MODELING IP-ATM GATEWAY USING M/G/1/N QUEUE

*Mahbub Hassan, **Ruhul Sarker, ***Mohammed Atiquzzaman

*School of Computer Science & Software Engineering
Monash University, Caulfield, Melbourne VIC 3145 Australia
E-mail: mahbub@insect.sd.monash.edu.au

**School of Computer Science
Australian Defence Force Academy, Canberra, ACT 2600, Australia

***Department of Electrical & Computer Engineering
The University of Dayton, Dayton, Ohio 45469-0226, USA
E-mail: atiq@engr.udayton.edu

ABSTRACT

Beacuse of the growing investment in both IP and ATM networks, designing efficient IP-ATM gateways for interconnecting these two networks has become very important. One of the main design issues is buffer dimensioning at the gateway for optimum performance. Modeling IP-ATM gateways helps obtaining the optimum design. However, due to very complex nature of an IP-ATM gateway, only models for gateways with infinite buffers have been studied previously. In this paper, an M/G/1/N model for studying the performance of IP-ATM gateway with finite buffers is presented. The performance of IP-ATM gateway as a function of buffer size is studied through analysis of the proposed model. The accuracy of the proposed model is verified through simulation.

1 INTRODUCTION

ATM is being designed as a high-speed, multiservice network to support voice, video and data in the same network. It is a connection oriented packet switched network and provides Quality of Service (QoS) guarantees to connections or virtual circuits (VCs). Because of the high speed and quality of service guarantees of ATM, a popular use of ATM is to interconnect TCP/IP networks using IP-ATM gateways as shown in Figure 1. One of the main functions of the IP-ATM gateway is to set up and release ATM VCs for transmitting IP packets over ATM networks. The efficiency of the interconnection of TCP/IP and ATM depends to a large extent on the performance of the IP-ATM gateway.

Since setting up and releasing ATM VCs for each IP packet may considerably slow down the gateway, an *inactivity timer* is used in the gateway for managing the VC release. IP packets arriving during the VC set-up and transmission of packets over the VC are queued in a buffer in the gateway. An inactivity timer is set and the VC is kept open after the last IP packet is transmnitted from the buffer untill the arrival of a new packet. The VC is released only if the timer expires (no packets arrive within the timeout period).

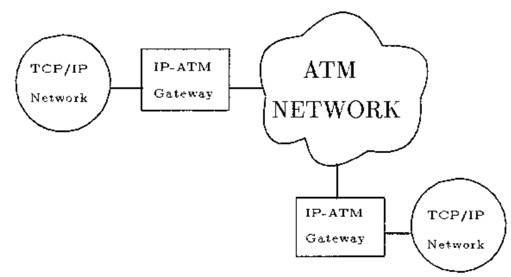

Figure 1: Interconnection of TCP/IP and ATM networks.

A number of parameters such as the ATM-VC bandwidth utilisation, packet loss probability, the number of VC setups per second etc. need to be studied to design efficient IP-ATM gateways. Due to the complexity in modeling the gateway, most of the previous studies [1, 2] have been carried out using simulation. A first modeling attempt in studying the performance of an IP-ATM gateway with infinite buffers at the gateway has been reported in [3]. By comparison of simulation and analytical models, it was shown that the model produced very accurate results. Although the model presented in [3] has laid the groundwork for such studies, an IP-ATM gateway, in practice would contain a finite amount of buffer space. Since such a model does not currently exist in the literature, the objective of this paper is to develop an analytical model of an IP-ATM gateway with a finite buffer size. The model would then be used to study the performance of the gateway. In order to validate our analytical model, we developed a simulation model and then compared the results obtained from the model with those obtained from simulations.

The rest of the paper is organised as follows. The performance measures for an IP-ATM gateway are described in Section 2. The modeling and analysis of

0-7803-4984-9/98/$10.00 ©1998 IEEE.

an IP-ATM gateway using an $M/G/1/N$ queue to obtain the expressions for the performance measures are presented in Section 3. Section 4 describes the simulation model used for verifying the expressions obtained in Section 3. Some numerical results obtained from the analysis are compared with the simulation results in Section 5. Finally a conclusion is presented in Section 6.

2 PERFORMANCE MEASURES

We are interested in estimating the following performance measures for an IP-ATM gateway as a function of the inactivity timeout T, set-up time C, and the buffer capacity of the gateway.

ATM VC set-up rate γ is calculated as the number of set-ups per unit time. ATM VC set-up rate directly affects the amount of processing load on the IP-ATM gateway and the ATM network and may also affect the cost of maintaining the ATM connection if there is per set-up charge imposed by the ATM network service provider.

ATM bandwidth utilisation U is calculated as the ratio of the time the ATM VC is transmitting data to the total duration of connection time. Since bandwidth is often considered an expensive resource, it is important to measure U of an IP-ATM gateway.

Mean packet delay W in the IP-ATM gateway is obtained as the mean of the time spent by a packet in the buffer plus the time taken by a packet to be transmitted over the ATM VC. Mean queue length L in the IP-ATM gateway is obtained as the expected number of packets waiting in the buffer plus any being transmitted over the ATM VC. Finally, we measure the packet loss (due to buffer overflow) probability $\mathcal{P}$ in the IP-ATM gateway. In the following section, we obtain the expressions for the above performance measures by modeling and analysing an IP-ATM gateway with finite buffers using an $M/G/1/N$ queue.

3 $M/G/1/N$ QUEUING MODEL

An IP-ATM gateway with finite buffers is modeled as an $M/G/1/N$ queue as shown in Figure 2. A finite buffer of size $N-1$ (the total capacity of the system is N) is associated with a single server. The server here represents an ATM VC. The packet arrival process to the queue is Poisson with a mean arrival rate of λ. The service time has a general distribution G with a mean of $\frac{1}{\mu}$, where μ is the mean service rate. The service time in this model corresponds to the transmission time of an IP packet over the ATM VC. Note that for a constant bandwidth ATM VC and a constant packet size, G reduces to a deterministic distribution. The packets from the buffer are served in a first-in-first-out (FIFO) order. Packets arriving at a full buffer are dropped and are lost from the system.

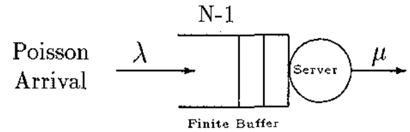

Figure 2: $M/G/1/N$ queueing model.

The server serves all packets in the buffer exhaustively until the buffer becomes empty. After serving the last packet in the buffer, the server waits T time units anticipating the arrival of more packets. If no packet arrives within T time units, the server is turned off (the server goes to a vacation corresponding to the release of an ATM VC) to conserve resources such as ATM VC bandwidth. For an IP-ATM gateway, T corresponds to an inactivity timeout.

If the server is off when a packet arrives, it is turned on (ATM VC set-up process begins) immediately. However, there is a service set-up period of C time units for the server. In other words, the server takes C time units from the moment it is turned on to the beginning of the service of the packet(s) waiting in the buffer. For an IP-ATM gateway, C corresponds to the set-up time of an ATM VC.

The above queueing model is analysed in the rest of Section 3 to obtain the performance measures described in Section 2.

3.1 Standard $M/G/1/N$ Queue

The $M/G/1/N$ queuing model presented in the previous section differs from a standard $M/G/1/N$ queue in that it lets the server take a vacation delayed by T and there is a service set-up time of C. The presence of the variables T and C makes the analysis very complex. Nevertheless, we can perform the analysis of this complex queue by utilising some of the results available in the literature for a standard $M/G/1/N$. In this section, the results for a standard $M/G/1/N$ queue are documented; these results will be used in Sections 3.2 to 3.4 to analyse the more complex queue with T and C.

The $M/G/1/N$ queuing system goes through two alternating periods: (i) busy periods when the server is either serving (the ATM VC is tranmitting packets) or the server is being setup (ATM VC is being setup) and (ii) idle periods when the system is empty and the server is either turned on (the ATM VC remains established) waiting for packets or it is in vacation (ATM VC has been released).

The expected length of a busy period for the standard $M/G/1/N$ system, where the server starts a busy period V with only one packet in the queue, is given by [4]:

$$E(V) = \frac{1-\rho^{N+1}}{\mu(1-\rho)} \tag{1}$$

where $\rho = \frac{\lambda}{\mu}$. By definition, the expected number of packets in the system (mean queue length) is:

$$L = \sum_{n=0}^{N} n P_n \tag{2}$$

where P_n is the probability of n packets in the system. From L, the expected waiting time (mean packet delay) W in the system is obtained by *Little's formula* as:

$$W = \frac{L}{\lambda'} \tag{3}$$

where λ' is the effective mean packet arrival rate to the system defined as the mean number of packets that are admitted to the system in unit time. It is obtained as:

$$\lambda' = \lambda(1 - P_N) \tag{4}$$

The packet loss probability $\mathcal{P}$ due to buffer overflow is equivalent to the probability of the system being full (N packets in the system) which is derived as:

$$\mathcal{P} = P_N \tag{5}$$

The probability of n packets in a standard $M/G/1/N$ system is given by[5]:

$$P_n = \begin{cases} c\hat{p}(n), & 0 \leq n < N \\ 1 - \frac{1-c(1-\rho)}{\rho}, & n = N \end{cases} \tag{6}$$

where $\hat{p}(n)$ is the probability of having n customers in the standard $M/G/1/\infty$ system and

$$c = \left\{1 - \rho\left[1 - \sum_{n=0}^{N-1} \hat{p}(n)\right]\right\}^{-1} \tag{7}$$

Due to the enormous complexity involved in deriving the exact solution for $\hat{p}(n)$, some approximations are often used. A popular approximation, called *diffusion approximation* [4, 5], is:

$$\hat{p}(n) = \begin{cases} 1 - \rho, & n = 0 \\ \rho(1-\hat{\rho})(\hat{\rho})^{n-1} & n \geq 1 \end{cases} \tag{8}$$

where

$$\hat{\rho} = \exp\left[\frac{2(\lambda - \mu)}{\lambda + \mu K_s}\right] \tag{9}$$

K_s is the square of the variation coefficient of the service time which is obtained as[5]:

$$K_s = \sigma^2 \mu^2 \tag{10}$$

where σ^2 is the variance of the service time.

The above results for a standard $M/G/1/N$ queue are used in the following sections to analyse the complex queue with T and C to derive the expressions for γ, U, L, W and $\mathcal{P}$.

3.2 Derivation of γ

In this section, we analyse the complex $M/G/1/N$ queue with T and C to obtain the set-up rate γ as[3]:

$$\gamma = e^{-\lambda T} \beta \tag{11}$$

where β is the rate of idle period (number of times the system becomes empty or idle in unit time). For $\rho < 1$, β can be obtained as (see Appendix for detailed derivation):

$$\beta = \frac{\lambda}{1 + Ce^{-\lambda T}\lambda + \frac{\lambda(1+K)(1-\rho^{N+1})}{\mu(1-\rho)}} \tag{12}$$

where $K = e^{-\lambda T} \times \min(N-2, C\lambda)$. The correctness of Eqn. (12) is verified by the fact that by taking the limit of β for $N \to \infty$ (for $\rho < 1$) results in the same expression for β as obtained for the $M/G/1/\infty$ queue with T and C in [3].

3.3 Derivation of U

Once the rate of idle period β is known, the bandwidth utilisation U can be obtained as [3]:

$$U = \frac{\rho}{\rho + \beta\{e^{-\lambda T}T + \frac{1}{\lambda}[1 - e^{-\lambda T}(\lambda T + 1)]} \tag{13}$$

3.4 Derivation of L, W and $\mathcal{P}$

In this section, we obtain the expressions for L, W and $\mathcal{P}$ of the complex $M/G/1/N$ queue with T and C by approximating it to an equivalent standard $M/G/1/N$ queue. The approximation is described next.

The difference between the complex $M/G/1/N$ queue with T and C and the standard $M/G/1/N$ queue is that some packets experience an additional set-up time C before the usual service time in the complex queuing model. Therefore, the complex queue can be approximated to an equivalent standard queue by deriving the effective mean service rate μ_{eff} and hence the effective load $\rho_{\text{eff}} = \frac{\lambda}{\mu_{\text{eff}}}$ and the variance σ^2_{eff} of the effective service time. Expressions for all other variables of the approximated queue remains the same as given in Section 3.1. It is expected that $\mu_{\text{eff}} \leq \mu$ and $\rho_{\text{eff}} \geq \rho$ because of the set-up time.

Once μ_{eff}, σ^2_{eff} and ρ_{eff} are obtained, one can estimate the performance measures L, W and $\mathcal{P}$ of the complex queue by replacing ρ with ρ_{eff} in Eqns. (6), (7) and (8), μ with μ_{eff} in Eqns. (9) and (10), and σ^2 with σ^2_{eff} in Eqn. (10). The derivation of μ_{eff} and σ^2_{eff} are presented below.

As shown in [3], the effective mean service rate μ_{eff} for the delayed vacation system is obtained as:

$$\mu_{\text{eff}} = \frac{\mu}{1 + \mu P_{\epsilon} C} \tag{14}$$

where P_e is the proportion of packets that experience a service set-up time. It can be seen from Eqn. (14) that $\mu_{\text{eff}} \leq \mu$ which is intuitively expected as the effective service time increases due to the set-up time. P_e is obtained as [3]:

$$P_e = \frac{e^{-\lambda T}\beta}{\lambda'} \tag{15}$$

It is interesting to note that for zero set-up time ($C = 0$), $\bar{\mu} = \mu$ and hence the complex $M/G/1/N$ queue reduces to the standard $M/G/1/N$ queue irrespective of the control parameter T. Similarly, since $\lim_{T\to\infty} P_e = 0$, the complex $M/G/1/N$ queue reduces to a standard $M/G/1/N$ queue irrespective of the set-up time C for very large inactivity timeout T.

By definition, the variance σ^2_{eff} of the effective service time can be obtained as [6]:

$$\sigma^2_{\text{eff}} = E(X^2) - [E(X)]^2 \tag{16}$$

where $E(X)$ and $E(X^2)$ are the first and second moments of the effective service time and are obtained as [3]:

$$E(X) = \frac{1}{\mu} + P_e C \tag{17}$$

$$E(X^2) = P_e E(X_e^2) + (1 - P_e)E(X_n^2) \tag{18}$$

where $E(X_e^2)$ is the second moment of the service time of the packets experiencing a set-up time and $E(X_n^2)$ is the second moment of the service time of the packets that do not experience the set-up time. The values of $E(X_e^2)$ and $E(X_n^2)$ depend on the distribution of the effective service time. For deterministic service time distribution (G is deterministic), it can be shown [3]:

$$E(X^2) = \frac{1}{\mu^2} + \frac{\gamma C^2}{\lambda} + \frac{2\gamma C}{\mu\lambda} \tag{19}$$

The second moment of the effective service time for other service time distributions can be obtained in a similar fashion.

The following section describes a simulation study conducted to verify the accuracy of the above model. In Section 5, some numerical results are analysed using the above model to study the performance of an IP-ATM gateway as a function of buffer capacity and inactivity timeout. Analytical results are compared with the ones obtained from the simulation experiments to assess the effectiveness of the model.

4 SIMULATION

We simulated an IP-ATM gateway where packet arrivals from the IP network to the gateway is generated using exponentially distributed interarrival times with a mean of 0.02 second. This gives an average arrival rate of 50 packets per second. All IP packets have a constant size of 1000 bytes and the bandwidth of the ATM VC is 0.5 Mbps. This simulates a constant (deterministic) service time distribution with a service rate equals to 62.5 packets per second.

parameter	*value*
mean IP packet interarrival time	0.02 seconds
IP packet size	1000 bytes
ATM VC bandwidth	0.5 Mbps
ATM VC set-up time	0.05 seconds

Table 1: Values of simulation parameters.

The simulation used a buffer of finite capacity at the ATM interface as shown in Figure 2. Packets arriving when the ATM VC is busy transmitting a packet or when the ATM VC is being set-up, are queued at the tail of the FIFO buffer if there is space in the buffer, otherwise they are dropped. The packet loss probability is obtained as the ratio of packets lost to total packets arrived.

The ATM VC transmits packets from the head of the buffer until the buffer becomes empty. After transmitting the last packet from the buffer, an inactivity timer is set with a timeout value of T seconds. If a packet arrives before the timer expires, the timer is cleared and the packet is transmitted. However, if no packets arrive within T seconds, the timer expires and the VC is released. If an arriving packet at an empty system finds that the VC has been released, a VC set-up process is initiated which takes 0.05 seconds. Values of various simulation parameters are summarised in Table 1. The results obtained from the simulation are presented and compared with the ones obtained from analysis in the following section.

5 RESULTS

In this section, some numerical results are analysed to study the impact of the buffer capacity and the timeout value of the inactivity timer in the IP-ATM gateway on the performance measures. Results obtained from the model are compared with the ones obtained from simulation to verify the accuracy of the model.

5.1 Impact of Buffer Capacity

In this section, we use the modeling and simulation results to study the impact of the buffer capacity in the IP-ATM gateway on the performance measures described in Section 2.

ATM VC set-up rate γ as a function of N obtained from simulation and Eqn. (11) are shown in Figure 3. It can be seen that set-up rate decreases as N is increased. This is because with smaller buffer capacity, some packets are lost in the gateway and hence it becomes idle more often, requiring more frequent set-ups..

ATM VC bandwidth utilisation U as a function of N obtained from simulation and Eqn. (13) are shown in Figure 4. As expected, the figure shows that utilisation increases as N is increased. However, there is a "knee" in the utilisation graph; it increases rapidly upto the "knee" after which the increase is mimimal. Therefore, it can be concluded that there is an optimum buffer size beyond which the gain in bandwidth utilisation is not significant. For the parameters selected in Figure 4, the optimum buffer capacity is about 6.

In Figures 3 and 4, the analytical results obtained from the model remains within 5% of the simulation results. This verifies the accuracy of the Eqns. (11) and (13) in estimating γ and U.

Figures 5 and 6 compare the results obtained from analysis with the ones obtained from simulation for L and W as a function of N. There is a good match between the simulation and analytical results for $N < 7$. The difference that appears between the simulation and analytical results for larger N is mainly due to the approximation error introduced by the ***diffusion approximation*** in approximating the probability distribution of the expected number of packets in the system in Eqns. (8) and (9).

To observe the amount of error introduced by the *diffusion approximation*, the probability that the system is full (equivalent to the packet loss probability) is plotted in Figure 7. It can be seen that the *diffusion approximation* introduces a noticeable error. Therefore, it may be concluded, that our model is expected to provide better results for L, W and $\mathcal{P}$ if a better approximation for the probability distribution of the expected number of packets for an $M/G/1/\infty$ is found. Deriving a better approximation is left as a future exercise.

5.2 Impact of Inactivity Timeout

In this section, numerical results are analysed to study the impact of inactivity timeout T in the IP-ATM gateway on the performance measures γ and U.

Figure 8 shows the set-up rate as a function of the inactivity timeout T. As expected, the set-up rate approaches zero as T is increased. However, there is a "knee" in the graph; set-up rate decreases rapidly upto the "knee" beyond which the decrease is negligible. In Figure 8, the "knee" is noticed at about $T = 0.08$.

Figure 9 shows the ATM VC bandwidth utilisation U as a function of the inactivity timeout T. Again, a "knee" is observed. As T is increased , the ATM

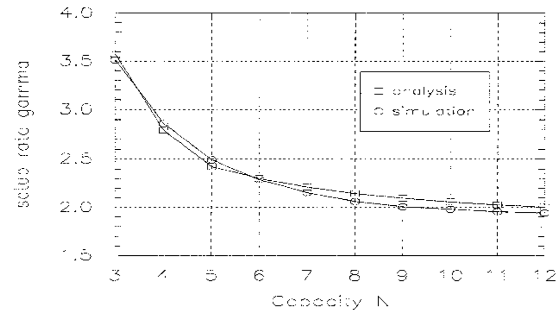

Figure 3: ATM VC setup rate as a function of buffer capacity for $C = 0.05$ and $T = 0.02$.

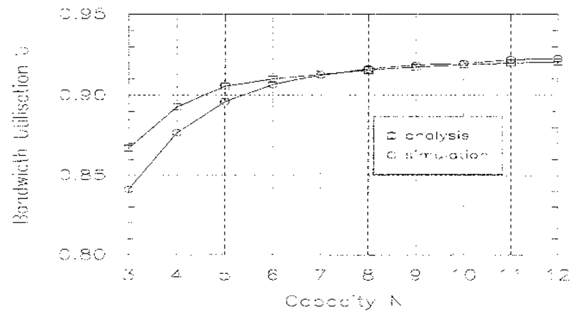

Figure 4: ATM bandwidth utilisation as a function of buffer capacity for $C = 0.05$ and $T = 0.02$.

VC remains open without any data being transmitted for larger periods. Hence, bandwidth utilisation decreases significantly as T is increased up to a certain point (around $T = 0.08$ in this case), after that U decreases very slowly and approaches a minimum value.

Once again the accuracy of Eqns. (11) and (13) are verified by the close match of the analytical results with the simulation results in Figures 8 and 9.

6 CONCLUSION

An IP-ATM gateway with finite buffers has been modeled using an $M/G/1/N$ queue. The model allows us to study a range of performance measures for the IP-ATM gateway as a function of the buffer size

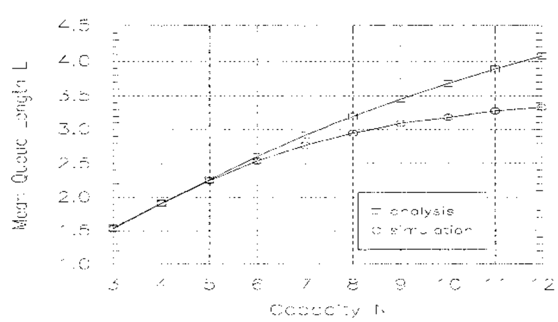

Figure 5: Mean queue length as a function of buffer capacity for $C = 0.05$ and $T = 0.02$.

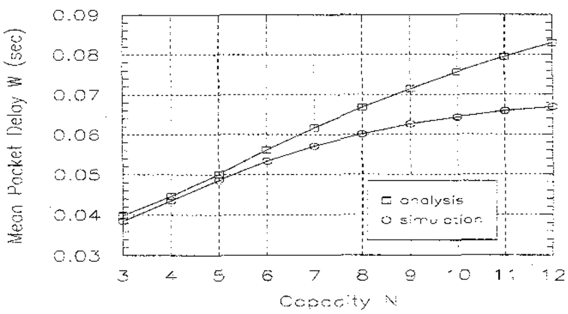

Figure 6: Mean packet delay as a function of buffer capacity for $C = 0.05$ and $T = 0.02$.

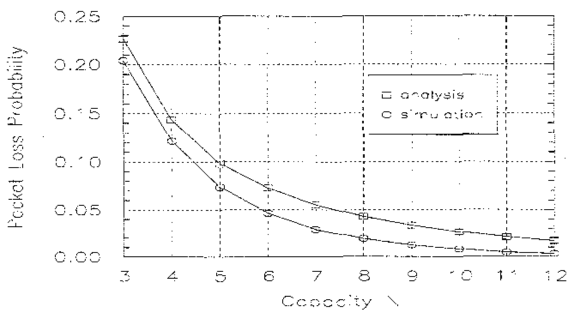

Figure 7: Packet loss probability as a function of buffer capacity for $C = 0.05$ and $T = 0.02$.

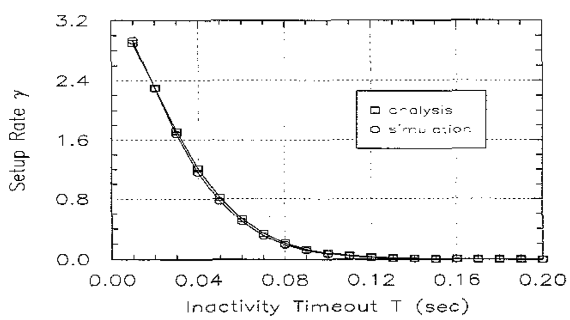

Figure 8: ATM VC setup rate as a function of the inactivity timeout for $C = 0.05$ and $N = 6$.

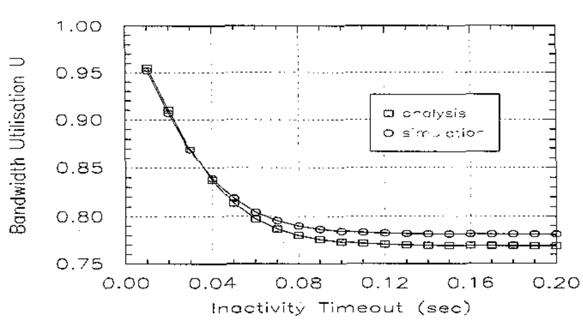

Figure 9: Bandwidth utilisation as a function of the inactivity timeout for $C = 0.05$ and $N = 6$.

which helps buffer dimensioning. The accuracy of the model is verified using simulations.

We have found that the VC set-up rate and packet loss rate decreases and bandwidth utilisation increases as the buffer capacity at the gateway is increased until a certain point is reached, after which they are insensitive to the increase in the buffer capacity. We have also seen that the VC set-up rate and VC bandwidth utilisation decreases exponentially with an increase in the timeout value of the inactivity timer. The above results enable one to optimise several factors, such as buffer capacity, VC bandwidth utilisation and inactivity timer value.

APPENDIX : DERIVATION of β

The rate of idle period β is simply $\beta = \frac{p_0}{\frac{1}{\lambda}}$, where $\frac{1}{\lambda}$ is the mean idle period and p_0 is the fraction of time the server remains idle and is obtained as:

$$p_0 = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + Ce^{-\lambda T} + E(V_{\text{eff}})} \tag{20}$$

where $Ce^{-\lambda T}$ is the mean set-up time for all busy periods[3] and $E(V_{\text{eff}})$ is the effective mean busy period of the complex queue obtained as $E(V_{\text{eff}}) = (1+K)E(V)$, where $E(V)$ is the mean busy period of the standard $M/G/1/N$ queue as given by Eqn. (1) and K is the mean number of packets that are queued for service during the start of a busy period for the complex queue and can be obtained as: $K = e^{-\lambda T} \times \min(N-2, C\lambda)$.

References

[1] S. Keshav, C. Lund, S. Phillip, N. Reingold, and H. Saran. An Empirical Evaluation of Virtual Circuit Holding Policies in Ip-Over-ATM Networks. *IEEE JSAC*, 13(8):1371–1382, October 1995.

[2] A. Schmidt and R. Campbell. Internet Protocol Traffic Analysis with Applications for ATM Switch Design. *Computer Communications Review*, 23(2):39–52, April 1993.

[3] M. Hassan and M. Atiquzzaman. A Delayed Vacation Model of an M/G/1 Queue with Setup Time and Its Application to SVCC-Based ATM Networks. *IEICE Transactions on Communications*, E80-B(2):317–323, February 1997.

[4] J. Medhi. *Stochastic Models in Queueing Theory*. Academic Press, 1991.

[5] E. Gelenbe and G. Pujolle. *Introduction to Queueing Networks*. John Wiley and Sons, 1987.

[6] H.A. Taha. *Operations Research An Introduction*. MacMillan, fourth edition, 1987.