State University of New York at Stony Brook
College of Engineering and Applied Sciences

Technical Report No. 768

Integrated Multimedia Personal Communications with Asymmetric Services for Mobile Users in Cellular Systems

by

Guodong Zhang and Stephen S. Rappaport

Department of Electrical and Computer Engineering
State University of New York
Stony Brook, New York 11794-2350

e-mail: gzhang@sbee.sunysb.edu, rappaport@sunysb.edu

Date: January 4, 1999

Integrated Multimedia Personal Communications with Asymmetric Services for Mobile Users in Cellular Systems

Guodong Zhang and Stephen S. Rappaport
Department of Electrical and Computer Engineering
State University of New York
Stony Brook, NY 11794-2350
email: gzhang@sbee.sunsyb.edu, rappaport@sunysb.edu

Abstract: Wireless personal communication networks are evolving to provide integrated multimedia services to mobile users. An important issue in their design is the accommodation of various services each having predefined Quality-of-Service (QoS) requirements. Efficient resource utilization can be achieved using a medium access scheme that allows various traffic components to be statistically multiplexed and call admission control to prevent overload. Some multimedia services have asymmetric resource needs. For example, web browsing generally requires more downlink bandwidth than uplink bandwidth. We present a medium access scheme and a call admission control algorithm for guaranteed QoS provisioning in asymmetric integrated multimedia personal communication networks. An overall performance analysis model is developed. Performance characteristics are calculated and presented for an example system.

I. INTRODUCTION

Advances in digital signal processing/compression technologies, low-power VLSI, and network infrastructure have enabled rapid development of wireless integrated multimedia services. In wireline networks, ADSL (Asymmetric Digital Subscriber Line) technology provides asymmetric uplink and downlink bandwidths to satisfy users' needs. It can be predicted that wireless networks will provide similar services to mobile users in the near future. Thus, the issue of meeting each service class' requirements for Quality-of-Service (QoS) emerges as a

---

The research reported in this paper was supported in part by the U.S. National Science Foundation under Grant No. NCR 94-15530. General research support from Hughes Network Systems is gratefully acknowledged.

challenging problem. Recently, there has been much research effort on the QoS of personal communications systems. Medium access schemes are discussed in [1], [2] and call admission control aspects are treated in [3]. However, little or nothing has appeared which considers both aspects of the problem for the QoS provisioning of wireless systems.

This paper suggests a call admission control algorithm and a medium access scheme. In this approach, an arriving new (hand-off) call first has to be admitted (accommodated) to a spatial cell to get service. The call's requirements of resources and QoS will be declared to the system upon arrival. The system will accommodate a call only if adequate resources are available and the QoS requirements of all admitted sessions could be met if this newly arriving session were to be admitted. The call admission algorithm uses call admission regions in a suitably defined state space to determine whether or not to accommodate a session in a spatial cell. In general, new calls and hand-off calls have different admission criteria so that priority access can be accommodated. After admission, a call will use some resources dynamically by competing with other admitted calls according to the medium access scheme that the system uses. Some specific resources may be dedicated to each admitted session. The medium access scheme proposed here is a modified version of R&R-DSA (Request and Report Dynamic Slot Allocation) that was put forth in our previous work [4]. The scheme, which was inspired by [5], employs a time division format and efficiently utilizes the channel resources by taking the activity of each traffic class into consideration.

Extending the framework for admission control [6], [7], [8], [9], that has been developed in recent years, we develop a model to analyze the overall performance of the integrated system. The earlier framework is expanded in several aspects. First, asymmetric services are integrated into the system, so both the uplink and downlink resources have to be considered. Secondly, the medium access scheme is taken into consideration in the framework. Thirdly, the possible system states are not only subject to the constraint of resource limits and quotas, but also to the constraintsof QoS requirements.

The system model is described is defined in section II of the paper, where system state is defined. A call admission algorithm is put forth in section III. In section IV, an example system is considered and traffic models are described. In section V, a medium access scheme, which considers the activity characteristics of different traffic components, is proposed and its performance is evaluated by simulation. In section VI, the analytical framework is used to calculate admission control performance. In addition, the QoS performance of the system is

analyzed. For the example system, numerical results are obtained and discussed in section VII. Conclusions are summarized in section VIII.

II. SYSTEM MODEL AND STATE DESCRIPTION

A. System Model

For convenience, we consider a homogeneous personal communications system covered by cells. Here, the word cell is used in its generic sense to describe a spatial zone serviced by a base station. It can be a sector, microcell, or macrocell, or satellite beam.

We assume that in the system there are G types of mobile platforms, labeled by g=1, 2, ..., G. No more than one call can be supported by a platform at any given time. Platform types differ primarily in their mobility characteristics. In each cell, there are K types of uplink resource and L types of downlink resource used to support calls. There are $R_U(k)$, k=0, 1, ..., K-1, units of uplink resource of type k and $R_D(l)$, l=0, 1, ..., L-1, units of downlink resources of type l at each cell.

The system will accommodate a variety of services. We assume that there are I types of services, labeled by i=1, 2, ..., I. The new i-type call origination rate from a noncommunicating g-type platform is $\Lambda(g,i)$. The number of noncommunicating platforms in any cell is denoted $\nu(g,0)$. This number is assumed to be large compared with the maximum number of calls that can be simultaneously supported in a cell. Thus, the origination rate of i-type calls from g-type platforms in any cell is $\Lambda_n(g,i) = \Lambda(g,i) \times \nu(g,0)$. The problem is how to accommodate these three types of traffic while assuring that their respective QoS requirements are met during session "lifetimes." We present a solution that uses call admission control and a medium access scheme for admitted sessions.

In order to give priority to hand-off calls, the system has stricter admission criteria for new calls than for hand-off calls. A cut-off priority scheme is used as specified in the discussion of call admission control in Section III. For any call of type i, we assume that the unencumbered call (session) duration is a ned random variable, $T(i)$, having a mean $\overline{T}(i)=1/\mu(i)$. The dwell time in a cell for a g-type platform is a ned random variable, $T_D(g)$, with a mean $\overline{T}_D(g)=1/\mu_D(g)$. Generalizations can be considered which nevertheless maintain the memoryless property needed for analytical tractability [7], [10], [11].

B. State Description

Consider a single cell. We define the state (of a cell) by a sequence of nonnegative integers. This can be conveniently written as G n-tuples [7], [8]

$$\begin{array}{l} v_{11}, v_{12}, \ldots, v_{1i}, \ldots, v_{1I} \\ v_{21}, v_{22}, \ldots, v_{2i}, \ldots, v_{2I} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ v_{g1}, v_{g2}, \ldots, v_{gi}, \ldots, v_{gI} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ v_{G1}, v_{G2}, \ldots, v_{Gi}, \ldots, v_{GI} \end{array} \qquad (1)$$

where v_{gi} $\{g=1, 2, \ldots, G; i=1, 2, \ldots, I\}$ is the number of platforms of type g that have a call of type i in progress. It is convenient to order the states using an index $s=0, 1, 2, \ldots, s_{\max}$. Then the state variables v_{gi} can be shown explicitly dependent on the state. That is, $v_{gi}=v(s,g,i)$. Let $\Phi(s)$ denote the G n-tuples in (1) when the cell is in state s.

Since guaranteed QoS provisioning is considered, permissible system states are subject to constraints of QoS requirements as well as resource requirements that were considered previously [6-9], [10], [11]. The medium access scheme is also taken into consideration in this model. Admitted sessions use resources dynamically in a statistically multiplexed manner. Asymmetric services are considered in the model, so both downlink and uplink QoS and resources have to be considered. These features are reflected in the model and the constraints that define permissible states.

There are two kinds of QoS metrics. One reflects the fidelity and timeliness of transmission for sessions that have been **admitted to a cell**. Packet loss probability and average packet delay metrics are of this kind. They are indicative of the competition for media access among sessions that have already been admitted. We call these **transmission** QoS metrics. For a state s, the number of admitted sessions of type i, ($i=1, 2, \ldots I$) calls is known, so *transmission* QoS metrics of an i-type call when the system is in state s is determined. The other kind, **connection** QoS metrics, reflects connectivity for a session. The probability that an admitted session is interrupted during its lifetime because of hand-off failure (forced termination probability) is an example. Blocking probability and hand-off failure probability are also of this

kind. **Connection** QoS metrics are quantities that are averaged over the state space and have no meaning at a particular state.

Generally, QoS metrics may depend on both call type and platform type. Let $\overrightarrow{QoS(s,g,i)}$ be a vector whose components denote QoS performance of i-type calls on g-type platforms in state s. As we discussed, $\overrightarrow{QoS(s,g,i)}$ does not include connection QoS metrics, it only includes transmission QoS. $\overrightarrow{QoS(s,g,i)}$ is determined by the state of the cell and the medium access scheme jointly. We define $[QoS(s)]$ as the array of transmission QoS performance metrics in state s. This is given by

$$[QoS(s)]=\begin{bmatrix} \overrightarrow{QoS(s,1,1)} & \overrightarrow{QoS(s,1,2)} & \ldots, & \overrightarrow{QoS(s,1,i)} & \ldots, & \overrightarrow{QoS(s,1,I)} \\ \overrightarrow{QoS(s,2,1)} & \overrightarrow{QoS(s,2,2)} & \ldots, & \overrightarrow{QoS(s,2,i)} & \ldots, & \overrightarrow{QoS(s,2,I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overrightarrow{QoS(s,g,1)} & \overrightarrow{QoS(s,g,2)} & \ldots, & \overrightarrow{QoS(s,g,i)} & \ldots, & \overrightarrow{QoS(s,g,I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overrightarrow{QoS(s,G,1)} & \overrightarrow{QoS(s,G,2)} & \ldots, & \overrightarrow{QoS(s,G,i)} & \ldots, & \overrightarrow{QoS(s,G,I)} \end{bmatrix} \quad (2)$$

We define $\overrightarrow{QoS_0(g,i)}$ as the required transmission QoS for i-type calls on g-type platforms to provide satisfactory services. Let $[QoS_0]$ denote the array of transmission QoS requirements. This is given by

$$[QoS_0]=\begin{bmatrix} \overrightarrow{QoS_0(1,1)} & \overrightarrow{QoS_0(1,2)} & \ldots, & \overrightarrow{QoS_0(1,i)} & \ldots, & \overrightarrow{QoS_0(1,I)} \\ \overrightarrow{QoS_0(2,1)} & \overrightarrow{QoS_0(2,2)} & \ldots, & \overrightarrow{QoS_0(2,i)} & \ldots, & \overrightarrow{QoS_0(2,I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overrightarrow{QoS_0(g,1)} & \overrightarrow{QoS_0(g,2)} & \ldots, & \overrightarrow{QoS_0(g,i)} & \ldots, & \overrightarrow{QoS_0(g,I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overrightarrow{QoS_0(G,1)} & \overrightarrow{QoS_0(G,2)} & \ldots, & \overrightarrow{QoS_0(G,i)} & \ldots, & \overrightarrow{QoS_0(G,I)} \end{bmatrix} \quad (3)$$

We use the symbol "$\prec\cdot$" to mean "satisfy". When used between commensurate arrays the comparison is between corresponding elements. Remember that the QoS performance in state s is subject to the constraints of QoS requirements. Thus, admission control must assure that for any state, s,

$$[QoS(s)] \prec\cdot [QoS_0]. \quad (4)$$

The traffic generated by each type of service varies from frame to frame, the resources used by each session change dynamically. Since calls of the same type but on different types of platforms have different hand-off rates, the resource requirements of some resources (such as call

supervising processors) that are needed to provide service may depend on platform type. Let $r_U(s,g,i,k)$ denote the amount of uplink resource k required by i-type calls on a g-type platforms to maintain predefined transmission QoS requirements when the cell is in state s. Similarly, let $r_D(s,g,i,l)$ denote the amount of downlink resource l required by i-type calls on g-type platforms to maintain predefined transmission QoS requirements when the cell is in state s. We assume that $r_U(s,g,i,k)$ and $r_D(s,g,i,l)$ are proportional to $v(s,g,i)$. Let $r_U(g,i,k)$ denote the amount of uplink resource k required by a single admitted i-type call on a g-type platform to maintain its transmission QoS requirement and $r_D(g,i,l)$ denote the amount of downlink resource l required by a single admitted i-type call on a g-type platform to maintain its transmission QoS requirement. Then

$$r_U(s,g,i,k) = r_U(g,i,k) \cdot v(s,g,i), \quad k=0, 1, 2, \ldots, K\text{-}1$$

$$r_D(s,g,i,l) = r_D(g,i,l) \cdot v(s,g,i), \quad l=0, 1, 2, \ldots, L\text{-}1 \qquad (5)$$

The amounts of uplink resource k and downlink resource l required by i-type calls in state s to maintain their transmission QoS requirements are given respectively by

$$r_U(s,*,i,k) = \sum_{g=1}^{G} r_U(s,g,i,k), \quad k=0, 1, 2, \ldots, K\text{-}1$$

$$r_D(s,*,i,l) = \sum_{g=1}^{G} r_D(s,g,i,l), \quad l=0, 1, 2, \ldots, L\text{-}1 \qquad (6)$$

The amounts of uplink resource k and downlink resource l required by g-type platforms in state s to maintain their transmission QoS requirements are given respectively by

$$r_U(s,g,*,k) = \sum_{i=1}^{I} r_U(s,g,i,k), \quad k=0, 1, 2, \ldots, K\text{-}1$$

$$r_D(s,g,*,l) = \sum_{i=1}^{I} r_D(s,g,i,l), \quad l=0, 1, 2, \ldots, L\text{-}1 \qquad (7)$$

The total amounts of uplink resource k and downlink resource l required in state s to maintain all admitted sessions' transmission QoS requirements are given respectively by

$$r_U(s,*,*,k) = \sum_{i=1}^{I} \sum_{g=1}^{G} r_U(s,g,i,k), \quad k=0, 1, 2, \ldots, K\text{-}1$$

$$r_D(s,*,*,l) = \sum_{i=1}^{I} \sum_{g=1}^{G} r_D(s,g,i,l), \quad l=0, 1, 2, \ldots, L\text{-}1 \qquad (8)$$

Recall, that there are $R_U(k)$ units of uplink resource k and $R_D(l)$ units of downlink resource l at each cell. The total amounts of required resources in any state are subject to resource limits. Thus,

$$r_U(s,*,*,k) \le R_U(k), \quad k=0, 1, \ldots, K\text{-}1$$
$$r_D(s,*,*,l) \le R_D(l), \quad l=0, 1, \ldots, L\text{-}1 \qquad (9)$$

In addition, resource quotas could be applied to the system which limits the amount of uplink resource k and downlink resource l that are required by g-type platforms to maintain transmission QoS requirements and requires

$$r_U(s,g,*,k) \le R_U{}^P(k,g), \quad k=0, 1, 2, \ldots, K\text{-}1$$
$$r_D(s,g,*,l) \le R_D{}^P(l,g), \quad l=0, 1, 2, \ldots, L\text{-}1 \qquad (10)$$

or limits the amount of uplink resource k and downlink resource l that are required by i-type calls to maintain transmission QoS requirements and requires

$$r_U(s,*,i,k) \le R_U{}^C(k,i), \quad k=0, 1, 2, \ldots, K\text{-}1$$
$$r_D(s,*,i,l) \le R_D{}^C(l,i), \quad l=0, 1, 2, \ldots, L\text{-}1 \qquad (11)$$

Additional constraints that enforce cut-off priorities as mentioned above can be applied to the system.

Permissible states correspond to those sequences in the form of (1), for which all constraints, such as QoS requirements, resource limits and quotas, are met. The set of permissible states Φ and state transition probabilities are affected by the medium access scheme as well as call admission control.

III. CALL ADMISSION CONTROL

The purpose of call admission control (CAC) is to assure that transmission QoS is guaranteed for all sessions that are admitted for service. The call admission control is based on whether the network can provide the required QoS for all sessions that have (already) been admitted to the cell, as well as for the arriving session. Recall that there are I types of traffic in the system, indexed by i=1, 2, …, I. The system has different QoS criteria of call admission for *new* calls and *hand-off* calls of the same call-platform types (i.e., same g,i). Since $\overrightarrow{QoS_0(g,i)}$ is the required transmission QoS for i-type calls on g-type platforms to provide satisfactory services. Then, the basic transmission QoS requirements, $\overrightarrow{QoS_0(g,i)}$, is used as the call admission criteria

for i-type *hand-off* calls on g-type platforms. From the mobile user's point of view, the forced termination of a call in progress is more annoying and less desirable than blocking of a new call attempt. Therefore, in the call admission control scheme, stricter call admission criteria are used for new calls to favor hand-off calls. Let $\overrightarrow{priority(g,i)}$ be a vector whose components denote the differences between call admission criteria for i-type new calls on g-type platforms and call admission criteria for i-type hand-off calls on g-type platforms, that is, the priority given to hand-off calls over new calls in admission. Thus, $\overrightarrow{QoS_0(g,i)} + \overrightarrow{priority(g,i)}$ is used as the call admission criteria for i-type new calls on g-type platforms. We define the array $[priority]$ as the priority given to hand-off calls over new calls in admission.

$$[priority]=\begin{bmatrix} \overrightarrow{priority(1,1)} & \overrightarrow{priority(1,2)} & \cdots & \overrightarrow{priority(1,i)} & \cdots & \overrightarrow{priority(1,I)} \\ \overrightarrow{priority(2,1)} & \overrightarrow{priority(2,2)} & \cdots & \overrightarrow{priority(2,i)} & \cdots & \overrightarrow{priority(2,I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overrightarrow{priority(g,1)} & \overrightarrow{priority(g,2)} & \cdots & \overrightarrow{priority(g,i)} & \cdots & \overrightarrow{priority(g,I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overrightarrow{priority(G,1)} & \overrightarrow{priority(G,2)} & \cdots & \overrightarrow{priority(G,i)} & \cdots & \overrightarrow{priority(G,I)} \end{bmatrix} \quad (12)$$

The system may accommodate some services, such as video, which require very high use of resources compared with speech service. Because of this the blocking and hand-off failure probabilities of call types that use a lot of resources may be inordinately high compared to the other services. It may desirable for the system to use some *quotas* on the admission of *low use sessions* in order to assure that service requirements of *high use sessions* are met. If quotas are used in this way, some low use sessions may be denied admission even if there are: 1) resources available to serve them at their time of arrival; and, 2) serving them would not otherwise interfere with the guaranteed QoS of sessions *that are already admitted.* To enforce such a quota, we assume that there are, among the I types of services in the system, N_l types of *low use* services, indexed by $i=i_1, i_2 \ldots i_{N_l}$. Then, if $R_U{}^C(k)$ and $R_D{}^C(l)$ denote the respective uplink and downlink resource quota limits on low use services, the quota constraints can be expressed as

$$\sum_{i\in(i_1,i_2,\ldots,i_{N_l})} r_U(s,*,i,k) \le R_U{}^C(k), \quad k=0, 1, 2, \ldots, K\text{-}1$$

$$\sum_{i\in(i_1,i_2,\ldots,i_{N_l})} r_D(s,*,i,l) \le R_D{}^C(l), \quad l=0, 1, 2, \ldots, L\text{-}1 \quad (13)$$

The call admission control procedure is described as below with the use of Fig. 1:

1) Upon the arrival of a new call or hand-off call, the base station determines whether or not the QoS requirements of all admitted sessions and the resources quota could be met if the new session were to be accommodated. For this purpose, we calculate call admission regions for the system. For *i*-type calls on *g*-type platforms, there are two call admission regions: a new call admission regions and a hand-off call admission regions. Every point in the hand-off call admission region corresponds to a possible state of the system in which admission criteria $[QoS_0]$ and the quota in (13) can be met for all admitted sessions if an *i*-type hand-off call on *g*-type platform is accommodated. Similarly, every point in the new call admission region corresponds to a possible state of the system in which admission criteria $[QoS_0]+[priority]$ and the quota in (13) can be met for all admitted sessions if an *i*-type new call on *g*-type platform is admitted. Let $REG_n(g,i)$ and $REG_h(g,i)$ denote the new call admission region and the hand-off call admission region for *i*-type calls on *g*-type platforms respectively. The admission regions for each type of call on each type of platform can be calculated and stored in the base station. When a new *i*-type call on a *g*-type platform arrives, the base station can decide whether or not to admit the call by checking whether the current state is in $REG_n(g,i)$. Similarly, the admission decision for a *(g,i)*-handoff call is based on whether the current state is in the $REG_h(g,i)$.

2) A new call admission or a hand-off admission of a specific class is assigned to the carrier which has the least traffic load of this specific class of traffic. Ties are broken by random selection. By doing this, the traffic load can be more or less uniformly distributed among all the carriers and a user of any class will enjoy the same Quality-of-Service as any other user of the same class.

IV. EXAMPLE SYSTEM AND TRAFFIC MODELS

A. Example System

As an example, we consider a personal communication network with a TDMA FDD structure and a spatially cellular layout based on a fixed channel assignment (FCA). We assume that the communication system is homogeneous. Specifically, in each cell there are C_{up} uplink frequency carriers and C_{down} downlink frequency carriers. Each carrier is divided into TDMA frames with duration of F sec per frame. Each frame is subdivided into short mini-slots and S payload slots. Each payload slot can transmit an ATM type cell with data payload of 48 bytes.

For convenient illustration, we consider three typical types of service: speech, video, and asymmetric traffic. In order to achieve guaranteed QoS provisioning, the activity characteristics of each admitted traffic type should be thoroughly examined and taken into consideration. Admitted speech, video and asymmetric traffic components are characterized in the following paragraphs.

B. Traffic Models:

In the following description of traffic models, the word "cell" means ATM type cell with a payload of 48 bytes, not a region served by a base station. In later sections, the meaning of "cell" (ATM or spatial) can be distinguished from the context.

1. Speech Traffic:

The activity of a speech session follows a pattern of alternating talkspurts and silence gaps. Speech cells are generated during talkspurt periods at the rate of one cell per frame. Each cell needs a payload slot to be transmitted. During silence gaps, a speech session has nothing to transmit. It is assumed that the duration of a talkspurt is a negative exponentially distributed (ned) random variable with a mean of $\overline{T_t}$ and the duration of a silence gap is ned random variable with a mean of $\overline{T_s}$ [1]. These mean durations are long compared with the duration of a frame. A speech session needs equal amounts of downlink and uplink resources. Since speech traffic is delay-sensitive, there is a maximum acceptable transmission delay associated with speech traffic. Let M_s denote the maximum acceptable transmission delay of a speech cell. Any speech cell that exceeds the maximum transmission delay will be discarded. Speech traffic can tolerate moderate packet loss. To provide acceptable QoS, a packet loss probability that is less than 1% is required [12].

2. Video Traffic:

The video traffic is of variable bit-rate (VBR) type. A video session generates one or more video cells every frame. Each video cell needs one payload slot to be transmitted. The number of ATM cells generated by a video session generally varies from frame to frame. Usually, applications such as low-bit-rate video conferencing can be modeled using a N_v-state discrete-time Markov chain [2], [12]. The state is the number of ATM cells generated by a video session in one frame. The state transition probabilities of the Markov chain can be obtained from

the coding of a video sequence with H.261 [13] or H.263 [14] coding standards. The uplink resources that a video session needs is dependent on its cell generation rate, while the downlink resources that a video session needs is dependent on the other party's cell generation rate. The other party's video session also follows the same traffic model. So a video session needs equal amounts of uplink and downlink resources. Like speech traffic, video traffic is also delay-sensitive traffic. Let M_v denote the maximum acceptable delay of a video cell. Subjective tests indicate that eyes are not as delay-sensitive as ears. Therefore, generally M_v is larger than M_s. On the other hand, video traffic is more sensitive than speech traffic to packet loss. If any video ATM cell exceeds its maximum transmission delay, it will be discarded. To provide acceptable QoS, a packet loss probability that is less than 0.1% is required.

3. Asymmetric Traffic:

There are many applications, such as video on demand, web browsing and file download, that need more downlink resources than uplink resources. These asymmetric traffic components cannot be summarized in single traffic model. Here, we model one type of asymmetric traffic: web-type traffic, as an example. Usually (both wireline and wireless) web-type sessions need more downlink (server to client) resources than uplink resources. In wireline networks, ADSL (Asymmetric Digital Subscriber Line) technology can provide users with speeds of up to 8Mbps (downlink) and 800Kbps (uplink). If wireless personal communication networks provide similar services to mobile terminals, for web-type sessions, the required downlink resources may be 5~10 times that of uplink resources.

To initiate a web-type session, a call setup for Internet access (admission) to the base station is required. A web-type session will be admitted only if the system can meet its requirements of resources and QoS. Different from today's technology, the session will be admitted directly to the service. Otherwise, this session will be blocked and we assume that no resources will be used by it. An admitted web-type session is defined as the collection of events that occur since a mobile user gets connected to the Internet until the user is disconnected from the Internet. During this period, delay-insensitive traffic may be generated when the user is browsing the web, downloading files, etc., while delay-sensitive traffic may be generated when the user is playing some music and video, etc. Applications that need equal amounts of uplink and downlink resources, such as Internet interactive speech and video conferencing, are not included in the web-type traffic component modeled in this example.

We model a web-type session's ***downlink*** traffic as a discrete-time Markov chain with three states: *Delay-Sensitive Busy* , *Delay-Insensitive Busy* and *Idle* as shown in Fig. 2. Thus, a web-type session's downlink traffic consists of alternating (delay-sensitive or insensitive) *Busy* and *Idle* periods. Downlink web-type cells are generated only during the busy periods at a fixed rate of CR_{down} (>1) cells per frame. We use index *B1* for Delay-insensitive Busy and *B2* for Delay-sensitive Busy. We assume that the durations of *Delay-Insensitive Busy*, *Delay-Sensitive Busy* and *Idle* period are ned random variables with means of $\overline{T_{B1}}$, $\overline{T_{B2}}$ and $\overline{T_I}$ respectively. Usually, $\overline{T_{B1}}$, $\overline{T_{B2}}$ and $\overline{T_I}$ are much larger than *F*, the duration of a frame. As shown in Fig. 2, the transition probability from *Delay-Insensitive Busy* state to *Idle* state is denoted by $P_{B1,I}$. This is the probability that a delay-insensitive busy period will end in a frame. Similarly, $P_{B2,I}$ denotes the transition probability from *Delay-Sensitive Busy* state to *Idle* state, and $P_{I,B}$ denotes the transition probability from *Idle* state to (either delay-insensitive or delay-sensitive) *Busy* state. These transition probabilities are determined by values of $\overline{T_{B1}}$, $\overline{T_{B2}}$, $\overline{T_I}$ and *F*. Let P_{ins} denote the probability that the next period is a delay-insensitive busy period given that the idle period ends in this frame. We define a cycle as a busy period and the following idle period.

We model the web-type session's ***uplink*** traffic as a discrete-time Markov chain with three states: *Delay-Sensitive Busy* , *Delay-Insensitive Busy* and *Idle* in a similar way. The differences in comparison with ***downlink*** traffic are: 1) web-type cells are generated only during the busy periods at the rate of CR_{up} ($<CR_{down}$) cells per frame; 2) the mean duration of *Delay-Insensitive Busy* period $\overline{T_{B1}}'$ is shorter than downlink's $\overline{T_{B1}}$, the mean duration of *Delay-Sensitive Busy* period $\overline{T_{B2}}'$ is shorter than downlink's $\overline{T_{B2}}$, and the mean duration of *Idle* period $\overline{T_I}'$ is longer than $\overline{T_I}$; and, 3) the probability that the next period is a delay-insensitive busy period given that the idle period ends in this frame, P_{ins}' , is larger than P_{ins}. These differences arise because of the characteristics of web-type traffic. For example, in a web-type session, a short inquiry in the uplink may evoke the download of a file or the transfer of a video file.

It is important to note that as shown in Fig. 2, the traffic generated in one cycle is either delay-sensitive traffic or delay-insensitive traffic, but not a mixture of the two. So generally there are delay-sensitive cycles and delay-insensitive cycles in a given session as shown in Fig. 3. There is a maximum acceptable transmission delay M_W associated with web-type delay-sensitive

cells. To provide acceptable QoS, a cell loss probability that is less than 0.1% is required for the web-type delay-sensitive traffic. There is no maximum delay for web-type delay-insensitive cells. So no cell will be dropped because of the excessive delay. To provide acceptable QoS, an average cell delay that is less than half of the mean idle duration, i.e. $\frac{1}{2}\overline{T_I}$ ($\overline{T_I}$), is required for the downlink (uplink) web-type delay-insensitive traffic.

V. Medium Access Scheme and Performance Evaluation

A. Medium Access Scheme

The medium access scheme has a significant influence on the traffic performance and system capacity. It must provide the ability to accommodate various traffic components at their predefined QoS levels, while allowing a reasonably efficient utilization. Here, the only resource that is statistically multiplexed by different admitted sessions is payload slots. At the head of the TDMA frame, there are a fixed number of mini-slots. Usually, a request (or report) packet is 1~2 bytes and an ATM type payload is 48 bytes. So the duration of a payload slot is much larger than the duration of a mini-slot. The size of a mini-slot in the system depends on the traffic parameters, such as maximum delay and number of states in video Markov chain model. The details of request and report packets will be described in the operation of the medium access scheme. In this scheme, we assume that there are enough mini-slots in each frame so that every admitted speech, video, and web-type session is assigned a mini-slot to transmit their request (or report) packets during the session. Thus, there are no collisions of either request or report packets. A diagram of the medium access scheme is shown in Fig. 4. The operation of the medium access scheme is described as follows:

1) A speech session that enters the talkspurt state will transmit a request packet to the base station to reserve a payload slot to use for the talkspurt period. A counter at the speech user records the number of times the speech session has failed its current request for reservation of its talkspurt. The speech user's ID and the count are included in the speech request packet. When there are not enough payload slots to accommodate all the speech requests, priority will be given to requests with larger counter values. An unsuccessful speech talkspurt request is ***retransmitted in the next frame*** as before until it either succeeds or exceeds the maximum delay constraint. Recall that a speech session generates cells at a constant rate during a talkspurt state. So if a speech talkspurt request exceeds the maximum delay, all subsequent speech packets that were

generated by the user and buffered in the user's local buffer during the request period will be discarded and the counter will be reset to zero. The next *new speech packet* will act as the beginning of the talkspurt period and will initiate a new request. A speech session that has made a successful reservation is allocated one payload slot for its whole talkspurt. When a speech session competes with a video session for a payload slot allocation, the video session is given priority over the speech session. Video traffic, however, can not use the payload slots that are already reserved by speech sessions. So the presence of admitted speech sessions affects the performance of video sessions

2) To deal with the VBR nature of video traffic, a report mechanism used. The video cells generated in the *first* frame are not transmitted immediately. The video user will send a report packet (to the base station) including the video user's ID and the number of video cells that are ready for transmission. Upon receiving all the report packets in the current frame, the base station knows the total number of video cells (from all video sessions) that are ready for the next frame. If there are not enough payload slots to accommodate all of these cells, some cells are randomly selected by the base station to utilize all the available payload slots. In the next frame, the cells that were selected by the system will be transmitted. Each of the other cells, if any, will be stored in the respective user's buffer. Also, each video session will report again the total number of ready cells (including newly generated cells in this frame). The video cells that exceed the maximum delay will be discarded by respective users. The same operation is repeated for subsequent cells in video sessions.

3) A web-type session that enters the *busy* (either delay-sensitive or delay-insensitive) period will send a request packet to the base station to reserve payload slots for its *busy* period. Let WDS denote the phrase "web-type delay-sensitive", and WDI denote the phrase "web-type delay-insensitive". The web-type user's ID and (for WDS busy period only) a counter that contains the number of times this web-type user has failed in its current request to obtain the needed resources are included in the request packet. No counter is used for WDI busy period. When there is a conflict between WDS and WDI busy periods, priority is always given to WDS components. The ties between WDS busy periods are broken by higher counter values, while the ties between WDI busy periods are broken randomly. If this is a WDI busy period, then only the payload slots that are not used by speech, video and WDS traffic will be assigned. The reservation made by a WDI busy period may be preempted by video, speech and WDS traffic. When such a reservation is preempted, the WDI user will send a request packet to the base station every frame until its

reservation is granted. A WDS busy period will have the same priority as speech traffic. If it competes with a video session, the video session is given priority. Video traffic, however, can not use the payload slots that are already reserved by WDS busy periods. If it competes with a speech session, choosing one of them randomly breaks the tie. An unsuccessful WDI request will be retransmitted as before until it succeeds, while an unsuccessful WDS request will be retransmitted as before until it either succeeds or exceeds the maximum delay. If the maximum delay is exceeded, subsequent packets which were generated by the session and stored in the user's buffer during the request period will be discarded by respective users and the counter will be reset to zero. The next *new cell(s)* will act as the beginning of the WDS busy period and will initiate a new request. For the downlink, when there are not enough payload slots, a session may get a partial reservation. That is, the amount of reserved resources is less than the amount of needed resources. In this case, the session will send request packets until it gets the full reservation.

4) If there are still some available payload slots after allocations are made for video, speech traffic and busy web-type periods, the base station will randomly select some WDI cells that are buffered in users' buffers to utilize these available slots. Thus, sometimes a WDI period may use more slots than it reserved.

5) The timing of the medium access scheme is shown in Fig. 5, in which τ is the one way propagation delay. After receiving all the request and report packets, the base station processes them and sends acknowledgments to mobile users. The time that is needed to transmit an acknowledgment from the base station to mobile users is denoted by T_{ACK}. The processing time at the base station is denoted by T_{p1} and similarly, T_{p2} denotes the processing time of an acknowledgment at a mobile user. We assume that in one frame there are N mini-slots, each with duration of σ. The system uses a TDMA FDD structure, so a mobile user can transmit and receive at the same time. To assure that all these requests/reports are processed and acknowledged in time, we must have

$$N\sigma + 2\tau + T_{ACK} + T_{p1} + T_{p2} \le F \qquad (14)$$

B. Performance Evaluation of the Medium Access Scheme:

As described in the medium access scheme, video traffic has priority when competing with speech and web-type traffic. The number of payload slots that can be used by video traffic, however, is constrained by the number of payload slots that are *already reserved* by speech and

WDS traffic. In turn the number of payload slots that will be reserved by speech and WDS traffic is also affected by the number of video sessions that. The traffic model of a video session is a Markov chain with many states. The video cells that do not get payload slots can be buffered until the maximum delay is exceeded. A downlink web-type session may get a partial reservation. These make the analysis more complex. An analytically tractable performance model for the medium access scheme is difficult to devise because of the complexity of the access scheme and the traffic characteristics of speech, video and asymmetric web-type services. (An analytically tractable traffic performance model was devised and used for the admission control aspects of the problem. This will be described subsequently).

We used SIMSCRIPT II.5 to build a discrete-time Monte-Carlo simulation model to evaluate the performance of the medium access scheme.

1. System and Traffic Model Parameters: The choices of system and traffic parameters are summarized in Table 1. For speech traffic, we used typical values of mean talkspurt and silence gaps that have appeared in the literature [1] for $\overline{T_t}$ and $\overline{T_s}$. For video traffic, we used a 10-state Markov chain model. That is the number of generated cells per frame varies from 1 to 10. While this is fewer than that in some models in the literature [2], [12], our simplified simulation model is sufficient to reveal how the medium access scheme deals with the VBR characteristics of video traffic.

Table 1: Parameters of example system and traffic model

Parameter	Symbol	Value
Frame length (msec)	F	10
Number of payload slots in one frame	S	15
Number of uplink and downlink frequency carriers	C_{up}, C_{down}	2, 3
Mean duration of a speech talkspurt (sec)	$\overline{T_t}$	1.0
Mean duration of a speech silence gap (sec)	$\overline{T_s}$	1.35
Maximum acceptable delay for speech cells (msec)	M_s	40
Number of states in the Markov chain of video traffic	N_v	10
Maximum acceptable delay for video cells (msec)	M_v	60
Cell generation of downlink web-type traffic (cell/frame)	CR_{down}	5
Cell generation of uplink web-type traffic (cell/frame)	CR_{up}	1
Mean busy durations of downlink WDI and WDS (sec)	$\overline{T_{B1}}$, $\overline{T_{B2}}$	2, 2
Mean busy durations of uplink WDI and WDS (sec)	$\overline{T_{B1}'}$, $\overline{T_{B2}'}$	0.5, 0.5
Mean idle duration of downlink web-type traffic (sec)	$\overline{T_I}$	5
Mean idle duration of uplink web-type traffic (sec)	$\overline{T_I'}$	6
Probability that an idle period is followed by a WDI period in downlink	P_{ins}	0.75
Probability that an idle period is followed by a WDI period in uplink	P'_{ins}	0.9
Maximum acceptable delay of WDS traffic (msec)	M_w	60

Let $P_{i,j}$ denote the transition probability from state i to state j. The state transition probability matrix P is given by

$$P=\begin{bmatrix} 0.133 & 0.20 & 0.10 & 0.067 & 0.05 & 0.033 & 0.05 & 0.067 & 0.10 & 0.20 \\ 0.20 & 0.133 & 0.20 & 0.10 & 0.067 & 0.05 & 0.033 & 0.05 & 0.067 & 0.10 \\ 0.10 & 0.20 & 0.133 & 0.20 & 0.10 & 0.067 & 0.05 & 0.033 & 0.05 & 0.067 \\ 0.067 & 0.10 & 0.20 & 0.133 & 0.20 & 0.10 & 0.067 & 0.05 & 0.033 & 0.05 \\ 0.05 & 0.067 & 0.10 & 0.20 & 0.133 & 0.20 & 0.10 & 0.067 & 0.05 & 0.033 \\ 0.033 & 0.05 & 0.067 & 0.10 & 0.20 & 0.133 & 0.20 & 0.10 & 0.067 & 0.05 \\ 0.05 & 0.033 & 0.05 & 0.067 & 0.10 & 0.20 & 0.133 & 0.20 & 0.10 & 0.067 \\ 0.067 & 0.05 & 0.033 & 0.05 & 0.067 & 0.10 & 0.20 & 0.133 & 0.20 & 0.10 \\ 0.10 & 0.067 & 0.05 & 0.033 & 0.05 & 0.067 & 0.10 & 0.20 & 0.133 & 0.20 \\ 0.20 & 0.10 & 0.067 & 0.05 & 0.033 & 0.05 & 0.067 & 0.10 & 0.20 & 0.133 \end{bmatrix} \quad (15)$$

All these parameters were chosen for the purpose of determining example numerical results.

2. Simulation Results

Simulations were run under various scenarios with different numbers of *admitted* speech, video and web-type sessions. Uplink and downlink traffic components were simulated separately. We define the performance of a session as the poorer performance of uplink and downlink. For some parameter choices, the performance is constrained by uplink resources – for others it is constrained by downlink resources. The cell loss probabilities of video, speech and web-type delay-sensitive traffic are shown in Fig. 6-8 for various numbers of admitted video, asymmetric web-type sessions and speech sessions respectively. The average packet delay of web-type delay-insensitive traffic in the same scenarios is shown in Fig. 9. From Fig. 6-9, we can see the range of admitted video, web-type and speech sessions that the system can accommodate with the given transmission QoS requirements. The call admission regions, $REG_n(g,i)$ and $REG_h(g,i)$, can be determined by respectively comparing the performance characteristics in Fig. 6-9 with the new call admission criteria and the hand-off call admission criteria.

VI. QUALITY OF SERVICE ANALYIS OF THE SYSTEM

Using the proposed medium access scheme and call admission control a call will only be admitted if the system can guarantee its QoS requirements for the lifetime of its session, while at the same time assuring that the QoS requirements for sessions that are already in progress are met. To develop a performance analysis model for call admission control, we use the basic framework put forth in [6], [7], [8], [9].

A. Driving Processes and State Transition Flow:

The state of a cell changes with time. Let s, x_{ni}, x_{ci}, x_{hi} and x_{di} denote permissible states of a cell in the following discussions. Underlying random processes as follows drive the state transitions:

1. Generation of new calls of i-type

A transition into state s, due to a new i-type call arrival on a g-type platform when the cell is in the state x_{ni}, will cause the state variable $v(x_{ni}, g, i)$ to be increased by 1. A permissible state

x_{ni} is a predecessor state of s for i-type new call arrivals on g-type platforms, if x_{ni} is in the new call admission region $REG_n(g,i)$, and the state variables are related by

$$\begin{aligned} &v(x_{ni},g,i)=v(s,g,i)-1 \\ &v(x_{ni},z_1,z_2)=v(s,z_1,z_2), \quad z_1 \neq g \\ &v(x_{ni},z_1,z_2)=v(s,z_1,z_2), \quad z_2 \neq i. \end{aligned} \tag{16}$$

If $\Lambda_n(g,i)$ denotes the average arrival rate per cell of i-type new calls from g-type platforms, the corresponding transition flow is given by

$$\gamma_n(s,x_{ni})=\Lambda_n(g,i). \tag{17}$$

2. Completion of calls of i-type

A transition into state s, due to an i-type call completion on a g-type platform when the cell is in the state x_{ci}, will cause the state variable $v(x_{ci},g,i)$ to be decreased by 1. Thus a permissible state x_{ci} is a predecessor state of s for i-type call completions on g-type platforms, if the state variables are related by

$$\begin{aligned} &v(x_{ci},g,i)=v(s,g,i)+1 \\ &v(x_{ci},z_1,z_2)=v(s,z_1,z_2), \qquad z_1 \neq g \\ &v(x_{ci},z_1,z_2)=v(s,z_1,z_2), \qquad z_2 \neq i \end{aligned} \tag{18}$$

The corresponding transition flow is given by

$$\gamma_c(s,x_{ci})=\mu(i) \times v(x_{ci},g,i) \tag{19}$$

3. Hand-off arrivals of i-type

Let $\Lambda_h(i)$ be the average rate at which hand-off arrivals of i-type service impinge on the cell, and F_{gi} denote the fraction of arrivals that are g-type platforms. A transition into state s, due to an i-type hand-off arrival on a g-type platform when the cell is in the state x_{hi}, will cause the state variable $v(x_{hi},g,i)$ to be incremented by 1. Thus a permissible state x_{hi} is a predecessor state of s for i-type hand-off arrivals on g-type platforms, if state x_{hi} is in the hand-off call admission region $REG_h(g,i)$, and the state variables are related by

$$\begin{aligned} &v(x_{hi},g,i)=v(s,g,i)-1 \\ &v(x_{hi},z_1,z_2)=v(s,z_1,z_2), \qquad z_1 \neq g \end{aligned}$$

$$v(x_{hi}, z_1, z_2) = v(s, z_1, z_2), \qquad z_2 \neq i \tag{20}$$

The corresponding transition flow is given by

$$\gamma_h(s, x_{hi}) = \Lambda_h(i) \times F_{gi} \tag{21}$$

4. Hand-off departures of *i*-type

A transition into state *s*, due to a hand-off departure of *i*-type on a *g*-type platform when the cell is in the state x_{di}, will cause the state variable $v(x_{di}, g, i)$ to be decreased by 1. Thus a permissible state x_{di} is a predecessor state of *s* for *i*-type hand-off departures on *g*-type platforms, if the state variables are related by

$$\begin{aligned} v(x_{di}, g, i) &= v(s, g, i) + 1 \\ v(x_{di}, z_1, z_2) &= v(s, z_1, z_2), \qquad z_1 \neq g \\ v(x_{di}, z_1, z_2) &= v(s, z_1, z_2), \qquad z_2 \neq i \end{aligned} \tag{22}$$

The corresponding transition flow is given by

$$\gamma_d(s, x_{di}) = \mu_D(g) \times v(x_{di}, g, i) \tag{23}$$

B. Flow balance equations

From the equations given above, the total transition flow into state *s* from any permissible state *x* can be expressed by

$$q(s, x) = \gamma_n(s, x) + \gamma_c(s, x) + \gamma_h(s, x) + \gamma_d(s, x) \tag{24}$$

in which $s \neq x$, and flow into a state *s* has been taken as a positive quantity. The total flow out of state *s* is denoted $q(s, s)$, and is given by

$$q(s, s) = -\sum_{\substack{k=0 \\ k \neq s}}^{s_{max}} q(k, s) \tag{25}$$

To find the statistical equilibrium state probabilities for a cell, we write the flow balance equations for the states. These comprise $s_{max}+1$ simultaneous equations for the unknown state probabilities, $p(s)$. They are of the form

$$\begin{aligned} &\sum_{j=0}^{s_{max}} q(i, j) \cdot p(j) = 0, \quad i = 0, 1, 2, \ldots, s_{max} - 1 \\ &\sum_{j=0}^{s_{max}} p(j) = 1 \end{aligned} \tag{26}$$

in which, for $i \neq j$, $q(i,j)$ represents the net transition flow into state i from state j, and $q(i,i)$ is the total transition flow out of state i. These equations express that in statistical equilibrium, the net probability flow into any state is zero and the sum of the probabilities is unity. The index, i, in (26) can run up to $s_{\max}$ to provide a redundant set that is helpful in numerical computation.

C. Determination of Hand-off Arrival Parameters

In the above analysis, we assumed that the parameters $\Lambda_h(i)$ (the average hand-off arrival rate of i-type calls) and F_{gi} (the fraction of i-type hand-off arrivals that are g-type platforms) are given. Actually, these parameters are implicit functions of the driving processes and must be determined from the dynamics of the processes themselves. The iterative approach described in [6] is used. Let $\Delta_h(g,i)$ denote the average hand-off departure rate of i-type calls on g-type platforms. It is given by

$$\Delta_h(g,i) = \sum_{s=0}^{s_{\max}} \mu(g) \cdot v(s,g,i) \cdot p(s) \tag{27}$$

Then the total departure rate of i-type calls is

$$\Delta_h(i) = \sum_{g=1}^{G} \Delta_h(g,i) \tag{28}$$

The fraction of i-type hand-off departures that are on g-type platforms is given by

$$F_{gi}' = \frac{\Delta_h(g,i)}{\Delta_h(i)} \tag{29}$$

A hand-off departure of an i-type call on a g-type platform will cause a hand-off arrival of an i-type call on a g-type platform to another cell. Therefore for a homogeneous system in statistical equilibrium, the average hand-off arrival and departure rates per cell must be equal. Thus, we have

$$F_{gi} = F_{gi}' \tag{30}$$

and

$$\Lambda_h(i) = \Delta_h(i) \tag{31}$$

D. Quality-of-Service metrics

When the statistical equilibrium state probabilities and transition flows are found, the following QoS metrics can be calculated.

1. Blocking probability

The blocking probability for an i-type new call on a g-type platform is the average fraction of new i-type calls on g-type platforms that are denied access to the system. Blocking of an i-type new call on a g-type platform occurs if the system is not in the new call admission region $REG_n(g,i)$. We define the following set of states

$$B_{gi} = \{s : s \notin REG_n(g,i)\} \tag{32}$$

Then the blocking probability for i-type calls on g-type platforms is

$$P_B(g,i) = \sum_{s \in B_{gi}} p(s) \tag{33}$$

2. Hand-off failure probability

The hand-off failure probability for i-type calls on g-type platforms is the average fraction of i-type hand-off attempts on g-type platforms that are denied access to the system. Failure of i-type hand-off attempts on g-type platforms occurs if the system is not in the hand-off call admission region $REG_h(g,i)$. We define the following set of states

$$H_{gi} = \{s : s \notin REG_h(g,i)\} \tag{34}$$

Then the hand-off failure probability for i-type hand-off calls on g-type platforms is

$$P_H(g,i) = \sum_{s \in H_{gi}} p(s) \tag{35}$$

3. Forced termination probability

A more important QoS metric than hand-off failure probability is forced termination probability, $P_{FT}(g,i)$. This is defined as the probability that an i-type session on a g-type platform that is admitted will be interrupted during its lifetime because of hand-off failure.

If we let $\alpha(g,i)$ denote the probability that an i-type call on g-type platform will make a hand-off attempt and will fail on that attempt. Similarly, let $b(g,i)$ denote the probability that an i-type call on a g-type platform will make a hand-off attempt and succeed. Using the Markovian properties of the model we have

$$\alpha(g,i) = \mu_D(g) \cdot P_H(g,i) / (\mu(i) + \mu_D(g)) \tag{36}$$

and

$$b(g,i) = \mu_D(g) \cdot (1 - P_H(g,i)) / (\mu(i) + \mu_D(g)) \quad (37)$$

Assuming independent hand-off attempts we get

$$P_{FT}(g,i) = \sum_{j=0}^{\infty} \alpha(g,i) \cdot (b(g,i))^j \quad (38)$$

Summing (39) and using (37), (38) we get

$$P_{FT}(g,i) = \mu_D(g) \cdot P_H(g,i) / (\mu(i) + \mu_D(g) \cdot P_H(g,i)) \quad (39)$$

VII. DISCUSSION OF NUMERICAL RESULTS

The analysis approach was used to consider the example system in section IV. We assume that the example system uses the call admission scheme in section III and the medium access scheme in section V. The same system and traffic model parameters as in the performance evaluation of medium access scheme in section V are used. We index video, web-type and speech traffic by 1, 2, 3 respectively.

In the call admission scheme, cell loss probabilities less than 0.1%, 0.1% and 1% are required for *hand-off* video, WDS and speech traffic components, respectively. An average cell delay less than 2.5 sec ($\frac{1}{2}\overline{T_I}$) is required for the *hand-off* WDI traffic component. In our example, all the transmission QoS requirements for video, web-type and speech traffic depend only on call type. Thus, $[QoS_0]$ can be written as

$$[QoS_0] = [\overrightarrow{QoS_0(1)}, \overrightarrow{QoS_0(2)}, \overrightarrow{QoS_0(3)}]$$
$$= [(0.1\%), (0.1\%, 2.5 \text{ sec}), (1\%)] \quad (40)$$

Similarly, $[priority]$ can be written as

$$[priority] = [\overrightarrow{priority(1)}, \overrightarrow{priority(2)}, \overrightarrow{priority(3)}] \quad (41)$$

Recall that $\overrightarrow{priority(i)}$ denotes the priority given to i-type *hand-off* calls over i-type *new* calls. Specifically, $\overrightarrow{priority(1)}$ denotes the priority for *hand-off video calls* in terms of cell loss probability, $\overrightarrow{priority(2)}$ denotes the priority for *hand-off web-type calls* in terms of WDS' cell loss probability and WDI's average cell delay, and $\overrightarrow{priority(3)}$ denotes the priority *for hand-off speech calls* in terms of cell loss probability. In order to give priority to hand-offs, the admission criteria applied to *new calls* will be stricter than that applied to hand-off calls. We study the system under three different admission criteria for new calls. The three different priorities are given as

$$[priority_A]=[(-0.08\%), (-0.08\%, -1.50 \text{ sec}), (-0.8\%)] \quad (42)$$

$$[priority_B]=[(-0.05\%), (-0.05\%, -1.25 \text{ sec}), (-0.5\%)] \quad (43)$$

$$[priority_C]=[(-0.03\%), (-0.03\%, -0.83 \text{ sec}), (-0.3\%)] \quad (44)$$

Call admission policy, C_A, consists of the call admission criteria for hand-off calls, $[QoS_0]$, and the call admission criteria for new calls, $[QoS_0]+[priority_A]$. Similarly, we have call admission policies C_B and C_C. The quota on *low use* services in (13) is taken to be 20 payload slots in uplink and 35 payload slots in the downlink.

The blocking probability and forced termination probability of speech traffic under different call admission parameters are shown as functions of call origination rate in Figs. 10 and 11 respectively. There is a tradeoff between blocking probability and forced termination probability. Figs. 12 and 13 show the blocking probability and forced termination probability of video traffic under different call admission parameters respectively. Figs. 14 and 15 show the blocking probability and forced termination probability of web-type traffic under different call admission parameters on similar plots.

The increase in blocking probability as stricter QoS_n is chosen to favor the hand-off calls can be observed. Blocking probability is a "prior-admission" QoS and forced termination probability has to be guaranteed after admission. Thus, the trade-off between blocking probability and forced termination probability is desirable to the guaranteed QoS provisioning. Suppose that in the system the call origination rate for platform type 1 and 2 is less than 2.25×10^{-4} calls/sec and call admission policy C_A is used. Then, all traffic's QoS requirement on forced termination probabilities can be guaranteed. Their QoS requirements on maximum delay and cell loss probabilities are guaranteed by using the medium access scheme in section V.

For call origination rate even higher than 2.25×10^{-4} calls/sec, forced termination probabilities of all admitted sessions can still be guaranteed by using call admission criteria that is stricter than C_A. But in this way, blocking probability for new sessions will be unacceptably high such that this system is not practical. That means for given system layout the combined call admission control and medium access can only guarantee QoS provisioning for a given range of offered traffic. For heavier offered traffic, more system resource, such as payload slots, is needed to accommodate them.

VIII. CONCLUSIONS

We proposed a combined media access and call admission control for guaranteed QoS provisioning in multimedia personal communication networks with asymmetric services. We found that an appropriate combination of medium access and call admission control can be used to achieve guaranteed QoS provisioning for various admitted traffic components in multimedia personal communication networks. An overall performance analysis model was developed. Example performance characteristics are calculated and presented. Comparing the performance of video, web-type and speech traffic, we can see that it is most difficult to provide guaranteed QoS to video traffic in the multimedia personal communication networks. Comparing the performance of video sessions on different platform types, we can see that it is much more difficult to provide guaranteed QoS to video sessions on high mobility platforms than to video sessions on low mobility platforms. If only video sessions on low mobility platforms were to be supported by the system, the range of offered traffic in which the system can accommodate with guaranteed QoS provisioning to all admitted sessions would be extended much.

REFERENCES

[1] S. Nanda and D. J. Goodman, " Performance of PRMA: A Packet Voice Protocol for Cellular Systems," *IEEE Trans. Veh. Technol.*, Vol. 40, No. 3. Aug. 1991.

[2] G. Anastasi, D. Grillo, and L. Lenzini, "An Access Protocol for Speech/Data/Video Integration in TDMA-Based Advanced Mobile Systems," *IEEE JSAC.*, Vol. 15., No. 8., Oct. 1997.

[3] A. S. Acampora and M. Naghshineh, "Control and Quality-of-Service Provisioning in High-Speed Microcellular Network," IEEE Personal Communications Mag., Vol. 1, No. 2, Second Quarter, 1994.

[4] G. Zhang and S. S. Rappaport, "Call Admission, Medium Access and Guaranteed Quality-of-Service Provisioning for ATM-Based Wireless Personal Communication Networks," Technical Report #759, College of Engineering and applied Sciences, State University of New York, Stony Brook, NY 11794.

[5] D. Raychaudhuri and N. D. Wilson, " ATM-based transport architecture for multiservices wireless personal communication networks," *IEEE JSAC*, Vol. 12, No. 8, Oct. 1994.

[6] D. Hong and S. S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Non-Prioritized Hand-off Procedures," *IEEE Trans. Veh. Technol.*, Aug. 1986, Vol. 35, No. 3, pp. 77-92.

[7] S. S. Rappaport, "Blocking Hand-off and Traffic Performance Analysis for Cellular Communication Systems with Mixed Platforms," *IEE Proceedings*, Part I, Communications, Speech and Vision, October 1993, Vol. 40, No. 5, pp. 389-401.

[8] S. S. Rappaport, and C. Purzynski, "Prioritized resource assignment for mobile cellular systems with mixed services and platform types," *IEEE Trans. Veh. Technol.*, Vol. 45, No. 3, Aug. 1996.

[9] Y. Park and S. S. Rappaport, "Cellular Communication Systems with Voice and Background Data," in D. J. Goodman and D. Raychaudhuri (Eds.) 'Mobile Multimedia Communications,' (Plenum Press, New York, 1997), pp. 33-42.

[10] P. V. Orlik and S. S. Rappaport, "Traffic Performance and Mobility Modeling of Cellular Communications with Mixed Platforms and Highly Variable Mobilities," *Proceedings of IEEE, Special Issue on Mobile Radio Centennial*, vol. 86, no. 7, July 1998, pp. 1464-1479. See also Proceedings of IEEE, vol. 86, no. 10, October 1998, p. 2111.

[11] P. V. Orlik and S. S. Rappaport, "A Model for Teletraffic Performance and Channel Holding Time Characterization in Wireless Cellular Communication with General Session and Dwell Time Distributions," *IEEE JSAC*, Vol. 16, No. 5, June 1998.

[12] Z. J. Haas and D. A. Dyson, "On the Performance of the Dynamic Packet Reservation Medium Access Scheme in the Presence of Fading," *VTC*'98.

[13] CCITT Recommendation, "Video codec for Audio/Visual Services at px64kps," H.261 CCITT, July 1990.

[14] ITU-T, "Video coding for Low Bit Rate Communication," Draft ITU-T Recommendation H.263, Dec. 5, 1995.

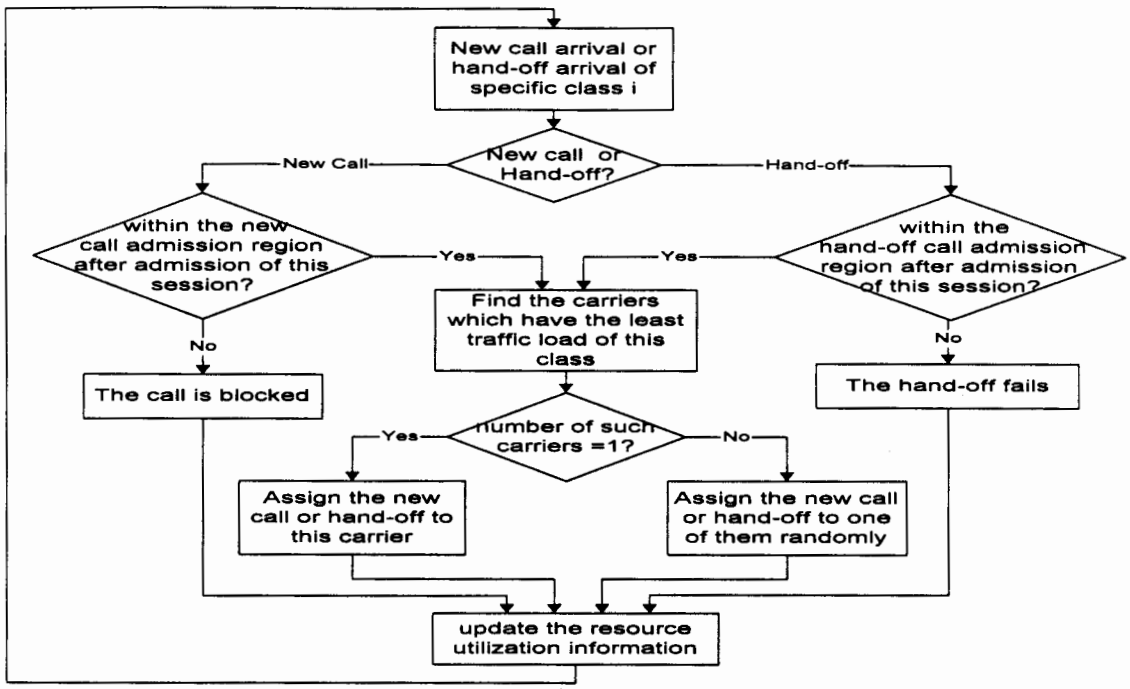

Fig. 1: The call admission control

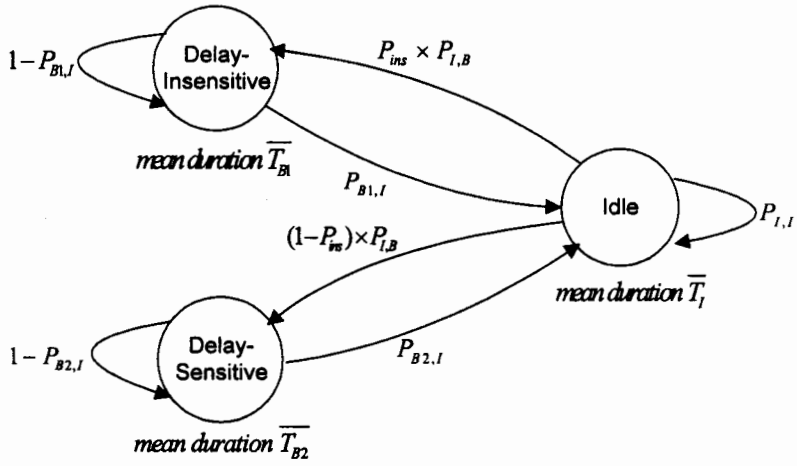

Fig. 2: The web-type traffic model at *Busy* and *Idle* periods level: A web-type session consists of alternating busy and idle periods.

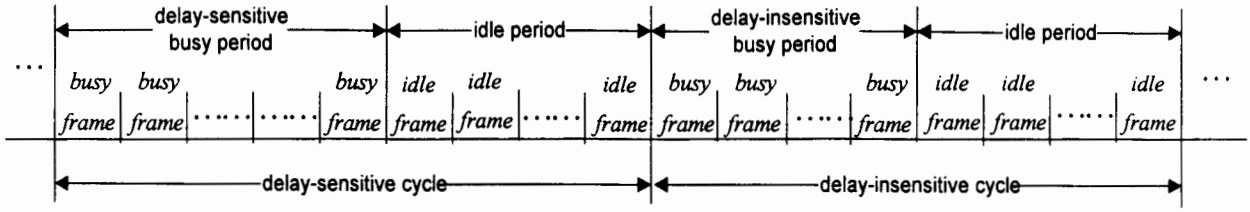

Fig. 3: The web-type (uplink or downlink) traffic model from the frame's point of view.

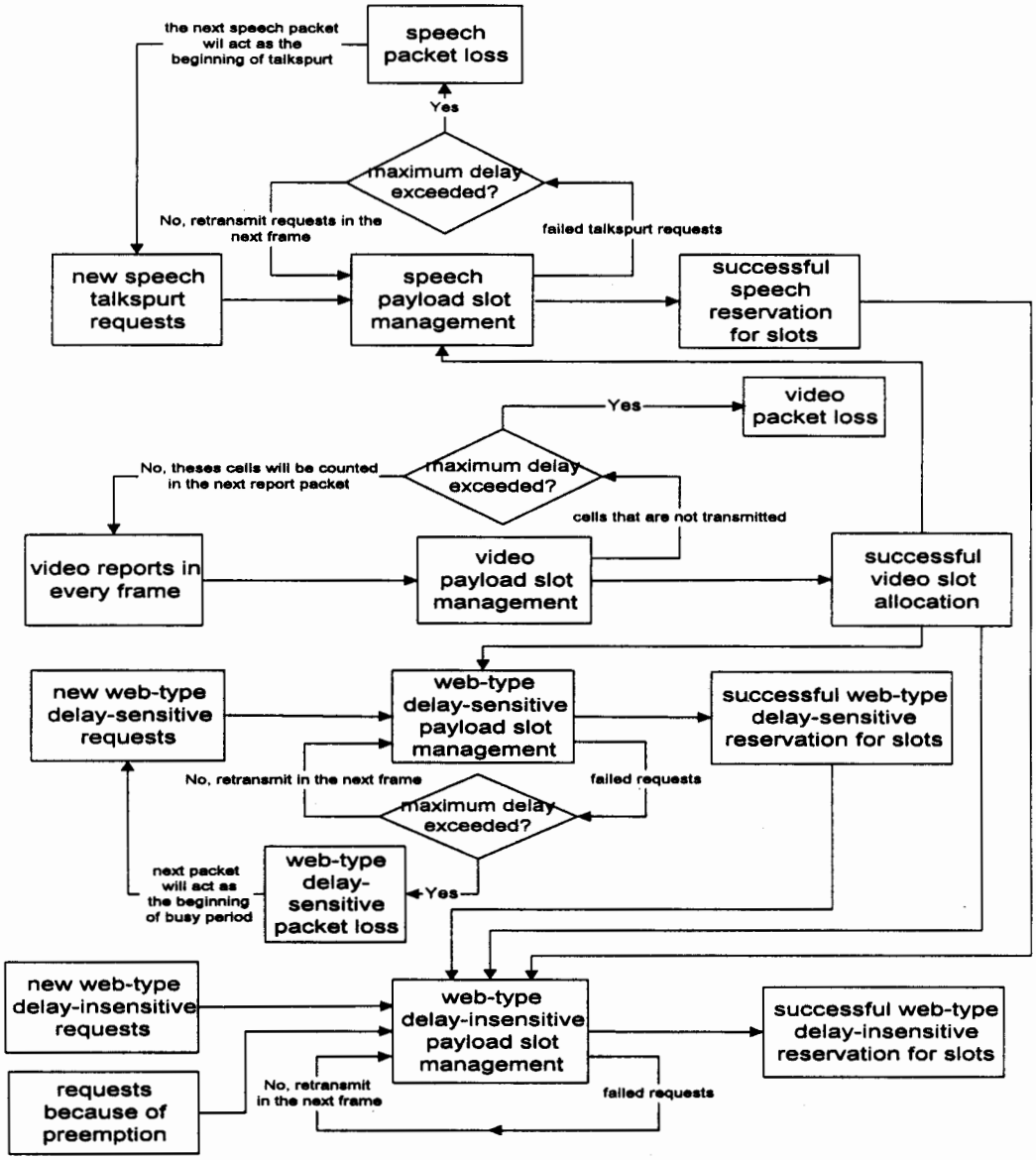

Fig. 4: The Medium access scheme.

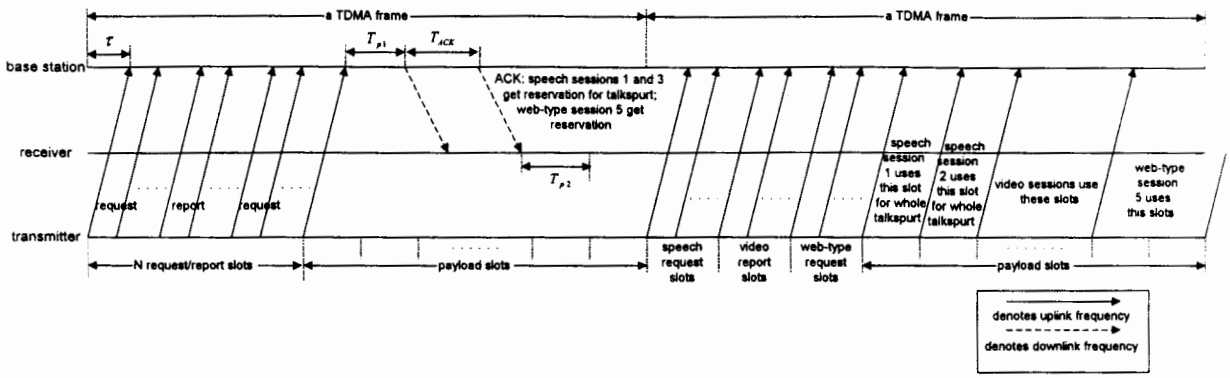

Fig. 5: The timing of the medium access scheme.

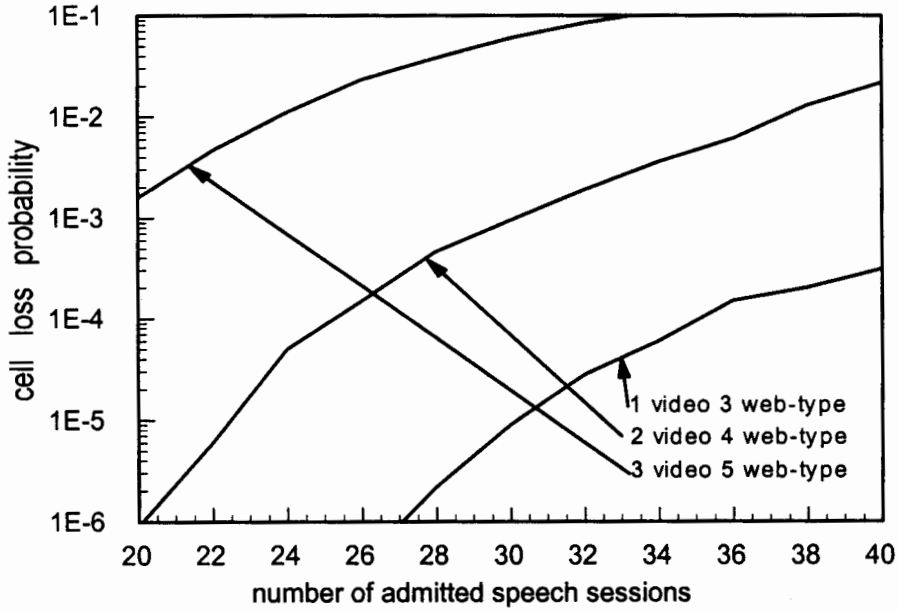

Fig. 6: Cell loss probability of video traffic.

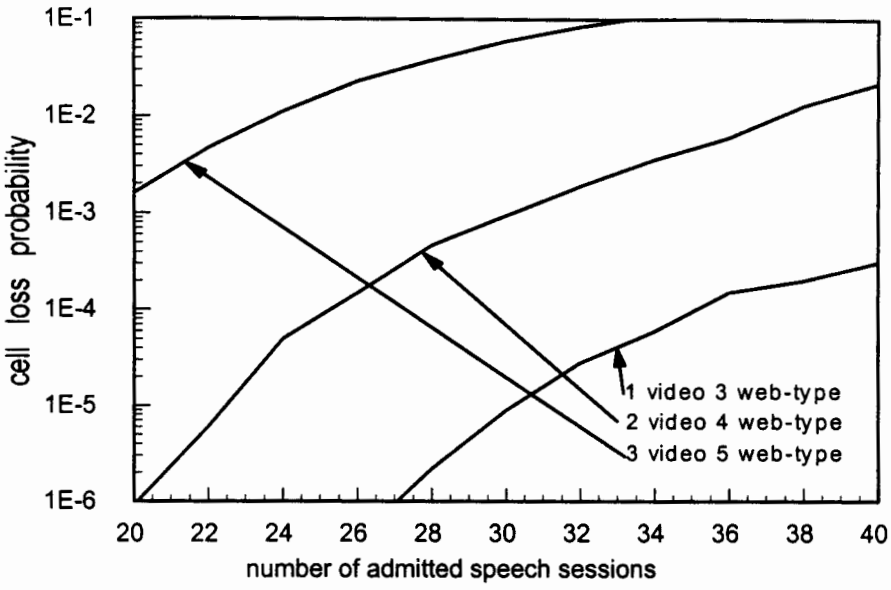

Fig. 7: Cell loss probability of speech traffic.

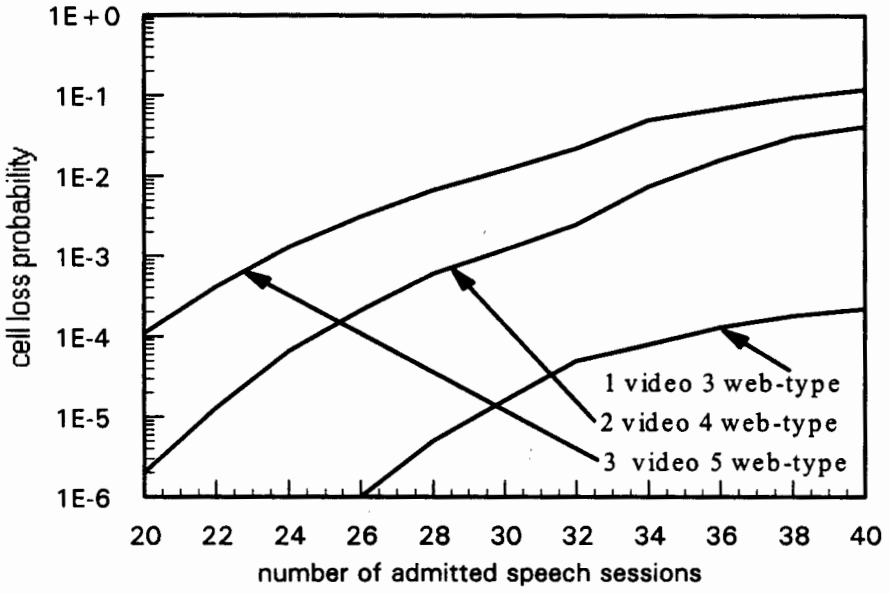

Fig. 8: Cell loss probability of web-type delay-sensitive traffic.

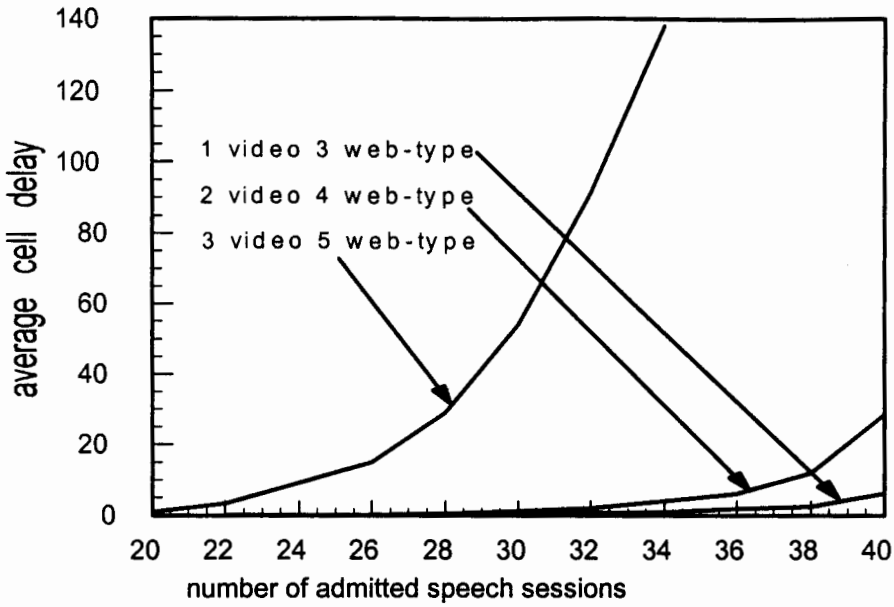

Fig. 9: The average cell delay of web-type delay-insensitive traffic.

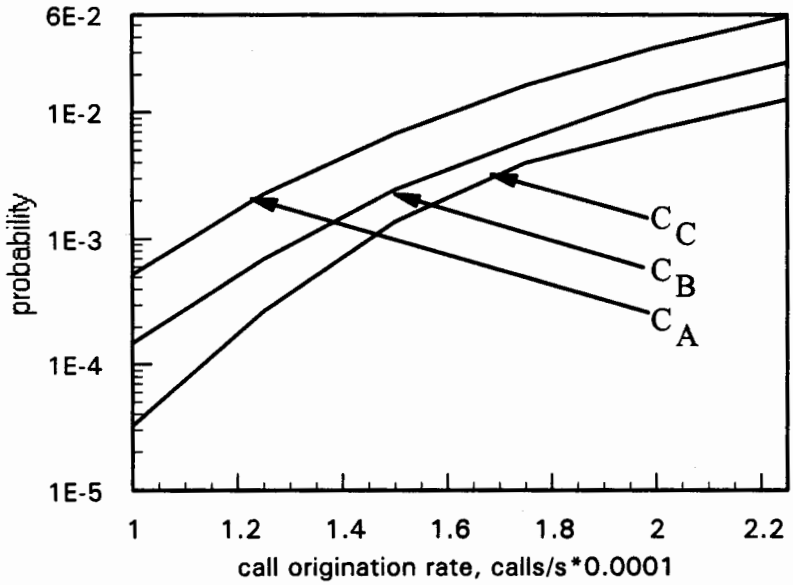

Fig. 10: Speech sessions' blocking probability.

S=15, G=2, I=3, C_{up}=2, C_{down}=3, $\nu(1,0)=\nu(2,0)$=200, $\Lambda(1,i)/\Lambda(2,i)=1.0$, $\Lambda_n(g,1)/\Lambda_n(g,3)=0.1$, $\Lambda_n(g,2)/\Lambda_n(g,3)=0.5$, $\overline{T}(1)$=100s, $\overline{T}(2)$=200s, $\overline{T}(3)$=100s, $\overline{T}_D(1)$=100s, $\overline{T}_D(2)$=500s.

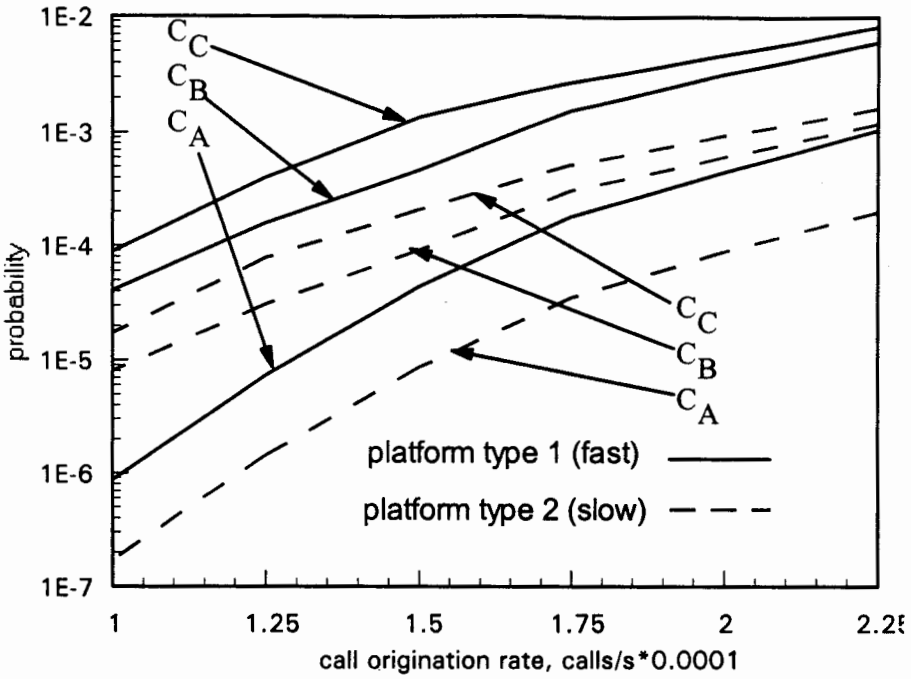

Fig. 11: Speech sessions' forced termination probability.

S=15, G=2, I=3, C_{up}=2, C_{down}=3, $\nu(1,0)=\nu(2,0)$=200, $\Lambda(1,i)/\Lambda(2,i)=1.0$, $\Lambda_n(g,1)/\Lambda_n(g,3)=0.1$, $\Lambda_n(g,2)/\Lambda_n(g,3)=0.5$, $\overline{T}(1)$=100s, $\overline{T}(2)$=200s, $\overline{T}(3)$=100s, $\overline{T}_D(1)$=100s, $\overline{T}_D(2)$=500s.

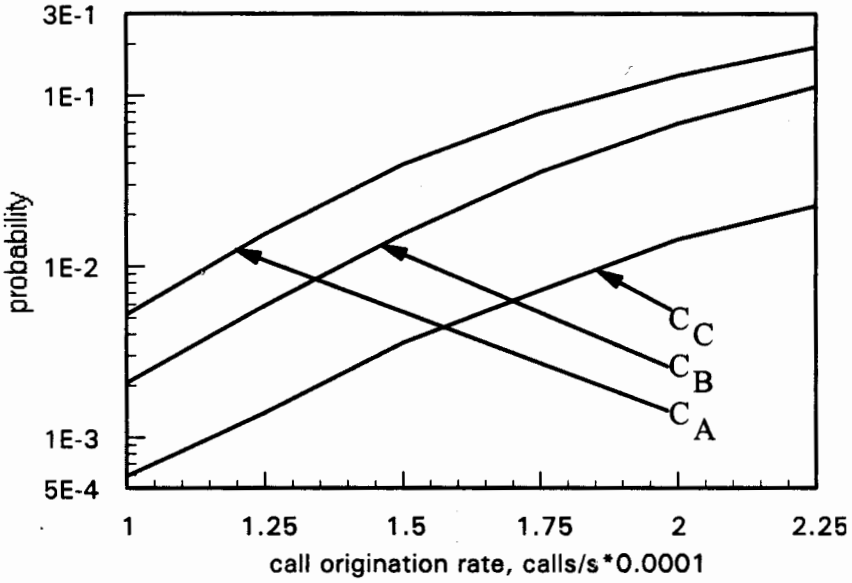

Fig. 12: Video sessions' blocking probability.

S=15, G=2, I=3, C_{up}=2, C_{down}=3, $\nu(1,0)=\nu(2,0)$=200, $\Lambda(1,i)/\Lambda(2,i)=1.0$, $\Lambda_n(g,1)/\Lambda_n(g,3)=0.1$, $\Lambda_n(g,2)/\Lambda_n(g,3)=0.5$, $\overline{T}(1)$=100s, $\overline{T}(2)$=200s, $\overline{T}(3)$=100s, $\overline{T}_D(1)$=100s, $\overline{T}_D(2)$=500s.

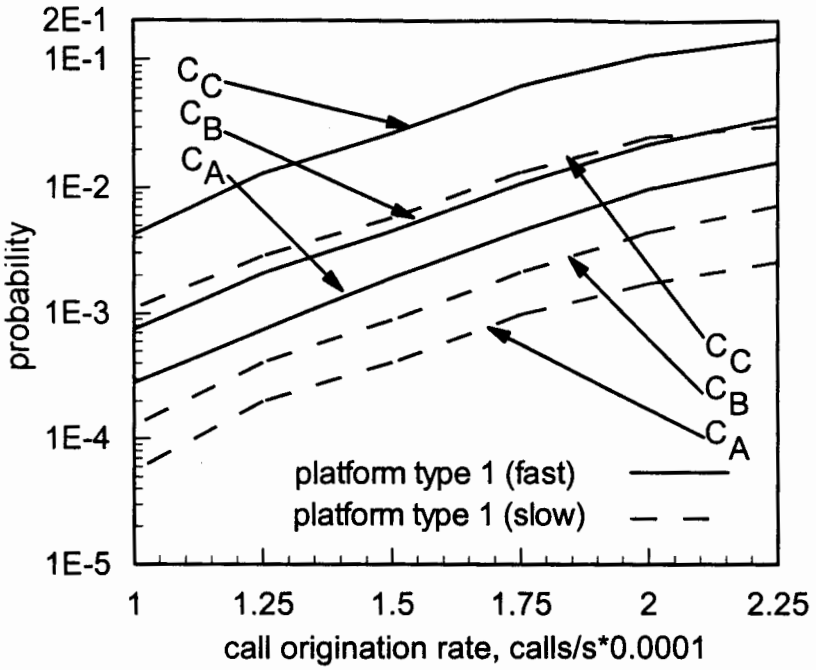

Fig. 13: Video sessions' forced termination probability.

S=15, G=2, I=3, C_{up}=2, C_{down}=3, $\nu(1,0) = \nu(2,0)$=200, $\Lambda(1,i)/\Lambda(2,i) = 1.0$, $\Lambda_n(g,1)/\Lambda_n(g,3) = 0.1$, $\Lambda_n(g,2)/\Lambda_n(g,3) = 0.5$, $\overline{T}(1)$=100s, $\overline{T}(2)$=200s, $\overline{T}(3)$=100s, $\overline{T}_D(1)$=100s, $\overline{T}_D(2)$=500s.

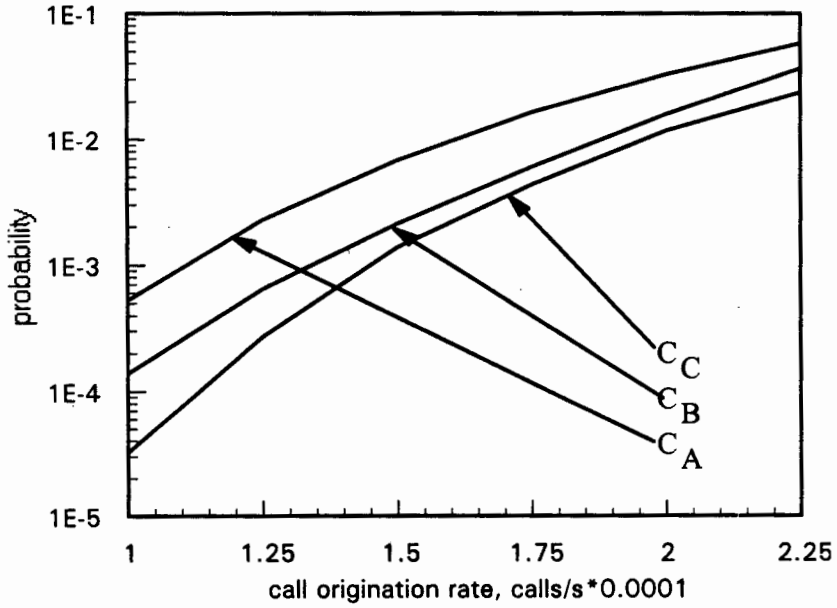

Fig. 14: Web-type sessions' blocking probability.

S=15, G=2, I=3, C_{up}=2, C_{down}=3, $\nu(1,0) = \nu(2,0)$=200, $\Lambda(1,i)/\Lambda(2,i) = 1.0$, $\Lambda_n(g,1)/\Lambda_n(g,3) = 0.1$, $\Lambda_n(g,2)/\Lambda_n(g,3) = 0.5$, $\overline{T}(1)$=100s, $\overline{T}(2)$=200s, $\overline{T}(3)$=100s, $\overline{T}_D(1)$=100s, $\overline{T}_D(2)$=500s.

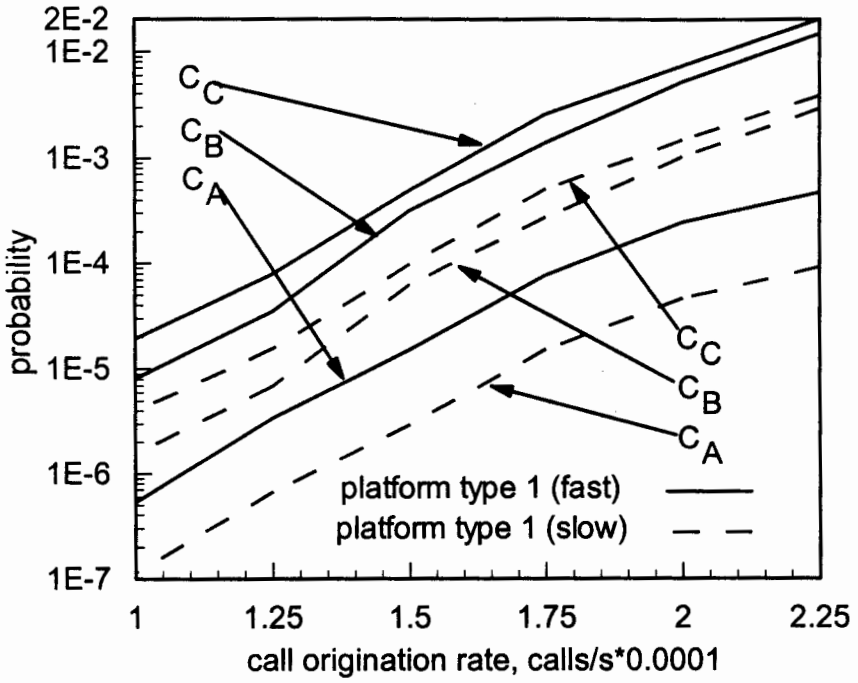

Fig. 15: Web-type sessions' forced termination probability.

S=15, G=2, I=3, C_{up}=2, C_{down}=3, $\nu(1,0) = \nu(2,0)$=200, $\Lambda(1,i)/\Lambda(2,i) = 1.0$, $\Lambda_n(g,1)/\Lambda_n(g,3) = 0.1$, $\Lambda_n(g,2)/\Lambda_n(g,3) = 0.5$, $\overline{T}(1)$=100s, $\overline{T}(2)$=200s, $\overline{T}(3)$=100s, $\overline{T}_D(1)$=100s, $\overline{T}_D(2)$=500s.