Achievable QoS for Multiple Delay Classes in Cellular TDMA Environments*

K. M. Tong and Mounir Hamdi
Department of Computer Science
Hong Kong University of Science and Technology
Kowloon, Hong Kong

Abstract In a real-time wireless TDMA environment, every packet generated by applications has a deadline associated with it. If the system cannot allocate enough resources to serve the packet before the deadline, the packet would be dropped. Different applications have different delay requirements that should be guaranteed by the system so as to maintain some given packet dropping probabilities. In this paper, a single-cell system traffic of multiple delay classes is mathematically analyzed, and it is proved to be independent of the scheduling algorithm used, for all work-conserving earliest-due-date (WC-EDD) scheduling algorithms. The dropping requirements of all individual applications are guaranteed using deadline-sensitive ordered-head-of-line (DSO-HoL) priority schemes. Verification of the model is shown through extensive simulations.

1 Introduction

Wireless communication is becoming increasingly important in recent decades. Services which a wireless network can provide are evolving from analog voice communication to digital multimedia services including video, voice and data [7]. Typically, a wireless network consists of mobile devices, base stations, and the backbone network. A single base station can only cover a limited geographical area, or cell, and the mobile devices inside the area communicate with the base station using some radio frequencies in a shared manner. To enable communications between mobile devices of different cells, the base stations need to be connected, usually via a fast, wired backbone network so that the packets from the source mobile can be forwarded to the destination cell and transmitted to the receiving mobile.

Within a cell, all mobile hosts share the transmission medium under a certain Medium Access Control (MAC) scheme. Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) are two examples. To support end-to-end Quality-of-Service (QoS), both wired and wireless media must be considered. The wired backbone is required to deliver the packets according to the QoS requirements and select a good route to achieve the aim. In wireless networks, QoS can be classified into two categories, call level and packet level. Call level QoS defines the call blocking probability and handoff dropping probability, whereas at packet level, or MAC level, QoS defines the packet dropping probability and delay tolerance.

Compared to the wired network, it is more challenging to provide QoS guarantees in the wireless network because of resource limitation, error characteristic and mobility. In [10], the authors claim the necessity of QoS support in wireless networks due to their fundamental differences from the wired counterparts, and QoS renegotiation and adaptation are required for wireless applications. From this point of view, supporting QoS in wireless networks is a collaboration between the network side and the application side.

Protocols like IEEE 802.11, have been proposed in recent years to support bandwidth from 1 Mbps to 11 Mbps, which is comparable to wire-line networks. But still, the QoS scheduling algorithm running at the base station should be aware of the diversity of the requirement of each application, and schedule sufficient bandwidth to the applications efficiently.

Scheduling the available bandwidth to different services is a non-trivial task. Some applications, e.g. video, require a huge amount of bandwidth and are very delay sensitive, whereas some others, e.g. voice, require a relatively small amount of bandwidth with less strict delay requirements. These applications should all be satisfied if the system has enough resources to support and maintain the QoS of each individual application.

This paper focuses on the packet level QoS in a TDMA wireless system. Following the suggestions of [11, 12], we adopt a TDMA MAC for the wireless scheme with on-demand scheduling of resources. Then, mathematical bounds on the delay and packet dropping probability in a single cell are derived.

The rest of the paper is organized as follows. In Section 2, related work is presented, followed by the description of the model for single-cell wireless network in section 3. Section 4 derives the mathematical bounds for the packet dropping probability in a single cell along with extensive simulation results.

2 Previous Work

In a wireless TDMA environment, any packet can be dropped mainly because of the following reasons: Not enough resources

*This research work was supported in part by a Hong Kong Research Grant Council Grant.

0-7803-6596-8/00/$10.00 © 2000 IEEE

for sending a critical packet; wireless channel quality; or collision of packets from different hosts. Typically, there are three ways to lower the packet dropping rate: Admission control; forward error correction; or a good resource scheduling algorithm.

In [1], the authors propose a call admission control for the Packet Reservation Multiple Access Protocol with Dynamic Allocation, PRMA/DA. Simulation is given to show the performance on the delay and loss rate of the protocol. [2] analyses a new MAC protocol D-RAMA and also simulates the performance on delay and loss rate. Comparative results are drawn based on the simulation performance of some other MAC protocols.

An analytical approach is used in [3], where the achievable QoS, in terms of system dropping probability and individual dropping probability, in a general TDMA system is discussed. The authors use the traffic distributions and the delay requirement of the applications to calculate the best QoS the system can achieve. It is shown that the system residual traffic and the system dropping rate are independent on the scheduling algorithm used. Individual dropping rate is also derived, and based on the individual dropping rate. Using the achievable region, an admission algorithm is suggested so that if a new application is beyond the achievable region, which means that the system cannot deliver the required QoS to the new application, the application would not be admitted.

For [4, ?], the error characteristic of the wireless channel is also modeled. Similar to [3], region of achievable QoS is calculated, with some enhancement in the scheduling algorithm. Under the scheduling algorithm, the packet dropping probabilities would be lowered.

However, in [3, 4, ?], the maximum possible delay for each packet is 2 TDMA frames. This limits the divergences of services which can be supported by the network. It would be more flexible if the system can support multiple delay classes not limited to 2 TDMA frames because of the diversity of the delay requirements for different applications.

The key contribution of this paper is to extend [3, 4, ?] in such a way that it supports any delay requirement. Under this extension, the system would be more realistic and flexible.

3 Wireless Network Model

TDMA is one of the common ways for sharing a channel in wireless networks. In TDMA, time is divided into fix-sized frames. A frame can be divided further into slots, which can be fix-sized or variable-sized. When an application is trying to send a packet over the air, it must ensure there are free slots available so as to avoid collisions with other packets. One way to ensure the availability of slots is by contention. Another way is by allocation in the base station (BS). Typically, there is a queue in every mobile host (MH) for holding ready-to-send packets. If the BS informs a MH that the number of slots available to the MH, the MH would select some packets from the queue for transmission. If real-time applications are to be supported, the queue would be a priority queue such that the MH would select the most critical packets first when the slots are available. Figure 1 shows the general model under consideration.

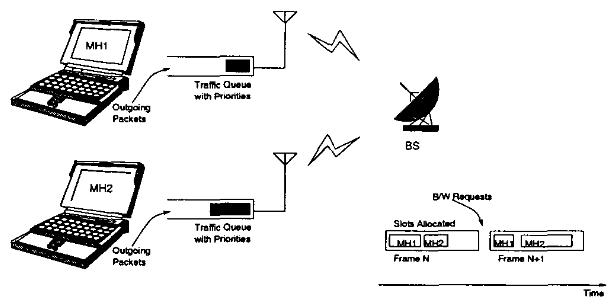

Figure 1: A general TDMA system

Each MH is assumed to make requests to the BS at frame boundaries. The BS would schedule the available bandwidth to different MHs according to the degree of importance.

In this paper, a generalization of the traffic model used in [?], which supports multiple delay requirements for different applications, is proposed as shown in Figure 2. We assume all packets generated by a particular application have the same deadline, which is specified in terms of number of frames. An application may generate very time-sensitive packets which must be sent within the next frame or it must be dropped otherwise. Other applications, which can tolerate a longer delay, may still drop their packets if they are not serviced within, for instance, 4 frames. We define:

- A class N packet would be dropped if it cannot get serviced in the next N frames.
- A class N application would only generate class N packets.

N is basically the maximum tolerable delay parameter of an application which the network should guarantee. As a class N application generates class N packets, a class N packet may become a class $N-1$ packet if it cannot be serviced in the current frame, since the maximum delay tolerable of the packet is changed from N frames to $N-1$ frames. This kind of class $N-1$ packet is called the *residual packet*.

An application i is said to be of class c_i if it would only generate new packets that must be serviced within c_i frames. We define $\lambda_i(n)$ to be the number of new packets from application i entering the system at the beginning of frame n. We also define *residual packets* to be packets generated in previous frames which are neither expired (dropped) or serviced (transmitted). Let $r_i^c(n, f)$ be the number of residual packets at the beginning of frame n from application i that must be serviced within the

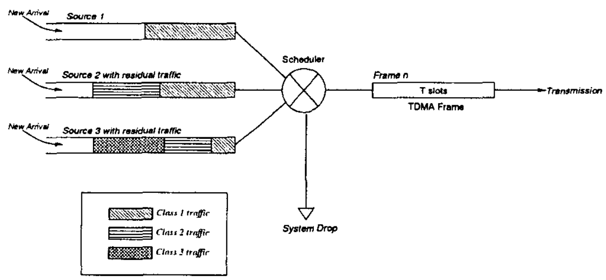

Figure 2: System model. The scheduler is used to allocate the slots to different applications.

next c frames, where f is the scheduling algorithm running at the base station.

The new arrival of all class c application form the system arrival of class c, defined as:

$$N_c(n) = \sum_{i,st.c_i=c} \lambda_i(n) \tag{1}$$

The total residual packets of class c in the system is:

$$R_c^f(n) = \sum_{i,st.c_i=c} r_i^c(n, f) \tag{2}$$

The total number of class c packets in the system is defined as:

$$\Lambda_c^f(n) = N_c(n) + R_c^f(n) \tag{3}$$

4 Single Cell QoS Guarantee

4.1 Work Conserving Earliest Due Date

As seen in the previous section, the amount of system residual traffic depends on the scheduling algorithm f. In order to be able to analyze the traffic, let $\mathcal{F}$ be a set of all work-conserving earliest-due-date (WC-EDD) policies. Work-conserving means no slots would be wasted and left idle whenever there are packets waiting to be transmitted, and earliest-due-date priority allocates resource to serve packets with the nearest deadline.

It is a well-known result that WC-EDD policy would minimize system packet dropping rate by maximizing the throughput of the whole system [6]. When the scheduling algorithm $f \in \mathcal{F}$ is applied in the BS, system dropping rate would be the lowest among other scheduling algorithms.

Suppose there are N classes of traffic, $\Lambda_1^f(n), \ldots, \Lambda_N^f(n)$ at frame n, competing for T slots. Let the corresponding service vector be $\left[a_1(n), a_2(n), \ldots, a_N(n)\right]$, which indicates the number of slots allocated to each traffic class. WC-EDD policies determine a_i's in the following way:

$$\begin{aligned} a_1(n) &= \min\{\Lambda_1^f(n), T\} \\ a_2(n) &= \min\{\Lambda_2^f(n), T - a_1(n)\} \\ &\vdots \\ a_N(n) &= \min\{\Lambda_N^f(n), T - \textstyle\sum_{i=1}^{N-1} a_i(n)\} \end{aligned} \tag{4}$$

4.2 Analysis of System Traffic

4.2.1 Independence of Scheduling Algorithm

Theorem: System traffic is independent of scheduling algorithm $f \in \mathcal{F}$.

Proof. For $f \in \mathcal{F}$, we can write the residual traffic of class c in the follow form:

$$R_c^f(n+1) = \begin{cases} \max\left(\sum_{i=1}^{c+1} \Lambda_i^f(n) - T, 0\right) & \text{if } \sum_{i=1}^{c} \Lambda_i^f(n) < \\ \Lambda_{c+1}^f(n) & \text{if } \sum_{i=1}^{c} \Lambda_i^f(n) \geq \end{cases} \tag{5}$$

Let the initial frame be 0. At frame 0, there are no residual traffic by definition. Therefore:

$$\begin{aligned} R_c^f(0) &= 0 \\ \Lambda_c^f(0) &= N_c(0) \end{aligned} \tag{6}$$

From (5), $R_c^f(1)$ depends on $\Lambda_i^f(0)$, $1 \leq i \leq c+1$. But in fact, all $\Lambda_i^f(0)$'s are zero and independent on f. Therefore, $R_c^f(1)$ is also independent on f.

By induction, $R_c^f(n)$ is independent on f. As from (3), the system traffic $\Lambda_c^f(n)$ is also independent on f. We can conclude that the system traffic is independent on which particular WC-EDD scheduling algorithm being used. The superscript f can be eliminated and the system traffic can be rewritten as:

$$\begin{aligned} R_c^f(n) &\triangleq R_c(n) \\ \Lambda_c^f(n) &\triangleq \Lambda_c(n) \end{aligned} \tag{7}$$

Suppose there are N traffic classes in the system, let $\mathbf{S}(n) = \left[\Lambda_1(n), \Lambda_2(n), \cdots, \Lambda_N(n)\right]$ be the system status at frame n.

If $\mathbf{S}(n)$ is known, residual traffic $R_i(n+1)$'s can be deduced using (5). $\mathbf{S}(n+1)$ can be expressed as follows:

$$\mathbf{S}(n+1) = \begin{bmatrix} R_1(n+1) + N_1(n+1) \\ R_2(n+1) + N_2(n+1) \\ \vdots \\ R_{N-1}(n+1) + N_{N-1}(n+1) \\ N_N(n+1) \end{bmatrix} \tag{8}$$

Given $\mathbf{S}(n) = [i_1, i_2, \cdots, i_N]$, if there exists x such that $\sum_{r=1}^{x} i_r < T$ and $\sum_{r=1}^{x+1} i_r \geq T$, then the residual traffic $R_j(n+1)$ would be:

$$R_j(n+1) = \begin{cases} \max\left(\sum_{r=1}^{j+1} i_r - T, 0\right) & \text{if } j \le x, \\ i_{j+1} & \text{if } j > x \end{cases} \quad (9)$$

By substituting (9) into (8), the transition probability from $\mathbf{S}(n)$ to $\mathbf{S}(n+1)$ can be expressed in the following form:

$$P\Big(\mathbf{S}(n+1) = [j_1, j_2, \cdots, j_N] \Big| \mathbf{S}(n) = [i_1, i_2, \cdots, i_N]\Big)$$

We can easily show that the system exhibits Markov property in vector space [5]. However, if there are N traffic classes in the system, the transition matrix is N-dimensional. If stationary processes are assumed for all traffic classes, the distribution can be obtained either by solving the N-dimensional transition matrix, or calculated by iteration. Iteration provide a quick approximate solution where the accuracy can be adjusted by changing the number of iterations that apply to equation (??).

4.2.2 System Packet Dropping Probability

Packet dropping, by definition, would be introduced if there are class 1 packets that are not serviced within the current frame. Therefore, system packet dropping probability is directly related to system class 1 traffic. The expected system dropping rate follows directly from the definition.

$$b_S = \mathrm{E}\Big(\Lambda_1(n) - T \Big| \Lambda_1(n) > T\Big) \times P\Big(\Lambda_1(n) > T\Big) \quad (10)$$

The expected system dropping rate describes the number of packets that are dropped per frame. Therefore, the expected system dropping probability can be obtained by dividing the expected system dropping rate by the average bandwidth required.

4.2.3 Individual Packet Dropping Probability

It is not enough to only guarantee the system packet dropping probability. Individual packet dropping probability for each application is important as well. However, the individual dropping probability can vary depending on the scheduling algorithm.

The expected individual dropping rate is defined as:

$$b_i = \mathrm{E}\Big(\Lambda_1^i(n) - T \Big| \Lambda_1^i(n) > T\Big) \times P\Big(\Lambda_1^i(n) > T\Big) \quad (11)$$

where Λ_1^i is the class 1 traffic induced by application i.

4.2.4 Deadline-Sensitive Ordered-Head-of-Line Priority

In [3, ?], a scheduling algorithm, Deadline-Sensitive Ordered-Head-of-Line Priority (DSO-HoL), is used to deliver the target individual packet dropping probability. DSO-HoL priority scheme belongs to the WC-EDD class policy. Under DSO-HoL, every application is given a priority, or service order.

Suppose there are three applications, namely, S_1, S_2, and S_3. If the service order is $\pi_{S_1} = 2, \pi_{S_2} = 3, \pi_{S_3} = 1$, then S_3 has a higher priority than S_1, and S_1 has higher priority than S_2. Application x has a higher priority if π_x is smaller. DSO-HoL services all class 1 packets in the order $\{\pi_{i_1}, \pi_{i_2}, \cdots, \pi_{i_n}\}$. For all class 2 packets, the service order is also $\{\pi_{i_1}, \pi_{i_2}, \cdots, \pi_{i_n}\}$, and so on for every traffic class.

4.2.5 Individual Traffic Distribution

All applications in the system, S, can be divided into two sets: g, the priority group, and $S - g$, the non-priority group. Since packets from g would be serviced first, for every class, resource is allocated to packets from g. Therefore class i traffic generated from g, denoted by Λ_i^g, can be calculated by the following equations:

$$\Lambda_i^g(n+1) = N_i^g(n+1) + R_i^g(n+1) \quad (12)$$

The subset traffic of class i depends on the system traffic of more urgent classes c, $1 \le c \le i$, and the subset traffic of the next class $i+1$ at previous frame.

(13) shows the expectation of packet dropping probability of the priority group g.

$$b_g = \mathrm{E}\Big(\Lambda_1^g(n) - T \Big| \Lambda_1^g(n) > T\Big) \times P\Big(\Lambda_1^g(n) > T\Big) \quad (13)$$

Given that DSO-HoL priority is employed, the packet dropping probability of each application can be found by first considering the highest priority application and then the first two highest priority applications, and so on:

$$\begin{aligned} b_1 &= b_{\{1\}} \\ b_2 &= b_{\{1,2\}} - b_{\{1\}} \\ &\vdots \\ b_n &= b_{\{1,2,\cdots,n\}} - b_{\{1,2,\cdots,n-1\}} \end{aligned} \quad (14)$$

Define f_i be a function representing the scheduling order such that $f_i(1)$ would be the highest priority application, and $f_i(n)$ would be the lowest priority application. One particular f_i induces one particular dropping pattern, the packet dropping vector is defined to be $\mathbf{d}^{f_i} = \left[b_{f_i(1)}, b_{f_i(2)}, \cdots, b_{f_i(n)}\right]$.

There are $n!$ different ways of ordering for n application, therefore there are $n!$ distinct functions, $f_1, f_2, \cdots, f_{n!}$, which induce $n!$ dropping vectors $\mathbf{d}^{f_1}, \mathbf{d}^{f_2}, \cdots, \mathbf{d}^{f_{n!}}$.

4.3 Scheduling Target QoS

Let α_i be the probability that the scheduling order f_i is used. Since f_i would induce a dropping vector $\mathbf{d}^{f_i}$, the average dropping vector can be calculated as:

$$\mathbf{d} = \sum_{i=1}^{n!} \alpha_i \mathbf{d}^{f_i} \qquad \text{where } \sum_{i=1}^{n!} \alpha_i = 1 \tag{15}$$

To schedule the target QoS, all α_i's should be calculated. Since all $\mathbf{d}^{f_i}$ can be calculated for a given priority scheme f_i, α_i's can be deduced by using a linear programming technique. Once α_i's are found, the target QoS can be scheduled in the way that a priority scheme f_i is employed with probability α_i, and another scheme f_j is employed with probability α_j. This probabilistic DSO-HoL priority is called *mixed DSO-HoL priority*.

4.4 Packet Level Call Admission Control

Each application is required to specify the traffic distribution, the delay class to which it belongs, and the maximum packet dropping rate. Based on the traffic distribution and the delay class, corresponding dropping vectors are known under different priority orderings. The admission controller can calculate whether such α_i's exist so that the packet dropping rates of all applications can be satisfied. If it cannot find such α_i's, this implies that the new connection would cause delay violations, either to the existing applications, or the new connection itself. Such a new connection would be simply dropped.

5 Numerical Examples

In this section, simulation results are presented for checking the validity and accuracy of the model of the previous section.

5.1 Traffic Models and System Parameters

Two source types, voice and video (MPEG-1), are used for the analysis. For each source, the distribution would be given to the system, as well as the delay class it belongs to. Assume each TDMA slot represents a 32 Kbps channel. A single voice channel can be modeled by an ON-OFF process with a maximum transmission rate of 64 Kbps [9]. MPEG is a common video encoding method used for services like Video-on-Demand (VoD) [8]. The distributionof the video sources represents an MPEG-1 source with the mean transmission rate of 0.648 Mbps.

We assume the video source is more time-sensitive than the voice source, let the MPEG-1 source to be of class 1 and the Voice source be class 3.

We assume the system bandwidth is 1 Mbps, that corresponds to 32 available slots ($T = 32$), each slot represents 32 Kbps. Suppose there are 2 applications in the system, 1 MPEG-1 stream and 1 voice application.

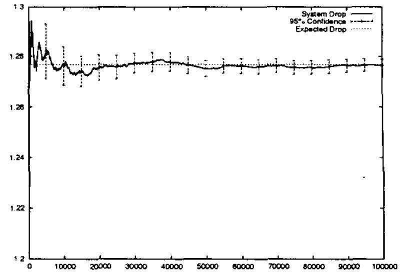

Figure 3: Time-averaged System Dropping Rate under {Voice, MPEG} priority

5.2 Results

Under the previously described system, the expected system drop b_S is calculated to be 1.27683 *packets per frame*. Individual dropping rates under different DSO-HoL priority are also calculated as shown in Table 1.

Service Order	Voice drop	MPEG drop	Sys. dro
1: {Voice, MPEG}	0	1.27683	1.276
2: {MPEG, Voice}	0.0015283	1.2753	1.276
Mix: 1:(50%) 2:(50%)	0.0007641	1.276065	1.276

Table 1: Calculated Values of Dropping Rates

5.2.1 Simulation Results

Independence of Scheduling Algorithm

Since system dropping is independent of the scheduling algorithm, simulation is run under service order 1: {Voice, MPEG}, and service order 2: {MPEG, Voice}. The expected values are shown in Table 1 and are verified in Figures 3, 4, 5, 6 and 7.

As shown in Figures 3 and 4, *System Drop* denotes the time-averaged result of the system dropping rates under different priority schemes. Notice that the system dropping rates simulated quickly converge to the expected value and stabilized, whatever the scheduling algorithm used.

Individual Dropping Rate

Individual dropping rates under different priority schemes are simulated. Under {Voice, MPEG} service order, the voice source has the priority and no dropping would be incurred. The dropping rate for MPEG-1 source would equal to the system dropping rate.

If {MPEG, Voice} priority is used, both sources would have some packets dropped. Although the value for voice dropping is very low, the expected value still fits into the 95% confidence interval, with the accuracy of less than 5%.

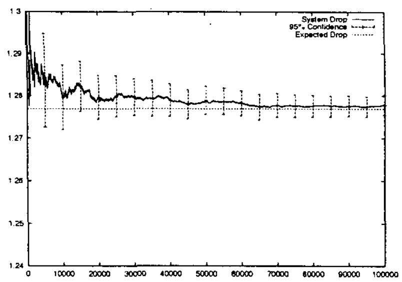

Figure 4: Time-averaged System Dropping Rate under {MPEG, Voice} priority

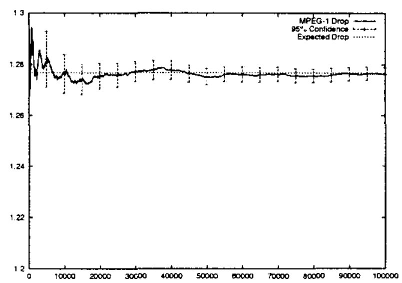

Figure 5: MPEG Dropping Rate under {Voice, MPEG} priority

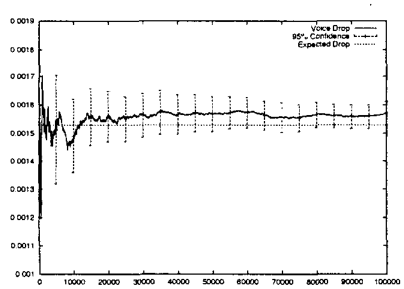

Figure 6: Voice Dropping Rate under {MPEG, Voice} priority

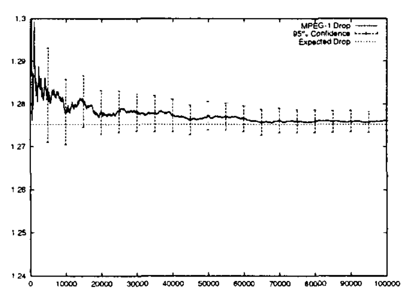

Figure 7: MPEG Dropping Rate under {MPEG, Voice} priority

6 Conclusion

Supporting QoS in a wireless network is a challenging task. In this paper, a single-cell system traffic of multiple delay classes is mathematically analyzed, and it is proved to be independent of the scheduling algorithm used, for all work-conserving earliest-due-date (WC-EDD) scheduling algorithms. The dropping requirements of all individual applications are shown to be guaranteed using deadline-sensitive ordered-head-of-line (DSO-HoL) priority schemes. Verification of the model is illustrated through extensive simulations. We are currently extending the model for mutiple cells where we can guarantee the call-level QoS using handoff dropping probability.

References

[1] J.G. Kim, I. Widjaja. Connection admission control for PRMA/DA wireless access protocol. *1997 IEEE Int. Perf., Comput. and Commun. Conf.,* 1997, pp.476-82.

[2] G.J. Santivanez, J.R. Boisson. D-RAMA: a new deterministic MAC protocol for wireless multimedia communications. *The 8th IEEE Int. Symp. PIMR*, 1997, pp.1043-8.

[3] J.M. Capone, I. Stavrakakis. Delivering QoS Requirements to Traffic with Diverse Delay Tolerances in a TDMA Environment. *IEEE Trans. on Net.,* Feb 1999.

[4] J.M. Capone, I. Stavrakakis. Achievable QoS in an Interference/Resource-Limited Shared Wireless Channel. To appear in IEEE JSAC.

[5] K.M. Tong. Achievable QoS for Multiple Delay Classes in TDMA Cellular Environments. *MPhil Thesis,* Computer Science Department, Hong Kong University of Science and Technology, August 1999.

[6] S.S. Panwar, D. Towsley, J.K. Wolf. Optimal scheduling policies for a class of queues with customer deadlines. *Journal of the ACM,* Oct. 1988, pp.832-44.

[7] A. Acampora, M. Naghshineh. Control and Quality-of-Service Provisioning in High-Speed Microcellular Networks. *IEEE Personal Commun.,* 1(2), 2nd Quarter 1994.

[8] P.R. Jelenkovic, A.A. Laznr, N. Semret. The effect of multiple time scales in MPEG video streams on queueing behavior. *IEEE JSAC,* Aug. 1997, pp.1052-71.

[9] P.T. Brandy. A Model for On-Off Speech Patterns in Two-way Conversation. *Bell Systems Technical Journal,* 48:885-890, Sep. 1969.

[10] Srivastava M, Mishra PP. On quality of service in mobile wireless networks. *IEEE 7th Int. Workshop on Network Support for Digital Audio and Video,* 1997, pp.147-58.

[11] M. Karol, Z. Liu, K. Eng, An Efficient Demand-Assignment Multiple Access Protocol for Wireless Packet Networks. *Wireless Networks,* 1(4):267-279. Dec 1995.