Connection Admission and Resource Allocation for Real Time and Non Real Time Traffic in a WiMAX System

A Thesis Submitted to the Faculty of Indian Institute of Technology
in Partial Fulfillment of the Requirements for the award of the degree of
Bachelor of Technology

Major Subject: Electronics and Communication Engineering

Sarabjot Singh (06010237)
Thesis Advisor: Prof. Sanjay K. Bose

Department of Electronics and Communication Engineering

Indian Institute of Technology (IIT)
Guwahati, India

April, 2010

Certificate

This is to certify that the work contained in this thesis entitled "Connection Admission and Resource Allocation for Real Time and Non Real Time Traffic in a WiMAX System" is a bonafide work of Sarabjot Singh (Roll number: 06010237), carried out in the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.

Professor Sanjay K. Bose
Department of Electronics & Communication Engineering,
Indian Institute of Technology (IIT) Guwahati,
Guwahati 781039, Assam, India.

Acknowledgments

First and foremost I would like to express my deepest gratitude and appreciation to my thesis advisor Prof. Sanjay Kumar Bose for his excellent guidance and generous support. Prof. Bose has been a source of endless creative ideas throughout my research work. Without him, finishing this thesis would not have been possible. Often times, I have realized, how fortunate I am to have an advisor who takes the pain of editing and proofreading my drafts through so many iterations.
I am thankful to Prof. Ma Maode of NTU, Singapore for his valuable inputs during the research. My sincere thanks to all the faculty members of the evaluation committee for their inputs during evaluation which helped improving the work.

I am also grateful to my friends in IITG, for making my undergraduate experience both memorable and fun. I would also like to thank all the lab assistants and research scholars of the ECE department for creating a positive atmosphere, ideal for conducting research.

Sarabjot Singh
Department of Electronics and Communication Engineering,
Indian Institute of Technology (IIT), Guwahati.

ABSTRACT

The advent of broadband wireless promises quality communications over the wireless channel. For broadband wireless access, the 802.16 standard is expected to provide high-speed data access to subscribers. For the 802.16 WiMAX standard providing high-speed data access, we propose a joint Call Admission Control (CAC) and Bandwidth Allocation (BA) approach for QoS support along with separate resource allocation algorithm for real time and non-real time traffic. We propose both Conservative and Non-Conservative strategies where the Conservative CAC guarantees the QoS requirements for all classes of traffic but is more restrictive and less efficient than the Non-Conservative CAC. Connection admission and bandwidth allocation for rtPS sources with different priorities are also outlined. Both analytical models and simulations are used to study the performance of the proposed schemes.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

In recent years, demands for high-speed internet access and multimedia services for residential and business customers have increased greatly. IEEE has proposed a new IEEE 802.16 standard for Broadband Wireless Access systems. This Worldwide Interoperability for Microwave Access (WiMAX) standard is an air interface for Fixed Broadband Wireless Access Systems and has been ratified by IEEE as a Wireless Metropolitan Area Network (WMAN) technology [1]. WiMAX aims at providing broadband wireless last-mile access in a Metropolitan Area Network. Its main advantages are fast deployment which results in cost savings, and the ability to reach very crowded or rural areas without the need for any wired infrastructure. It also provides Quality of Service (QoS) to support different kinds of real-time application in wireless networks. QoS can provide different priority to different users or data flows or guarantee a certain level of performance to data flows in accordance with requests from the application program or based on the Internet Service Provider's (ISP) service policy. The frequency band supported by the standard ranges from 2 to 66 GHz. The MAC layer supports both point to multi-point (PMP) and mesh mode operations. PMP is a centralized architecture where all traffic between clients or subscriber stations (SS) are controlled by a base station (BS). In the mesh mode, traffic can be routed directly between SSs or between the SS and the BS. The PMP mode is mainly applied in WiMAX, where all data traffic is controlled by the serving BS. Traffic direction is used to distinguish two types of data channels: uplink (UL) channel where data is sent from the SSs to the BS, and downlink (DL) channel where data bursts are sent from BS to all SSs. In both modes, the MAC layer is designed to support quality of service (QoS) in order to enhance end user performance parameters in terms of bandwidth, latency, jitter and reliability. Although the IEEE 802.16 specifications [2] also defines multiple access signaling mechanisms, the radio resource management issues such as bandwidth allocation (BA) and connection admission control (CAC) are still open. Connection admission control is used to limit the number of connections in the network. It works together with the bandwidth allocation mechanism (which allocates available radio resources among the ongoing and the incoming connections) so that the QoS performances of both types of connections can be maintained at the target level. The CAC blocks unwanted calls in order to guarantee the QoS of existing calls and reduces the buffer needed for packet scheduling. To ensure the QoS of high priority services further, packet scheduling grants the channel for service according to their respective priorities. The IEEE 802.16 standard supports five service classes, namely Unsolicited Grant Service (UGS), Real Time Polling Service (rtPS), Non-Real Time Polling Service (nrtPS), Extended Real Time Polling Service (eRTPS) and Best Effort (BE) Service [3], with specific QoS requirement for each service class. The first two classes are used for real-time fixed-size and non-fixed packet size respectively. The rest are used for non real-time data traffic. In general, the bandwidth allocation of WiMAX from Base Station (BS) to Subscriber Station (SS) will be performed under two modes: grant per connection (GPC) and grant per SS (GPSS). In the GPC mode, a BS scheduler treats each connection request from the SS independently and the bandwidth is explicitly granted to each connection. The SS transmits according to the order specified by the BS. In the GPSS mode, the BS scheduler treats all the connections from a single SS as one unit and grants bandwidth to SS. A separate scheduler at SS determines the service priority and duration for its connections in the granted slot.

Several researchers have studied Connection Admission Control (CAC) and Bandwidth Allocation (BA) separately [3],[4],[11] but works which consider the CAC and BA jointly for a WiMAX system are scarce. This report addresses this along with analytical and simulation studies on the performance of the proposed systems. The rest of the report is organized as follows. Section 2 gives a basic description of resource allocation and admission control policies proposed in the literature for the IEEE 802.16 system. Section 3 provides the description of the joint CAC and the fair resource allocation algorithm proposed by us. A discrete time Markov chain based analysis is also presented and the corresponding analytical and simulation results are compared in Section 4. We extend our description to incorporate real time sources with differential QoS requirements in Section 5 by formulating an appropriate joint CAC and BA algorithm and demonstrate its performance through simulation studies. Section 6 concludes the report.

2. Background

2.1 WiMAX MAC Overview

The Media Access Control (MAC) protocol of IEEE 802.16 is connection oriented and each connection is identified by a 16-bit Connection Identification Number (CID), which is given to each subscriber station (SS) in the initialization process. The transmissions are divided either by Time Division Duplex (TDD) or Frequency Division Duplex (FDD) method. In the downlink (DL) direction, connections are usually multicast, but unicast can also be supported. The SSs use Time-Division-Multiple-Access (TDMA) on the uplink and transmit back to the base station (BS) in specific allocated time slots. This means that connections from the SSs to the BS are always unicast. Thus the CID plays an important identification role in the uplink (UL) channels allowing the BS to identify the SS that sent the MAC PDUs in the UL direction. Unlike ETHERNET or WiFi networks, the 48-bit MAC address does not play any role in the transmission but serves only as an equipment identifier.

QoS guarantees are made possible through the QoS differentiation that may be needed by different types of service flows that might operate in such a broadband wireless network. IEEE 802.16 defines the following five types of service flow with distinct QoS requirement:

- Unsolicited Grant Services (UGS): designed to support Constant Bit Rate (CBR) services such as voice applications.
- Real-Time Polling Services (rtPS): designed to support real-time services that generate variable size data packets on a periodic basis, such as MPEG video.
- Non-Real-Time Polling Services (nrtPS): designed to support non-real-time and delay tolerant services that require variable size data grant burst types on a regular basis such as FTP.
- ertPS: derived from UGS and rtPS and designed for service flows of variable size data packets on a periodic basis, such as VoIP with silence suppression.
- Best Effort (BE): designed to support data streams that do not require any guaranteed QoS such as HTTP.

IEEE 802.16 is a centrally controlled protocol but can also operate in Mesh mode. In the former case the BS controls the uplink bandwidth allocation and the SSs request transmission opportunities in the uplink channel. In the Mesh mode, traffic can be routed through SSs and distributed scheduling algorithm is used with one node (SS) taking on the role of the Mesh BS. In the centrally controlled method there are two ways to contend for a transmission opportunity. The first is to transmit in periodic intervals and the second is to contend with the other SSs transmitting request for grants. The BS collects all the requests and therefore has sufficient information about the bandwidth requests. The scheduler can therefore assign an appropriate number of data minislots to accommodate the requests. This information is passed to the SSs through the MAP message, which describes the way the upstream bandwidth is assigned to each SS. The DL and UL subframes are included in the frame. An example of this is shown in Figure

1. In the UL contention period, collisions might occur as when two or more SSs place their request PDUs in the same minislot. Moreover, since the SSs cannot listen directly to the upstream, the correct request will be acknowledged only in the next MAP message. The collided requests are repeated until they are successfully received by the BS. To control such collisions, IEEE 802.16 makes use of a binary exponential back-off algorithm similar to the one used for CSMA-CD in ETHERNET. Because of this contention based access, this protocol cannot guarantee the access delay. IEEE 802.16 takes care of real time applications (VoIP, Video on demand etc.) by assigning unsolicited bandwidth grants and polling. The use of polling is essential because these applications should receive service on an isochronous basis.

Bandwidth allocation in IEEE 802.16 can be made in two ways - either by grant per connection (GPC) or by grant per Service Station (GPSS). In the first case, each grant is associated with a specific connection. In this case, whenever several connections of a SS are polled or granted transmission opportunities, multiple entries are appropriately set in the UL-MAP message. The main disadvantage of this approach is that it creates additional overhead. In the GPSS approach, the SS is given a single grant for all its connections. Then the local scheduler in the SS decides how to allocate the transmission opportunities to each connection. In doing this the SS must respect the QoS requirements of its connections. In both modes the bandwidth requests are issued on a per connection basis.

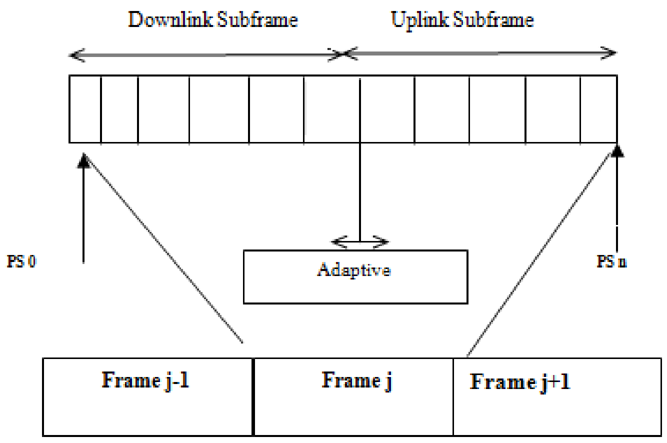

Figure 1: Time Division Duplexing in IEEE 802.16 frame

2.2 Related Work

The bandwidth allocation and scheduling approaches proposed for WiMAX may be classified as homogenous, hybrid or opportunistic algorithms [3]. The first two types use various legacy allocation approaches. For example, the scheduling algorithm of [5] assigns a fixed bandwidth for UGS, uses Earliest Deadline First (EDF) technique for rtPS, Weighted Fair Queuing (WFQ) for nrtPS and equal distribution for BE. A token bucket based conservative approach for CAC is also proposed. These are improved in [6] by a hybrid scheme that uses EDF for SSs of the rtPS class and WFQ for SSs of nrtPS and BE classes. In [7], an adaptive queue-aware algorithm is proposed for uplink bandwidth allocation and rate control mechanisms in a SS for polling services in a GPC system. Here, rate control is also used to limit the transmission rate of the connections under the polling service class so that the overall QoS performance can be controlled. However, [7] treats real-time and non real-time services identically and does not adequately exploit QoS factors like maximum latency in its scheduling. In [4] a queue based scheduling algorithm for real-time and non real-time traffic at the Medium Access Control (MAC) layer is proposed for resource sharing between real-time and non real-time traffic depending on their queue size and latency requirements. Performance of the EDF scheduling algorithm for BWA networks is evaluated in [8]. Tsai et.al [9] propose an uplink scheduling algorithm and a token bucket based Call Admission Control (CAC) algorithm. The scheme in [10] accepts each service based on its Minimum Reserved Traffic Rate and thus cannot satisfy the bandwidth request of real-time VBR services. A TCP aware CAC mechanism is proposed in [11] for packet switched wireless networks, which relies on the elastic behavior of TCP with respect to changing bandwidth conditions for admitting new connections into the system. However, this kind of mechanism is more suited for admission of BE connections which have no strict QoS requirements. In [12]-[13] degradation based CAC have been proposed which aims at better handover service by minimizing the handoff connection dropping probability.

3. Proposed Resource Allocation Algorithm and Connection Admission

We consider a single BS serving multiple connections from SSs through a TDMA/TDD access mode. We use the terms bandwidth and capacity interchangeably as the number of PDUs that can be transmitted in one frame. For modeling convenience we assume all PDUs (packets) to be of the same size. (Extending the proposed schemes for variable PDU lengths can be easily done.) We consider two strategies, Conservative and Non-Conservative, for CAC. An active UGS source s is assumed to generate U_s packets per frame. Each rtPS source specifies both the mean and maximum number of packets it generates per frame, i.e. R_l & R_{lmax} for rtPS source l. Each nrtPS source specifies the mean number of packets it generates per frame, i.e. NR_m for nrtPS source m. Let there be S UGS sources, L rtPS and M nrtPS sources already present in the system and let C be the maximum number of packets that can be carried in a frame.

3.1 Connection Admission

3.1.1 Conservative CAC

This guarantees the QoS for UGS, rtPS and nrtPS traffic. The packet transmission requests of UGS and rtPS are satisfied on a per frame basis while that of nrtPS are satisfied on a long term basis. The conditions for admitting a new rtPS or nrtPS source is given below. (We assume that UGS admission has already been done. No admission control is required for BE connections since they do not get QoS support).

rtPS Source: A new rtPS source (R_{new}, R_{newmax}) is admitted if

$$\sum_{s=1}^{S} U_s + \sum_{l=1}^{L} R_l + \sum_{m=1}^{M} NR_m + R_{new} \le C \qquad (1)$$

and

$$\sum_{s=1}^{S} U_s + \sum_{l=1}^{L} R_{l\max} + R_{new\max} \le C \qquad (2)$$

nrtPS Source: A new nrtPS source (NR_{new}) is admitted if

$$\sum_{s=1}^{S} U_s + \sum_{l=1}^{L} R_l + \sum_{m=1}^{M} NR_m + NR_{new} \le C \qquad (3)$$

3.1.2 Non-Conservative CAC

Here, a rtPS packet need not be carried in the immediate next frame. For the K-frame buffering case, a rtPS packet arriving in frame *j* can be sent in frames *j+1* to *j+K+1* but will be lost otherwise. For this we waive (2) and admit a rtPS flow even if only (1) is met. However, the rtPS application must be one which can tolerate some packet loss, e.g. streamed video or audio. A new nrtPS source must still meet (3) to be admitted.

3.2 Bandwidth Allocation

Once admitted into the system, the connections (UGS/rtPS/nrtPS) are allocated resources in the priority UGS>rtPS>nrtPS. The Conservative CAC ensures that all UGS and rtPS packets arriving during frame *j* will be transmitted in frame *j+1*. To fairly allocate the remaining resource packav (i.e. what is left after UGS and rtPS allocation) among the M nrtPS connections, the following algorithm is used. Here λ_i, *i= 1... M,* packets are the respective demands in the current frame of each nrtPS connection and NR_i, *i =1…M*, are their mean requirements (which are to be met on a long term basis).

<u>Algorithm for Bandwidth Allocation in a frame</u>

1. SET $\lambda_{prop-i} = NR_i$ for $i = 1\ldots M$
2. SET $\lambda_{tot} = \sum_{i=1}^{M} \lambda_{prop-i}$
3. $\lambda_{av-i} = (packav \times \lambda_{prop-i}) / \lambda_{tot}$ for $i = 1\ldots M$
4. $\lambda_{init-i} = \lambda_i$ for $i = 1\ldots M$
5. Check if $packav > 0$ and $\lambda_{tot} > 0$; otherwise GOTO 8
6. SET $packav = 0$ and $\lambda_{tot} = 0$
7. For each $i = 1\ldots M$ if $\lambda_{prop-i} > 0$ and $\lambda_i \le \lambda_{av-i}$ then set $\lambda_{prop-i} = 0$, $packav = packav + (\lambda_{av-i} - \lambda_i)$ and $\lambda_i = 0$ else if $\lambda_{prop-i} > 0$ and $\lambda_i > \lambda_{av-i}$ then set $\lambda_i = \lambda_i - \lambda_{av-i}$; if *i*=*M* GOTO 2
8. END
9. $\lambda_{alloc-i} = \lambda_{init-i} - \lambda_i$ for $i = 1\ldots M$

The above algorithm allocates resources fairly to the nrtPS connections. Since different nrtPS connections have different arrival rates of packets with different means, the allocation of the available resource is done in proportion to their mean requirements. Step 3 assigns each nrtPS connection with its fair share of the available resource, packav, as λ_{av-i}. However, some connections may demand more while some may demand less than their respective fair share in a particular frame, even though they conform to the long term average requirement. The algorithm tackles this by allocating resources unused by connections which demand less to connections which demand more and does this again in a proportionally fair manner. Steps 4-7 do this iteratively until either there is no available resource to allocate or there are no unmet demands left. Note that connection *i* gets bandwidth $\lambda_{alloc-i}$ λ_{all} allocated to it at the end of the execution of the above algorithm.

3.3 Analytical Model

This system is analyzed using a discrete-time Markov chain to calculate the average delay of MPDUs and mean losses, if any, in the system. We consider a PMP network where the generated traffic stays within the network. For allocation by the BS, the individual rtPS and nrtPS queues merge into a global rtPS queue and a global nrtPS queue respectively at the BS. Our analytical model differentiates between the cases where no buffering is allowed for the rtPS packets and one where the rtPS packets may be buffered for one or more frames.

3.3.1 Conservative CAC without Buffering

Here, the rtPS packets coming in a frame are always transmitted in the next frame. Let N_i^{nr} be the number of packets in the global nrtPS queue (i.e. the state of the Markov Chain) at the end of the frame *i*. For frame *i*, let ra$_i$ and rd$_i$ be the number of rtPS packets arriving and leaving, respectively. Let nra$_i$ and nrd$_i$ be the respective number of nrtPS packets arriving and leaving in frame *i* as shown in Figure 2 Then

$$N_{i+1}^{nr} = \max\left(N_i^{nr} - \max(P - ra_i, 0), 0\right) + nra_{i+1} \qquad (4)$$

where P is the total number of rtPS and nrtPS packets that can be carried in a frame (i.e. capacity excluding the UGS packets).

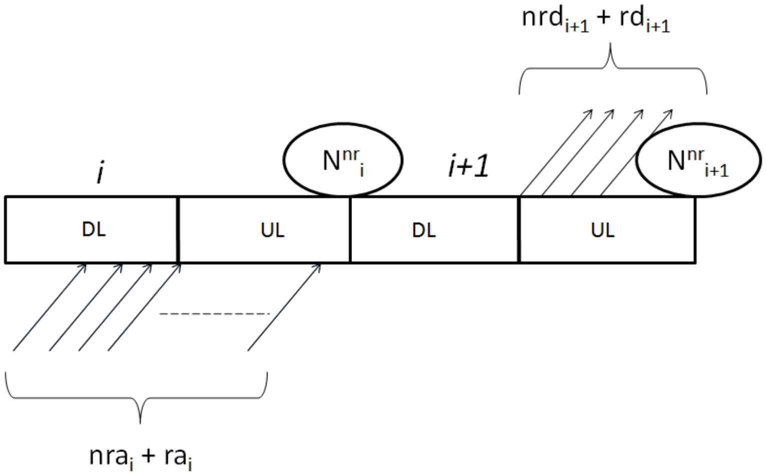

Figure 2: State Transitions with no buffering for rtPS packetss

The state transition probabilities P$_{i,j}$ (for transition from state i to state j) are calculated using the packet arrival probabilities to the nrtPS and rtPS global queues. The equilibrium state probabilities π_j, mean number N$_N$ in the global nrtPS queue and their average delay D$_N$ may then be computed as follows.

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{i,j} \quad \text{and} \quad \sum_{j=0}^{\infty} \pi_j = 1 \qquad (5)$$

$$N_N = \sum_{j=0}^{\infty} j\pi_j \qquad (6)$$

$$D_N = N_N / \lambda_{nr} \qquad (7)$$

Here, $\lambda_{nr} = \sum_{i=1}^{M} NR_i$ is the sum of the mean offered packets per frame by all the M nrtPS connections.

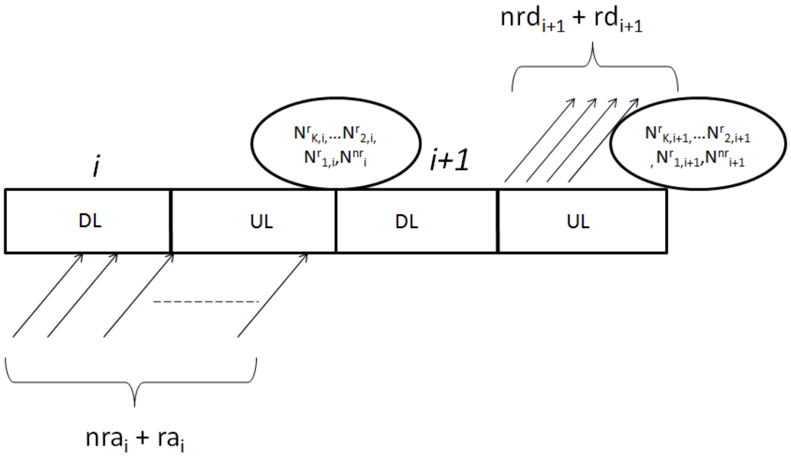
nrd$_{i+1}$ + rd$_{i+1}$
N$^r_{K,i,}$...N$^r_{2,i,}$ N$^r_{1,i,}$N$^{nr}_i$
i
i+1
N$^r_{K,i+1,}$...N$^r_{2,i+1}$,N$^r_{1,i+1,}$N$^{nr}_{i+1}$
DL
UL
DL
UL
nra$_i$ + ra$_i$

Figure 3: Transitions with *K* frame buffering for rtPS packets

3.3.2 Non-Conservative CAC with Finite Buffering

Figure 3 considers the general case where a rtPS packet can be buffered for a maximum of K frames before it is discarded. Let $(N^r_{K,i},......,N^r_{1,i},N^{nr}_i)$ be the system state at the end of frame *i* where N^{nr}_i is the number of MPDUs in the nrtPS queue and $N^r_{l,i}$ are those MPDUs in the rtPS queue which will be discarded if not transmitted within the next *l+1* frames. Following the conventions defined earlier, the state at the end of the *i+1* frame can be found as -

$$N^r_{l,i+1} = \max\left(N^r{}_{l+1,i} - \max\left(0, P - \sum_{n=1}^{l} N^r_{n,i} \right), 0 \right) \quad l = 1,.,K-1$$

$$= \max\left(ra_i - \max\left(0, P - \sum_{n=1}^{l} N^r_{n,i} \right), 0 \right) \qquad l = K \qquad (8)$$

$$N^{nr}_{i+1} = N^{nr}_i + nra_{i+1} - \min\left(N^{nr}_i, \max\left(0, \mathrm{P} - \sum_{n=1}^{K} N^r_{n,i} - ra_i \right) \right) \qquad (9)$$

Given the arrival distribution in a frame for each rtPS and nrtPS queue, the arrival distribution to the global rtPS and nrtPS queues can be found. It can then be used to calculate the transition probabilities of the Markov Chain assuming that service is provided in a FCFS manner for K-frame buffering. Note that K=0 corresponds to the Conservative CAC and K=1 corresponds to the Non-conservative CAC case where the rtPS packets which cannot be sent in their next frame are buffered and tried for at most one more frame. The K=1 Non-Conservative CAC is studied further through subsequent analysis and simulations. Note that for this K=1 case, the state descriptor will be $(N_{1,i}^{r}, N_{i}^{nr})$ as per our earlier notation. The transition probability $P_{ij,kl}$ for the (i,j)→(k,l) transition may be found by considering each transition event. (The details are given in the Appendix). The equilibrium state probabilities $\pi_{n,m}$ are computed as follows.

$$\pi_{n.m} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi_{i,j}P_{ij,nm} \qquad \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\pi_{n,m} = 1 \qquad (10)$$

We compute Loss_R as the average number of packets discarded per frame for rtPS packets and the mean delay D_N for nrtPS packets as follows.

$$Loss_R = \sum_{i=P+1}^{\infty}\sum_{j=0}^{\infty}(i-P)\pi_{i,j} \qquad (11)$$

$$N_N = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} j\pi_{i,j} \qquad (12)$$

$$D_N = \frac{N_N}{\lambda_{nr}} \qquad (13)$$

4. Analytical and simulation results

The proposed CACs and resource allocation algorithm have been studied through analysis and simulations. For simulations, a discrete-event simulator was run until statistically significant results have been obtained. To simulate the two types of scenarios mentioned in the previous section we implemented a discrete event simulator in MATLAB. The number of arrivals to each of the rtPS and nrtPS queues in each frame was uniformly distributed and their arrival times were also uniformly distributed throughout the frame period. The emptying of rtPS and nrtPS queues in succeeding frames was done using the resource allocation algorithm as developed in the previous sections. The simulator works in a gated manner, packets arriving in frame *f* can be transmitted no earlier than frame *f+1*. Statistically significant results from the simulator were ensured by running the simulator for sufficient number of frames of the order of 10^6.

Our studies focus on the rtPS and nrtPS flows as UGS traffic gets a fixed number of packets in the UL of each frame and the BE traffic gets essentially what remains unused in UL. (We did not consider ertPS traffic in our model.) We look at the average packet delay for the nrtPS traffic for both types of CAC. For the Non-Conservative CAC, we present the results on the packet loss for rtPS traffic. For analytical computations, the infinite Markov Chain of Section 3 is truncated to a finite one which is large enough so that increasing it further does not change the numerical results significantly.

4.1.1 Conservative CAC

We simulate a WiMAX PMP scenario with single BS and variable number of SSs. Each SS is associated with a single type of connection (UGS/rtPS/nrtPS) on the UL. The total number of packets that can be carried in a UL subframe is set at 60 with the 10 ms frame divided equally between up-link and down-link. The total number of requests from all UGS connections is fixed at 15 packets per frame so that a maximum of 45 rtPS and nrtPS packets can be carried in a frame. The nrtPS and rtPS packets are assumed to be of equal length of 100 bytes at UL data rate of 10 Mbps. The number of rtPS connections is fixed at 4 with each requesting a mean of 5 packets per frame. Two types of rtPS sources are studied: Type 1 with packets arriving per frame distributed as ~U[1,9] and the less bursty Type 2 with ~U[4,6]. (~U[a,b] is the uniform distribution over (a, b).) Packets arrive from each nrtPS source with distribution ~U[0, 4] and the number of nrtPS connections are varied to study the system with varying load. Note that the uniform distribution was chosen for numerical convenience. Any distribution can be chosen for each individual rtPS/nrtPS source and the arrival distribution to the respective global queues can be analytically found. Figure 4 shows the average delay for nrtPS sources with varying traffic intensities (rtPS and nrtPS) normalized to the capacity available for rtPS and nrtPS (i.e. 45 in this case). This shows that, for the same offered traffic, the average nrtPS delays are higher when the rtPS traffic is burstier. It may also be noted that the analytically obtained results are very close to the simulation results with some deviations only at very high traffic loads.

We further plot in Figure 5 the average delays of nrtPS packets with rtPS sources (Type3 and Type4) having the same variance as the earlier case but with increased mean arrival rate at 10 packets per frame (~U [6, 14] and ~U[9, 11] respectively). Similar dependence of average nrtPS delays as in the earlier case on the variance of rtPS sources was observed in this case. This shows that the variance of the rtPS traffic has a dominant effect on the nrtPS delays irrespective of the mean at a given intensity of traffic.

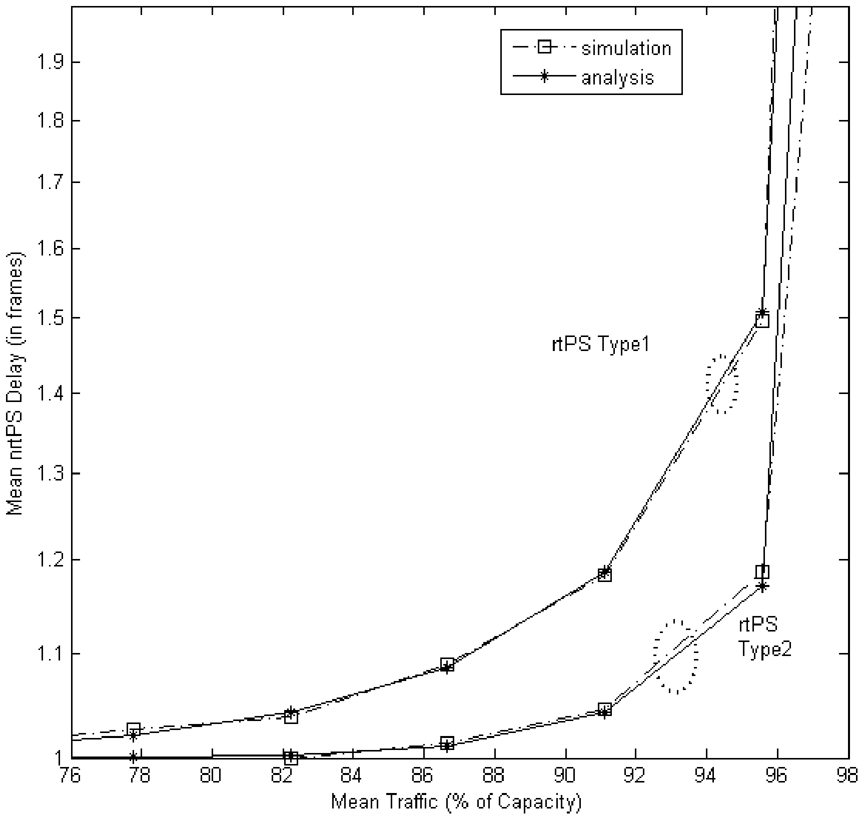

Figure 4: Average nrtPS Delays with Increasing Traffic Intensity

4.1.2 Non-Conservative CAC

We consider the same system as described earlier where in each frame, a nrtPS source generates packets with distribution ~U[0, 4] whereas a rtPS source generates packets with ~U[0, 10]. We consider two cases (a) no frame buffering (i.e. *K*=0, the rtPS packets that cannot be carried in the next frame are lost) and (b) one frame buffering (i.e. *K*=1, rtPS packets can be buffered and tried for one more frame). The number of nrtPS sources is fixed at 4 while the number of rtPS sources is varied as per what is allowed by this CAC. Note that the maximum rtPS traffic may be higher than the capacity in this case as the CAC only takes the mean traffic into account. Figure 6 shows the average nrtPS delays for *K*=0 and 1 with varying normalized maximum rtPS traffic. As expected, increasing the rtPS traffic increases the nrtPS delay since system load increases. The nrtPS delays are higher when the frame buffering of rtPS is increased (i.e. from *K*=0 to *K*=1). This occurs because the one frame buffering case allows more rtPS traffic to be carried by the system – part of which may otherwise be lost when *K*=0.

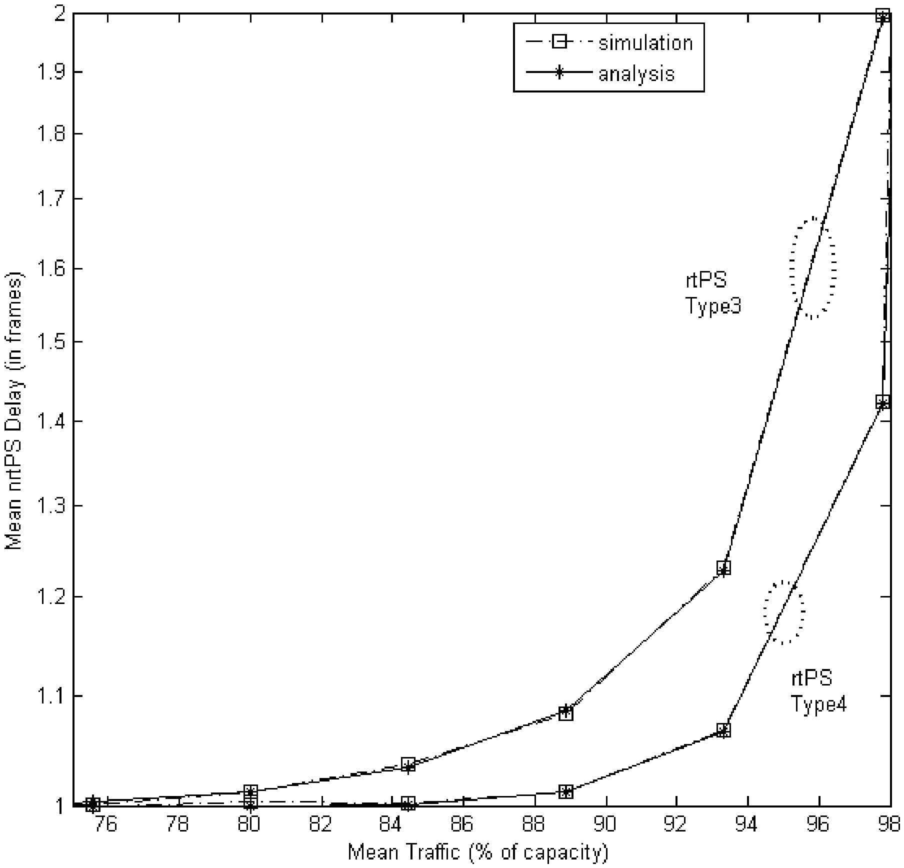

Figure 5: Average nrtPS Delays with increasing Traffic Intensity with higher rtPS means

Our results also show that as the ratio of maximum rtPS traffic to maximum available capacity (after taking care of the fixed UGS traffic) ranges from 1.1 to 1.6, the rtPS mean packet loss percentages range from approximately 10^{-2} to nearly 1 for *K*=0 but from merely 10^{-14} to 10^{-6} for *K*=1 as shown in Figure 7. It is clear that the packet loss is drastically reduced by merely incorporating one-frame buffering in the Non-Conservative CAC. This makes this CAC a practical choice in a real system where the rtPS channels are being used for multimedia applications. With just one frame buffering, the Non-Conservative CAC has a low enough packet loss to be acceptable for typical multimedia applications but will allow the system to accommodate more rtPS sources than the Conservative CAC. The one frame buffering does increase the mean rtPS delay. However, this will be at the most one frame period more and may be tackled easily by adjusting the playback delays at the respective destinations. This is shown in Figure 7 which plots the mean excess rtPS delays – this is the mean delay excluding the inherent one frame delay.

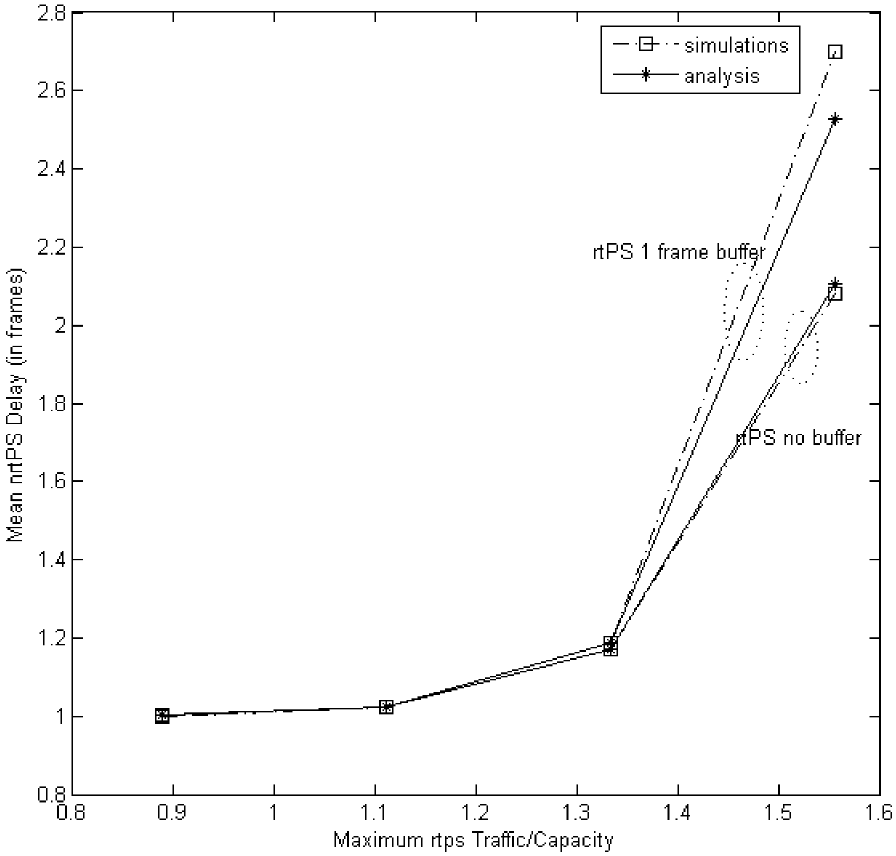

Figure 6: **Average nrtPS Delays with varying Maximum rtPS Traffic**

We use our simulator to plot a contour of operating points i.e. number of nrtPS and rtPS connections that can exist in the system under different loss constraints on real time packets. In these plots a rtPS source is ~U[1, 9] and a nrtPS source is ~U[0, 4] with system capacity being 45 packets (nrtPS and rtPS) per UL subframe. The given system can have a maximum of 9 rtPS sources and 22 nrtPS sources at any given time. Figure 8 plots the contour when no buffering is allowed for the rtPS packets for different loss constraints. It is important to note that the contours are plotted with maximum allowed loss to be 0%, 1%, 5% and 10% in each case. The 1 and 2 frame buffering cases are plotted in Figure 9. The allowed losses in both the cases are limited to 1%. These plots show that increasing the allowed loss and allowing more buffering for the real time traffic, allows more rtPS sources to be admitted into the system leading to better system utilization. Since the loss is still within acceptable limits (i.e. loss≤5% or 10%) the system will still be a practical one to use for real-time multimedia traffic.

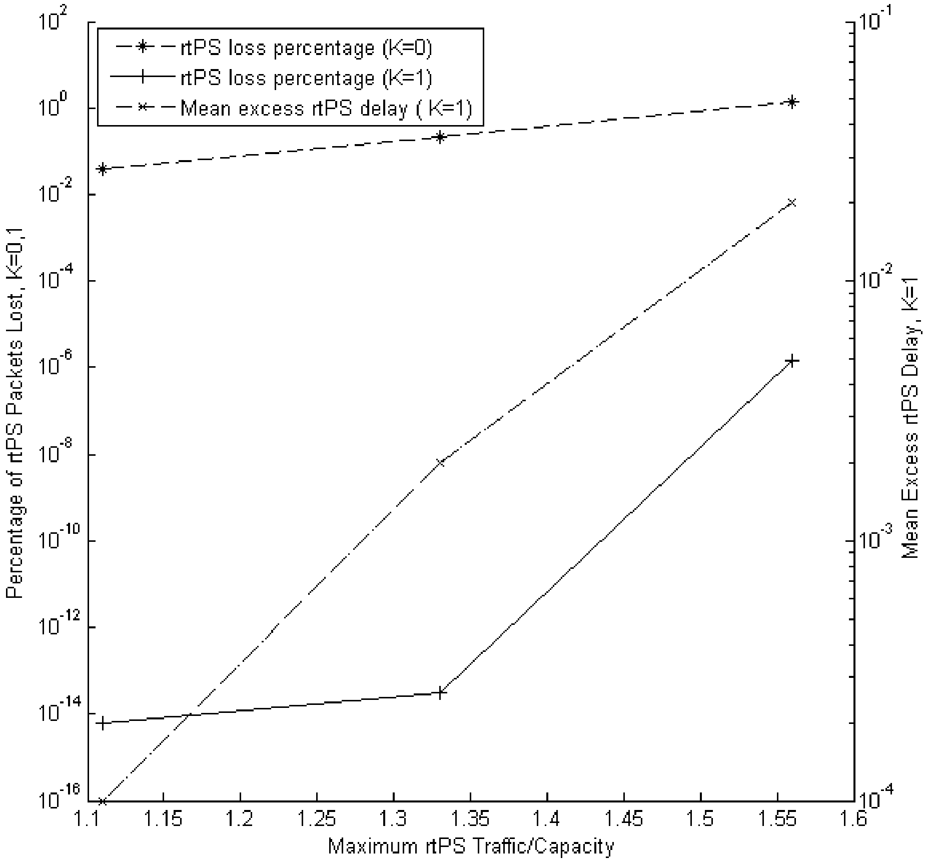

Figure 7: Mean rtPS Excess Delays and Packet Loss Percentage with varying Maximum rtPS Traffic

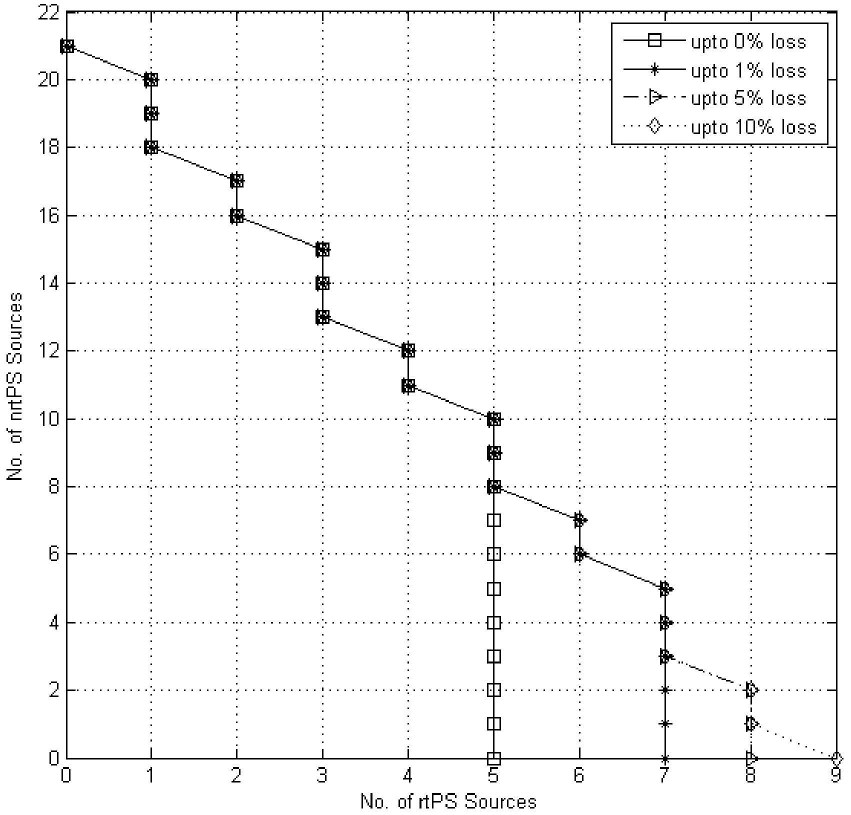

Figure 8: Contour of Operating Regions for Specified Packet Loss when no frame buffering is allowed

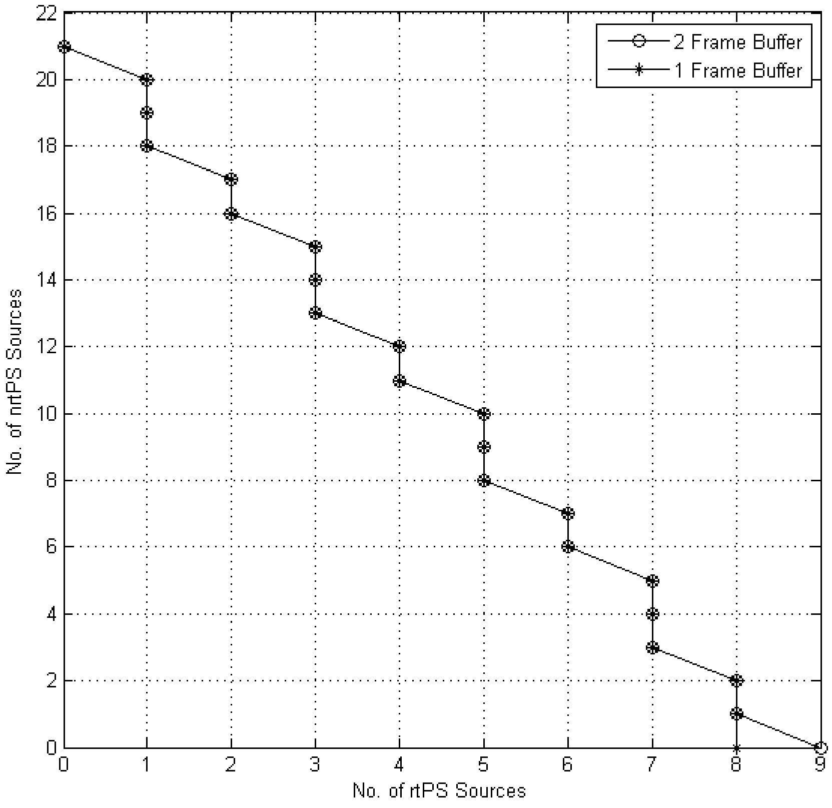

Figure 9: Contour of Operating regions for up to 1% loss where variable buffering is allowed

5. Bandwidth Allocation and Connection Admission for rtPS sources with differential QoS requirements

Our earlier model assumed all rtPS sources in the network to be of the same priority and with similar QoS requirements. In this section, we consider the operation of our proposed approach when we differentiate the QoS requirements of rtPS sources on the basis of the maximum delay the respective sources can tolerate. This situation is expected to arise when the sources represent different VBR sources with varying QoS requirements

5.1 Bandwidth Allocation

We consider a system which has N different types of rtPS sources differentiated on the basis of the maximum *excess delay**[1] that they can tolerate. We label the sources from 0 to N-*1*. Unlike the previous case (where we assumed a common global queue for all the rtPS sources), this system will have J separate queues Q_0 to $Q_{J\text{-}1}$ ($J \leq N$) where the new arrivals from the corresponding N sources would enter, where the packets entering Q_i are from rtPS sources which can tolerate upto i frames of excess delay. This has been shown in Figure 10. A particular queue Q_i is considered to be subdivided into $i+1$ parts P_0 to P_i. The packets in P_0 of each queue are those which cannot tolerate anymore delay and will be discarded if not transmitted in the current frame. Similarly packets belonging to P_l of each queue can be transmitted over the next l frame and will be discarded thereafter.

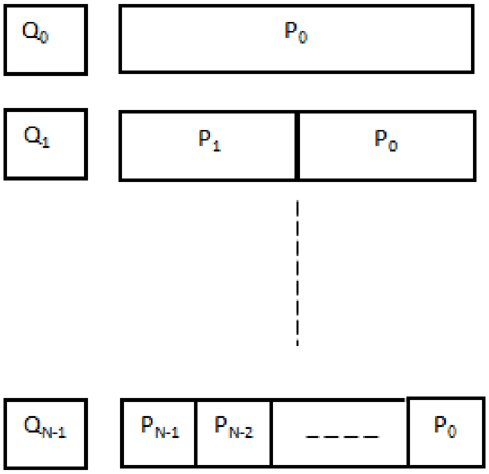

Figure 10: Priority Structure

Based on the above formulation, the packets arriving from rtPS sources which can tolerate up to K frames of excess delay will enter the queue Q_K and will be kept in P_K

[1] *Excess delay is the total delay a packet suffers excluding the inherent* 1-*frame delay.*

from next frame. After each subsequent frame, the packets not transmitted in each queue will be transferred from P$_{l}$ to P$_{l\text{-}1}$ for l=2 to K. The bandwidth allocation to each queue is done as per each sub-part of each queue – i.e. in the order of first to the packets of P$_0$ of each Q$_i$ (in the order of Q$_0$ to Q$_N$), then to P$_1$ of each Q$_i$ (in the order of Q$_0$ to Q$_N$) and continuing in similar fashion until either all packets in all queues are served or the bandwidth available for allocation is exhausted. To handle the situation where rtPS sources at the same priority level have different mean arrival rates, we can divide the available bandwidth to a particular P$_i$ of a Q$_k$ in proportion of the mean requirements of the individual connections.

5.2 Connection Admission

Following the notation of our earlier analysis, each rtPS source is classified on the basis of the mean and maximum number of packets it generates per frame, i.e. R$_l$ & R$_{l\ \max}$ for rtPS source *l*. Let there be *S* UGS sources, *L* rtPS and *M* nrtPS sources currently present in the system and let *C* be the maximum number of packets that can be carried in a frame. Out of the *L* rtPS sources let *l*=1 to *L1* indexed rtPS sources be the one which have the maximum excess delay requirement of zero frames. Then for a new rtPS source (R_{new}, $R_{new\ max}$) to be admitted, the following must be satisfied

$$\sum_{s=1}^{S} U_s + \sum_{l=1}^{L} R_l + \sum_{m=1}^{M} NR_m + R_{new} \le C \qquad (14)$$

In addition, if the rtPS source to be admitted is one with a zero excess delay requirement, then the following condition should also be satisfied.

$$\sum_{s=1}^{S} U_s + \sum_{l=1}^{L1-1} R_{l\max} + R_{new\max} \le C \qquad (15)$$

In essence, we propose using conservative CAC for highest priority rtPS sources (with zero excess delay requirements) and non-conservative CAC for the other (lower priority) rtPS sources which can tolerate higher excess delays.

5.3 Performance analysis and Simulations Results

5.3.1 Variation with traffic load

For our performance studies, we consider a system with three types of rtPS sources, i.e. ones which can tolerate either 0 (*type* 0), 1 (*type* 1) or 2 (*type* 2) frames of excess delay. Assume that the maximum number of packets of rtPS and nrtPS sources (excluding the UGS) that can be carried in a frame is fixed at 75. Consider a system with 2 *type* 0 rtPS sources, 3 of *type* 1 and 4 of *type* 2 rtPS sources. Each *type* 0 rtPS source has a mean arrival rate of 15 packets per frame with distribution ~U [0, 30]. The mean arrival rate of *type* 1 and *type* 2 rtPS sources are varied according to what is allowed by the CAC. In Figure 11 and Figure 12, the varying traffic intensities (normalized traffic load) are obtained by varying the distribution of *type* 1 rtPS sources from ~U[0,10] to ~U[3,13] and that of *type* 2 rtPS sources from ~U[0,4] to ~U[3,7]. Figure 11 shows the mean excess delay of rtPS sources of *type* 1 and *type* 2 with varying traffic intensities where the ordered pair on the plots shows the respective mean arrival rates of respective sources at particular traffic intensity. As seen in this figure, there is evident differentiation between the excess delays of *type* 1 and *type* 2 rtPS sources in accordance with the priority structure assigned in bandwidth allocation. Figure 12 shows the rtPS packet loss percentages with varying traffic intensities (normalized traffic load). Packet losses for *type* 2 rtPS sources are higher because of their lower priority (as compared to *type* 1 packets) in bandwidth allocation. The loss percentages at lower intensities are zero and hence cannot be shown in the plot. (It may be noted that *type* 0 packets will not suffer any loss as the corresponding sources are admitted based on the *conservative CAC*.)

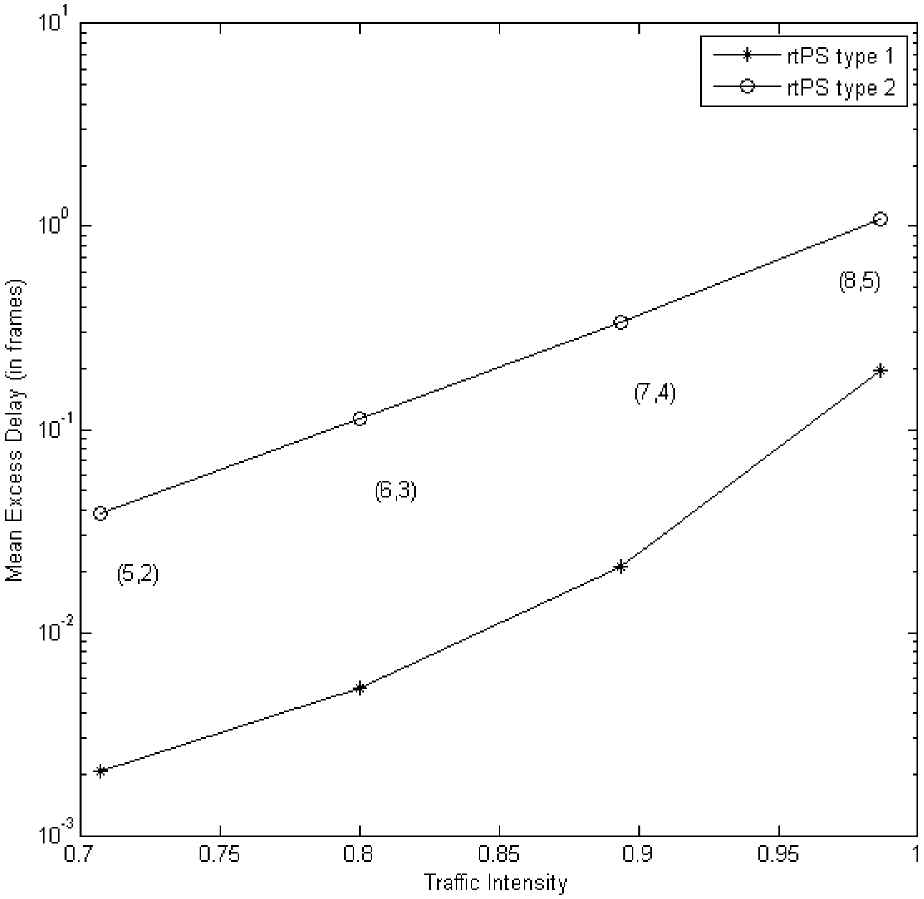

Figure 11 : Mean Excess Delays with varying Traffic Intensity in the System

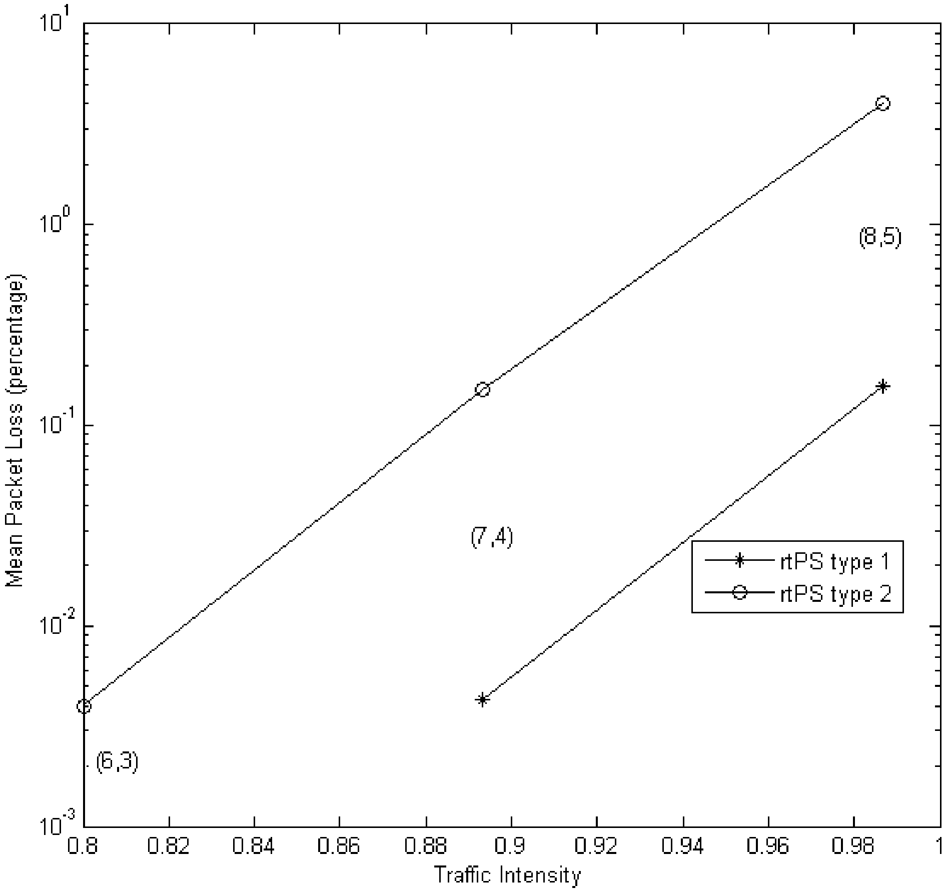

Figure 12 : Mean Packet Loss (percentage) with varying Traffic Intensities

5.3.2 Variation with burstiness of high priority traffic

We again consider a system where there are three types of rtPS sources- sources which can tolerate 0 (*type* 0), 1 (*type* 1) and 2 (*type* 2) frames of excess delay respectively. In this case, we assume that the maximum number of packets of rtPS and nrtPS sources (excluding the UGS) that can be carried in a frame is 55. Consider a system with two *type* 0 rtPS sources, three of *type* 1 and four of *typ*e 2 rtPS sources. Both *type* 0 and *type* 2 rtPS sources have a mean arrival rate of 5 packets per frame. *Type* 1 and *Type* 2 rtPS sources have the arrival distributions ~ U [0,16] and ~U[3,7] respectively, while for *type* 0 sources the packet arrival distribution is ~U[5-*c*,5+*c*] where *c* is varied in our simulations. Figure 13 shows the mean excess delays of *type* 1 and *type* 2 rtPS sources with varying variance of the *type* 0 rtPS sources. As the rtPS *type* 0 sources become burstier, the excess delays for other type of rtPS sources tend to increase. Figure 14 shows the rtPS packet loss percentages with varying variance of *type* 0 rtPS sources.

Packet losses for *type* 2 rtPS sources are higher because of the lower priority assigned to them while allocating the bandwidth, as compared to *type* 1 packets.

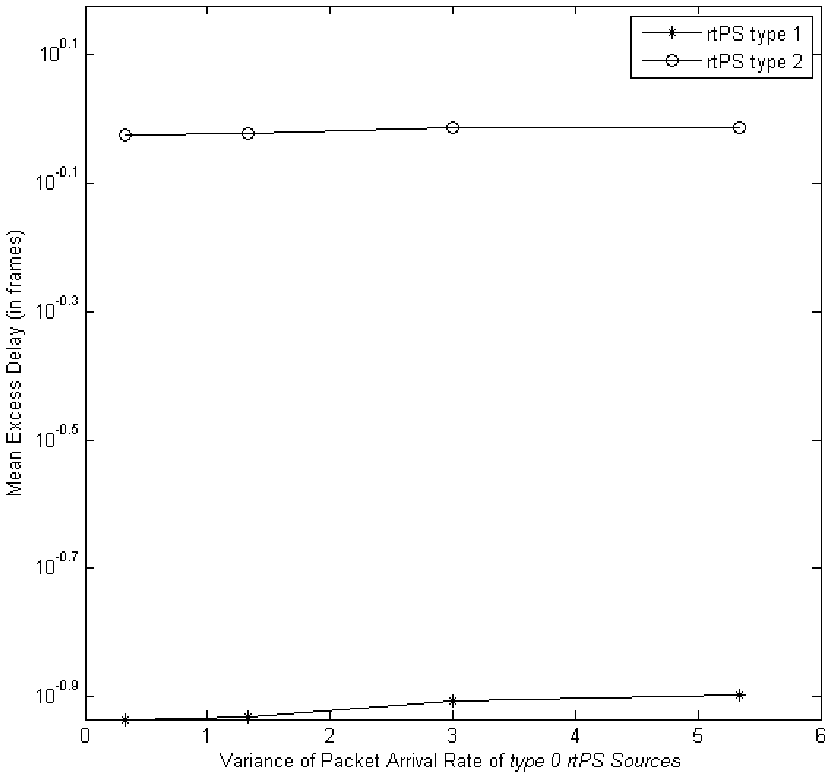

Figure 13 : Mean Excess Delays with varying Burstiness of type 0 rtPS Traffic

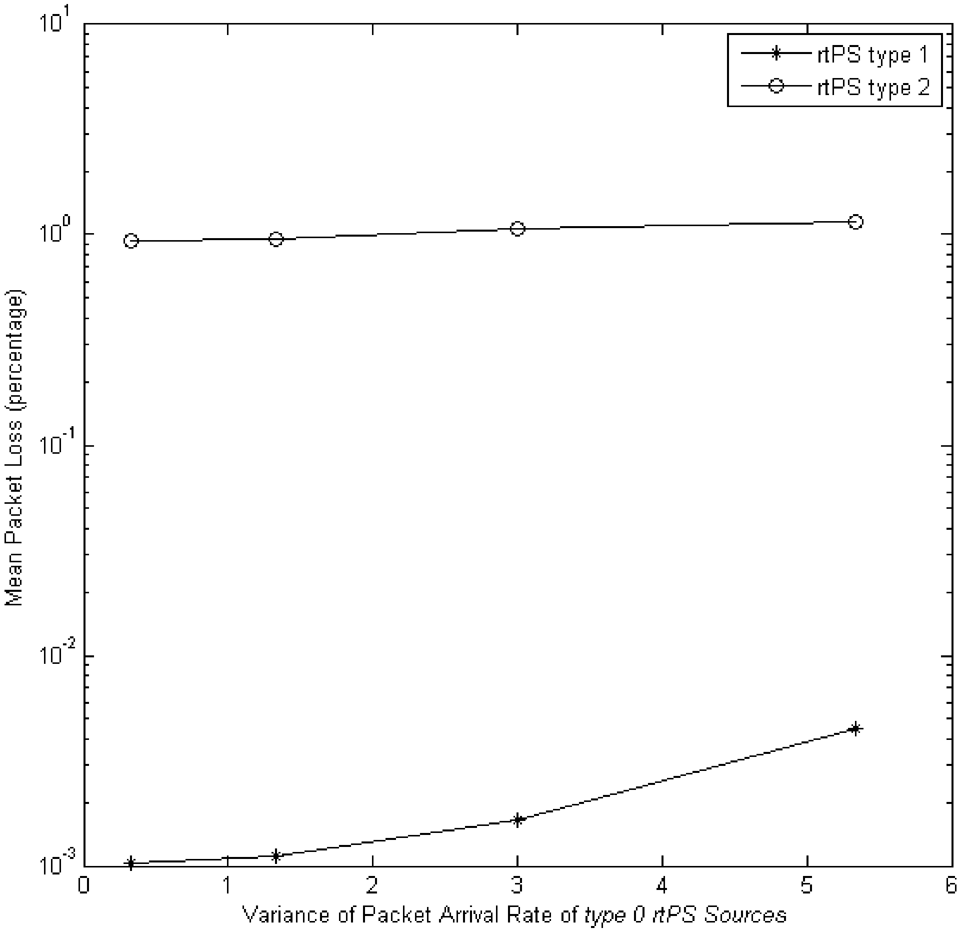

Figure 14 : Mean Packet Loss (percentage) with varying Burstiness of *type* 0 rtPS Sources

5.3.3 Performance analysis with a MPEG video source model

As a MPEG video source is a classic example of real time multirate video source we tried to test our CAC and BA with a simulated MPEG source model. This was done by feeding MPEG video packets into the proposed system where these were generated following the model proposed in [14]. This video source model is summarized below.

5.3.3.1 MPEG Model

According to the video traffic model of [14], the compression pattern for encoding is fixed as shown in Figure 15.

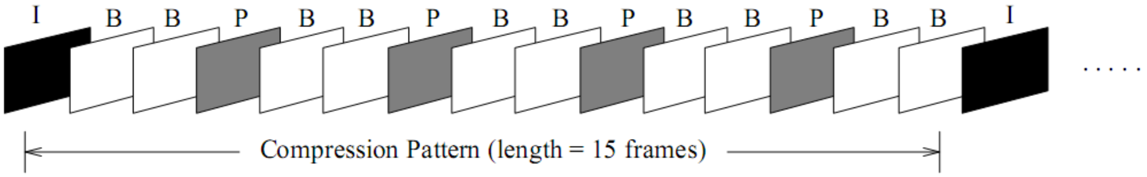

Figure 15: The Compression Pattern used to Generate the Video Stream

The length of each of I, P or B frames (intra-frames, predictive or bidirectional respectively) is modeled as a lognormal distribution (in terms of ATM cells). The corresponding parameters of the lognormal distribution are given in Table 1. During a scene length the same realization of the length of I frame was used. The length of a scene is given as a gamma distribution in terms of I frames with mean of ~10 I frames. For our model in order to limit the maximum length in terms of number of packets, the corresponding lognormal distribution was truncated at a point such that the probability of frame length being greater than the truncation point is of the order of 10^{-2}.

Frame Type	μ	σ
I	6.0188	0.4566
P	3.7380	0.5961
B	2.8687	0.2657

Table 1: The Parameter of the Lognormal Distribution

5.3.3.2 Simulation Results

In order to fit the MPEG packets into our scheme we treated packets generated from an I frame as the highest priority, the packets from P frames with next highest priority and the packets from B frames with the lowest priority. The lowest priority packets (i.e. from the B frames) were discarded whenever they exceeded two frames of excess delay. Packets from the P frames were discarded whenever they exceeded one frame of excess delay. Conservative CAC was used for the I frame packets guaranteeing that they were carried in the system without loss (if they are admitted by the CAC).

The simulation setup was similar to our earlier settings. A total of about 65 MPEG video packets can be carried in an uplink frame. The number of video sources was varied as allowed by the CAC from 3 to 6 and corresponding losses (if any) and mean excess delays were noted at varying traffic intensities. The excess delay of P frame packets and B frame packets with varying video traffic intensity for this system has been shown in

Figure 16. The maximum percentage packet losses of P frame and B frame under the traffic intensity variation were seen to be 0.0001% and 3% respectively. No losses of packets belonging to I frame occurred due to the highest priority assigned to them for resource allocation. This simple model effectively illustrates that a priority scheme like the one proposed by us is a good candidate for carrying MPEG video traffic in the proposed system.

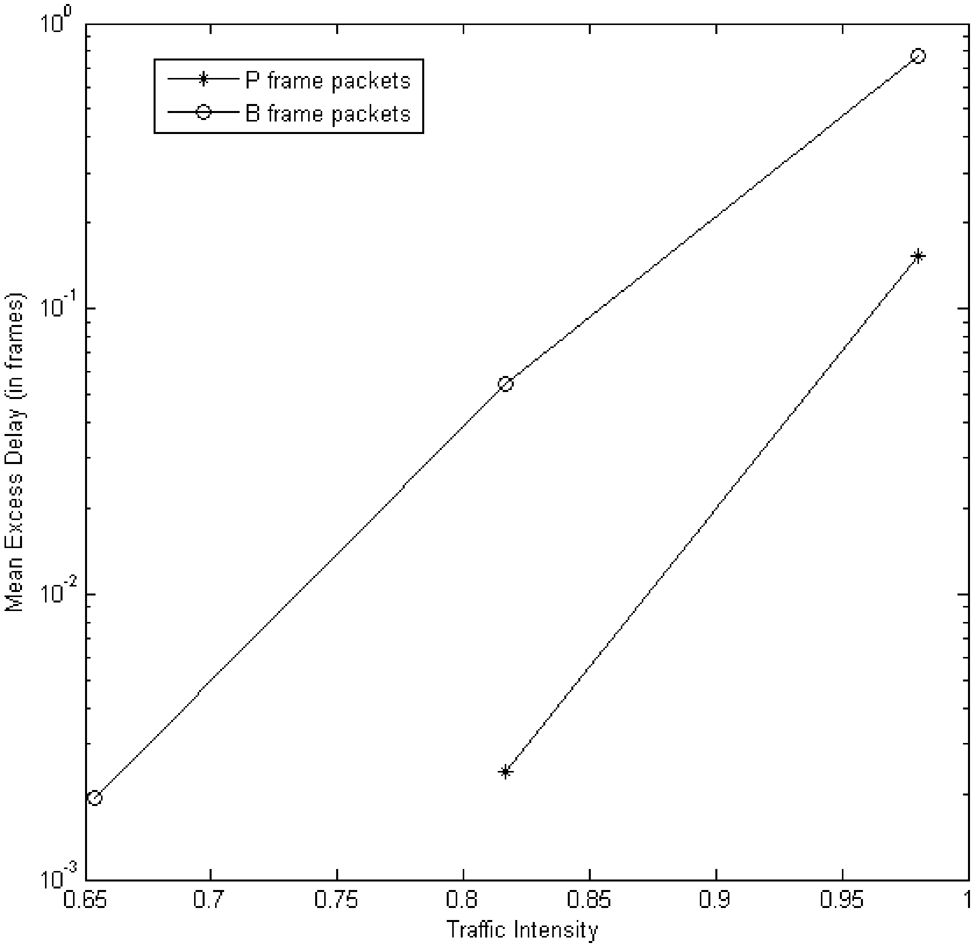

Figure 16 : Mean Excess Delays of P and B Frame Packets with varying Traffic Intensities

6. Conclusion

We have analyzed the proposed Connection Admission Control (CAC) and Bandwidth Allocation (BA) mechanisms using a discrete Markov chain approach and simulations. The analysis results are very close to the simulation results. These not only verify our analysis but also validate the simulation model. We have also proposed a resource allocation scheme that can provide better resource allocation with fairness among non real-time traffic streams proportional to their respective QoS requirements. Packet level performances like average delays and loss rate are also shown for different mixes of traffic combinations and for the two different versions of CAC proposed. The results showed that better network resource utilization can be achieved if the delay and loss tolerances for real-time traffic streams are relaxed. The Non-Conservative Connection Admission Control proves to be better in terms of overall network resource utilisation with only minor tradeoff on the delay and loss performance. We have also proposed an extension to the resource allocation and connection admission algorithms by incorporating real time sources with differential QoS requirements. Simulations results show that the with the outlined approach of BA and CAC the QoS requirements of individual rtPS sources can be met with small loss percentages and high network utilisation. The proposed scheme is validated by feeding with a model of an actual VBR source.

7. References

[1] C. Eklund et al., “IEEE Standard 802.16: A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access,” IEEE Commun. Mag., Jun. 2002, pp. 98–107

[2] IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems," IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001), pp. 1-857, 2004

[3] P. Dhrona, “A Performance Study of Uplink Scheduling Algorithms in Point to Multipoint WiMAX Networks,” M.S. thesis, School of Computing, Queens Univ., Kingston, ON, 2007

[4] Raghu, K.R., Bose, S.K., Maode Ma, "Queue based scheduling for IEEE 802.16 wireless broadband," Information, Communications and Signal Processing, 2007 6th International Conference on , vol., no., pp.1-5, 10-13 Dec. 2007.

[5] K. Wongthavarawat and A. Ganz, “Packet Scheduling for QoS support in IEEE 802.16 broadband wireless access systems,” Intl. J. Communication Systems, vol. 16, Issue 1, Feb 2003, pp. 81-96

[6] K. Vinay, N. Sreenivasulu, D. Jayaram, and D. Das, "Performance evaluation of end-to-end delay by hybrid scheduling algorithm for QoS in IEEE 802.16 network," Proc. of IFIP Intl. Conf. Wireless and Optical Communications Networks, 2006

[7] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks," IEEE Trans. on Mobile Computing, vol. 5, Issue 6, June 2006, pp. 668-679

[8] N. Ruangchaijatupon, L. Wang, and Y. Ji, ”A study on the performance of scheduling schemes for broadband wireless access networks,” in Proc. International Symposium on Communications and Information Technology (ISCIT’06),Oct. 2006.

[9] T.Tsai, C.Jiang and C.Wang, “CAC and Packet Scheduling Using Token Bucket for IEEE 802.16 Networks”, Journal of Communications, vol.1, no2., pp.30-37, May 2006.

[10] J. Sun, Y. Yau, and H. Zhu, Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems. In Proceedings of the IEEE 63rd Vehicular Technology Conference, pp. 1221-1225.

[11] Xinbing Wang, Do Young Eun and Wenye Wang, “A Dynamic TCP-Aware Call Admission Control Scheme for Generic Next Generation Packet-Switched Wireless Networks,” IEEE Transactions on Wireless Communications, vol.6, no.9, Sept. 2007, pp.3344 – 3352

[12] Yin Ge and Geng-Sheng Kuo, “An Efficient Admission Control Scheme for Adaptive Multimedia Services in IEEE 802.16e Networks,” Proceedings of IEEE 64th Vehicular Technology Conference, VTC-2006 Fall, pp. 1–5, Sept. 2006.

[13] Liping Wang, Fuqiang Liu, Yusheng Ji and Nararat Ruangchaijatupon, “Admission Control for Non-pre-provisioned Service Flow in Wireless Metropolitan Area Networks,” Proceedings of Fourth European Conference on Universal Multiservice Networks, ECUMN ’07, pp.243–249, Feb. 2007

[14] M. Krunz and H. Hughes, “ A Traffic Model for MPEG-Coded VBR Streams” Proceedings of ACM SIGMETRICS Conference on the Measurement and Modeling of Computer Systems, pp 47-56, 1995

8. APPENDIX

Computer Algorithm for constructing Markov Chain and calulating state probabilities based on packet arrival distributions for 1 frame buffering for rtPS packets:

1. State Space for global rtPS and nrtPS queues are initialized as (*R*, *S*)
2. For each state (*i, j*), for $i \in (0,1,2,.....,R)$ and $j \in (0,1,2,.....,S)$, arrivals (*n*,*m*) to the rtPS and nrtPS queues are given by the arrival probability distributions P_{rt} and P_{nrt}. As an example, for uniform rtPS arrival distribution: ~U[*a*,*b*], P_{rt} (*n*)= 1/(*b*-*a*) for $a \le n \le b$.
3. The next state at the beginning of UL frame boundary, is determined by current state (*i, j*), arrivals *n, m* and the total number of MPDUs that can be dequeued i.e. *C* . The next state is determined as:

$$Nxtst\_rtPS = \min\left(\max\left(n - \max\left(0, C - i\right), 0\right), R\right)$$

$$Nxtst\_nrtPS = \min\left(j + m - \min\left(j, \max\left(0, C - i - n\right)\right), S\right)$$

4. The state transition probability from (*i, j*) to (*Nxtst_rtPS*, *Nxtst_nrtPS*) is given by $p_{ij,Nxtst-rtPS\ Nxtst-nrtPS} = P_{rt}(n) \times P_{nrt}(m)$
5. For each such state (*i, j*) all such state transitions possible are scanned and from the corresponding transition probabilities, a transition probability matrix is constructed $P_{tr} = \begin{pmatrix} p_{00,00} & \cdots & p_{00,RS} \\ \vdots & \ddots & \vdots \\ p_{RS,00} & \cdots & p_{RS,RS} \end{pmatrix}$
6. Then equilibrium state probability vector is $\pi = \left[\pi_{00}\ \pi_{01} \cdots \pi_{RS}\right]$. The Markov chain equation can be written as $\pi = P_{tr}\pi$ which gives a system of linear equations that can be solved with the normalizing conditionto get the respective state probabilites.
   The Markov chain equations are of the form

$$\begin{aligned}
&\pi_{00} = p_{00,00}\pi_{00} + p_{01,00}\pi_{01} + \ldots + p_{RS,00}\pi_{RS} \\
&\pi_{01} = p_{00,01}\pi_{00} + p_{01,01}\pi_{01} + \ldots + p_{RS,01}\pi_{RS} \\
&\vdots \\
&\pi_{R-1S-1} = p_{00,R-1S-1}\pi_{00} + p_{01,R-1S-1}\pi_{01} + \ldots + p_{RS,R-1S-1}\pi_{RS} \\
&1 = \pi_{00} + \pi_{01} + \ldots + \pi_{RS}
\end{aligned}$$