Performance Guarantees on ATM Networks

Cheng Tang, Shree Murthy, Darrell D. E. Long[†]
Baskin Center for Computer Engineering & Information Sciences
University of California
Santa Cruz, CA 95064

ABSTRACT Recent developments in ATM technology has made multiplexing of a wide range of traffic with diverse performance requirements important. The statistical multiplexing of several traffic types such as voice, video and data can lead to network congestion thus violating the quality of service (QOS) guarantees. In order to meet the performance requirements of the applications an efficient transport protocol and congestion control mechanism becomes necessary.

We propose a transport mechanism for guaranteeing application required QOS requirements over an ATM network. Leaky-bucket congestion control scheme is used as the traffic policing mechanism. Source characterization of voice and video are modeled as Markov Modulated Poisson Processes (MMPP) and the proposed transport mechanism is evaluated by simulation on a single and multihop network with multimedia traffic.

The performance of the scheme is examined as a function of source characteristics and the effect of statistical multiplexing is also investigated. The results indicate that the leaky-bucket scheme and statistical multiplexing are more effective for bursty traffic. This calls for a dynamic control mechanism to achieve optimal system utilization with guaranteed QOS.

1 Introduction

Advances in high-speed packet-switching and fiber optic technologies have led to the development of Broadband Integrated Services Digital Network (B-ISDN) with transmission speeds ranging from hundreds of Mb/s to even several Gb/s and bit-error rate from 10^{-6} to less than 10^{-9} [**?**]. Unlike traditional networks, B-ISDN networks based on asynchronous transfer mode (ATM) technology are required to support traffic generated by a wide range of applications. These applications will have very diverse traffic characteristics ranging from bursty, variable-rate sources, such as voice and variable-rate coded video, to smooth, constant bit rate sources [**?**] and have diverse performance requirements. In ATM based networks, congestion, defined as the phenomenon during which the networks can not satisfy the quality of service (QOS) for some sources, becomes more important than bit errors rates. The need to provide end-to-end QOS guarantees while still taking advantage of the resource gains offered by a statistically multiplexed transport mechanism remains an important, yet largely unsolved problem [**?**].

The importance of congestion control in ATM networks is manifested in a large body of literature [**?**, **?**, **?**, **?**]. Some of these mechanisms include *moving window* [**?**], *jumping window* [**?**], and *leaky bucket* [**?**, **?**]. These mechanisms are all classified as Preventive Congestion Control (PCC) schemes. They first run an admission control algorithm to assure the guaranteed end-to-end QOS to both an arriving connection and *all* existing connections. The admission algorithm takes *source characteristics* and the QOS as input provided by the user, as well as the current network state such as loading and congestion, and estimates the so-called *equivalent bandwidth* [**?**] for an "affordable" connection. After a connection has been established, a *traffic policing function* such as leaky bucket is employed to ensure that the actual traffic adhere to their value agreed during the connection set up.

Several difficulties exist for the PCC schemes to achieve maximum flexibility and economical utilization of the network resource. First, the source traffic is too complicated to be described completely by *Markov Modulated Poisson Process* (MMPP) models. Second, while traffic at the access point of the network may be reasonably well-approximated by such models [**?**], it is still unknown whether this is also true for a connection's traffic when it is deep within the network, having passed through several multiplexors. Third, characteristics of some traffic are not known until the actual communications take place and the possibility for the user to intentionally misbehave still exists.

We propose a general purpose transport protocol which can support multimedia traffic such as data, voice, and

[†] Supported in part by the Office of Naval Research under Grant N00014–92–J–1807.

video over ATM based networks. To ensure that the application required QOS parameters namely, cell loss, delay and delay jitter are satisfied, admission and leaky bucket congestion control mechanism is enforced at the media access control (MAC) layer. We present our transport layer in § 2 and discuss source traffic characterization for data, voice, and video in § 3. A simulation model which aims at evaluating performance of the leaky bucket mechanism in a multihop network with multimedia traffic is developed in § 4. This simulation model allows us to investigate the impact of various source characteristics and statistical multiplexing on the leaky-bucket scheme. We discuss simulation results and performance in § 5 and present our conclusion in § 6.

2 Transport Layer

The function of the transport layer is to provide an end-to-end guaranteed delivery and to ensure that the QOS parameters such as the throughput, delay and jitter requirements of the applications are satisfied all through the connection.

We support both connectionless and connection-oriented services at the transport layer. The user is required to specify the traffic characteristics and the type of the service desired. A PCC scheme (leaky bucket) is adopted as it is more efficient than the traditional window-based flow control schemes [**?**]. This necessitates the negotiation of the user QOS parameters at the time of connection establishment. The QOS parameters are negotiated with its peer for each of the connections and connection is set-up if these requirements are satisfied. The traffic arrival pattern and its type (voice, video or data) are specified at the transport level itself. This information is used to classify the traffic into several classes at the MAC level. Each traffic type is translated into its corresponding traffic class at the network level. For connectionless traffic, admission control is not required. These packets have the lowest priority at the MAC layer and thus has the highest probability of being dropped during congestion.

3 Source Characterization

To evaluate the performance of multiplexing heterogeneous traffic sources over ATM based medium, source characterization and accurate modeling of different traffic types is required. Source characterization describes the stochastic processes that can be used to model various traffic-generating sources. The traffic types are modeled depending on the arrival rates of each of the traffic classes. This describes the individual sources as well as the aggregate traffic. The performance measures such as the throughput, average delay, jitter, queue length and cell loss probability can be obtained from this characterization.

3.1 Data Traffic

Typically, the generation of data from a single source is well characterized by a Poisson arrival process [**?**]. It has been observed that interactive packet interarrivals closely match a constant plus exponential arrival. For interactive data transmission, a single cell may be generated at a time. For bulk data transmission, such as file transfer, a large number of cells may be generated at once (batch arrivals). Packets could be either constant or variable length. In ATM networks, since the packet size is fixed and is relatively small compared to the size of the data packets, each data packet will be broken down into multiple cells. The Poisson arrival assumption of data traffic makes it easier to model with the existing models of bursty traffic.

3.2 Voice Traffic

The burstiness of voice traffic makes a Poisson process unsuitable for modeling it. However, in principle, we can approximate the superposition of a number of bursty sources as a Poisson stream. By modeling voice as MMPP allows us to model sources with time-varying behavior while keeping the related analysis tractable. MMPP is a doubly stochastic Poisson process where the system arrival rate is determined by the state of a continuous-time Markov chain [**?**]. The arrival process makes transitions among a finite number of states by a Markov process. The characterization of the superposition of identical ON/OFF processes using an MMPP is due to Heffes and Lucantoni [**?**].

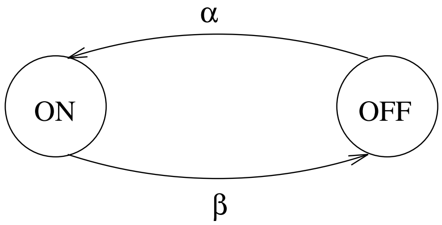

Figure 1: Model for voice traffic

Voice traffic can be modeled as an alternating burst of talk spurt and silences. During each burst, voice cells are generated periodically and during silent periods, no cells are generated. The burst of voice packets is represented by the *ON* state whereas *OFF* state indicates silence period (Figure **??**) [**?**].

Each voice source is modeled as a ON/OFF process. Mean transition rate from the OFF state to the ON state (also called as the birth rate) is given by

$$\alpha = \frac{pF_p}{(1-p)L_B} \tag{1}$$

Mean transition rate from ON state to the OFF state (also known as the death rate) is given by

$$\beta = \frac{F_p}{L_B} \tag{2}$$

Cell emission rate in the ON state

$$\Lambda = \frac{F_p}{L_l} \tag{3}$$

where

- L_l is the length of ATM cell payload (48 bytes)
- F_p is the peak bit rate
- p, the activity factor is the ratio of bit rate and F_p
- L_B is the mean burst length

3.3 Video Traffic

Applications supporting video requires large bandwidth and need to meet the quality constraints of both cell jitter and cell loss probability. Even though video traffic generates correlated cell arrivals as in voice, its statistical nature is quite different from that of the voice source. Video traffic is produced by encoding subsequent video frames which are generated at a rate of 30 frames per second for full-motion video. Video signal exhibits *spatial* as well as *temporal* correlation.

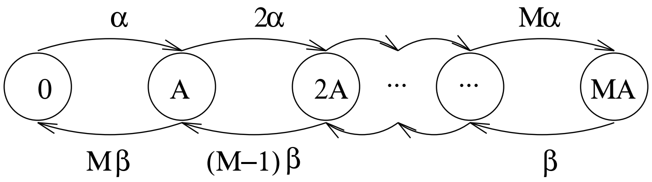

Figure 2: Traffic model for video

To model the video traffic, a quantized bit rate process is assumed [**?**] that is, the arrivals $\lambda(n)$ will take values which are integral multiples of a specific value called the quantization step A. The continuous rate will be sampled at Poisson points and the bit rate becomes a *continuous time process*. This property makes the video to be modeled as a *discrete-state continuous-time Markov process* [**?**] with the parameter $\lambda(t)$. The model is given in Figure **??**.

The transitions from one state to another in the Figure **??** are assumed to be like a birth-death process [**?**]. The rate assignments are based on the assumption that a process in a low activity state is more likely to transit to a higher activity state than to an even lower activity state and vice versa, one quantization step at a time.

For the given bimodal model, the birth rate α, death rate β, and the quantization step A are given by

$$\beta = \frac{a}{1 + \frac{NE^2[\lambda_N]}{MVar[N]}}$$

$$\alpha = a - \beta$$

$$A = \frac{Var[N]}{E[\lambda_N]} + \frac{E[\lambda_N]}{M}$$

where, $E[\lambda_N]$ is the mean arrival rate, $Var[N]$ is the variance and a is a constant which is determined by the proper tuning of the parameters [**?**].

The structure of the Figure **??** is exactly that of the superposition of M identical ON/OFF sources of the voice model.

4 Simulation Model

The model of the simulated system consists of a number of hosts connected to each other through a high-speed ATM switch. These hosts can be individual nodes or a subnet with a hierarchy of nodes. Each host will be generating traffic independent of other hosts in the system and the traffic type can be data, voice, or video.

Figure **??** depicts the multihop topology simulated in this paper. The simulation was performed first on a single hop topology which consists of only one central switch and a number of hosts. Then the simulation was extended to a multihop topology by connecting two switches. The main purpose for choosing such a simple topology is to evaluate the effect of statistical multiplexing on the change of source

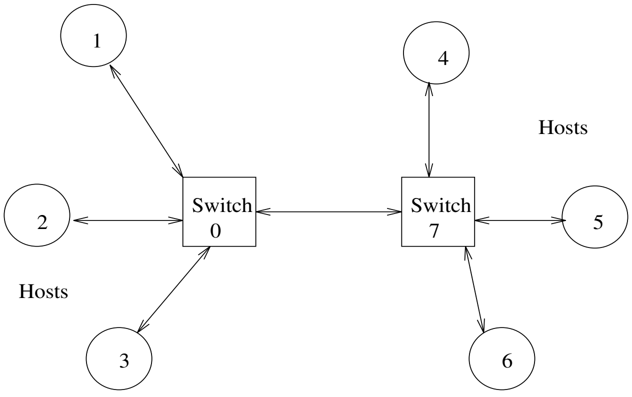

Figure 3: Simulated network topology

characterization while keeping routing complexity to a minimum. To that end, a static source routing table is used at each host.

4.1 Source Parameters

Table **??** summarizes the source characteristics which are used to drive the simulation. Given the source traffic parameters, such as *peak bit rate* F_p, *utilization* p, and *burst length* L_B, the typical values of the birth rate α, death rate β, and cell emission rate Λ in the ON state for voice can be computed from equations **??**, **??**, and **??**. Video frames are generated at a rate of 30 frames/sec with the peak rate of video source as 3.6 Mb/s [**?**]. The desired loss probability due to buffer overflow is given by ϵ and the probability of running out of tokens when source is active is given by ξ.

Table 1: Characteristics of Source Traffic

Traffic Type	F_p (Mb/s)	p	L_B (sec)	ϵ	ξ
Voice	0.064	0.649	0.002841	10^{-5}	10^{-2}
Video	28.8	0.0872	0.34	10^{-5}	10^{-2}
Data	16	0.2	0.001	10^{-7}	10^{-2}

The arrival process of each burst is assumed to be exponentially distributed around their mean birth and death rates for voice and video. The cell arrivals within each burst are uniformly distributed.

4.2 Traffic Control at Host

When a connection-oriented source becomes active, its traffic type is identified; its source characteristics (F_p, p, and L_B) and its QOS requirements (cell loss rate) are translated into the network parameters (equivalent bandwidth). A connection set-up message is then sent to its peer at destination to reserve the network resources along the way. Once the connection is established, leaky bucket parameters, token generation rate and bucket size, for the source are computed based on its delay requirements and transmission continues accordingly. When a source finishes a session, a disconnect message is sent to its destination and the resources (bandwidth and buffer space) will be released by the source, destination and all the intermediate nodes. Figure **??** gives a schematic view of implementation details at each host.

Traffic generation, connection admission and leaky bucket flow control is implemented at each host by the following modules:

- **Source:** simulates traffic arrival process for data, voice, and video according to their characteristics shown in Table **??**.

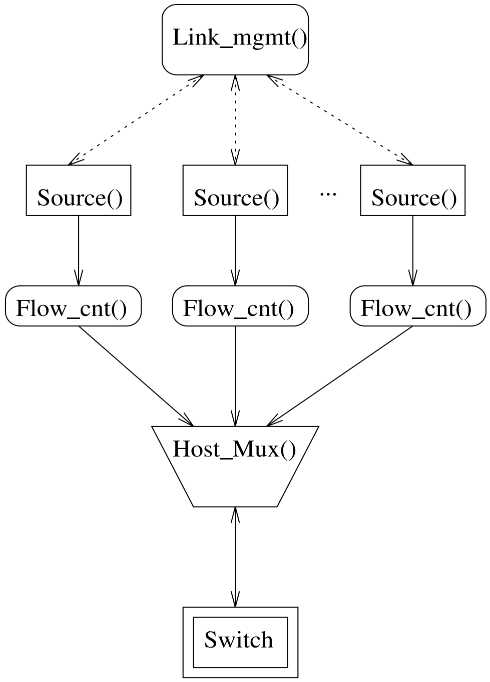

Figure 4: Traffic control at each host

- **Link_mgmt:** monitors the available bandwidth and performs admission control.
- **Flow_Cntl:** implement leaky bucket mechanism and regulate each source traffic according to their traffic characteristics and QOS requirements.
- **Host_MUX:** performs statistical multiplexing on the traffic generated by host sources and transmits cells to switch.

Packets from the source to the destination are routed through the switch. A priority mechanism has been adopted

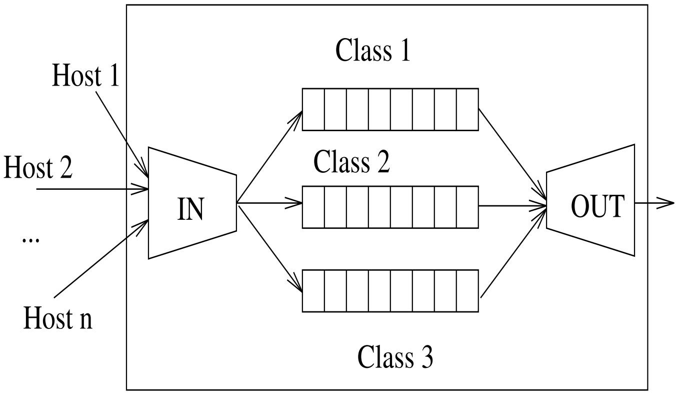

Figure 5: Switch scheduling scheme

by the switching node to accommodate several traffic classes. The incoming traffic is classified as Class 1, 2 and 3 traffic (Figure **??**). We use static priority scheme where, the delay sensitive traffic is given higher priority. A round-robin scheme is used between Class 1 and Class 2 traffic. Two threshold values $threshold_1$ and $threshold_2$ determine access to Class 3 traffic. If the average queue length of Class 2 traffic is less than $threshold_1$ and the queue length of Class 3 is greater than $threshold_2$, Class 3 packet will be transmitted. For our experiments, we have chosen the threshold values to be 2 and 20 packets respectively. The total queue length of each class is restricted to 30 packets.

5 Performance

In this section, we present some of the theoretical bounds obtained from the analytical modeling and compare it with the results obtained by simulation. The model described in § 4 has been simulated using *CSIM*, a C-based simulation tool [**?**].

5.1 Single Hop Network

Table **??** gives the simulation results for a single hop topology where three hosts are connected by a single switch and the traffic is homogeneous, *i.e,* either data, or voice, or video. In such simple environment, the expected delay introduced by traffic policing function at host is estimated roughly by a simple queueing model (M/D/1) for a Poisson arrival process. The observed delay shown in Table **??** agrees well with the analytical estimation, suggesting that the PCC leaky-bucket mechanism works well for the homogeneous traffic with characteristics given in Table **??**. However, the overall low system utilization even for overloaded situation is also observed.

The total link bandwidth (150 Mb/s in our simulation) can accommodate at the most 5.56 video sources (F_p = 28 Mb/s, yields equivalent bandwidth of 26.7 Mb/s) even with statistical multiplexing. When five video sources are active at host, the subsequent video traffic has to wait for enough bandwidth to become available, wasting about 15 Mb/s. Moreover, even the bandwidth is reserved at source link, the traffic may be aborted due to the insufficient bandwidth of remote link on the path. Taking account of all these, plus the overhead introduced by control packets, the low utilization seems inevitable. This result indicates that multiplexing different type traffic on to a ATM link should increase the system utilization. For example, with 15 Mb/s residual bandwidth, theoretically, 3 data sources or 289 voice sources can be multiplexed.

Table 2: Simulation results with homogeneous traffic
Hosts: 3 Sources: 15

Traffic Type	Expected Delay	Observed Delay	Observed Loss	System Usage
Voice	0.345	0.556	0.0	2.3%
Video	4.9	6.865	1.8x10^{-5}	54.7%
Data	1.199	1.239	7.6X10^{-7}	29.5%

Typically, in an ATM network, several traffic types are multiplexed on to a single channel. We study the performance of multiplexing traffic types on to a single medium. Table **??** shows the basic performance metrics such as the delay, jitter and system utilization for multimedia traffic. Queueing analysis of our model for multimedia traffic becomes very complicated and hence no reliable analytical estimate on delay and loss rate is available. In our simulation, the traffic type of each source is generated randomly. As indicated by Table **??**, the performance guarantee for each type of traffic is enforced and a congestion free network is maintained by the control mechanism. The observed delay, delay jitter, and the packet loss rate are all within expected limit. The system utilization varies with traffic pattern, especially with the number of video traffic at each node. It can be observed that by reducing the number of active video sources, the performance of voice and data improves slightly but the utilization decreased by almost 30%.

5.2 Multihop Network

The multihop network under consideration is shown in Figure **??**. Here also, the simulation conditions are similar to the case of single-hop network. The results of homogeneous and multimedia traffic are summarized in the Tables **??** and **??** respectively.

In the homogeneous traffic case, the low utilization for voice and data is due to their low bandwidth and load. For video, utilization depends on session destination. In Table **??**, Class 3 sources refer to the traffic that do not have QOS guarantee. The large delay is due to its lower priority on scheduling queue at switches. The system utilization is 28.33% when the Class 3 sources are not active and is increased to 53.87% when they are active.

5.3 Impact of burst length

The effect of the average burst length on the mean waiting time and the percentage utilization for multimedia traffic in multihop network was observed. The average burst length of video was varied while keeping rest of the parameters fixed. Since video traffic requires much more bandwidth than voice and data, varying its source characteristics will have the most impact on the overall system perfor-

Table 3: Simulation results with multimedia traffic for Single-hop network

Hosts: 3 Sources: 5 System Utilization: 39.69%

Traffic Type	Num	Observed Delay	Variance	Observed Loss
Voice	5	0.83	0.04	0.0
Video	4	2.43	0.06	5.8x10^{-5}
Data	6	55.3	0.23	0.0

Hosts: 3 Sources: 5 System Utilization: 27.6%

Traffic Type	Num	Observed Delay	Variance	Observed Loss
Voice	6	0.736	0.03	0.0
Video	3	1.09	0.04	2.2x10^{-5}
Data	6	27.3	0.15	0.0

Hosts: 3 Sources: 15 System Utilization: 63.33%

Traffic Type	Num	Observed Delay	Variance	Observed Loss
Voice	14	4.09	1.64	0.0
Video	16	28.02	0.4	6.2x10^{-5}
Data	15	75.69	0.2	3.2x10^{-7}

Table 4: Simulation results with homogeneous traffic

Traffic Type	Observed Delay	Variance	Observed Loss	System Usage
Voice	10.16	0.41	0.0	0.02%
Video	43.31	0.09	0.0	12.2%
Data	24.32	0.07	0.0	2.6%

Table 5: Simulation results with multimedia traffic

Traffic Type	Num	Observed Delay	Variance	Observed Loss
Voice	29	8.78	1.11	0.0
Video	24	22.542	1.56	0.0
Data	25	58.44	0.02	0.0
CLASS 3	9	777.2	23.1	0.14

Table 6: Effect of statistical multiplexing on performance

(I) Host and Switch Multiplexing. Utilization: 28.9%

Traffic Type	Num	Observed Delay	Variance	Observed Loss
Voice	25	10.701	0.05	0.0
Video	27	48.874	1.18	0.0
Data	26	25.175	0.25	0.0

(II) Switch Multiplexing only. Utilization: 36.3%

Traffic Type	Num	Observed Delay	Variance	Observed Loss
Voice	25	12.074	0.05	0.0
Video	27	42.786	0.79	0.0
Data	26	6.20	0.01	0.0

mance. Figures 6 and 7 show the effect of waiting time and utilization respectively.

The delay of video traffic increases with the increase in the average burst length. This result indicates that when source becomes less bursty (longer burst length), the leaky bucket scheme becomes less effective. However, the system utilization (Figure 7), increases as video sources become less bursty.

5.4 Impact of Idle periods

Idle periods between bursts also determine the burstiness of a source. For a fixed burst length, the longer the idle period, the more bursty is the source. The effect of the variation in the idle period of video on delay and utilization was observed. The results are summarized in Figures 8 and 9 respectively.

The results show that the delay of video source decreases with the increase in idle period while it has little effect on the utilization. This confirms that the current PCC scheme works better for bursty traffic. It has been reported that statistical multiplexing can achieve more bandwidth gain when traffic is more bursty [**?**].

5.5 Effect of Statistical Multiplexing

In the simulations, statistical multiplexing has been used at hosts and switch. Simulation was also conducted by disabling statistical multiplexing at host as though there were a separate channel for each source. The results are summarized in Table **??**. Notice that delay on data traffic is affected most. Since data traffic in our simulation is the most bursty source, the results suggest that the characteristics of a bursty traffic is more sensitive to statistical multiplexing. Similar observation has been reported by Kurose *et al.* [**?**].

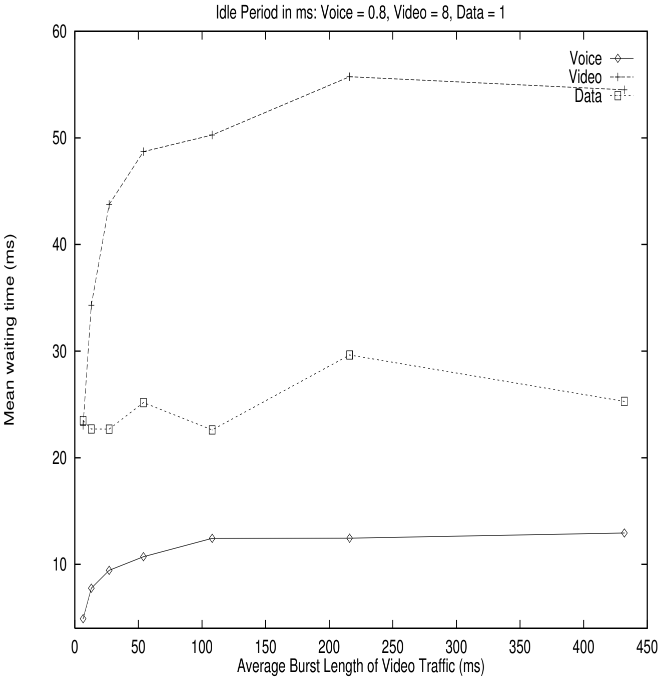

Figure 6: Impact of burst length on delay

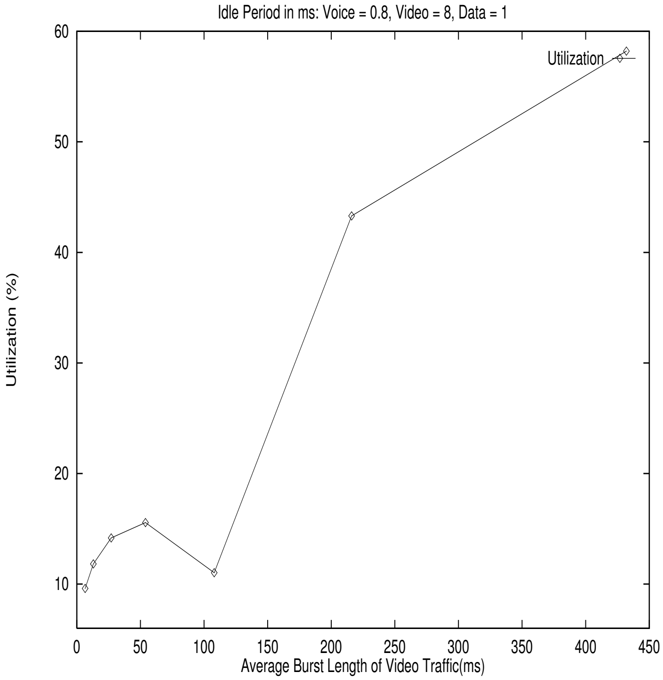

Figure 7: Impact of burst length on utilization

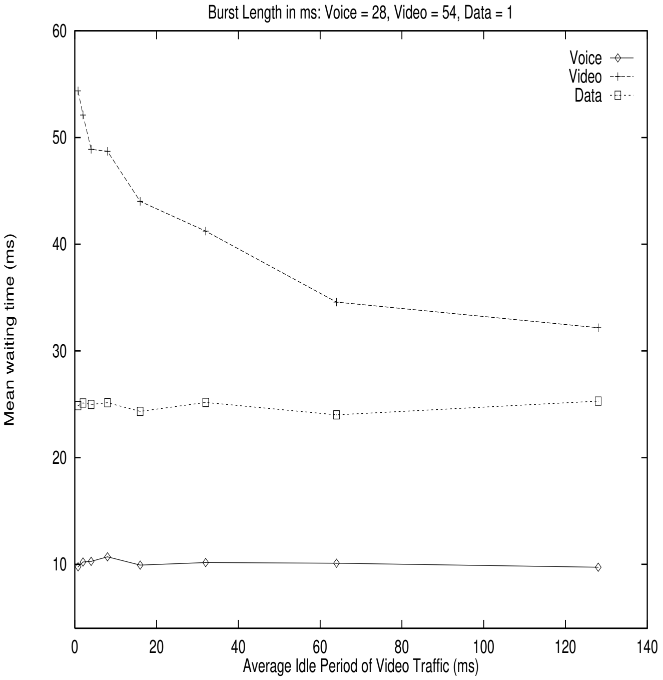

Figure 8: Impact of idle period on delay

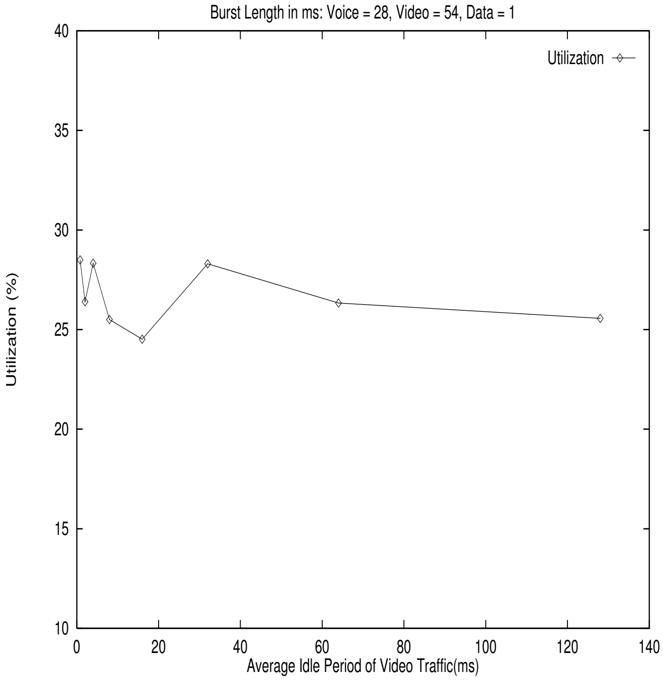

Figure 9: Impact of idle period on utilization

6 Conclusion

We have presented the analytical models for characterizing data, voice and video traffic on a ATM based technology. A transport protocol model that supports both connectionless and connection-oriented traffic has been suggested. We have adopted the leaky bucket mechanism to satisfy the application required quality of service parameters. The basic performance metrics such as the delay, delay jitter, and system utilization are evaluated using simulations.

This study indicates that source characterization is essential for the PCC schemes to predict performance and guarantee the QOS of multimedia applications. It is also important to facilitate the design of an efficient bandwidth management scheme. For homogeneous network traffic, the simulation results match the analytical estimation for the exponential arrival assumption. For multimedia traffic applications, the leaky bucket scheme proves to be effective at the expense of overall low system utilization (63% being the maximum observed). The simulation results also indicates that the system utilization depends not only on the network load but also on the traffic pattern.

In a multihop network with multimedia traffic, simulation results indicate that the leaky-bucket scheme and statistical multiplexing are more effective for bursty traffic. This observation suggests that a control protocol should be able to identify bursty sources and perform statistical multiplexing and equivalent bandwidth allocation on more bursty traffic. Bandwidth allocation according to peak rate and nonstatistical multiplexing would be more efficient for uniform traffic stream.