The constant c corresponds to the intercept in the path loss model and is assumed to be 28.5 dB when distance is in meters [18].

The slow shadow fading is modeled by independent log-normal variables. To account for the spatial correlation of the shadows, we assume the model proposed by Gudmundson [19], where log-normal shadowing is modeled as a Gaussian white noise process that is filtered by a first-order low-pass filer

$$\Psi_{l+1(dB)} = \zeta \Psi_{l(dB)} + (1-\zeta)\vartheta_l, \tag{16}$$

where $\Psi_{l(dB)}$ is the mean envelope or mean-squared envelope expressed in decibels, that is experienced at location l, ϑ_l is a zero-mean Gaussian random variable with the standard deviation of 8 dB, and ζ is a parameter that controls the spatial correlation of the shadows. Every T seconds, the spatial correlation factor ζ for a mobile that is traveling with velocity v is calculated as

$$\zeta = \zeta_D^{(vT/D)}, \tag{17}$$

where ζ_D represents a shadow correlation between two points separated by a spatial distance of D m. In our simulation, ζ_D is set to 0.82 for a distance of 100 m, based on the experiments by Gudmundson [20]

Taking into account the shadowing, the interference power received from an interfering base station j by a user i at location l is

$$S_{ji} = \frac{P * c}{d_{ji}^{3.5}} 10^{\frac{\Psi_l}{10}}. \tag{18}$$

B. Assumptions and Parameter Defaults

To simulate a very large PCS network, the authors in [21] advocate a wrap-around topology. This approach eliminates the boundary effects in an un-wrapped topology. Thus, we simulate our PCS network using a wrapped mesh topology with 25 squared cells. Each cell is surrounded by two rings of base stations so that a significant fraction of interference is captured. We make the following assumptions in our simulations:

A1. The movement of the mobile users is based on a two-dimensional random walk model, that is, the mobiles can travel in any direction in a plane with an equal probability. The speed of a mobile is chosen randomly below $SP^{\max}$. The default $SP^{\max}$ is set to 100 km/hour, unless otherwise specified. Initial mobiles are generated randomly and uniformly across the cells, and can appear anywhere with an equal probability. After a mobile is initiated (i.e., a mobile subscribes to the system), its location is tracked even when it is inactive.

A2. The default diameter of a cell is 2 km, and all the base stations are assumed to use the same transmission power of 15 W. For each base station, 20 % of the power is assigned to pilot channel, 70 % of the power is assigned to traffic channel [22], and the remaining power is assigned to other control channels. For an active session, closed-loop power control is simulated to maintain the minimum bit energy to noise density ratio Eb/Io at some pre-determined target level, 5 dB for voice and 1 dB for data. The spread bandwidth B_w is 3.84 MHz, and the thermal noise $N_o B_w$ is set to -105 dBm, derived from [22]. The self-interference factor h_{ki} in Equation 7 is set to 0.01, reflecting the transmitter and receiver non-linearities.

A3. Connection requests are generated according to a Poisson distribution at a rate that varies with the required simulation load and the number of subscribers. Each connection's lifetime is exponentially-distributed with mean interval of one minute. Unless otherwise specified, 70 % of the total traffic is voice, with activity factor 0.5 and rate chosen randomly from the rate set of 10.2 kb/s, 6.7 kb/s, 5.9 kb/s and 4.95 kb/s. The remaining 30 % of the traffic is data, with activity factor 1 and rate chosen randomly from the rate set of 14.4 kb/s, 28.8 kb/s, 57.6 kb/s, 64 kb/s, 128 kb/s, and 384 kb/s [23].

A4. The handoff parameters are set as: $T_{\text{ADD}} = -13$ dB, $T_{\text{DROP}} = -15$ dB, $T_{\text{TDROP}} = 2$ s, derived from [22]. The default prediction threshold T_{TPREDICT} is set to 0.85 T_{ADD}.

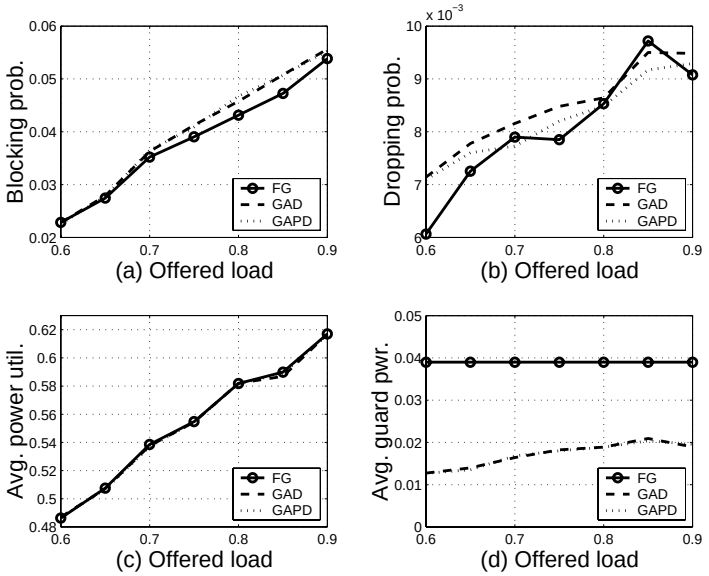

Fig. 1. Performance metrics of capacity management for FG, GAD and GAPD: (a) new call blocking probability; (b) handoff call dropping probability; (c) average cell power utilization; (d) average cell guard capacity.

A5. The target handoff dropping probability is set to 0.01 [24], the adaptation step for the predicted power scaling factor α is set as 0.08, and the range of α is constrained to [0.05, 1]. The prediction window W_p is set as one second for all the cells. The adaptation steps for guard capacity of GAD and minimum guard capacity of GAPD are set to the same value of 0.00025, the long-term control interval LW_p for adjusting minimum guard capacity in GAPD and interval for adapting guard capacity of GAD are both set as 20 seconds. The guard capacity as a fraction of total traffic power is constrained to be below 0.30.

VII. Performance Evaluation

In this section, we present a detailed performance evaluation of three schemes: 1) fixed guard capacity (FG), 2) guard capacity adaptation based on handoff dropping probability (GAD), and 3) guard capacity adaptation based on prediction and handoff dropping probability (GAPD). The central problem considered in this paper is to set aside the right amount of guard capacity so as to obtain a good trade-off between call quality (low handoff dropping probability) and availability (low call blocking probability and high cell power utilization). Accordingly, we use the following performance metrics:

- *New Call blocking probability* - the number of new calls blocked as a fraction of the number of new arrivals received.
- *Handoff dropping probability* - the number of handoff calls blocked as a fraction of handoff calls received.
- *Average cell power utilization* - the power consumed by the active sessions as a fraction of the total traffic power available.
- *Average cell guard power fraction* - the average fraction of traffic power set aside as guard capacity.

These metrics are shown as functions of offered load. The offered load is defined as the average number of mobiles in a cell normalized with the maximum number of mobiles a cell can support, which is calculated based on the average data rate due to different types of traffic combinations.

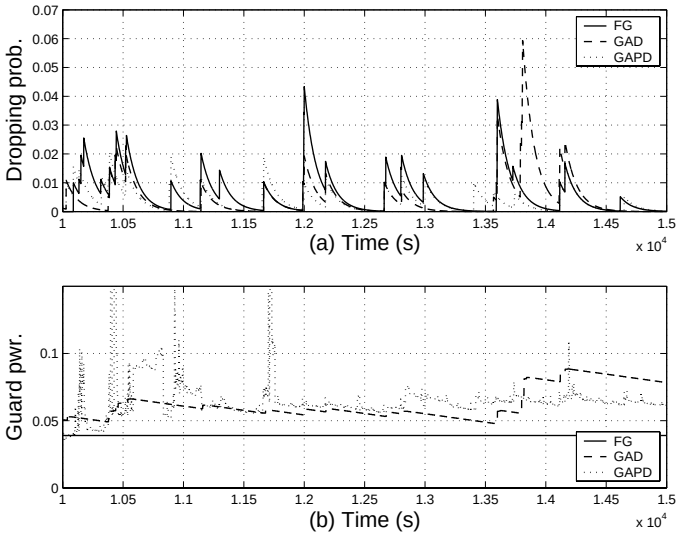

Fig. 2. Time variation of handoff dropping probability and guard power for GAD and GAPD at offered load 0.85.

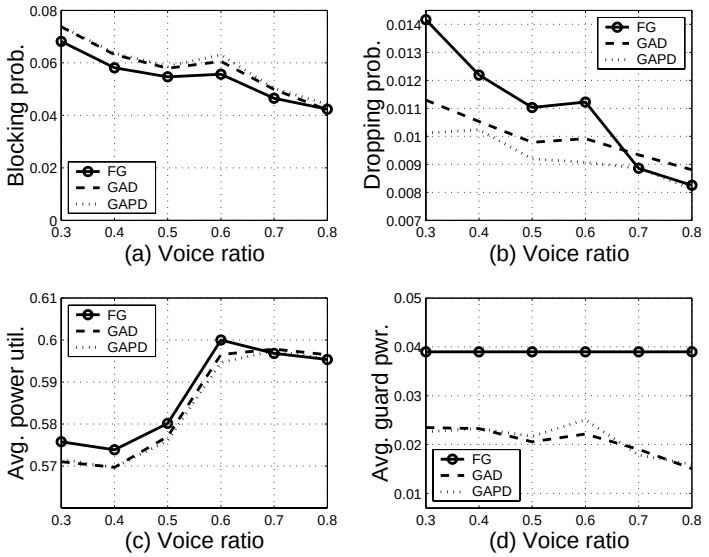

Fig. 3. Performance with the variation of voice traffic ratio: (a) new call blocking probability; (b) handoff call dropping probability; (c) average cell power utilization; (d) average cell guard capacity.

In the next section, we compare the performance of the three schemes under default parameter settings. We then examine the impact of voice ratio (varying the voice-data traffic mix) and user mobility speed on the performance of these schemes. We also examine the effect of cell size and base station power on these schemes, as these parameters effectively alter the load patterns. Finally, we examine the robustness of the GAD and the GAPD schemes to changes in the various control parameters.

A. Comparison

We first compare the basic performance of three schemes, FG, GAD, and GAPD. Clearly, the performance of the FG scheme depends on the amount of fixed guard capacity: the optimal guard capacity depends on the size and power of the cell, and the traffic pattern (voice/data ratio, mobility, etc.). Under the default conditions, the optimal guard capacity is about 0.039; we used this value in the simulations of FG. The performance of the FG scheme is then similar to that of the GAD and GAPD schemes, in terms of blocking probability (except at high loads), handoff dropping probability, and cell power utilization (Figs. 1 (a), (b) and (c) respectively). All three guard capacity schemes are able to maintain the handoff dropping probability at or below the target level, which is achieved at the cost of a smaller increase of new call blocking probability. However, since FG only works optimally for one particular traffic/mobility/load configuration, its performance under different conditions is generally much worse than that of GAD and GAPD as will be seen in later sections.

Fig. 1 (a) shows that GAD and GAPD have slightly higher new call blocking probabilities as compared to FG at high load. This is because GAD and GAPD rely on an iterative process (guard capacity adjustment based on handoff predictions and/or dropping probability in previous period), and are inherently more conservative in reserving guard capacity during instantaneous high handoff loads. Fig. 1 (d) shows that GAD and GAPD have much smaller average guard capacities than FG; this is because both schemes reduce their guard capacity during periods of low handoff load, although they have comparable or slightly higher guard capacities during periods of high handoff load.

The instantaneous handoff dropping probabilities and guard capacities are shown in Figs. 2 (a) and (b), which are based on a snapshot of a single cell between 10000 seconds and 15000 seconds. Even though GAPD is shown to control the instaneous dropping probability around the target level, burst dropping is still seen with GAD scheme. Prior to and during bursts of high handoff load (indicated by peaks in the dropping probability traces), GAPD is seen to adapt the guard capacity more actively than GAD (because of handoff prediction), and avoid very high peak dropping probability during these bursts. Guard capacity reservation is different from capacity reservation. Guard capacity is used when the remaining capacity is low to prevent the new calls from being accepted and hence give handoff calls priority. But the guard capacity is only used by a handoff call at admission control time and is released immediately after the admission control is completed for the call. The same guard capacity can then be reused for other handoff calls. Therefore, even though the amount of guard capacity of GAPD is not significantly higher than that of GAD between time 1.35×10^4 and 1.4×10^4, it is effective enough to reduce the dropping probability burst.

Note that at time 1.1×10^4 the dropping probability of GAPD does not drop immediately, although there is a high guard power allocation. When the remaining capacity of a cell is smaller than that of the guard capacity, even though guard capacity is reserved, it may take some time for a cell to accumulate enough capacity (e.g., released when some existing connections leave) for admission control purpose. Also, the aggregation scheme assumed by GAPD may lead to some delay in guard capacity reservation. However, GAPD is not designed to keep handoff blocking rate to zero, but rather to well control the handoff blocking probability to avoid high burst.

B. Effect of Voice Ratio and User Mobility

We study the impact of traffic patterns by varying the ratio of voice traffic to data traffic, and by varying the maximum mobile speed $SP^{\max}$. Other parameter values are at default levels; in particular, the offered load is 0.85.

Voice connections generally have lower data rates and a smaller range of data rates than data traffic, and voice traffic is less bursty than data traffic. If the total offered load is the same, a larger fraction of voice traffic allows for better multiplexing and hence

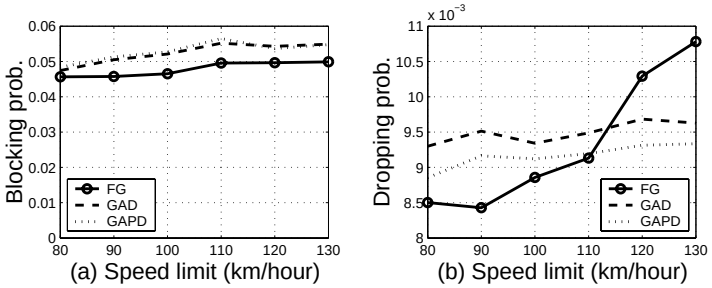

Fig. 4. Probability of new call blocking and handoff call dropping with variation of the allowable mobile speed limit $SP^{\max}$.

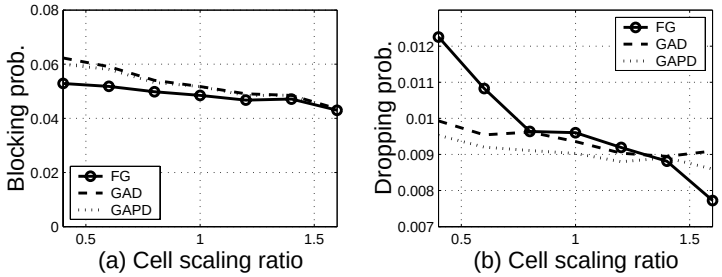

Fig. 5. Probabilities of new call blocking and handoff call dropping with the variation of cell size.

more efficient resource usage. Therefore, for all three schemes, both the new call blocking probability and handoff dropping probability decrease as the voice ratio increases.

Fig. 3 (b) shows that when the voice ratio decreases (making the overall traffic burstier), the GAD and GAPD scheme are generally able to keep the handoff dropping probability below the target level (0.01), by reserving more guard capacity (Fig. 3 (d)). The FG scheme, on the other hand, cannot keep the dropping probability within the target level at small voice ratios. The significant handoff performance improvement of GAD and GAPD come at the cost of a slightly lower cell power utilization due to the higher guard capacity, and correspondingly slightly higher new call blocking probability (Fig. 3 (c),(a)).

The dropping probability under GAD is always somewhat higher than that under GAPD, with the largest difference between the two schemes (11.3%) occurring at the smallest voice ratio of 0.3. Since both schemes reserve approximately the same average guard capacity even at small voice ratios(Fig. 3 (d)), the advantage under GAPD under the most dynamic conditions evidently comes from the fast handoff prediction based adaptation.

In our simulations, the speeds of the mobiles are randomly generated between zero and a maximum speed $SP^{\max}$. Increasing $SP^{\max}$ increases user mobility, and therefore increases the frequency of handoffs. With a fixed guard capacity, the increased frequency of handoffs with speed limit results in the handoff dropping probability being 8% higher than the target probability at the highest speed $SP^{\max}$ (Fig. 4 (b)), while GAD and GAPD maintain the dropping probability at less than the target (0.01) at all speed limits. Note that the advantage under GAD and GAPD would be even greater at higher offered load.

C. Effect of Cell Size

In this experiment, we study the sensitivity of the three guard schemes to the change of physical cell size (i.e., the distance between base stations), while keeping the equal average traffic arrival rate for each cell.

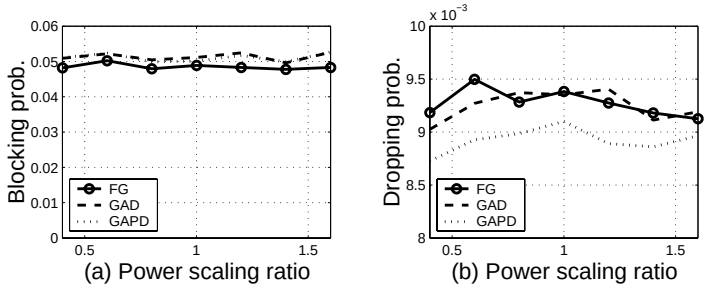

Fig. 6. Probabilities of new call blocking and handoff call dropping with the variation of base station power at offered load 0.85.

A decrease in cell size effectively increases the frequency of handoffs. Changing the cell size also has two conflicting effects on the mobile power requirement. On the one hand, when the cell size is reduced, interference increases, tending to increase the power requirement of a mobile. On the other hand, since the mobiles are on average closer to the base station, there is lower path loss, tending to reduce the transmission power requirement. Therefore, in general, changing the cell size effectively changes the load pattern.

Fig. 5 shows that, for all three schemes, the dropping probabilities increase as the cell size decreases. The dropping probability of FG scheme is up to 23 % higher than the target when the cell size is scaled down by 0.4. Both the GAD and GAPD schemes better maintain the target handoff dropping rate as the cell size varies (at the cost of a slightly higher new call blocking rate), with GAPD having a somewhat lower dropping rate and comparable new call blocking rate over the entire size range.

D. Effect of Base Station Power

In this experiment, we vary the base station power by scaling the default power of each cell, while keeping the equal average traffic arrival rate for each cell.

As with cell size, changing the base station power has two conflicting effects: increasing the total transmission power of cells tends to increase interference and the power requirement to maintain signal to interference ratios, and hence tends to increase the effective load, but there is also more available power to handle the increased load. Since the traffic is generated randomly across all the cells, our simulations indicate that for all three schemes, there are no significant changes in the dropping probabilities as the BS power varies. The dropping rates of GAD and FG are comparable, while that of GAPD is somewhat smaller. The new call blocking probability of GAD and GAPD are slightly higher than that of FG.

E. Effect of Control Parameters

In this section, we study the effect of various control parameters on the performance of GAD and GAPD. The parameters we consider are: the target handoff dropping probability B^*_{hf}; the size of the prediction threshold, T_{PREDICT}, and prediction window, W_p, in GAPD; the length of the long-term adaptation period LW_p; the guard power scaling factor α; and the adaptation step σ. We vary one parameter at a time, while keeping the other parameters at default values. The offered load is fixed at 0.85.

1) Target handoff Dropping Probability: We vary the target handoff dropping probability in this simulation. Fig. 7 shows that both GAPD and GAD are able to keep the dropping probability

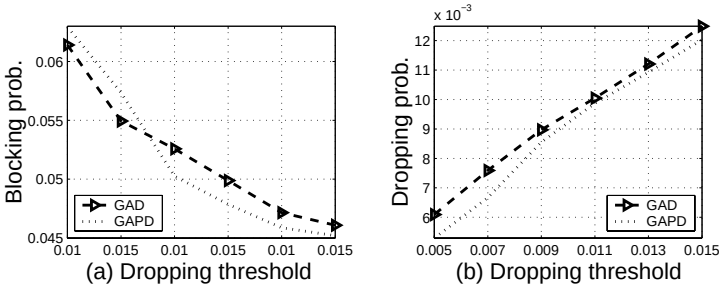

Fig. 7. Probability of new call blocking and handoff call dropping with the variation of handoff dropping threshold B^*_{hf}.

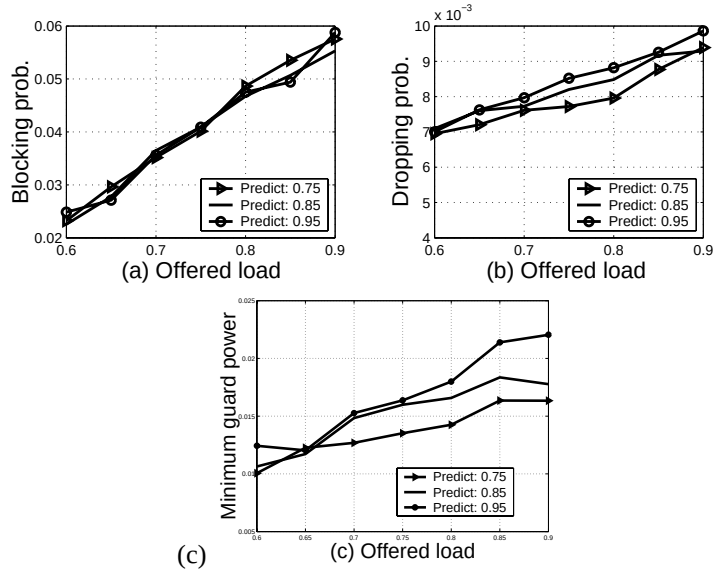

Fig. 8. Probability of new call blocking and handoff call dropping, as well as minimum guard capacity variation at different prediction thresholds

at or below the target for target values higher than 0.008, with GAPD having somewhat lower dropping and new call blocking rates. For target dropping probabilities less than 0.008, GAD with the default control parameters can no longer keep the dropping rate within the target, while GAPD is able to do so, at the cost of slightly higher new call blocking probability.

2) Prediction Threshold $T_{PREDICT}$ in GAPD: In this section, we look at the effect of the T_{PREDICT} threshold in GAPD, which defines the power threshold at which a handoff is predicted. By default, T_{PREDICT} is set as $0.85T_{\text{ADD}}$. We show simulation results for three different values of T_{PREDICT} (as a fraction of T_{ADD}): 0.75, 0.85, and 0.95.

Setting a lower prediction threshold results in earlier handoff predictions. This allows the target cell more time to accumulate guard capacity, and results in a lower dropping rate, as seen in Fig. 8 (b). On the other hand, a lower prediction threshold causes guard capacity to be held longer. It also causes more withdrawn predictions, which again result in unnecessary guard capacity as well as more signaling overhead. This would tend to increase the new call blocking probability; however, the longer term adjustment of minimum guard capacity in GAPD prevents excessive guard capacity from being held too long. This is seen in Fig. 8 (c): as the prediction threshold is reduced, the minimum guard capacity at a given load decreases, countering the increase of prediction-based guard capacity. This prevents the new call blocking probability from increasing significantly as the prediction threshold is reduced (Fig. 8 (a)), and makes the GAPD scheme robust to the setting of the prediction threshold.

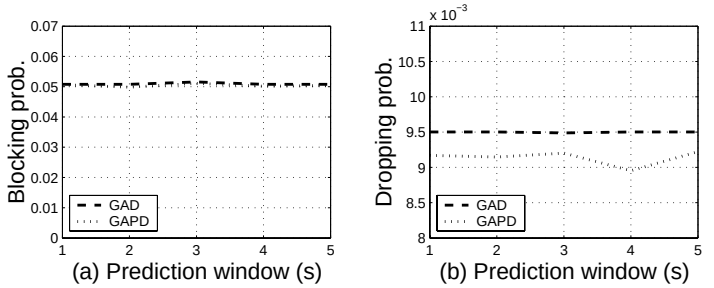

Fig. 9. Probability of new call blocking and handoff call dropping with the variation of prediction window W_p.

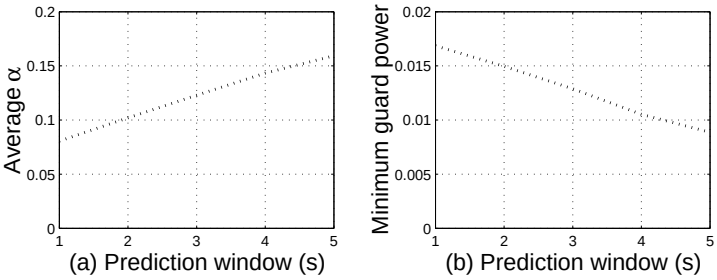

Fig. 10. Average prediction scaling factor α and minimum guard power fraction with the variation of prediction window W_p.

3) Prediction Window Length W_p in GAPD: The prediction length controls the interval at which a cell sends aggregated handoff power predictions to a neighboring cell. A larger W_p means longer delays between making handoff predictions and signaling them, which reduces the time for the target cell to accumulate guard capacity. This would tend to increase handoff dropping probability. However, Fig. 9 shows that handoff dropping probability and new call blocking probability remain well-controlled over the entire range of W_p settings. This is likely due to several reasons. While handoff predictions are delayed longer with larger W_p, withdrawn predictions are also delayed longer, so that the corresponding excess guard capacity is held longer. Also, the scaling factor α, which controls the fraction of the predicted handoff power actually added to (or subtracted from) the guard capacity, is adjusted in response to handoff dropping probability. As shown in Fig. 10 (a), α indeed increases with the increase in W_p, causing guard power reservation to change more sharply in response to predictions. Fig. 10 (b) shows that the minimum guard capacity decreases at the same time, reflecting the more abrupt prediction-based adaptation of guard capacity.

4) Long-term Control Period LW_p: The LW_p parameter sets the interval at which the guard capacity is adjusted in GAD, and the minimum guard capacity is adjusted in GAPD, in order to maintain the handoff dropping probability at its target value. Fig. 11 shows that the dropping probability of GAD increases much

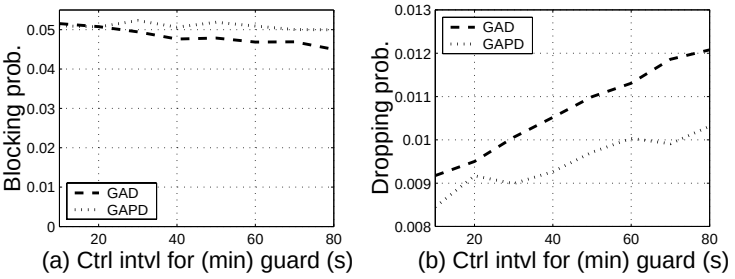

Fig. 11. Probability of new call blocking and handoff call dropping with the variation of long-term control period LW_p.

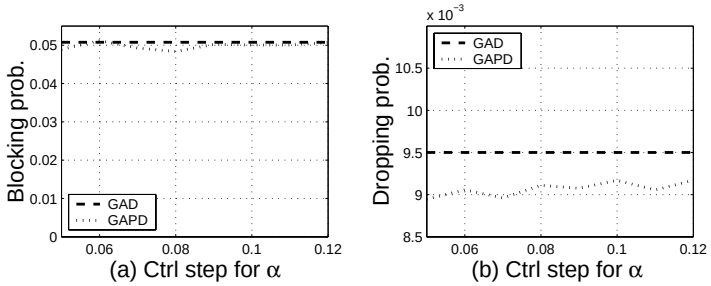

Fig. 12. Probability of new call blocking and handoff call dropping with the variation of adaptation step for α.

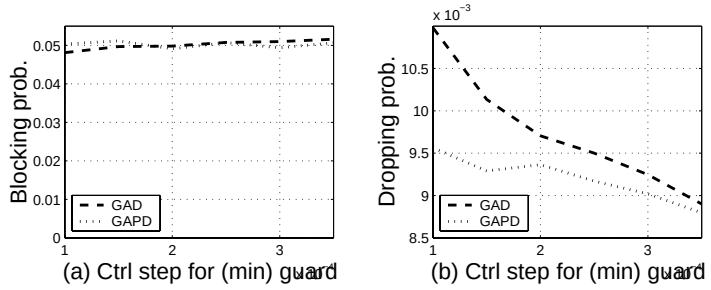

Fig. 13. Probability of new call blocking and handoff call dropping with the variation of adaptation step for minimum guard capacity of GAPD and guard capacity of GAD.

more sharply than that of GAPD with the increase of the control period, and exceeds the target when the control period is larger than 30 seconds. The additional (prediction-based) adaptation mechanism in GAPD enables it to keep the dropping rate below the target for much larger values of the control period, up to 70 seconds, at the cost of only slightly higher new call blocking probability. At a control period of 70 seconds, the dropping probability of GAPD is about 20 % lower than that of GAD, while the new call blocking is only 6.6% higher.

5) Adaptation Step for α: In this simulation, we vary the control step for the adaptation power scaling factor α. As the adaptation step increases, the fraction of the predicted handoff power used in guard capacity adaptation can be adjusted faster. However, changes in the dynamic guard capacity adaptation are compensated over time by the longer-term minimum guard capacity adaptation. Fig. 12 shows that the adaptation step does not have a big impact on either the handoff dropping rate or the new call blocking rate.

6) Adaptation Step for Minimum Guard Capacity of GAPD and Guard Capacity of GAD: In this simulation, we vary the control step for minimum guard capacity adaptation of GAPD and guard capacity adaptation of GAD. As expected, Fig. 13 shows that as the guard adaptation step decreases, the handoff dropping probability increases for both schemes. If the adaptation step is too small, GAD can no longer maintain the handoff dropping rate below the target level, while GAPD can do so. As the adaptation step for the minimum guard capacity is reduced, GAPD is able to make its prediction-based guard capacity adaptation more aggressive by increasing α, and is thus able to maintain the handoff dropping rate at the cost of a small increase in the new call blocking rate.

VIII. Signaling Overhead

In this section, we study the signaling cost due to handoff predictions. Generally, the control in a CDMA system is very complicated, and signaling required for managing soft handoff consists only a very small part of the total control cost of the whole system. Instead of describing all the control signals in CDMA system, we only consider the additional signaling needed for handoff predictions, with reference to that associated with soft handoff management. Unless specified otherwise, the prediction window size in this section is set to one second.

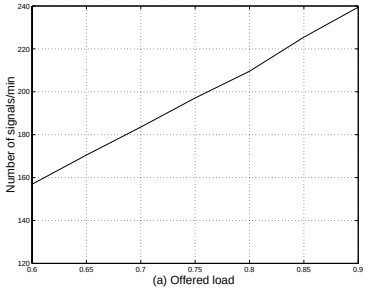

Fig. 18. Signaling overhead for handoffs with the change of offered load.

We have described the soft handoff initiation process in Section V-A. The signaling flows for handoff initiation and termination between a mobile and its serving base station and signaling in the back-haul network are shown in Fig. 14 and 15 (See [25]). Fig. 16 shows the signaling flow when a handoff request is rejected. For handoff prediction, we have introduced a soft handoff *prediction threshold* T_{PREDICT}, above which a mobile signals its serving base station for approaching the predicted cell. On the other hand, a signal for prediction cancellation needs to be sent to the serving base station if a mobile detects that handoff predicted previously will not be performed. Once a serving base station receives the prediction (or prediction cancellation) signal from a mobile, it will inform the predicted target base station so that necessary processing (e.g., guard capacity adjustment) can be done before the handoff is performed. To reduce the signaling cost, we have proposed an aggregation scheme in Section V-A.1. Fig. 17 shows the signals required due to soft handoff predictions. Note that the signaling flows for predicting a soft handoff and for canceling a handoff prediction are similar, with the pilot strength measured differently and the aggregated resource predicted increased upon predicting a handoff and reduced upon canceling a prediction.

The signal format for soft handoff prediction is very simple. Similar to soft handoff monitoring, from a mobile to its serving base station, only Pilot Strength Measurement information needs to be sent. Between a serving base station and the prediction target base station, only the total amount of power adjustment predicted for the target base station during a prediction window needs to be sent. Since prediction is only for improving handoff performance, no acknowledgment signal is required. To avoid relying on the detailed control signal format of different CDMA system standard, for evaluating the control overhead, we only consider the number of signals used. Since the control signal for handoff prediction is very simple, the overhead comparison based on the number of signals is more conservative. In what follows, we study the impact of various factors on the prediction overhead based on the signal flows shown in Fig. 14, 15, 16 and 17.

A. Impact of Prediction Threshold and Window Size

Fig. 18 shows the variation of the number of handoff signals with the offered load. As expected, the signaling overhead for handoff management increases due to the increasing number of handoffs at higher load.

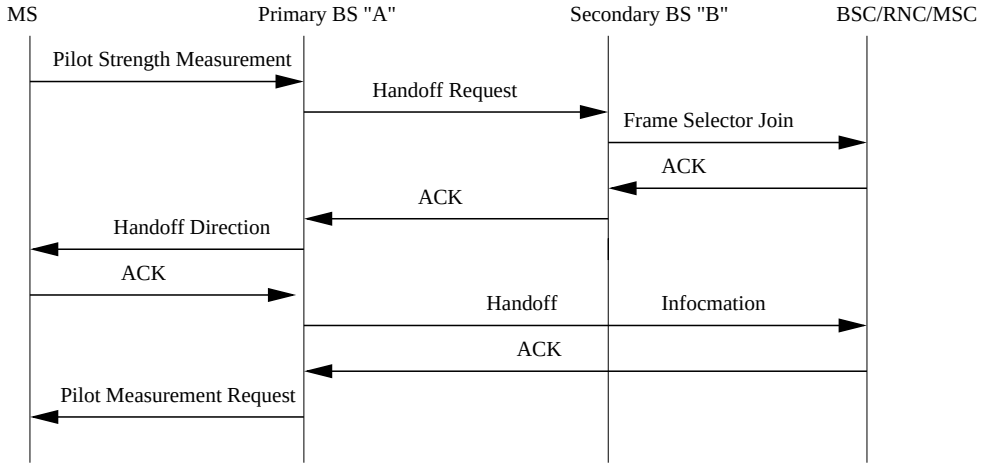

Fig. 14. Signaling for adding a soft handoff leg (between a mobile and B).

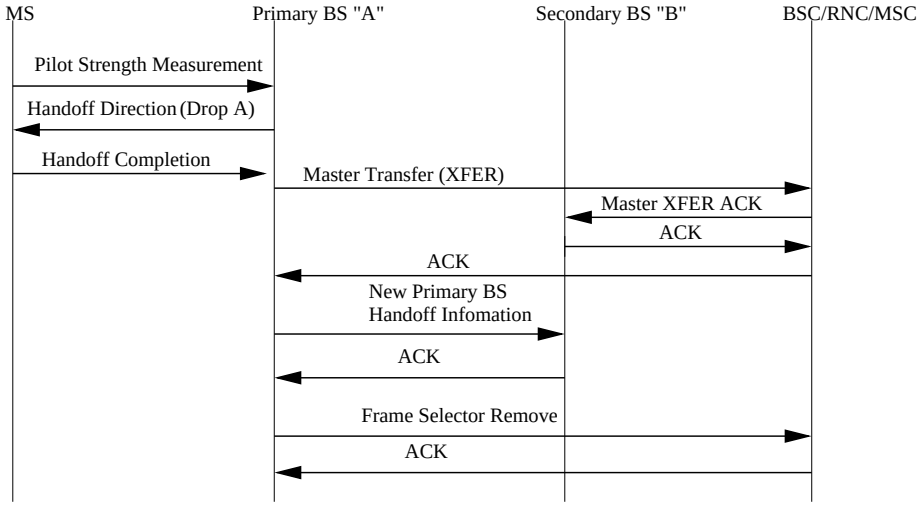

Fig. 15. Signaling for dropping a soft handoff leg (between a mobile and A) .

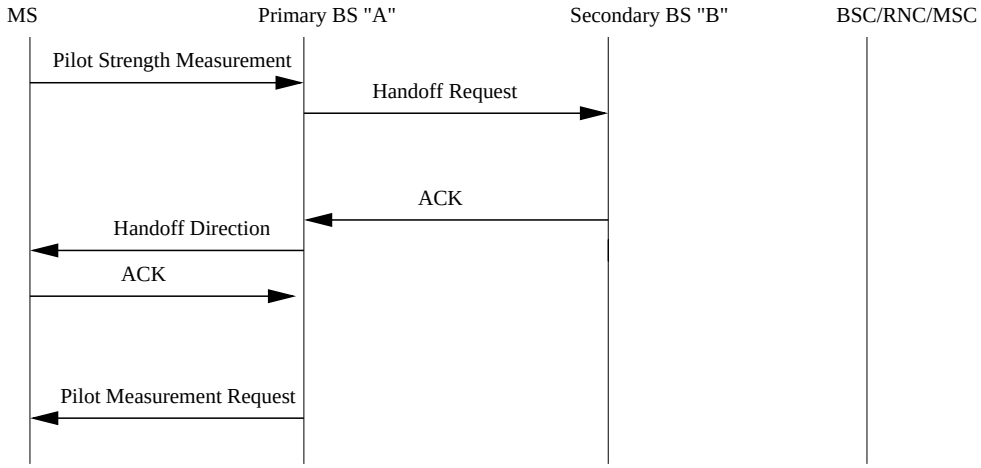

Fig. 16. Signaling when a soft handoff request is rejected (between a mobile and B).

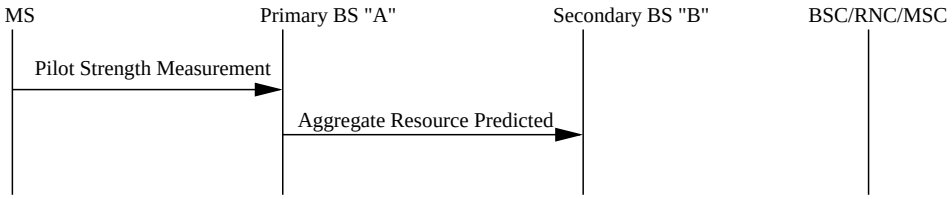

Fig. 17. Signaling for soft handoff prediction and cancellation.

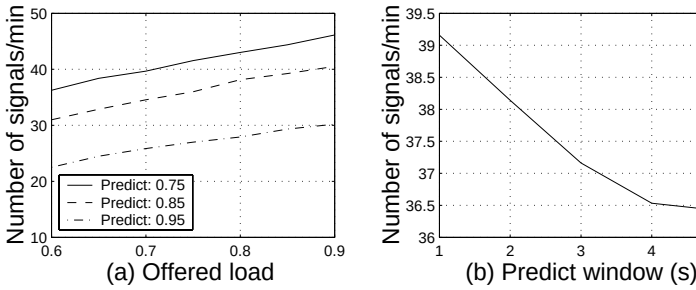

Fig. 19. Signaling overhead for handoff predictions with the change of prediction threshold (a) and size of prediction window at offered load 0.85 (b).

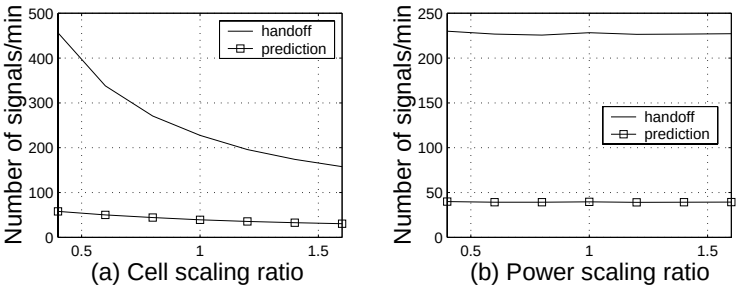

Fig. 21. Signaling overhead for handoff predictions with the variations of cell scaling ratio (a) and base station power scaling ratio (b) at offered load 0.85.

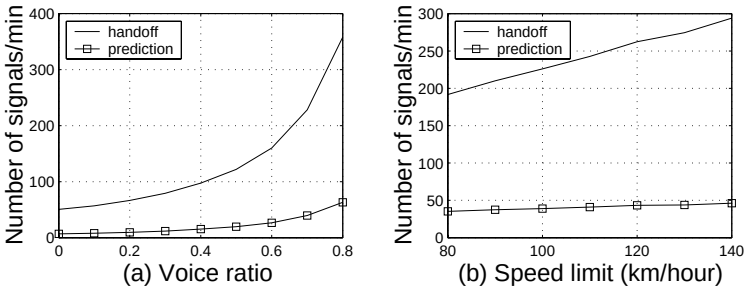

Fig. 20. Signaling overhead for handoff predictions with the variations of voice traffic ratio (a) and maximum mobile speed (b) at offered load 0.85.

Prediction threshold will impact the handoff prediction rate. A lower prediction threshold results in earlier handoff predictions, but also causes more withdrawn predictions, and hence results in higher signaling overhead as confirmed by Fig. 19 (a). Even though the signaling overhead for predictions also increases with the load, the overhead and the overhead increasing rate is much lower than that of the handoffs.

As the prediction window increases, the signaling overhead reduces due to the higher signal aggregation ratio (Fig. 19 (b)). However, the overhead reduction rate is not as big as expected within our investigated prediction window range. This is because the aggregation of prediction messages is performed by the serving base station of the mobiles, and only the messages managed by the same serving base station and targeted for the same neighboring cell can be aggregated. Since each cell can have multiple neighboring cells (e.g., With squared cell in our simulation, each cell has eight closest neighbors), the ratio for message aggregation is not only dependent on the window size but will be reduced as the number of neighbors increases.

B. Impact of Voice Ratio and Maximum Mobile Speed

Fig. 20 (a) indicates that the signaling overhead for both handoffs and predictions increase exponentially with the increase of voice ratio, while the overhead due to predictions increases much slower than that of handoffs. If the total offered load is kept the same, the increase of voice ratio will lead to the increase of the total number of active connections in a cell, and hence results in more handoffs as well as more handoff predictions and prediction withdrawn. In addition, Fig. 14, 15 and 16 show that the signaling load for a successful handoff is much higher than that of a failed handoff. Since the number of successful handoffs will increase as the voice ratio increases (due to the reduction of handoff dropping rate as shown in Fig. 3 (b)), the overall increasing rate of handoff load is much higher than the increasing rate of voice.

As expected, the increase of maximum mobile speed leads to the increase of user mobility, and therefore increases the frequency of handoffs. Correspondingly, the signaling overheads for both predictions and handoffs increase. Again, the signaling overhead for handoffs increases at higher speed (Fig. 20 (b)).

C. Impact of Cell Size and BS Power

As indicated in Section VII, the change of cell size and base station power will change the traffic patterns. An increase in cell size leads to the decrease of the frequency of handoffs. Therefore, the signaling overhead for both handoffs and predictions decreases as the cell size increases, and the overhead of handoffs decreases faster as shown in Fig. 21 (a). The variation of base station power, however, does not change the signaling overhead significantly (Fig. 21 (b)).

IX. SUMMARY

We have presented two schemes (GAD and GAPD) for managing downlink CDMA radio resources that maintain on-going call quality by minimizing call-dropping during handoffs, without over-penalizing new arrivals. In both schemes, the guard capacity of a cell is dynamically adjusted so as to maintain the handoff dropping rate at or below a target level. In the GAPD scheme, there is an additional, frequent adjustment of the guard capacity based on a novel soft handoff prediction mechanism, which aggregates prediction decisions and acts in concert with the pilot

power-based handoff detection mechanism to reduce signaling overhead. The emphasis of this work has been to develop simple and robust mechanisms that do not assume knowledge of traffic and mobility patterns, and can work over a wide range of system and control parameters.

In our simulations, we study the performance of the GAD and GAPD schemes, and also a scheme with fixed guard capacity (FG), in which the amount of guard capacity can be tuned offline with optimal parameters for a given set of system parameters and traffic conditions. The performance of FG is comparable with the performance of the GAD and GAPD schemes under the default conditions. However, FG has significantly higher handoff dropping probability (up to 23%) than GAD and GAPD as we vary the ratio of voice and data traffic, user mobility, and cell size. GAD and GAPD are both able to maintain the handoff dropping rate below the target value over a wide range of traffic, system and control parameters, with only small effects on the blocking rate of new calls. However GAPD performs significantly better than GAD under certain conditions, because its predictive control allows it to respond more quickly, and its dual control can compensate for prediction errors more effectively. GAPD is better able to control the handoff dropping rate under dynamic traffic conditions (e.g., due to bursty data traffic), and is also more robust to a wide range of control parameter values. We have also studied the signaling cost due to soft handoff predictions. Our results indicate that the additional signaling cost is much smaller than that needed to manage soft handoffs, and the overhead is well controlled under different conditions.

REFERENCES

[1] E. C. Posner and R. Gurin, "Traffic policies in cellular radio that minimize blocking of handoff calls," in *Proc. 11th ITC*, (Kyoto, Japan), pp. 294–298, Sept. 1985.

[2] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, pp. 77–92, Aug 1986.

[3] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. J. Weaver, and C. E. I. Wheatley, "On the capacity of a cellular CDMA system," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 303–312, May 1991.

[4] A. M. Viterbi and A. J. Viterbi, "Erlang capacity of a power controlled CDMA system," *IEEE Journal on Selected Area in Communications*, vol. 11, pp. 892–900, Aug 1993.

[5] Z. Liu and M. E. Zarki, "SIR-Based call admission control for DS-CDMA cellular systems," *IEEE Journal on Selected Area in Communications*, vol. 12, Apr 1994.

[6] S. Su, J. Chen, and J. Huang, "Performance analysis of soft handoff in CDMA cellular networks," *IEEE Journal on Selected Area in Communications*, vol. 14, pp. 1762–1769, Dec 1996.

[7] Y. Ma, J. Han, and K. Trivedi, "Call admission control for reducing dropped calls in code division multiple access (CDMA) cellular systems," in *Proc. of Infocom*, (Tel Aviv, Israel), Mar 2000.

[8] F. D. Priscoli and F. Sestini, "Fixed and adaptive blocking thresholds in CDMA cellular networks," *IEEE Personal Communications*, pp. 56–63, Apr 1998.

[9] J. W. Chang and D. K. Sung, "Adaptive channel reservation scheme for soft handoff in DS-CDMA cellular systems," *IEEE Transactions on Vehicular Technology*, vol. 50, pp. 341–353, Mar 2001.

[10] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 257–271, Apr 2002.

[11] A. Viterbi, *CDMA principles of spread spectrum communications.* Addison-Wesley, 1995.

[12] W. Park, Y. Kwon, and D. Lee, "A CAC scheme with code and interference limits on the forward link in CDMA cellular network," in *Proc. IEEE MILCOM'2001*, (McLean, VA), Oct 2001.

[13] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Area in Communications*, vol. 14, pp. 711–717, May 1996.

[14] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multiemdia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 1–12, Feb 1997.

[15] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks," *Computer commun. review*, vol. 28, pp. 155–166, Sept. 1998.

[16] C. Huang and R. Yates, "Downlink admission control strategies for CDMA systems in a Manhattan environment," in *Vehicular Technology Conference*, 1998.

[17] R. Vaccaro, *Digital control, a state space approach.* McGraw Hill, 1998.

[18] P. E. Mogensen, P. Eggers, C. Jensen, and J. B. Andersen, "Urban area radio propagation measurements at 955 and 1845 Mhz for small and micro cells," in *IEEE Global Commun. Conf.*, (Phonix, AZ), pp. 1297–1302, Dec 1991.

[19] M. Gudmundson, "Correlation moble for shadow fading in mobile radio systems," *Electron. Lett*, vol. 27, pp. 2145–2146, Nov 1991.

[20] M. Gudmundson, "Analysis of handover algorithms," in *IEEE Veh. Technol. Conf.*, (Saint Louis, MO), pp. 537–541, May 1991.

[21] Y.-B. Lin and V. Mak, "Eliminating the boundary effect of a large-scale personal communication service network simulation," *ACM Transactions on Modeling and Computer Simulation*, vol. 1, no. 2, pp. 165–190, 1994.

[22] W. E. A. 3rd Generation Partnership Project 2, "1xEV-DV evaluation methodology - addendum (v6)," 2001.

[23] 3rd Generation Partnership Project, "Gommon test environments for user equipment," *3GPP TS 34.108 V3.8.0*, Jun 2002.

[24] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," *ACM Baltzer Wireless Networks Journal*, vol. 3, no. 1, pp. 29–41, 1997.

[25] V. Garg, *Wireless network evolution, 2G to 3G.* Prentice Hall, 2002.

PLACE PHOTO HERE

Xin Wang received her BS and MS degrees in telecommunications engineering and wireless communications engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1990 and 1993, respectively, and her PhD degree in electrical engineering from Columbia University, New York, NY, in 2001. From 2001 to 2003, she was a Member of Technical Staff in the area of mobile and wireless networking at Bell Labs Research, Lucent Technologies, New Jersey. She is currently an Assistant Professor in the department of Computer Science and Engineering of the State University of New York at Buffalo, Buffalo, New York. Her research interests include modeling and analysis of mobile and wireless networks, integrated network infrastructure design and performance enhancement across network layers, applications and heterogeneous networks, network and mobility management, QoS, signaling and control, as well as adaptive network services and applications.

PLACE PHOTO HERE

Ram Ramjee received his B.Tech in Computer Science and Engineering from the Indian Institute of Technology, Madras, and his M.S. and Ph.D. in Computer Science from University of Massachusetts, Amherst. He has been at Bell Labs, Lucent Technologies since 1996, where he is currently a Distinguished Member of Technical Staff. His research interests include protocols, architecture, and performance issues in wireless and high speed networks. He is also an adjunct faculty at the Electrical Engineering Department of Columbia University where he teaches graduate courses in wireless networks. Dr. Ramjee is an area editor of ACM Mobile Communications Review, an associate editor of IEEE Transactions on Mobile Computing and a technical editor of IEEE Wireless Communications Magazine. He has published over 30 papers in premier conferences and journals and holds 11 U.S. patents.

PLACE PHOTO HERE

Harish Viswanathan received the B. Tech. degree from the Department of Electrical Engineering, Indian Institute of Technology, Madras, India in 1992 and the M.S. and Ph.D. degrees from the School of Electrical Engineering, Cornell University, Ithaca, NY in 1995 and 1997, respectively. He is presently with Lucent Technologies Bell Labs, Murray Hill, NJ. His research interests include information theory, communication theory, wireless networks and signal processing.