# Approximate Analysis of Coupled Queueing in ATM Networks

Erol Gelenbe, *Fellow, IEEE* Anoop Ghanwani

*Abstract*— **We discuss a scheduling discipline for multiclass traffic in ATM network nodes. The scheduler provides minimum bandwidth guarantees for each class of traffic and is 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 provides a very good estimate for the mean response time of customers in each queue. We also demonstrate the applicability of the analysis to a system with $n \geq 2$ queues.**

*Keywords*— **Coupled queues, packet scheduling, ATM networks.**

## I. 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 [3]. 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 [7] and the references therein. Algorithms for bandwidth allocation and quality of service (QoS) in ATM networks with multiple classes of traffic are discussed in [5][6].

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[1] 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:

E. Gelenbe is with the School of Computer Science at University of Central Florida, Orlando, FL. Email: erol@cs.ucf.edu.

A. Ghanwani is with Nortel Networks, Billerica, MA. Email: aghanwan@nortelnetworks.com. At the time of this work, he was a Ph.D. candidate in the Department of Electrical and Computer Engineering at Duke University, Durham, NC.

[1]The words "customer" and "cell" are used interchangeably since we are analyzing an ATM system.

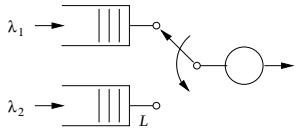


Fig. 1. A dynamic priority queueing system with two classes of traffic

- 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 of this scheme was proposed in [2] 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.

## II. 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 $\lambda_i$. Let $\alpha_i$ 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

$$\alpha_i = \lambda_i E[S_i], \quad (1)$$

where $S_i$ is a random variable which denotes the the number slots between the time a class $i$ customer becomes 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 the amount of 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. Then, $S_i$ is approximated by a geometrically distributed random variable with means

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

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

and second moments

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

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

Substituting Eq. (2) and Eq. (3) in Eq. (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 $\alpha_1$ or $\alpha_2$. 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 [1] 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 these results provide an accurate estimate of the mean response time for the high priority traffic class. However, the results are not as good for the low priority traffic class. This result can be explained as follows. The calculation of the mean response time requires the mean and the second moment of the service time, $S_i$. In the approximate analysis presented above, $S_i$ is a random variable on the infinite set of positive integers. For the exact case, however, $S_1$ is a random variable on the finite set $\{1, 2\}$, and $S_2$ is a random variable on the finite set $\{1, 2, \ldots, L+1\}$. Therefore, while the mean service time using the approximation is quite accurate, the variance tends to be less accurate, especially as the load on the system is increased and/or when $L$ is small. The numerical value of the inaccuracy in the variance of the service time is smaller for the first class of traffic because of the nature of the queueing discipline, especially for larger $L$.

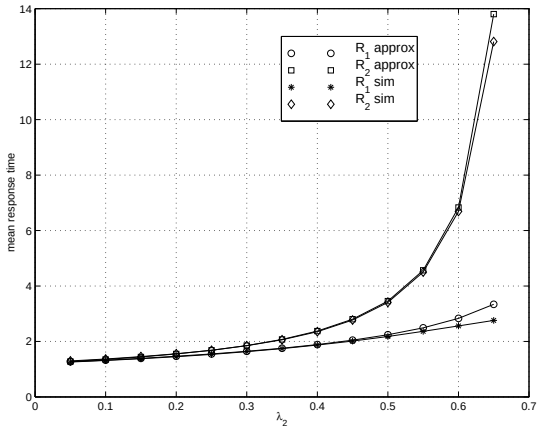


Fig. 2. Results for $\lambda_1 = 0.3$, $L = 1$

In order to get around the inaccuracy of this method for Class 2 traffic, we use of the conservation law applicable to $M/G/1$ queues when the service discipline is work-conserving. The law states that [4]:

$$\sum_i \rho_i W_i = \frac{\rho W_0}{1-\rho}, \quad (6)$$

where $W_0 = \sum_i \frac{\lambda_i E[S_i^2]}{2}$. In our case, the RHS of Eq. (6) becomes $\frac{(\lambda_1+\lambda_2)^2}{2(1-\lambda_1-\lambda_2)}$. $W_2$ 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 2–3. In each case, the traffic load on the high priority queue is a constant value (30% 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

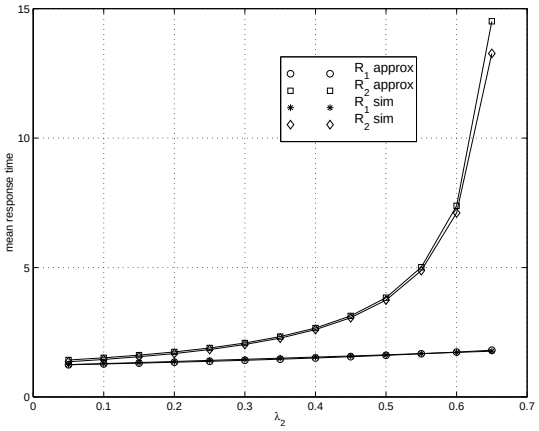


Fig. 3. Results for $\lambda_1 = 0.3$, $L = 3$

when the system is light to moderately loaded (up to 60–70% load). The approximation tends to produce less accurate results in cases where the $L$ is very small and the load is high (Figure 2). This is likely due to the fact that in this instance, the queues are highly coupled and the approximation, based on decoupling, therefore yields inaccurate results. In fact, a system with $L = 1$ is essentially equivalent to a polling system.

## III. Analyzing a System With Multiple Queues

The queueing analysis presented in Section II may be used for analyzing systems with more than two queues. The procedure is as follows. Consider a 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 $\gamma_i$. For a stable system, we also require $\sum_{i=1}^{n} \gamma_i \leq 1$. The solution of the system is obtained by the following steps:

- **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 II 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:

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

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

Thus, a system with $n$ queues is solved by successively solving $(n-1)$ 2-queue systems. Figure 4 illustrates the above procedure for a system with 3 queues. The result of applying this method to a system with three queues is

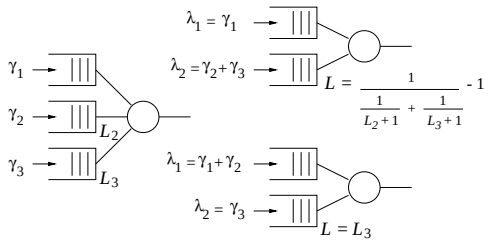


Fig. 4. Analyzing a system with $n$ queues requires the solution of $(n-1)$ systems with two queues

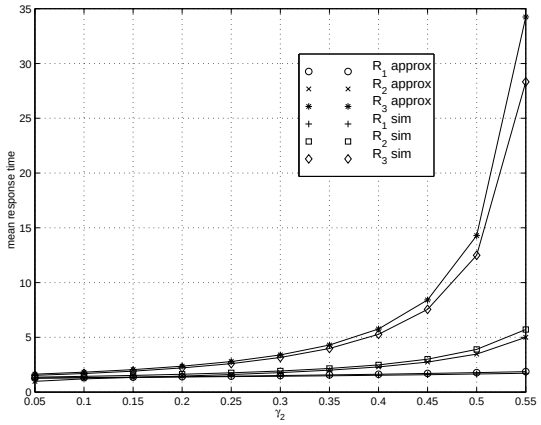


Fig. 5. Results for $\gamma_1 = 0.2$, $\gamma_3 = 0.2$, $L_2 = 4$, $L_3 = 6$

presented in Figure 5. Again, the analysis is found to yield very good results, except when the load on the system is very high.

## IV. 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 decoupled for the purpose of analysis. We also demonstrate how this approach may be used to analyze systems with more than two queues. The 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


[1] E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems*. Academic Press, 1980.

[2] T.-Y. Huang and J.-L. C. Wu. Performance analysis of a dynamic priority scheduling method in ATM networks. *IEE Proceedings-I*, 140(4):285–290, August 1993.

[3] R. Jain. Congestion control and traffic management in ATM networks: Recent advances and a survey. *Computer Networks and ISDN Systems*, 28(13):1723–1738, October 1996.

[4] L. Kleinrock. *Queueing Systems, Volume II: Computer Applications*. John Wiley & Sons, 1976.

[5] K. Sriram. Dynamic bandwidth allocation and congestion control schemes for voice and data multiplexing in wideband packet technology. *Proceedings of the IEEE ICC*, pages 1003–1009, April 1990.

[6] K. Sriram. Methodologies for bandwidth allocation, transmission scheduling, and congestion avoidance in broadband ATM networks. *Computer Networks and ISDN Systems*, 26:43–59, 1993.

[7] H. Zhang. Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE*, 83(10):1374–1396, October 1995.