# Queueing Study of a 3-Priority Policy with Distinct Service Strategies

Randall Landry, *Student Member, IEEE,* and Ioannis Stavrakakis, *Senior Member, IEEE*

*Abstract*—**In this paper, a discrete-time, single server, 3-queue system is presented and analyzed. A distinct service strategy, namely the consistent-gated (c-G), 1-limited (L) and Head-of-Line (HoL), is applied to each of the queues (c-G/L/HoL policy). It is shown that this queueing system provides for an accurate analytical model for a DQDB station, as well as a means for an approximate evaluation of the correlation associated with key traffic processes in that network. In addition, the developed queueing system could be useful for the modeling of the queueing behavior of an ATM link shared by high-priority, low priority and control traffic. Through an asymptotic analysis under heavy low-priority traffic, the worst case performance for the high priority traffic is determined. Furthermore, it is illustrated that the asymptotic analysis provides for a potentially tight delay bounding technique. Finally, the delay performance of the developed queueing system is compared to that of a similar system in which one of the queues receives 1-limited service and the other two exhaustive ($HoL^-/L/HoL^+$ policy).**

## I. Introduction

DISCRETE-TIME queueing models are frequently developed for the study of packetized communication networks [1]–[4]. In particular, the evaluation of new switching technologies such as ATM (Asynchronous Transfer Mode) and network standards like DQDB (Distributed Queue Dual Bus) and FDDI (Fiber Distributed Data Interface), is heavily dependent upon the development of sophisticated queueing models, which frequently involve multi-priority service disciplines [4]–[19].

Token-passing protocols such as FDDI are often modeled as cyclic polling systems, in which a single server attends to queues in a fixed and cyclic order. The service strategy at a queue can usually be described by one of three basic policies and their variations: *limited* service, in which a specified number of customers are served during each visit by the server; *gated* service, in which case only customers present at the queue upon arrival of the server are considered for service; or *exhaustive* service, in which case the server remains at the queue until it becomes empty. Detailed surveys on recent work in polling systems may be found in [15], [16].

The exact analysis of polling systems has, for the most part, been limited to a few special cases [18]. Specifically, for polling systems with cyclic order of service and either purely gated or purely exhaustive service policies, detailed analyses exist for an arbitrary number of queues. However, as the order of service becomes noncyclic and/or service policies among the queues become mixed, exact analysis becomes difficult at best. In [19], a discrete-time pseudoconservation law for mean waiting times is derived for a system of queues with mixed cyclic-service strategies and nonzero switch-over times between queues.

A discrete-time system of two queues served according to two distinct policies has been considered in [20]. Queue $Q^L$ (associated with low-priority customers) receives limited service while queue $Q^H$ (associated with high-priority customers) receives consistent-gated service. The service policy in [20] is characterized as *consistent* in the sense that it guarantees that no low-priority customer will be served before a high-priority customer that arrived at an earlier time (or simultaneously). A continuous-time system of two queues with mixed exhaustive and k-limited services has been studied in [21].

In this paper, a discrete-time system of three queues ($Q^L$, $Q^H$ and $Q^P$) with mixed service policies is considered. Note that the adopted service policy was originally introduced in continuous time in [9], and referred to as the *quasi-gated* discipline in [10], [11]. Although the resulting queueing system could be viewed as a step toward the consideration of more than two queues with mixed services, its study has been motivated by the following major potential applications.

The three-queue system with the adopted mixed service policies is shown here, as well as in [10], [11], to model accurately the queueing behavior of a station in the DQDB metropolitan area network (MAN). In fact, the only point of approximation is introduced by the approximate modeling of the arrival processes to each of the three queues; otherwise, the three-queue system is an exact model for the queueing behavior of a DQDB station. This work marks the first time that correlated arrivals are considered for one of the key network traffic processes, thereby increasing the model's accuracy. Detailed discussion on this application, which is motivated by the work in [9]–[11], may be found in Section III.

In addition to the DQDB application, the mixed service policy of the developed three-queue system is an appealing candidate for implementing network resource sharing or bandwidth allocation in an ATM environment [6], [7]. The two-priority service policy mentioned earlier is inadequate in capturing the diversified service requirements in the integrated traffic ATM environment. In such an environment,

Manuscript received December 7, 1992; revised August 23, 1993; approved by IEEE/ACM Transactions on Networking Editor Hideaki Takagi. This work was supported by the National Science Foundation under grant NCR-9011962.

The authors are with Department of Electrical Engineering and Computer Science, University of Vermont, Burlington, VT 05405 (email: rlandry@emba.uvm.edu) (email: stavraka@uvm.edu).,

IEEE Log Number 9214671.

a limited service policy would be meaningful for *delay-insensitive* traffic which, if not properly controlled, could temporarily monopolize network resources, causing severe degradation of the service provided to *delay-sensitive* traffic. An asymptotic analysis (Section V) establishes the inherent capability of the limited service policy in protecting network resources against "misbehaving" delay-insensitive traffic; long file transfers could represent such traffic. On the other hand, a gated service policy for delay-sensitive traffic guarantees that all the accumulated delay-sensitive traffic will be served before the network resources become available to the delay-insensitive traffic. At the same time the *accumulation horizon*, or the time interval over which delay-sensitive traffic is not being served, is controlled by the provision of limited service to delay-insensitive traffic. Depending upon the structure of each class of traffic, a specific type of limited service could be determined so as to provide a required quality of service to delay-sensitive traffic; video and voice represent such traffic.

In the integrated traffic ATM environment, a third major class of traffic is assumed to be present. This type of traffic consists of critical, network control and reservation information, the delay of which adds significantly to the deterioration of the overall service provided by the network. The volume of this *control* traffic is usually small, but its service requires immediate availability of the network resources, which would otherwise be used for the service of delay-insensitive and delay-sensitive traffic. The network resources can be assumed to adequately serve the control traffic and, thus, the evaluation of the quality of service provided to this third type of traffic is not an issue. What is of interest in this case is the evaluation of the induced degradation of service provided to the delay-insensitive and, in particular, the delay-sensitive traffics. The resulting queueing model for the resource-sharing policy described above, would be that of a three-queue system with distinct services, or, equivalently, that of a two-queue system with distinct services and vacations.

In the next section the three-queue system with distinct services is described and analyzed. The applicability of this system for the modeling of the queueing behavior of a DQDB station is presented in Section III. Numerical results for the general three-queue system, as well as the DQDB stations, are presented in Section IV. In view of the fact that the three-queue system with distinct services exactly describes the service policy of a DQDB station, the accuracy of the results suggests that the adopted modeling of key traffics is a well performing one and, thus, could be used in other DQDB studies. An asymptotic analysis under heavy delay-insensitive (low-priority) traffic is presented in section V, and the worst case performance for delay-sensitive (high-priority) traffic is established. In addition, it is illustrated that the asymptotic result could be used as a bounding technique which could generate very tight bounds on the numerically obtained performance bounds. Finally, a comparison between the proposed three-queue system with distinct services and a three-queue system with distinct services which treats the delay-sensitive traffic as control traffic, is presented in the last section.

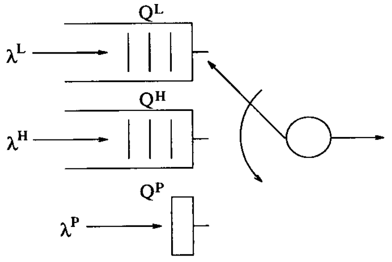


Fig. 1. The discrete-time queueing system.

## II. Analysis Of The Three-Queue System

### A. Description of the Queueing System

In this sub-section the three-queue system is described in detail (Fig. 1). Time is assumed to be slotted and the service time of all customers deterministic and equal to the duration of one slot. Customers will be called packets in the rest of the paper. Packet arrivals and completion of service are assumed to occur at the slot boundaries.

The packet arrival processes to $Q^L$ and $Q^H$ are assumed to be Bernoulli processes with rates $\lambda^L$ and $\lambda^H$, respectively. The packet arrival process to $Q^P$ is assumed to be correlated, delivering at most one packet per slot. The latter process is described by a two-state Markov chain; one packet is delivered to $Q^P$ when the chain is in state 1 and no packet is delivered when the chain is in state 0. The mean number of packet arrivals per slot is given by the steady-state probability that the chain is in state 1, and denoted by $\lambda^P$. The fact that this process delivers at most one packet per slot, together with the fact that these arrivals receive Head-of-Line priority (see below), implies that a buffer capacity of one packet is sufficient for the dimensioning of $Q^P$. $Q^L$ and $Q^H$, on the other hand, are assumed to be of infinite capacity.

It should be noted that the assumptions on the arrival processes considered above are not to be viewed as critical, since they may be relaxed in a number of ways. For instance, the analysis is directly applicable to the case in which the arrival process to $Q^H$ is a general independent and identically distributed (i.i.d.) process with arbitrarily distributed batch size over a time slot, and/or the arrival process to $Q^L$ is an i.i.d. process with geometrically distributed batch size. In both cases, some of the equations will be slightly modified. A similar analysis procedure may be followed when the arrival process to $Q^H$ is correlated, resulting in increased complexity of the numerical solution.

Packets arriving to $Q^P$, $Q^H$ and $Q^L$ will be called P-packets, H-packets and L-packets, respectively. The service policy for the three-queue system, denoted by c-G/L/HoL, is described by the following rules.

(a) The system is work conserving (WC) and nonpreemptive (NP). Non-preemption is guaranteed by the deterministic service time of one slot and by the fact that packet arrivals (and service completion) are assumed to occur at the end of a slot.

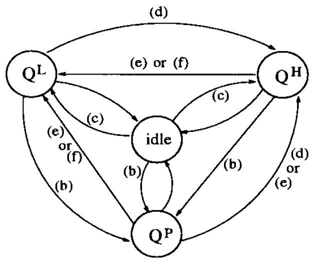


Fig. 2. State diagram for the c-G/L/HoL policy.

(b) $Q^P$ is served according to the HoL priority policy. That is, P-packets begin receiving service immediately. In view of the adopted arrival process to $Q^P$, the previous implies that P-packets suffer a delay of one time slot (service time).

(c) If an H-packet (L-packet) arrives to a previously empty system, the server visits and starts serving $Q^H$ ($Q^L$) at the beginning of the next slot. If an L-packet and an H-packet arrive simultaneously to a previously empty system, the server switches to $Q^H$ and serves the H-packet. It then operates on the system as described below.

(d) The server switches from $Q^L$ to $Q^H$ after serving the L-packet at the head of $Q^L$, provided that $Q^P$ remains empty. Otherwise, the server switches to $Q^P$ and remains there as long as P-packets are present. It then switches to $Q^H$ to begin service of the H-packets.

(e) Upon leaving $Q^L$, the server closes a gate in $Q^H$. If $Q^L$ was left nonempty, the server serves all H- packets present at the time the gate was closed. If, however, a P-packet arrives following the completion of service of the L-packet or one of the H-packets, the server temporarily suspends service to $Q^H$ by switching to $Q^P$ and remaining there until no P-packets are present. The server then switches back to $Q^H$ and resumes service of H-packets. When all of the H-packets present at the time the gate was closed have been served, the server switches back to $Q^L$, providing that $Q^P$ is empty.

(f) If upon closing the gate in $Q^H$, no L-packets were present in $Q^L$, the server will serve all H-packets that arrive prior to or over the same slot as the next L-packet (consistency property). The server then switches back to $Q^L$, providing that $Q^P$ is empty. (Again, the service of H-packets will be suspended any time a P-packet is found in the system.)

The above service policy can also be described in terms of the state diagram in Fig. 2, where states represent the position of the server and transitions are made at slot boundaries according to rules (a)–(f).

The resulting queueing system is identical to that presented in [10], [11] when the arrival process to $Q^P$ is Bernoulli. A first attempt at analysis is made in [11] by providing loose bounds on the mean packet delay of customers corresponding to L-packets. The objective in the sequel is to provide an exact result for mean packet delays of all classes of customers in the three priority queueing system, where arrivals to $Q^P$ follow a more general first-order Markov process.

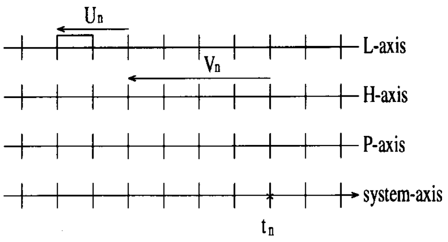


Fig. 3. Discrete-time axes used to represent system.

### B. Delay Analysis of the Queueing System

The analysis of the three-queue system is carried out by following an approach similar to that developed in [20] for the study of a two-queue system with distinct policies. Let $\{S_n : n \in \mathbf{N}\}$ denote the sequence of time instants (defined at the slot boundaries) at which the system is empty; $\mathbf{N}$ represents the set of natural numbers. It is easy to establish that $\{S_n : n \in \mathbf{N}\}$ is a renewal sequence. The random variables $X_n$, $n \geq 1$, which represent the length (in slots) of the $n^{th}$ renewal cycle, are independent and identically distributed (i.i.d.) with mean value $\bar{X}$ given by

$$\bar{X} = \frac{1}{1-\rho}\,, \tag{1}$$

which is easily proven using the Regeneration Theorem (Theorem 6–7 in [23]). $\rho = \lambda^L + \lambda^H + \lambda^P$ denotes the system utilization; under stability ($\rho < 1$), $\bar{X} < \infty$.

Let $C_n^i$ denote the cumulative delay of the $i$-packets that arrived (and were served) over the $n^{th}$ cycle, $i \in \{H, L, P\}$, $n \geq 1$. $\{C_n^i : n \in \mathbf{N}\}$ is a regenerative process with respect to the renewal process $\{S_n : n \in \mathbf{N}\}$. The expected value of the cumulative delay of the $i$-packets will be denoted by $E\{C_n^i\} = \bar{C}^i$. The subscript $n$ may be dropped since the definition of a regenerative process implies that $\{C_n^i\}_{n=0}^{\infty}$ is stochastically equivalent to $\{C_{n+\delta}^i\}_{n=0}^{\infty}$ for all $n,\delta \in \mathbf{N}$. By invoking the Regeneration Theorem once again, the mean delay of a $i$-packet can be obtained from,

$$D^i = \frac{\bar{C}^i}{\lambda^i \bar{X}}\,, \tag{2}$$

where $\lambda^i$ is the arrival rate of the $i$-packets.

The remainder of this section is focused on the calculation of $D^i$, $i \in \{H, L, P\}$, $n \geq 1$. To facilitate the description of the analysis, a multiple discrete-time axis is introduced as shown in Fig. 3. The system axis describes the evolution of the system at the server output, and the $i$-axis is a fictitious time axis used to mark the arrival of $i$-packets, where $i \in \{L, H, P\}$.

Let $\{t_n\}_{n \geq 0}$ denote the time instants at which the server is ready to switch to $Q^L$ in order to serve the L-packet at the head of that queue (if such a packet is present). Define $V_n$, $0 \leq V_n < \infty$, to be a random variable describing the length (in slots) of the *unexamined* interval on the H-axis at a given time instant $t_n$. $V_n$ represents an unexamined interval in the sense that none of the H-packets which arrived over the interval $V_n$ have been considered for service by time $t_n$. Let $U_n$, $0 \leq U_n < \infty$, be a random variable such that $U_n + V_n$ describes the distance from $t_n$ to the arrival time of the L-packet at the head of $Q^L$ (oldest packet in $Q^L$). Finally, let $\{r_n\}_{n \geq 0}$ denote a stochastic process embedded at $\{t_n\}_{n \geq 0}$ with state space $\{(i,j) : 0 \leq i,j < \infty\}$, where $i$ and $j$ are the values of $U_n$ and $V_n$ at the current time instant $t_n \in \{t_n\}_{n \geq 1}$. Due to the fact that switching instants $t_n$ can only occur when $Q^P$ is empty, together with the fact that packet arrivals to $Q^L$ and $Q^H$ are independent over consecutive slots, it is easy to show that $\{r_n\}_{n \geq 0}$ is a Markov chain embedded at $t_n \in \{t_n\}_{n \geq 1}$.

Let $f_L(\cdot)$ $[f_H(\cdot)]$ denote the Bernoulli probability mass function (PMF) for L-packet [H-packet] arrivals. Let $\{Z_k\}_{k \geq 0}$ denote the first-order Markov chain describing the P-packet arrivals. A P-packet arrives to $Q^P$ during the $k^{th}$ slot if $Z_k = 1$; no packet arrives if $Z_k = 0$. Let $G$ be a random variable that describes the length (in slots) of a burst of consecutive packet arrivals to $Q^P$ beginning at some time $k$, given that $Z_{k-1}$ was equal to 0. The random variable $G$, $0 \leq G < \infty$, has the following PMF, denoted by $f_G(\cdot)$.

$$f_G(l) = Pr(G = l) = \begin{cases} p_{00} & \text{for } l = 0 \\ p_{01} \cdot p_{11}^{l-1} \cdot p_{10} & \text{for } l \geq 1 \; , \end{cases} \tag{3}$$

where $p_{ij} = Pr(Z_k = j/Z_{k-1} = i)$, $\forall k \geq 1$.

In order to proceed with the analysis, the following quantities need to be defined.

1) $[X(i,j)$:] A random variable describing the amount of time (in slots) it takes the system to move from state $(i,j)$, at time $t_n$, to empty (the next renewal epoch from $\{S_n : n \in \mathbb{N}\}$); $i,j \geq 0$, $i + j \neq 0$. Its expected value will be denoted by $\bar{X}(i,j)$.
2) $[X(0,0)$:] A random variable describing the length of the time interval between two consecutive instants when the system is empty. Notice that $X(0,0)$ is equal to the cycle length, $X$, defined earlier but should not be interpreted as $X(i,j)$ - defined above - evaluated at $(i,j) = (0,0)$, since this is not defined. Let $\bar{X}(0,0)$ denote its expected value.
3) $[C^L(i,j)$:] A random variable describing the cumulative delay of all L-packets which arrived (and were served) over the interval $X(i,j)$, for $i,j \geq 0$. Its expected value will be denoted by $\bar{C}^L(i,j)$.
4) $[C^H(i,j)$:] Same as $C^L(i,j)$ applied to H-packets.
5) $[A_k$:] A random variable measuring the elapsed time between the earliest L-packet arrival to $Q^L$ and the current time, over an unexamined interval of $k$ slots; $0 \leq A_k \leq k$.
6) $[B_k$:] Same as $A_k$ applied to H-packets.
7) $[H_k$:] A random variable describing the number of H-packets arrived over $k$ slots. Let $h(k,j)$, $0 \leq j \leq k$, denote its PMF which is given by the $k$-fold convolution of $f_H(\cdot)$. Note that when $f_H(\cdot)$ is Bernoulli, $h(k,j)$ is given by the binomial distribution.
8) $[\tilde{G}_k$:] The sum of $k$ i.i.d. random variables $G_i$, $i = 1, 2, \ldots, k$, where $G_i$ denotes the $i^{th}$ potential burst of P-packets arriving during the interval $(t_n, t_{n+1})$, for all $t_n \in \{t_n\}_{n \geq 1}$. Let the PMF of $\tilde{G}_k$ be denoted by $\tilde{g}(k,j)$, $0 \leq j < \infty$, which is given by the $k$-fold convolution of $f_G(\cdot)$.

The procedure used to calculate the mean cumulative delay of $i$-packets over a renewal cycle, $\bar{C}^i$, where $i \in \{L, H\}$, will first be used to compute $\bar{X}(0,0) = \bar{X}$. Although the mean cycle length can be computed exactly from (1), it is more convenient to present this analytic approach for the derivation of the quantities $X(i,j)$, $i,j \geq 0$, which will then be used to compute $\bar{X}(i,j)$, $i,j \geq 0$. The same approach will lead to the computation of $\bar{C}^L$ and $\bar{C}^H$, which are needed to compute mean packet delays in (2). Note that $D^P = 1$, as explained earlier.

The length of the time interval between two consecutive instants when the system is empty is given by,

$$X(0,0) = X$$
$$= \begin{cases} 1 & \text{if } A_1 + B_1 + G_1 = 0 \\ 1 + G_1 + X(0, G_1) & \text{if } A_1 + B_1 = 0, G_1 > 0 \\ 1 + G_1 + X(1, G_1) & \text{if } A_1 = 1, B_1 = 0 \\ 2 + \tilde{G}_2 + X(A_1, 1 + \tilde{G}_2) & \text{if } B_1 = 1 \end{cases} \tag{4}$$

where for $i \geq 1$, $j \geq 0$, see (5) below, and for $i = 0$, $j \geq 1$, see (6) below.

Equation (4) can be explained by referring to Fig. 4, which represents a realization of the system given that $B_1 = 1$. The system is empty at time instant $t$, and (4) represents the amount of time (in slots) that passes until the next such time instant. If no packets arrive to the system during the time slot $(t, t')$, then the system is empty at time $t'$ and $X = 1$. If, however,

$$X(i,j) = \begin{cases} 1 & \text{if } A_{i+j} + B_{j+1} + G_1 = 0 \\ 1 + G_1 + X(0, G_1) & \text{if } A_{i+j} + B_{j+1} = 0, G_1 > 0 \\ 1 + G_1 + H_{j+1} + \tilde{G}_{H_{j+1}} + X(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \tag{5}$$

$$X(0,j) = \begin{cases} 0 & \text{if } A_j + B_j = 0 \\ X(1, A_j - 1) & \text{if } A_j > B_j, A_j > 0 \\ 1 + G_1 + X(A_1, B_j + G_1) & \text{if } B_j \geq A_j, B_j > 0 \; . \end{cases} \tag{6}$$

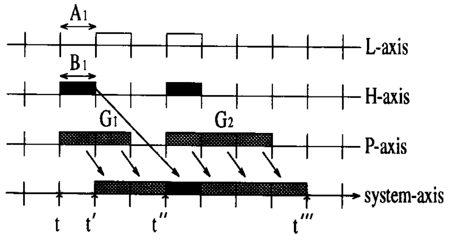


Fig. 4. Sample realization for derivation of (4) with $B_1 = 1$.

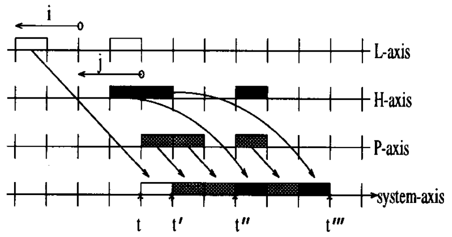


Fig. 5. Sample realization for $i \geq 1$.

only a burst of P-packets arrives during $(t, t')$, corresponding to the event $A_1 + B_1 = 0, G_1 > 0$, then the P-packet(s) receive service beginning at $t'$. The time instant at which service to this burst of P-packets is completed, $t''$, belongs to the sequence $\{t_n\}_{n\geq 1}$ of potential switching instants to $Q^L$ as defined earlier. At time $t''$ the unexamined interval on the $H$-axis is equal to $G_1$ and the state of the system is defined as $(0, G_1)$. If an L-packet arrived during $(t, t')$ but no H-packet arrived, in which case $A_1 = 1, B_1 = 0$, then the system is in state $(1, G_1)$ at time $t''$. Notice that under this scenario, $G_1$ can equal zero if no burst of P-packets arrives over $(t, t')$, in which case $t''$ coincides with $t'$. If an H-packet arrives during $(t, t')$, that is, if $B_1 = 1$ (which is the case in Fig. 4), then it will receive service beginning at time $t''$. After completing service to the H-packet, at time $t'' + 1$, the server begins serving the burst of P-packets that may have arrived over the slot $(t'', t''+1)$. Service to this burst $(G_2)$ is completed at time $t'''$ when the system is in state $(A_1, G_1+1+G_2) = (A_1, 1+\tilde{G}_2)$. Notice that if no burst arrives during $(t'', t'' + 1)$ ,that is, if $G_2 = 0$, then $t'''$ coincides with $t'' + 1$. Equations (5) and (6) represent the values of $X(i, j)$ for $i, j \geq 0$ when $i + j \neq 0$.

Equation (5) represents the amount of time required for the system to pass from state $(i, j)$ to empty when $i \geq 1$. Referring to Fig. 5, the system is in state $(i, j)$ at time $t \in \{t_n\}_{n\geq 1}$. Note that whenever $i \geq 1$, there exists an L-packet at the head of $Q^L$ awaiting service. This packet receives service which is completed at time $t' = t + 1$. If at time $t'$ no L-packets have arrived over the interval $i + j$ $(A_{i+j} = 0)$, no H-packets have arrived over the unexamined interval of length $j + 1$ $(B_{j+1} = 0)$, and a burst of P-packets has not arrived over the time slot $(t, t')$ then the system is empty and $X(i, j) = 1$. If, however, $A_{i+j} + B_{j+1} = 0$, but a burst of P-packets began arriving over the time slot $(t, t')$, then this burst will receive service over the interval $(t', t' + G_1)$. In this case, the system is in state $(0, G_1)$ at time $t'' = t' + G_1$, which represents a switching instant, and $X(i, j) = 1 + G_1 + X(0, G_1)$. If, as is the case for the realization of Fig. 5, $A_{i+j} + B_{j+1} \neq 0$, then at time $t''$ the service of H-packets that may have arrived over the interval of length $j + 1$ will commence. During the transmission slot of each of the $H_{j+1}$ H-packets, the arrival of a burst to $Q^P$ will cause the server to suspend service to $Q^H$ until the burst has been served. Clearly, the contribution of these bursts to $X(i, j)$ will be equal to $G_1 + G_2 + \cdots + G_{H_{j+1}} = \tilde{G}_{H_{j+1}}$. When service to all $H_{j+1}$ packets is completed at time $t''' \in \{t_n\}_{n\geq 1}$, the system is in state $(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1})$ and $X(i, j) = 1 + G_1 + H_{j+1} + \tilde{G}_{H_{j+1}} + X(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1})$. The derivation of (6) is very similar to (5) and hence will not be discussed here.

By applying the expectation operator to (4)–(6), the following infinite dimensional system of linear equations is obtained.

$$\bar{X}(i,j) = a(i,j) + \sum_{i'=0}^{\infty}\sum_{j'=0}^{\infty} b(i,j,i',j')\bar{X}(i',j') \; i,j \geq 0 \quad (7)$$

The constants $a(i, j)$ and the coefficients $b(i, j, i', j')$ for $i, j, i', j' \geq 0$ are derived in Appendix A.

A lower bound on $\bar{X} = \bar{X}(0, 0)$, call it $\bar{X}_{lo}(0, 0)$, can be obtained by solving a finite number of the equations in (7), in which case the following finite system of linear equations is obtained.

$$\bar{X}_{lo}(i,j) = a(i,j) + \sum_{i'=0}^{N_1}\sum_{j'=0}^{N_2} b(i,j,i',j')\bar{X}_{lo}(i',j') \qquad 0 \leq i \leq N_1, 0 \leq j \leq N_2 \quad (8)$$

It can be shown [24] that the system in (8) yields solutions, $\bar{X}_{lo}(i, j)$, which satisfy $\bar{X}_{lo}(i, j) \leq \bar{X}(i, j)$ and $\lim_{N_1, N_2 \to \infty} \bar{X}_{lo}(i, j) = \bar{X}(i, j)$ for $0 \leq i \leq N_1$ and $0 \leq j \leq N_2$. The above establishes the fact that $\bar{X}_{lo}(0, 0)$ can be made arbitrarily close to $\bar{X} = \bar{X}(0, 0)$ by solving a sufficiently large system of equations in (8), which is verified through numerical results.

Lower bounds on $\bar{C}^L$ and $\bar{C}^H$ can also be obtained by following the approach detailed above. The following equations for $C^L(i, j)$ and $C^H(i, j)$, $i, j \geq 0$, are easily derived in view of equations (4)–(6).

$$C^L(0,0) = \begin{cases} 0 & \text{if } A_1 + B_1 = 0 \\ C^L(1, G_1) & \text{if } A_1 = 1, B_1 = 0 \\ C^L(A_1, 1 + \tilde{G}_2) & \text{if } B_1 = 1 \end{cases} \quad (9)$$

$$C^L(i,j) = \begin{cases} i + j & \text{if } A_{i+j} + B_{j+1} + G_1 = 0 \\ i + j + C^L(0, G_1) & \text{if } A_{i+j} + B_{j+1} = 0, G_1 > 0 \\ i + j + C^L(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \quad (10)$$

where for $i \geq 1$, $j \geq 0$, see (10) on the previous page, and for $i = 0$, $j \geq 1$,

$$C^L(0,j) = \begin{cases} 0 & \text{if } A_j + B_j = 0 \\ C^L(1, A_j - 1) & \text{if } A_j > B_j, A_j > 0 \\ C^L(A_1, B_j + G_1) & \text{if } B_j \geq A_j, B_j > 0 \ . \end{cases} \quad (11)$$

The cumulative delay over a renewal cycle for H-packets is given by (12)–(14) below.

The derivation of (9)–(14) is similar to that of (4)–(6). Nevertheless, an explanation of (13) is included here since it contains quantities, $L_m$ and $K$, that have yet to be introduced. Recall that $C^H(i,j)$ represents the cumulative delay of all H-packets that arrive over the interval $X(i,j)$. For $i \geq 1$ and $j \geq 0$, $C^H(i,j)$ will be zero if no packets arrive ($A_{i+j} + B_{j+1} + G_1 = 0$) by time $t'$ (see Fig. 5). If only a burst of P-packets arrive by $t'$, then no contribution to $C^H(i,j)$ has yet been made and $C^H(i,j)$ will be equal to $C^H(0, G_1)$, which describes the cumulative delay of H-packets over the interval $X(0, G_1)$. If some H-packets have arrived by time $t'$, that is if $H_{j+1} > 0$, their contribution to $C^H(i,j)$, up until $t'$, will be equal to $\sum_{m=1}^{H_{j+1}} L_m$, where $L_m$ denotes the amount of time that the $m^{th}$ H-packet, which arrived over the interval of length $j+1$, has been waiting in $Q^H$ at time $t'$. Their contribution to $C^H(i,j)$ over the interval $(t', t''')$ is given in (13) by $K$, which is equal to $(1 + \tilde{G}_1) + (2 + \tilde{G}_2) + \ldots + (H_{j+1} + \tilde{G}_{H_{j+1}}) = \frac{1}{2}H_{j+1}(H_{j+1}+1) + \tilde{G}_{\frac{1}{2}H_{j+1}(H_{j+1}+1)}$. The remainder of the contribution to $C^H(i,j)$ corresponds to that associated with the time interval $X(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1})$ and is given by $C^H(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1})$.

By taking expectations for (9)–(14), two sets of equations of the same form as (7) are obtained for $\bar{C}^L(i,j)$ and $\bar{C}^H(i,j)$, $i,j \geq 0$, with coefficients, $b(i,j,i',j')$, identical to those in (7). Only the constants, denoted $a^L(i,j)$ and $a^H(i,j)$, are different from $a(i,j)$. These constants are derived in Appendix A. Tight lower bounds on $\bar{C}^L = \bar{C}^L(0,0)$ and $\bar{C}^H = \bar{C}^H(0,0)$ can be obtained by solving truncated versions of the corresponding infinite dimensional systems of linear equations. These lower bounds, denoted by $\bar{C}^L_{lo}$ and $\bar{C}^H_{lo}$, are then substituted into (2) to obtain lower bounds on the mean packet delay of L-packets and H-packets, respectively. Finally, upper bounds on the mean packet delays of H-packets and L-packets can be computed from

$$D^i_{up} = \frac{1}{\lambda^i}\left[\lambda D^{FIFO} - \sum_{j=1, j\neq i}^{N} \lambda^j D^j_{lo}\right] . \quad (15)$$

This relation is a direct application of a well-known conservation law [22] for nonpreemptive priority systems supporting N priority classes. $D^{FIFO}$ denotes the mean packet delay in the equivalent FIFO (First In-First Out) system. The equivalent FIFO system is assumed to be identical to the three-queue system presented in this paper, except that the service policy is FIFO. Since the service time of all customers is deterministic and equal to one slot, the arrival rate $\lambda$ is equal to the system utilization, usually denoted by $\rho$. $D^{FIFO}$ can be calculated by [26]

$$D^{FIFO} = 1 + \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} \lambda^i \lambda^j \left(1 + \frac{\gamma_i}{1-\gamma_i} + \frac{\gamma_j}{1-\gamma_j}\right)}{(1-\lambda)\lambda} , \quad (16)$$

where $\gamma_i$ denotes the burstiness coefficient of the $i^{th}$ arrival process and $N$ denotes the number of independent arrival streams present in the system (in this case, 3). For the two Bernoulli processes, $\gamma_L = \gamma_H = 0$. For the arrival process $\{Z_k\}_{k\geq 0}$, the burstiness coefficient is defined as $\gamma_P = p_{11} - p_{01}$, where $p_{kj}$ denotes the probability that $\{Z_k\}_{k\geq 0}$ moves from state $k$ to state $j$, $k, j, \in \{0, 1\}$.

## III. Application Of The c-G/L/HoL Priority Policy To DQDB Modeling

A useful application of the queueing system analyzed in the previous section is in modeling the queueing behavior of a station in the DQDB MAN, which will be described briefly in order to facilitate the discussion of the model. For a complete description of the DQDB medium access protocol, the reader is refered to [25].

The DQDB network consists of two high speed unidirectional buses carrying information in opposite directions (Fig. 6). The network users (stations) are distributed along the buses and are capable of transmitting information to, or receiving information from, any network station. For

$$C^H(0,0) = \begin{cases} 0 & \text{if } A_1 + B_1 = 0 \\ C^H(1, G_1) & \text{if } A_1 = 1, B_1 = 0 \\ 1 + G_1 + C^H(A_1, 1 + \tilde{G}_2) & \text{if } B_1 = 1 \end{cases} \quad (12)$$

where for $i \geq 1$, $j \geq 0$ ,

$$C^H(i,j) = \begin{cases} 0 & \text{if } A_{i+j} + B_{j+1} + G_1 = 0 \\ C^H(0, G_1) & \text{if } A_{i+j} + B_{j+1} = 0, G_1 > 0 \\ \sum_{m=1}^{H_{j+1}} L_m + K + C^H(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \quad (13)$$

where the quantities $L_m$ and $K$ are described, and for $i = 0$, $j \geq 1$,

$$C^H(0,j) = \begin{cases} 0 & \text{if } A_j + B_j = 0 \\ C^H(1, A_j - 1) & \text{if } A_j > B_j, A_j > 0 \\ B_j + C^H(A_1, B_j + G_1) & \text{if } B_j \geq A_j, B_j > 0 \end{cases} \quad (14)$$

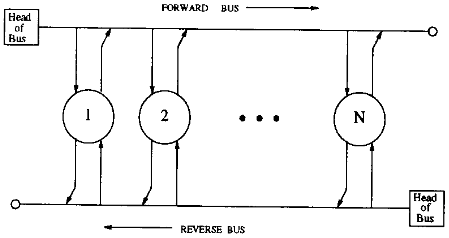


Fig. 6. The DQDB network.

the topology shown in Fig. 6, station $i, 1 \leq i \leq N$, uses the forward bus to transmit information to its downstream stations $i+1, \ldots, N$ and receive information from its upstream stations $1, 2, \ldots, i-1$ ; it uses the reverse bus to transmit information to its upstream stations and receive information from its downstream stations. Each bus is equipped with a terminal station, called the head of the bus, which continuously generates fixed size slots that propagate downstream past each station before being discarded at the end of the bus. Since the media access protocol is identical for both network buses, packet transmissions only along the forward bus will be considered.

At any given time a DQDB station may be either busy or idle; it is busy if a packet is queued in its buffer (to be transmitted on the forward bus), and it is idle otherwise. A busy station must notify all upstream stations by registering a request on the reverse bus. Each station continuously counts requests on the reverse bus by incrementing its request counter (REQ-CNTR) whenever a set busy bit is detected. An idle station also decrements its REQ-CNTR whenever an empty slot is detected on the forward bus, since this slot will be used to satisfy a downstream request.

When a station passes from the idle state to the busy state, it immediately downloads the value of the REQ-CNTR into the countdown counter (CD-CNTR) and resets the REQ-CNTR to zero (remember that the station must also send a request on the reverse bus). The busy station then decrements its CD-CNTR whenever an empty slot is detected on the forward bus; the station continues counting requests by incrementing the REQ-CNTR as before. When the value of the CD-CNTR reaches zero, the station is permitted to transmit into the next empty slot on the forward bus, since all downstream requests counted before the tagged station became busy have been satisfied. When the packet is delivered to the forward bus, the station enters the idle state. If another packet has been waiting in the buffer, the station enters the busy state instantaneously. Notice that the above procedure prevents any station from having more than one outstanding request at any given time, which could cause monopolization of the network by a few stations.

Performance evaluation of the DQDB MAN based on analytical models has proven to be a difficult task, and therefore, most of the work in this area has been based on simulation studies [27]. In [28], a throughput analysis is performed for networks operating with and without the bandwidth balancing (BWB) mechanism, under the assumption that stations are heavily loaded. In this paper, no BWB mechanism is assumed to be present. A two-queue system served under the quasi-gated policy, which was originally formulated in continuous-time [9], was adopted in [10], [11] for the development of a DQDB simulator, as well as the derivation of rather loose bounds on mean packet delays at a DQDB station. This two-queue system, however, does not allow for effective modeling of the busy slot process on the forward bus, which is a result of packet transmissions from upstream stations, since this process is incorporated into the model through a Bernoulli server availability. The busy slot process imposes an additional delay on packets arriving to the station being modeled since the station cannot transmit a packet in a busy slot or satisfy a downstream request (by allowing an empty slot to pass). The effect of the busy slot process on the queueing behavior of a station is approximately modeled in [12] by merging the request and busy slot processes to describe high priority traffic in the c-G/L policy, as well as introducing a second stage queueing system into the model. The three-queue system served according to the c-G/L/HoL priority policy allows the busy slot process to be introduced into the model for a station in a straightforward manner, as described below.

Let $Q^L$ model the queue of the tagged DQDB station under study; L-packets represent information that the tagged station wishes to transmit on the forward bus. H-packets arriving to $Q^H$ will model the arrival of downstream requests on the reverse bus, while P-packets model the arrival of busy slots on the forward bus.

If, upon transmitting a packet, the tagged station's queue is empty, rule (f) of the c-G/L/HoL priority policy (Section II) guarantees that the DQDB medium access protocol will be followed exactly. That is, all downstream requests (H-packets) that are registered prior to or over the same time slot as the next local arrival (L-packet) will be satisfied (by allowing an empty slot to pass) before the station can transmit it's own packet. Note that a station fails to satisfy a downstream request whenever a busy slot passes on the forward bus, which is consistent with the queueing model since H-packets will never be served when a P-packet is present in the system.

If, on the other hand, upon transmitting a packet, the tagged station's queue remains nonempty, all pending requests present following the transmission must be served before the station transmits it's next packet. This order of events is guaranteed by rule (e) of the service policy (Section II). Again, packet transmissions and the service of requests will be delayed whenever a busy slot passes on the forward bus (P-packet arrives to $Q^P$).

Clearly, the c-G/L/HoL priority policy captures the queueing behavior of a DQDB station. However, the accuracy of such a model is heavily dependent upon assumptions regarding the nature of traffic on the buses. In particular, the busy slot process on the forward bus exhibits significant dependencies among consecutive slots. For this reason, Bernoulli arrivals have been shown to be insufficient in modeling this process [12], [13]. A first-order Markov process is a simple model that has been used in [13] to capture some of the correlations in the busy slot process. A high degree of accuracy is achieved for a wide range of loads under correlated busy slot arrivals,

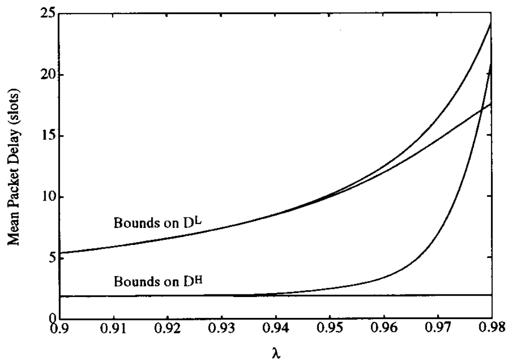


Fig. 7. Delay results for $\lambda > 0.9$.

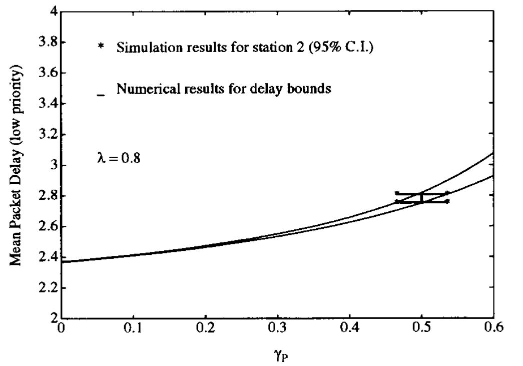


Fig. 8. Delay results for DQDB station 2 as a function of $\gamma_P$.

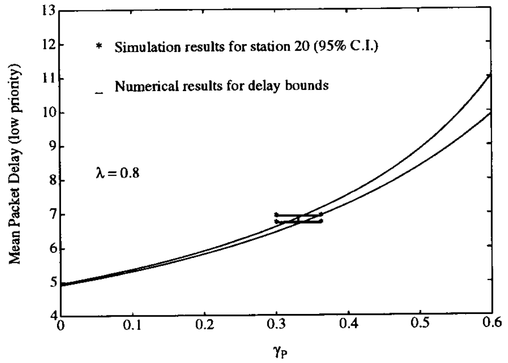


Fig. 9. Delay results for DQDB station 20 as a function of $\gamma_P$.

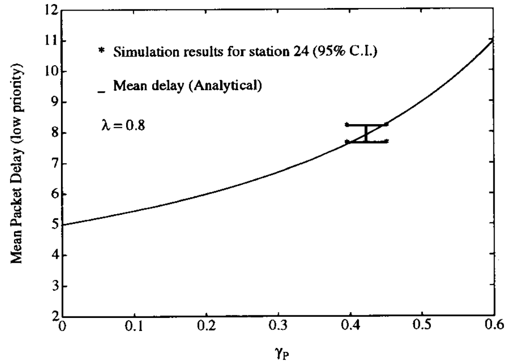


Fig. 10. Delay results for DQDB station 24 as a function of $\gamma_P$.

which is captured by the arrival process for HoL customers in the c-G/L/HoL priority policy. In the sequel, it is shown that the degree of flexibility introduced by adopting such an arrival process to $Q^P$ provides for the ability to accurately describe the delay performance of a DQDB station.

## IV. Numerical Results

In this section, numerical results for upper and lower bounds on the mean packet delay of each priority class are presented. Applicability of the queueing system to DQDB modeling is also examined by obtaining mean delay results for stations located at opposite ends of a bus in a DQDB network of 25 stations.

The performance of the queueing system, in terms of mean packet delays for H-packets and L-packets, is illustrated in Fig. 7. The results in Fig. 7 are derived for $\lambda^H = \lambda^P = 0.2$ and $\gamma_P = 0$, in which case the arrival process $\{Z_k\}_{k\geq 0}$ is Bernoulli. The mean packet delay is plotted as a function of the total load, $\lambda$, which varies with the arrival rate of L-packets, $\lambda^L$. The bounds on the delay have been obtained by solving a truncated at $N_1 = 60$, $N_2 = 10$ version of the infinite dimensional system of linear equations and setting the unknowns outside the truncation region equal to the closest unknown on the boundary $\mathbf{B}$ of the solution region. Details on this bounding approach may be found in [20], as well as other bounding techniques. The new coefficients $\tilde{b}(i,j,i',j')$, for $(i',j') \in \mathbf{B}$, are derived in Appendix B.

Figs. 8–10 present results for three stations in a DQDB network that are modeled in terms of the c-G/L/HoL priority policy. The network consists of $N = 25$ stations. The arrival rate ($\lambda^L$) to the local queue ($Q^L$) of the tagged station $i$ is proportional to the number of downstream stations. The arrival rate of requests from downstream stations is given as $\lambda^H = \sum_{j=i+1}^{N} \lambda_j$, where $\lambda_j$ denotes the local packet arrival rate to station $j$. Similarly, the arrival rate of busy slots on the forward bus is given by $\lambda^P = \sum_{j=1}^{i-1} \lambda_j$. All results are obtained for total load $\lambda = \lambda^H + \lambda^L + \lambda^P = 0.8$, which is considered to be a nominal load for DQDB operation, and $N_1 = 10, N_2 = 70$.

In Figs. 8–10, the mean packet delay, $D^L$, of the tagged station is plotted as a function of $\gamma_P$, which represents the burstiness coefficient of the busy slot process, modeled by $\{Z_k\}_{k\geq 0}$. Notice that as $\gamma_P$ increases, the mean delay bounds become loose, which should be expected since the process $\{Z_k\}_{k\geq 0}$ generates packet bursts ($G$) of greater length and states $(i,j)$, for large values of $j$, are expected to be visited by the system more frequently. Consequently, the truncation effect increases and the bounds become looser. Notice that the delay bounds shown in Figs. 8 and 9 fall within the 95% confidence intervals obtained by simulations of a uniformly spaced network with inter-station spacing of two slots. In both cases, the Bernoulli approximation ($\gamma_P = 0$) for the busy slot process in the analytical model produces results that are significantly lower than simulation results, which justifies the use of correlated arrivals to model this process. Note that the discrete event simulator used here described the DQDB MAN exactly and, therefore, the busy slot and request processes evolved "naturally", and did not require modeling.

In order to understand Fig. 10, it must be made clear that in a dual-bus network consisting of $N$ stations, only $N-1$ stations transmit information on a single bus. Specifically, the station at the end of a bus has no downstream stations to communicate with, thereby eliminating the need to send requests on the reverse bus. Clearly, this means that $\lambda^H = 0$ in the queueing model for station $N-1$, which is station 24 in this case. This, together with the fact that the delay of P-packets is deterministic and equal to one slot, implies that (15) and (16) can be used to determine the exact result for $D^L$ at station $N-1$, which explains why only one curve is present in Fig. 10.

Fig. 10 suggests that station 24 can be modeled accurately by choosing $\gamma_P$ from the interval [0.38,0.44]. Similarly, Fig. 9 implies that station 20 can be modeled sufficiently well by choosing $\gamma_P$ from the interval [0.3, 0.4]. This trend is reasonable, in an intuitive sense, since it is expected that stations at the end of a bus should see a more highly correlated busy slot process on the forward bus. For this reason, the results for station 2 (Fig. 8) seem to be inconsistent, since this station is best modeled by choosing $\gamma_P$ from the interval [0.45, 0.53]. This inconsistency, however, is a result of the Bernoulli model for the request process on the reverse bus, which should also contain some dependencies among consecutive slots. Consequently, it appears that the process $\{Z_k\}_{k\geq 0}$, for stations at the head of the forward bus, needs to be made "artificially" bursty to account for the additional packet delay induced by a correlated request process. However, by applying a burstiness coefficient of 0.4, which is in the range of $\gamma_P$ for stations at the end of the bus, only a 4.7% deviation from simulation results is observed at station 2. This suggests that for a given set of network operating conditions, a value of $\gamma_P$ that is constant with respect to the station index could be used to produce satisfactory delay results for all stations. A more analytical study of $\gamma_P$, as a function of a station's position along the bus and the network size, is the subject of future research and beyond the scope of this paper.

## V. Asymptotic Analysis Under Heavy L-Packet Traffic

The purpose of this section is to quantify the claim that $D^H$ is upper bounded for all values of $\lambda^L$, as long as $\lambda^H + \lambda^P < 1$. Derivation of this upper bound $\hat{D}^H_{up}$ proceeds by first observing that whenever $\lambda^L \geq 1 - (\lambda^H + \lambda^P)$, there will always be work present at $Q^L$ in the steady state. Therefore, upon completion of service to $Q^H$ and $Q^P$, the server visits $Q^L$ with probability one and serves exactly one packet. In other words, $U_n \geq 1$, $\forall n \geq 1$. Consequently, the state space of the Markov chain $\{r_n\}_{n\geq 0}$, embedded at $\{t_n\}_{n\geq 0}$, can be reduced to a single dimension $\{j : 0 \leq j < \infty\}$, where $j$ is the value of $V_n$ at the current time instant $t_n \in \{t_n\}$. Note that the renewal sequence $\{S_n\}_{n\geq 0}$ is now defined as the sequence of time instants at which $Q^H$ and $Q^P$ are simultaneously empty.

In view of the above discussion, it is easy to establish that the length of a renewal cycle, and the cumulative delay of H-packets are determined by modifying (5) and (13), respectively to yield, for $j \geq 0$, see (17) and (18) below.

Again, by taking expectations and solving the corresponding systems of linear equations, $\bar{X}_{lo}$ and $\bar{C}^H_{lo}$ are determined. Since the state space of the Markov chain $\{r_n\}_{n\geq 0}$ has been reduced to a single dimension, it is no longer computationally difficult to solve a sufficient number of equations to ensure that $\bar{X}_{lo}$ ($\bar{C}^H_{lo}$) converges to $\bar{X}$ ($\bar{C}^H$). In fact, for all results generated in this section, the system size $N_2$, or the number of linear equations solved, was increased until $\bar{X}_{lo}$ ($\bar{C}^H_{lo}$) was numerically equivalent to $\bar{X}$ ($\bar{C}^H$). The resulting upper bound on $D^H$ given $\lambda^P$ and $\lambda^H$ is given by,

$$\hat{D}^H_{up} = \frac{\bar{C}^H}{\lambda^H \bar{X}} \,. \tag{19}$$

The upper bound computed in (19) holds for all values of $\lambda^L$, as long as $\lambda^H + \lambda^P < 1$. An important observation regarding this upper bound on $D^H$ is that, for high system loads, such as those in Fig. 7, $\hat{D}^H_{up}$ is much tighter than the upper bound given by the conservation law in (15). Consequently, by using $D^H_{lo}$ calculated in section II, and the upper bound $\hat{D}^H_{up}$, equation (15) can be used to compute extremely tight bounds on $D^L$. This is illustrated in Fig. 11 which represents delay results for the system that is identical to the one presented in Fig. 7.

## VI. Some Comments On The Performance Of The c-G/L/HoL Policy

As previously mentioned, one of the advantages of the c-G/L/HoL priority policy is that it guarantees L-packets some degree of fairness by ensuring that an L-packet is served during every switching cycle. It is therefore expected that the mean delay suffered by L-packets will increase if $Q^H$ is granted HoL priority over $Q^L$. This can be illustrated by comparing

$$X(j) = \begin{cases} 1 & \text{if } B_{j+1} + G_1 = 0 \\ 1 + G_1 + X(G_1) & \text{if } B_{j+1} = 0, G_1 > 0 \\ 1 + G_1 + H_{j+1} + \tilde{G}_{H_{j+1}} + X(H_{j+1} + \tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \tag{17}$$

and,

$$C^H(j) = \begin{cases} C^H(G_1) & \text{if } B_{j+1} = 0 \\ \sum_{m=1}^{H_{j+1}} L_m + K + C^H(H_{j+1} + \tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \tag{18}$$

where $L_m$ and $K$ remain as given in (13).

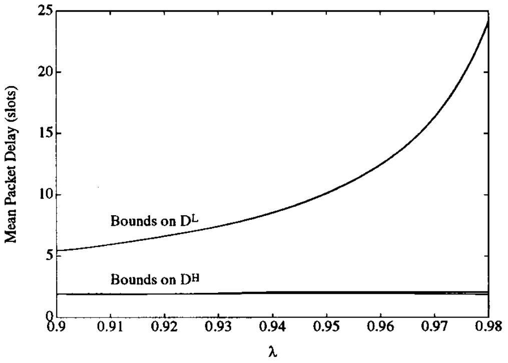


Fig. 11. Delay results using $\hat{D}^H_{up}$ for system in Fig. 7.

the c-G/L/HoL policy to a policy in which $Q^P$ continues to receive HoL priority over the other queues ($HoL^+$), $Q^L$ continues to receive limited service, but $Q^H$ is granted HoL priority ($HoL^-$) over $Q^L$. The resulting system is consistent as defined earlier and will be denoted $HoL^-/L/HoL^+$. The delay performance of the new system is easily established as follows.

Let $\tilde{D}^{HoL}$ denote the mean packet delay of an arbitrary packet arriving to $Q^H$ or $Q^P$, and let $\tilde{D}^P$ ($\tilde{D}^H$) denote the mean delay of P-packets (H-packets) in the new system. The delay of P-packets is deterministic and equal to one slot, as was the case in the c-G/L/HoL policy, and $\tilde{D}^{HoL}$ can be computed from (16) to yield,

$$\tilde{D}^{HoL} = 1 + \frac{\lambda^P \lambda^H \left(1 + \frac{\gamma_P}{1-\gamma_P}\right)}{(1 - \lambda^{HoL})\lambda^{HoL}}\,, \tag{20}$$

where $\lambda^{HoL} = \lambda^P + \lambda^H$. Finally, $\tilde{D}^H$ and $\tilde{D}^L$ can be computed by using the conservation law in (15).

Delay performances of the two consistent queueing disciplines are compared in Table I as a function of $\lambda^H$, where $\lambda^L = \lambda^P = 0.2$ and $\gamma_P = 0$. As expected, for all values of $\rho$ in Table I, the inequalities $D^L < \tilde{D}^L$ and $\tilde{D}^H < D^H$ hold true. It is also clear that $D^L > D^H$ and $\tilde{D}^L > \tilde{D}^H$, which implies that

$$0 < \frac{\tilde{D}^H}{\tilde{D}^L} < \frac{D^H}{D^L} < 1\,. \tag{21}$$

The ratio of the smallest delay to the largest delay for each queueing discipline can be seen as a measure of fairness similar to that suggested in [18]. That is, the fairness of the queueing policy can be measured by the magnitude of the appropriate ratio in (21), which implies that the c-G/L/HoL policy is more fair than the $HoL^-/L/HoL^+$ service policy.

Obviously, by gating the service to $Q^H$ and thereby relieving pressure to $Q^L$, H-packets are forced to pay a price in the form of increased delay. In many instances however, this may be a desirable trade-off, especially since the mean delay of H-packets remains relatively small even for high loads.

TABLE I
PACKET DELAYS (IN SLOTS) FOR $HoL^-/L/HoL^+$ AND c-G/L/HoL FOR VARIOUS VALUES OF OFFERED LOAD $\rho$

| $\rho$ | $\gamma^H$ | $\tilde{D}^L$ | $\tilde{D}^H$ | $D^L_{lo}$ | $D^L_{up}$ | $D^H_{lo}$ | $D^H_{up}$ |
|---|---|---|---|---|---|---|---|
| 0.6 | 0.2 | 2.17 | 1.33 | 1.95 | 1.95 | 1.55 | 1.55 |
| 0.7 | 0.3 | 3.07 | 1.40 | 2.51 | 2.51 | 1.77 | 1.77 |
| 0.8 | 0.4 | 5.00 | 1.50 | 3.66 | 3.66 | 2.17 | 2.17 |
| 0.9 | 0.5 | 11.33 | 1.67 | 7.80 | 7.84 | 3.06 | 3.08 |
| 0.94 | 0.54 | 20.26 | 1.77 | 14.21 | 14.85 | 3.77 | 4.01 |

## VII. CONCLUSIONS

In this paper, a three-queue system with mixed service policies has been presented and analyzed in discrete time. While section II provided an exact analysis for mean packet delays of each priority class, the numerical results produced tight upper and lower bounds on these quantities by solving a truncated version of an infinite dimensional system of linear equations.

The proposed queueing system was shown to accurately model the queueing behavior of an arbitrary station in the IEEE 802.6 (DQDB) MAN. One of the significant contributions to DQDB modeling is the inclusion of correlated arrivals to model the busy slot process on the forward bus. A Bernoulli model for the busy slot process, which has been used in the past, was shown to produce delay results significantly lower than those obtained by simulation. One limitation of the proposed model for a DQDB station is the fact that the request process on the reverse bus is assumed to be uncorrelated. Although this assumption does not significantly effect the delay results for stations situated near the end of the forward bus, it does account for lower than expected delays for stations near the head of the bus.

The c-G/L/HoL policy is also potentially applicable to bandwidth allocation in integrated traffic environments such as ATM, since it provides HoL service to network control traffic. An asymptotic analysis under heavily loaded low priority traffic established the worst-case packet delay for delay-sensitive traffic, represented by H-packets. The resulting upper bound on $D^H$ was shown to produce extremely tight bounds on the delay of L-packets as well as H-packets.

Finally, delay performance of the c-G/L/HoL queueing policy was examined with respect to a similar three-queue system in which arriving H-packets receive HoL priority over L-packets. It was illustrated that, by gating the service of H-packets in the c-G/L/HoL policy, L-packets are ensured a degree of fairness not provided by the $HoL^-/L/HoL^+$ policy. This fairness, however, comes at the expense of increased delay to H-packets, which may be an acceptable tradeoff in many cases, depending on quality of service requirements for this traffic class.

## APPENDIX A

By applying the expectation operator, $E\{\cdot\}$, to both sides of (4)–(6) the following equations are obtained.

$$\begin{aligned}\bar{X}(0,0) =&1 + E\{G\} + Pr(B_1 = 1)[1 + E\{G\}] + Pr(A_1 = 0)Pr(B_1 = 0)\\ &\sum_{j'=1}^{M^P} Pr(G = j')\bar{X}(0,j') + Pr(A_1 = 1)Pr(B_1 = 0)\sum_{j'=0}^{M^P} Pr(G = j')\bar{X}(1,j')\\ &+ Pr(B_1 = 1)\sum_{i'=0}^{1} Pr(A_1 = i')\sum_{k=0}^{2\cdot M^P} Pr(\tilde{G}_2 = k)\bar{X}(i', k+1)\end{aligned}$$

where for $i \geq 1,\ j \geq 0,$

$$\begin{aligned}\bar{X}(i,j) =&1 + E\{G\} + (j+1)\lambda^H E\{G\} + (j+1)\lambda^H + \sum_{j'=1}^{M^P} Pr(G = j')Pr(A_{i+j} = 0)\\ &Pr(B_{j+1} = 0)\bar{X}(0,j') + Pr(A_{i+j} = 0)\sum_{k=1}^{j+1} Pr(H_{j+1} = k)\\ &\sum_{k'=0}^{(k+1)M^P} Pr(\tilde{G}_{k+1} = k')\bar{X}(0, k+k') + \sum_{i'=1}^{i+j} Pr(A_{i+j} = i')\sum_{k=0}^{j+1} Pr(H_{j+1} = k)\\ &\sum_{k'=0}^{(k+1)M^P} Pr(\tilde{G}_{k+1} = k')\bar{X}(i', k+k')\end{aligned}$$

and for $j \geq 1,$

$$\begin{aligned}\bar{X}(0,j) =&[1 + E\{G\}]Pr(B_j \geq A_j, B_j > 0) + \sum_{k=1}^{j} Pr(A_j = k)Pr(B_j < k)\bar{X}(1, k-1)\\ &+ \sum_{k=1}^{j} Pr(B_j = k)Pr(A_j \leq k)\sum_{l=0}^{M^P} Pr(G = l)\sum_{i'=0}^{1} Pr(A_1 = i')\bar{X}(i', k+l)\end{aligned}$$

where $M^P$ is the maximum burst size for P-packet arrivals. To facilitate a numerical solution, $M^P$ must be finite and, in this work, is set equal to $max(N_1, N_2)$. In view of the above equations, the nonzero constants $a(i,j)$ and coefficients $b(i,j,i',j')$ for $i,j,i',j' \geq 0$ are given by,

$$a(0,0) = \left[1 + \lambda^H\right]\left[1 + \frac{p_{01}}{p_{10}}\right]$$

$$b(0,0,1,0) = \lambda^L\left[1 - \lambda^H\right]f_G(0)$$

$$b(0,0,0,j') = \left[1 - \lambda^H\right]\left[1 - \lambda^L\right]f_G(j') + \lambda^H\left[1 - \lambda^L\right]\tilde{g}(2, j'-1) \qquad 1 \leq j' \leq N_2$$

$$b(0,0,1,j') = \lambda^H\lambda^L\tilde{g}(2, j'-1) + \lambda^L\left[1 - \lambda^H\right]f_G(j') \qquad 1 \leq j' \leq N_2$$

For $i \geq 1,\ j \geq 0,$

$$a(i,j) = 1 + \frac{p_{01}}{p_{10}} + [j+1]\lambda^H\left[1 + \frac{p_{01}}{p_{10}}\right]$$

$$\begin{aligned}b(i,j,0,j') =&h(j+1,0)d(i+j,0)f_G(j') + d(i+j,0)\\ &\sum_{k=1}^{min\{j+1,j'\}} h(j+1,k)\tilde{g}(k+1, j'-k) \qquad 1 \leq j' \leq N_2\end{aligned}$$

$$b(i,j,i',j') = d(i+j,i')\sum_{k=0}^{min\{j+1,j'\}} h(j+1,k)\tilde{g}(k+1, j'-k) \qquad 1 \leq i' \leq min\{N_1, i+j\}, 0 \leq j' \leq N_2$$

where

$$d(k,j) = Pr(A_k = j) = \begin{cases} [1 - f_L(0)][f_L(0)]^{k-j} & \text{if } j > 0 \\ [f_L(0)]^k & \text{if } j = 0 \end{cases}$$

and for $j \geq 1$,

$$a(0,j) = \left[1 + \frac{p_{01}}{p_{10}}\right]\tilde{s}(j) \; b(0,j,1,0) = d(j,1)h(j,0)$$

$$b(0,j,0,j') = \sum_{k=1}^{min\{j,j'\}} s(j,k)f_G(j'-k)\left[1 - \lambda^L\right] \qquad 1 \leq j' \leq N_2$$

$$b(0,j,1,j') = p_{ab}(j,j'+1) + \lambda^L \sum_{k=1}^{min\{j,j'\}} s(j,k)f_G(j'-k) \qquad 1 \leq j' \leq j-1$$

$$b(0,j,1,j') = \lambda^L \sum_{k=1}^{min\{j,j'\}} s(j,k)f_G(j'-k) \qquad j \leq j' \leq N_2$$

where,

$$\begin{aligned} p_{ab}(j,k) &= Pr(A_j = k)Pr(B_j < k) \\ &= [1 - f_L(0)][f_L(0)]^{j-k}[f_H(0)]^{j-k+1} \qquad j \geq 1 \;, 1 \leq k \leq j \end{aligned}$$

$$\begin{aligned} s(j,k) &= Pr(A_j \leq k)Pr(B_j = k) \\ &= [1 - f_H(0)][f_H(0)]^{j-k}[f_L(0)]^{j-k} \qquad j \geq 1 \;, 1 \leq k \leq j \end{aligned}$$

$$\tilde{s}(j) = \sum_{k=1}^{j} s(j,k) \qquad j \geq 1$$

The constants generated by taking expected values of (9)–(11) are given by,

$$a^L(0,0) = 0$$

$$a^L(i,j) = i + j \qquad i \geq 1 \;, j \geq 0$$

$$a^L(0,j) = 0 \; j \geq 1$$

The corresponding constants generated by (12)–(14) are given by,

$$a^H(0,0) = \lambda^H \frac{p_{01}}{p_{10}} + \lambda^H$$

$$a^H(0,j) = \sum_{k=1}^{j} s(j,k)k \qquad j \geq 1$$

$$a^H(i,j) = \frac{1}{2}j[j+1]\lambda^H + \left[\frac{1}{2}[j+1]\lambda^H + \frac{1}{2}E\{H_{j+1}^2\}\right]\left[1 + \frac{p_{01}}{p_{10}}\right] \qquad i \geq 1, j \geq 0$$

## APPENDIX B

The increased coefficients $\tilde{b}(i,j,i',j')$ for $(i',j') \in \mathbf{B}$ are given by,

$$\begin{aligned}\tilde{b}(0,j,0,N_2) &= \sum_{j'=N_2}^{\infty} b(0,j,0,j') = \sum_{j'=0}^{\infty} b(0,j,0,j') - \sum_{j'=0}^{N_2-1} b(0,j,0,j') \\ &= \sum_{j'=0}^{\infty}\sum_{k=1}^{j} s(j,k) f_G(j'-k) f_L(0) - \sum_{j'=0}^{N_2-1} b(0,j,0,j') \\ &= f_L(0)\tilde{s}(j) - \sum_{j'=0}^{N_2-1} b(0,j,0,j') \qquad 1 \le j \le N_2\end{aligned}$$

$$\tilde{b}(0,j,1,N_2) = \sum_{j'=0}^{j-1} p_{ab}(j,j'+1) + f_L(1)\tilde{s}(j) - \sum_{j'=0}^{N_2-1} b(0,j,1,j') \qquad 1 \le j \le N_2$$

$$\tilde{b}(i,j,0,N_2) = d(i+j,0)[1 - h(j+1,0)f_G(0)] - \sum_{j'=0}^{N_2-1} b(i,j,0,j') \qquad 1 \le i \le N_1,\ 0 \le j \le N_2$$

$$\tilde{b}(i,j,N_1,0) = h(j+1,0)f_G(0)[1 - d(i+j,0)] - \sum_{i'=0}^{N_1-1} b(i,j,i',0) \qquad 1 \le i \le N_1,\ 0 \le j \le N_2$$

$$\tilde{b}(i,j,N_1,j') = \sum_{k=0}^{min\{j+1,j'\}} h(j+1,k)\tilde{g}(k+1,j'-k) - \sum_{i'=0}^{N_1-1} b(i,j,i',j') \qquad 1 \le i \le N_1,\ 0 \le j \le N_2,\ 1 \le j' \le N_2 - 1$$

$$\tilde{b}(i,j,i',N_2) = d(i+j,i') - \sum_{j'=0}^{N_2-1} b(i,j,i',j') \qquad 1 \le i \le N_1,\ 0 \le j \le N_2,\ 1 \le i' \le N_1 - 1$$

$$\tilde{b}(i,j,N_1,N_2) = 1 - \sum_{i'=0}^{N_1}\sum_{j'=0}^{N_2} \tilde{b}(i,j,i',j') - d(i+j,0)h(j+1,0)f_G(0) \qquad 1 \le i \le N_1,\ 0 \le j \le N_2,\ i'+j' \ne N_1+N_2.$$

## REFERENCES

[1] M. Reiser, "Performance evaluation of data networks," *Proc. IEEE*, vol. 70, Feb. 1982.

[2] H. Kobayashi, "Discrete-time queueing systems," *Probability Theory and Computer Science*, G. Louchard and G. Latouche, ed. London, United Kingdom: Academic Press, 1983, pp. 53–85.

[3] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice Hall, 1992.

[4] I. Stavrakakis, "Statistical multiplexing under non-i.i.d. packet arrival processes and different priority policies," *Performance Evaluation J.* vol. 12, pp. 181–189, 1991.

[5] I. Stavrakakis and S. Tsakiridou, "Analysis of integrated services TDM with correlated traffic," in *Proc. INFOCOM'92*, Florence, Italy, 1992

[6] K. Sriram, "Methodologies for bandwidth allocation, transmission scheduling, and congestion avoidance in broadband ATM networks," *Proc. GLOBECOM'92*, Orlando, FL, 1992. (To appear in *Computer Networks* and *ISDN Systems J.*)

[7] H. Kroner, G. Hebuterne, P. Boyer, and A. Gravey, "Priority management in ATM switching nodes," *IEEE J. Select. Areas Commun.*, vol. 9, Apr. 1991.

[8] P. Potter and M. Zukerman, "Analysis of a discrete multipriority queueing system involving a central processor serving many local queues," *IEEE J. Select. Areas Commun.*, vol. 9, Feb. 1991.

[9] C. Bisdikian, "A queueing model with application to bridges and the DQDB (IEEE 802.6) MAN," Res. Rep. RC 15218, Thomas J. Watson Research Center, IBM Corporation, Yorktown Heights, NY, Dec. 1989.

[10] C. Bisdikian, "A quasi-gated service discipline model for a DQDB data station", in *Proc. 2nd ORSA Telecommun. Workshop*, Boca Raton, FL, Mar. 1992 (Also, Res. Rep. RC 15587, Thomas J. Watson Research Center, IBM Corporation, Yorktown Heights, NY, Mar. 1990.)

[11] C. Bisdikian, "A queueing model for a data station within the IEEE 802.6 MAN," in *Proc. IEEE 17th Local Computer Networks Conf.*, Minneapolis, MN, Sept. 13–16, 1992.

[12] I. Stavrakakis and R. Landry, "Delay analysis of the DQDB MAN based on a simple model," in *Proc. ICC'92*, Chicago, IL, pp. 154–158.

[13] M. Conti, E. Gregori, and L. Lenzini, "On the approximation of the slot occupancy pattern in a DQDB network," in *Proc. INFOCOM '92*, Florence, Italy, pp. 518–526.

[14] H. Takagi, *Analysis of Polling Systems*, Computer Systems Series. Cambridge, MA: MIT Press, 1986.

[15] H. Takagi, "Queueing analysis of polling models: An update," in *Stochastic Analysis of Computer and Communication Systems*, H. Takagi, ed. Amsterdam, The Netherlands: North-Holland, 1990, pp 267–318.

[16] D. Grillo, "Polling mechanism models in communication systems—Some application examples," in *Stochastic Analysis of Computer and Communication Systems*, H. Takagi, ed. Amsterdam, The Netherlands: North-Holland, 1990, pp 659–698.

[17] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation*, vol. 1. Amsterdam, The Netherlands: North-Holland, 1991.

[18] O.J. Boxma, "Analysis and optimization of polling systems," in *Queueing Performance and Control in ATM*, ITC-13. Amsterdam, The Netherlands: Elsevier, 1991, pp 173–183.

[19] O. J. Boxma and W. D. Groenendijk, "Waiting times in discrete-time cyclic-service systems," *IEEE Trans. Commun.*, vol. COM-36, pp. 164–170, Feb. 1988.

[20] I. Stavrakakis, "A considerate priority queueing system with guaranteed policy fairness," in *Proc. INFOCOM'92*, Florence, Italy, 1992. (Also Res. Rep., CSEE/92/04–01, Computer Science and Electrical Engineering Dept., University of Vermont, Apr. 1992.)

[21] T. Ozawa, "Alternative service queues with mixed exhaustive and k-limited service," *Performance Evaluation J.*, vol. 12, pp. 165–175, 1990.

[22] R. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[23] D. Heyman and M. Sobel, *Stochastic Models in Operations Research*, vol. I. New York: McGraw-Hill, 1982.

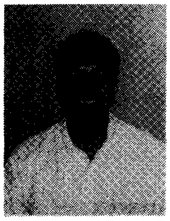
[24] L. Kantorovich and V. Krylov, *Approximate Methods of Higher Analysis*. New York: Inter-science, 1958.

[25] Distributed Queue Dual Bus (DQDB) Subnetwork of a Metropolitan Area Network (MAN), IEEE 802.6 Standard, Dec. 1990.

[26] A. Viterbi, "Approximate analysis of time synchronous packet networks," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 879–890, Sept. 1986.

[27] B. Mukherjee and C. Bisdikian, "A journey through the DQDB literature," *Performance Evaluation*, special issue on Modeling Techniques for High-Speed Telecommunication Networks, vol. 16, pp. 129–158, Dec. 1992.

[28] E. L. Hahne, A. K. Choudhury, and N. F. Maxemchuk, "DQDB networks with and without bandwidth balancing," *IEEE Trans. Commun.*, vol. 40, pp. 1192–1204, July 1992.



**Randall Landry** (ACM'91/S'91) received the B. S. degree in electrical engineering from the University of Southern Maine, Portland, ME, in 1990 and the M. S. degree in electrical engineering from the University of Vermont, Burlington, VT, in 1992. He is currently a Ph.D. candidate and a graduate research assistant in the Department of Electrical Engineering and Computer Science at the University of Vermont.

His research interests include queueing theory, performance analysis of communication systems, resource allocation in high-speed packet networks, and the modeling of broadband integrated services traffic.

**Ioannis Stavrakakis** (S'85–M'89–SM'93) received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1983, and the Ph. D. degree in electrical engineering from the University of Virginia, in 1988.

Since 1988, he has been an Assistant Professor in the Department of Electrical Engineering and Computer Science, University of Vermont. His research interests are in stochastic system modeling, teletraffic analysis and discrete-time queueing theory, with primary focus on the design and performance evaluation of Broadband Integrated Services Digital Networks (B-ISDN).

Dr. Stavrakakis is a member of the IEEE Communications Society, Technical Committee on Computer Communications. He has organized and chaired sessions, and has been a technical committee member, for conferences such as GLOBECOM, ICC and INFOCOM.