A Three-Priority Queueing Policy with Application to DQDB Modeling *

Randall Landry† Ioannis Stavrakakis‡
Department of Electrical Engineering and Computer Science
University of Vermont
Burlington, VT 05405
Tel: (802) 656-8506 FAX: (802) 656-0696

Abstract

In this paper, a discrete-time priority queueing policy is presented and analyzed. The 3-queue system provides distinct service strategies, namely the consistent-gated (c-G), 1-limited (L) and Head-of-Line (HoL), to each of the queues. The proposed service policy is potentially applicable to the modeling of network resource allocation policies in integrated services communication networks. In particular, 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.

1 Introduction

The performance 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) for the metropolitan area is often dependent upon the formulation and study of sophisticated priority queueing models [1]-[4]. The discrete-time priority queueing system presented and analyzed in this paper can be seen as an extension of the consistent gated/limited (c-G/L) priority policy studied in [6] that supports two priority classes of customers (packets). Low priority customers (L-packets) and high priority customers (H-packets), in the c-G/L service policy, arrive to two distinct queues denoted Q^L and Q^H, respectively. The server provides gated service to Q^H in the sense that it serves only H-packets that are present upon switching to Q^H. After serving these H-packets, the server switches to Q^L and provides service to a group of L-packets containing the one at the head of that queue (limited service). The service policy is *consistent* [6] in that it guarantees that no L-packet will be served before an H-packet that arrived at an earlier time (or simultaneously).

The priority policy presented in the next section allows for the introduction of a third class of customers that are guaranteed service during the time slot following their arrival. These customers (denoted P-packets) receive Head of the Line (HoL) priority in the resulting system. Their respective queue will be denoted Q^P. Clearly, this system is less considerate to both L-packets and H-packets than the c-G/L priority policy since the server will now suspend a visit to either queue in order to accomodate the arrival of a P-packet. The new policy, which is analyzed for mean packet delays in section 3, is a consistent service policy and will be referred to as the consistent gated/limited priority policy with head of line service (c-G/L/HoL).

The c-G/L/HoL priority policy is an appealing candidate for implementing network resource sharing or bandwidth allocation in an ATM environment since it can accomodate delay-insensitive (L-packets), delay-sensitive (H-packets), and network control/reservation (P-packets) traffic.

In a more specific application, the mixed service policy of the proposed three-queue system is shown 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 modeling 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 application is discussed in detail in section 4, and the model's accuracy is evaluated in section 5 by comparing the analysis to simulation results.

*Research supported by the National Science Foundation under grant NCR-9011962.

†email : rlandry@uvm.edu

‡email : stavraka@uvm.edu

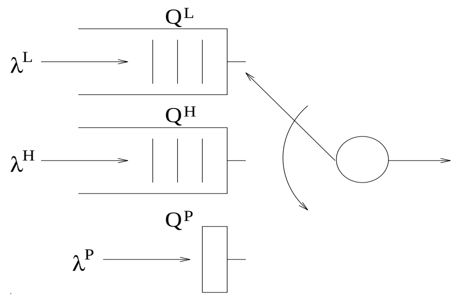

Figure 1: The discrete-time queueing system

2 The Queueing System

In this section the queueing system (Fig. 1) is described in detail. Time is assumed to be slotted and the service time of all packets deterministic and equal to the duration of one slot. 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. 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. This process generates bursts of consecutive packet arrivals whose lengths are geometrically distributed. The fact that this process delivers at most one packet per slot, together with the fact that these arrivals receive HoL priority, 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. The c-G/L/HoL priority policy is defined by the following set of rules.

(a) The system is work conserving (WC) and non-preemptive (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.

(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 non-empty, the server serves all H- packets present at the time the gate was closed. (The service of H-packets will be interrupted, however, whenever a P-packet is found in the system.) The server then switches back to Q^L, provided 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, provided that Q^P is empty. (Again, P-packets enjoy HoL priority at all times.)

3 Analysis of the c-G/L/HoL Policy

The analysis of the three-queue system is carried out by following an approach similar to that developed in [6] for the study of a two-queue system with distinct policies. Let $\{S_n : n \in \mathbf{N}\}$, where $\mathbf{N}$ represents the set of natural numbers, denote the sequence of time instants (defined at the slot boundaries) at which the system is empty. It is easy to show that $\{S_n : n \in \mathbf{N}\}$ is a renewal sequence, where S_n denotes the n^{th} renewal epoch. The random variables X_n, $n \geq 1$, which denote 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

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

which is easily proven using the Regeneration Theorem (Theorem 6-7 in [5]). ρ denotes the system utilization; under stability ($\rho < 1$), $\bar{X} < \infty$. The mean cycle length, $\bar{X}$, will be used shortly to determine lower bounds on the mean delay of both H-packets and L-packets. Note that the delay for P-packets is deterministic and equal to one slot.

Let the cumulative delay of the i^{th} priority packets that arrived (and were served) over the n^{th} cycle

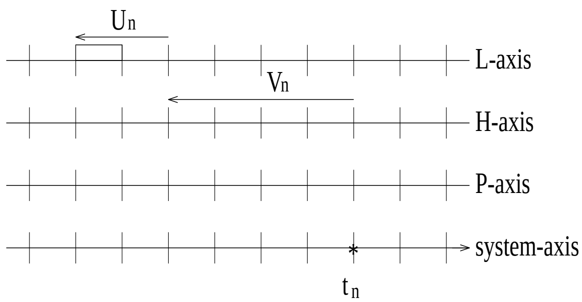

Figure 2: Discrete-time axes used to represent system

be denoted by C_n^i, for $n \geq 1$. $\left\{C_n^i : n \in \mathbf{N}\right\}$ 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^{th} priority packets will be denoted by $E\left\{C_n^i\right\} = \bar{C}^i$. By invoking the Regeneration Theorem once again, the mean delay of a i^{th} priority packet can be obtained from,

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

where λ^i is the arrival rate of the i^{th} priority packets.

The computation of $\bar{C}^i$ for the c-G/L/HoL policy, requires the solution of an infinite number of linear equations. However, a lower bound on $\bar{C}^i$, call it $\bar{C}^i_{lo}$, is obtained by solving a finite number of these equations. Hence, a lower bound on D^i, D^i_{lo}, is obtained by substituting $\bar{C}^i_{lo}$ into (2).

Finally, an upper bound on D^i, D^i_{up}, can be obtained for a NP-WC system supporting K priority classes by using [6],

$$D^i_{up} = \frac{1}{\lambda^i}\left[\lambda D^{FIFO} - \sum_{j=1, j\neq i}^{K} \lambda^j D^j_{lo}\right] \qquad . \tag{3}$$

This relation is the result of a well-known conservation law[2] that is independent of the priority policy.

In order to describe the queueing system presented in Section 2, it is useful to represent the discrete-time axis as shown in Fig. 2, where the system axis describes the evolution of the system at the server output. 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 service of H-packets found upon the n^{th} visit to Q^H is completed *and* Q^P is empty. Note that, t_n represents a time instant at which the server is allowed to switch to Q^L in order to serve one L-packet (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. Hence, the value of V_n represents the amount of time (in slots) that has elapsed since the gate was closed in Q^H for the n^{th} time. H-packets which arrived after $t_n - V_n$ have not received 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 $\{r_n\}_{n\geq 0}$ define the Markov chain 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 $t_n \in \{t_n\}_{n\geq 1}$.

As mentioned earlier, the arrival of L-packets and H-packets are described by Bernoulli processes whose arrival rates will be denoted by λ^L and λ^H, respectively. Let $f_L(\cdot)$ $[f_H(\cdot)]$ denote the Bernoulli probability mass function (PMF) for L-packet [H-packet] arrivals. The arrival of P-packets, on the other hand, is described by a first-order Markov chain, denoted by $\{Z_k\}_{k\geq 0}$, with state space $\{0,1\}$. A P-packet arrives to Q^P during the k^{th} slot if $Z_k = 1$; no packet arrives otherwise. The arrival rate of P-packets, λ^P, is equal to the steady-state probability that the chain is in state 1. 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 $f_G(\cdot)$.

$$f_G(l) = Pr(G=l) = \begin{cases} p_{00} & \text{for } l = 0 \\ p_{01} \cdot p_{11}^{l-1} \cdot p_{10} & \text{for } l \geq 1, \end{cases} \tag{4}$$

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 are needed.

- $X(i,j)$: 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). Its expected value will be denoted by $\bar{X}(i,j)$.
- $X(0,0)$: 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. Let $\bar{X}(0,0)$ denote its expected value.
- $C^L(i,j)$: 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 $\bar{C}^L(i,j)$.

$C^H(i,j)$: Same as $C^L(i,j)$ applied to H-packets.

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; $0 \leq A_k \leq k$.

B_k: Same as A_k applied to H-packets.

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 $f_H(\cdot)$.

$\tilde{G}_k$: 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 bounds for the mean cumulative delay over a renewal cycle, will first be used to compute $\bar{X}(0,0) = \bar{X}$. 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).

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

$$X(0,0) = \begin{cases} 1 & \text{if } A_1, B_1, G_1 = 0 \\ 1 + G_1 + X(0, G_1) & \text{if } A_1, B_1 = 0, G_1 > 0 \\ 1 + G_1 + X(1, G_1) & \text{if } A_1 = 1, B_1 = 0 \\ 2 + \tilde{G}_2 + X(A_1, 1 + \tilde{G}_2) & \text{if } B_1 = 1 \end{cases} \quad (5)$$

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

$$X(i,j) = \begin{cases} 1 & \text{if } A_{i+j}, B_{j+1}, G_1 = 0 \\ 1 + G_1 + X(0, G_1) & \text{if } A_{i+j}, B_{j+1} = 0, G_1 > 0 \\ 1 + G_1 + H_{j+1} + \tilde{G}_{H_{j+1}} + \\ \quad X(A_{i+j}, H_{j+1} + \tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \quad (6)$$

and for $i = 0$, $j \geq 1$,

$$X(0,j) = \begin{cases} 0 & \text{if } A_j, B_j = 0 \\ X(1, A_j - 1) & \text{if } A_j > B_j, A_j > 0 \\ 1 + G_1 + X(A_1, B_j + G_1) & \text{if } B_j \geq A_j, B_j > 0 \end{cases} \quad (7)$$

Equation (5) can be explained by referring to Fig. 3, which represents a realization of the system given that $B_1 = 1$. The system is empty at time instant t, and equation (5) 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

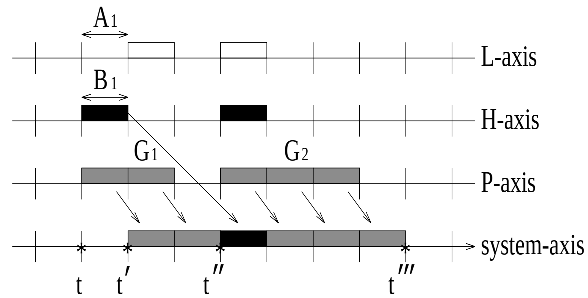

Figure 3: Realization for derivation of (5) with $B_1 = 1$.

(t, t'), then the system is empty at time t' and $X = 1$. If, however, 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. 3), 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 (6) and (7) are easily derived in light of the above discussion.

By applying the expectation operator to equations (5)-(7), the following infinite dimensional system of linear equations is obtained.

$$\bar{X}(i,j) = a(i,j) + \sum_{i'=0}^{\infty} \sum_{j'=0}^{\infty} b(i,j,i',j') \bar{X}(i',j') \quad (8)$$

The constants $a(i,j)$ and the coefficients $b(i,j,i',j')$ for $i, j, i', j' \geq 0$ are derived in [7].

A lower bound on $\bar{X} = \bar{X}(0,0)$, call it $\bar{X}_{lo}(0,0)$, may be obtained by solving a truncated version of the system of equations in (8), in which case $0 \leq i, i' \leq N_1$ and $0 \leq j, j' \leq N_2$, for $N_1, N_2 < \infty$. It is easy to show that $\bar{X}_{lo}(0,0)$ increases monotonically to $\bar{X}(0,0) = \bar{X}$ as $N_1, N_2 \to \infty$.

Lower bounds on $\bar{C}^L$ and $\bar{C}^H$ can also be obtained by following the approach detailed earlier. The following equations for $C^L(i,j)$ and $C^H(i,j)$, $i,j \geq 0$, are easily derived in view of equations (5)-(7).

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

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

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

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 the following set of equations.

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

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

$$C^H(i,j) = \begin{cases} C^H(0,G_1) & \text{if } A_{i+j}, B_{j+1} = 0 \\ \frac{1}{2} j H_{j+1} + K + & \\ \quad C^H(A_{i+j}, H_{j+1}+\tilde{G}_{H_{j+1}+1}) & \text{otherwise} \end{cases} \quad (13)$$

where $K = (1+\tilde{G}_1)+(2+\tilde{G}_2)+\ldots+(H_{j+1}+\tilde{G}_{H_{j+1}}) = \frac{1}{2}H_{j+1}(H_{j+1}) + \tilde{G}_{\frac{1}{2}H_{j+1}(H_{j+1})}$.
For $i = 0$, $j \geq 1$,

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

By taking expectations for equations (9)-(14), two sets of equations of the same form as (8) 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 (8). Only the constants, denoted $a^L(i,j)$ and $a^H(i,j)$, are different from $a(i,j)$. These constants are also derived in [7].

As was the case with the mean cycle length, tight lower bounds on $\bar{C}^L = \bar{C}^L(0,0)$ and $\bar{C}^H = \bar{C}^H(0,0)$ may be obtained by solving truncated versions of the corresponding infinite dimensional systems of linear equations. These lower bounds, call them $\bar{C}^L_{lo}$ 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 (3). D^{FIFO} can be computed by using [9],

$$D^{FIFO} = 1 + \frac{\displaystyle\sum_{i=1}^{N} \sum_{j=i+1}^{N} \lambda^i \lambda^j \left(1 + \frac{\gamma_i}{1-\gamma_i} + \frac{\gamma_j}{1-\gamma_j}\right)}{(1-\lambda)\lambda}, \quad (15)$$

where γ_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, $k, j, \in \{0,1\}$.

4 Application 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 [8].

The DQDB network consists of two high speed unidirectional buses carrying information in opposite directions. The network users (stations) are distributed along the buses and are capable of transmitting information to, or receiving information from, any network station. A station uses the forward bus to transmit information to its downstream stations and receive information from its upstream stations; it uses the reverse bus to transmit information to its upstream stations and receive information from its downstream stations. Since the medium 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.

A two-queue system served under a version of the c-G/L priority policy, which was originally formulated in continuous-time [10], was adopted in [11] for the development of a DQDB simulator. This two-queue system, however, does not allow for effective modeling of the busy slot process on the forward bus. Busy slots passing a station on the forward bus are a result of packet transmissions from upstream stations. This 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 affect 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 2) 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 non-empty, 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 2). 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, which is captured by the arrival process for HoL customers in the c-G/L/HoL priority policy.

5 Numerical Results

Figures 4-6 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 (λ^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 λ_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. The mean delay bounds have been obtained by solving a trun-

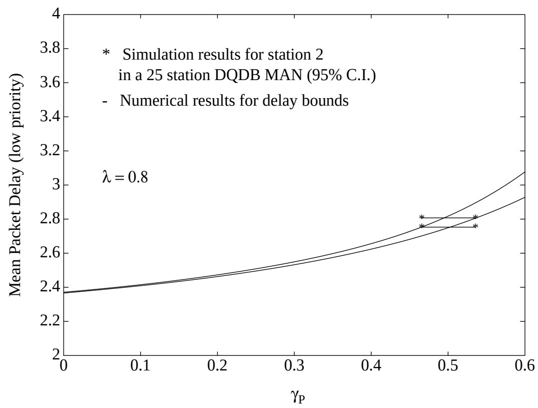

Figure 4: Delay Results for DQDB station 2 of 25

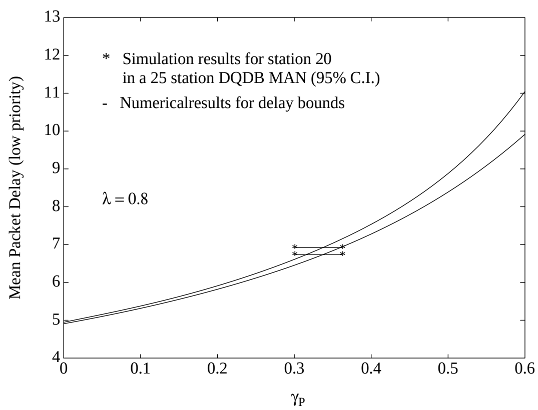

Figure 5: Delay Results for DQDB station 20 of 25

cated at $N_1 = 10$, $N_2 = 70$ 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 technique, as well as a derivation of the new coefficients $b(i,j,i',j')$, for $(i',j') \in \mathbf{B}$, may be found in [7].

In Figures 4-6, the mean packet delay, D^L, of the tagged station is plotted as a function of γ_P, which represents the burstiness coefficient of the busy slot process, modeled by $\{Z_k\}_{k\geq 0}$. Notice that as γ_P increases in Figures 4 and 5, the mean delay bounds become looser, 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 becomes significant and the bounds become loose. Also notice that the delay bounds shown in Figures 4 and 5 fall

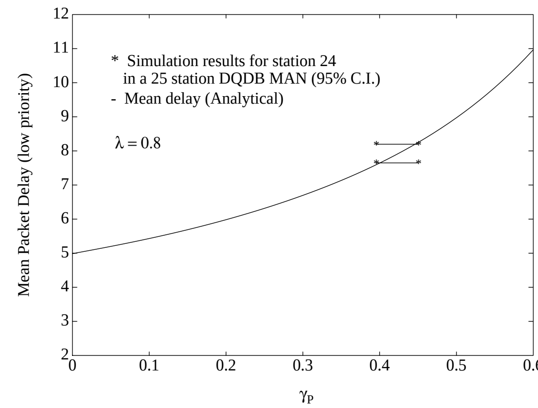

Figure 6: Delay Results for DQDB station 24 of 25

within the 95% confidence intervals obtained by simulations for a uniformly spaced network with interstation spacing of 2 slots. In both cases, the Bernoulli approximation ($\gamma_P = 0$) for the busy slot process produces results that are significantly lower than simulation results, which justifies the use of correlated arrivals to model this process.

Before considering Fig. 6, note 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 equation (3) 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. 6.

Fig. 6 suggests that station 24 can be modeled accurately by choosing γ_P from the interval [0.38,0.44]. Similarly, Fig. 5 implies that station 20 can be modeled sufficiently well by choosing γ_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. 4) seem to be inconsistent, since this station is best modeled by choosing γ_P from the interval [0.45,0.53]. This inconsistency, however, may be attributed to 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 γ_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 a possible strategy for delay analysis of a DQDB MAN, using the c-G/L/HoL policy, is to choose a γ_P that produces satisfactory results for station $N-1$, and use that value for all stations in the network. A more analytical study of γ_P, as a function of a stations position along the bus and the network size, is the subject of future research and beyond the scope of this paper.

6 Conclusions

In this paper, a three-queue system with mixed service policies has been presented and analyzed in discrete time. The proposed queueing system was shown to accurately evaluate the delay performance of a station in the DQDB MAN. The significant contribution to DQDB modeling lies in the fact that the three-priority queueing system exactly describes the service provided to stations in a DQDB network. In fact, the modeling of the busy slot process on the forward bus and the request process on the reverse bus represent the only points of approximation in the proposed model. A first-order Markov chain was adopted in an attempt to capture the correlations present in the busy slot process and improve the model's accuracy. A Bernoulli model for this process, which has been used in the past, was shown to produce delay results significantly lower than those obtained by simulation.

Finally, it should be noted that although the request process is assumed to be i.i.d., the approach described here can be used, at the expense of numerical complexity, to analyze the queueing system with correlated H-packet arrivals, which are used to model downstream requests. The preceding results, however, suggest that the delay performance of the network can be accurately evaluated through the adopted arrival processes.

References

[1] L. Kleinrock, *Queueing Systems*, Vol. I&II, John Wiley & Sons, 1976.

[2] R. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice Hall, 1989.

[3] I. Stavrakakis, "Statistical Multiplexing under Non-i.i.d. Packet Arrival Processes and Different Priority Policies", Performance Evaluation Journal 12 (1991), pp. 181-189.

[4] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation*, Vol. 1, North-Holland, 1991.

[5] D. Heyman, M. Sobel, *Stochastic Models in Operations Research*, Vol. I, McGraw-Hill, Inc., 1982.

[6] I. Stavrakakis, "Delay Bounds on a Queueing System with Consistent Priorities", Research Report CSEE/92/04-01, CSEE Dept., University of Vermont, April 1992; also submitted for publication.

[7] R. Landry, *Analytical Models for the Performance Evaluation of the DQDB MAN*, Masters Thesis, University of Vermont, Oct. 1992.

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

[9] A. Viterbi, "Approximate Analysis of Time Synchronous Packet Networks", IEEE Journal on Selected Areas in Communications, Vol. SAC-4, Sept. 1986, pp. 879-890.

[10] C. Bisdikian, "A Queueing Model with Application to Bridges and the DQDB (IEEE 802.6) MAN", Research Report RC 15218, IBM Thomas J. Watson Research Center, Yorktown Heights, Dec. 1989.

[11] C. Bisdikian, "A Quasi-Gated Service Discipline Model for a DQDB Data Station", 2^{nd} ORSA Telecommunications Workshop, March 1992, Boca Raton, FL; also, Research Report RC 15587, IBM Thomas J. Watson Research Center, Yorktown Heights, March 1990.

[12] I. Stavrakakis, R. Landry, "Delay Analysis of the DQDB MAN Based on a Simple Model", ICC'92, Chicago, IL, pp. 154-158.

[13] M. Conti, E. Gregori, L. Lenzini, "On the Approximation of the Slot Occupancy Pattern in a DQDB Network", INFOCOM '92, Florence, Italy, pp. 518-526.