Approximate Models for the Study of Nonstationary Queues and Their Applications to Communication Networks

Sandeep Sharma

Alcatel Network Systems
2912, Wake Forest Rd.,
Raleigh, NC 27609

David Tipper

ECE Dept., 221-B Riggs Hall
Clemson University,
Clemson, SC 29634

ABSTRACT The modelling of ATM networks, and the effect of nonstationary traffic patterns, is of considerable interest to the research community. Traditional simulation techniques can be employed to study small network models. However, the complexity of current network models, and the amount of time expended in simulating the network has forced researchers to look at alternate modelling techniques. In this paper, we present numerical methods based techniques to model multi-class Markovian traffic, with and without priority, in ATM networks. The use of these methods is validated, under stationary and nonstationary traffic patterns, with simulation. The models developed are quite accurate, and provide substantial saving in computational time over the simulations done to validate the model. An illustrative application of the models to an ATM network is presented.

1 Introduction

Computer and telecommunication networks are evolving into Broadband Services Digital Networks (BISDN) which will support a wide variety of services such as voice, data, and video applications utilizing a single common network. Asynchronous Transfer Mode (ATM) has been identified by CCITT as the preferred transport technique for BISDN. A distinguishing feature of ATM is packet switching using small fixed length packets. Since BISDNs will support various classes of multimedia traffic with different bit rates and quality of service requirements, traffic in the network is expected to be very bursty and nonstationary at times. Also, it has been noted in [2]- [9] that due to the different time scales involved in traffic control and resource management, it is useful to know the dynamic nonstationary behaviour of the network.

Recently, a great deal of research has been done in the area of analytical and numerical solutions for nonstationary queuing models [10]- [5]. Most techniques use the Chapman-Kolmogorov (C-K) forward differential equations for modelling queue behaviour [12]. The major drawback for this method is that the number of differential equations needed to model a queue depends on the state space for that particular queue. Another common method for looking at queue behaviour is Jensen's method, also referred to as Randomization or Uniformization [5]. Jensen's method is compared with the Chapman-Kolmogorov(C-K) method in [10]. Both Jensen's Method and the C-K differential equations become computationally expensive as the state space increases. Other alternatives like Fluid Flow Models [3]- [12] have been suggested by researchers. These methods lend themselves to ease of computation, and are accurate within reasonable bounds. Tipper et. al. [12] and the references therein, provide a description for modelling different queues using fluid flow approximations. Queues with different classes of traffic can also be modelled using the fluid flow approximations.

In this paper, fluid flow models for mean behaviour of the multi-class M/D/1 queue, with and without priority, are presented. These models are simple and computationally inexpensive.

The remainder of the paper is laid out as follows. In section 2, the derivations for the fluid flow models of an M/D/1 queue, with multiple classes of traffic, with and without priority are given. In section 3, we use the results of section 2 to model an ATM virtual circuit or virtual path traversing three nodes. Section 4 concludes the paper.

2 Multi-class traffic modelling

In this section, fluid flow model equations are derived for the M/D/1 queue with multiple classes of traffic, with and without priority. In doing so, we follow the approach used in [12], where fluid flow models have been presented for the M/M/1 queue, the M/D/1 queue, and the M/M/1 queue with multiple classes of traffic. We proceed as follows: We first give a general background of mean value fluid flow models. We then derive equations for the M/D/1 queue with multiple classes of traffic, with and without priority, and apply the results to a specific example.

2.1 Fluid flow models: Background

Let $X(t)$ be a state variable denoting the ensemble average number in in the system in an arbitrary queuing model at time t. Let $f_{in}(t)$ and $f_{out}(t)$ be the ensemble average of flow in and flow out of the system respectively. Let $\dot{X}(t) = \frac{dX(t)}{dt}$ be the rate of change of the state variable with respect to time.

0-7803-0950-2/93/$3.00©1993IEEE

The rate of change of the state variable can then be written as:

$$\dot{X}(t) = -f_{out}(t) + f_{in}(t). \quad (1)$$

Equations of this kind have been used in the literature, and are commonly referred to as fluid flow equations [12]- [1]. To tailor this equation to a queuing system, we define C, λ, and $\frac{1}{\mu}$ as the queue server capacity, average arrival rate, and the mean packet size respectively. Assuming that the queue capacity is unlimited, $f_{in}(t)$ is just the arrival rate λ. The flow out of the system, $f_{out}(t)$, can be related to the ensemble average utilization of the queue ρ by $f_{out}(t) = \mu C\rho$. We assume that the utilization of the link, ρ, can be approximated by the function $G(X(t))$, which represents the ensemble average utilization of the link at time t as a function of the state variable. Hence, we can represent the queue by the following nonlinear differential equation:

$$\dot{X}(t) = -\mu C G(X(t)) + \lambda. \quad (2)$$

The form of the utilization function, G(X(t)), depends on the queuing system under study and upon the data available. If statistical data is available, this function can be empirically formulated. This, however, is not generally the case and G(X(t)) is normally determined by matching the steady state queuing theory results with Equation 2.

Standard numerical integration techniques can be iteratively applied to solve Equation 2 [12].

2.2 The M/D/1 queue with multiple classes of traffic, no priority

The fluid flow equation for this model is given as [12]:

$$\dot{X} = -\mu C[(X+1) - \sqrt{X^2+1}] + \lambda. \quad (3)$$

Equation 3 can now be generalized for multiple classes of traffic. With K classes of traffic, Let X_T, λ_T, X_i and λ_i denote the total number in system, the mean aggregate arrival rate to the queue, number of class i packets in the system, and arrival rate of the i th class of packets respectively. We can then rewrite Equation 3 as:

$$\dot{X}_T = -\mu C[(X_T+1) - \sqrt{X_T^2+1}] + \lambda_T. \quad (4)$$

(Equation 4 is possible, because we can add up Poisson arrival streams into one aggregate arrival process, and we are not concerned about the class of packet currently undergoing service [6].) Let $G_i(X_i, X_T)$ be the average utilization of the server for class i traffic, as a function of the total number in the system and the number of class i packets in the system. The fluid flow equation for the i th class traffic can then be written as:

$$\dot{X}_i = -\mu C G_i(X_i, X_T) + \lambda_i. \quad (5)$$

At steady state, with arrival rate λ_T, the average number in the system, X_T, is [6]:

$$X_T = \frac{\lambda_T}{\mu C} + \frac{\lambda_T^2}{2\mu^2 C^2(1-\frac{\lambda_T}{\mu C})}. \quad (6)$$

We can also write X_i, the state variable representing the average queue length of the i th class traffic as:

$$X_i = \frac{\lambda_i[2\mu C - \lambda_T]}{2\mu C[\mu C - \lambda_T]}. \quad (7)$$

At steady state, (i.e $\dot{X}_T = 0$) from Equation 4 we get: $\lambda_T = \mu C[(X_T+1) - \sqrt{X_T^2+1}]$. Similarly, from Equation 5, at steady state (i.e $\dot{X}_i = 0$) we get: $\lambda_i = \mu C G_i(X_i, X_T)$. Substituting these relationships into Equation 7, and simplifying further, we get:

$$G_i(X_i, X_T) = \frac{2X_i[\sqrt{X_T^2+1} - X_T]}{\sqrt{X_T^2+1} - (X_T - 1)}. \quad (8)$$

Finally, substituting Equation 8 in Equation 5 we get:

$$\dot{X}_i = -\mu C[\frac{2X_i[\sqrt{X_T^2+1} - X_T]}{\sqrt{X_T^2+1} - (X_T-1)}] + \lambda_i. \quad i = 1, 2, \cdots, K \quad (9)$$

Thus, Equation 9 is the set of fluid flow equations describing the queue length dynamics of each class i traffic, in an M/D/1 queue with multiple classes of traffic.

The accuracy of equation 9 has been verified by extensive comparison with simulation, and the results can be found in [11]. A typical comparison is shown in Figure 1, for two classes of traffic with initial conditions $X_1(0) = 0$, $X_2(0) = 0$, and parameters $\mu C = 1$, $\lambda_1 = 0.2$, $\lambda_2 = 0.25$. The simulation results shown were generated using the ensemble average approach of [7]. Note that three curves are given for the simulation results, the estimate of the mean number in the system as determined by 10,000 simulation runs and two curves corresponding to the 95 % confidence interval on the estimate. One can clearly see that the fluid flow model results match the simulation results closely, and, as expected, the steady state values are the same for both cases.

2.3 The M/D/1 queue with multiple classes of traffic and non-preemptive priority

In this section, we extend the model above to include the effects of non-preemptive priority, specifically, we assume that a class i packet has priority over a class $(i+1)$ packet. Proceeding along the lines of the no-priority case, the total number in system of the queue $X_T = \sum_{i=1}^{K} X_i$ is given by Equation 6. Again, this equation is possible because the Poisson inputs can all be added up, irrespective of the packet currently undergoing service. The fluid flow equations for each class, and for the total queue length are given by Equations 4 and 5 as before. Due to the complexity of the problem, the equations for the priority case are derived for 2 classes of traffic. However a similar approach could be extended for any number of priorities. For the case of two priorities, the queue lengths of class 1 and 2 packets can be written as [6]:

$$X_1 = \frac{\frac{\lambda_1}{\mu C}[\frac{\lambda_2 - \lambda_1}{\mu C} + 2]}{2(1-\frac{\lambda_1}{\mu C})}. \quad (10)$$

and:

$$X_2 = \frac{\frac{\lambda_2}{\mu C}[2 - \frac{\lambda_1+\lambda_2}{\mu C} - \frac{2\lambda_1}{\mu C} + \frac{2(\lambda_1+\lambda_2)\lambda_1}{\mu^2 C^2}]}{2(1 - \frac{\lambda_1+\lambda_2}{\mu C})(1 - \frac{\lambda_1}{\mu C})}. \quad (11)$$

The fluid flow equations for each class are given by:

$$\dot{X}_1 = -\mu C G_1(X_1, X_T) + \lambda_1. \quad (12a)$$

$$\dot{X}_2 = -\mu C G_2(X_2, X_T) + \lambda_2. \quad (12b)$$

Following the same derivation procedure as before (i.e. set $\dot{X}_1 = 0$, $\dot{X}_2 = 0$ and $\dot{X}_T = 0$ in 12, and 4, and then substituting in 10, 11 and 6 and solving), we get:

$$\dot{X}_1 = -\mu C[\frac{A - \sqrt{C^2 + 16X_1}}{4}] + \lambda_1. \quad (13)$$

and

$$\dot{X}_2 = -\mu C[\frac{4B - A - \sqrt{A^2 + 16X_1}}{4}] + \lambda_2. \quad (14)$$

where $A = (3X_1 + X_2 + 3 - \sqrt{(X_1 + X_2)^2 + 1})$ and $B = (X_1 + X_2 + 1 - \sqrt{(X_1 + X_2)^2 + 1})$. Details of the derivation can be found in [11].

To verify the accuracy of these equations, different cases were simulated with different traffic patterns. Details can be found in [11] In a typical example, the arrival rate of the priority cells was assumed to be nonstationary, with rates $\mu C = 1$, $\lambda_1 = 0.5 + 0.4\sin(0.2(t + 20))$, and $\lambda_2 = 0.2$ with initial conditions $X_1(0) = X_2(0) = 0$. Figure 2 plots the behaviour of the number in the system. Note that the sinusoidal behaviour of the high priority cells considerably effects the low priority cell occupancy also. The analytical result models the high priority occupancy fairly well, but constantly over and undershoots the simulation curve for the low priority cells.

3 Effect of nonstationarity in ATM networks

In this section we study the interaction of two classes of traffic in an ATM network, one stationary, and one nonstationary. We assume each queue in an ATM 'virtual path/circuit' can thus be modelled as an $M/D/1$ queue, with two classes of traffic. The departure process from the queues (which in turn is part of the arrival process for the down stream node), with deterministic service, is not Markovian. However, we assume that the load of the virtual circuit is small compared to the local traffic, which we have modelled as Markovian. Therefore, the combination of both classes of traffic is mostly dominated by the local traffic, and is assumed to be Markovian. This assumption is validated by simulation, as presented in the results. Refer to Figure 3 where an end-to-end virtual circuit or virtual path is modelled. This virtual circuit competes for resources (bandwidth etc.) along with the 'background traffic' which is local to each node. It is of interest to determine the behaviour of the virtual circuit at each of the nodes when the background traffic at the first node exhibits different types of nonstationarity. All the queues can be described by the M/D/1 differential equation model with two classes of traffic derived earlier. For analyzing this model, we define the following terms: Let $X_{i1}(t)$, $X_{i2}(t)$, $G_{i1}(t)$, $G_{i2}(t)$, $\lambda_{i1}(t)$, $\lambda_{i2}(t)$, and μC be the average number in the system, average utilization, average arrival rates, average service rates for class I and class II traffic at node i respectively. Let $X_{iT}(t)$ be the total number of packets at node i respectively. Let $\lambda_{12}(t) = \lambda_{vc}(t)$ be the arrival rate of the virtual circuit to the first queue. This virtual circuit passes through the second queue, into the third queue, where it terminates. The departure rate of the virtual circuit from the first queue will be $\mu C G_{12}(t)$, the average utilization of the first queue server in serving the virtual circuit packets. This departure rate then becomes the input to the second queue with a deterministic propagation delay of D time units. Hence $\lambda_{22}(t) = \mu C G_{12}(t - D)$ is the arrival rate of class II traffic to the second queue. Similarly, the average departure rate of class II traffic from the second queue is $\mu C G_{22}(t)$, which, with a time delay of D units, is the average arrival rate for class II traffic to the third queue, i.e $\lambda_{32}(t) = \mu C G_{22}(t - D)$. Notice that the queues are inherently coupled with each other. This is due to the virtual circuit which utilizes resources along all the three queues. We study the behaviour of the various queue lengths when the background traffic to the first queue exhibits different characteristics.

Knowing the average arrival and departure rates for both classes of traffic at all the three queues, we use the results derived in this section for the M/D/1 queue with multiple classes of traffic to write the following differential equations for the three queues (G_{ij}'s are as derived in section 2):

$$\begin{aligned}
\dot{X}_{11}(t) &= -\mu C G_{11}(t) + \lambda_{11}(t). \\
\dot{X}_{12}(t) &= -\mu C G_{12}(t) + \lambda_{12}(t) \quad (i.e \lambda_{vc}(t)). \\
\dot{X}_{21}(t) &= -\mu C G_{21}(t) + \lambda_{21}(t). \\
\dot{X}_{22}(t) &= -\mu C G_{22}(t) + \mu C G_{12}(t - D). \\
\dot{X}_{31}(t) &= -\mu C G_{31}(t) + \lambda_{31}(t). \\
\dot{X}_{32}(t) &= -\mu C G_{32}(t) + \mu C G_{22}(t - D).
\end{aligned}$$

To validate the accuracy of this model, a simulation study was conducted under both stationary and nonstationary traffic conditions. The results of a typical validation study are now given. Additional validation results can be found in [11].

The arrival rate of the background traffic to both the second and third queue was kept at 0.5. The arrival rate for the virtual circuit originating at the first queue, was kept at 0.1. Service time for all the three queues was assumed to be one time unit and each queue was initially empty. The propagation delay, D, was 10 time units. The arrival rate of the background traffic to the first queue was computed by $\lambda_{11}(t) = 0.5 + 0.4\sin(0.2(t + 20))$. The number in system of the different queues are plotted in Figures 4- 6. The virtual circuit traffic is affected all the way to the third node. This supports our belief that congestion at any node in the network spreads to downstream nodes also. Note that the analytical model tracks the simulation results. We achieve significant computational savings also. For example, a typical simulation study like the one in Figure 4, written in SIMAN, re-

quired about 85 minutes, whereas the numerical integration of the differential equations, using the fourth-order Runge-Kutta routine in MATLAB required about 6 minutes on a Sun SPARC station.

Having validated the model with simulation, we now consider different types of nonstationarity for the background traffic at the first node, and investigate how it affects the downstream nodes. Results for one such nonstationarity are now presented. For comparison purposes, we plot the curves for both finite and zero propagation delay. All arrival and service rates are as before. When the propagation delay is considered, we assume it to be $D = 10$ time units.

The background traffic at the first node was assumed to be constant at 0.5 for some time, after which it experiences an overload burst which raises the arrival rate to 1.5 for some time. All other arrival rates remaining the same, the arrival rate of the background traffic to the first queue is modelled as follows:

$$\lambda_{11}(t) = \begin{array}{ll} 0.5 & 0 \leq t < 50 \\ 1.5 & 50 \leq t \leq 125 \\ 0.5 & 125 < t \end{array}$$

Figures 7- 9 plot the number in the system at the different network nodes. At the first queue, following the change of arrival rate, the number in system for the background traffic increases rapidly. This causes the virtual circuit packets to be delayed at the first node, resulting in the number of class II packets to also increase. An interesting phenomena is observed at the second and third queues. As the virtual circuit packets get delayed at the first queue, the arrival rate of the virtual circuit at the second queue actually decreases. This causes the background traffic queue length to also decrease, as more bandwidth is available due to the decreasing virtual circuit arrival rate. Notice that as soon as the first queue 'burst period' ends, the backlogged packets of the virtual circuit are transmitted, causing the arrival rate of the virtual circuit at the second and third queue to increase, and hence the queue length suddenly increases, after which it comes back to its steady state value.

4 Conclusions

In this paper, we present derivations for the M/D/1 queue with multiple classes of traffic. The methodology used is general and can be followed for other queuing systems. Similar equations for the M/D/1 queue operating under a prioritized environment are also developed. These equations are validated using simulation under different traffic conditions. The importance of these equations stems from the fact that the state space dependence of traditional differential equation models is removed. Instead, for a queue with i classes of traffic, only i differential equations are necessary. This maps into a tremendous saving of computation and allows us to study the transient behaviour of a queuing network without actually simulating it. An illustration of the use of the model in studying nonstationary effects in ATM networks is presented.

References

[1] Agnew, C. *Dynamic Modelling and Control of Congestion Prone Systems*, Operations Research, vol. 24, no. 3, pp. 400-419, 1976.

[2] Evans, S. *Performance and Control in B-ISDN: Transient Behaviour and Time Scales*, in Proceedings I.T.C. 7th Specialist Seminar, Adelaide, Australia, Oct., 1990.

[3] Filipiak, J. *Real Time Network Management*, North Holland, 1991.

[4] Glynn, P. W. *Diffusion Approximations*, in Handbooks in Operations Research and Management Science - Vol. 2, pp. 145-196, North Holland , 1990.

[5] Grassmann, W. K. *Computational Methods in Probability Theory*, in Handbooks in Operations Research and Management Science - Vol. 2, pp. 199-250, North Holland, 1990.

[6] Gross, D., Harris, C. M. *Fundamentals of Queuing Theory*, 2nd ed, John Wiley & Sons, New York.

[7] Lovegrove, W., Hammond, J. L., Tipper, D. *Simulation Methods for Studying Nonstationary Behaviour of Computer Networks*, IEEE JSAC, Vol. 8, pp. 1696-1708, December 1990.

[8] Odoni, A. R., Roth, E. *An empirical investigation of the transient behaviour of stationary queuing systems*, Oper. Res., Vol. 31, no. 3, May-June 1983.

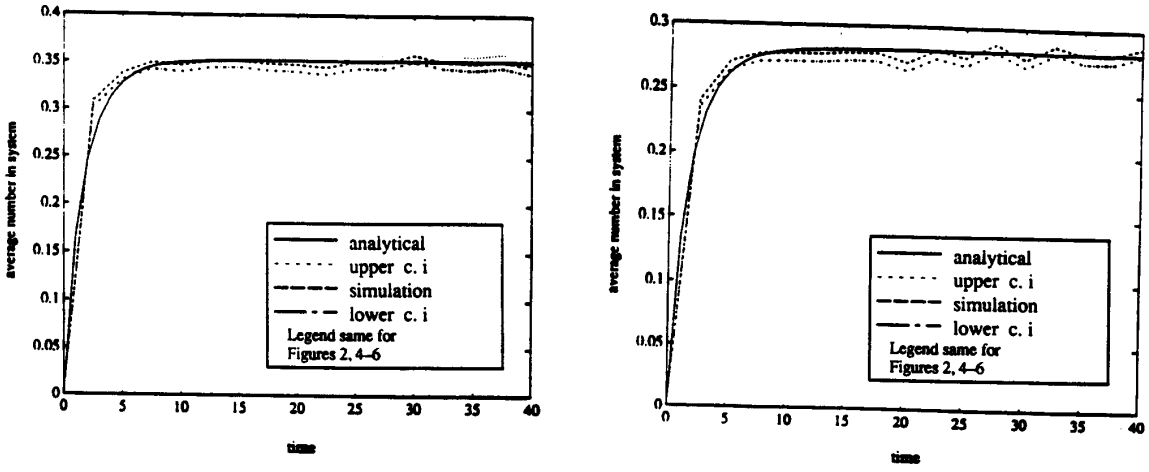
[9] Pitsillides, A., Lambert, J., Li, N., Steiner, J. *Dynamic Bandwidth Allocation in Communication Systems: An Optimal Control Approach*, in Proceedings IEEE International Conference on Systems Engineering, Kobe, Japan, Sept. 1992.

[10] Reibman, A., Trivedi, K. *Numerical Analysis of Markov Models* in Comput. Operations Research, Vol. 15, No. 1. 1988.

[11] Sharma, S. *Approximate Models for the Study of Nonstationary Queues, and their Application to Communication Networks.* Master's Thesis, Clemson University, Dec. 1992,

[12] Tipper, D., Sundareshan, M. K. *Numerical Methods for Modeling Computer Networks Under Nonstationary Conditions*, in IEEE-JSAC, Vol. 8, No. 9, Dec. 1990.

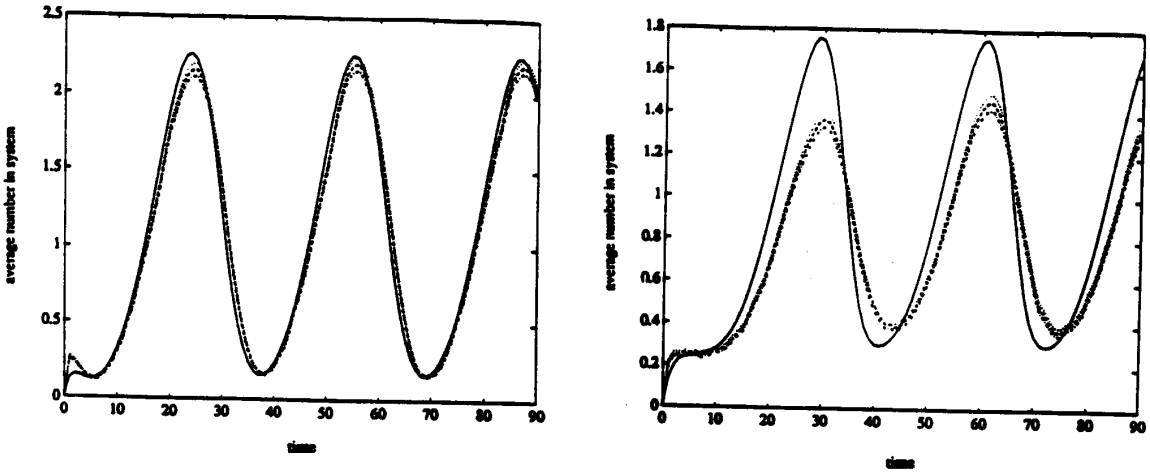
[13] Weiss, A. A., Mitra, D. *A Transient Analysis of a Data Network, With a Processor-Sharing Switch*, in AT&T Technical Journal, Sept 1988.



(a) Behaviour of Class I traffic

(b) Behaviour of Class II traffic

Figure 1: Validation of the M/D/1 model with two classes of traffic, no priority.



(a) Behaviour of Class I traffic

(b) Behaviour of Class II traffic

Figure 2: Validation of the M/D/1 model with two classes of traffic, with priority - nonstationary loads

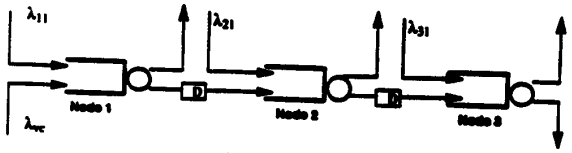
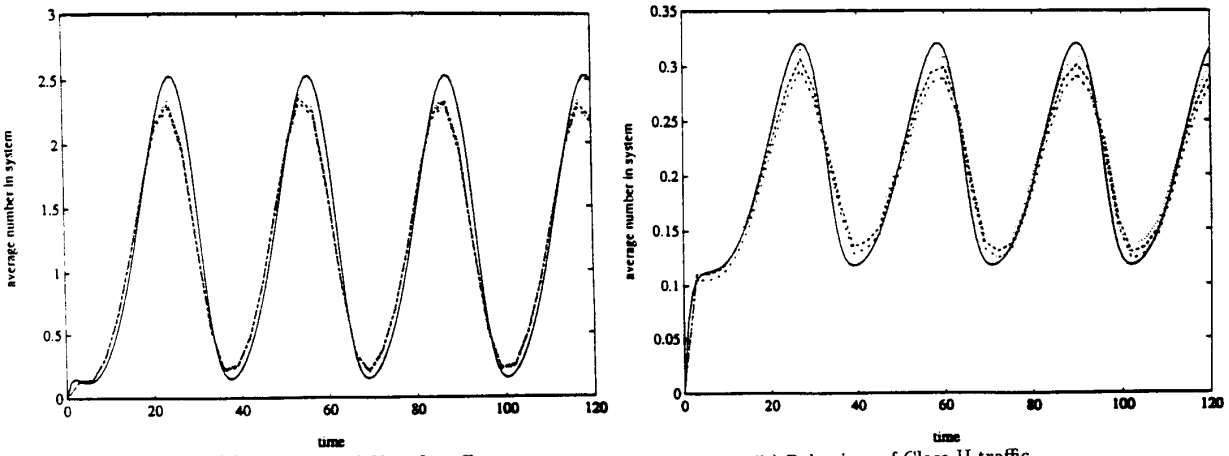
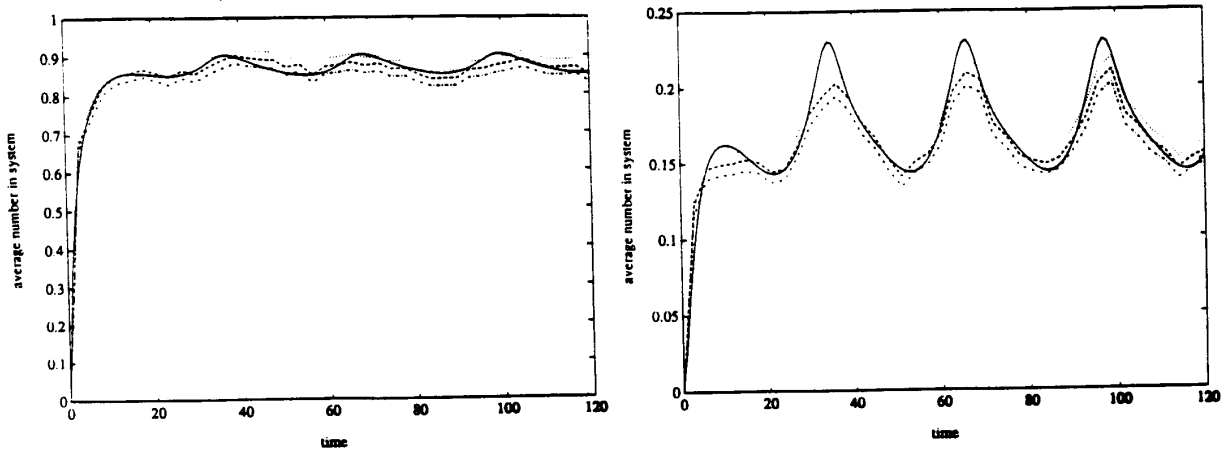


Figure 3: Model of end-to-end virtual circuits in B-ISDN networks



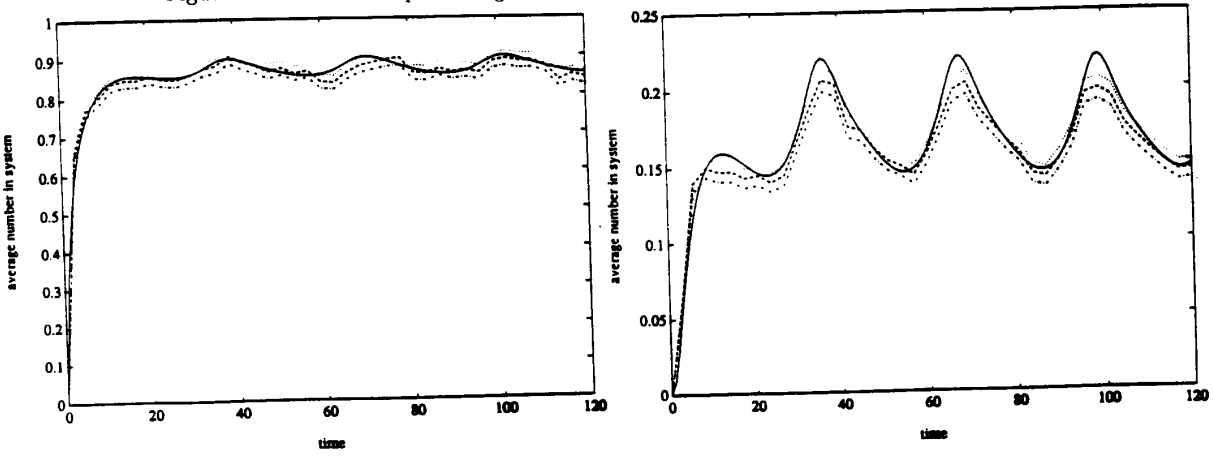
(a) Behaviour of Class I traffic (b) Behaviour of Class II traffic

Figure 4: Validation of queue lengths for nonstationary arrival process - Node 1



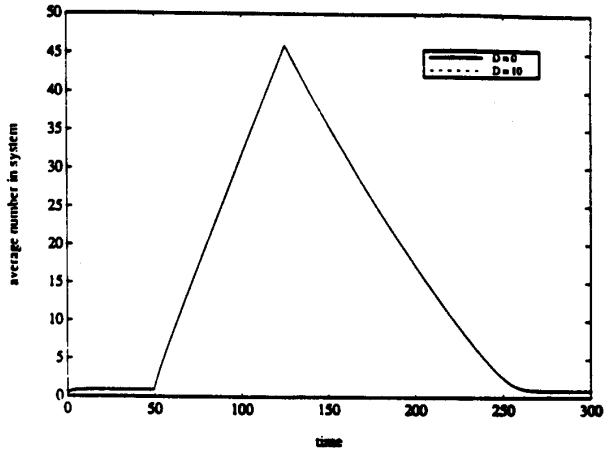
(a) Behaviour of Class I traffic (b) Behaviour of Class II traffic

Figure 5: Validation of queue lengths for nonstationary arrival process - Node 2

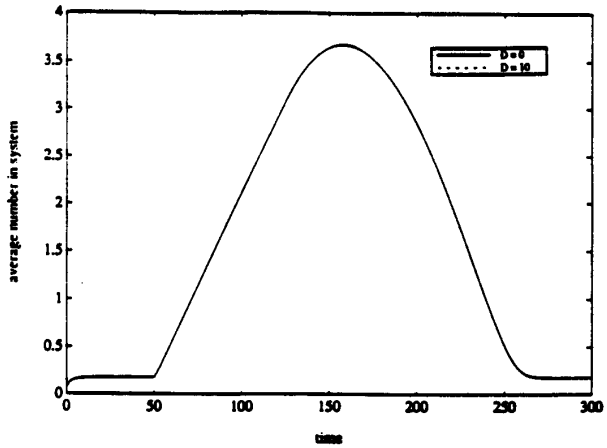


(a) Behaviour of Class I traffic (b) Behaviour of Class II traffic

Figure 6: Validation of queue lengths for nonstationary arrival process - Node 3

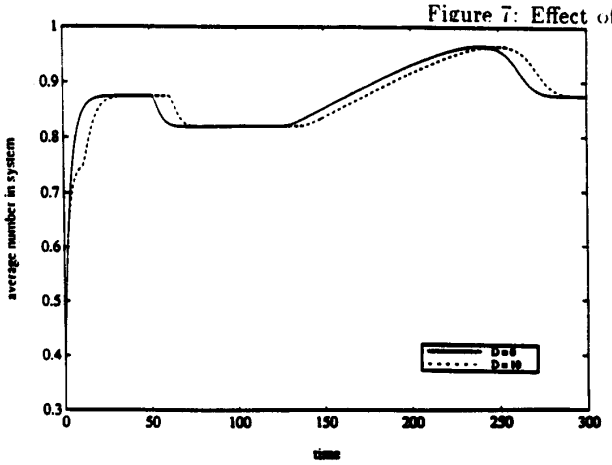


(a) Behaviour of Class I traffic

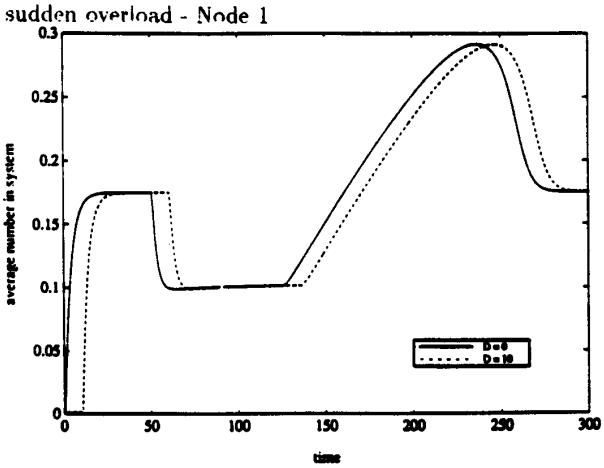


(b) Behaviour of Class II traffic

Figure 7: Effect of sudden overload - Node 1

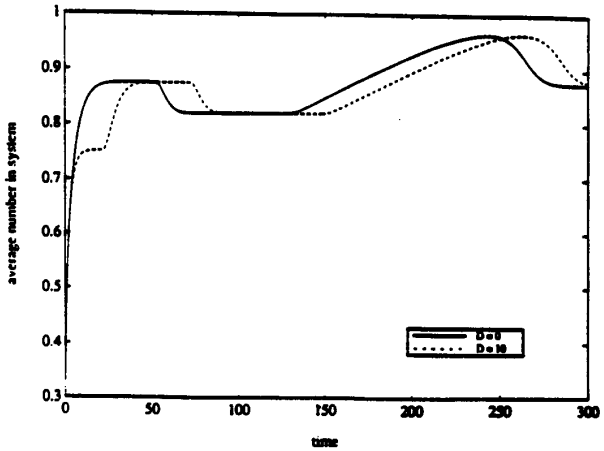


(a) Behaviour of Class I traffic

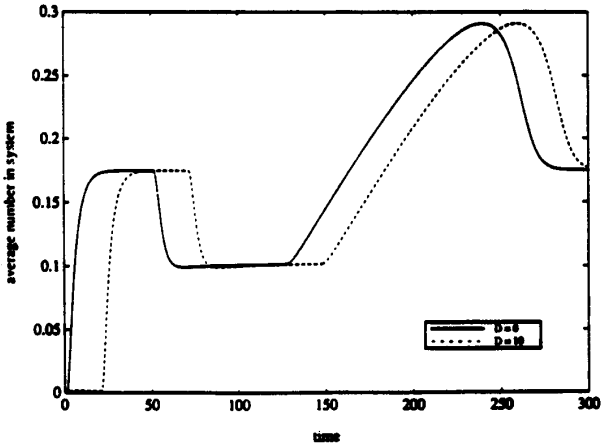


(b) Behaviour of Class II traffic

Figure 8: Effect of sudden overload - Node 2



(a) Behaviour of Class I traffic



(b) Behaviour of Class II traffic

Figure 9: Effect of sudden overload - Node 3