19

Approximate Analysis of a Dynamic Priority Queueing Method for ATM Networks

Anoop Ghanwani
Internetworking Technology, IBM Corporation
Box 12195, Research Triangle Park, NC 27709, USA
Tel: +1-919-254-0260
Fax: +1-919-254-5483
Email: anoop@raleigh.ibm.com

Erol Gelenbe
Department of Electrical and Computer Engineering, Duke University
Box 90291, Durham, NC 27708, USA
Tel: +1-919-660-5442
Fax: +1-919-660-5293
Email: erol@ee.duke.edu

Abstract

A scheduling discipline for multiple classes of traffic in an ATM network is discussed and analyzed. The scheduler has the desirable property of providing minimum bandwidth guarantees for each class of traffic. Its simplicity makes it particularly well suited for high speed implementation. The scheme is a modification of static head-of-line priority queueing, and was originally presented in a slightly different form by Huang and Wu. We begin by considering a system with two queues which is analyzed by decoupling the system into separate $M/G/1$ queues. The analysis is found to provide a very good estimate for the mean response time of customers in each queue. The applicability of the analysis to a system with multiple queues is also demonstrated.

Keywords

Coupled queues, ATM networks, scheduling disciplines

Performance of Information and Communication Systems U. Körner & A. Nilsson (Eds.)

1 INTRODUCTION

In asynchronous transfer mode (ATM) networks, data are transported in fixed size 53 byte cells. The ATM Forum has standardized many classes of service for users' traffic based on the loss and delay requirements of various applications (Jain 1996). In order to meet the service requirements for each class of traffic, it is necessary to provide a scheduling algorithm to decide which class receives service when the server becomes free. Many scheduling algorithms have been proposed and analyzed, ranging from simple scheduling disciplines such as static priority and round robin, to more sophisticated algorithms such as weighted fair queueing and its variants. A discussion of scheduling disciplines for high speed networks may be found in (Zhang 1995) and the references therein.

We consider a priority queueing system with two classes of traffic. A counter is associated with the low priority queue which is incremented whenever a high priority cell is served and a low priority cell is waiting for service. The counter is reset whenever a cell from the low priority queue is served. High priority customers* have non-preemptive priority over low priority customers except when the counter has reached a predefined threshold L. In that case, the head-of-line cell of the low priority queue is served and the counter is reset. The counter may be thought of as a measure of the "impatience" of the cell waiting at the head of the low priority queue. The behavior of the scheduler is completely described as follows:

- If both queues are empty, the server remains idle until a cell arrives to the system.
- If the low priority queue is empty, and there are jobs in the high priority queue, a job from the high priority queue is scheduled for service.
- If the high priority queue is empty, and the low priority queue has cells, then a low priority cell is scheduled for service and the counter is reset.
- If both the queues have customers waiting then:
  - If the value of the counter is less than L, a cell from the high priority queue is scheduled for service, and the value of the counter is incremented by 1.
  - If the value of the counter is equal to L, a cell from the low priority queue is scheduled for service and the counter is reset.

The instantaneous priority of a traffic class depends on the value of L and the arrival rate for each class. This yields a closely coupled queueing system where the degree of coupling depends on L. A closed form solution for the exact mean response time of this system does not exist. A generalized version

*The words "customer" and "cell" are used inter-changeably since we are analyzing an ATM system.

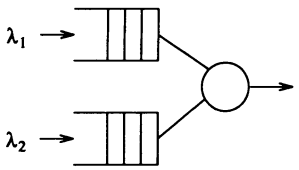

Figure 1 Priority queueing system with two classes of traffic

of this scheme was proposed in (Huang *et al.* 1993) for a system with n priority queues, each having a counter associated with it. When a counter reaches the threshold L_i, $1 \leq i \leq n$, the cell at the head of that queue is scheduled for transmission in the next slot provided no other higher priority queue's counter has exceeded the threshold. The algorithm incurs very little processing overhead; yet it avoids the problem of "starving" lower priority traffic. Our scheme is slightly different in that the first queue does not have an "impatience" counter.

Many adaptive schemes based on static priority and round-robin which attempt to overcome the drawbacks of each have been proposed. Kleinrock (Kleinrock 1976) proposes a model where the instantaneous priority depends on a variable parameter. A model with p classes is considered, each having a parameter b_p associated with it ($0 \leq b_1 \leq b_2 \leq \ldots \leq b_p$). The priority of a class i customer, which arrived at time τ, at time t is then given by $(t - \tau)b_i$. Lim and Kobza (Lim *et al.* 1988) propose a scheme referred to as head-of-line priority with jumps (HOL-PJ). They consider a model with p classes of traffic. Class i has non-preemptive priority over class j if $i < j$. However, a customer has an upper limit on the amount of time it spends in a given queue. If that limit is exceeded, the customer joins the end of the next higher priority queue. Ozawa (Ozawa 1990) studied a system with two queues where the high priority queue receives exhaustive service and the service of the low priority queue is K-limited. Lee and Sengupta (Lee *et al.* 1993) propose and analyze a model with two classes of traffic in an ATM network. The system is serviced using the round-robin service discipline between classes. A threshold L may be defined for either class. If the queue length for the class exceeds the threshold, cells from only that class will be serviced until the queue length falls below the threshold; it then reverts back to round-robin. The analysis of coupled queueing systems such as the ones described above typically involves transform based analysis often leading to numerical solutions. Closed form solutions hard to achieve without making simplifying approximations about the behavior of the system. Our focus is a very simple method for approximating the mean response time of the system described above.

The remainder of this paper is organized as follows. In Section 2, the system with two queues is analyzed and the results are compared with the mean response time obtained from discrete event simulation. Section 3 shows how the analysis may be extended to a multi queue system. Again results are presented

to compare the analytical approximation with simulation. The conclusions of our work are presented in Section 4.

2 NOTATION AND ANALYSIS

We use a time-slotted model where the duration of a slot is the time required to service a single cell. The arrival process at queue i is assumed to be Poisson with rate parameter λ_i. Let α_i be the stationary probability that queue i is busy, i.e. that there are cells either in service or waiting to be served. Let q_i be the stationary conditional probability that the head-of-line cell in queue i receives service given that both high and low priority queues have cells waiting to be served. We use the suffix 1 to denote the high priority traffic class and suffix 2 to denote the low priority traffic class. We make the following approximation to account for the behavior of the scheduler. When both queues are busy, the low priority queue will on average receive service 1 out of every $(L+1)$ slots. Therefore, we can set $q_2 = \frac{1}{L+1}$ and $q_1 = 1 - q_2$. Then, the probability that queue i is busy is given by

$$\alpha_i = \lambda_i E[S_i], \tag{1}$$

where S_i is a random variable which denotes the number of slots between the time a class i customer gets to the head-of-line, to the time when it leaves the system. Note that S_i consists not only of the amount of time that the server will be kept busy by the cell, but also includes the time that the cell spends at the head of the queue waiting to access the server. In other words, S_i is the sum of access time and service time, where access time is a random variable which accounts for the time that the cell waits before it gets access to the server, and the service time is a single slot. Let k be the number slots that a cell spends at the head-of-line position before getting service. We approximate S_1, S_2 by geometrically distributed random variables with means:

$$E[S_1] = \sum_{k=0}^{\infty}(k+1)(\alpha_2 q_2)^k(1-\alpha_2 q_2) = \frac{1}{1-\alpha_2 q_2}, \tag{2}$$

$$E[S_2] = \sum_{k=0}^{\infty}(k+1)(\alpha_1 q_1)^k(1-\alpha_1 q_1) = \frac{1}{1-\alpha_1 q_1}, \tag{3}$$

and second moments:

$$E[S_1^2] = \sum_{k=0}^{\infty}(k+1)^2(\alpha_2 q_2)^k(1-\alpha_2 q_2) = \frac{1+\alpha_2 q_2}{(1-\alpha_2 q_2)^2}, \tag{4}$$

$$E[S_2^2] = \sum_{k=0}^{\infty}(k+1)^2(\alpha_1 q_1)^k(1-\alpha_1 q_1) = \frac{1+\alpha_1 q_1}{(1-\alpha_1 q_1)^2}. \tag{5}$$

Substituting (2) and (3) in (1), we can write $\alpha_1 = \frac{\lambda_1}{1-\alpha_2 q_2}$ and $\alpha_2 = \frac{\lambda_2}{1-\alpha_1 q_1}$. These equations may be solved simultaneously to yield a quadratic equation in either α_1 or α_2. The root of interest can be found by using the additional criterion $\lambda_i \leq \alpha_i \leq 1$. Note that since the service time of a customer is a single slot, it is required that $\lambda_1 + \lambda_2 < 1$ for stability.

This approximate analysis allows us to decouple the system into separate queues, each with its own arrival rate and service time. To compute the mean waiting time, we apply standard results for an $M/G/1$ queueing system (Gelenbe *et al.* 1980) separately to each queue as follows:

$$W_i = \frac{\lambda_i E[S_i^2]}{2\,[1-\lambda_i E[S_i]]}.$$

The mean response time is then $R_i = W_i + E[S_i]$. From comparison with simulation, we find that using these results, an accurate estimate of the mean response time for the high priority traffic class is obtained. However, the results are not as good for the low priority traffic class. We therefore make use of the conservation law applicable to $M/G/1$ queues when the service discipline is work-conserving. The law states that (Kleinrock 1976):

$$\sum_i \rho_i W_i = \frac{W_0}{1-\rho}, \tag{6}$$

where $W_0 = \sum_i \frac{\lambda_i E[S_i^2]}{2}$. In our case, the RHS of Equation (6) becomes $\frac{\lambda_1+\lambda_2}{1-\lambda_1-\lambda_2}$. W_2 is then computed as:

$$W_2 = \frac{\frac{(\lambda_1+\lambda_2)^2}{2(1-\lambda_1-\lambda_2)} - \lambda_1 W_1^*}{\lambda_2},$$

where $W_1^* = R_1 - 1$. The mean response time for the low priority queue is then $R_2 = W_2 + 1$.

The mean response times using the analytical approximation are compared with results from discrete event simulation in Figures 2–7. In each case, the traffic load on the high priority queue is a constant value (either 30% or 50% of the server capacity); the load on the low priority queue is varied from very light until a value which saturates the system.

The figures indicate that the approximation yields very accurate response times for most of the cases tested. In most instances, the error between the analytical and simulation results is less than 10%. It performs especially well when the system is light to moderately loaded (up to 60–70% load). The

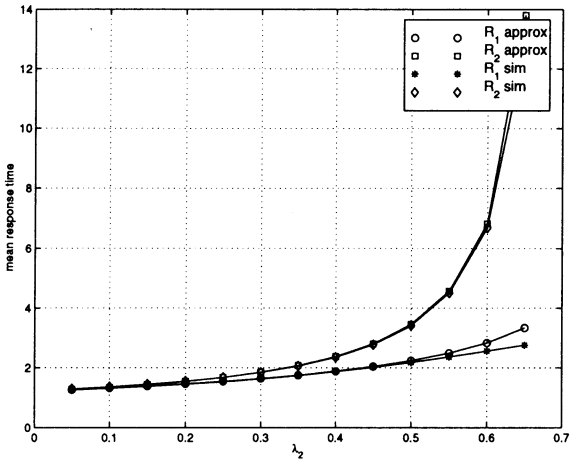

Figure 2 Results for $\lambda_1 = 0.3$, $L = 1$

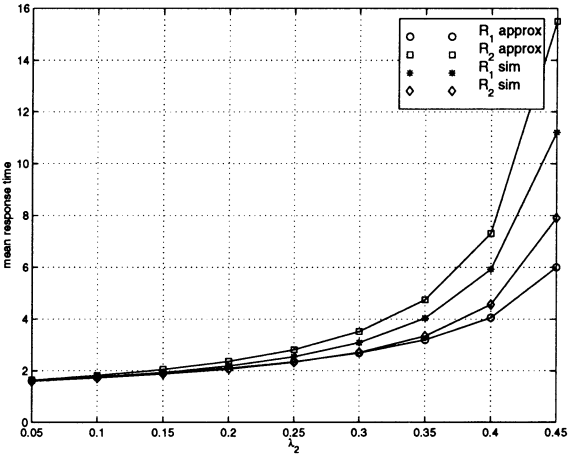

Figure 3 Results for $\lambda_1 = 0.5$, $L = 1$

approximation tends to produce less accurate results in cases where the L is very small and the load is high (Figure 3). This is likely due to the fact that in this instance, the queues are highly coupled, and the approximation based on decoupling yields inaccurate results. In fact, a system with $L = 1$ is essentially equivalent to a polling system.

3 ANALYZING A SYSTEM WITH MULTIPLE QUEUES

The queueing analysis presented in Section 2 may be used for analyzing systems with more than two queues. The procedure is as follows. Consider a

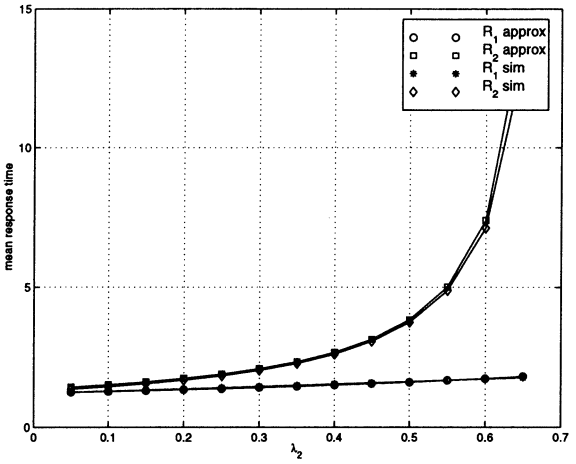

Figure 4 Results for $\lambda_1 = 0.3$, $L = 3$

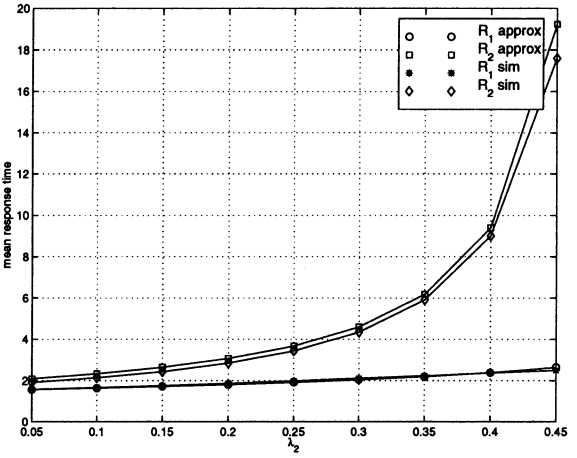

Figure 5 Results for $\lambda_1 = 0.3$, $L = 3$

system of n queues. Queues 2 through n each have a threshold $L_i \in Z^+$. In order to be able to guarantee a minimum bandwidth of $\frac{1}{L_i+1}$ for class i, it is required that $\sum_{i=2}^{n} \frac{1}{L_i+1} < 1$. We assume that the arrival process at queue i is Poisson with rate parameter γ_i. For a stable system, we also require $\sum_{i=1}^{n} \gamma_i \leq 1$.

First, the system is solved by reducing it to a two queue system — the first queue and all the others put together. For this case, using the notation defined in the previous sections, the arrival rates for the two queues are given by: $\lambda_1 = \gamma_1$, $\lambda_2 = \sum_{i=2}^{n} \gamma_i$. For this system with two queues, the value of L for the second queue is given by $L = \frac{1}{\sum_{i=2}^{n} \frac{1}{L_i+1}} - 1$. The two queue system is

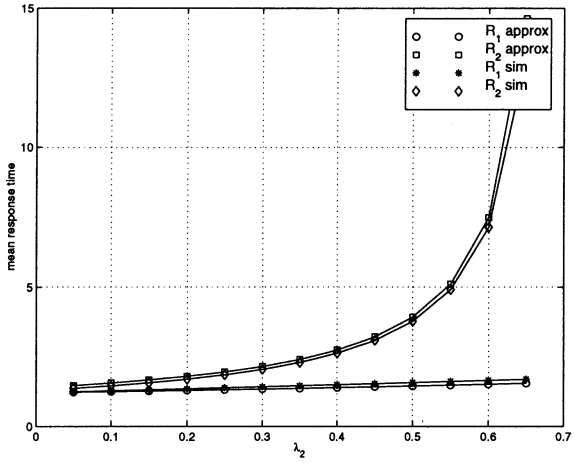

Figure 6 Results for $\lambda_1 = 0.3$, $L = 5$

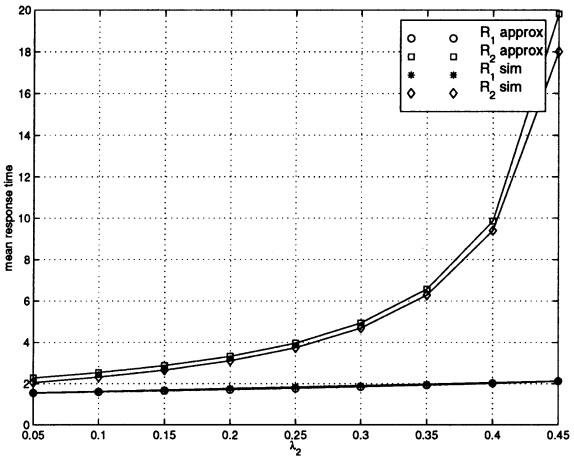

Figure 7 Results for $\lambda_1 = 0.5$, $L = 5$

then solved using the method outlined in Section 2 to yield the mean response times for the first queue, and the mean response time for all the other queues. Next, we go through the same procedure described above with the first two queues corresponding to one queue and all the others corresponding to the the second queue. This will yield the mean response time for the first two queues combined, and the mean response time for the rest of the queues. Then, using the law of conservation, we can compute the mean response time for the second queue by itself. In this way, the 2 queue system must be solved $n - 1$ for an n queue system yielding the mean response time each class. The following steps summarize the procedure for solving a system with n queues.

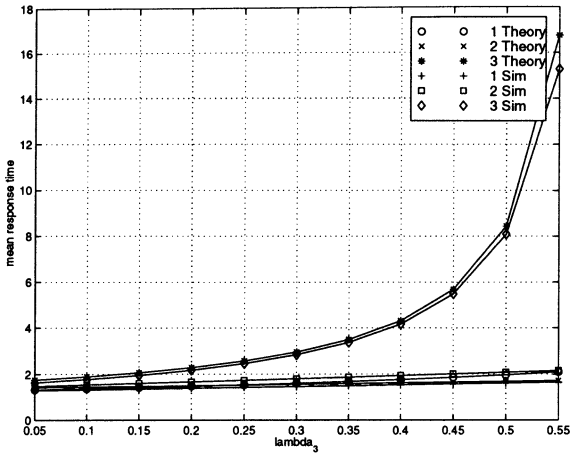

Figure 8 Results for $\gamma_1 = 0.2$, $\gamma_2 = 0.2$, $L_2 = 4, L_3 = 4$

- **Step 1**. Set $m \leftarrow 0$.
- **Step 2**. $m \leftarrow m + 1$.
- **Step 3**. Create a two queue system with parameters:

$$\lambda_1 = \sum_{j=1}^{m} \gamma_j, \quad \lambda_2 = \sum_{j=m+1}^{n} \gamma_j, \quad L = \frac{1}{\sum_{j=m+1}^{n} \frac{1}{L_j+1}} - 1.$$

- **Step 4**. Use the analysis of Section 2 to compute the mean response time R_1 and R_2 for the two queue system.
- **Step 5**. The mean response time for queue m is:

$$R'_m = \begin{cases} R_1 & \text{if } m = 1; \\ R_1 - \sum_{j=1}^{m-1} (R'_m - 1)\gamma_j & \text{otherwise.} \end{cases}$$

- **Step 6**. If $m < n - 1$, go to Step 2, else the mean response time for the n^{th} queue is given by $R'_n = R_2$.

Results of this method for a system with three queues is presented in Figures 8 and 11. Again, we see that the approximation is very good except when the equivalent value for L is small and the load on the system is high. In the scenario of Figure 9, the value of L in the first iteration is $\frac{1}{\frac{1}{5}+\frac{1}{5}} - 1 = 1.5$.

4 CONCLUSIONS

An adaptive queueing discipline for ATM network nodes with two classes of traffic is analyzed. An approximation is used in which the two queues are

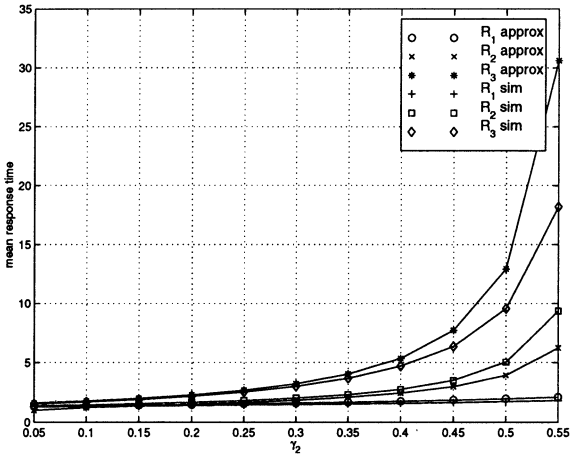

Figure 9 Results for $\gamma_1 = 0.2$, $\gamma_3 = 0.2$, $L_2 = 4, L_3 = 4$

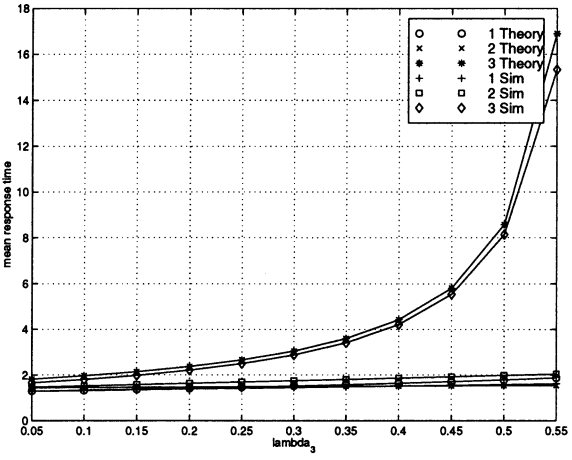

Figure 10 Results for $\gamma_1 = 0.2$, $\gamma_2 = 0.2$, $L_2 = 4$, $L_3 = 6$

decoupled for the purpose of analysis. We also demonstrate how this approach may be used to analyze systems with more than two queues. The analytical approximation is compared with results from discrete event simulation and is found to work very well under a variety of traffic conditions for systems with two and three queues.

REFERENCES

Gelenbe, E. and Mitrani, I. (1980) *Analysis and Synthesis of Computer Systems.* Academic Press.

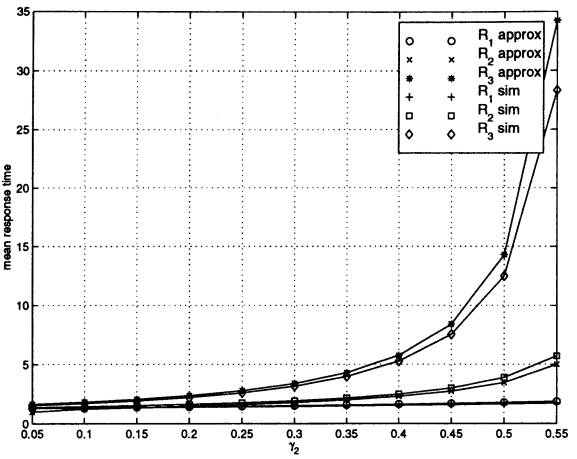

Figure 11 Results for $\gamma_1 = 0.2$, $\gamma_3 = 0.2$, $L_2 = 4$, $L_3 = 6$

Huang, T.-Y and Wu, J.-L. C. (1993) Performance analysis of a dynamic priority scheduling method in ATM networks. *IEE Proceedings-I*, **140**, 285–290.

Jain, R. (1996) Congestion control and traffic management in ATM networks: Recent advances and a survey. *Computer Networks and ISDN Systems*, **28**, 1723–1738.

Kleinrock, L. (1976) *Queueing systems, Volume II: Computer applications.* John Wiley and Sons.

Lee, D.-S. and Sengupta, B. (1993) Queueing analysis of a threshold based priority scheme for ATM networks. *IEEE/ACM Transactions on Networking*, **1**, 709–717.

Lim, Y. and Kobza, J. (1988) Analysis of a delay-dependent priority discipline in a multiclass traffic packet switching node. in *Proc. IEEE INFOCOM.*

Ozawa, T. (1990) Alternating service queues with mixed exhaustive and K-limited service. *Performance Evaluation*, **11**, 165–175.

Zhang, H. (1995) Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE*, **83**, 1374–1396.

5 BIOGRAPHY

Anoop Ghanwani received the Bachelor of Engineering in Electronics and Telecommunications Engineering from the Govt. College of Engineering, Pune, India in 1992. He received the Master of Science in Electrical Engineering from Duke University in 1995, and is presently enrolled in the doctoral program. Since August 1996, he has been working as Staff Engineer with the Internetworking Technology department at IBM in the Research Triangle Park, NC, USA. His research interests include routing, scheduling and bandwidth man-

agement in high speed networks.

Erol Gelenbe is the Nello L. Teer Jr. Professor of Electrical and Computer Engineering at Duke University, and is also Professor of Computer Science and of Psychology-Experimental. He has authored four books on queueing systems and computer and communication system performance, and some 100 journal papers. His former doctoral students are active in academic and industrial research in Europe and the US. His honors include *Fellow of the IEEE (1986), Chevalier de l'Ordre du Merite (France, 1992), Dott. Ing. "Honoris Causa" of the University of Rome (Italy, 1996), Grand Prix France Telecom (French Academy of Sciences, 1996), Science Award of the Parlar Foundation (Turkey, 1995).* Erol's interests cover computer-communication networks and distributed systems, computer performance analysis, artificial neural networks and image processing. In the area of networks, recent work has included CAC in ATM, as well new product form queueing networks. His applied work since 1993 includes designing search algorithms in probabilistic environments, novel algorithms for explosive mines, automatic target recognition, brain imaging and video compression. Currenty his research is funded by the Computational Neurosciences Program of the Office of Naval Research, the U.S. Army Research Office, the Multidisciplinary University Research Initiative on Demining (MURI-ARO), and IBM. Erol is an Associate Editor of several journals including *Acta Informatica, Proceedings of the IEEE, Telecommunication Systems, Performance Evaluation, Journal de Recherche Opérationnelle, Information Sciences, Simulation Practice and Theory, and RESIM: Réseaux et Systèmes Multimedia.*