On Maintaining Multimedia Session's Quality in CDMA Cellular Networks Using a Rate Adaptive Framework

Ehab S. Elmallah
Department of Computing Science
University of Alberta
Edmonton, T6G 2E8, Canada
ehab@cs.ualberta.ca

Mrinal Mandal
Department of Electrical and Computer Engineering
University of Alberta
Edmonton, T6G 2V4, Canada
mandal@ece.ualberta.ca

Abstract

In [4] the authors have developed call admission control and adaptive bandwidth allocation schemes for serving multimedia connections in cellular wireless networks with fixed cell capacity. The architecture considers an adaptive networking framework where the bandwidth of multimedia calls can be dynamically adjusted, and the proposed admission method works by enforcing an upper bound on the cell overload probability. In this paper we consider a similar adaptive framework, and devise call admission control and bandwidth allocation strategies to serve multimedia connections in a CDMA-based 3G cellular network. The architecture aims at maintaining the session's quality during both intra-cell and inter-cell user movements by limiting the cell overload probability. A novel aspect of our work is a method for exploiting a priori knowledge of user mobility patterns to estimate the cell overload probability after some prescribed prediction interval. Important properties of the devised method are proved analytically. Compared to a non-predictive admission control scheme, the obtained results show that the proposed scheme achieves a lower forced termination probability, and higher throughput while consuming less base station transmission energy.

1. Introduction

Provisioning multimedia services to mobile subscribers of both third generation (3G) and newer generation (NG) wireless networks in a cost effective and profitable way is widely regarded as a crucial step toward successful deployment of such networks. Prominent among such services is the class of multimedia streaming that enables content playback on mobile terminals that are generally energy and storage limited. Current standardization efforts of such services by major partnership projects (e.g., 3GPP and 3GPP2) have addressed the basic functional characteristics and requirements of the streaming services class for 3G CDMA-based networks (e.g., CDMA 2000 and W-CDMA) [6]. In such CDMA-based architectures, however, multiple access interference (MAI) plays a major role in resource management that is known to further complicate the designs (see, e.g., [1, 3, 7, 8, 9, 13] and the references therein). Our focus in this paper is on designing call admission control (CAC), and adaptive bandwidth allocation (ABA) schemes that can be used to give assurance that the admitted streaming connections will continue receiving acceptable levels of service (e.g., a minimum acceptable data rate) over the duration of a complete streaming session (e.g., a few minutes long), and also during both intra-cell and inter-cell user movements. Maintaining the session's quality of service (QoS) during user movements is mentioned as an interesting problem in [2] for which bandwidth based reservation techniques can be applied. The approach taken in this work, however, emphasizes the design of a suitable mobility prediction mechanism, and a matching CAC module that accepts new connection requests only if the predicted cell overload probability is below a certain threshold value.

For prior work, we note that many of the currently well established results on maintaining QoS during inter-cell user movement (handoff) concern voice only traffic (see, e.g., [12], and the surveyed results therein). On the other hand, research on supporting heterogeneous traffic (voice, and relatively high data rate traffic) in 3G systems is much more diversified in scope, mechanisms, performance measures, and end goals. The brief literature review given below considers whether the proposed designs use primarily scheduling only mechanisms, CAC only mechanisms, or combined scheduling and CAC mechanisms.

Examples of previous work on devising scheduling mechanisms include the work of [3, 13]. In [3], the authors evaluate the performance of various scheduling algorithms for serving user requests on the downlink. The size of each request (e.g., file length of an HTTP request) is assumed to be known a priori. In contrast to our work here, however,

channel conditions in [3] are assumed to be time-invariant during the entire period of serving any single request. In [13] the authors consider dynamic bandwidth allocation for heterogeneous traffic on the uplink based on weighted fair scheduling. User requests and data rate allocation are done during well defined time slots. The devised scheme aims at achieving good utilization of the available bandwidth as well as fairness in allocating the bandwidth. The devised scheme can also guarantee delay bounds on a session's flow if the flow conforms to the parameters of a leaky bucket regulator, and the cell load allows the base station to allocate bandwidth more than the token arrival rate parameter. As can be seen, the above approaches do not address the problem of serving the admitted flows at some data rate during the entire session time while taking mobility into account.

In a second direction, [7] devises a CAC mechanism for serving uplink traffic flows with different QoS requirements. Admission requests are made by mobile stations at the end of certain time slots; each request is assumed to be sent to the base station together with the desired bandwidth, mobile's transmission power limit, and a required outage probability ([7] assumes no link level error recovery protocol is used). The devised CAC mechanism accepts only if the requested performance values can be satisfied during the following time slot (with no particular consideration of the potential changes in the base station transmission power requirements as users move within the cell). More recently, [11] used a control theoretic approach for designing a CAC that dynamically adjusts the capacity of guard channels in CDMA cellular networks so as to maintain the handoff dropping rate at a target level. Simulation on voice dominated traffic is used to show the resulting performance.

The closest previous work to this paper is [4] where the authors devise a combined CAC and ABA scheme for serving streaming connections in networks with fixed cell capacity. The results of [4] deal with an adaptive multimedia networking framework where the bandwidth of an ongoing multimedia connection can by dynamically adjusted several times during the connection's lifetime. The proposed CAC scheme works by predicting the probability that the cell under consideration will run into an overload condition (after some prescribed time interval), if a newly arriving connection request is accepted. The work of [14] also considers the development of a rate adaptive framework for fixed capacity cellular networks.

Compared to the case of fixed cell capacity networks, predicting the cell overload probability in the presence of MAI for CDMA networks is more challenging. In particular, it is insufficient to take the number of active streaming users in the cell as the only measure of cell load, rather a more detailed account of the base station transmission power requirement of different classes of users should be considered. The approach taken in our work here is based on classifying the active streaming users according to the amount of base station transmission power required to provide service at an acceptable data rate. The classification is based on the average large scale path loss experienced by each user. To this end, we assume that the geographical area of the cell under consideration is partitioned into r regions, $r \geq 2$, called *rings* hereafter. Each ring is viewed by the mobility model as a homogeneous area where the average large scale wireless path loss from the serving base station to each point of interest falls within some pre-determined range (e.g., [70 dB, 110 dB]) that uniquely distinguishes that particular ring in the cell. In general, subdividing the cell (or a sector) into a large number of rings results in a more accurate model at the expense of increased computational requirements of the CAC scheme.

The rest of the paper is organized as follows. Section 2 discusses different aspects of the system model and introduces the particular mobility model used in the analysis. Sections 3 and 4 outline the general structure of the proposed CAC scheme. Section 5 describes the main functions executed by the ABA module. In section 6, the performance of the proposed mechanisms are evaluated using simulation which is followed by the conclusions in Section 7.

2. Network, Service, and Mobility Models

Network Architecture. Throughout the paper, we consider a UMTS-like cellular network operating in the frequency division duplex (FDD) mode where each user streaming flow is served by a dedicated data channel. The newly added functionality of the proposed CAC and ABA modules are assumed to be embedded in the radio network control (RNC) part. Multimedia streaming requests are assumed to be served by an adaptive server capable of choosing suitable video streaming parameters (e.g., video resolution, frame rate, and encoding parameters) in response to possible requests from the wireless network to vary the currently allocated bandwidth to the stream. Such server can be hosted by the public land mobile network (PLMN) provider, and attached to the gateway GPRS support node (GGSN) of the core network (CN) module of the cellular network.

Service Model of the Streaming Class. Similar to the framework of [4], we assume that the streaming QoS class is characterized by a set $\{R_{min}, R_{sat}, R_{max}, P_{admit}\}$ of parameters, where: $R_{min} \leq R_{sat} \leq R_{max}$ (bps), and the range $[R_{min}, R_{max}]$ defines the *acceptable* downlink data rates for the streaming service, R_{sat} defines a *satisfactory* data rate for any connection in the class, and P_{admit} is a QoS admission control probability representing the desirability level of not degrading a connection data rate below R_{sat}. A connection stream served at (or above) R_{sat} for the entire connection duration is viewed as a *satisfied* con-

nection. Allocating an acceptable rate below R_{sat} results in a *degraded* service. We denote by $\rho_{sat,i}$ the fraction of time the ith connection is satisfied, and use ρ_{sat} to denote the average (over all class connections served during some time epoch of the day) of the achieved $\rho_{sat,i}$ satisfaction ratios. We take ρ_{sat} as one component of the objective criterion that the resource management strives to maximize.

Conversely, if at some instant the available base station transmission power becomes insufficient to satisfy each admitted stream at, or above, R_{min} (e.g., due to user mobility) then allocating a data rate below R_{min} is assumed to result in an unacceptable playback performance. To evaluate the worst case performance of the devised mechanisms, we assume that a stream that cannot be continued at an acceptable rate is forcibly terminated. Furthermore, to account for the impact of service termination on the network provider's revenue, we use the average *effective* throughput where the average is taken over all traffic streams that have been served to completion at acceptable data rates (that is, excluding the streams that have been forcibly terminated).

Mobility Model and Transition Diagrams. A novel aspect of our work is a method that exploits a priori knowledge of both intra-cell and inter-cell user mobility patterns in making admission decisions. To this end, we formalize the concept of mobility transition diagrams. The concept adopts a probabilistic mobility model (similar to the model used in [4] and [5]) to the case where users move among rings within the same cell, as well as between different cells. The model uses the following assumptions.

The arrival of streaming connection requests to the target cell is assumed to be Poisson with rate λ connections per second, and the time duration of each connection is exponentially distributed with average $1/\mu$ seconds. For connections initiated in ring $i \in [1, r]$ of the target cell, we refer to the time duration the connection is served in that particular ring as ring-i *residence* time. After that time duration, the user may depart to another ring. Here, we assume that ring-i residence time is exponentially distributed with average $1/\delta_i$ (δ_i may be viewed as the departure rate from ring i). Handoff traffic from neighbouring cells to the target cell is modelled using a similar setup: that is, we view the handoff traffic as coming from a new ring (denoted ring 0 in Fig. 1 below) lying outside the target cell, and assume that ring-0 residence time is exponentially distributed with average $1/\delta_0$ (here, δ_0 is the handoff rate from the neighbouring cells to the target cell).

In addition, user movement across rings during a certain epoch of the day is modeled by a transition diagram where rings and transitions are represented by nodes and directed edges, respectively (as illustrated in Fig. 1). Here, we use $p_{i,j}$, $i \neq j$, $i, j \in [0, r]$, to denote the probability that an active user moves from ring i to ring j after spending the

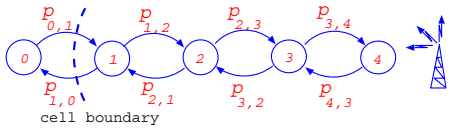

Figure 1. A mobility transition diagram

residence time mentioned above. The transition probabilities out of any ring $i \in [1, r]$ satisfies $\sum_{i \neq j} p_{i,j} = 1$. For the special case of $i = 0$ (i.e., for ring 0 representing the neighbouring cells), we have $\sum_{j=1}^{r} p_{0,j} \leq 1$.

The probability that a test connection initiated in ring i at time t will continue to be served in the same ring during the prescribed prediction interval $t_{predict}$ is denoted $p_{r,i}$; thus, $p_{r,i} = e^{-(\mu+\delta_i)t_{predict}}$. In addition, the probability that a test connection initiated in ring i will depart to another ring during the next $t_{predict}$ time units is denoted $p_{d,i}$ (here, $p_{d,i} = 1 - e^{-\delta_i t_{predict}}$). Thus, corresponding to each (i, j) transition in the mobility diagram, the product $p_{d,i}p_{i,j}$ is the probability that a test connection initiated in ring i will depart to ring j during the next $t_{predict}$ time units.

Modelling for Worst Case Analysis. We assume that during the lifetime of any connection, user mobility is modeled by taking into account at most one transition. Such assumption can be used to construct different behavioural models during a prediction interval $[t, t + t_{predict}]$. The design of the proposed CAC, however, relies on overestimating the required base station transmission power during the prediction interval. To this end, we adapt the above assumption as follows. If an active user moves away from the base station by traversing rings $(i, i-1, i-2, \cdots, j)$ then that user behaviour contributes to the (i, j)-transition. A similar worst case rule is used to model handoff users moving from some inner ring $i > 1$ to some neighbouring cell. Conversely, if an active user moves toward the base station by traversing rings $(i, i+1, \cdots)$ then that user contributes to the $(i, i+1)$-transition. Moreover, all incoming handoff traffic are assumed to be transitions into the outermost ring of the cell. In addition, the proposed CAC assumes that each mobile station experiences the maximum observed average large scale path loss in its respective ring during any such prediction interval.

3. Call Admission Control

The proposed CAC scheme combines a SINR-based admission control scheme (see, e.g., [2]) with a mechanism to predict the state of the target cell after a prescribed prediction interval; the main steps of the scheme are formulated using the following definitions. The state of the cell at any instant t is captured by an occupancy distribution se-

quence $N = (n_0, n_1, \cdots, n_r)$, where n_0 is the number of the streaming class users that may perform handoff to the target cell, and for $i \in [1, r]$, n_i is the number of active ring-i users in the target cell. An occupancy distribution sequence N is an *overload* distribution if the base station cannot allocate sufficient transmission power to serve each mobile at the distinguished rate R_{sat} assuming that each mobile experiences the maximum observed average large scale path loss in its respective ring. In addition, we denote the set of all overload distributions by $\mathcal{OL}$. Given that the cell is in state N at sometime t, we define the *overload* probability, denoted $P_{N,\mathcal{OL}}$, as the probability that the cell will be in some overload state at time $t + t_{predict}$.

Given an occupancy distribution N associated with a set $M = \{1, 2, \ldots, m\}$ of users in the target cell (i.e., $m = \sum_{i=1}^{r} n_i$), determining whether N is an overload distribution requires that the RNC estimates the minimum amount of base station transmission power that should be allocated to serve each user. We denote such vector of required transmission power levels by $\mathbf{P_{tx}} = (P_{tx,1}, P_{tx,2}, \ldots, P_{tx,m})$. In addition, we denote the total amount of the base station transmission power allocated for serving the streaming QoS class by $\mathcal{P}_T$. Determining $\mathbf{P_{tx}}$ (if one exists) is carried out by solving a linear system of equations where the equation associated with mobile receiver $u \in M$ has the following form (see, for example, [2, 10, 15])

$$E_b/I_0 = \frac{(W/R_u)(P_{tx,u}/L_u)}{\gamma\, I_{intra\_cell} \;+\; I_{inter\_cell} \;+\; \eta_0 W} \qquad (1)$$

where E_b/I_0 is a target bit energy to interference and noise ratio that should be achieved at each receiver (e.g., 7 dB), W is the chip rate of the CDMA air interface, R_u is the transmission data rate to mobile u, $P_{tx,u}$ is the amount of serving base station transmission power allocated to mobile u, L_u is the average path loss from the base station of the target cell to mobile u, $\gamma \in [0, 1]$ is a fraction reflecting the loss of orthogonality due to multipath, I_{intra_cell} is the interference power received by mobile u from the serving base station, I_{inter_cell} is the total interference power received by mobile u from the neighbouring base stations (a sum of log-normally distributed random variables), and η_0 is the noise power spectral density at the base station. A positive solution $\mathbf{P_{tx}}$ of the resulting system of equations is a feasible power allocation if it does not violate the base station capacity constraint:

$$\sum_{u=1}^{m} P_{tx,u} \leq \mathcal{P}_T. \qquad (2)$$

Given the above framework, the scheme works as follows. Upon arrival of a new streaming connection request, the CAC computes the occupancy distribution vector N that results if the newly arrived request is admitted. The RNC then tests whether there exists a feasible power allocation $\mathbf{P_{tx}}$ to serve all in-cell connections of the distribution N at the R_{sat} data rate. If there is no feasible allocation, the base station rejects the request. We recall from Section 2 that a forced termination occurs if the network cannot serve a connection at, or above, R_{min}. By using R_{sat} above, the scheme tries to avoid accepting a new request by degrading the service to some users. Else (if a feasible solution exists), the procedure uses the method described in Section 4 to decide whether the following admission condition is satisfied:

$$P_{N,\mathcal{OL}} \leq P_{admit} \qquad (3)$$

where P_{admit} is a QoS admission control probability for the streaming class set by the network provider. Finally, the RNC admits the request if condition (3) is satisfied.

4. Computing the Overload Probability

A core aspect of the devised CAC scheme is the procedure used to decide whether condition (3) is satisfied for a given distribution N. The design problem can be stated as follows. Given a target cell with the following parameters: $N = (n_0, n_1, \ldots n_r)$: an occupancy distribution vector that may result if a newly arrived call is admitted at sometime t, $\{p_{r,i} : \; i \in [0, r]\}$: the set of ring residence probabilities associated with the interval $[t, t+t_{predict}]$, $\{p_{d,i} : \; i \in [0, r]\}$: the set of ring departure probabilities associated with the interval $[t, t+t_{predict}]$, and $\{p_{i,j} : \; i \neq j, i, j \in [0, r]\}$: the set of possible transition probabilities in the diagram, the objective is to develop a procedure for deciding whether condition (3) holds for N. Additionally, the sought procedure is required to provide sufficiently fast response time that enables timely processing of incoming requests to any single cell, or a cluster of cells.

The above problem is computationally challenging, given the large number of possible occupancy distributions. A commonly used design paradigm to meet the stringent fast response time requirement is to organize the underlying computations into two distinct phases: an *offline* phase that may perform substantial computations and produces its output in a suitably encoded form for use in a second (*online*) phase. Subsequently, the online phase (performed by the CAC in real time) uses the encoded data from the first phase to take decisions at the required speed.

Notation. To describe the procedure we first introduce some notation. We denote by E_i^+ the set of (i, j)-transitions leaving ring i in the mobility diagram together with a self transition, denoted (i, i), for in-cell rings (i.e., for $i \geq 1$). A self transition is used to represent streaming connections that are initiated at time t in some ring inside the target cell and continue to be served in the same ring during the interval $[t, t + t_{predict}]$. In Fig. 1, e.g., we have $|E_0^+| = 1$,

$|E_4^+| = 2$, and $|E_i^+| = 3$ for $i \in [1,3]$. Furthermore, given a ring occupancy distribution $N = (n_0, n_1, \ldots n_r)$ at time t, we denote by $N_i^+ = (n_{i,j} : (i,j) \in E_i^+)$, $i \in [0,r]$, a possible distribution of users following the transitions defined by E_i^+. Thus, any such sequence N_i^+ satisfies $\sum_{(i,j)\in E_i^+} n_{i,j} \leq n_i$.

The aggregate sequence $N^+ = (N_0^+, N_1^+, \ldots, N_r^+)$ then denotes a possible distribution of users that may move along all transitions in $\{E_0^+, E_1^+, \ldots, E_r^+\}$ during the interval $[t, t + t_{predict}]$. For example, in Fig. 1, if $N = (7,2,3,4,4)$ is a feasible occupancy distribution at time t, and users follow the transitions according to the distributions: $N_0^+ = (n_{0,1} = 2)$, $N_1^+ = (n_{1,0} = 1, n_{1,1} = 0, n_{1,2} = 1)$, $N_2^+ = (n_{2,1} = 2, n_{2,2} = 0, n_{2,3} = 0)$, $N_3^+ = (n_{3,2} = 1, n_{3,3} = 1, n_{3,4} = 2)$, and $N_4^+ = (n_{4,3} = 1, n_{4,4} = 2)$, then the resulting new occupancy distribution N' equals $(6,4,2,2,4)$, where for example, at time $t + t_{predict}$ ring 2 new population is given by $n_{1,2} + n_{2,2} + n_{3,2} = 2$, and one user out of the active $n_2 = 3$ users at time t has completed a connection during the prediction interval. Finally, we denote by$M(n_0, p_0, n_1, p_1, n_2, p_2, \ldots)$ the value of the multinomial distribution where n_i is the number of occurrences of the ith outcome, and p_i is the probability that the ith outcome occurs.

Computation of the Overload Probability. Theorem 1 outlines a method for computing the overload probability for any given mobility diagram and test distribution N; the running time is polynomial in the number of users (but exponential in the number of rings). The method requires the computation of certain multinomial coefficients. For simplicity of presenting the timing analysis, we assume that such numerical coefficients are precomputed and stored in suitable lookup tables.

Theorem 1. Given a non-overload occupancy distribution $N = (n_0, n_1, \ldots n_r)$, the overload probability $P_{N,\mathcal{OL}}$ can be computed in time $O\left(\prod_{i=0}^{r} n_i^{|E_i^+|}\right)$.

Proof. Given a distribution $N = (n_0, n_1, \ldots n_r)$ that does not cause an overload condition at time t, computing $P_{N,\mathcal{OL}}$ can be implemented (exhaustively) as follows. Denote by $\mathcal{N}^+$ the set of all possible sequences $N^+ = (N_0^+, N_1^+, \ldots, N_r^+)$ in which each sequence N^+ leads to an overload distribution N' at time $t + t_{predict}$ starting from N. We note that the probability of taking all transitions in any such sequence N^+, given that the system is in state N, denoted $P_{N^+|N}$, is given by

$$P_{N^+|N} = \prod_{i=0}^{r} P_{N_i^+|n_i} \tag{4}$$

where each factor $P_{N_i^+|n_i}$ is the probability that all transitions in the sequence N_i^+ occur given n_i users in ring i. Moreover, each factor $P_{N_i^+|n_i}$ is given by a multinomial expression. For our running example of Fig. 1, we have

$$P_{N_2^+|n_2} = M(\ n_{2,1},\ p_{d,2}p_{2,1},\ n_{2,3},\ p_{d,2}p_{2,3},\ n_{2,2},\ p_{r,2},\ n_2 - \textstyle\sum_{j=1}^{3} n_{2,j},\ 1 - p_{r,2} - p_{d,2}).$$

By definition, the required probability is

$$P_{N,\mathcal{OL}} = \sum_{N^+ \in \mathcal{N}^+} P_{N^+|N}. \tag{5}$$

The result then follows by observing an upper bound on the size of the set $\mathcal{N}^+$. ■

Encoding Safe Distributions. We call an occupancy distribution sequence N $(R_{sat}, t_{predict})$-safe (or *safe* for short) if N is not an overload distribution that satisfies the admission condition $P_{N,\mathcal{OL}} \leq P_{admit}$. In addition, we denote the set of all safe distributions by $\mathcal{S}$. The proof of Theorem 1 outlines a computational method that can be used to generate the set $\mathcal{S}$. The offline phase can then compute and store $\mathcal{S}$ in a suitable data structure that supports searching in constant time (e.g., a B+-tree). The online phase then simply searches the data structure mentioned above for any given distribution N under test. A more efficient design, however, strives to find a compact representation of $\mathcal{S}$. One approach to achieve this goal is to use the *maximal* members of $\mathcal{S}$, as defined next. For two occupancy distributions $X = (x_0, x_1, \ldots x_r)$ and $Y = (y_0, y_1, \ldots, y_r)$ we write $X \leq Y$ (or $X < Y$) if for each possible index $i \in [0,r]$ we have $x_i \leq y_i$ (respectively, $x_i < y_i$). We call a distribution $X \in \mathcal{S}$ maximal if there is no distribution $Y \in \mathcal{S}$ such that $X \leq Y$. Theorem 2 (part 1) below provides a key result that permits the use of maximal safe distributions to perform CAC decisions. In particular, the result implies that for a maximal safe distribution Y, and another distribution X: if $X \leq Y$ then X is necessarily safe, moreover, if $X > Y$ then X is necessarily unsafe,

Theorem 2.

1. For any pair of distributions X and Y, if $X \leq Y$ and $Y \in \mathcal{S}$ then $X \in \mathcal{S}$.
2. The CAC can be implemented to accept and reject any test distribution in constant time.

Proof.
Part 1. Let X and Y be two feasible distributions where $X \leq Y$ (thus, $x_i \leq y_i$ for each $i \in [0,r]$). We show that $P_{X,\mathcal{OL}} \leq P_{Y,\mathcal{OL}}$. To this end, we map disjoint events that cause X to yield overload distributions into disjoint events that cause Y to yield overload distributions. To describe the mapping, for each possible ring i, we partition

the y_i active users in Y into two classes: the *green*, and the *yellow* users, where x_i users of y_i are green, and the remaining $y_i - x_i$ users are yellow. A possible event accounted for by equation (5) corresponds to a possible sequence $X^+ = (X_0^+, X_1^+, \ldots, X_r^+)$ that causes X to yield an overload distribution. The sequence X^+ of transitions when followed by the green users of Y (combined with any behaviour of the yellow users) leads to a set of overload distributions. Moreover, such set of distributions is obtained from Y uniquely through the transitions of X^+. The proof then follows by noting that $P_{X^+|X} = \prod_{i=0}^{r} P_{X_i^+|x_i}$, and for each ring i, $P_{X_i^+|x_i}$ is exactly the probability that the green users of Y take all transitions in X_i^+ while the yellow users behave in any possible way.

Part 2. The set of maximal distributions of $\mathcal{S}$ can be stored in a B^+-tree structure with at most $r+1$ levels. Such a tree requires $O(r)$ time to search; this time requirement is constant not depending on the number of users in the distribution N under test. Hence, the CAC can decide whether a given distribution N is less than or equal to some stored distribution in constant time, as required. ■

In summary, the set of maximal safe distributions provides a compact representation of the space of all safe distributions that enables efficient processing of admission requests.

5. Adaptive Bandwidth Allocation

The proposed ABA scheme aims at allocating bandwidth so as to maximize the sum of the average effective throughput and the average satisfaction ratio. Feasible settings of the allocated data rates (at each decision making instant) are constrained by $\mathcal{P}_T$, and the requirement of allocating sufficient transmission power to each mobile to ensure that the receiver can overcome the experienced MAI and achieve the desired bit error rate. We now address the design of three basic heuristic functions required by the ABA module.

Function `terminate()` is invoked in severe situations where there is insufficient base station transmission power to continue serving each of the admitted streams with an acceptable data rate. In this case, one (or more) streams are chosen for termination. The function works by iteratively terminating streams until a feasible allocation of base station transmission power to all remaining mobiles exists. To maximize the average effective throughput, at each iteration the function selects for termination a connection that has received the least amount of traffic thus far.

Function `rate_reduce()` is invoked in less severe situations when $\mathcal{P}_T$ suffices to serve each of the admitted streams with an acceptable data rate after reducing the currently allocated data rate of a subset of connections. To maximize the objective criterion, the function starts by lowering the rate for connections that score high according to the ordered pair $(R_i, \rho_{sat,i})$. That is, if at some instant, two connections are allocated the same data rate R_i but have different satisfaction ratios then the connection that has enjoyed a higher satisfaction ratio is considered first for bandwidth degradation. Additionally, if two connections are allocated different data rates, then the one with higher data rate is considered first for bandwidth degradation. The heuristic aims at maximizing the class average satisfaction ratio ρ_{sat}.

Finally, function `rate_increase()` is invoked when more transmission power becomes available to the base station as a result of either successful completion of a stream, forced termination of a stream, or an active user handoff to a neighbouring cell. The function attempts to maximize the objective function by favouring streams that score lowest according to the ordered pair $(R_i, \rho_{sat,i})$. In summary, the terminate function aims at maximizing the effective throughput, whereas the later two functions aim at maximizing the average satisfaction ratio.

6. Single Cell Analysis

Simulation Parameters. We illustrate the effectiveness of the proposed mechanisms using the mobility transition diagram of Fig. 1. The average length of the exponentially distributed multimedia streaming connection is set to $1/\mu = 500$ seconds. The CAC uses a prediction interval of $t_{predict} = 60$ seconds. For the streaming QoS class, we use $R_{min} = 96$ Kbps, $R_{sat} = 192$ Kbps, $R_{max} = 256$ Kbps (after convolution coding), and vary P_{admit} over the values in $\{0.1, 0.2, 0.3\}$. To test the system under relatively high mobility conditions, we set the average residence time of a user in any ring $i \in [1,4]$ to $1/\delta_i = 100$ seconds. Thus, the ring-i residence probability is $p_{r,i} = e^{-(\mu+\delta_i)t_{predict}} \approx 0.487$, and the departure probability $p_{d,i} = 1 - e^{-\delta_i t_{predict}} \approx 0.451$ (thus, $1 - p_{r,i} - p_{d,i} \approx 0.062$ is the probability that a connection terminates in the same ring it has been initiated in). For a ring $i \in [2,3]$, we assume that users depart to one of the two neighbouring rings with equal probability. To simplify the analysis, we assume that users do not leave the cell, and new users do not enter the cell.

For the target cell and air-interface parameters, we assume that the streaming QoS class is allocated a total base station transmission power ($\mathcal{P}_T$) of 50 Watts in a target cell of radius 1000 meters. Furthermore, we assume standard W-CDMA chip rate of 4.096 Mcps, convolution coding rate= $1/3$, noise spectral density (η_0)= -174 dBm, orthogonality factor (γ)= 0.2, and E_b/I_0 requirement= 7 dB. Large scale path loss is predicted using the log-normal shadow fading model with exponent= 4, and standard deviation

$= 4$ dB. The offered traffic (in Erlang) is the product of the average connection arrival rate to the cell and the average connection duration time. In contrast, the admitted load is calculated based on the average rate of admitted connections. In the experiments, the offered load is varied from moderate to heavy load.

Results and Discussion. We now compare the performance of the proposed predictive CAC (for $P_{admit} = 0.1, 0.2,$ and 0.3) with a non-predictive CAC that admits new connections based on the current state of the cell at the decision making instants. We start by examining the achieved forced termination probability, and the achieved user satisfaction ratio (ρ_{sat}) as the primary user oriented performance measures. Figures 2.a and 2.b show the obtained performance measures; other measures that are found to exhibit monotonic and slowly changing distributions are summarized in Table 1 using their mean (μ) and standard deviation (σ) values.

For the forced termination probability, we remark that setting P_{admit} of the predictive CAC to any of the three test values mentioned above rarely result in terminating an admitted stream when the offered load changes in the test range of $[19, 32]$ Erlang; the resulting termination probabilities are consistently near zero values (cf. Table 1 for $P_{admit} = 0.2$). In contrast, for the non-predictive CAC, Fig. 2.a illustrates that such probability increases significantly as the offered load increases, where in the worst case, the system terminates on the average more than 40% of the admitted connections.

Table 1. Measures described by μ and σ

Non-predictive CAC Results:	μ	σ
ρ_{sat}	0.45	0.03
Wasted Energy (Watt-sec)	1.85	0.32
Predictive CAC Results:		
ρ_{sat} ($P_{admit} = 0.2$)	0.0	0.07
Forced Termination Probability ($P_{admit} = 0.3$)	0.0	0.001

The results on the achieved average satisfaction ratio show a similar trend. Namely, the predictive CAC consistently achieves nearly perfect ρ_{sat} values for all of the three test settings of P_{admit}, over the $[19 - 32]$-Erlang offered load range. In contrast, Table 1 shows that on average the admitted streams operate in a degraded mode about 45% of the time.

The above behaviour is attributed to the fact that the predictive CAC scheme admits considerably less traffic streams than the non-predictive CAC. Indeed, Fig. 2.b shows that the load admitted by the non-predictive scheme often exceeds twice the load admitted by the predictive CAC for all three test settings of P_{admit}. The figure also implies that the predictive CAC results in a higher call blocking probability than the non-predictive CAC. Similar trade offs between the call blocking probability and the forced termination probability have been reported in the literature.

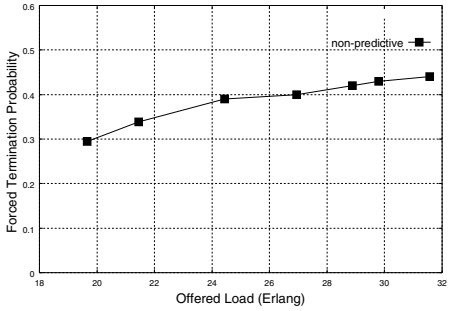

Figure 2.a Forced termination probability

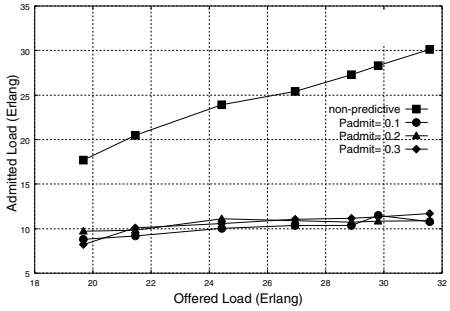

Figure 2.b Admitted load

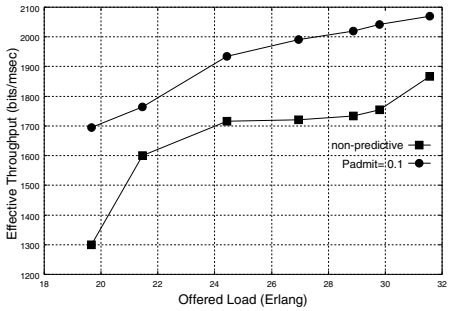

Figure 2.c The average effective throughput

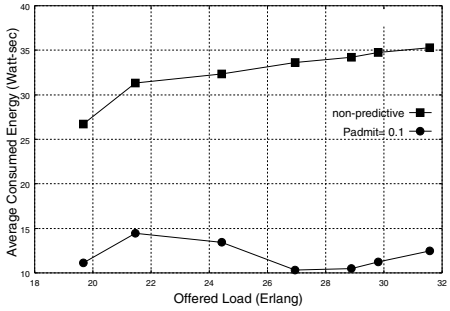

Figure 2.d The average consumed energy

For network provider oriented measures, we note that Fig. 2.c on the achieved average effective throughput shows

a rather unintuitive result: the predictive CAC ($P_{admit} = 0.1$) consistently outperforms the non-predictive CAC with respect to the achieved effective throughput. Results on using $P_{admit} = 0.2$ and 0.3 are omitted from the diagram, however, they show further improvements over the $P_{admit} = 0.1$ case. Likewise, results on the average base station energy incurred in transmitting data streams that have been served to completion (Fig. 2.d) shows that the predictive CAC ($P_{admit} = 0.1$) consistently uses less than half the energy used by the non-predictive CAC in delivering the completed flows (again, similar results holds for $P_{admit} = 0.2$ and 0.3).

From the above, one may conclude that although the predictive CAC rejects more connections than the non-predictive CAC, the former achieves superior performance in terms of increased average effective throughput, and reduced average base station power consumption. Finally, we note that the non-predictive CAC wastes an average of 3.7% of the total base station energy (1.85 Watts-sec) on the forcibly terminated flows (see Table 1), whereas the predictive CAC results in a negligible wasted energy consumption.

7. Concluding Remarks

To date, various aspects of managing network resources to accommodate multimedia traffic in 3G networks have been investigated. The work done in this paper emphasizes the need to provision multimedia streaming services at relatively high data rates while maintaining session's quality during user movements. The performance of the proposed predictive admission control scheme, combined with an adaptive rate allocation scheme, show significant performance gains compared to using an admission scheme based on the state of the cell at admission time. The obtained results motivate further investigation on the performance of a more elaborate scheme that incorporates the proposed scheme with a suitable end-to-end framework for provisioning quality of service to the multimedia streaming class. We leave such investigation as a possible future research direction.

ACKNOWLEDGMENT

This research is supported by NSERC Canada and the Canadian Institute for Telecommunications Research (CITR).

References

[1] S.-E. Elayoubi, T. Chahed, and G. Hébuterne. Admission control in UMTS in the presence of shared channel. *Computer Communications*, 27:1115–1126, 2004.

[2] L. Jorguseski, J. Farserotu, and R. Prasad. Radio resource allocation in third-generation mobile communication systems. *IEEE Communications Magazine*, 39:117–123, February 2001.

[3] N. S. Joshi, S. R. Kadaba, S. Patel, and G. S. Sundaram. Downlink scheduling in CDMA data networks. In *MobiCom 2000*, August 2000.

[4] T. Kwon, Y. Choi, C. Bisdikian, and M. Naghshineh. QoS provisioning in wireless/mobile multimedia networks using an adaptive framework. *Wireless Networks*, 9:51–59, 2003.

[5] B. Li, L. Yin, K. Wong, and S. Wu. An efficient and adaptive bandwidth allocation scheme for mobile wireless networks using an on-line local estimation technique. *Wireless Networks*, 7(2):107–116, March 2001.

[6] H. Montes, G. Gomez, J. Paris, and R. Cuny. Deployment of IP multimedia streaming services in third-generation mobile networks. *IEEE Wireless Communications*, 9:84–92, October 2002.

[7] D. Shen and C. Ji. Admission control of multimedia traffic for third generation CDMA network. In *INFOCOM 2000*, volume 3, pages 1077–1086, March 2000.

[8] S. Singh and S. K. Tripathi. A time-slotted-cdma architecture and adaptive resource allocation method for connections with diverse QoS guarantees. *Wireless Networks*, 9:497–494, September 2003.

[9] M. Soleimanipor, W. Zhuang, and G. Freeman. Optimal resource management in wireless multimedia wideband CDMA systems. *IEEE Transaction son Mobile Computing*, 1(2):143–160, April-June 2002.

[10] A. Viterbi. *CDMA: Principles of Spread Spectrum Communications*. Addison-Wesley, 1995.

[11] X. Wang, R. Ramjee, and H. Viswanathan. Adaptive and predictive downlink resource management in next generation CDMA networks. In *INFOCOM 2004*, volume 4, pages 2754–2765, 2004.

[12] S. Wu, K. 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*, 10(2):257–271, April 2002.

[13] L. Xu, X. Shen, and J. W. Mark. Dynamic bandwidth allocation with fair scheduling for WCDMA systems. *IEEE Wireless Communications*, pages 26–32, April 2002.

[14] F. Yu, V. Wong, and V. Leung. A new QoS provisioning method for adaptive multimedia in cellular wireless networks. In *INFOCOM 2004*, volume 3, pages 2130–2141, 2004.

[15] J. Zander and S.-L. Kim. *Radio Resource Management for Wireless Networks*. Artech House Publishers, 2001.