Congestion Control Policies for IP-based CDMA Radio Access Networks

Sneha K. Kasera
School of Computing
University of Utah
Salt Lake City, UT 84112
Email: kasera@cs.utah.edu

Ramachandran Ramjee, Sandra Thuel
Bell Laboratories
Lucent Technologies
Holmdel, NJ 07733
{ramjee, thuel}@bell-labs.com

Xin Wang
Computer Science and Engineering
SUNY at Buffalo
NY 14260
xwang8@cse.buffalo.edu

Abstract

As CDMA-based cellular networks mature, the current point-to-point links used in connecting base stations to network controllers will evolve to an IP-based Radio Access Network (RAN) for reasons of lower cost due to statistical multiplexing gains, better scalability and reliability, and the projected growth in data applications. In this paper, we study the impact of congestion in a best-effort IP RAN on CDMA cellular voice networks. We propose and evaluate three congestion control mechanisms, *admission control*, *diversity control*, and *router control*, to maximize network capacity while maintaining good voice quality. We first propose two new enhancements to CDMA call admission control that consider a unified view of both IP RAN and air interface resources. Next, we introduce a novel technique called diversity control that exploits the soft-handoff feature of CDMA networks and drops selected frames belonging to multiple soft-handoff legs to gracefully degrade voice quality during congestion. Finally, we study the impact of router control where an active queue management technique is used to reduce delay and minimize correlated losses. Using simulations of a large mobile network, we show that the three different control mechanisms can help gracefully manage 10-40% congestion overload in the IP RAN.

I. Introduction

Cellular wireless networks have become an indispensable part of the communication infrastructure. CDMA is an important air-interface technology for cellular wireless networks. It has been selected for implementation in both the North American and European 3G standards. Traditionally, in these wireless access networks, the base stations are connected to radio network controllers or base station controllers by point-to-point (usually T1/E1) links. These links, also called backhaul links, are expensive and their use imposes an on-going cost on the service providers. In such networks, reliability comes at high price: by replication of links or controllers. As CDMA-based cellular networks mature, the current point-to-point links will evolve to an IP-based Radio Access Network (RAN). Replacing the point-to-point links between the base stations and the radio network controllers by an IP RAN has the following advantages:

- *Cost* - Point-to-point links including T1 links are expensive and cannot be shared. An IP network will benefit from statistical multiplexing gains and could be shared with other wireless and wireline applications.
- *Scalability and Reliability* - Replacing point-to-point links by a distributed IP network will provide alternate paths to more than one network controller, thereby improving reliability

and scalability. For example, it is shown in [1] that adding a selected few paths between base stations and network controllers results in majority of the gains in resiliency to failures.

- *Data Applications* - Increasingly, a large number of IP-based "data applications" including web browsing, email, streaming and packetized voice (voice over IP) are being offered in wireless networks. Hence wireless access networks must support IP traffic. An IP RAN efficiently addresses this eventuality.

While the use of an IP RAN results in the above advantages, mechanisms must be designed to control IP RAN congestion. Congestion occurs when the offered traffic exceeds the engineered IP RAN capacity. There are essentially three approaches to control and avoid congestion. First, the network can be over-provisioned or peak-provisioned so that congestion never occurs. Although simple, this is not a practical solution because access network bandwidth is still very expensive compared to core network bandwidth. According to the financial firm Friedman Billings Ramsey, it costs 14,000 times more to carry a bit of information one mile on a local network versus a long distance network [2]. Second, one can reserve resources in the access network. While several research efforts have focused on this problem (e.g., [3], [4]), inaccurate resource estimation due to dynamic load patterns and/or mobility, variations in the wireless environment, and the wide range of application characteristics makes it a very hard problem to solve. Besides, even though various reservation schemes have been proposed and implemented in routers, these approaches are yet to be widely deployed in current IP networks. The third approach is to assume a best-effort IP RAN and use properly designed policies to control and avoid congestion. This is the focus of our paper.

We study the impact of congestion in a best-effort IP RAN on CDMA cellular voice networks. Congestion introduces loss and delay jitter in the user traffic. Uncontrolled loss and delay jitter could drastically reduce the voice quality. Therefore congestion control techniques are essential in maintaining good voice quality. We focus on the voice application for two reasons: a) current cellular networks are predominantly used for voice transmission; and b) voice has tighter delay and loss requirements than data (where retransmission is an option).

We propose and evaluate three congestion control mechanisms to maximize network capacity while maintaining good voice quality: *admission control, diversity control,* and *router control*. Call admission control in current CDMA cellular voice networks is restricted to controlling the usage of air interface resources. We first propose two new enhancements to CDMA call admission control that consider a unified view of both IP RAN and air interface resources to adequately match the number of voice users to the engineered capacity. The principle underlying both schemes is regulation of the IP RAN load by adjusting the admission control criterion at the air-interface. Next, we introduce a novel technique called diversity control that exploits one of the unique features of CDMA, namely macro-diversity or soft-handoff. A cellular user in soft-handoff transmits and listens to multiple base stations simultaneously. During IP RAN congestion, our diversity control technique allows dropping of selected frames belonging to potentially redundant soft-handoff legs, thereby reducing congestion gracefully while maintaining adequate voice quality. Last, we study the impact of router control in the form of active queue management [5]. IP routers using a drop tail mechanism during congestion could produce high delays and bursty losses resulting in poor voice quality. Use of active queue management at the routers reduces delays and loss correlation, thereby improving voice quality during congestion. Using simulations of a large mobile network, we evaluate the behavior of the three different control mechanisms and show how these techniques help manage congestion in the IP RAN gracefully. To our knowledge this is the first paper to consider the impact of congestion in IP RAN on CDMA network performance.

The rest of the paper is structured as follows. In Section II, we present an overview of the problem. In Section III, we present two call admission control algorithms that take into account both the air interface and IP RAN resources to regulate incoming traffic. In Section IV, we present our diversity control techniques that selectively drop soft-handoff legs to control congestion. In Section V, we present our router control technique using active queue management. In Section VI, we present our simulation results demonstrating the benefits of all three congestion control techniques. Section VII contains the impact of our policies on network provisioning. Section VIII contains future directions for enhancing our research and finally in Section IX, we present our conclusions.

II. Problem Setting

In this section we describe the components of a CDMA wireless access network that uses an IP RAN and identify our problem space. Figure 1 shows a wireless access network with mobile devices communicating with base stations over wireless links. The base stations communicate with the rest of the voice or data network through the access network controllers (ANCs) (also

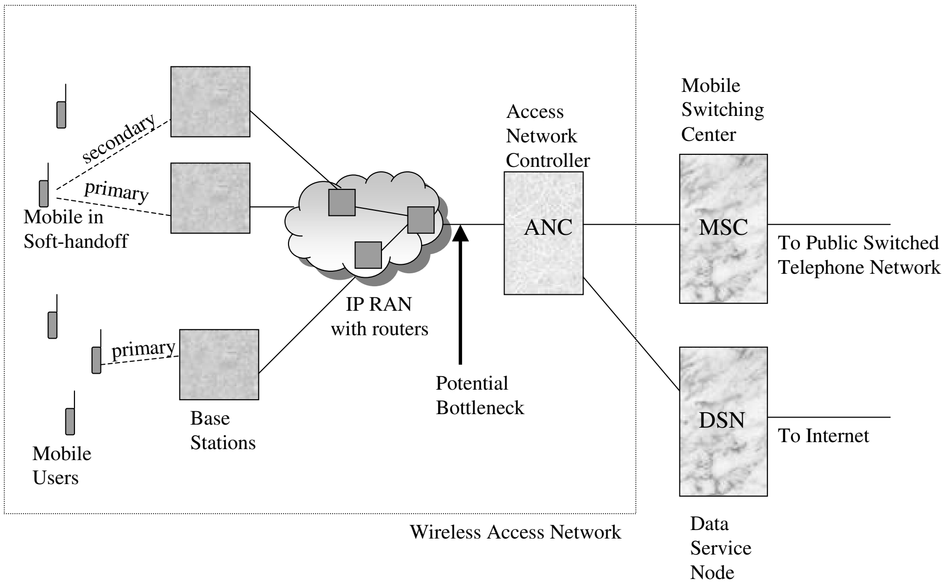


Fig. 1. CDMA Wireless Access Network with IP RAN

called Radio Network Controller, RNC, in 3G UMTS [6], and Base Station Controller, BSC, in CDMA2000 [7]). Note that this part of the network is common to both wireless voice and data traffic. The network separates only beyond the ANC where voice frames are forwarded to the MSC (PSTN) and data frames are forwarded to the Data Service Nodes (Internet). Each base station typically communicates with hundred or more mobiles and each ANC typically controls several tens of base stations. An ANC performs two main wireless functions, *frame selection* and *reverse outer loop power control*. Frame selection exploits one of the key properties of a CDMA network, namely, soft-handoff. In soft-handoff, a mobile communicates with more than one base station simultaneously. Soft-handoff helps reduce interference on the wireless link thereby increasing CDMA capacity. When in soft-handoff, a mobile receives multiple frames in the downlink direction (also called forward link) and combines them to construct a single voice frame. In the uplink direction (also called the reverse link) the ANC receives multiple frames from the mobile. It performs the frame selection function which involves selecting the frame with the best quality among the ones it receives. If the frames from all the different legs of a call in soft-handoff call do not arrive within a preset time interval (20ms in the case of CDMA2000), the ANC forwards the current best frame to the network. In other words, a late frame is treated as if it were a dropped frame and thus, controlling delay in the access network is extremely important.

In addition to frame selection, the ANC also performs reverse outer loop power control, in which it sets target signal to noise ratio (E_b/I_t) for each mobile at each base station. The target (E_b/I_t) is set such that the target frame error rate for the flow after frame selection is maintained below a preset limit (such as 1%). Each base station receives the target E_b/I_t for each mobile device from the ANC and instructs the devices to increase or decrease their transmission power.

We now describe the operation of the IP network between the base station and the ANC. On receiving voice frames from different mobiles, a base station aggregates several voice frames into an IP packet and sends them out towards the ANC. Voice frames are typically only few tens of bytes in length. Their aggregation helps in reducing IP header overhead. Voice frames belonging to the different legs of soft-handoff are transmitted by different base stations and hence arrive at the ANC on different IP packets. On receiving IP packets from the base stations, an ANC demultiplexes the voice frames and performs frame selection and outer loop power control functions and forwards the best voice frame uplink. Voice frames also contain power information that is used by the ANC for outer loop power control. Therefore packet loss, and hence loss of voice frames, due to RAN congestion could result in imperfect outer loop power control. This could cause the power consumption in a cell to be higher than its expected value and thereby reduce the overall capacity. Thus, controlling loss in the access network is very important[1].

The link leading to the ANC is likely to become a bottleneck during congestion because this link carries the aggregate traffic of several tens of base stations. While the link will be engineered to take into account the statistical multiplexing gains of this aggregation, offered traffic could temporarily exceed the engineered capacity of the link due to hot spots or other reasons. We require mechanisms to respond to these temporary congestion events in a graceful manner. This congestion response is the focus of the rest of the paper. While we focus only on the reverse path from base stations towards the ANC, most of the techniques described here can be applied to the forward path as well.

In summary, in this paper, we will study the effect of a single bottleneck link in the common path from the base stations towards their ANC for aggregated CDMA voice traffic arriving as IP packets and propose solutions to control congestion gracefully.

[1]Although, we do not study the imperfections in outer loop power control due to loss of power control information, our congestion control mechanisms do control IP RAN loss.

III. Call Admission Control

CDMA systems are typically interference limited and rely on the processing gain[2] to be able to operate at a low signal-to-interference ratio. In order to minimize the interference, the mechanisms adopted by the CDMA systems control the power emitted on each channel (in either direction) to keep the signal-to-interference ratio (SIR) at a receiver at a target value. When the power limit of a base station is reached, the SIR can no longer be maintained at the target level, and calls serviced by the base station are blocked or dropped. Hence, call admission is closely tied to power control. The call admission control in current CDMA systems is restricted to controlling the usage of air interface resources. The point-to-point access links between the base stations and the ANC are expected to be loss-free. The presence of an IP RAN adds a new dimension to this process since a lossy or congested best-effort IP RAN could result in high losses and delays, thereby reducing the voice quality. In this section, we describe two schemes to enhance the air interface call admission control algorithms that also take the IP RAN loss into consideration when deciding on admitting new calls. We first present the call admission control mechanism that considers the air interface resources only and then describe our two enhancements to admission control that consider a unified view of both the IP RAN and the air interface resources.

A. Admission Control at Air Interface (CAC)

We consider a single base station that is serving a geographical area called a *cell*. The relationship between received power at the base station and cell load, and an associated admission control threshold in terms of the allowed received power is described as follows. The total interference at a mobile includes the interference from the mobiles of the same cell and the neighboring cells, and from thermal noise. For a CDMA cell with M signals, one for each mobile, there are $M-1$ interferers from the same cell. If we assume perfect power control on the reverse link and that the signals transmitted from all the mobiles arrive at the base station with the same received power S, the ratio of signal bit energy E_b to total interference and thermal noise power spectral density I_t can be expressed as [8]

[2]Processing gain is defined as the ratio of transmitted signal bandwidth to the data signal bandwidth (information rate).

$$\frac{E_b}{I_t} = B_w \frac{S/R}{(M-1)v_f S(1+f_I) + N_0 B_w}, \tag{1}$$

where v_f is the channel activity factor, f_I represents the interference due to other cells, R represents the mobile transmission rate, B_w is the spreading bandwidth, and N_0 is the thermal noise power spectral density. Given a target E_b/I_t, with the processing gain represented as $G_p = B_w/R$, the number of mobiles M a cell can admit is a function of the received power S:

$$M = 1 + G_p \frac{1}{(E_b/I_t)v_f(1+f_I)} - \frac{N_0 B_w}{S v_f (1+f_I)}. \tag{2}$$

The total received power can be represented as $P_{\text{total}} = v_f(1+f_I)SM + N_o B_w$. Using this expression to obtain S as a function of the total received power and inserting it into Equation 2, we can determine the number of mobiles M a cell can admit when the total received signal power is P_{total} by the following equation:

$$M = \frac{1 + G_p \frac{1}{(E_b/I_t)v_f(1+f_I)}}{1 + \frac{N_o B_w}{P_{\text{total}} - N_o B_w}}. \tag{3}$$

When the total received power is restricted to P^{max}, the maximum number of connections that can be admitted by a cell is obtained from Equation 3 by replacing P_{total} with P^{max}. This forms the basis of the air interface admission control algorithm. Note that this algorithm does not take the quality of the IP RAN into account. In the next subsection, we enhance this algorithm to also consider access network load.

We end this subsection by describing the relationship between the total received power and the cell load. The cell load, represented by ρ, is simply determined by the ratio M/M^{pole} , where M^{pole} is the maximum cell capacity or the *pole capacity*. The maximum cell capacity, M^{pole}, can be obtained by setting $P_{\text{total}} \to \infty$. Therefore, $M^{\text{pole}} = 1 + G_p \frac{1}{(E_b/I_t)v_f(1+f_I)}$. Thus, the total received power and cell load are related by the following equation:

$$\frac{P_{\text{total}}}{N_o B_w} = \frac{1}{1-\rho}. \tag{4}$$

Equation 4 is used for call admission control enhancements in the next subsection.

B. Enhanced admission control

In this section, we propose *two enhanced admission control algorithms which respond to changes in the access network load by regulating the admission control criteria at the air interface, thereby indirectly adjusting the load entering the access network*. Instead of using the maximum allowable received power, P^{max}, to perform admission control, each base station uses a variable received power threshold, P^{adm}, that is periodically calculated by the ANC. The calculations are based on the air interface capacity and cell load, as well as IP RAN loss rate.

Our enhanced admission control schemes are called *Maximum-Power-based* call admission control (CAC-MPC) and *Usage-based* call admission control (CAC-UC). Both schemes share the same basic approach: the measured packet loss rate is used to calculate a power scaling factor α, which is then used to scale the admission control threshold, P^{adm}, for the air interface, thereby throttling incoming traffic and reducing congestion. We first present the general methodology that we use to adaptively quantify the loss-rate and then present our two admission control schemes.

1) General Methodology: The purpose of admission control in the IP RAN is to keep the packet loss rate within a target level, so that the quality of voice transmission can be maintained. We use a feedback control strategy based on a constrained integral control law [9]. In this scheme, the packet transmission between base stations and the ANC is monitored by the ANC. Since all the uplink traffic will go through the ANC, the ANC can detect the packet loss in the IP RAN. The power scaling factor α is calculated periodically by comparing the monitored loss rate with a target loss rate. With the measured loss rate of the IP RAN represented as L and the target loss rate as L^*, the scaling factor α for a period n is calculated as:

$$\alpha_n = \min\{\max\{\alpha_{\text{min}}, \alpha_{\text{n-1}} - \sigma \frac{L - L^*}{L^*}\}, \alpha_{\text{max}}\}, \tag{5}$$

where the parameter σ controls the adaptation speed of the scaling factor, and α_{min} and α_{max} are the minimum and maximum values allowed of the scaling factor.

For an integral controller such as ours, higher σ leads to a faster response. However, higher values of σ can cause larger oscillation and even instabilities. Also, if $|\sigma \frac{L-L^*}{L^*}|$ is too large, α will be set to a value that could scale up P^{adm} causing undesired air interface control. In order to obtain tighter control, we constrain α within the range $[\alpha_{\text{min}}, \alpha_{\text{max}}]^3$.

2) Maximum-Power-based Control (CAC-MPC): In our Maximum-Power-based control, the admission control threshold power P^{adm} is obtained by scaling down the maximum allowed receiving power P^{max} by α, when the measured loss rate exceeds the target value:

$$P^{\text{adm}} = \alpha P^{\text{max}}. \tag{6}$$

Combining Equation 3, 4 and 6, the total number of mobiles a cell can admit is given by:

$$M^{\text{adm}} = \frac{M^{\text{pole}}}{1 + \frac{N_o B_w}{\alpha P^{\text{max}} - N_0 B_w}} = M^{\text{pole}}(1 - \frac{1 - \rho^{\text{max}}}{\alpha}), \tag{7}$$

where ρ^{max} is the maximum load allowed corresponding to the maximum allowed received power P^{max}. Hence, the total load allowed in a cell is $\rho^{\text{adm}} = 1 - \frac{1-\rho^{\text{max}}}{\alpha}$. Since the load cannot be negative, the range $[\alpha_{\text{min}}, \alpha_{\text{max}}]$ for the scaling factor is $[1 - \rho^{\text{max}}, 1]$.

Note that Equation 7 can set the allowed cell load to zero. We use a safe-guard mechanism to set a lower limit on the threshold power according to a predetermined planned load. This also ensures fairness since only cells with higher than planned load will be throttled back when the IP RAN is congested. In this case, we set

$$P^{\text{adm}} = \max\{P^{\text{plan}}, \alpha P^{\text{max}}\}, \tag{8}$$

where P^{adm} is the planned limit on threshold power.

3) Usage-based Control (CAC-UC): In Usage-based control, the admission control threshold power P^{adm} at a cell is calculated based on the current load of the cell as well as the IP RAN loss rate. The control principle is that the cells with higher load should bear a larger share of the total load reduction. When the loss rate in the IP RAN is higher than the target value, P^{adm} is obtained by scaling down the current received power P^{cur} of a cell, with the scaling factor α obtained using Equation 5. Since this can result in a much smaller power threshold relative to the maximum allowable power P^{max}, the power threshold should not be set back to P^{max} immediately after the loss is restored to the target level. P^{adm} is scaled up progressively, using

[3] With the range constraints, care must be taken that α 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 α can be prevented from being absorbed into the extreme states if $\sigma < \frac{\alpha_{\text{max}} - \alpha_{\text{min}}}{\epsilon}$.

a scaling factor α greater than one. The usage-based power control algorithm is described below:

$$P^{\text{adm}} = \begin{cases} \max\{P^{\text{plan}}, \alpha \frac{\rho^{\text{plan}}}{\rho^{\text{cur}}} P^{\text{cur}}\} & \text{if} \quad \alpha < 1 \text{ or } L > L^*. \\ \min\{\alpha P^{\text{adm}}, P^{\text{max}}\}, & \text{otherwise}. \end{cases}$$

In the above equation, ρ^{plan} and ρ^{cur} represent separately the planned cell load and current cell load. In order to ensure fairness to a lightly loaded cell when power is being scaled down, and also to avoid an excessively slow response following a period of low loss (when the initial value of α is greater than one), the load ratio of a cell (relative to the planned load) is used as an additional scaling factor. The use of the two factors together, instead of using either the load ratio or the scaling factor α separately, results in faster response to congestion and also keeps the loss close to the target level. Furthermore, as with the maximum power based scheme, we add a safe-guard mechanism to place a lower limit on the threshold power, and prevent penalizing cells operating below the planned load.

Detailed performance evaluation of these two admission control algorithms is presented in Section VI.

IV. Diversity Control

Recall that CDMA supports macro-diversity or soft-handoff (SHO) in which a mobile user transmits and listens to multiple base stations simultaneously. The uplink packets received at the multiple base stations are forwarded to the ANC for frame selection and the best frame is forwarded into the wired network. Diversity allows mobile users to smoothly transition their connections from one cell to the next without loosing connectivity or suffering service degradations as is typical in hard-handoff scenarios.

In typical cellular networks, mobiles can be in SHO with up to six base stations at once with one primary leg and up to five secondary legs. These multiple legs constitute what is known as the mobile's active set. Field measurements indicate that CDMA voice users tend to operate in SHO mode almost half of the time, with an average of 1.5 legs per call. However, field experience also suggests that the time interval in which a user actually needs more than one leg tends to be relatively short and that the primary leg is typically the best quality leg. This indicates that there may be extended periods of time where a user has more than one leg but the primary leg might have been sufficient, resulting in some unnecessary redundant traffic in the wireless access network. Clearly, how often this occurs depends heavily on the aggressiveness or conservativeness of the active set management policy and relative differences in voice quality across the diverse legs due to the air link quality. For wireless access networks where the above field observations hold true, one could expect non-primary legs to be redundant often. This redundancy can be exploited using a technique we call diversity control.

Diversity control selectively discards uplink voice radio frames from potentially redundant secondary legs for some mobile users at the base stations in such a way that the voice quality is not noticeably degraded while reducing traffic in the IP RAN to manage congestion. Key challenges in diversity control are the service degradation and restoration policies, which dictate how users are chosen for and freed from diversity control, respectively.

We propose and evaluate two diversity control policies, referred to as the service-degradation policy (SDP) and the frame-discard policy (FDP). SDP implements a binary service-level model for each mobile user, wherein the cells tag SHO users as being in a degraded or non-degraded state; mobiles with degraded service must discard all their secondary legs. FDP, on the other hand, drops secondary legs for SHO users randomly, each time a packet is sent by the cell.

In both policies, each cell periodically sets a diversity control target based on the estimated packet loss rate in the IP RAN[4]. The diversity control target for the n^{th} control period is a drop probability denoted by P_n, computed by a first-order autoregressive model given as:

$$P_n = \max\{\min\{0, P_{\text{n-1}} + \sigma \frac{L - L^*}{L^*}\}, 1\}, \tag{9}$$

where σ, L, and L^* represent the adaptation speed factor, measured loss rate and target loss rate, respectively, as defined for the admission control scaling factor in Equation 5. Although this equation is very similar to Eq. 5, the drop probability is limited to a value range between 0 and 1 and the variance term to adjust for the measured loss rate error uses a positive value for the adaptation speed parameter σ. This ensures that the drop probability increases as the measured loss rate increases above the target and that it decreases otherwise. σ controls how smoothly drop probabilities change, leading to higher P_n variances as σ increases. While higher values of

[4]Packet loss rate estimate is conveyed to the base stations by the ANC.

σ lead to faster reaction to changes in rate loss, the higher P_n variances may sometimes lead to a relatively high mean drop probability. It is possible to obtain fast reactivity and a low mean drop probability without introducing high variance by adjusting the drop probability in a non-linear (e.g., exponential) fashion. However, a discussion of this topic is beyond the scope of this paper.

Once the drop probability is computed, the SDP and FDP policies differ on how it is used to discard the traffic for non-primary legs. In SDP, the drop probability represents a service degradation threshold for each cell. This threshold indicates what percentage of SHO users need to be in a degraded service state for the duration of the control period and users are randomly selected for degradation. In FDP, the drop probability represents the frame drop probability for each cell. Every time a cell assembles an aggregated IP packet to send to the ANC, it randomly drops secondary voice frames according to the frame drop probability. Note that the key difference between SDP and FDP is that SDP does not consider the number of SHO legs each user has in its selection process and it forces the discard of all of a degraded user's non-primary legs.

Restoration is the counterpart to the degradation policies. In SDP, a degraded user is restored at a control period arrival when its service-level status is toggled to non-degraded. As a result, a diversity control period constitutes the minimum amount of time a degraded user must wait prior to being restored. In FDP, on the other hand, restorations occur on a per-leg, per packet-basis, as previously discarded legs are re-enabled to transmit. Consequently, FDP exerts a finer grain control by randomizing the selection of frames to be discarded on every single packet transmission, which occurs in the millisecond timescale rather than on the second timescale where the SDP diversity control period operates.

Since SDP explicitly affects per-user state, diversity control comes into play on call handoff events, where a change in a user's active set needs to be managed. For instance, if a degraded user looses all its non-primary legs on a handoff, should this user be restored? If, on the other hand, the handoff would add more legs should the new legs be degraded? For simplicity, we assume that degraded users are restored if they loose all their secondary legs and that degraded users are prohibited from adding any more legs due to handoff. FDP would be affected by call handoffs if they were to trigger the re-computation of cell frame dropping probabilities but for simplicity, we make no such assumption.

Diversity control, by itself, might not be sufficient to recover from congestion, particularly in situations where the instantaneous amount of diversity to be exploited is low or the congestion is very high. As a result, it is advantageous to apply diversity control in conjunction with load control, to ensure that they jointly achieve the goal of maintaining good call quality (by minimizing the average user frame error rate, FER).

In Section VI, we evaluate the potential benefits of diversity control and discuss the impact of a joint diversity- and load-control on the CDMA network.

V. Router Control Using Active Queue Management

When congested, a router typically follows a *drop tail* policy where packets arriving at the router are queued as long as there is space to buffer them and dropped otherwise. Even though it is simple to implement, the drop tail policy poses two important problems.

- If the buffer size is large, the queuing delays can be very high. Given the frame selection deadline of 20ms for CDMA2000 networks at the ANC, these delays would cause the transmitted packets to arrive at the ANC too late to be of any use. However, if the routers could provide the required delay bound then this problem does not exist.
- The packet drops tend to be bursty. In a CDMA system frames from calls in soft-handoff arrive in different IP packets at the IP bottleneck link at about the same instant. If the loss is bursty, the multiple frames associated with different legs of a soft-handoff call could be dropped at the same time. When the number of calls in soft-handoff is large and multiple frames associated with different legs of the soft-handoff are good (meaning that it will suffice even if any one of these frames is received at the ANC), the bursty loss would increase the probability of frame error.

Active queue management (AQM) [5] is a form of router control that attempts to provide congestion control by monitoring the congestion state of a router queue and pro-actively dropping packets before the buffers become full and queuing delays become too high. Some of the AQM policies (for e.g., [10]) drop packets with a certain probability to avoid bursty loss.

It appears that *the random dropping and tight delay features of AQM policies are an excellent fit for the unique delay deadline and soft-handoff requirements of a CDMA access network.* For example, an AQM policy could reduce the queuing delays in the routers so that the voice frames could be received at the ANC by the deadline. An AQM policy could also help prevent all the frames associated with different legs of a soft-handoff call from being dropped. It should be noted that proposing or designing an efficient AQM policy is not within the scope of this paper. Rather, we want to study the impact of router control using any reasonable AQM policy.

In order to examine the benefits of AQM, we study the use a variant of the RED AQM policy called SRED, first proposed in the context of signaling overload control in [11]. This policy uses a timer-based approach in which the bottleneck queue length is measured and an exponentially weighted moving average, Q_n, is computed every T time units. Q_n is compared to a minimum and a maximum queue length, denoted Q_{min} and Q_{max} respectively, and a fraction of packets that could be allowed into the queue is computed. The function for determining the fraction allowed in the (n+1)'th interval, f_{n+1} is described below:

$$f_{n+1} = \begin{cases} f_{\min}, & Q_n \geq Q_{\max} \\ 1, & Q_n \leq Q_{min} \\ \max\left(f_{\min}, \frac{Q_{\max}-Q_n}{Q_{\max}-Q_{\min}}\right), & \text{otherwise} \end{cases} .$$

f_{min} is the minimum fraction allowed in a given time interval. It is set to a very small value. Once the fraction allowed is computed, a deterministic algorithm, first proposed by Hajek [12] and shown to perform well in [11], is used to drop incoming packets. Hajek's deterministic algorithm is described below. A variable r is first initialized to 0, then the forward/drop decision procedure described below is used.

$r := r + f_n$.
If $r \geq 1$
    $r := r - 1$
    forward packet
else drop packet.

In the next section, we evaluate the performance impact of all three congestion control techniques on the IP RAN.

VI. Performance Study

In this section, we study the performance of our congestion control policies in a CDMA IP RAN with a single bottleneck. We use a custom designed simulator. The simulator consists of two parts. The first part simulates user mobility, call generation, call termination and soft-handoffs. The second part of the simulator uses the traces generated from the first part and simulates the

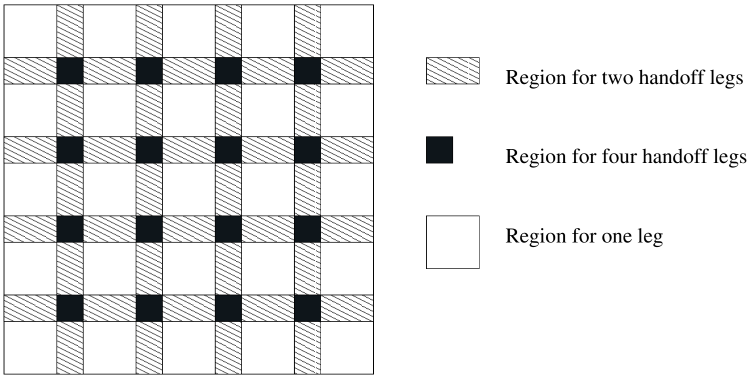


Fig. 2. Soft-Handoff Regions for a 5 X 5 cellular network

generation and aggregation of voice frames, IP packet transmission through a single bottleneck link, and frame selection at the ANC.

A. Simulation Model

We now describe our model and the various parameters that we use in our simulation study.

1) Mobility Model: In order to simulate a very large PCS network we use a wrap around technology as advocated by the authors in [13]. This approach eliminates the boundary effects in an un-wrapped topology. We simulate our PCS network using a wrapped mesh topology with the number of cells ranging from 25 to 64. In order to simulate soft-handoffs, we assume that cells are rectangular and all neighboring cells (top, left, bottom, right) overlap. This results in regions in the network with one, two, and four soft-handoff legs as shown in Figure VI-A.1. A similar approach using rectangular cells but with overlap between two neighboring cells only (resulting in regions with one and two soft-handoff legs) has been used in [14] and [15]. We assume that the mobile user spends an exponential amount of time in each region with average residence times in regions with one, two and four legs distributed with the ratio 10:4:2.5. The movement of the mobile users is based on the two-dimensional random walk model used in [16]. In this model, the mobile users move to one of their neighboring cells with equal probability. In our simulation, while the user always moves from a one-leg to a two-leg region and from a four-leg to a two-leg region, we bias the movements from two leg regions, such that there is a 80%

chance of going to a one leg region and only a 20% chance of going to a four-leg region. This movement behavior coupled with the average residence times in the different regions gives us an average active soft-handoff leg value of 1.44 . This is between the values of 1.2 and 1.7 average legs as presented in [17] and within the range evaluated for a nominal propagation environment in [18]. Mobiles are generated randomly and uniformly across the cells. In order to load a cell, we generate call arrivals for each mobile user. The call rate varies with the required simulation load of the cell. The arrival rate per user is determined as follows.

We start with the assumption that call arrivals at each user are based on a Poisson process with a mean arrival rate λ and the call holding times are exponentially distributed with a mean value $1/\mu$. Now, if the call holding times were always smaller than the call inter-arrival times then λ could be simply determined as a function of the required cell load and the number of users. Interestingly, even if we pick the mean arrival rate λ to be less than the mean call holding rate μ, inter-arrival times can be smaller than call holding times. In such situations, a call arrival event does not result in an actual call. In other words, a mobile user initiating a call at time t_o cannot make another call sooner than t_o plus the call holding time. In order to take the call holding time into account in generating the required call load we use the following analysis.

Let X be a random variable that represents the call holding time. Let Y be another random variable that represents the time between when a call is over until the time the next call is made. Therefore, the interval between two calls (T) is given by

$$T = X + Y.$$

As noted earlier, $E[X] = 1/\mu$. Also, since the call arrival events are generated from a Poisson distribution, the inter-arrival time is exponentially distributed. Now, due to the memoryless property of exponential distributions, the time of a call arrival starting at an instance t_o is independent of the time of a call arrival starting at an instance $t_o + X$. This implies that $E[Y] = 1/\lambda$. Therefore,

$$\begin{aligned} E[T] &= E[X] + E[Y] \\ &= 1/\mu + 1/\lambda \end{aligned}$$

The effective mean call arrival rate, λ_{eff}, is $1/E[T]$.

For determining the mean call arrival rate, we first find the λ_{eff} as a function of the simulation load of a cell and the number of mobile users. Next, we use the above relationship to obtain λ

as a function of λ_{eff}. In summary, although call arrival events are generated at every user with a mean arrival rate λ, not all arrivals result in an actual call. Effectively, calls are generated at a rate λ_{eff} which is less than λ.

2) Air Interface Parameters: The simulation parameters adopted for the air interface are as follows. We use three sectors per cell in our simulation, and the sectorization gain is set to 2.55 [19]. This amounts to a sectorization efficiency of 0.85 for a three sector antenna. We set the ratio of bit energy to noise density E_b/I_t to 7 dB, the spreading bandwidth B_w to 1.23 MHz, thermal noise of a cell N_oB_w to -108 dBm, voice rate to 9.6 kb/s and activity factor to 0.5, and other cell interference parameter f_I to 0.67. The maximum allowable cell load and engineered load for each cell are 0.9 and 0.7 respectively. The RAN is provisioned to support the total engineered load of all the cells. While the modeling of the dynamic quality behavior of soft-handoff user legs remains an open research issue, we assumed a simple model where weights are assigned to each user leg, adding up to 1.0 across all legs and biasing the weight of the primary leg to be highest.

In order to isolate the effects of congestion in the RAN, we assume that the bottleneck is only in the IP RAN and there is enough capacity at the air interface. Based on the call arrival and handoff events in the trace, we generate voice frames every 20ms for each active call. These frames are sent to the appropriate base stations. The base stations then aggregate these frames into IP packets which are then sent to the ANC through the bottleneck link. The provisioning in the RAN is such that when the load of all the cells is 0.7, the planned load, and the bandwidth utilization of the bottleneck link in the RAN is 0.90. The maximum drop-tail router delay is set to 20ms.

The scaling factor for load control in the RAN is calculated using Equation 5. For both maximum-power-based load control and usage-based load control, the typical parameters are 0.01 for the targeted packet loss rate at the RAN, 0.01 for the scaling factor adjustment parameters σ, and 0.1 for the minimum value of α. The maximum value of α is set to one for maximum-power-based load control, and 1.3 for usage-based load control. The typical control interval time period when parameters are updated is five seconds for both admission control schemes.

The parameters used in our simulation are summarized in Table I.

Parameters	Values
E_b/I_t	7dB
B_w	1.23MHz
N_oB_w	-108dBm
Sectorization Efficiency	0.85
Voice Rate	9.6Kbps
Inter-frame Separation	20ms
v_f	0.5
f_I	0.67
ρ^{max}	0.9
ρ^{plan}	0.7
Maximum Drop-Tail Delay	20ms
α_{min}	0.01
CAC-MPC:α_{max}	1
CAC-UC:α_{max}	1.3

TABLE I

SIMULATION PARAMETERS.

B. Performance Metrics

In order to study the impact of our control policies on the CDMA network with an IP RAN, we use the following three performance measures.

- *Call blocking rate* - Call blocking rate is defined to be the number of calls blocked over the number of calls received. This measure shows the direct impact of IP RAN capacity limitations on CDMA network capacity.
- *Frame Error Rate (FER)* - The frame error rate is defined to be the number of frames not received at the ANC over the number of frames sent by the mobile users. We study both average and instantaneous FER.
- *Burst Size* - Burst size is defined to be the number of consecutive frames with errors.

The frame error rate and burst size together represent a good measure of user voice quality while the call blocking rate is a good measure of network efficiency. We now look at our three congestion control techniques and study the impact of congestion on these performance measures.

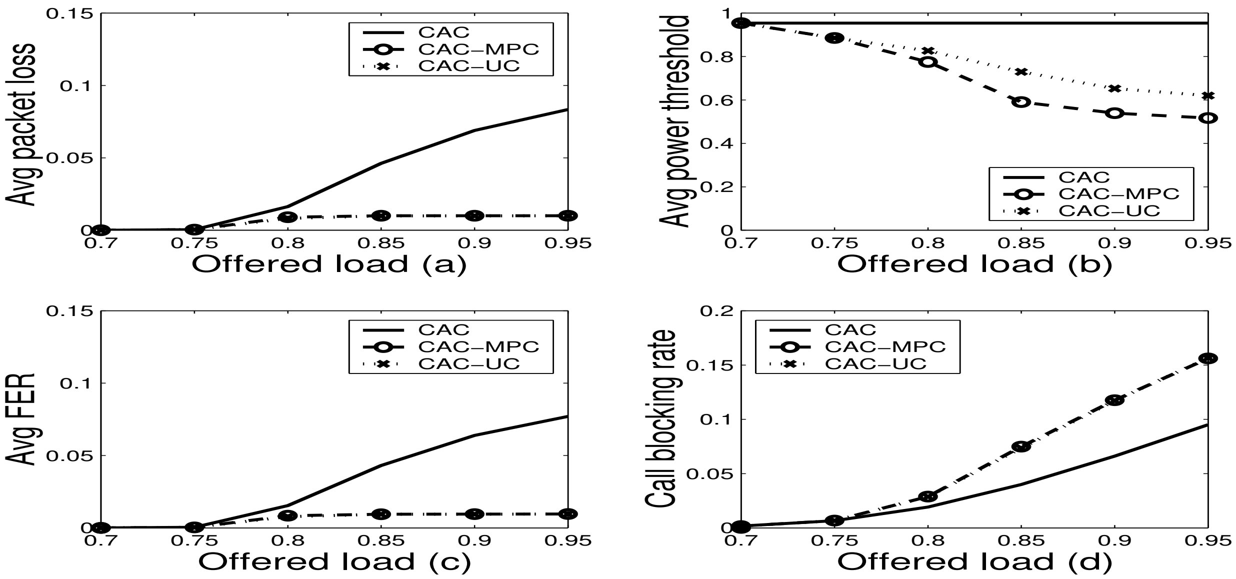


Fig. 3. Performance metrics of CAC, CAC-MPC, and CAC-UC as a function of offered load: (a) average packet loss rate; (b) average power threshold; (c) average frame error rate; (d) average call blocking rate.

C. Performance of Call Admission Control Algorithms

We compare the performance of three different admission control schemes – the basic scheme that controls admission in the air interface (CAC) by using a received power threshold, and the two enhancements (CAC-MPC and CAC-UC) that measure packet loss rate to regulate the admission control power threshold. In addition to the call blocking rate and the frame error rate performance measures, we also evaluate the performance of the three schemes with respect to IP packet (aggregated frames) loss rate and average admission control power threshold values since these parameters impact the admission control algorithms directly. The average power threshold value is normalized with respect to the maximum allowable received power in each cell, and averaged over all cells. We also study the impact of load control parameters including the target loss rate, L^*, the load control interval, and scaling factor σ.

We compare the performance of the three schemes using average values of the performance measures over the entire duration of the simulation. This is followed by a comparison of the temporal behavior of the three schemes. Last, we examine the sensitivity of our enhanced call admission control schemes to different system parameter values.

1) Performance Comparison of CAC, CAC-MPC, and CAC-UC: Fig. 3 (a) and (c) show the variation of the average packet loss rate and received voice frame error rate as a function of the offered load. The loss rate is almost zero until the offered load exceeds 0.7, the planned load. Both maximum-power-based control and usage-based control are seen to maintain the average loss rate at the targeted level, 1%, which is up to 8 times lower than the performance of basic CAC. Fig. 3 (a) and (c) also show that the average frame error rate and the average packet loss rate behave almost identically with changing load. We find similar behavior in other simulations as well.

When the access network is congested, the two enhanced admission control schemes reduce the threshold power for admission control. As a result, new calls from heavily loaded cells are blocked. Fig. 3 (d) shows that the blocking rate of all the schemes increases almost linearly as the load increases beyond 0.8. As expected, the blocking rates of the enhanced admission control schemes are higher than that of the basic CAC scheme. In practice, due to the self-regulating behavior of reverse outer loop power control, the ANC would instruct the mobiles to increase their power levels since it typically assumes that losses only occur in the air interface. This would result in a higher blocking rate than the blocking rate shown in the figure for the basic CAC scheme.

Fig. 3 (b) plots the average admission power levels for the schemes. It can be seen that CAC-UC allows up to 20% higher average admission power than CAC-MPC. This is because CAC-UC responds faster to the changing power level in each overloaded cell. However, due to the non-linear relationship between the admission control power threshold and the air interface capacity, the relative decrease in blocking rate is much smaller than the relative increase in the admission control power. The call blocking probability for CAC-MPC and CAC-UC is seen almost identical in Fig. 3 (d).

Fig. 4 shows the probability of occurrence of successive frame errors denoted by the burst size parameter. We find that both the load control schemes keep the burst size low with the CAC-UC scheme performing slightly better than the CAC-MPC scheme. As expected, the basic CAC scheme does not perform as well, showing higher burst losses.

2) System Dynamics: In order to understand and compare the temporal behavior of the three schemes, we use a snapshot of the simulation between 7500 seconds and 8000 seconds. Figs. 5 (a) and (b) depict the variation with time of offered cell load, the power threshold for admission

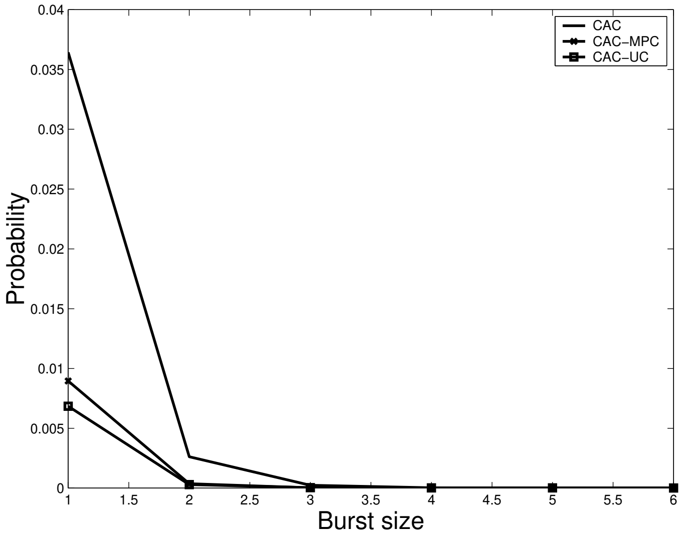


Fig. 4. Probability vs Burst Size, offered load = 0.85, for CAC, CAC-MPC and CAC-UC

control, and packet loss rate at two values of average offered cell load, 0.8 and 0.9.

As before, maximum-power-based load control results in a lower (more conservative) power threshold; consequently, the threshold needs to be adjusted less often and less drastically as compared to usage-based load control. At the higher offered load, both load control schemes apply a lower power threshold, while CAC-UC also adjusts the threshold more frequently than CAC-MPC. Both the load control schemes, effectively control the packet loss rate keeping it stable and close to the target. Figs. 5 (a) and (b) also show that CAC-UC and CAC-MPC have similar behavior of packet loss variation with time.

The instantaneous frame error rate, which directly impacts voice quality, is shown in Fig. 6 (a) and (b), again in the time-window between 7500 and 8000 seconds. Similar to the instantaneous loss rate, the frame error rate is well-controlled under both load control schemes. At light to medium loads, the frame error rate is reduced faster by usage-based control than by maximum-power-based control, but the former scheme also results in larger and more frequent oscillations. The frame error rate variation is smaller at higher loads, since the admission control power threshold is better-controlled.

3) Sensitivity to Target Loss Rate, Control Interval, and Load Adaptation Speed: We study the system performance at three different values of the target loss rate: 0.008, 0.01, and 0.03. Fig. 7(a) shows that the proposed load control schemes can effectively guarantee a range of voice

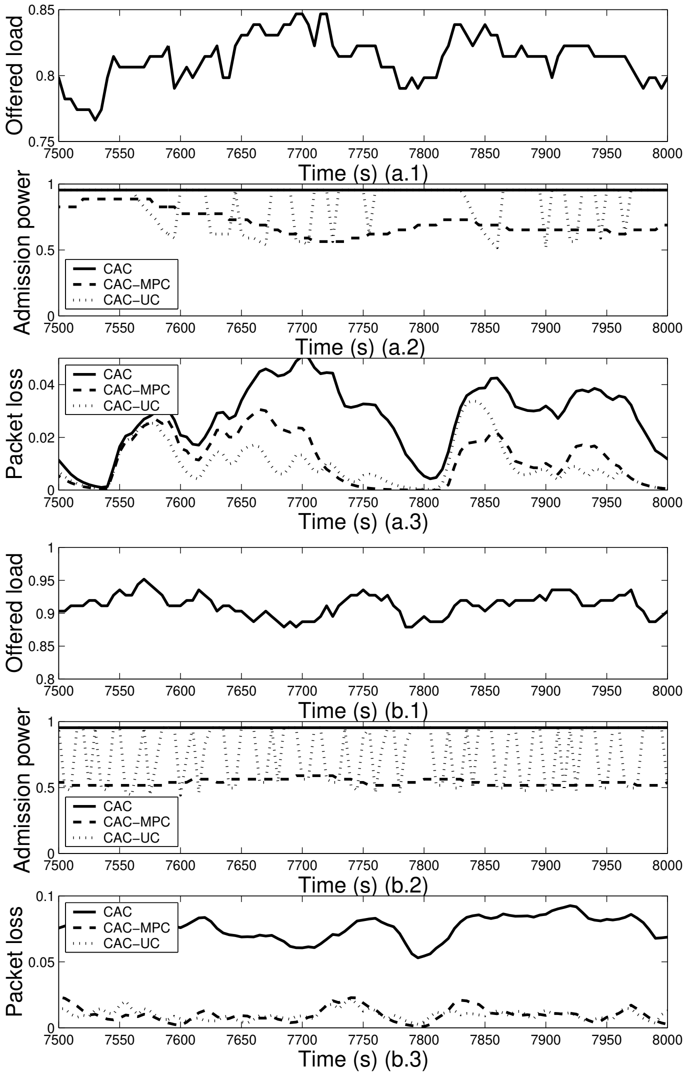


Fig. 5. Time variation of offered load, power threshold and packet loss rate for CAC, CAC-MPC and CAC-UC: (a) at offered load 0.8; (b) at offered load 0.9

frame error rates (and similar IP packet loss rates). Fig. 7 (b) shows that blocking increases as the target loss rate is lowered (thereby increasing voice quality). As before, CAC-UC allows a higher power threshold for admission control than CAC-MPC, and correspondingly has a slightly lower call blocking rate.

In order to study the impact of the control interval at which packet loss is estimated and power thresholds are computed by the ANC, we perform simulations for three values of the control interval: one second, five seconds and 10 seconds. Fig. 8 (a) shows that more frequent

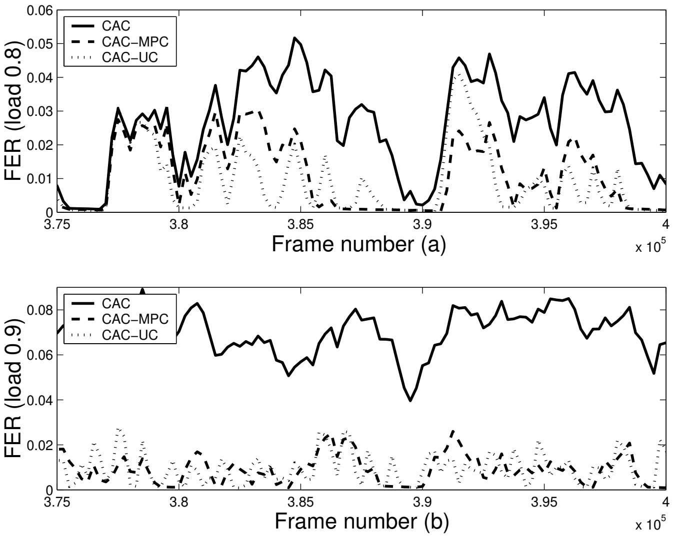


Fig. 6. Time variation of the received voice frame error rate: (a) at offered load 0.8; (b) at offered load 0.9.

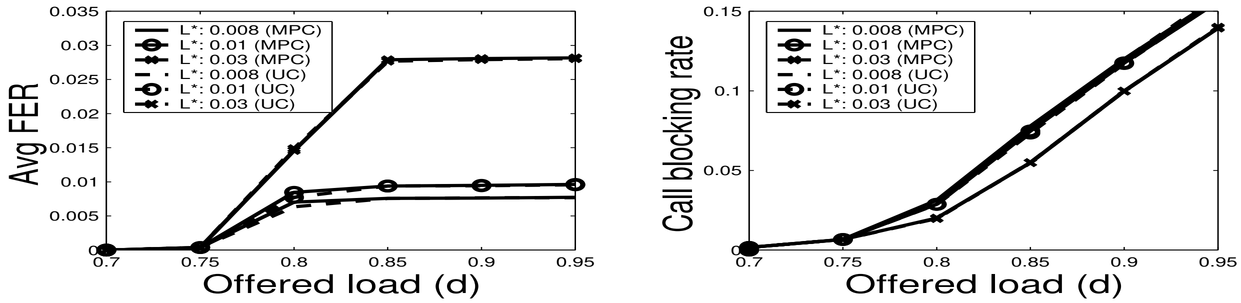


Fig. 7. Performance metrics of CAC, CAC-MPC, and CAC-UC as a function of offered load at target loss rate 0.008, 0.01, and 0.03: (a) average frame error rate; (b) average call blocking rate.

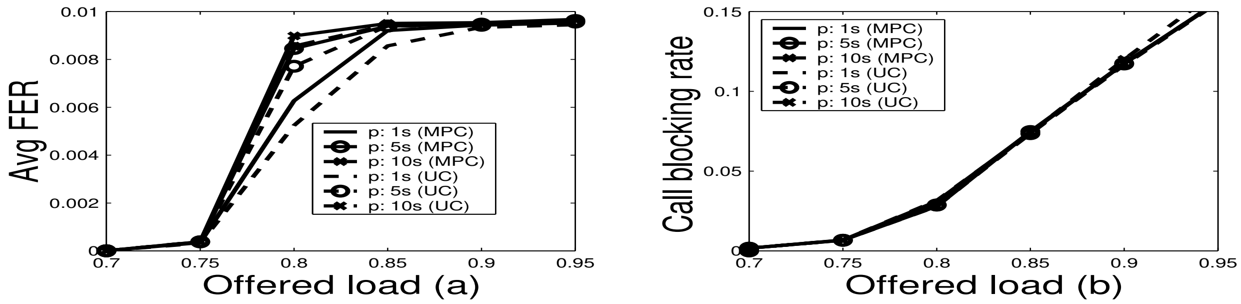


Fig. 8. Performance metrics of CAC, CAC-MPC, and CAC-UC as a function of offered load with control interval 1s, 5s, and 10s: (a) average frame error rate; (b) average call blocking rate.

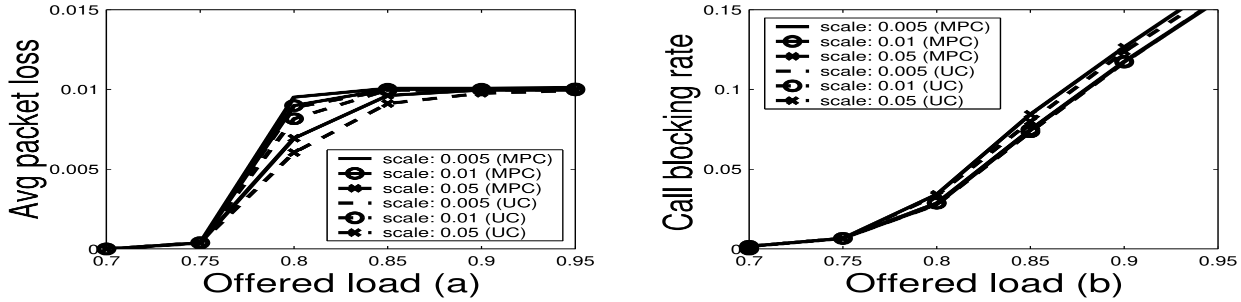


Fig. 9. Performance metrics of CAC, CAC-MPC, and CAC-UC as a function of offered load with control factor σ at 0.005, 0.01, and 0.05: (a) average packet loss rate; (b) average call blocking rate.

load control leads to lower average loss rate for both control schemes at moderately high loads. At very high loads, this effect becomes less important, because the large difference between the current and target values of the packet loss rate results in a very aggressive load reduction within a single control interval. Fig. 8 (b) shows that the blocking rate does not change appreciably with the tested control intervals.

The control factor σ in Equation 5 controls the adjustment rate of the load control parameter α with respect to change in loss. Hence it controls the rate at which the RAN load is adjusted towards the target value. We examine the system performance with different values of σ. Figure 9 shows how the performance of the call admission control algorithms as a function of σ. A higher value of σ is seen to result in lower average loss rate at moderately high loads only. The effect of σ on the overall blocking rate is small. Our simulations also indicate that a higher value of σ reduces the frame error rate faster, but results in more oscillatory behavior. In our simulations, a value of 0.01 for σ achieves good response to changing load conditions with minimal oscillations.

D. *Impact of Diversity Control*

We conduct a set of simulations to evaluate the performance of the service-degradation and frame-discard policies (SDP and FDP) for diversity control introduced in Section IV.

Fig. 10(a) shows frame error rates (FER) while Fig. 10(b) depicts call blocking results. As the mean offered cell load varies from 70% to 95%, both figures compare: a) the performance of call admission control (CAC) only, against b) the performance gains of adding usage-based admission control (CAC+UC= CAC-UC), and c) the additional gains of adding diversity control(CAC+UC+DIVC= CAC-DIVC). The diversity control period is set to 1 second,

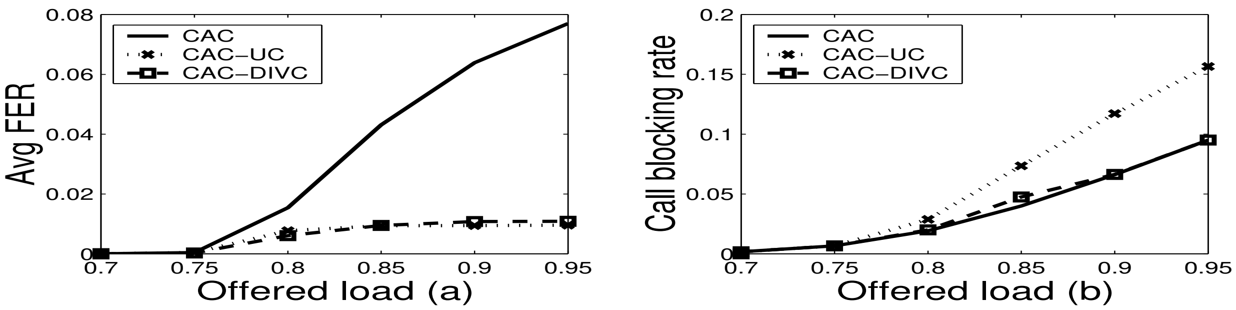


Fig. 10. Performance of CAC and the incremental gains of diversity and load control with respect to: a) mean user FER; b) call blocking rate.

as preliminary sensitivity studies suggest this to be a reasonable balance between control overhead and reactivity. The admission control period is set to 5 seconds, as done in all previous admission control experiments. The drop probability is computed by Equation 9.

The DIVC curves shown are for the frame-discard policy, FDP, as performance for the service-degradation policy, SDP, is virtually indistinguishable from FDP. Interestingly, we observe that although both policies differ on how they discard frames for certain mobile users, the diversity selection effect at the ANC masks these differences, resulting in an uplink traffic stream with the same FER in both cases. This suggests that both streams perform comparably well in discarding the legs of least quality.

Fig. 10(b) shows that the diversity and call admission controls exhibit comparable call blocking rate at all loads (note that the CAC-DIVC and CAC curves are almost identical). This demonstrates that the enhanced admission control is rarely triggered, if at all, so any call blocking observed is due to the basic CAC scheme and not due to IP RAN capacity constraints. Diversity control operates at the relatively low call blocking rates of CAC while maintaining the FER at or below the target level of 1%. Thus, diversity control delivers both low blocking and the desirable FER resulting in a graceful adaptation to congestion.

Intuitively, by discarding frames at the cells that would have likely been redundant in the diversity selection process at the ANC, the packet loss rate in the RAN is reduced. This keeps the FER below the trigger point where load control would have been needed to increase call blocking in order to reduce packet loss. For instance, at a load of 95%, using load control only to reduce congestion requires about a 6% increase in the call blocking rate to maintain a target

FER below 1%. CAC, on the other hand, operates at low blocking rates but does not react to any fluctuations in RAN congestion state resulting in large increases in FER as loss in the RAN increases at higher loads. This illustrates an important point: provided that soft-handoff legs have a high probability of being redundant most of the time, diversity control results in a win-win by reducing loss in the RAN and also maintaining low call blocking rates.

Recall that the aggressiveness with which diversity control reacts to changes in measured rate loss is controlled by the adaptation speed parameter σ. We observe that a value of 0.10 for σ, used for the simulation results shown in Fig. 10, provides a good balance between fast responsiveness and mean drop probability. Though the variance was reasonable, efforts to further reduce it by lowering the value of σ would make the control's response sluggish, causing the load control to be triggered and subsequently increase the call blocking rate. For instance, at a load of 95%, values of 0.05 and 0.01 for σ yielded call blocking rates of 11.7% and 13.2%, respectively (contrasting with the 9.5% observed for $\sigma = 0.10$). This clearly illustrates that while the joint operation of diversity and load control is effective at maintaining the target FER, the trigger point for load control depends on the aggressiveness with which the diversity control is tuned to react to congestion episodes.

It is desirable to use a fast-acting diversity control whenever possible in order to avoid or minimize the increase in call blocking rates that will ensue when the load control is triggered. However, regardless of how aggressive the diversity control policy is, its capabilities are limited by the average amount of redundant data available for discard at the cells. When the average load increases beyond the limits of what diversity control can manage, load control should be applied to increase call blocking rates in order to ensure acceptable quality levels are maintained. The application of both load and diversity controls, in fact, achieves the joint goal of maximizing network call capacity while minimizing voice quality degradations during congestion episodes in the IP RAN. Note that the benefits of a joint diversity- and load-controlled RAN are not visible in Fig. 10 because the offered load range depicted is well within the limits of what diversity control can independently manage. Simulation experiments with both controls operative confirm that diversity control has a dominant effect in this load range.

In conclusion, since diversity control exploits traffic redundancy, it can be a very effective congestion control technique by adapting to congestion without either increasing the frame error rate or blocking. However, its performance depends crucially on the average amount of redundancy per user (i.e., number of SHO legs) as well as on the ability to select low-quality legs to be discarded.

E. Impact of Router Control Using AQM

We now study the impact of router control using active queue management (AQM) by turning AQM on and off in the router. The AQM policy used in these simulations is described in Section V.

We first consider the scenario where a router guarantees that packets do not experience any delay more than a pre-specified delay threshold. The router queue is set to a maximum length such that the queuing delay never exceeds the delay threshold. Any packets arriving when the queue is full are dropped in a drop tail manner. If this delay threshold is set to the time by which the packets are expected to arrive at the ANC then all packets arrive in time at the ANC. We find that turning on AQM in this scenario does not help improve the CDMA network performance. In fact, both the blocking rate and average frame error rate are slightly worse than the scenario when AQM is turned off. There are two reasons for this behavior. First, the router is able to bound the delays of packets within the required threshold. An attempt to reduce the delay causes unnecessary loss. Proactively dropping packets does not speed up the load control because load control operates at much larger time scales (5 seconds). Second, in our experiments, the probability of multiple frames associated with different legs of a soft-handoff being good is low. This implies that random packet dropping does not necessarily ensure that a good frame is received at the ANC. Hence, there is no drastic improvement in the frame error rate.

When the router is not able to provide the required delay guarantee (for example, to accommodate data bursts), we find that turning on AQM helps improve performance considerably, especially when no additional controls such as load control or diversity control are applied at the ANC. We conduct an experiment in which the required delay threshold at the ANC is 20ms but the router queues up packets with a maximum delay of 30ms and uses SRED as the AQM policy. Fig. 11 (a) shows the frame error rate as a function of frame number. Here, only the basic CAC scheme is being used. The average cell load offered to the IP RAN is 85% of the maximum cell load. This is much more than the load (70%) that the line capacity is provisioned to handle. We see that without AQM, the frame error rate is very high and close to 100% but with AQM, the frame error rate is brought down to less than 10%. This is because the bottleneck

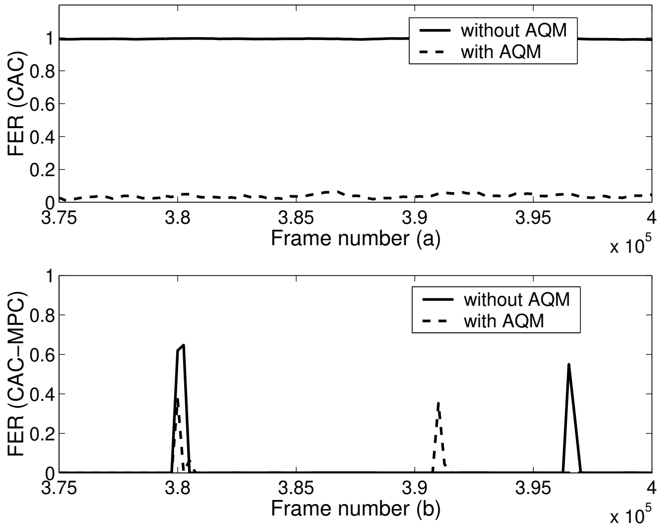
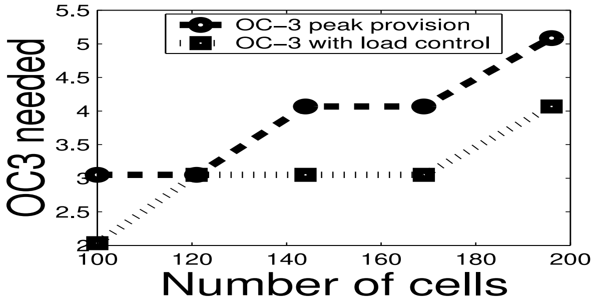


Fig. 11. FER vs Frame Number, Offered Load = 85%, Maximum Router Queuing Delay = 30ms (a) CAC (b) CAC-MPC

queue becomes full very fast and packets do not get processed fast enough. Every packet arriving after the queue is full or nearly full suffers a delay greater than 20ms and arrives at ANC too late. We also find that the average frame error rate over the entire simulation time is reduced from 98% to 4%, when using AQM with the basic CAC scheme.

Fig. 11 (b) shows the frame error rate as a function of frame number when maximum power-based load control is used with AQM turned on and AQM turned off. We find that without AQM, high delays in router queue due to drop tail packet dropping cause large spikes in frame error rate. The magnitude of these spikes is brought down when AQM is applied. We observe that the spikes are several hundred frames wide. This is because our load control policy controls admission of new calls only and does not drop existing calls. The existing calls could continue to contribute to the congestion for several seconds. In our experiments using AQM and CAC-MPC, we also find that the call blocking rate is brought down from 15.7% to 13.4% and the average frame error rate during the entire duration of the simulation is reduced from about 2% to 1%. As expected, our experiments with other maximum queuing delay values show that as the maximum router queuing delay value is increased (i.e., the drop tail buffer size is increased) the improvement due to AQM becomes even higher.

In summary, router control using an AQM policy that ensures smaller delays will result in better CDMA network performance.



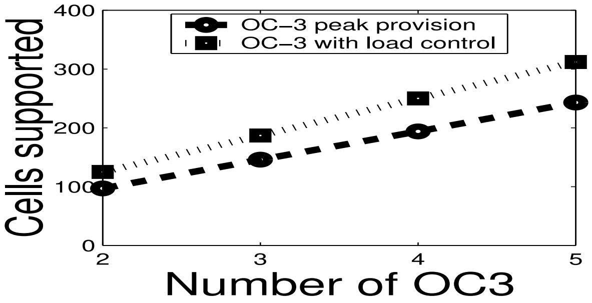


Fig. 12. Performance of CAC and the incremental gains of diversity and load control with respect to: a) mean user FER; b) call blocking rate

VII. Impact on Network Provisioning

Voice transmissions generally have high quality requirements and peak provisioning could be used for supporting voice calls in the IP radio access network. However, the dynamic variations in call traffic patterns may result in under-utilization of a peak provisioned IP RAN. Additionally, the rapid growth of wireless traffic and the evolution of radio technology may quickly render the resources in an existing peak-provisioned IP RAN in-sufficient.

In previous sections, we have shown that our proposed congestion control algorithms could effectively control the network load and maintain the expected voice quality under various network conditions. In this section, we study the effect of efficient load control on network provisioning. We consider cells that have three sectors, and each sector has three carriers. The air interface parameters are the same as in Section VI, and the planned air interface load to support is also set at 0.7. We first vary the number of cells connecting to one ANC and compare the number of OC-3[5] links needed to support the required voice quality. Next we vary the number of OC-3 links in the IP RAN, and compare the number of cells that can be supported.

Fig. 12 (a) depicts respectively the number of OC-3 links needed to support the traffic from various number of cells to the ANC using peak provisioning and efficient provisioning enhanced with our load control algorithms. With the same target voice quality, the provisioning with our algorithms is seen to require up to 20% less OC-3 links. This reduces the IP RAN bandwidth cost significantly.

[5] OC-3 stands for Optical Carrier Level 3 and corresponds to 155.52 Mb/sec [20].

With an IP RAN provisioned in advance, Fig. 12 (b) shows the number of cells that can be supported with and without load control algorithms enabled. With load control, about 30% more cells can be supported by an existing IP RAN provisioning. In other word, if the number of cells connecting to an ANC is fixed, up to 30% more traffic is allowed for each cell without the need to upgrade the IP RAN.

VIII. Future Directions

In this section we present and discuss some future directions for enhancing our research.

A. Congestion Control for Data

One interesting difference between voice and data with respect to the frame selection function of the ANC is that while the best voice frame is always forwarded, the best data frame is only forwarded if the cyclic redundancy check (CRC) is successful. Thus, one can further improve diversity control techniques by dropping data frames that fail the CRC at the base station. Currently, we are studying the impact of these techniques for IP RAN congestion control in the presence of wireless data traffic.

B. Congestion in the Downlink Direction

The focus of this paper has been congestion in the uplink direction. While call admission control and router control techniques are applicable equally in both directions, the use of diversity control in the downlink direction is not as straightforward. This is because, in CDMA systems, while *frame selection* from different legs is performed in the uplink direction, *frame combining* (using maximal rate combining) is performed in the downlink direction at the mobile device. This requires that all base stations transmit the same frame at the same time in the downlink direction. Thus, dropping of frames from selected legs during soft-handoff using diversity control could result in higher error rates for a given mobile device but this also reduces the overall interference level on the wireless downlink and could result in better reception for other mobile devices. A detailed study of diversity control in the downlink direction is necessary to determine its efficacy. It is important to note that in the downlink direction IP multicast (e.g. source specific multicast [21]) could be used to transmit the frames from the ANC to multiple base stations, thereby reducing the usage of the link connecting the ANC and the router.

C. Wireless RAN

So far, we have assumed that the radio access network is comprised of wired links with fixed capacities. However, it might be advantageous in certain locations to connect base stations to a network controller through fixed wireless backhaul links (possibly, multihop for range extension). The emerging IEEE 802.16 standard [22] with range up to 30 miles is one suitable wireless RAN technology. In this case, capacity of the RAN is no longer fixed and congestion in the RAN could also be caused by variations in the wireless link capacity potentially due to interference or atmospheric effects. An adaptive solution, as proposed in this paper, is essential in combating these dynamic congestion events in the wireless backhaul.

IX. Conclusions

In this paper, we studied the problem of congestion control in the IP RAN of a CDMA wireless access network and examined three control techniques called admission control, diversity control, and router control.

In the case of call admission control, we presented two enhanced admission control mechanisms. These schemes make use of a variable power-based admission control threshold at the air interface to regulate load entering the RAN. Simulations show that both algorithms are able to control packet loss rate and frame error rate in the RAN at a desired level. The usage-based control scheme achieves a higher power-threshold and hence slightly lower call-blocking than the maximum-power-based scheme, but the latter scheme results in a smoother, less oscillatory control. Both algorithms were also shown to be robust to the control parameters. In the case of diversity control, we proposed two novel policies called Service Degradation Policy (SDP) and Frame Discard Policy (FDP) for gracefully adapting to congestion. The results were very promising. Both policies were shown to be able to adapt to significant congestion without either increasing the frame error rate or blocking. In the case of router control, we evaluated an active queue management policy called SRED and found that router control that ensures low delays is essential for achieving low frame error rate during congestion.

In the future, issues related to data traffic, downlink congestion, and wireless RANs as discussed in Section VIII, need careful examination.

ACKNOWLEDGMENT

The authors thank Tom LaPorta, Srinivas Kadaba, Ganesh Sundaram for useful discussions on the topic.

REFERENCES

[1] T. Bu, M.C. Chan, and R. Ramjee, "Connectivity, performance, and resiliency of ip-based cdma radio access," in *Proceedings of IEEE Infocom (to appear)*, 2004.

[2] Kevin Maney, "Future not so bright for telecoms," in *USA Today*, July 2002.

[3] Anup Kumar Talukdar, B. R. Badrinath, and A. Acharya, "Mrsvp: A resource reservation protocol for an integrated services network with mobile hosts," *Wireless Networks*, vol. 7, no. 1, 2001.

[4] G. Heijenk, G. Karagiannis, V. Rexhepi, and L. Westberg, "Diffserv resource management in ip-based radio access networks," in *Proceedings of WPMC*, Aalborg, Denmark, September 2001.

[5] B. Braden et al, "Recommendations on queue management and congestion avoidance in the internet," Request for Comments 2309, Network Working Group, April 1998.

[6] 3G Partnership Project, "Release 99," .

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

[8] V. Garg, *Wireless Network Evolution 2G to 3G*, Prentice-Hall, 2001.

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

[10] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, August 1993.

[11] S. Kasera, J. Pinheiro, C. Loader, M. Karaul, A. Hari, and T. LaPorta, "Fast and robust signaling overload control," in *Proceedings of ICNP*, Riverside, CA, November 2001.

[12] B. Hajek, "External splitting of point processes," *Mathematics of Operations Research*, vol. 10, pp. 543–556, 1985.

[13] Y.-B. Lin and V.W. 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.

[14] J.K. Kwon and D.K. Sung, "Soft handoff modeling in cdma cellular systems," in *Proceedings of Vehicular Technology Conference*, May 1997.

[15] D.K. Kim and D.K. Sung, "Characterization of soft handoff in cdma systems," in *Proceedings of Vehicular Technology Conference*, June 1999.

[16] Y.-B. Lin and W. Chen, "Impact of busy lines and mobility on cell blocking in pcs networks," *International Journal of Communication Systems*, , no. 9, pp. 35–45, 1996.

[17] M. Wallace and R. Walton, "Cdma radio network planning," in *Third Annual International Conference on Universal Personal Communications*, Sept. 1994.

[18] S. Hong and I-T Lu, "Soft handoff parameter optimization in various propagation environments," in *Proceedings of Vehicular Technology Conference*, 2001.

[19] Qualcomm, *The CDMA network engineering handbook*, Qualcomm, 1993.

[20] C. Partridge, *Gigabit Networking*, Addison-Wesley, 1994.

[21] H. Holbrook and B. Cain, "Source-specific multicast," Inernet Draft draft-ietf-ssm-arch-04.txt, Oct. 2003.

[22] IEEE 802.16-2001, *IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, Apr. 2001.