Multiple Access and Bandwidth Reservation in Wireless ATM Local Networks

By

Zhiguang Zhang

A dissertation submitted to the Graduate Faculty in Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

The City University of New York

1998

UMI Number: 9820598

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

This manuscript has been read and accepted for the Graduate Faculty in Engineering in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

11/24/97
Date

Chair of Examining Committee

11/24/97
Date

Executive Officer

Professor Joseph Barba

Professor Ibrahim Habib

Professor Myung Lee

Professor Constantine Manikopoulos

Supervisory Committee

The City University of New York

ABSTRACT

Multiple Access and Bandwidth Reservation in Wireless ATM Local Networks

by

Zhiguang Zhang

Advisors: Professor Ibrahim Habib, Professor Tarek Saadawi

In a wireless asynchronous transfer mode (*ATM)* network, an efficient multiple access control (*MAC*) protocol should include provisions for reserving resources in order to provide multiple quality of service for different types of traffic. In this dissertation a bandwidth reservation multiple access (*BRMA*) protocol is proposed and analyzed. On the uplink channel, mini slots are deterministically assigned to each user to guarantee contention-free channel reservation. Data slots are dynamically assigned according to the service ratio to guarantee collision-free data transmission. The service quota is determined during the call admission phase such that different users can receive their requested quality of service. Performance analysis results prove the effectiveness of *BRMA* in supporting multimedia traffic that have diverse mixture of statistical characteristics.

ACKNOWLEDGMENTS

Four years ago, my mentor handed me an introductory paper on wireless ATM. At that time I knew almost nothing about the subject. Today I am very happy that I have made my way here.

I would like first to express sincere thanks to my mentor Professor Tarek Saadawi, who has provided me with both academic advice and financial assistance throughout my years of studying as a graduate student.

Many thanks also go to my academic advisor Professor Ibrahim Habib, who devoted much of his time to discussing details of this research with me and helping me in many other ways. Without his help, this work would not have been possible.

I would like to also thank the Electrical Engineering faculty and the examination committee for their support.

Finally, I would like to thank my financial sponsors and friends — the Salzman family. They gave me support and love at the time when I most needed them.

This dissertation is dedicated to my loving parents.

Contents

List of Tables

List of Figures

1. Introduction

1.1 Wireless Information Networks

Most people will associate wireless information networks with cordless or cellular phones. In fact, the meaning of the wireless information networks goes beyond these. The existing wireless networks include following two broad categories: the voice oriented isochronous networks and the data-oriented asynchronous networks, see [1]. The former includes the popular cordless phones that cover local area with low transmission powers and the cellular phones that cover wide area with high transmission powers. The latter includes the high-speed wireless *LANs* (*WLAN*) and the low-speed mobile data networks that serve wide areas. Historically, the wireless information networks has gone through generations of revolutions and evolution. The service of first generation, characterized by the analog voice transmission and limited network control, is still operational and still the main wireless service available in this country. One example is the Advanced Mobile Phone Service (*AMPS*) that was designed by AT&T and mandated by *FCC*. The second generation wireless information networks, characterized by digital transmission technologies and providing service to wide areas, are now being implemented and becoming operational worldwide. One example is the European Global System for Mobile Communication (*GSM*). The third generation of wireless information networks, aiming to provide integrated services over cordless and cellular phones, *WLANs* and the Internet multimedia, to wireless users anywhere and at any time, is under research worldwide.

1.2 Bandwidth for Wireless Communications

The problem of insufficient commercial wireless bandwidth has existed for a long time. In 1983, the first cellular system became operational in Chicago. Some cells already became saturated in less than a year, see [1]. According to the Cellular Telecommunications Industry Association (*CTIA*), in 1984 there were about 90,000 cellular subscribers in the United States. In 1994 the number increased to 17 million. Recently, in order to increase the capacity of commercial wireless networks, the *FCC* relinquished and auctioned some wireless bandwidth that was designated for government users to commercial users. Such bandwidth includes the spectrum between 1850 - 1910 MHz and 1930 - 1990 MHz relocated to licensed *PCS* and 1920 - 1930 MHz to unlicensed *PCS* (*UPCS*), see Figure 1. The high frequency band to the right of the Figure 1 can of course be explored. However such exploration is limited by the technology available today. For now, the wireless bandwidth resource is like precious real estate.

The insufficient availability of the wireless bandwidth has imposed more and more stringent requirements to the efficient utilization of the precious resource. A tremendous work has been done in wireless system infrastructure, coding, modulation technology and efficient multiple access technology to squeeze the most out of the limited wireless bandwidth.

1.3 Wireless Multiple Access Control Protocols

The effectiveness of a wireless system depends largely on the multiple access control (*MAC*) protocol that is used. The *MAC* technology has evolved from frequency division multiple access (*FDMA*) to time division multiple access (*TDMA*) and code division multiple access (*CDMA*). As the earliest analog *MAC* technology, the *FDMA* is still in use in some current wireless networks, such as the cellular phone systems in the United States. It does not utilize the wireless bandwidth as efficiently as the digital *TDMA* or *CDMA*. *TDMA* is a mature and well understood *MAC* technology and has been widely used for many years in wireless networks as well as in the wired networks. A few types of *TDMA* protocols exist now. One is the random access *TDMA* with or without control, such as the *Aloha*, *Slotted Aloha* and *CSMA/CD*. It is usually used in data communications. Another is centrally controlled *TDMA* which is mostly used in the isochronous voice communications. Unfortunately, neither type above can utilize the bandwidth very efficiently for variable bit rate (*VBR*) traffic that is usually encountered in the future wireless information networks. The random access *TDMA* cannot even provide satisfactory utilization of bandwidth to data traffic. For example, the Slotted Aloha can typically reach a maximum utilization of 36% due to the packet collisions. The traditional central controlled *TDMA*, on the other hand, does not provide high bandwidth utilization for live voice communication either. In most *TDMA* protocols with central control, each channel (a data slot in a frame) has to be assigned to a voice conversation during the whole connection period. Because the voice activity detectors are difficult to implement with these protocols, about 60% of bandwidth — corresponding to the silence periods of the voice conversations — can only be used to transmit data packets or even wasted.

To efficiently utilize the bandwidth, a Slotted Aloha reservation type of *TDMA* called packet reservation multiple access (*PRMA*) was proposed by Goodman et al [2] and widely studied by many others, see [3], [4], [5], [6] and [7]. Unlike *Slotted Aloha*, contentions for a time slot in *PRMA* occur only at the beginning of each talkspurt of the conversation. Unlike traditional *TDMA*, *PRMA* allows a voice user to reserve a slot only during each talkspurt rather than during the whole conversation. For voice conversations where the average talkspurt is in the order of a second, the utilization of the channel can reach 68% [3], higher than that of traditional *TDMA* (40%). Mitrou et al in [8] proposed and studied an improved version of *PRMA*. In this improved version a minimum portion of the available channel capacity is dedicated to the reservation channel. Slotted Aloha contentions occur in some slots that are further divided into mini-slots. As a result, throughput performance under high load conditions is improved.

We observe, however, that as the average length of talkspurt of the conversation decreases, the performance of this class of *Slotted Aloha / TDMA* reservation protocols degrades due to more frequent collisions at the beginning of each talkspurt. In the extreme case, when the user transmits one packet each spurt, above protocols perform just like *Slotted Aloha*. Therefore *PRMA* described in [2] and [3], though good for periodic and slow time-varying traffic such as voice, is not suitable for the highly time-varying traffic such as those encountered in multimedia *ATM* networks.

2. Bandwidth Reservation Multiple Access Protocol

2.1 Architecture of Wireless Local Systems

In wireless *TDMA* schemes, time is usually divided into frames and slots. A guard time is needed between two consecutive slots or frames to accommodate the effect of the propagation delays, see [9] and [10]. Partially for this reason in a high speed wireless *LAN* where *TDMA* is used, the propagation delay should be very small if high utilization is to be achieved. In other words, most high speed wireless *LANs* are usually designed to cover small areas called micro-cells or even pico-cells. In fact, in some applications the confinement of the signal to a small area is a desirable feature for wireless *LANs*, see [11]. In *LANs* that cover such small areas, practical propagation delay is in the order of a few microseconds and has no effect on the protocol performance.

Our protocol *BRMA* is a wireless protocol for such purpose. In order to provide a coverage of a wider geographical area, a backbone network or metropolitan area network (*MAN*) can be used to interconnect the base stations of the smaller cells, see Figure 2.

2.2 A Bandwidth Reservation Protocol for Wireless ATM Networks

In this section we describe the *BRMA* protocol from three perspectives: call admission, signaling and transmission scheduling.

2.2.1 Call Admission Control

In an *ATM* network, a call admission controller determines whether the network has enough resource to support a new call with requested *QoS*. The *QoS* parameters include peak-to-peak cell delay variation, maximum cell transfer delay, cell loss ratio, cell error ratio, cell block ratio, and cell misinsertion rate, see [12]. These *QoS* commitments are probabilistic in nature and are intended to be only a first order approximation of the expected performance that the network needs to provide. Even so, to support these *QoS* needs careful calculations of required bandwidth. There are many methods to determine the required bandwidth to support calls with *QoS*, see [13] — [18]. For example, the required bandwidth is studied with queuing or simulation analysis in [13] and [14]. In [15] and [16], the "effective bandwidth" method was used to solve the bandwidth allocation problems for individual and multiple calls. Many of these methods are rather effective for the studied traffic patterns.

However, it is very difficult to characterize and model the highly variable and correlated properties of the multimedia traffic in the *ATM* environment with just a few parameters. It is even more difficult to determine the required bandwidth for such traffic dynamically without using over complicated module. In this study, we propose to use the feed-forward back-propagation neural networks (*NNs*) to perform the call admission control functionality. The neural network call admission controller resides at the base station. There are *R* virtual paths supported over the wireless uplink. Each virtual path carries traffic of different arrival characteristics but share the similar *QoS* requirements. One neural network is used to estimate the capacity allocated to each virtual path. A decision maker calculates the total allocated bandwidth and determines if a new call should be admitted. Figure 3 shows the functional structure of the call admission controller for two virtual paths.

Traffic characteristics are measured with an arrival count process in consecutive constant time periods. As indicated in [19], the traffic correlation of most time varying videos can usually be captured by a count process with measure window containing 600 samples measured at consecutive video frames. In this study of *BRMA*, the base station obtains each sample by summing the total number of requested data slots per frame for this virtual path in the time period of a video frame. This number of requested data slots per frame is reported from each mobile user to base station every time frame in a mini slot, which will be explained in the subsection 2.2.3 when the transmission scheduling is detailed. These samples are pushed from the count process to the neural network via a shift register. Using these samples as inputs, the neural network calculates the capacity used by this virtual path and reports to the decision maker. To avoid complicated design, we divide the neural network into two levels. The 600 samples are grouped into 30 groups with each group containing 20 samples. The first level of neural networks calculates the capacity for each group of 20 samples and forwards it to the second level. When results of 30 groups reach the second level, the total capacity for the virtual path is calculated.

The decision maker takes the calculated capacity from all virtual paths as inputs, considers the multiplexing gain at the virtual paths level, total capacity, and the capacity already in use in a virtual path to determine if the new call should be accepted. More exactly, a new call is accepted if:

$$C_{vpi} + X \geq C_{ai} + C_{new} \tag{1}$$

where C_{vpi} is the total available capacity for the virtual path i, C_{ai} is the capacity already used in the virtual path i, C_{new} is the calculated capacity for the new call, and X is the multiplexing gain defined by:

$$X = \frac{\sum_{i=1}^{i=R} C_{vpi} - C_{link}}{R} \tag{2}$$

where R is the total number of virtual paths in the network and C_{link} is the total capacity in the wireless link.

This feed-forward back-propagation *NN* call admission controller needs initial off-line training based on extensive simulation results. Once properly trained, it can adaptively and promptly determine the required capacity for a new call of the type that is carried over that virtual path. The time to determine the capacity needed for a new call of learned traffic pattern is minimal, while the time to learn a new traffic pattern is no more than a time period of 600 video frames. Most parts of this neural network call admission controller is similar to the one that is described in detail in [19].

2.2.2 Signaling

The uplink signaling is performed on a separate frequency band using Slotted Aloha scheme. The downlink signaling is performed on the broadcast channel. When a mobile station requests a new call admission, it sends a packet in one of the call admission request slots based on *Slotted Aloha* contention scheme. If this call admission request packet successfully reaches the base station, the *NN* call admission controller at base station determines whether this new call can be supported with described traffic pattern and requested *QoS* requirements. If the base station determines to accept this call, it sets up the end-to-end connection, reserves resources such as bandwidth and buffers along the way, and notifies the requesting mobile station through the downlink broadcast channel. The admitted call then starts its transmission according to the procedure described subsection 2.2.3 bellow. If the base station decides to reject the new call due to insufficient available bandwidth, it returns a busy signal to the requesting mobile station through the broadcast channel. If the mobile station does not receive any acknowledgment from the base station after a time-out period, it considers the *Slotted Aloha* transmission unsuccessful and will try again after a random waiting time period. When the mobile station decides to terminate the call, it sends a request to the base station. Upon successfully receiving the request, the base station notifies the other end user, releases the allocated resources and then, acknowledges the mobile station through the broadcast channel, see figure 4. In this example, mobile stations *A* and *B* are already connected before *MS C* requests admission. Synchronization in the micro cell is managed by the base station. The time clock is transmitted through the downlink channel periodically from the base station to the mobile station, which adjusts its clock accordingly.

2.2.3 Transmission Scheduling

We consider an *RF* channel of high capacity (10 Mbps) shared by mobile multimedia terminals within a micro cell. Communications between base station and mobile stations are carried over uplink and downlink channels. The downlink channel is divided into data slots, each of which carries an *ATM* cell.

Besides signaling, the broadcast downlink channel is used to transmit data cells and *OAM* cells to mobile stations. In the uplink channel, time is divided into data slots, each of which carries an *ATM* cell. Figure 5 shows the structures of the uplink time frame, mini slot, and data slots. Each frame is composed of K data slots and N mini slots, with N equals to the number of mobile stations that are accommodated in the system. Each mini slot is deterministically assigned to a mobile station that has succeeded in establishing a connection during the call setup phase which was described in subsection 2.2.1. In each mini slot, a mobile user sends a mini packet which includes u_1 bits for station ID, u_2 bits for number of data slots requested in current frame, and u_3 bits of error check to the base station. We have:

$$\begin{aligned} u_1 &= \lceil \ln(N) \rceil \\ u_2 &= \lceil \ln(K) \rceil \end{aligned} \tag{3}$$

and u_3 can be just one (1) bit as parity check for the simplest case. At the end of the mini slots, the base station knows the total number of data slots requested in this frame and broadcasts the data slots assignment immediately. Each user transmits its data packet(s) in its assigned data slot(s) in this frame. Obviously, the protocol works only in very small areas, such as micro cells or pico cells, where the propagation delay is negligible, see Figure 6.

In each frame, if the total requested data slots exceed what are available, each mobile user is assigned at least a minimum number of slots which has been previously determined during call admission. For any mobile user, if the number of assigned data slots are less than what has been requested, then some packets will be queued up for transmission in the following frames. Delay sensitive packets will be dropped if they have waited longer than their accepted delay limits. Because the call admission and bandwidth allocation functionality will limit the number of connections, the *QoS* is maintained for all connections. The above data slots assignment procedure is repeated every frame. One may observe that the slots assignment is dynamically varied every time frame according to the mobile user's traffic arrival pattern.

When the available bandwidth is more than the sum of the individual minimum bandwidth, a service ratio is determined to keep fair *QoS* to all mobile users. As an example, we consider two groups of wireless users *A* and *B*: and the number of users in each group is N_A and N_B, respectively. Group *A* and *B* have different *QoS* requirements. The call admission controller determines that the minimum number of data slots per frame (K_A and K_B) are guaranteed to groups *A* and *B*. We call K_A/K_B the service ratio. The call admission controller guarantees that

$$K_A + K_B \leq K. \tag{4}$$

In figure 7, we show a typical scenario of the slot assignment where K=7 and $K_A / K_B = 3/4$. In any particular frame, the data slots used by group *A* and *B* may not be exactly K_A and K_B respectively. When group *A* cannot use up the K_A data slots, and when group *B* needs more than K_B data slots, group *B* may utilize the unused portion of the K_A data slots . This is true vice versa. Similarly but not shown in Figure 7, any slots unused by group *A* and group *B* can also be reallocated to the best-effort service which was not preassigned any slots during call setup. Obviously, such service discipline maximally utilize statistical multiplexing gain to enhance the channel's throughput.

In the frame structure, the value of *K* is a design parameter. If *K* is too large (the frame is very long), new arrivals have to wait for a long time to request data slots for transmission, therefore packets suffer from long delay. If *K* is too small (corresponding to a very short frame), mini slots overhead occupies a large portion of the bandwidth hence packets also suffer from long access delays. Therefore an optimal value of *K* exists. This value depends on the traffic condition, the *QoS* requirement and the system capacity. In Chapter 3, we will determine it by queuing analysis and simulations for voice and data traffic. In Chapter 4 we determine optimal *K* for more complex traffic (video, voice) by extensive simulations.

Of course, the number of groups of traffic that *BRMA* can support is by no means limited to two. For simplicity, we only consider the case of two types of traffic in this and following chapters.

3. Bandwidth Reservation for Voice and Data

In this Chapter, the *BRMA* performance is studied for voice and data traffic through both queuing analysis and simulations.

3.1 BRMA Performance for Voice

We first consider the case when voice conversations are the only traffic in the system.

3.1.1 Queuing Analysis

We assume that a voice packet has v bits payload, and that each packet has a header of h bits. Each mini-slot contains h_m bits. Each voice source generates voice packets at the peak rate of R_p kbps during the talkspurts and no packets during silence periods. Both talkspurts and silence period, are exponentially distributed with mean of T_A ms and T_S ms respectively. During talkspurts voice packets arrive every t_p ms. There are N stations connected through the base station. If the voice packets are not transmitted within D ms, they will be dropped. A satisfactory *QoS* indicates a packet loss rate of under 1%. Intuitively very long frames will lead to excessive packet loss due to prolonged delays, thus limiting the number of stations that can be supported. On the other hand, very short frames may cause the total throughput to degrade due to excessive frame overhead. The optimal number of data slots in a frame is the one that can maximize the throughput and guarantee *QoS* for a certain number of users.

Designating d as the data slot length, L_f as the length of a frame, we have:

$$L_f = K \times d + \frac{N \times h_m}{C}$$
$$d = \frac{h + v}{C} \tag{5}$$

In order to simplify the analysis, we express the maximum delay limit in terms of a finite-size buffer. Considering that the buffer space is equivalent to a waiting time of one data slot d, and that each frame contains a mini-slot header of and that there are frames in time D, we obtain the equivalent buffer size limit B:

$$B = \left\lfloor \frac{(D - \frac{D}{L_f} \times (L_f - K \times d))}{d} \right\rfloor \tag{6}$$

We observe the arrival process from N homogeneous voice sources at the beginning of each frame. This process can be characterized by a discrete-time Markov chain of dimension (N+1), where its state is defined by the number of conversations in talkspurt (x). In many papers such as [20] and [21], in which similar analyses were performed, frame length was chosen to be the packet inter-arrival time, t_p. In our study, this case is not applicable since the frame length is, indeed, different from the packet inter-arrival time.

Let $P(x_1,x_2)$ be the probability that at the beginning of the $(n+1)$th frame there will be x_1 sources in talkspurt given that at the beginning of the nth frame there are x_2 sources in talkspurt in the system.

Let $P_{A/S}$ be the transitional probability that a source will switch from talkspurt phase to silence phase during a single frame, whereas $P_{S/A}$ is the transitional probability that a source switches from silence phase to talkspurt phase during the same time period. Noticing that both talkspurt and silence periods are exponentially distributed with means T_A and T_S respectively, we have:

$$\begin{aligned} P_{A/S} &= 1 - e^{-\frac{L_f}{T_A}} \\ P_{S/A} &= 1 - e^{-\frac{L_f}{T_S}} \end{aligned} \tag{7}$$

Now we try to obtain an expression for $P(x_1,x_2)$. Two cases arise:

- For $x_1 \geq x_2$, for the number of sources in talkspurt in the system to change from x_2 to x_1, there must be exactly k sources in talkspurt changing to silence and exactly (x_1-x_2+k) silent sources changing to talkspurt in one frame, where $k = 0,1,2,\ldots,\min(x_2, N - x_1)$. Hence we have:

$$P(x_1,x_2) = \sum_{k=0}^{\min(x_2,N-x_1)} \binom{x_2}{k} \times P_{A/S}^{k} \times (1 - P_{A/S}) \times \binom{N - x_2}{x_1 - x_2 + k} \times P_{S/A}^{x_1 - x_2 + k} \times (1 - P_{S/A})^{N - x_1 - k} \tag{8}$$

- Similarly for x1 < x2 we have:

$$P(x_1,x_2) = \sum_{k=0}^{\min(x_a,N-x_2)} \binom{N - x_2}{x_2 - x_1 + k} \times P_{A/S}^{x_2 - x_1 + k} \times (1 - P_{A/S})^{x_1 - k} \times \binom{N - x_2}{k} \times P_{S/A}^{k} \times (1 - P_{S/A})^{N - x_2 - k} \tag{9}$$

For the N-state discrete-time Markov arrival process, we designate $P(A|x)$ as the conditional probability that there are exactly A packet arrivals during a frame given that at the beginning of the frame there has been x voice sources in the talkspurt. During time L_f, a source generates either ρ packets with probability $(L_f - t_p \times \rho) / t_p$ or $\rho + 1$ packets with probability $1 - (L_f - t_p \times \rho) / t_p$ respectively, where

$$\rho = \left\lfloor \frac{L_f}{t_p} \right\rfloor \tag{10}$$

For x sources in talkspurt, we have:

$$P(A|x)=\begin{cases}\begin{pmatrix}x\\A-\rho x\end{pmatrix}\left(\frac{L_f-t_p\rho}{t_p}\right)^{A-\rho x}\left(1-\frac{L_f-t_p\rho}{t_p}\right)^{x-(A-\rho x)} & 0\le A-\rho x\le x\\ 0 & otherwise\end{cases} \tag{11}$$

When the frame length L_f is exact integer multiples of t_p, i.e. $\rho = L_f/t_p$, each source in talkspurt will generate ρ packets during the frame. Therefore $P(A|x)$ will exactly become:

$$P(A|x)=\begin{cases}1 & A=x\times\rho\\ 0 & otherwise\end{cases} \tag{12}$$

Now we consider the queuing model. As in [20] and [21], the system observed at time units of L_f, is characterized by a two-dimensional discrete-time Markov chain process (see Figure 8). The state of this chain is defined by (q, x), where q is the total number of packets in the queue at the beginning of a frame while x is the total number of voice sources that are in talkspurt at the beginning of the frame. It follows that the queuing process evolves from nth frame to $(n+1)th$ frame according to the following:

$$q_{n+1}=\min\left(\left(q_n-K\right)^+ + A_n, B\right) \tag{13}$$

where $(\bullet)^+$ denotes the larger of its content or 0, A_n is the number of packets arriving during the nth frame.

For the above two-dimensional discrete-time *Markov* chain, let $\alpha(q_1, x_1|q_2, x_2)$ be transitional probability that the next state will be (q_1, x_1) given that the present state is

(q_2, x_2).

For $B \le K$:

$$\alpha(q_1, x_1 | q_2, x_2) = \begin{cases} P(x_1, x_2) \times P(q_1 | x_2) & 0 \le q_1 < B \\ \sum_{A=B}^{\infty} P(x_1, x_2) \times P(A | x_2) & q_1 = B \end{cases} \tag{14}$$

For $B > K$:

if $0 \le q_1 < B - K$

$$\alpha(q_1, x_1 | q_2, x_2) = \begin{cases} P(x_1, x_2) \times P(q_1 | x_2) & 0 \le q_2 < K \\ P(x_1, x_2) \times P(q_1 + K - q_2 | x_2) & K \le q_2 \le q_1 + K \\ 0 & q_1 + K < q_2 \le B \end{cases} \tag{15}$$

if $B - K \le q_1 < B$:

$$\alpha(q_1, x_1 | q_2, x_2) = \begin{cases} P(x_1, x_2) \times P(q_1 | x_2) & 0 \le q_2 < K \\ P(x_1, x_2) \times P(q_1 + K - q_2 | x_2) & K \le q_2 \le B \end{cases} \tag{16}$$

if $q_1 = B$:

$$\alpha(q_1, x_1 | q_2, x_2) = \begin{cases} P(x_1, x_2) \times \sum_{A=B}^{\infty} P(A | x_2) & 0 \le q_2 < K \\ P(x_1, x_2) \times \sum_{A=B+K-q_2}^{\infty} P(A | x_2) & K \le q_2 \le B \end{cases} \tag{17}$$

The $\alpha(q_1, x_1 | q_2, x_2)$ in the above equations assumes that the system is memory-less. This is accurate only when $\rho = L_f/t_p$. However for $\rho < L_f/t_p$ those equations are only approximations.

The steady-state probability (q,x) satisfies following equations:

$$\pi(q,x) = \sum_{x_2=0}^{N}\sum_{q_2=0}^{B}\pi(q_2,x_2)\times\alpha(q,x|q_2,x_2) \tag{18}$$

where the boundary condition is:

$$\sum_{x=0}^{N}\sum_{q=0}^{B}\pi(q,x) = 1 \tag{19}$$

Above equations can be solved with linear algebra to obtain $\pi(q,x)$. To compute the probability of packet loss, we have:

for $B > K$:

$$P_{Loss} = \sum_{x=0}^{N}\sum_{q=0}^{K}\sum_{A=B+1}^{\infty}\pi(q,x)P(A|x)(A-B) + \sum_{x=0}^{N}\sum_{q=K+1}^{B}\sum_{A=B+k-q}^{\infty}\pi(q,x)P(A|x)(q-K+A-B)/\overline{A} \tag{20}$$

whereas for $B \leq K$:

$$P_{Loss} = \left(\sum_{x=0}^{N} \sum_{q=0}^{K} \sum_{A=B+1}^{\infty} \pi(q,x) P(A|x)(A-B) \right) / \overline{A} \tag{21}$$

where $\overline{A}$ is the average number of packet arrivals in a frame. It is given by:

$$\overline{A} = \frac{N \times T_A \times L_f}{(T_S + T_A) \times t_p} \tag{22}$$

It is simple to see that the average queue length $\overline{q}$ at the beginning of each frame is:

$$\overline{q} = \sum_{x=0}^{N} \sum_{q=0}^{q=B} \pi(q,x) \times q \tag{23}$$

To obtain the average packet delay, we have to compute the average queue length Q during the frame as follows.

For $K \le \overline{q} + 1$, the average queue lengths at 0th, 1st, 2nd, ..., (K-1)th data slots are: $\overline{q}$, $\overline{q}$-1 + $\overline{A}$, $\overline{q}$ - 2 + 2 × $\overline{A}/K$, ..., $\overline{q}$ - (K-1) + (K-1)× $\overline{A}/K$ respectively. Averaging above values we have:

$$\begin{aligned} Q &= \frac{\left(\overline{q} + \overline{q} - 1 + \frac{\overline{A}}{K} + \overline{q} - 2 + 2 \times \frac{\overline{A}}{K} + \ldots + \overline{q} - (K-1) + (K-1) \times \frac{\overline{A}}{K} \right)}{K} \\ &= \frac{2\overline{q} + \frac{K-1}{K} \times \overline{A} - K + 1}{2K} \end{aligned} \tag{24}$$

Fore K > $\bar{q}$ + 1:

$$Q = \frac{\left(2\bar{q} - \lfloor\bar{q}\rfloor + \frac{\lfloor\bar{q}\rfloor \times \bar{A}}{K}\right)\left(\lfloor\bar{q}\rfloor + 1\right) + \left(\frac{\lfloor\bar{q}\rfloor \times \bar{A}}{K} + \bar{A}\right)\left(K - \lfloor\bar{q}\rfloor - 1\right)}{2K} \tag{25}$$

where $\lfloor\bullet\rfloor$ is the integer part of $\bullet$.

Noticing that only $\left(1 - P_{Loss}\right) \times \bar{A}$ packets actually contribute to queuing delay per frame, we apply *Little's formula* [22] to obtain the average voice packet waiting time, we have:

$$W_v = \frac{Q \times L_f}{\left(1 - P_{Loss}\right) \times \bar{A}} - d \tag{26}$$

3.1.2 Numeric Results and Discussions

The system has been simulated for approximately fifteen minutes of real time operation during which at least ten million packets were generated.

In our study of performance analyses and simulations when there only voice traffic, we try to compare our protocol with *PRMA*. Unless otherwise stated, the variables and their default values are given as listed in Table 1. Each data slot header is 2 bytes. In the following, we use the same values of parameters as those that were used in the *PRMA* performance study in [3] unless otherwise stated.

To assess the validity of the approximations, we compare, in Figure 9, the results of simulations and numerical analysis. The cell loss rate is compared among three cases. The first curve is obtained from the simulation in which packets are dropped whenever delay exceeds 32 ms. The second curve is the result of the simulation in which packets are dropped whenever the buffer (of length 40, according to equation (6)) is full. The last curve is the result of the numerical analysis. All three curves agree very closely and indicate that 39 stations (or 76% channel utilization) can be supported with cell loss rate less than the 1%. The frame length here is 16 ms.

Figure 10 compares the packet loss rate performance when all parameters are default values as given in Table 1. The results of simulation and analysis indicate that when the K is 29, the protocol reaches its best performance and 41 conversations can be supported. Comparatively *PRMA* can support 36 conversations, see [3].

Figure 11 shows the simulations result of the delay performance for different frame length. For the similar reasons mentioned above, too short or too long frame would lead to high packet delays. The optimal K is around 11.

The protocol performance for voice sources of different burstiness is shown in Figure 12. Simulations shows that when the average talkspurt time and average silence time are both one fourth of the default values, the packet loss performance of BRMA in a 10 Mbps system is only slightly different a system with the same *QoS* requirement. This indicates that the protocol yields high utilization and is insensitive to the burstiness of each traffic source, especially for high speed systems. For *PRMA*, however, when the average talkspurt time and average silence time decrease, more collision happens and cell loss rate increases hence lower utilization will result.

Figure 13 compares the cell loss rate between *PRMA* and *BRMA*. *PRMA* can at best reach a utilization of 64% for a 720 Kbps system and given *QoS*. *BRMA* can reach a utilization of 76% for the same system and QoS. When the system capacity is as high as 10 Mbps, *BRMA* can reach a utilization of 95% for the same *QoS*.

The system performance for *ATM*-size voice packets is similar, as demonstrated in Figure 14. When each voice source applies the *ATM* voice packet encoder, i.e., when t_p is 12 ms instead of 16 ms, and each packet payload is 48 bytes instead of 64 bytes, the protocol can reach a utilization of 93% in a 10 Mbps system. This slight decrease in utilization is due to the fact that for the same two bytes packet header less payload traffic is transmitted in each packet. There has been discussions on whether *ATM* technology should be directly introduced into the wireless channel — the result of lower channel efficiency is one major counter argument. Of course, the *ATM* technology will do the wireless network more good than harm. After all, the *QoS* feature of *ATM* is the ultimate solution to the transmission of multimedia traffic.

3.2 BRMA Performance for the superposition of Voice and Data

Now we study the performance of *BRMA* for superposition of both data and voice traffic.

We assume that data packets are allowed to transmit only after all voice packets are transmitted. Because the voice packets have higher priority than that of the data packets, their queuing performance is independent of the data traffic. Data packets are not subject to packet loss. They are simply delayed for a longer time when bandwidth is not currently available. Therefore we can use the method introduced previously to calculate the voice packets loss rate and average voice packet delay without concerning the data traffic at all. Assuming that the packet loss rate for the voice packet is negligibly low (much less than 1%), we can use an infinite buffer model to analyze average packet delay of the superposition of voice and data traffic. Finally the average delay of the data packets can be calculated. This independent treatment in the protocol performance analysis for the voice traffic is quite like the approach used in [23] by Wieselthier et al.

3.2.1 Queuing Analysis

The previous approach that uses N-state discrete-time *Markov* chain to model the arrival process is entirely feasible when the system buffer size B is not very large. When an infinite (or very large) buffer system is considered, the approach will become infeasible due to the excessive requirement of the computer memory that is needed for the calculation. In order to facilitate our calculation, we use a two-state discrete-time *Markov* Modulated *Poisson* Process (*MMPP*) to model the superposition of voice and data arrivals.

In [24] and [25] the *Markov* Modulated *Poisson* Process was used to model superposition of voice packets arrivals, by matching traffic characteristics such as the mean packet arrival rate, the variance-to-mean ratio of the number of arrivals, the long term variance-to-mean ratio of the number of arrivals, and the third moment of the number of arrivals. Data traffic can easily be incorporated into the *MMPP* model, because the data arrival process is also *Poisson*.

We consider that N - 1 homogeneous stations generate voice traffic as before, and one station generates data packet one at a time according to a *Poisson* process and each packet has the same size as the voice packets. This consideration that all data traffic come from one source does not sacrifice generality, because the superposition of multiple *Poisson* is still *Poisson*.

We characterize the superposition arrival process by a simple two-state *MMPP* model [24]. The *MMPP* model alternates between a High state (*H*) and a Low state (*L*). Let the mean sojourn time of each state be t_H and t_L respectively. Packets arrive during each state according to *Poisson* process with mean rate of λ_H and λ_L respectively. Notice that the duration of a typical frame is in the order of tens of milliseconds, whereas the sojourn times of the *MMPP* states are in the order of 0.5 seconds, see [24]. Hence it is reasonable to assure that during a typical frame, the *MMPP* model is stationary (i.e. does not alternate between the two states). Let the transitional probabilities from High to Low and from Low to High states be P_H and P_L respectively. It follows that

$$P_H = 1 - e^{-\frac{L_f}{t_H}}$$
$$P_L = 1 - e^{-\frac{L_f}{t_L}} \tag{27}$$

and *P(A/x)*, the probability that there are *A* arrivals in a frame given that the arrival state is *x* (*x=H, L*) is:

$$P(A|H) = \frac{e^{-\lambda_H \times L_f} \times (\lambda_H \times L_f)^A}{A!}$$
$$P(A|L) = \frac{e^{-\lambda_L \times L_f} \times (\lambda_L \times L_f)^A}{A!} \tag{28}$$

where λ_H and λ_L are the arrival rates at states of the *Markov* Chain.

Recall that P_{Loss} is evaluated using formula (20) and (21). This probability is not affected by the data traffic, as indicated in [26]. Lost packets should be subtracted from our *MMPP* arrival model because in this *MMPP* queuing analysis no packet loss is considered due to the infinite buffer size. The arrival rates of the states of the final *MMPP* arrival model are then:

$$\begin{aligned} \lambda_H &= (1 - P_{Loss})\lambda_{vH} + \lambda_d \\ \lambda_L &= (1 - P_{Loss})\lambda_{vL} + \lambda_d \end{aligned} \tag{29}$$

where the λ_{vH} and λ_{vL} are the arrival rates of the *Markov* model which does not include data traffic.

Now we consider the arising two-dimensional discrete-time queuing process at each frame length L_f. Let the state of the arising *Markov* chain be (q,x), where q is the number of packets in queue and x is the number of sources in talkspurt observed at the beginning of each frame. Let (q,x) be the steady state probability. We notice that this *Markov* chain is similar to the one solved in the previous section. However, there are two differences:

1. the buffer size is infinite. This can be approximated by a finite buffer with very large size B.

2. the arrival process has only two states.

While the $\alpha(q_1,x_1 \mid q_2,x_2)$ expressions remain the same as those presented earlier, $P(x_1,x_2)$ and $P(A/x)$ equations are now given by (27) and (28) instead of equations (7), (8), (9), (11) and (12).

Solving this queuing model as we did previously, we obtain $\pi(q,x)$.

Recall that the average number of packet arrivals in each frame for voice packets is evaluated by (22), now we have:

$$\overline{A} = \frac{(N-1) \times T_A \times L_f}{(T_S + T_A) \times t_p} + \lambda_d \times L_f \tag{30}$$

where, we have added the term to accommodate the data packets.

Then we can find the average packet delay by applying *Little's formula*:

$$W = \frac{Q \times L_f}{\overline{A}} - d \tag{31}$$

where Q is evaluated from equations (24) and (25).

Using equation (31) to evaluate W and using equation (26) to evaluate W_v, we can then obtain the average delay for data packets W_d as follows:

$$W_d = \frac{W - W_v \times (1 - R_d)}{R_d} - d \tag{32}$$

where R_d is percentage of the data traffic in the combined traffic:

$$R_d = \frac{\lambda_d \times (t_H + t_L)}{\lambda_H \times t_H + \lambda_L \times t_L} \tag{33}$$

3.2.2 Numeric Results and Discussions

In the numeric study, a large buffer (B = 600) is used to approximate the infinite buffer in our *MMPP* queuing analysis.

Figure 15 shows the packet delay performance of simulation results using the *MMPP* model in a 720 Kbps system when both voice and data traffic are present. Frame length equals 16 ms here. We note that *PRMA* behaves like a *Slotted Aloha* when the data traffic is generated according to Poisson process and each message is just one packet. The voice traffic is 45% of the capacity. Under these conditions, when the throughput is 36% of the channel capacity, *PRMA* will produce infinite delay. On the other hand, Figure 15 demonstrates that the delay for our protocol is only 45 ms. In Figure 16, the results of packet delay for the *MMPP* analysis and simulations are presented for voice and data traffic when the frame length is optimized (K = 29). Compared to the delay performance when frame length $L_f = t_p$ as indicated in Figure 15, the protocol with optimal frame length reduces the mean data packet delay remarkably, especially when the throughput is high.

Finally, Figure 17 shows the delay performance for data and voice packets for different values of K. The error of numerical analysis is reasonable.

In the above study we determine that *BRMA* can gracefully support voice and data traffic. We also find that *BRMA* is not very sensitive to the burstiness of each source, especially in high speed high load systems. Hence it has potential to support multimedia *ATM* traffic. We demonstrate that in spite of the frame overhead, *BRMA* achieves better performance than *PRMA* in terms of less packet delay and loss. Such improvement is due to the deterministic assignment of the request channels and due to the multiplexing gain in the high speed systems. Therefore the request channel is contention-free and the data channel is collision-free. A high multiplexing gain is achievable in the high speed systems.

Although we do not provide the numerical comparison between the conventional *TDMA* and *BRMA* here, the difference between these two is very clear, especially for multimedia traffic. Taking the traffic of voice conversation for example, conventional *TDMA* does utilize the voice activity detectors thus wastes over half of the total bandwidth. In other words, for high speed systems, *BRMA* can double the channel utilization for voice traffic as compared to conventional *TDMA*.

4. Bandwidth Reservation for Multimedia Traffic

As mentioned earlier, *BRMA* was proposed to support multimedia traffic for high speed high load wireless local networks where the propagation delays are very small. We study its performance for multimedia traffic in this chapter.

4.1 BRMA Supporting Two Classes of Traffic

Different class of traffic in the wireless *LAN* may demand different class of Quality of Service (*QoS*). In this chapter the study we will only consider two classes (or groups) of traffic, class *A* and class *B*. The multi-classes traffic case is just a natural extension of the concept.

We assume that the *QoS* is defined as follows. For class *A* traffic, a packet will be considered over-delayed and dropped if it is not transmitted after a certain time D_A. The *QoS* is satisfied if the packet loss rate is under R_A. Similar *QoS* is defined for class *B* traffic.

We further assume that the number of wireless users in group *A* and group *B* are N_A and N_B respectively. For *BRMA* to guarantee the required *QoS*, a minimum bandwidth in terms of slots per frame must be guaranteed to each group. We assume that in each frame K_A and K_B data slots are guaranteed to group *A* and *B* respectively. Apparently we have:

$$\begin{aligned} N &= N_A + N_B \\ K &= K_A + K_B \end{aligned} \tag{34}$$

The ratio $K_A : K_B$ is called service ratio here.

We need to determine the appropriate service ratio, the optimal frame length so that different *QoS* can be guaranteed to traffic group *A* and *B*. We need to study also the system performance for various kinds of traffic in this part of the study.

4.2 BRMA Supporting High Activity Video and Voice

High activity *VBR* video traffic is expected to be among the major traffic components in the wireless *ATM* networks. Due to its high burstiness and nonstationary property, *VBR* video traffic introduces extra difficulty in efficient channel access as well as effective bandwidth allocation. In this section, we investigate if *BRMA* can efficiently handle such traffic. Since so far there are no simple mathematical models that can accurately capture the behavior of the real time video traffic, we choose to use simulations as tools of performance evaluation.

Currently many multimedia services in which stored media objects can be retrieved on demand by end users adopt the popular international encoding standard Motion Picture Experts Group (*MPEG-I*) [27]. In *MPEG*-I, video is coded into a sequence of Intracoded frame (*I*), Predictive frame (*P*) and Bidirectional frame (*B*). We use the *MPEG*-I coded movie *Star Wars* as an example of high activity video traffic, which is obtained from the public domain of Bellcore. The movie generates a frame sequence of *I B B P B B P B B P B B*. there are frames every second, see [28] for detailed description of this set of data. In our study we use a 6-minute segment of the Star Wars movie to be a sample. Some statistics of this sample are shown in Table 2. We shift the sample by some frames to obtain additional video sources so that these sources are statistically similar. In each simulation, video and voice terminals are connected to the *BS* via a common wireless channel. Packets are generated by each terminal independently. Each packet is 53 bytes, including a 5-byte header. Each voice terminal generates packets in the way that was described in previous section except that the voice packet size is shorter (53 bytes instead of 72 bytes, accordingly packets comes more often). Video packets from each video terminal are generated uniformly in each video frame interval of 1/24 seconds. Each simulation runs 6 minutes in real time.

Video packet loss happens when a packet is delayed D_m or longer, while voice packet loss happens when it is delayed D_v or longer. Packet loss rates are calculated based on measurements of lost packets in simulations. The packet loss performance of *BRMA* in a 10 Mbps system is shown in Figure 18 where $D_m = 20$ ms and $D_v = 32$ ms are chosen. The normalized utilization is 69% and the service ratio K_{vidio}/K_{voice} is 6/1. When the frame length is around 150 slots (which is the optimal frame length because packet loss rates are minimized here), 7 video terminals and 260 voice terminals can be supported such that the packet loss rate for video is under 10^{-5} and for voice is under 10^{-2}, respectively. In traditional wireless *TDMA*, with which bandwidth is dedicated to each user based on peak rate, a 10 Mbps system can only support

$$N = \frac{10\,Mbps}{24x161726} \approx 2.6 \tag{35}$$

such high activity video connections with no voice connections, corresponds to a channel utilization of 10%. Hence a significant multiplexing gain has been achieved with *BRMA*.

Now we study the impact of the service ratio on the performance of the protocol. The service ratio here sets a flexible boundary between the resources allocated to video and voice traffic. Note that we do not consider allocating resources to "best-effort" services such as data traffic. Our primary interest is real-time services demanding *QoS* guarantees. In these simulations the optimal frame length is found to be 150 slots/frame (see Figure 18). As shown in Figure 19, for 7 video terminals and 310 voice terminals, a channel capacity of 10 Mbps can guarantee the above *QoS* when the service ratio is between 7.2/1 to 10/1. After conducting extensive simulations using various values for the service ratio, we have found that only the range reported in the figure provides acceptable performance in terms of packet loss rate. Note that the packet loss rate of the voice traffic is not as sensitive to changes in the service ratio as the video traffic does, despite the fact that it constitutes 45% of the total channel capacity. Although video traffic consists of only 31% of the available channel capacity, its packet loss rate is significantly affected by the service ratio. This is due to that fact that the superposition traffic from 310 voice terminals is quite smooth while the superposition traffic from 7 video terminals is rather bursty. Figure 20 shows the limits of the number of users that can be supported over a 10 Mbps channel, under various *QoS* requirements. Here, a K_{video}/K_{voice} service ratio of 6/1, and 150 slots/frame, are used to ensure packet loss rate of under 10^{-2} for voice, and under 10^{-5} for video. Obviously, as the packet delay limit is decreased, fewer users can be supported with the channel. Clearly, increasing the delay limits implies longer buffers and, hence, more effective statistical multiplexing gain.

4.3 BRMA Supporting Low Activity Video and Voice

Low activity video traffic such as video conferencing should have less stringent bandwidth requirement than its high activity counterpart. It usually produces less bit rates as well as fewer abrupt changes. Therefore for the same *QoS* requirement a higher utilization should be achievable. We use a 6-minute segment of an *MPEG* coded video of a "*lecture*" video sequence as a sample of low activity video in this study. Some statistics of the sample are also shown in Table 2. It is obvious that the sequence has much lower activity compared to the *Star Wars* movie. The simulation environment is very much similar to the one described in the previous section.

Figure 21 shows the achieved utilization for various *QoS* requirements. When the video traffic dominates the traffic load, for D_{voice} < 64 ms and *D*video < 40 ms, 25 voice users and 20 video users can be supported. The corresponding utilization is 63% of the channel capacity. Comparatively for the case of high activity video as shown in Figure 20, under the same *QoS* requirement, 25 voice users and 10 video users can be supported, corresponding to 43% utilization. Obviously the high activity video demands more bandwidth for the same *QoS*, and the same average arrival bit rate. One can conjecture that supporting high activity video would require "smoothing" its high burstiness using a low pass filter. This will eliminate the need to allocate bandwidth for high frequency component of the traffic, hence the channel utilization can be greatly enhanced.

4.4 BRMA Supporting High Activity Video and Ethernet Traffic

LAN Ethernet traffic is usually bursty. As indicated in [29], it exhibits self-similar properties and does not have the "*Poisson*-like" nature. In other words, *LAN* traffic does not get "smoother" when more sources are multiplexed. In fact, the burstiness of the traffic gets worse as the traffic load increases. For such unfavorable traffic pattern, can satisfactory multiplexing gain be achieved with *BRMA*? In this section we intend to answer this question. The *Ethernet* traffic trace that we use here was recorded at Bellcore on August 29, 1989 and made available to the public via *FTP*. We used two six-minute segments of it as two sample traces. We assumed that each trace was generated by 130 *Ethernet* users. The superposition of the two traces forms the double-load trace, which is then assumed to be generated 260 *Ethernet* user terminals.

We now consider a 12 Mbps system in which 8 high activity video terminals and 260 *Ethernet* terminals are connected. Figure 22 shows the performance of the *BRMA* for the superposition traffic. Video packets from 8 video terminals are generated as previously explained and the *Ethernet LAN* data packets are generated according to the double load trace described above. The service ratio is $K_{video}/K_{LAN} = 5/4$, and packet delay limits are $D_{video} < 20$ ms and $D_{LAN} < 50$ ms. Packet loss occurs when a packet is delayed beyond the limits. When the frame length is above 60 slots, minimum packet loss rate is achieved for *LAN* traffic. The packets loss rates for video, and *LAN* traffic, are approximately 7.0×10^{-6}, and 1.2×10^{-4}, respectively.

Also shown in figure 22, are the results of similar simulations where both the traffic load and the channel capacity are reduced by half. Now when the frame length is about 35 slots/frame, the minimum packet loss rates are 1.6×10^{-4} and 1.1×10^{-3} for video, and *LAN* traffic, respectively. As shown in the figure, the packet loss rates for both video and *LAN* traffic are reduced by about one order when both the channel capacity, and the traffic load, are doubled. We can, equivalently, conclude that for the same *QoS* requirement, the high speed system can achieve a higher channel utilization. This demonstrates that even for the self-similar *Ethernet LAN* traffic, good multiplexing gains are achievable: the protocol works better for a high speed, high load system.

5. Dynamic Bandwidth Partitioning for *VBR* Video Traffic

In this chapter, we study the bandwidth partitioning and a variation of *BRMA* for *VBR* video traffic that guarantees loss-less service.

The queuing behavior (such as maximum queue length, average queue length and channel utilization) of the system with finite buffer sizes will be studied. Different from the previous chapters, packet losses will not be permitted. Instead, different bandwidth allocations will be made to guarantee such loss-less service with short delay (finite buffers).

5.1 Introduction To Dynamic Bandwidth Partitioning

In [26] and [30], it was established that the link capacity requirement at each node is essentially captured by its low frequency input traffic filtered at a properly selected cutoff frequency.

For *VBR* video traffic, the low frequency component is determined by the long term correlation caused by scene changes, while the high frequency components is determined by the short term correlation. Since the time between scene changes is usually about 1 second, the link capacity can be updated every 1 second accordingly. Then the short term correlation can be smoothed out with finite buffers. Long term correlation, however, cannot be smoothed out with filters. Instead, dynamic bandwidth allocation or partitioning is required.

Dynamic bandwidth allocation can adjust the bandwidth allocated to one or more users. When total traffic increase, higher layer protocol should increase the total bandwidth allocated to the link, so that buffer overflow does not happen. While this technique is totally suitable to the wired link, it may not be applicable in the wireless channel where the total capacity is fixed.

Dynamic bandwidth partitioning, however, is applicable to both wired and wireless channels. The goal is to dynamically adjust the bandwidth allocation to each user while not to change the total channel capacity to achieve the similar results. In [26] and [30], it was proposed to partition the bandwidth dynamically using following method. In a system of capacity C, N users share the bandwidth according to the following equation:

$$Ci = \frac{T_i \times C}{\sum_{i=1}^{N} T_i} \tag{36}$$

where T_i is the prediction of the filtered bandwidth for user i. The purpose of this filtering is to get rid of the short term traffic correlation. The controller updates this bandwidth periodically so that the scene change of the video traffic is captured. The results of this method were compared with the ones of the protocols in which bandwidth partitioning were fixed, based on mean or peak traffic load respectively. We simulated the protocol for a system of 70% channel utilization. The results agreed with those provided in [26]. Further more we simulated the protocol performance for a system of 90% utilization. The results are shown in Figure 23 to 26. The improvement of the dynamic partitioning over the fixed partition was obvious: shorter buffers were needed and fairer service quality (in terms of maximum queue lengths and mean queue lengths) was achieved.

Such dynamic bandwidth partitioning can improve performance. However, since the bandwidth assigned to one connection cannot be shared by another before next bandwidth adjustment, at high channel utilization, the performance needs further improvement. We will show next that with *BRMA*, the protocol performance can be further improved due to the bandwidth sharing at frame level. In the following, we will should how the dynamic bandwidth partition concept can be applied with *BRMA*.

5.2 BRMA Bandwidth Partitioning For JPEG Video

Applying the dynamic bandwidth partitioning principle to *BRMA*, we consider following case.

Five users, each is transmitting a *JPEG* traffic, are sharing a wireless uplink channel of fixed capacity *C*. *BRMA* protocol is used to control multiple access. Each user is preallocated a certain amount of bandwidth. In other words, each user is given a "quota" in each frame. Within its quota, a user's request for bandwidth in each frame will always be granted by the base station. For those users needing more bandwidth than their quota, they can fairly share the quotas that are not used by other users. For heavy traffic load, no bandwidth will be wasted if one or more user have packets waiting in the queues. This is just basic *BRMA* principle.

The dynamic bandwidth partitioning principle comes in here. The quotas pre-allocated to each of the five users are not fixed for the whole connection. Instead, they are dynamically adjusted based on the traffic load. The quotas are adjusted about every half a second - a time period that is, short enough to catch the low frequency variation of the traffic load (therefore to catch the scene changes of the *JPEG* video), but long enough not to unduly burden the bandwidth partition calculation and processing.

Just as described in [26] and [30], the *JPEG* traffic were taken from the page 39, 40, 41, 55 and 56 of the movie *Star Wars* obtain from Bellcore's public domain. The video traces are filtered with cutoff frequency of 2π and sampled at time unit Δ, which is 0.14 seconds (100 slices). Four Δs of video ahead are predicted based on previous five samples. The prediction and bandwidth adjustment is supposed to be done within one sample unit time. The adaptive predictor uses a *RLS*-based algorithm that has the adaptive capability of tracking the statistical variation of the nonstationary *VBR* traffic. See [26] for detail descriptions.

We simulated the *BRMA* performance with above described dynamic bandwidth prediction and partitioning for system of 70% and 90% utilization. The results are also included in Figure 23 to 26. Simulation results of *BRMA* with fixed bandwidth partitioning based on mean and peak traffic are also given in the same figures for comparison.

In Figures 23 and 24 it is clear that at low utilization (70%) the system performance of *BRMA* in terms of fairness and maximum buffer size for the dynamic bandwidth partitioning is only slightly improved over the those of the *BRMA* with static bandwidth partitioning based on mean and peak traffic. This is because at light load, the basic *BRMA* already provides fairness control and reduces the required maximum buffer for lossless traffic due to the bandwidth sharing at frame level. The complexity of the dynamic bandwidth partitioning is therefore not quite justified.

At high load system, however, performance improvement is significant, as demonstrated in Figures 25 and 26. This is because when traffic and queues build up, the dynamic adjustment of the service quota provided in the dynamic bandwidth partitioning will have stronger effect. In this case, the static bandwidth partitioning scheme will not provide enough improvement.

To compare the results in Figures 23 and 24 to those in Figures 25 and 26, we find that at light load (70% utilization) the dynamic partitioning scheme introduced in [26] is actually better than the dynamic partitioning scheme with *BRMA*. Again this is not surprising: the overhead (mini slots) introduced in *BRMA* is not justified at light traffic load. At high traffic load (90% utilization), however, the frame level bandwidth sharing capability of *BRMA* far outweighs its burden of mini slots overhead. The *BRMA* with dynamic bandwidth partitioning provides reasonable improvement, compared to the dynamic bandwidth partitioning scheme studied in [26].

5.3 BRMA Bandwidth Partitioning For MPEG Video

The *MPEG* video, as discussed in Chapter 4, uses typically periodic traffic pattern *I B B P B B P B B P B B*. In most cases, *I* frames are much larger in sizes compared to *B* or *P* frames. *I* frames appears periodically every twelve frames. One period of the twelve frames is call a Group Of Picture (*GOP*). Due to the periodicity of the *I* frames, the power spectrum of the traffic spreads far into the high frequencies. The energy spreaded in the high frequencies is not as insignificant as in the case of *JPEG* traffic. Therefore the low pass filter applied in the *JPEG* case will not be able to filter out the short term dependency of the traffic without lossing too much information. In short, the technique used in last section is less efficient for *MPEG* traffic than *JPEG* traffic.

To accommodate the periodic characteristics of the *MPEG* video, we use a *GOP* as a bandwidth update period. Typically, as mentioned before, scene changes happens in a time of the order of a second. Considering a video frame is 1/24 seconds for *MPEG*, a *GOP* will be half a second. Therefore updating the bandwidth every *GOP* will be good enough to catch the changes of the video traffic flow due to the scene changes, yet does not unduly burden the signal processing of the protocol.

To deal with the very bursty *MPEG* traffic, we dynamically partition the bandwidth based on not only the predicted new arrivals, but also the residual queue length. Assuming again N users are sharing the wireless channel of capacity of C, at the bandwidth updating moment, user No. i has a queue length of q_i and a predicted arrival traffic T_i, then the bandwidth allocated to this user in next time period is described as:

$$C_i = \frac{(T_i + q_i) \times C}{\sum_{i=1}^{N} (T_i + q_i)} \tag{37}$$

We consider two cases for traffic arrivals. Firstly, we use the predictions of the summation of the twelve *MPEG* video frames as the traffic arrival parameter T_i.

Secondly, we use only the predictions of the *I* frames as the traffic arrival parameter T_i.

Simulations have been done for these two dynamic bandwidth partitioning with *BRMA* for five *MPEG* sequences taken from the *MPEG* movie *Star Wars* obtained from Bellcore's public domain. We assume that the video packets arrive uniformly in each video frame. As comparison, we also simulated the *BRMA* with fixed bandwidth partitioning based on mean and peak traffic arrivals. Results are shown in Figure 27 to Figure 30.

As we can see from the figures, dynamic bandwidth partitioning improves the *BRMA* performance in terms of fairness and maximum buffers sizes required for lossless traffic, especially at high traffic load. Also, the results of the dynamic bandwidth partitioning based on *I* frames are similar to those of the dynamic bandwidth partitioning based on sum a *GOP*. This is because that the maximum queue lengths are usually caused by the *I* frames which are much larger in size than *P* or *B* frames. Therefore the *I* frames roughly reflect the traffic load of the video over time.

So far we chose to update the bandwidth every 12 video frame (one *GOP*). If we need to further reduce the processing burden of the bandwidth calculation, we can also choose to update the bandwidth in a larger window time. Of course this is a trade off: larger window time will naturally reduce the effect of dynamic bandwidth partitioning because it is getting less "dynamic". In Figure 31 and Figure 32 we show the results of our simulations of *BRMA* with dynamic bandwidth partitioning at different bandwidth update window in multiple *GOP*s, at 90% utilization. We show two cases: use the prediction of the mean of *I* frame size or the prediction of the sum of *GOP* size, as the T_i in equation (37). Figure 31 shows the standard deviation of mean queue size (an index showing the fairness of service) of the five users for different bandwidth update window. As we indicated earlier, the service fairness will asymptotically decrease as the bandwidth update window gets longer. We also notice the dynamic bandwidth partitioning based on sum of *GOP* will be slightly more effective than that based on *I* frames along. After all, the sum of *GOP* is a better indication of traffic flow than the *I* frames along. In Figure 32 we show the maximum buffer size that is needed to guarantee the lossless service. We can see that when the bandwidth update window is more than 20 *GOPs*, the required buffer size to guarantee lossless service will approach that the static bandwidth partition based on mean traffic load (close to 8000 packets in buffer, as indicated in Figure 29).

6. An adaptive version of BRMA

6.1 The advantage of variable frame length

Many wireless *MAC* protocols share a common technique: fixed frame length. This technique has an important advantage: it simplifies the protocol. However, a frame length that is changing with the traffic variation will improve system performance. In [31], D. G. Jeong et al studied a reservation-based multiple access scheme with variable frame length (*RMAV*) for wireless data communications. Aiming at short delay under light load condition, and high throughput under heavy load condition, *RMAV* allows its frame to vary dynamically every frame. Frame length increases (decreases) by one slot in each frame if the traffic load increases (decreases). While this protocol does enhance the performance compared to fixed frame length protocols, we notice that the speed at which the frame length is adapted to the traffic is rather slow, since the frame length can be varied by only one slot each frame. Nevertheless, *RMAV* provides some insights into the advantages of variable frame length protocols.

As we indicated in our study on *BRMA*, the optimum frame length largely depends on the traffic load as well as the *QoS* requirement. Generally, the optimum frame length is small when the traffic load is light and is large when the traffic load is high. However, the value of this optimal frame length cannot be explicitly determined if the traffic pattern is not explicitly known in advance.

6.2 Adaptive BRMA

In order to enhance the protocol performance as well as to bypass the difficulty of determining the optimal frame length, we consider a version of *BRMA* in which the frame length is adaptive with respect to the traffic load. In this adaptive version of *BRMA*, the base station includes the number of slots that this frame will contain when the data slot assignment is broadcasted every frame. Therefore, any frame length change can be done in one frame's time (this is different from the *RMAV*). In order to support multimedia traffic, a maximum frame length K_{max} is needed to limit packet delay which is specified by the *QoS* of the traffic. In the adaptive version of *BRMA*, when the total number of packets waiting to be transmitted is B, the number of slots (K) in this frame is chosen as the smaller one of B and K_{max}, i.e.,

$$K = \min(B, K_{max}) \tag{38}$$

Apparently, under light traffic load, the number of packets waiting to be transmitted is small, the frame length is small so that the packets do not have to wait for a long time before being scheduled for transmission. The packet delay is therefore reduced. Under high traffic load, the number of packets waiting to be transmitted is large, the frame length is increased so that the effect of overhead can be reduced and a higher throughput can be achieved. In either case, there are no unused data slots in any frame. We simulate the performance of the adaptive *BRMA* and compare it with the non-adaptive version of *BRMA*. Simulation environment is the same as the one corresponding to figure 18. The results of simulations that are comparative to those shown in figure 18 are given in figure 33. When K_{max} is around 250, the packet loss rates for video and voice are 2.0×10^{-6} and 4.0×10^{-4}, much lower than the minimum packet loss rates 8.0×10^{-6} and 3.0×10^{-3} of the non-adaptive version. Shown in the figure 34 is the comparative delay performance between *BRMA* and Adaptive *BRMA*. Our simulations demonstrate that packet delay is also significantly reduced: when $K =$ 150 slots (or $K_{max} = 150$ slots for Adaptive *BRMA*), Adaptive *BRMA* reduces delays from 8.8 ms to 4.8 ms for voice and from 7.0 ms to 4.5 ms for video respectively. Hence we conclude that the Adaptive version does reduce packet loss rate and packet delay significantly.

Obviously, as K_{max} increases, the regulation effect of the service ratio tends to decrease. The resulting service discipline will approach that of a first in first out (*FIFO*). This is confirmed in figure 33: for large values of K_{max} (near or more than 400 slots), the packet loss rates for video and voice are asymptotically close to each other. In a multimedia environment, K_{max} should be chosen to accommodate the most stringent delay limit requirement. In this simulation, the video packet has a tighter delay limit requirement (under 20 ms) than the voice packet (under 32 ms). From figure 33, one can choose the frame length to be 250 slots (or about 10 ms), which is about half of the delay limit for the video traffic.

7. Conclusions and Discussions

This dissertation focuses on the performance of multiple access control protocols and proposes a novel bandwidth reservation multiple access (*BRMA*) protocol for wireless *ATM* local networks.

This protocol specifies the contention-free reservation channels and collision-free data transmission channels. High multiplexing gain in a high speed wireless system can be achieved due to high protocol efficiency and due to low level bandwidth sharing. With the incorporated admission control scheme, the service quota provides *QoS* to multimedia traffic. This is very important in the wireless *ATM* networks.

Mathematical analysis of modeled traffic and simulation studies of real-time multimedia traffic (voice, video and data) show that this protocol is very efficient. Compared with many random access protocols (such as *Aloha*, *Slotted Aloha*, *PRMA*, *CSMA/CD* etc), *BRMA* produces lower packet loss rate and shorter average packet delays. Compared with the fixed assignment access protocols *BRMA* has more scalability and results in higher channel utilization, especially for variable bit rate traffic. Further more, multiple *QoS* requirements can be accommodated with bandwidth reservation. In short, *BRMA* is a good choice of *MAC* protocol in high speed (over 10 Mbps) high load (over 70% utilization) wireless ATM networks.

An adaptive *BRMA* is also studied. We demonstrated that it has the capability to provide low packet delay in light traffic load and high utilization in heavy traffic load.

This protocol can have many potential applications in the wireless local networks where the propagation delay is trivial. It can be used to provide wireless users with access to the *ATM* backbone network services. Examples of applications may include the micro-cellular multimedia service systems, family communications systems in which many cordless equipment are sharing the access to the interfaces that are connected to the optical fiber networks. Of course it can also be applied to other emerging new applications.

Variables	Values
mean talkspurt duration	1.00 second
mean silent duration	1.35 seconds
mean data packet interarrival time	0.32 seconds
channel rate	360 kbps
speech coding rate	32 kbps
packet size	576 bits
data slot duration	1.6 ms
frame duration (PRMA)	16 ms
Number of slots per frame (PRMA)	10
Priority P_v:P_d (PRMA)	15:1
number of voice terminals M_v	5, 10, 15

Table 1 PRMA and BRMA System Parameters

	peak (bit/frame)	mean (bit/frame)	variance	peak/mean ratio
low activity video (lecture)	48,384	11,543	1.4e+12	4.2
high activity video (Star Wars)	161,726	16,679	3.6e+12	9.7

Table 2 Traffic Parameters for Video Sources

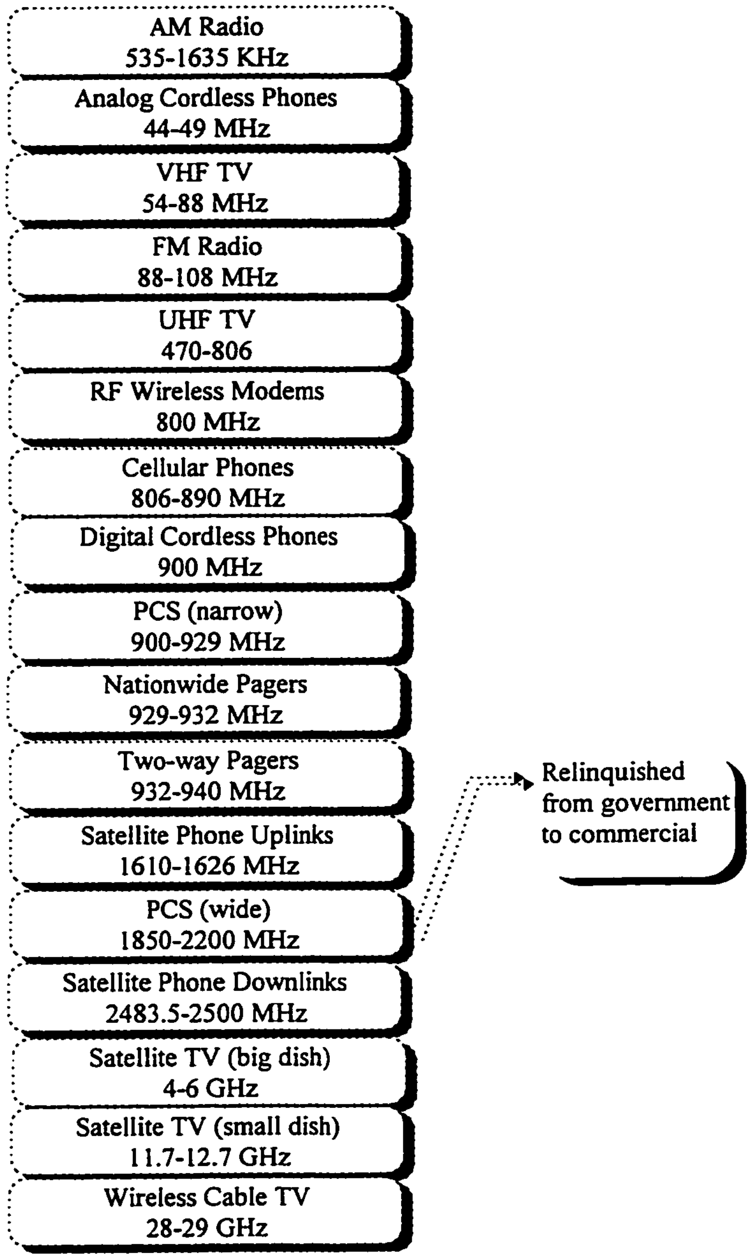

Figure 1. Wireless Bandwidth

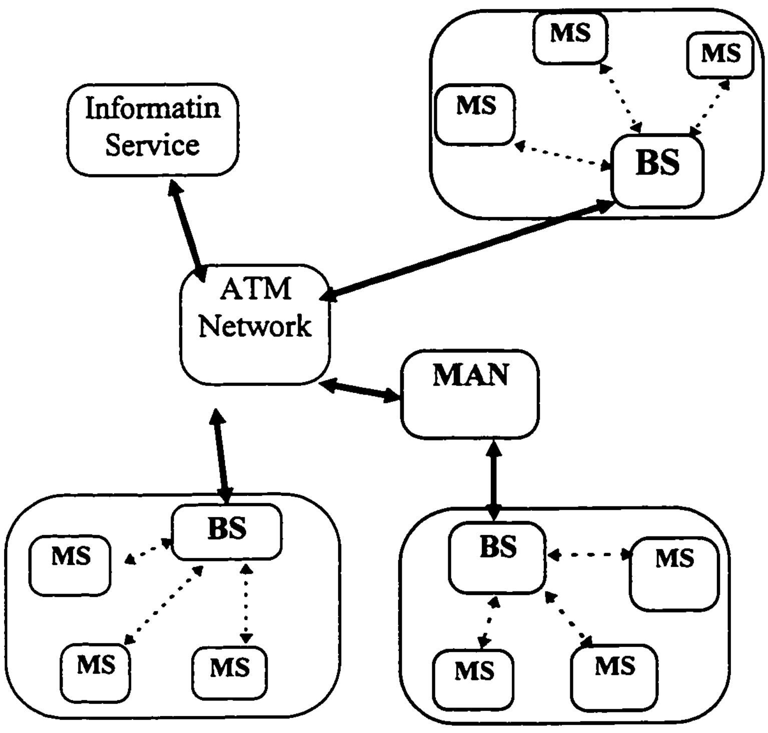

Figure 2 System Overview

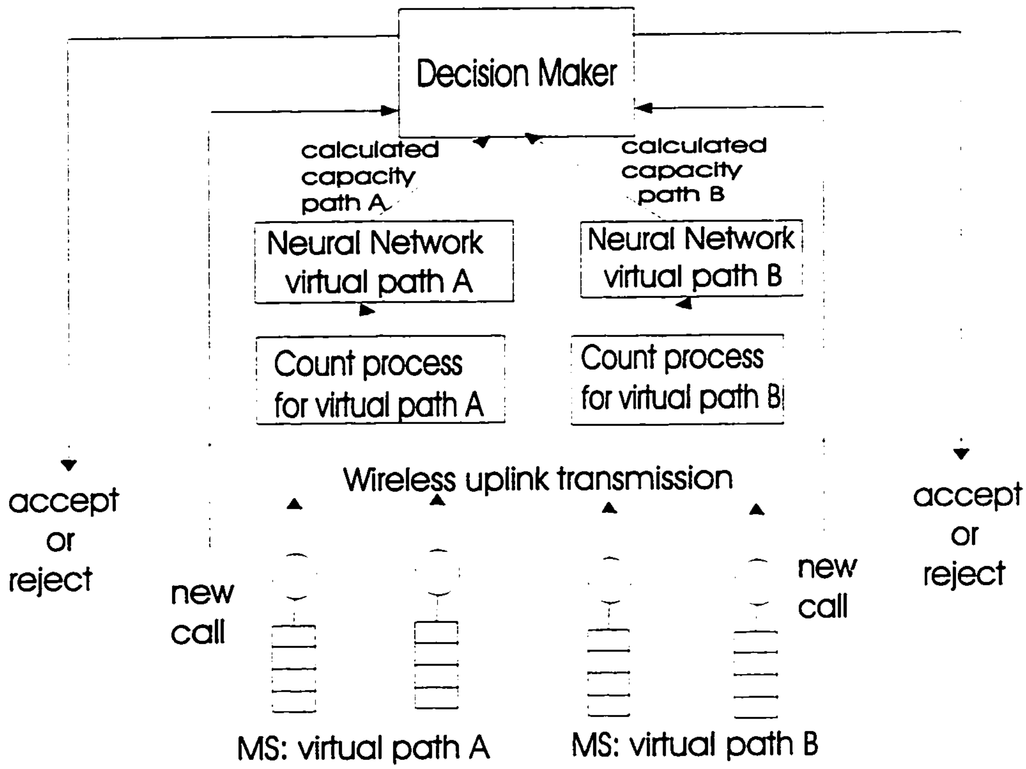

Figure 3 NN Admission Control

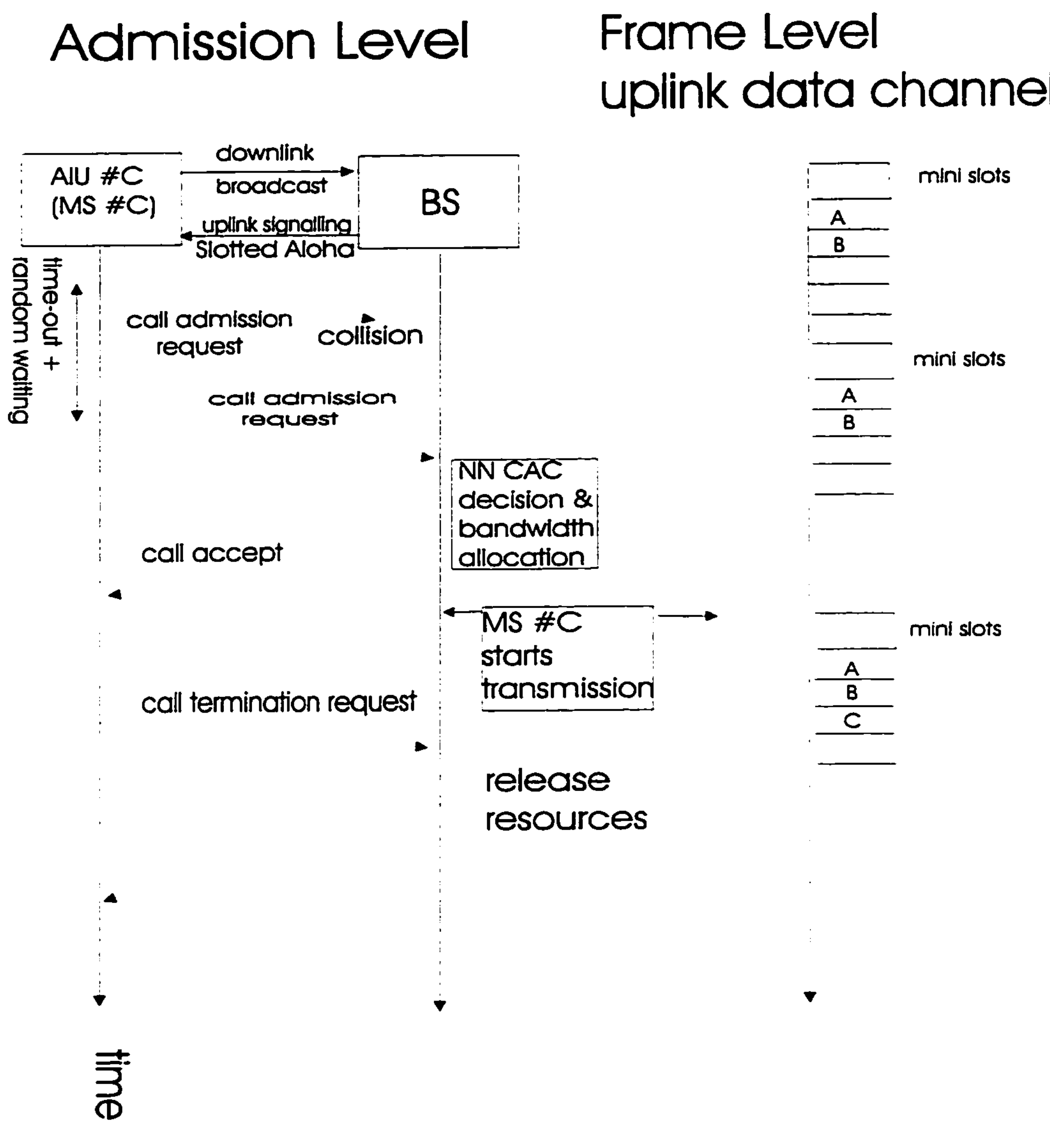

Figure 4 Signalling

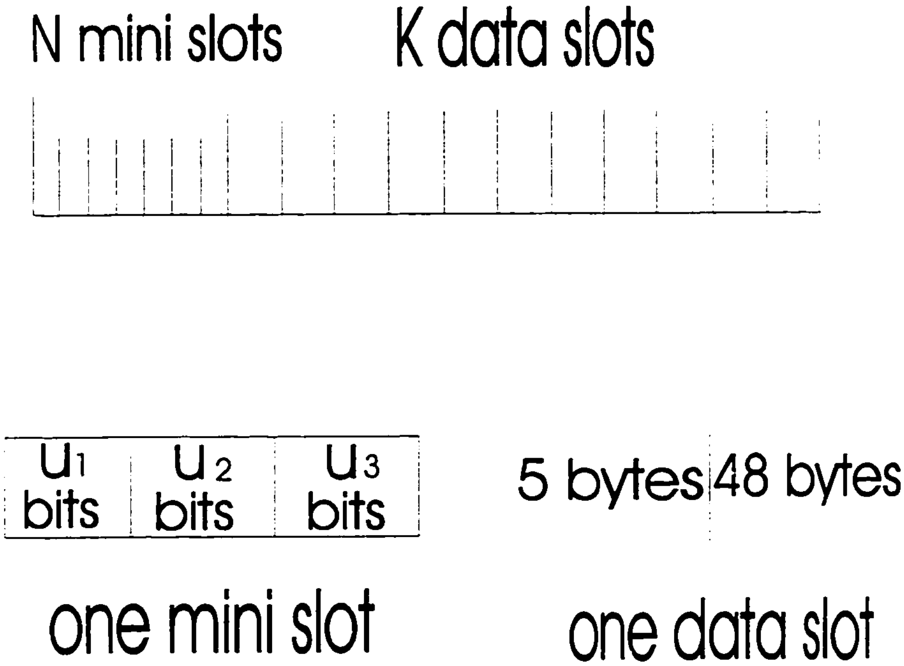

Figure 5 Frame Structure

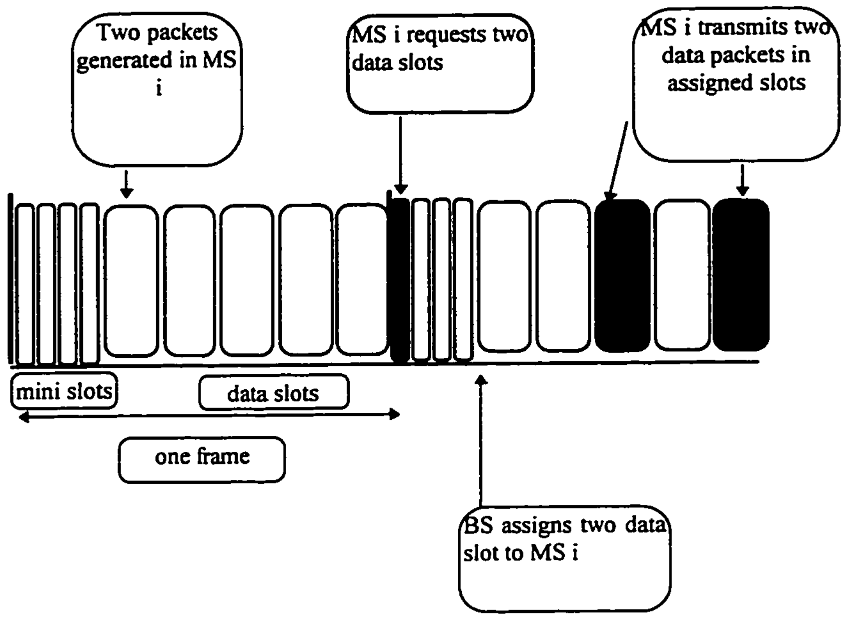

Figure 6 Transmision Scheduling

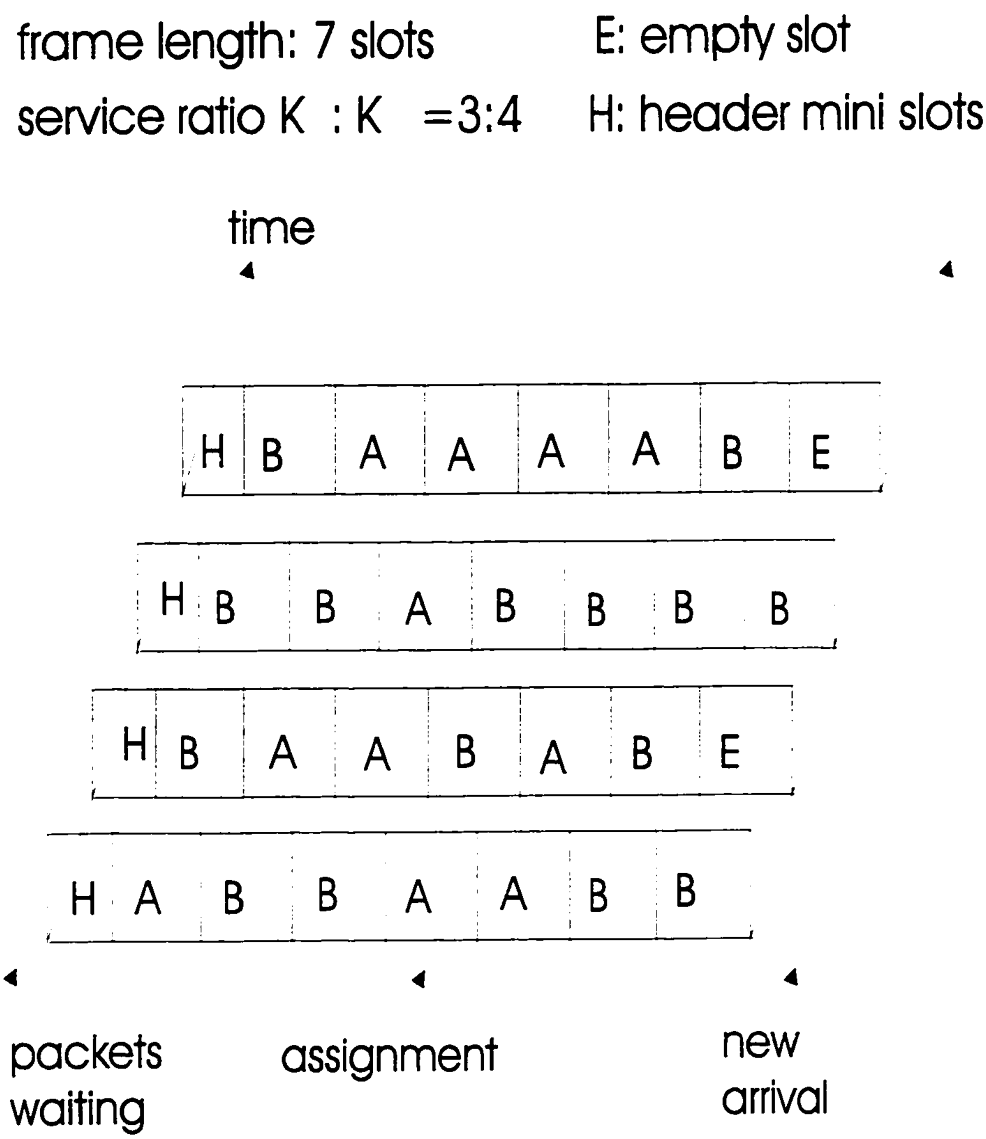

Figure 7 An Example of Data Slots Assignment

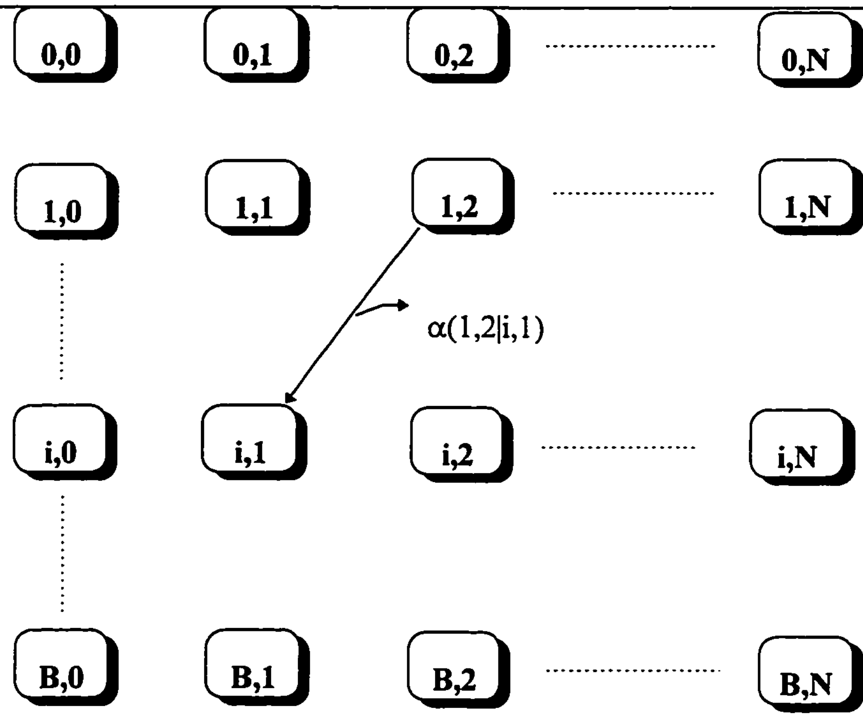

Figure 8 Two-State Markov Chain

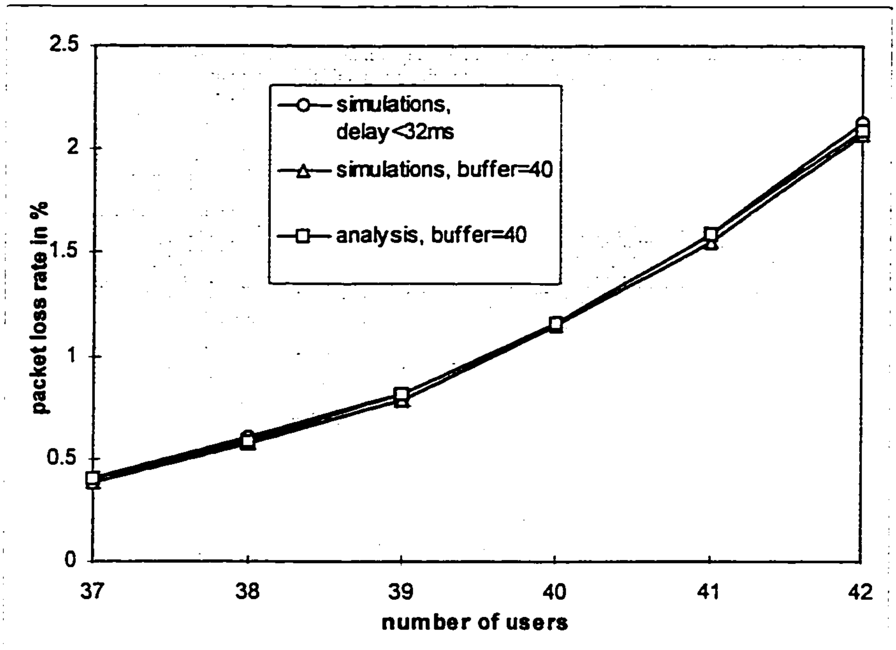

Figure 9 Packet Loss for BRMA

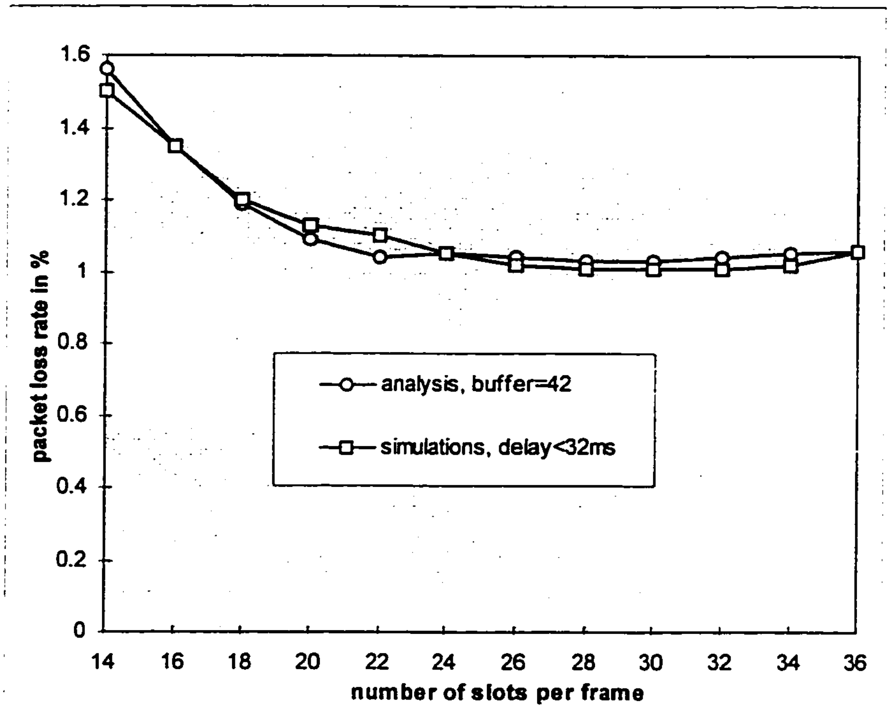

Figure 10 Packet Delay for Different Frame Length

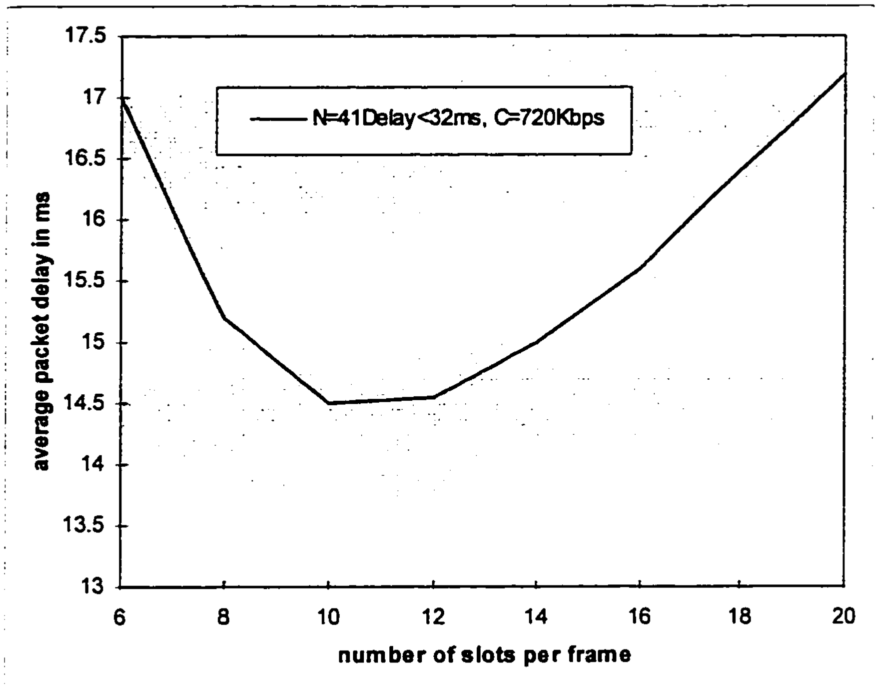

Figure 11 Delay performance for different frame length

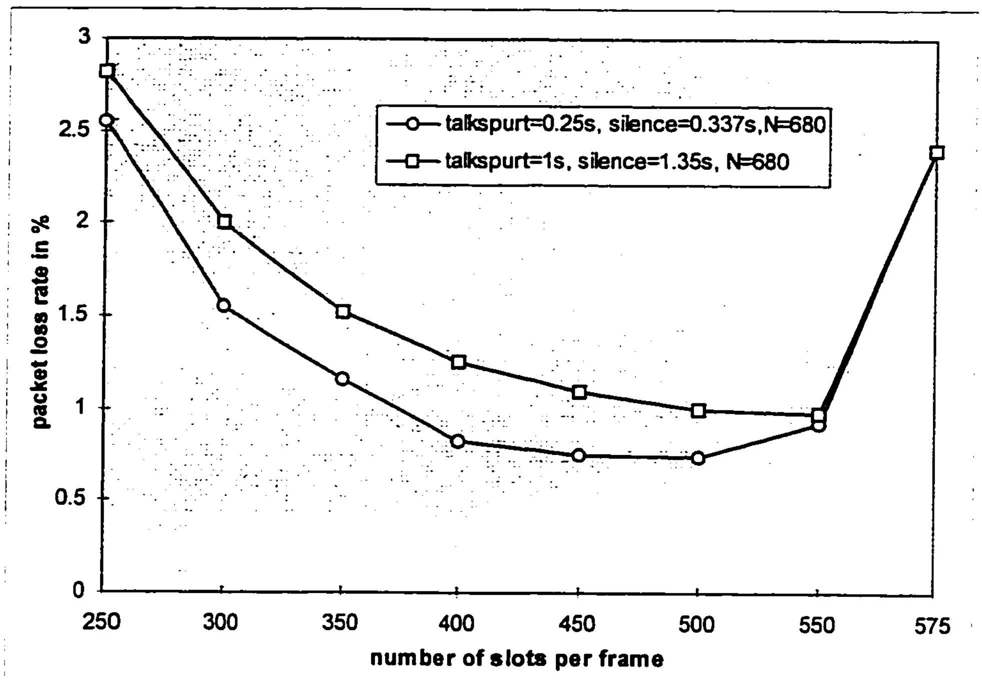

Figure 12 Packet Loss for Traffic of Different Burstiness

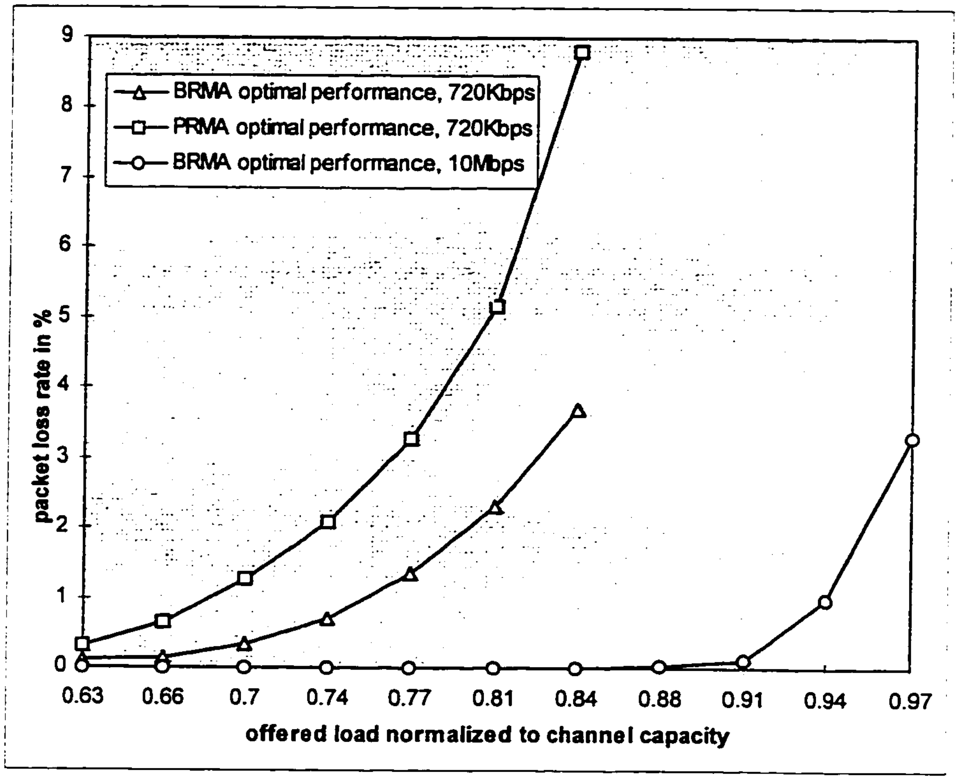

Figure 13 Packet Loss for PRMA and BRMA

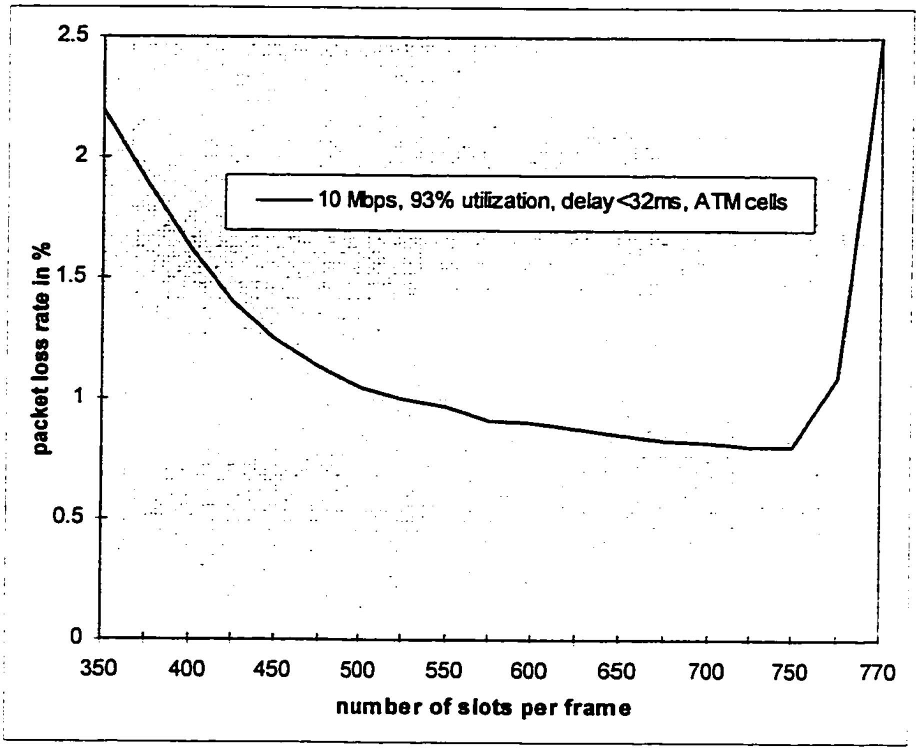

Figure 14 Packet Loss for ATM Size Voice Packet

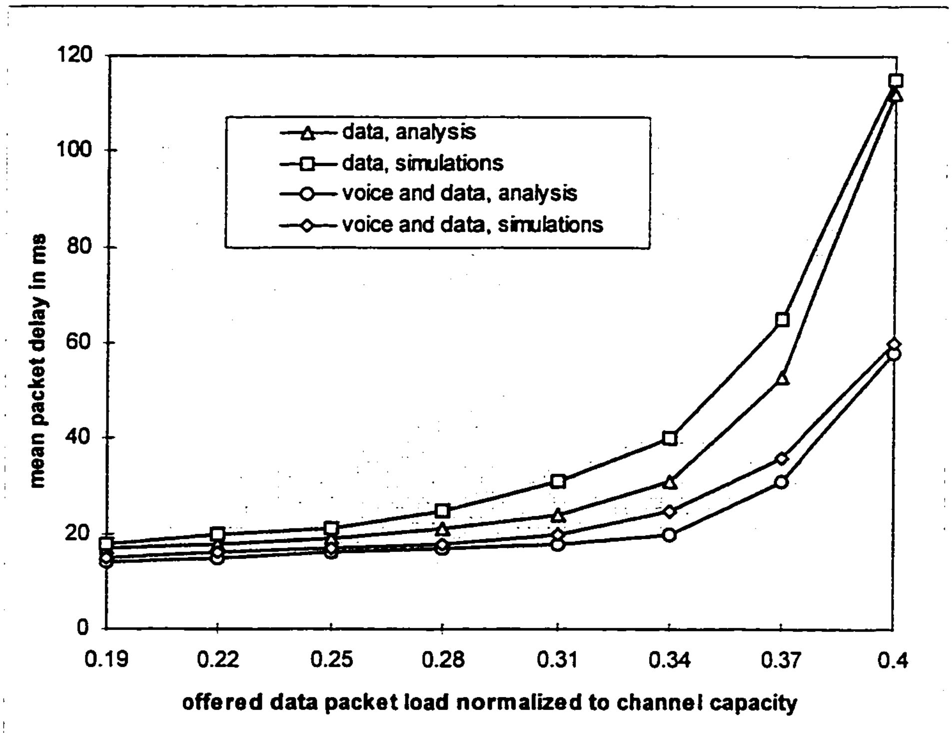

Figure 15 Data and Voice Delay Performance, $L_f = t_p$

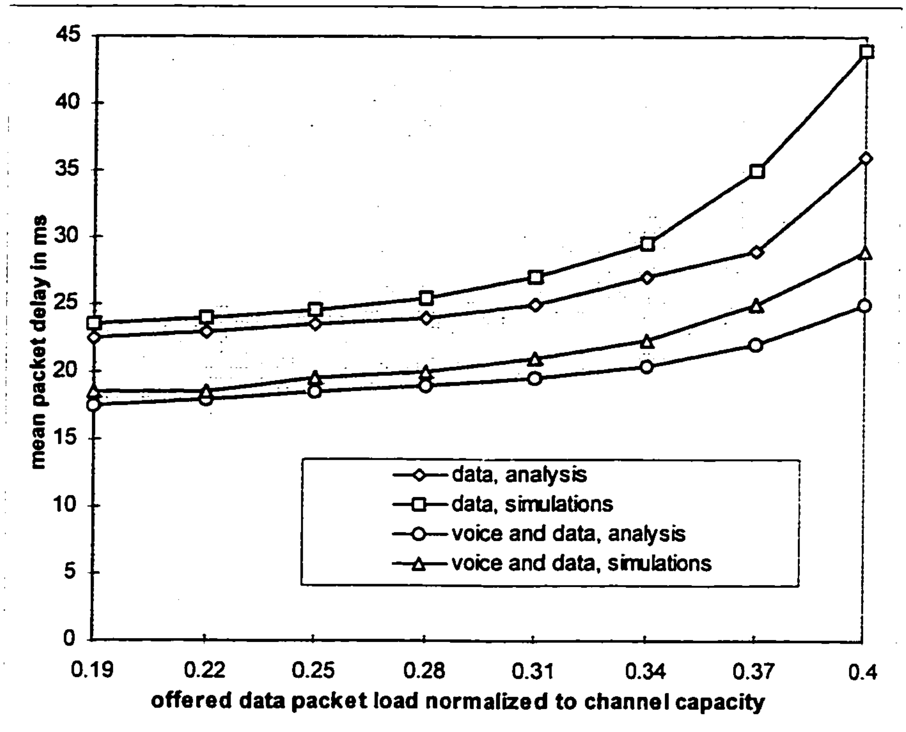

Figure 16 Voice and Data Delay Performance, K=29

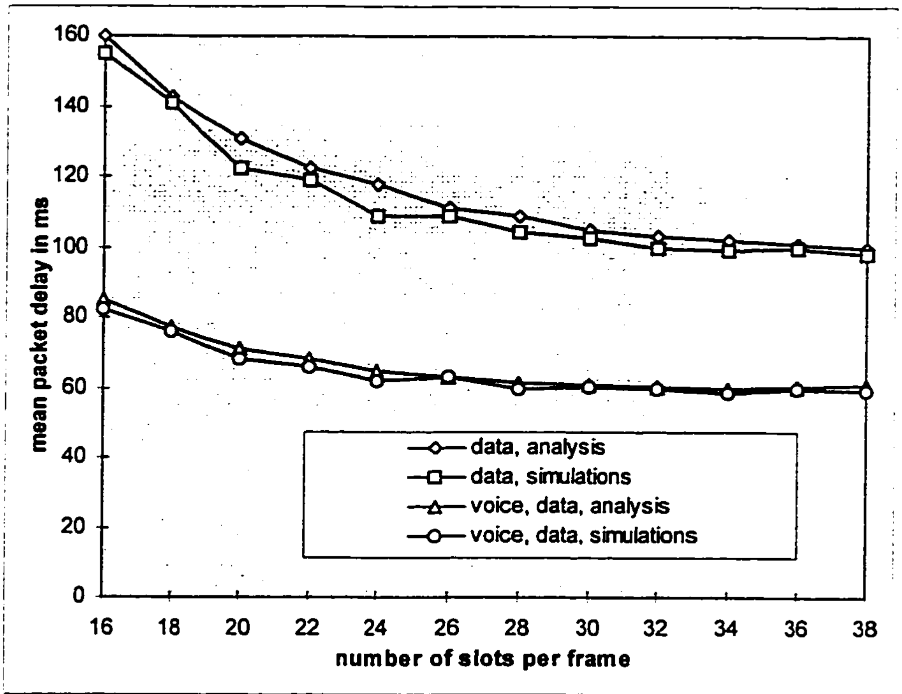

Figure 17 Voice and Data Delay for Different Frame Length K

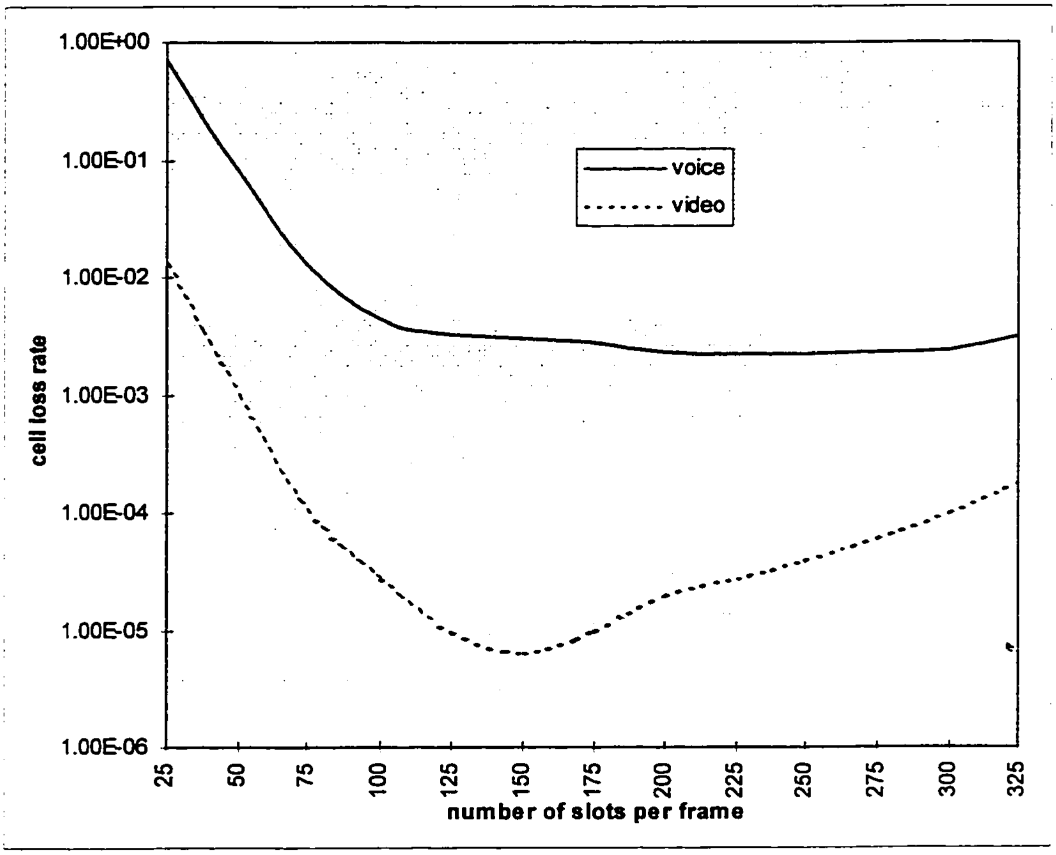

Figure 18 Packet Loss for Different Frame Length

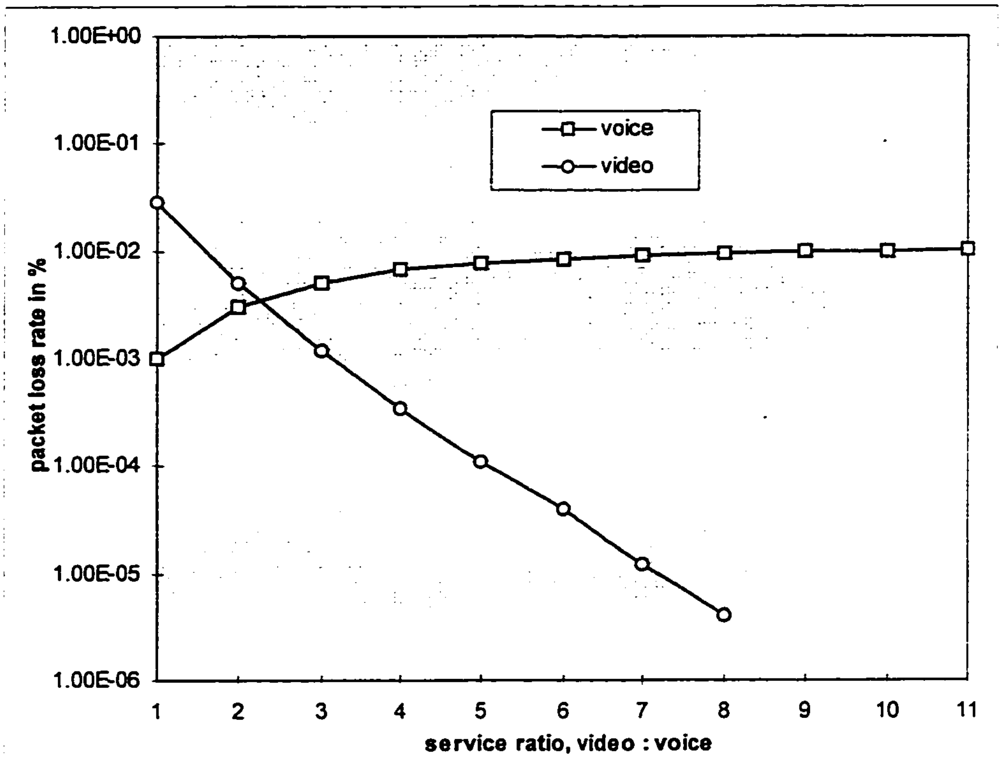

Figure 19 Packet Loss and Service Ratio

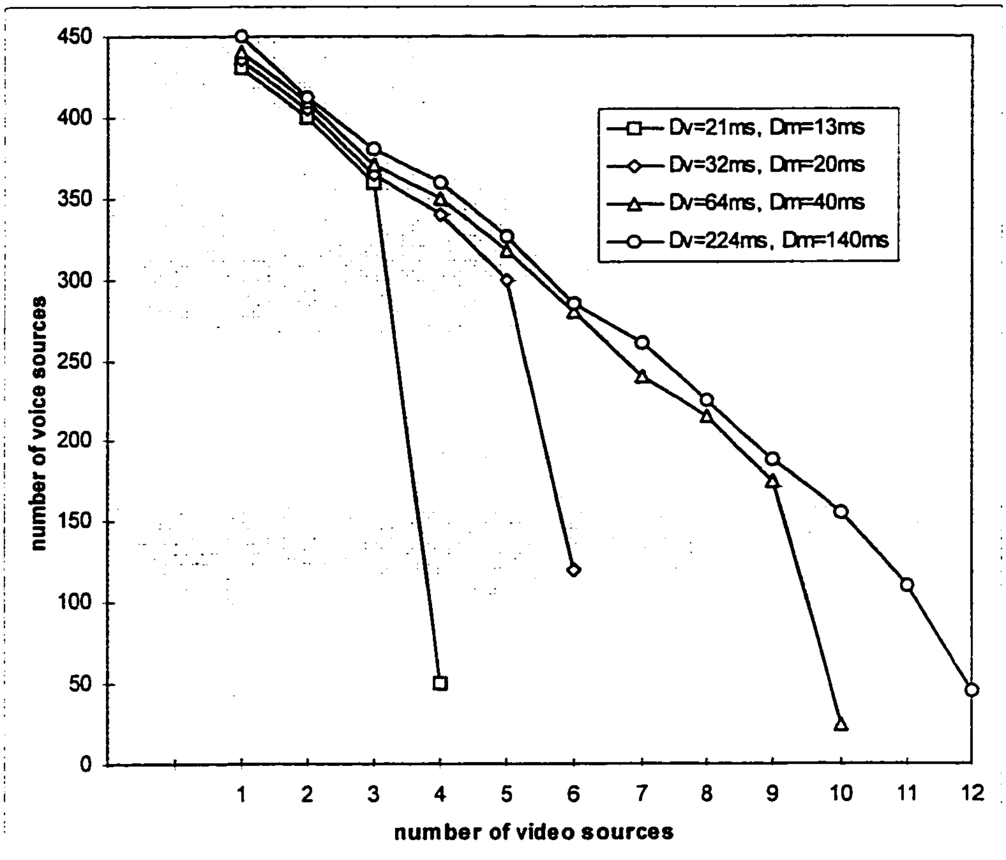

Figure 20 Utilization and QoS, High Activity Video

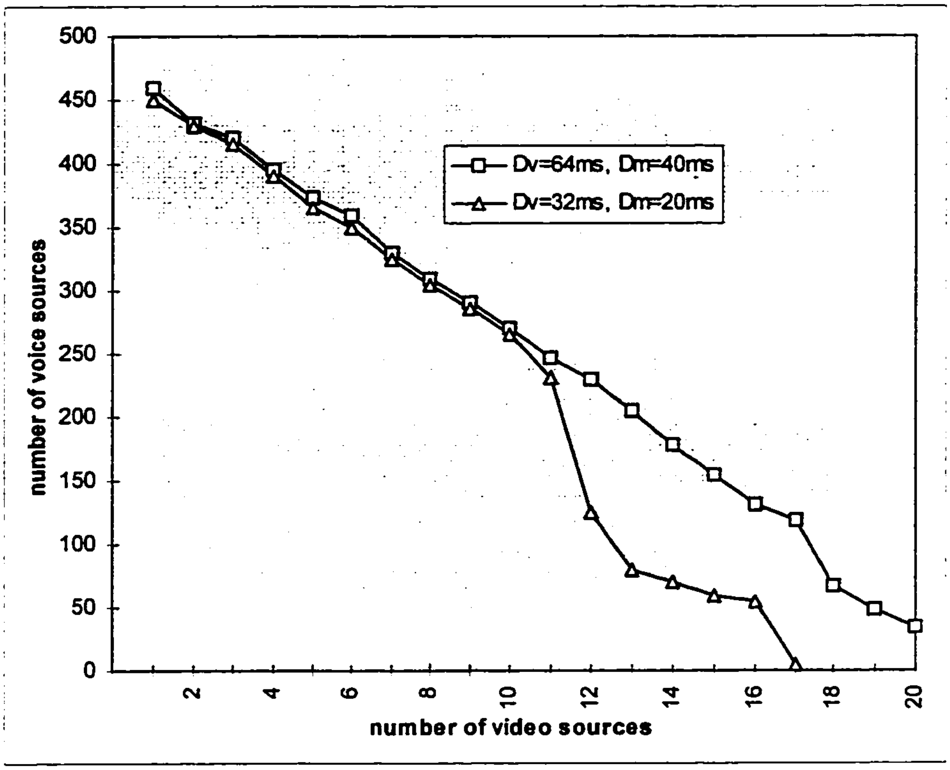

Figure 21 Utilization and QoS, Low Activity Video

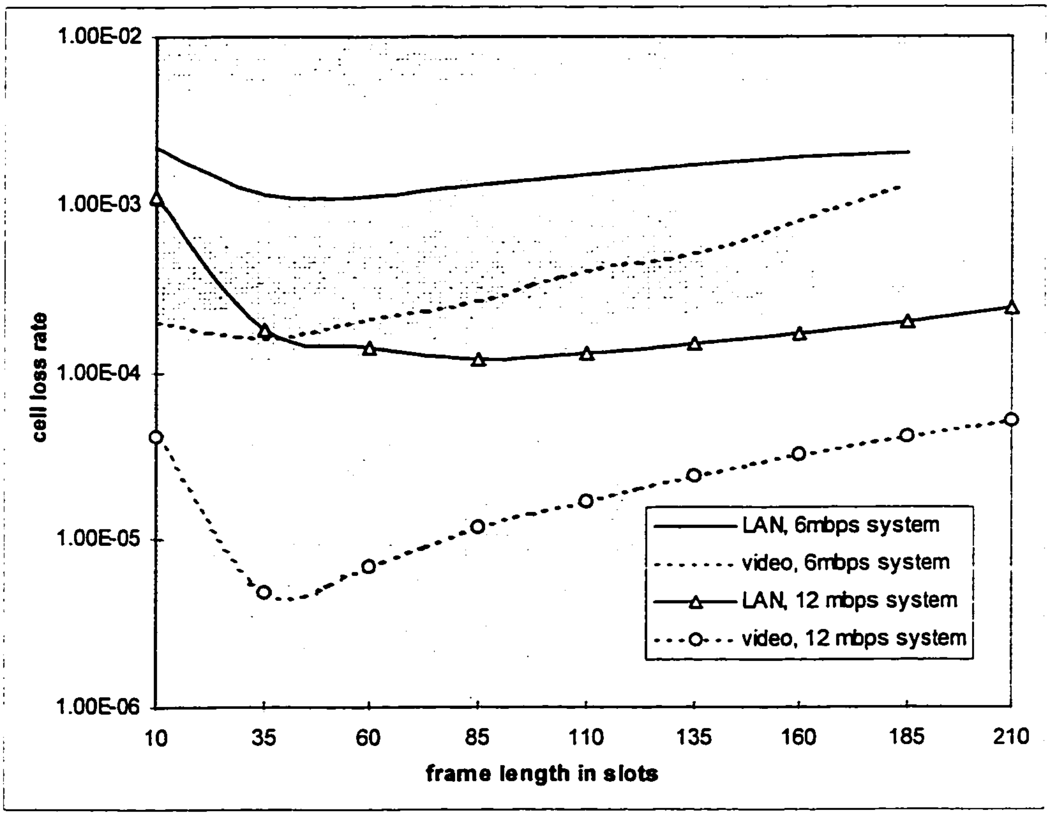

Figure 22 Packet Loss for Video and LAN Traffic (12 Mbps)

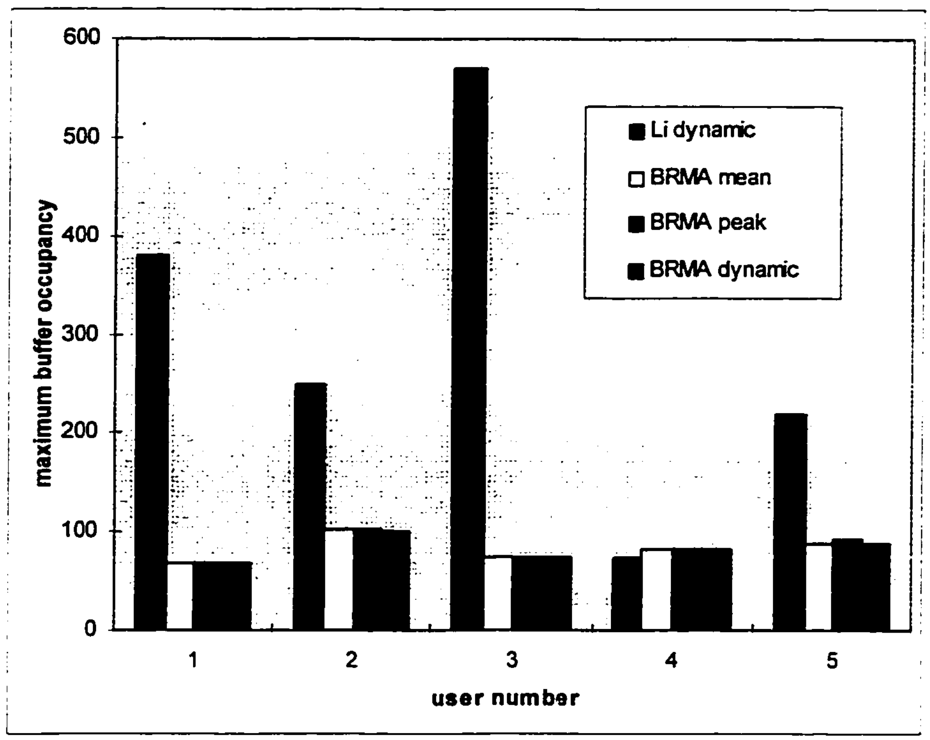

Figure 23 Maximum Queues for JPEG, 70% Utilization

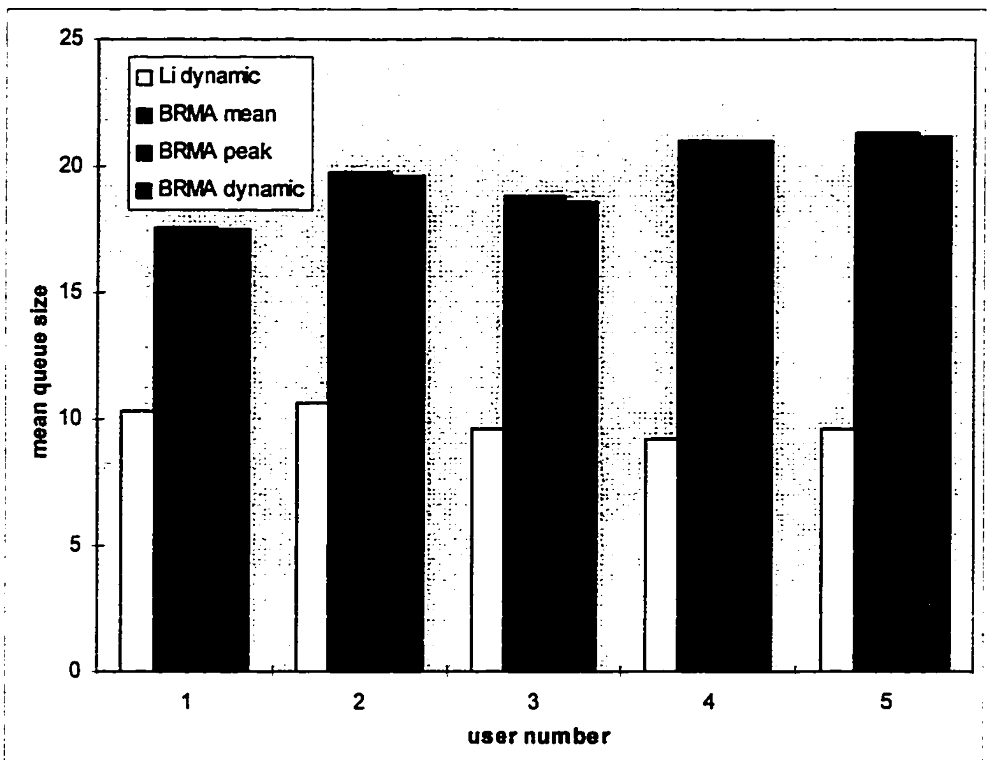

Figure 24 Mean Queue Size for JPEG, 70% Utilization

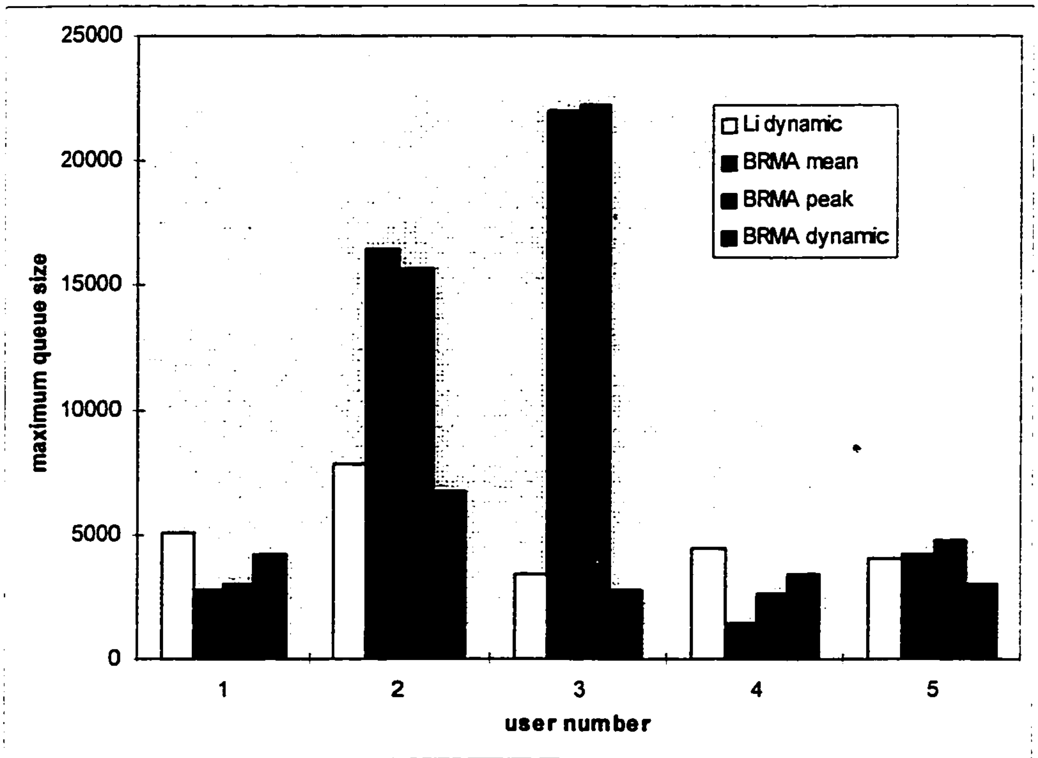

Figure 25 Maximum Queue Size for JPEG, 90% Utilization

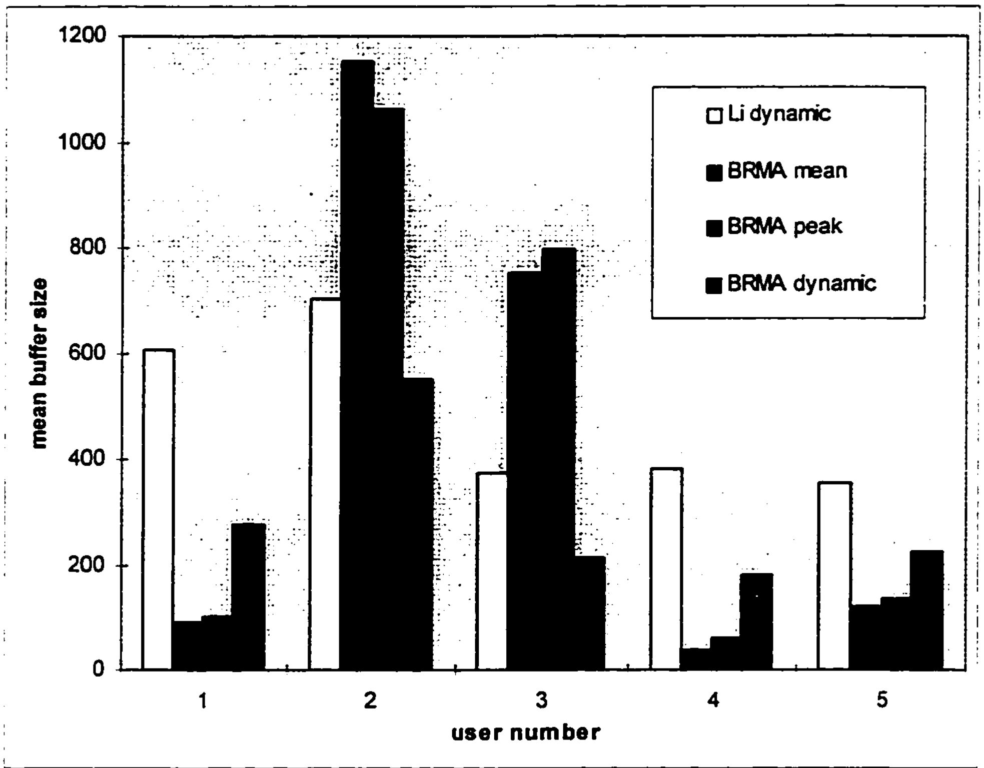

Figure 26 Mean Queue Size for JPEG, 90% Utilization

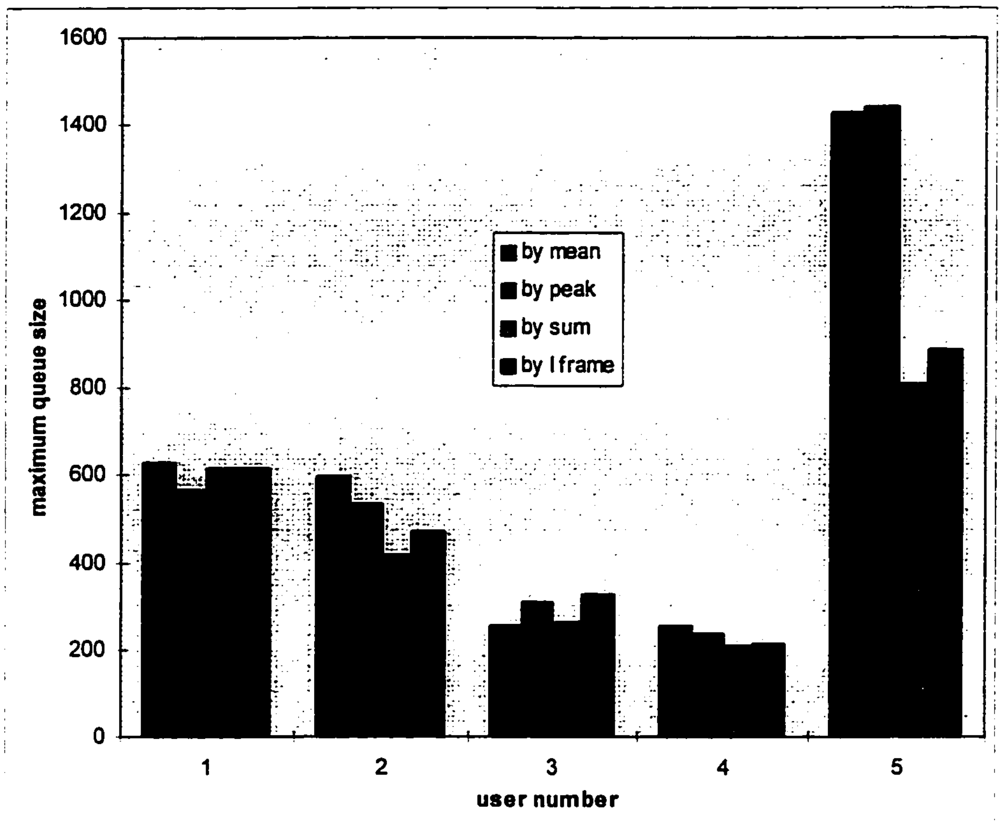

Figure 27 Maximum Queue Size for MPEG, 70% Utilization

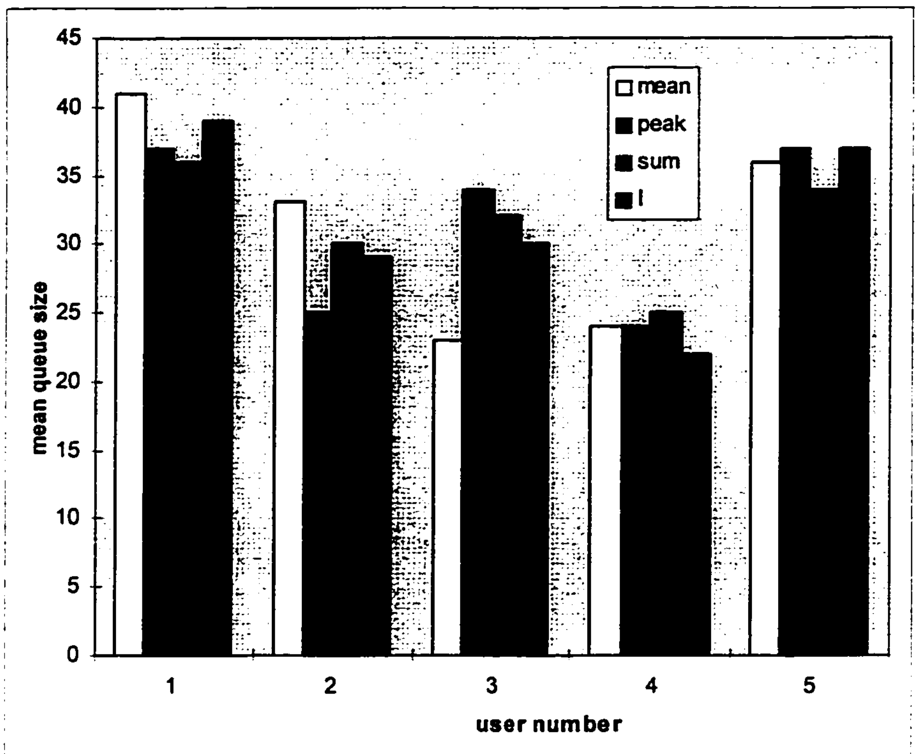

Figure 28 Mean Queue Size for MPEG, 70% Utilization

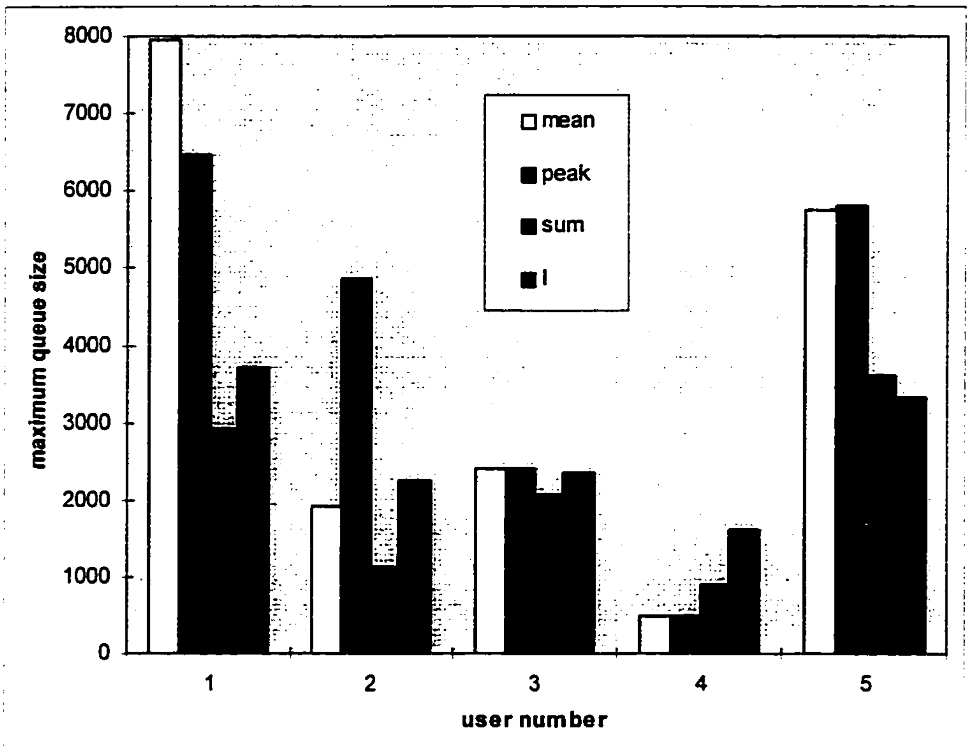

Figure 29 Maximum Queue Size for MPEG, 90% Utilization

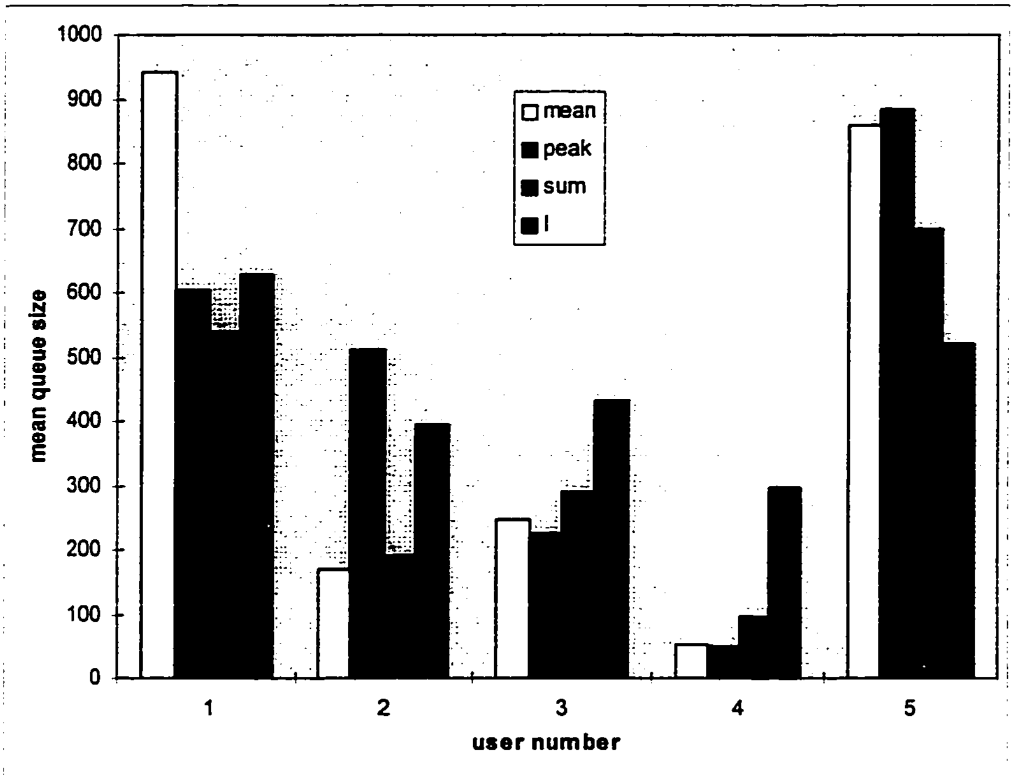

Figure 30 Mean Queue Size for MPEG, 90% Utilization

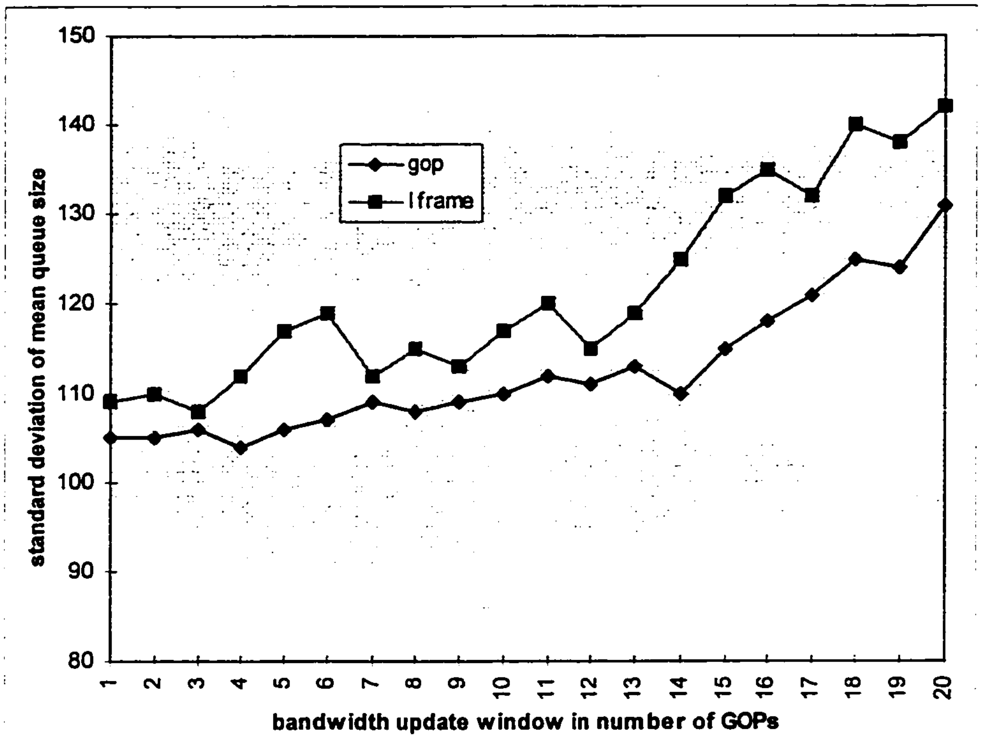

Figure 31 Bandwidth Update Window and Mean Queue Deviation

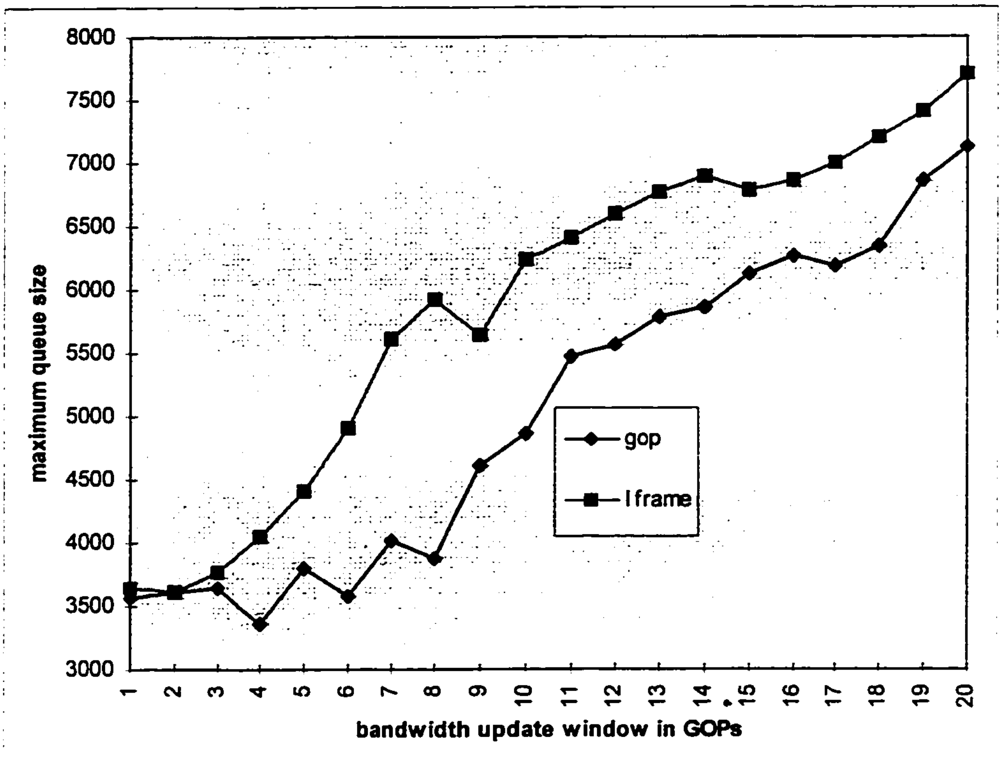

Figure 32 Bandwidth Update Window and Maximum Queue Size

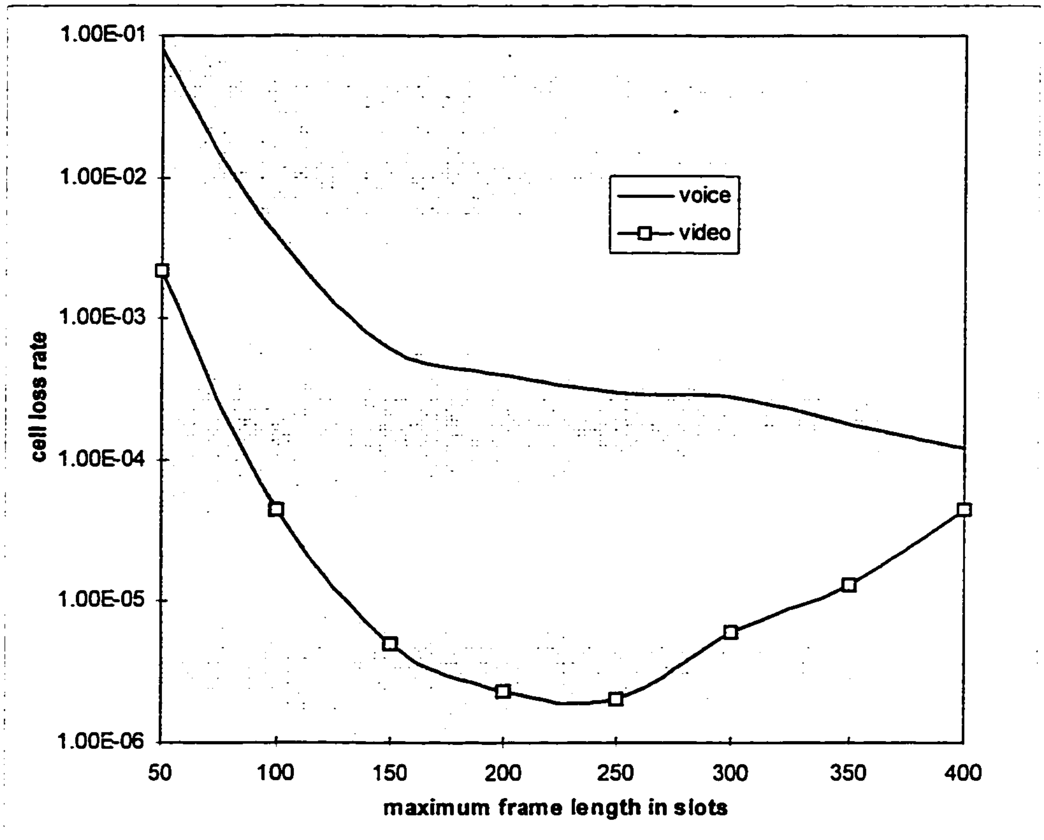

Figure 33 Packet Loss for Adaptive BRMA

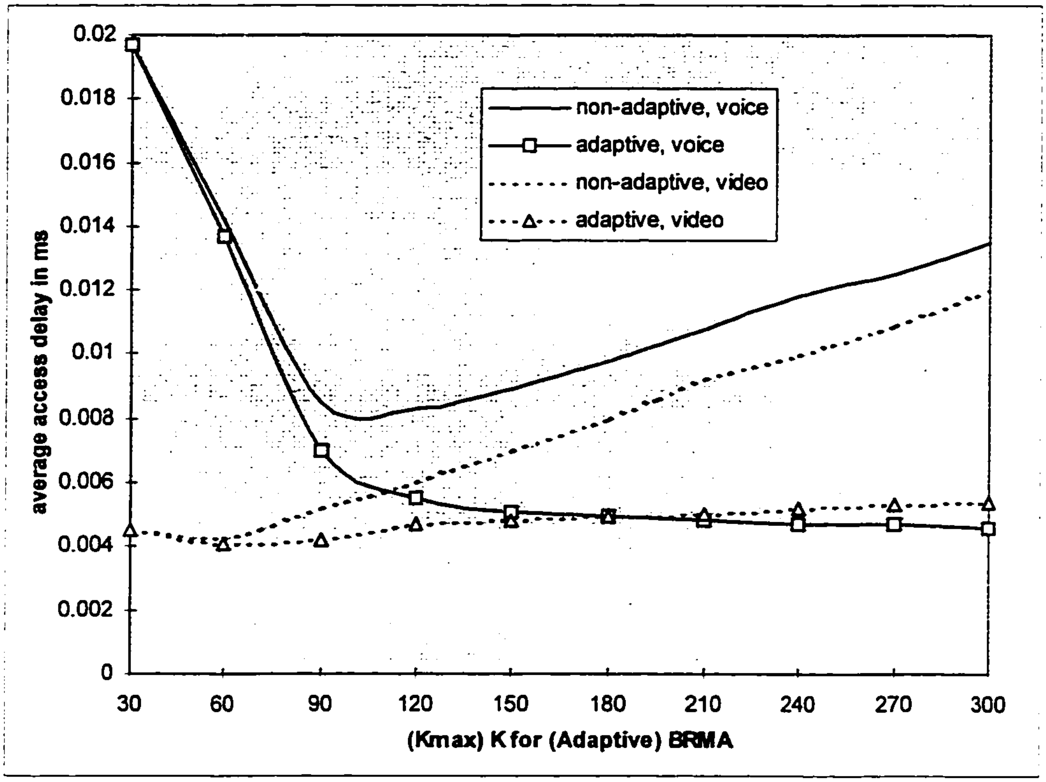

Figure 34 Packet Delay for Adaptive BRMA

Bibliography

[1] Kaveh Pahlavan and Allen H. Levesque, *Wireless Information Networks*, John Wiley Sons, INC. 1995

[2] J. Goodman, R. A. Valenzuela, K. T. Gayliard and B. Ramamurthi, "Packet Reservation Multiple Access For Local Wireless Communications," IEEE Trans. Comm., Vol. 37, No. 8, pp. 885--889, Aug. 1989

[3] J. Goodman and Sherry X. Wei, "Efficiency Of Packet Reservation Multiple Access," IEEE Trans. on Vehicular Tech., Vol. 40, No. 1, pp. 170--176, Feb. 1991.

[4] S. Nanda, "Analysis Of Packet Reservation Multiple Access: Voice And Data Integration For Wireless Networks," in Proc. GLOBECOM90, Dan Diego, CA, 1990, pp.1984--1988.

[5] Mario Frullone, Guido Riva, Paolo Grizios and Maria Missiroli, "Comparisons Of Multiple Access Schemes For Personal Communication Systems In A Mixed Cellular Environment," IEEE Trans. on Vehicular Tech., T. Vol. 43, No. 1, pp.99--109, Feb. 1994.

[6] Wai-choong Wong, "Packet Reservation Multiple Access In A Metropolitan Microcellular Radio Environment," IEEE JSAC., Vol. 6, pp 918 -- 925, Aug., 1993.

[7] Gang Wu, Kaiji Mukumoto and Akira Fukuda, "Analysis Of An Integrated Voice And Data Transmission System Using Packet Reservation Multiple Access," IEEE Trans. on Vehicular Tech., Vol. 43, No. 2, pp. 289--397, May 1994.

[8] N. M. Mitrou, Th. D. Orinos, and E. N. Protonotarios, " A Reservation Multiple Access Protocol For Microcellular Mobile-Communication Systems," IEEE Trans. on Vehicular Tech., Vol. 39, No. 4. pp. 340--350, Nov., 1990.

[9] David D. Falconer, Fumiyuki Adachi and Bjorn Gudmundson, "Time Division Multiple Access Methods For Wireless Personal

Communications," IEEE Comm. magazine, Vol.33, No.1, pp. 50--57, Jan. 1995

[10] Newman D. Wilson, Rajamani Ganesh, Kuriacose Joseph and Dipankar Raychaudhuri, "Packet CDMA Versus Dynamic TDMA For Multiple Access In An Integrated Voice/data PCN," IEEE JSAC, Vol. 11, No. 6, pp. 870--884, Aug. 1993

[11] Kaveh Pahlavan, Thomas H. Prober and Mitchell E. Chase, "Trends In Local Wireless Networks," IEEE Comm. magazine, pp. 88--95, March 1995

[12] ATM Forum Technical Committee, Traffic Management Workin Group, document ATM Forum/af-tm-0056.000.

[13] I. Habib and T. Saadawi, "Dynamic Bandwidth Allocation and Access Control of Virtual Paths in ATM Broadband Networks," in Proc. 4th IFIP Conf. High Performance Networks, Dec. 1992.

[14] H. Saito, M. Kawarasaki and H. Yamada, "An Analysis of Statistical Multiplexing in an ATM Transport Networks," IEEE JSAC, Feb. 1991.

[15] R. Guerin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High Speed Networks," IEEE JSAC, Dec. 1991.

[16] A. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," IEEE/ACM Trans. on Networking, June 1993.

[17] S. Chong, S. Li, and J. Ghosh, "Predictive Dynamic Bandwidth Allocation for Efficient Transport of Real-Time VBR Video over ATM", IEEE JSAC, Jan. 1995, pp. 12-23

[18] G. Choudhury, D. Lucantoni and W. Whitt, "Squeezing the Most Out of ATM," IEEE Trans. on Comm., Vol. 44, No. 2, Feb., 1996

[19] S. Youssef, I. Habib and T. Saadawi, "A Neurocomputing Controller for Bandwidth Allocation in ATM Networks", IEEE JSAC, Feb. 1997, pp. 191-199.

[20] Qiyi Zhu, Joseph G. Kneuer and Wei-Min Huang, "A Study Of Temporal Behavior Between The Packet Loss Bursts In Packet Switched Systems," IEEE Proc. ICC 1995

[21] San-Qi Li, "A General Solution Technique For Discrete Queuing Analysis Of Multimedia Traffic On ATM," IEEE Trans. on Comm., Vol. 39, No.7, pp.

1115--1132, July 1991

[22] Tarek N. Saadawi, Mostafa H. Ammar and Ahmed El Hakeem, *Fundamentals of Telecommunication Networks* , John Willey Sons Inc., 1994

[23] Jeffre E. Wieselthier and Anthony Ephremides, "Fixed- And Moveable-boundary Channel-access Schemes For Integrated Voice/data Wireless Networks," IEEE Trans. on Comm., vol. 43. No. 1, pp. 64--74, Jan. 1995

[24] Harry Heffes and David M. Lucantoni, "A Markov Modulated Characterization Of Packetized Voice And Data Traffic And Related Statistical multiplexer Performance," IEEE JSAC, Vol. SAC-4, No. 6, pp. 856--868, Sept. 1986

[25] Andrea Baiocchi, Nicola Blefari Melazzi, Marco Listanti, Aldo Roveri and Roberto Winkler, "Loss Performance Analysis Of An ATM Multiplexer Loaded With High-speed On-off Sources," IEEE JSAC, Vol. 9, No. 3, pp. 388--393, April 1991

[26] San-qi Li, Song Chong, Chia-lin Hwang and Xirong Zhao, "Link Capacity Allocation and Network Control by Filtered Input Rate in High Speed Networks," IEEE Proc. GLOBECOM93, pp. 744--750.

[27] WG 11, "MPEG I: Coded Representation of Picture, Audio and Multimedia/hypermedia Information," ISO/IEC JTC 1/SC 29, Nov. 1991

[28] M. W. Garrett, "Contributions Toward Real-Time Services on Packet Networks," Ph.D. Dissertation, Columbia University, May 1993

[29] W. Leland, M. Taqqu, W. Willinger and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," IEEE/ACM Trans. on Networking, Vol. 2, No. 1, Feb. 1994

[30] Song Chong, San-qi Li and Joydeep Ghosh, "Dynamic Bandwidth Allocation for Efficient Transport of Real-Time VBR Video over ATM," IEEE Proc. INFOCOM94, pp. 81- 90.

[31] D. G. Jeong, C. Choi and W. Jeon, "Design and Performance Evaluation of a New Medium Access Control Protocol for Local Wireless Data Communications," IEEE/ACM Trans. on Networking, Vol. 3, No. 6, pp. 742--752, Dec. 1995

IMAGE EVALUATION TEST TARGET (QA-3)

150mm

6″

APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989