A Framework for Call Admission Control and QoS Support in Wireless Environments

AbdulRahman Aljadhai[†]
Taieb F. Znati[†,§]
[†]Department of Computer Science
[§]Telecommunications Program
University of Pittsburgh
Pittsburgh, PA 15260
(jadhai, znati)@cs.pitt.edu

Abstract— **With the proliferation of wireless networks technologies, mobile users are expected to demand the same quality-of-service (QoS) available to fixed users. This paper presents a framework to support *predictive timed-QoS guarantees*, in wireless environments. The main components of this framework include a service model for QoS guarantees, a path predictability model, and a call admission control scheme. The unique feature of this framework is the ability to combine the path predictability model with the call admission control to determine the level of predictive timed-QoS guarantees that the network can provide over a predetermined interval of time. The performance of the call admission control scheme, in terms of the dropping ratio of hand-off calls and the blocking ratio of new calls, is presented.**

I. Introduction

Mobile users are expected to operate in an environment in which the quality of service (QoS) can vary significantly within and across different types of wireless networks, ranging from high speed indoor wireless LANs to very low speed outdoor wireless WANs. The continuous support of minimum QoS guarantees of multimedia applications poses a major challenge in heterogeneous, wireless environments [7], [10]. Such a support requires the development of a flexible and adaptable network resource management framework. The framework must provide efficient mechanisms to bridge the heterogeneity gap between different types of networks, resolve potential QoS mismatch as mobile units move from one coverage to another and dynamically accommodate applications with different QoS requirements in response to network performance degradation.

Several schemes have recently been proposed to support QoS provision in a wireless environment[3], [9], [13], [8], [1]. In [3], an application adaptation framework to support QoS negotiation, monitoring and notification is proposed. The framework defines *Adaptation Blocks* as sequence of segments that comprises the total execution time of the application. QoS re-negotiation is performed at the beginning of every Adaptation Blocks.

In [1], a *virtual connection tree* based scheme is proposed. In this scheme, a set of adjacent cells are grouped into a cell-cluster in a static fashion. Upon the admission of a call, the scheme pre-establishes connections between a root switch and each base station in the cell-cluster. The scheme builds cell-clusters in a static fashion regardless of the user mobility. This may result in an unnecessary resource overloading that may under-utilize the network resources and cause severe congestion.

The *shadow cluster* scheme supports a distributed call admission control based on the estimated bandwidth requirements in the *shadow cluster*[8]. A *shadow cluster* is a collection of base stations to which a mobile unit is likely to attach in the future. The call admission decision is made by all base stations within the *shadow cluster*. The scheme partitions the time into predefined intervals and verifies the feasibility of supporting a call over these intervals. The scheme requires the exchange of a large number of messages between base stations during each time interval to verify the feasibility of admitting a call. Moreover, the bandwidth estimates are calculated at the beginning of each time interval, while the admission decisions are made at the end of each time interval; therefore, admission of new calls is delayed for at least a time equal to these predefined intervals. Furthermore, the shadow cluster scheme, as the virtual connection tree scheme, lacks the mechanisms to predict the mobile's trajectory and determine the future cells that the mobile unit may visit.

In order to efficiently support QoS guarantees in a wireless network regardless of the users' mobility, an accurate estimation of the mobile's trajectory as well as the arrival and departure times for each cell along the path is required. Using these estimates, the network can determine if enough resources are available, in each cell along the mobile's path, to support the QoS requirements of the call. In this paper, we present a framework for *predictive timed-QoS guarantees* in wireless networks. The framework is designed to gracefully accommodate dynamic vari-

0-7803-5417-6/99/$10.00 ©1999 IEEE.

ations in network resources. The basic components of this framework are: (i) a predictive service model to support timed-QoS guarantees, (ii) a mobility model to determine the mobile's *most likely cluster* (MLC). The MLC represents a set of cells that are most likely to be visited by a mobile unit during its itinerary, and (iii) a call admission control model to verify the feasibility of supporting a call within the MLC.

The service model accommodates different types of applications by supporting *integral* and *fractional* predictive QoS guarantees over a predefined time-guarantee period. The MLC model is used to actively predict the set of cells that are most likely to be visited by the mobile unit. For each MLC cell, the mobile's earliest arrival time, latest arrival time and latest departure time are estimated. These estimates are then used by the call admission control to determine the feasibility of admitting a call by verifying that enough resources are available in each of the MLC cells during the time interval between the mobile's earliest arrival time and its latest departure time. If available, resources are then *reserved* for the time interval between the mobile's earliest arrival time and latest departure time, and *leased* for the time interval between the mobile's earliest and latest arrival times. If the mobile unit does not arrive before the lease expires, the reservation is canceled and the resources are returned to the pool of available resources. The unique feature of the proposed framework is the ability to combine the mobility model with the the call admission control model to determine the level of predictive timed-QoS guarantees that the network can provide to a call, and dynamically adjust these guarantees as the mobile unit moves from one cell to another.

The rest of the paper is organized as follows: Section II presents the *predictive timed-QoS guarantees* service model to support different types of applications in wireless environments. Section III describes the *most likely cluster* model. Section IV discusses the call admission control scheme used to handle new and hand-off calls. Performance evaluation of the proposed scheme to support predictive timed-QoS guarantees is presented in section V. Section VI presents the conclusion of this work.

II. Predictive Timed-QoS Guarantees Service Model

The support of deterministic QoS guarantees in wireless environment can only be achieved if resources are allocated for the duration of the call in all the cells which will be visited by the mobile unit. This is only feasible if exact knowledge of the mobile's path and the arrival and departure times to and from each cell along the path are available for the duration of the call. Acquiring such a knowledge, however, is difficult in a wireless environment characterized by a high degree of uncertainty both in resource availability and mobile mobility.

In order to provide efficient network support to multimedia applications in wireless environments, a balance between an acceptable level of QoS guarantees and a high level of network resource utilization is required. To achieve this balance, a *predictive timed-QoS guarantees* service model, is proposed. In this model, a predictive level of QoS support is guaranteed by reserving resources in advance in each MLC cell of the mobile unit. Furthermore, these reservations only extend for a time duration equal to the time interval the mobile unit is expected to spend within a cell starting from the time of its arrival to that cell until the time of its departure.

The predictive timed-QoS guarantees based service model is characterized by a set of application dependent parameters, namely a time guarantee period, T_G, a cluster-reservation threshold, τ, and a bandwidth-reservation threshold, γ. T_G specifies the time duration for which the required QoS level is guaranteed. τ defines the minimum percentage of the MLC cells that can support the required QoS level for the guarantee period T_G. γ represents the minimum percentage of the required bandwidth that must be reserved in every MLC cell.

To accommodate different types of applications, the service model provides two types of predictive services, namely *integral guaranteed* service and *fractional guaranteed* service. The integral guaranteed service guarantees that *all* MLC cells can support the requested bandwidth requirements for the lifetime of the call. In this case, T_G must be equal to the call duration and τ and γ are both equal to 100%. The fractional guaranteed service, on the other hand, ensures that at least τ percent of the MLC cells can support at least γ percent of the requested bandwidth requirements for the next T_G interval. A special case arises when either τ or γ are zeros. In this case, the service is referred to as *best-effort*.

III. Most Likely Cluster (MLC) Model

In order to efficiently support QoS guarantees in a wireless network, bandwidth must only be reserved in those cells that are most likely to be visited by the mobile unit for the duration of the residence time within each cell. This, however, requires a proper characterization of the "most likely to be visited cell" concept and an accurate estimate of the residence time of a mobile unit within that cell. The MLC model addresses these issues in a manner which takes into consideration the network efficiency and the predictive aspect of QoS support in the predictive timed-QoS guarantees framework.

Based on the MLC model, the "most likely to be visited" property of a cell is directly related to the position of the cell with respect to the estimated direction of the mobile unit. This likelihood is referred to as *directional probability*. Therefore, cells that are situated along the mobile unit's direction have higher directional probabili-

ties, and are more likely to be visited, than those that are situated outside of this direction.

At any point in time during the call, the MLC represents a collection of contiguous cells each of which is characterized by a directional probability that exceeds a certain threshold. For each MLC cell, the expected time of arrival and departure times of the mobile are estimated. Using these estimates, the feasibility of supporting the requested level of timed-QoS guarantees during the mobile's residence time within each cell along path is verified. In the following, we present the direction prediction method, the MLC formation scheme and the algorithm used to estimate the expected times of arrival and departure to a given cell within the MLC.

A. Direction Prediction Method

The method used to predict the mobile user's direction is based on the history of its movement. It is clear, however, that the prediction method used should not be "largely" affected by small deviations in the mobile direction. Furthermore, the method should converge rapidly to the new direction of the mobile unit.

To take the above properties into consideration, a first order autoregressive filter, with a smoothing factor α, is used[4]. More specifically, let D_0 be the current direction of the mobile unit when the call is made[1]. Notice that when the mobile is stationary within a cell it is assumed that the current cell is the only member of the MLC, so reservations are done only within the current cell. If D_t represents the observed direction of the mobile unit at time t, the predicted direction, $\tilde{D}_{t+1}$, at $t+1$ is obtained as follows:

$$\tilde{D}_{t+1} = (1-\alpha)\tilde{D}_t + \alpha D_t \tag{1}$$

In order to track the actual direction of the mobile unit more accurately, the smoothing factor α is computed as:

$$\alpha = c\frac{E_s^2}{\sigma_{s+1}} \tag{2}$$

where $0 < c < 1$, $E_s = D_s - \tilde{D}_s$ is the prediction error, and σ_s is the average of the past square prediction errors at time s. σ_s can be expressed as follows:

$$\sigma_{s+1} = cE_s^2 + (1-c)\sigma_s \tag{3}$$

The first order autoregressive filter used to predict future directions of the mobile unit, combined with the method used to adaptively compute the parameter α, guarantees that the predicted mobile direction is not effected by small deviations in the mobile direction. In addition, the method has the ability to quickly track down abrupt changes in the actual direction of the mobile unit.

[1]The network support can determine the current direction of a mobile unit based on base station measurements or using global positioning systems [11]

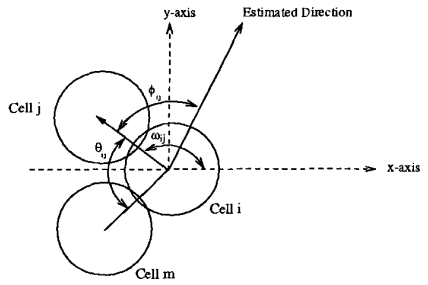

Fig. 1. Parameters used to calculate the directional probability.

B. Directional Probability Derivation

At any point in time t, the directional probability of any cell being visited next by a mobile unit can be derived based on the current cell, where the mobile resides, and the estimated direction $\tilde{D}_t$ of the mobile unit at time t. The basic property of this probability distribution is that for a given direction, the cell that lies on the estimated direction from the current cell has the highest probability of being visited in the future [2].

More specifically, consider a mobile unit currently residing at cell i coming from cell m, and let j, $j = 1, 2, ...,$ represent a set of adjacent cells to cell i. Each cell j is situated at an angle ω_{ij} from the horizontal axis passing by the center of cell i, as depicted in Figure 1. Furthermore, define the *directional path* from i to j as the direct path from the center of cell i to the center of cell j.[2]

Based on the directional path, the *directionality*, D_{ij}, for a given cell j can be expressed as:

$$D_{ij} = \begin{cases} \frac{\theta_{ij}}{\phi_{ij}}, & \phi_{ij} > 0 \\ \theta_{ij} & \phi_{ij} = 0 \end{cases} \tag{4}$$

where ϕ_{ij} is an integer representing the deviation angle between the straight path to destination and the directional path from i to j, while θ_{ij} represents the angle between the directional path from m to i and the directional path from i to j.

Based on its directionality D_{ij}, the directional probability, $P_{i\to j}$, of cell j being visited next by a mobile unit currently at cell i can be expressed as follows:

$$P_{i\to j} = \frac{D_{ij}}{\sum_k D_{ik}} \tag{5}$$

where k is a cell at the same *ring* as j with respect to i. A cell k is said to be at *ring* L with respect to cell i, if it is located L cells away from i.

For a given cell i, the directional probabilities, $P_{i\to j}$, provide the basis upon which MLCs are formed as the mobile units moves from one cell to another.

[2]Notice that the directional path concept is only used in the mathematical formulation to derive directional probabilities and does not infer any restrictions over the movement of the mobile unit within a cell. A mobile unit may travel from any point in a cell to any other point in an adjacent cell.

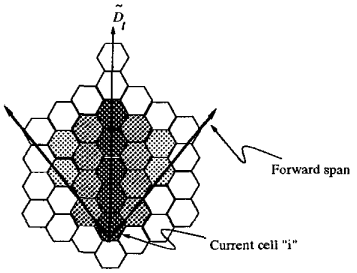

Fig. 2. Definition of the Most Likely Cluster.

C. MLC Formation

The MLC concept can be better explained based on a typical behavior of a mobile unit. Starting from the cell where the call originated, a mobile unit is expected to progress toward its destination. The unit, however, can temporarily deviate from its long-term direction to the destination, but is expected to converge back at some point in time toward its destination. Therefore, this mobility behavior can be used to determine the cells that are likely to be visited by a mobile unit. Define the forward span as the set of cells situated within an angle, δ_i, with respect to the estimated direction $\tilde{D}_t$ of the mobile unit. An instance of a forward span of a mobile unit currently at cell i progressing along the estimated direction $\tilde{D}_t$ is depicted in Figure 2.

Based on the directional probabilities and the definition of a forward span, the MLC of a given mobile unit, u, currently located at cell i, denoted as $\mathcal{C}_i^{\text{MLC}}(u)$, can be expressed as:

$$\mathcal{C}_i^{\text{MLC}}(u) = \{ \text{ cells } j \mid \phi_{ij} \leq \delta_i \ , \ j = 1, 2, ...\} \tag{6}$$

where ϕ_{ij} is the deviation angle between the straight path to destination and the directional path from i to j. The angle δ_i is defined such that $P_{i \to j} \geq \mu$, where μ represents a system defined threshold on the likelihood that cell is to be visited. More specifically, δ_i can be expressed as:

$$\delta_i = \max[\phi_{ij}] \text{ such that } P_{i \to j} \geq \mu \tag{7}$$

After determining the forward span the next step in the process of forming the MLC is to decide on the size of the *MLC window*. The MLC window is defined as the number of adjacent rings of cells to be included in the MLC.

The size of the MLC window has a strong impact on the performance of the scheme. Increasing the MLC window size, by including more rings, increases the likelihood of supporting the required QoS if the mobile moves along the predicted direction $\tilde{D}(t)$. On the other hand, if the mobile deviates from the predicted direction, increasing the MLC window size may not ensure the continued support of the call, as the mobile unit may move out from the MLC. Our approach is to reward users who move within the predicted direction by increasing their MLC window size up to a maximum, R_{max}. The value R_{max} depends on the value of the guarantee period T_G. Higher values of T_G result in larger values of R_{max}.

When the user deviates from the estimated direction, the MLC window size is decreased by an amount proportional to the degree of deviation. As a result, support of the predictable users' QoS requirements can be achieved with high probability, whereas users with unpredictable behavior do not unnecessarily overcommit the network resources. The proposed algorithm dynamically updates the size of the MLC window based on the observed movement patterns of the mobile users.

Define Dev_t to be the measure of the mobile's deviation with respect to the estimated direction at time t as:

$$\text{Dev}_{t+1} = \beta \cdot \text{Dev}_t + (1 - \beta)|\tilde{D}_t - D_t| \tag{8}$$

where $0 < \beta < 1$ and Dev_0 equals 0. The MLC window size, W_{MLC}, at time t can be derived as follows:

$$W_{MLC} = \min(R_{max}, \lceil (1 - \frac{\text{Dev}_t}{2\pi}) \cdot R_{max} \rceil). \tag{9}$$

The MLC window size is recalculated at every hand-off; therefore the window size shrinks and grows depending on the mobile behavior.

D. Expected Arrival and Departure Times

The call admission control scheme uses the expected cell residence time of a mobile unit to determine the feasibility of supporting its QoS requirements. The cell residence time within cell j for a mobile unit currently in cell i is characterized by three parameters, namely, expected earliest arrival time ($T_{\text{EA}}(i,j)$), expected latest arrival time ($T_{\text{LA}}(i,j)$) and expected latest departure time ($T_{\text{LD}}(i,j)$). Consequently, $[T_{\text{EA}}(i,j), T_{\text{LD}}(i,j)]$ is the expected residence time of the mobile unit within cell j. This interval is referred to as the *resource reservation interval* (RRI), while the interval $[T_{\text{EA}}(i,j), T_{\text{LA}}(i,j)]$ is referred to as the *resource leasing interval* (RLI), as depicted in Figure 3. Resources are reserved for the entire duration of RRI. However, if the mobile does not arrive to cell j before RLI expires, all resources are released and the reservation is canceled. This is necessary to prevent mobile units from holding resources unnecessarily.

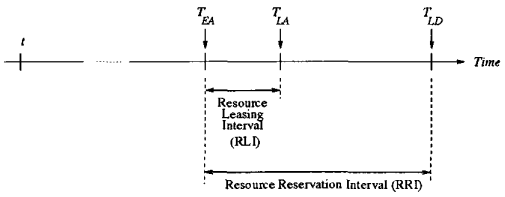

Fig. 3. Resource leasing and reservation intervals within a cell.

In order to derive these time intervals, one can adopt the method used in the the shadow cluster and consider all possible paths from the current cell to each cell in the cluster[8]. This method can be complex, since there are many possible paths that a mobile unit may follow to reach a cell. The approach taken in the predictive timed-QoS guarantees framework is based on the concept of most likely paths.

Consider a mobile unit, u, currently located at cell m, and let $\mathcal{C}_m^{\text{MLC}}(u)$ denote its MLC. Define $G = (V, E)$ to be a directed graph, where V is a set of vertices and E a set of edges. A vertex, $v_i \in V$, represents MLC cell i. For each cell i and j in $\mathcal{C}_m^{\text{MLC}}(u)$, an edge (v_i, v_j) is in E if and only if j is a *reachable direct neighbor* of i. Each directed edge (v_i, v_j) in G is assigned a cost $1-P_{i\to j}$, where $P_{i\to j}$ represents the directional probability as defined in Equation 5.

A path, Π, between MLC cells i and k, is defined as a sequence of edges, $(v_i, v_{i+1}), (v_{i+1}, v_{i+2}), \cdots, (v_{k-1}, v_k)$. Furthermore, the cost of a path, Π, between MLC cells i and k, is derived from the cost of its edges so that the least costly path represents the most likely path to be followed by the mobile. A *k-shortest paths* algorithm [5] is then used to obtain the set, $\mathcal{K}$, of *k-most likely paths* to be followed by the mobile unit.

For each path, $\Pi \in \mathcal{K}$, between MLC cell i and j, we define the *path residence time* as the sum of the residence time of each cell in the path. Let Π_{S} and Π_{l} in $\mathcal{K}$, represent the paths with the shortest and longest path residence time, respectively. Π_{S} is used to derive the expected earliest arrival time, $T_{\text{EA}}(i,j)$, while Π_{l} is used to derive expected latest arrival, $T_{\text{LA}}(i,j)$. More specifically, $T_{\text{EA}}(i,j)$ and $T_{\text{LA}}(i,j)$ can respectively be expressed as described as:

$$T_{\text{EA}}(i,j) = \sum_{k\in\Pi_{\text{S}}} R_T(k) \tag{10}$$

$$T_{\text{LA}}(i,j) = \sum_{k\in\Pi_{\text{l}}} R_T(k) \tag{11}$$

where, $R_T(k)$ represents the mean residence time in cell k. The value of $R_T(k)$ depends on whether cell k is the cell where the call originated or a cell to which a call was handed off. Assuming that the mobile units are evenly spread over a cell area of radius D, travel at a constant speed, V with a pdf $f(v)$, along a constant direction within the the cell, the value of $R_T(k)$ can be obtained using the standard methods described in [6], [12]:

$$R_T(k) = \begin{cases} \frac{8DE[1/V]}{3\pi} & \text{if } k \text{ is the originating cell} \\ \\ \frac{\pi D}{2E[V]} & \text{if } k \text{ is a hand-off cell} \end{cases} \tag{12}$$

where $E[V]$ represents the expected value of the mobile's speed. Similarly, the expected latest departure time, $T_{\text{LD}}(i,j)$, from cell j can be computed as follows:

$$T_{\text{LD}}(i,j) = T_{\text{LA}}(i,j) + R_T(j) \tag{13}$$

The estimates of $T_{\text{EA}}(i,j)$, $T_{\text{LD}}(i,j)$, and $T_{\text{LD}}(i,j)$ for a mobile u currently located at cell i are used to compute RLI and RRI for each cell $j \in \mathcal{C}_i^{\text{MLC}}(u)$. The call admission control uses these values to verify the feasibility of supporting u's call in each cell $j \in \mathcal{C}_i^{\text{MLC}}(u)$.

IV. Call Admission Control

The main objectives of the call admission control are: (i) to guarantee an uninterruptable service for admitted calls as they move from one cell to another, and (ii) to maximize the network resource utilization by reserving resources only where needed and within the expected residence time interval. These two objectives may conflict with each other, as guaranteeing uninterruptable service requires reserving resources in a large number of cells while maximizing resource utilization requires limiting resource reservation to those cells which are expected to be visited by the mobile unit. The proposed call admission control strikes a balance between these two extremes and guarantees, with high probability, uninterruptable service without unnecessarily sacrificing the network resource utilization.

Call admission control involves each cell within the MLC and is performed in a distributed fashion. For a call to be feasible the amount of bandwidth required to support the call must be available in every MLC cell. The bandwidth, however, is only *reserved* but not *allocated* to the call. Furthermore, the reservation is only held for the RRI period in every MLC cell.

A. Call Feasibility and Setup Procedure

Based on the requested service guarantees a call is accepted if two conditions are met: First, the bandwidth required to support the requested service guarantees must be available in the cell where the call originates. Second, the γ% of the required bandwidth must also be available for reservation in τ% of the MLC cells. If the decision is to admit the call, the required bandwidth is allocated in the current cell and γ% bandwidth is reserved in the τ% of the MLC cells.

Assume that each cell i has a total bandwidth of C_T^i units. At any given point in time, the unused bandwidth is denoted as $C_f^i(t)$, the allocated bandwidth is denoted as $C_a^i(t)$ and the reserved bandwidth is denoted as $C_r^i(t)$, where $C_T^i = C_f^i(t) + C_a^i(t)$. Both $C_a^i(t)$ and $C_r^i(t)$ change dynamically with time in every cell, as new calls are admitted and existing calls leave the system. Note that $C_a^i(t) + C_r^i(t)$ may be greater than C_T^i.

Let B_u be the bandwidth required by mobile unit u. A new call can be admitted to a cell if the cell can provide $\gamma \cdot B_u$ units of bandwidth for the time duration RRI.

Therefore, for a given cell j, $\gamma \cdot B_u^j$ units of bandwidth must be reserved to guarantee the QoS requirements of mobile unit u currently at cell i. Cell i informs other cells about these requirements by issuing a reservation request containing the following parameters:

$$[\gamma \cdot B_u^j, T_{\text{EA}}(i,j), T_{\text{LA}}(i,j), T_{\text{LD}}(i,j)], \tag{14}$$

Each cell in the MLC uses the reservation request to decide on whether the request can be accepted or rejected. Let $[C_f^j(t) \geq \gamma \cdot B_u]$ represent the decision of cell j to accept or reject a call. $[C_f^j(t) \geq \gamma \cdot B_u]$ can be written as:

$$[C_f^j(t) \geq \gamma \cdot B_u] = \begin{cases} 1, & \text{if } C_f^j(t) \geq \gamma \cdot B_u \\ & \text{for } [T_{\text{EA}}(i,j), T_{\text{LD}}(i,j)] \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

The values of $[C_f^j(t) \geq \gamma \cdot B_u]$ for all cells including i, where the call originates, are collected. Based on these values, cell i admits the call if the following holds:

$$\sum_{j \in \mathcal{C}_i^{\text{MLC}}(u)} P_{i \to j} \cdot [C_f^j(t) \geq \gamma \cdot B_u] \geq \tau \cdot \sum_{j \in \mathcal{C}_i^{\text{MLC}}(u)} P_{i \to j} \tag{16}$$

The above feasibility test uses the directional probability to ensure that cells that are most likely to be visited weigh heavier in the decision of accepting a call. For integral guarantees, the call is accepted if all cells in the MLC can reserve the required bandwidth. On the other hand, for fractional guarantees, a call can be accepted even if some MLC cells cannot reserve bandwidth for the RRI period, as long as τ% of the MLC cells can support the call.

When originating cell i accepts the call, it sends a message to all cells, that are able to reserve bandwidth, indicating that the reservation must now be performed. Every cell reserves the required bandwidth for the RRI duration and updates its available bandwidth accordingly. Notice, however, that cell j cancels the reservation if the mobile unit does not arrive prior to the expiration of the lease period RLI.

B. Prediction Conforming and Non-conforming Calls Handling Policy

Due to the changing behavior of the mobile units in a cellular network, calls may hand-off to cells that are not part of their MLC and calls may not arrive within their lease periods, RLI's. Therefore, we define the status of a call to be either *prediction conforming* or *prediction non-conforming*. When a call is accepted to the network, it is said to be *prediction conforming* as long as it visits cells that belong to its MLC and resides within each cell for the RRI period. A call that arrives to a cell earlier than its earliest arrival time or to a cell which is not part of its MLC is considered *prediction non-conforming*. The same status applies to a call that arrives to a cell after its lease period, RLI, expires. Notice also that an early call remains prediction non-conforming up to the beginning of its RRI period, after which it becomes prediction conforming. Furthermore, a call that remains within a cell beyond its RRI period is also considered prediction non-conforming. Consequently, a call may change status multiple times during its duration.

The call admission control gives higher priority to prediction conforming calls over prediction non-conforming calls. A prediction non-conforming call is dropped when a prediction conforming call arrives and cannot be accommodated by the available bandwidth. When more than one prediction non-conforming call exist within a cell, different policies can be used to decide which call is to be dropped. For instance, the lifetime of a call, bandwidth requirements or call priorities can be used to implement different dropping policies.

The basic steps of the call admission control algorithm used to verify the feasibility of accepting new and hand-off calls are presented next.

C. Call Admission Algorithm for New Calls

Consider a mobile unit u, with a bandwidth requirement B_u, seeking admission to cell i, and let $\mathcal{C}_i^{\text{MLC}}(u)$ its MLC at time t_0. The following steps are executed to verify the feasibility of supporting u's call request:

- Cell i verifies that its current available bandwidth can support the call, i.e. $C_f^i(t_0) \geq B_u$. If $C_f^i(t_0)$ is less than B_u, the call is rejected even if prediction non-conforming calls are in progress within the cell. This is to ensure that existing calls have higher priority over new calls.
- If the call is accepted, cell i sends the reservation request, described in Equation 14, to all cells in $\mathcal{C}_i^{\text{MLC}}(u)$.
- Upon receiving the reservation request, every cell replies based on its available bandwidth for the time interval RRI, as described in Equation 15.
- Upon receiving the reservation replies, cell i executes the feasibility test, described in Equation 16. If the decision is to accept the call, cell i declares the call to be prediction conforming, allocates B_u units of bandwidth to the call, and instructs all cells in $\mathcal{C}_i^{\text{MLC}}(u)$ to reserve $\gamma \cdot B_u$ units of bandwidth to support the call.
- Upon receiving the admission confirmation from cell i, every cell in $\mathcal{C}_i^{\text{MLC}}(u)$ reserves $\gamma \cdot B_u$ units of bandwidth for the RRI period.

D. Call Admission Algorithm for Hand-off Calls

When a hand-off occurs, the call admission control is invoked to ensure that the call can continue to be supported at its desired QoS level. Assume a mobile u, requiring B_u units of bandwidth, hands off to cell j from cell i.

Parameter	*Value*
Total number of cells	37
Total Bandwidth per cell	60kb/s
Required bandwidth Audio/Video	2/20kb/s
Speed	45-70 miles/hr
Cells' diameter	1 mile

TABLE I
SUMMARY OF SIMULATION PARAMETERS

The following steps are executed to verify the feasibility of supporting the QoS requirements of the hand-off call:

- Let j be a cell in $\mathcal{C}_i^{\text{MLC}}(u)$, and assume that u arrives to cell j during the lease period RLI:
  - If B_u units of bandwidth are available then j allocates these units to the hand-off call and admits it as prediction conforming.
  - If B_u units of bandwidth are not available, cell j checks to see if there are prediction non-conforming calls in cell j which when terminated cause u to be admitted. If B_u units of bandwidth, currently held by prediction non-conforming calls, can be freed, the corresponding calls are terminated and the bandwidth is allocated to u. The call is then admitted as prediction conforming.
  - If neither of the above is possible, the call is dropped.
- If the call arrives to cell j and either cell j is not in $\mathcal{C}_i^{\text{MLC}}(u)$ or does not arrive within its lease period, RLI, the call is admitted as prediction non-conforming if cell j can allocate B_u units of bandwidth. Otherwise, the call is dropped.
- If the call is accepted at cell j, $\mathcal{C}_j^{\text{MLC}}(u)$ is computed and reservation requests are sent to all cells in the new MLC. Cells over-reserve bandwidth for hand-off calls, if necessary, to increase the probability of supporting these calls in the future.

V. PERFORMANCE EVALUATION EXPERIMENT

The objective of this experiment is to study the performance of the call admission scheme with respect to three measures, namely *dropping ratio*, *blocking ratio* and *bandwidth utilization*. The dropping ratio represents the percent of calls that are dropped during hand-offs. This is a critical parameter for real-time traffic where dropping a call is undesirable from the user's perspective. The blocking ratio represents the percent of calls that are denied access to the network.

Two types of calls are studied, namely, audio and video. Each call is characterized by its own bandwidth requirement and mean call holding time. Table I lists a summary of the parameters used in the simulator. The call holding time T_H is exponentially distributed with mean 3 minutes for audio and video calls. The call arrival rate is assumed to be the same for all cells and is exponentially distributed

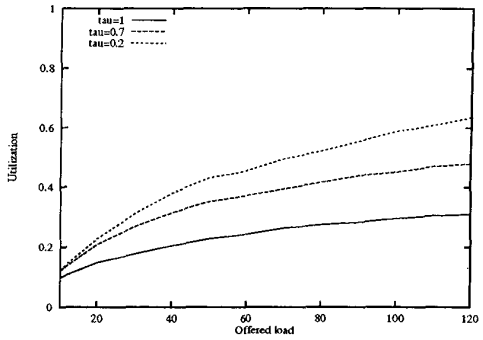

Fig. 4. Utilization vs. Offered load for three different values of τ (tau).

with mean λ. The speed of the mobile unit is uniformly distributed between 45 and 70 miles per hour. The speed changes at the boundaries of the cell, but stays fixed within the cell. The call residence time in the cell is computed based on the distance across the cell and the speed of the mobile unit. γ, the bandwidth-reservation threshold, is assumed to be 1 and τ is varied for different simulation experiments. The offered load is calculated as follows:

$$\text{Offered Load} = T_H \cdot \lambda(B_a \cdot P_a + B_v \cdot P_v) \qquad (17)$$

where B_a and B_v are the required bandwidth for audio and video, respectively. P_a and P_v are percentage of audio and video calls, respectively, and $P_a + P_v = 1$.

The study is conducted in a two dimensional cellular network where hand-offs can occur in any direction. Furthermore, we assume that the cells have a hexagonal shape. The simulated cellular network consists of 37 cells. In the experiment, each new call is assigned randomly a destination cell which can be the same cell where the call originates. In this case the user is assumed to remain in that cell for the entire call duration. Using the mobile's direction the MLC is constructed and bandwidth is reserved in the current cell and every cell in its MLC, following the two call admission algorithms presented in Section IV.

Figures 4, 5 and 6 show the performance of the MLC scheme as a function of the offered load for three values of τ. The traffic is 0.8 audio and 0.2 video. When τ is equal to 1 the dropping ratio is almost equal to zero and is not shown in Figure 5. With a cluster-reservation threshold (τ) equals to 0.7 the dropping ratio is maintained below 0.04. At this threshold less blocking ratio and more utilization is achieved than the 100% cluster-reservation threshold ($\tau = 1$). This shows that based on the expected traffic type, offered load and target dropping ratio, a value for the cluster-reservation threshold can be chosen to meet specific system design goals.

The call admission control presented in this paper is also compared to a *non-timed* call admission control that

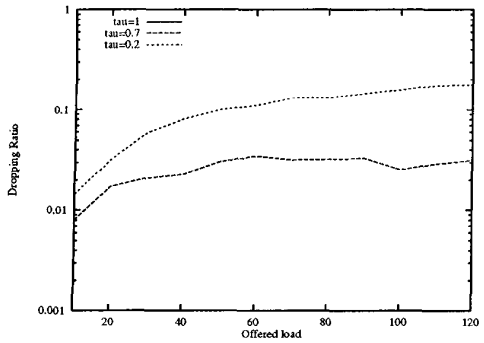

Fig. 5. Dropping ratio vs. Offered load for three different τ (tau) values.

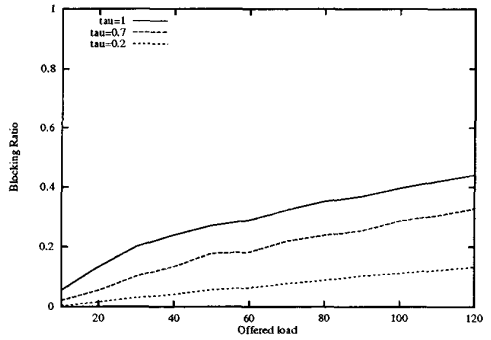

Fig. 6. Blocking ratio vs. Offered load for three different τ (tau) values.

reserve bandwidth in the MLC cells for the entire duration of the call. Figure 7 shows the results of comparing the proposed scheme, timed-QoS guarantees, with the non-timed approach. The timed-QoS guarantees scheme, which reserves bandwidth only during the mobile's expected cell residence time interval, shows higher overall bandwidth utilization than the non-timed scheme. Furthermore, reserving bandwidth only for those cells that are most likely to be visited reduces the blocking ratio of the MLC timed-guarantees scheme in comparison to the non-timed scheme. These results show that the proposed approach achieves a balance between guaranteeing an uninterruptable service for admitted calls and maximizing the utilization of the network resources

VI. Conclusion

This paper presented a framework for predictive timed-QoS guarantees based on a mobility model that estimates the cluster of cells that are most likely to be visited and the time interval during which these cells are visited. A distributed call admission control which verifies the feasibility of admitting new and hand-off calls was also described. Using simulation, the performance of the schemes used to support predictive timed-QoS guarantees was studied. It

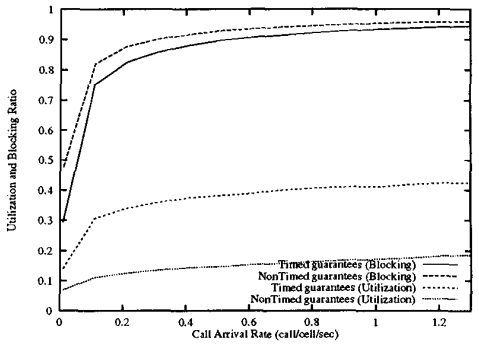

Fig. 7. Bandwidth utilization and Blocking ratio vs. Call arrival rate for timed and non-timed guarantees.

was shown that the schemes achieve a balance between guaranteeing uninterruptable service with a high probability and maximizing resource utilization.

References

[1] A. Acampora and M. Naghshineh. Control and quality-of-service provisioning in high speed microcellular networks. *IEEE Personal Communications*, 1(2):36–42, 1994.

[2] A. Aljadhai and T. Znati. A predictive bandwidth allocation scheme for multimedia wireless networks. In *Conference on Communication Networks and Distributed Systems Modeling and Simulation (CNDS '97)*, pages 95–100, Phoenix, Arizona, January 1997.

[3] V. Bharghavan and V. Gupta. A framework for application adaptation in mobile computing environments. In *Computer Software and Application Conference*, 1997.

[4] R. G. Brown. *Smoothing Forecasting and Prediction of Discrete Time Series.* Prentice hall, NJ, 1963.

[5] Sturat E. Dreyfus. An appraisal of some shortest-path algorithms. *Operations Research*, 17:395–412, 1969.

[6] D. Hong and S. Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Trans. on Veh. Tech.*, 35(3):77–92, August 1986.

[7] T. Imielinski and B. Badrinath. Mobile Wireless Computing: Challenges in Data Management. *Communication of ACM*, 37(10):18–28, October 1994.

[8] D. Levine, I. Akyildiz, and M. Naghshineh. Resource estimation and call admission algorithm for wireless multimedia using the shadow cluster concept. *IEEE/ACM Transactions on Networking*, 5(1):1–12, Feb. 1997.

[9] S. Lu, K. Lee, and V. Bharghavan. Adaptive service in mobile computing environments. In *Int. Workshop on QoS*, 1997.

[10] C. R. Perkins. Quality-of-Service Guarantees in Mobile Computing. Technical report, Internet-Draft, May 1996.

[11] J. H. Reed T. S. Rappaport and B. D. Woerner. Position location using wireless communications on highways of the future. *IEEE Communications Magazine*, 34(10):33–41, Oct. 1996.

[12] K. Yeung and S. Nanda. Channel management in microcell/macrocell cellular radio systems. *IEEE Transactions on Vehicular Technology*, 45(4):601–612, Nov. 1996.

[13] O. Yu and V. Leung. Adaptive resource allocation for prioritized call admission over an atm-based wireless pcs. *IEEE Journal on Selected Areas in Communications*, 15(7):1208–1225, Spt. 1997.