Delay Bounds on a Queueing System with Consistent Priorities

Ioannis Stavrakakis, *Senior Member, IEEE*

Abstract— **A discrete-time queueing system operating under a two-level, consistent priority service policy is studied in this paper. The consistency of the policy guarantees that no low priority customer will be served before a previously (or simultaneously) arrived high priority one. Unlike the well known head of the line priority policy (which is consistent), the considered policy provides for limited service to low priority customers, even in the presence of high priority ones. The proposed policy may be viewed as a consistent version of the straightforward gated/limited service priority policy. It may also be viewed as a compromise between the head of the line priority policy and the straightforward gated/limited priority service policy. The customer service time is assumed to be deterministic and equal to one time unit, which renders the queueing model applicable to a packetized communication network environment; potential relevant applications are presented. Based on renewal arguments, the theory of infinite dimensional linear equations and a work-conservation law, a general methodology is developed for the derivation of arbitrarily tight bounds on the induced mean packet delay.**

I. Introduction

Queueing systems are naturally formulated in communication networks, due to the statistical behavior of the information traffic and the sharing of the resources for increased efficiency. Discrete-time queueing models have been widely adopted for the analysis of packet communication networks, where packet processes are described by discrete time stochastic point processes [1–6]. Priority queueing systems have also been studied extensively in the past [6-12].

A queue supporting two classes of customers with different priorities may be described in terms of two distinct queues. Let $\mathcal{H}$-$\mathcal{Q}$ and $\mathcal{L}$-$\mathcal{Q}$ denote the high and the low priority queues, respectively. A priority service policy will be considered to be consistent if it does not allow for the service of low priority customers in the presence of earlier (or simultaneously) arrived high priority customers. The well known Head of the Line (HoL) priority policy is an example of a consistent policy. According to this policy, the server moves to $\mathcal{L}$-$\mathcal{Q}$ only if $\mathcal{H}$-$\mathcal{Q}$ is empty; it switches back to $\mathcal{H}$-$\mathcal{Q}$ as soon as this queue becomes non-empty.

Consistency may be a strongly desired property of a priority service policy. In certain applications, it may be inefficient or meaningless or even impossible to provide service to low priority customers arrived after (or simultaneously with) high priority ones, which are still in the queue. Such cases may appear in production lines, job schedulers in computer systems and in high speed Asynchronous Transfer Mode/Broadband-Integrated Services Digital Networks (ATM/B-ISDNs) regarding the transmission of real time traffic requiring preservation of the information cell sequence.

A potential problem with the HoL priority policy is that it might be penalizing unacceptably the low priority customers. A well known policy which is more considerate to the low priority customers is the straightforward gated/limited service (s-G/L) priority policy. According to this policy, only the customers found in $\mathcal{H}$-$\mathcal{Q}$ at the switching instant of the server to that queue, are served (gated service). Then, the server switches to $\mathcal{L}$-$\mathcal{Q}$ and provides for some limited service to that queue. Although this policy does favor the high priority customers it is not consistent, since high priority customers may be served after simultaneously (or at a later time) arrived low priority ones, even if only one customer limited service is provided to $\mathcal{L}$-$\mathcal{Q}$ at each server visit. For instance, if $\mathcal{L}$-$\mathcal{Q}$ is left empty upon switching to $\mathcal{H}$-$\mathcal{Q}$ and both a low and a high priority customers arrive after the switching instant and before the completion of the gated service to $\mathcal{H}$-$\mathcal{Q}$, then the low priority customer will be served before the high priority one, even if the latter had arrived before the former.

The queueing system considered in this paper is a modified version of the s-G/L policy, which makes this policy consistent. The new policy will be referred to as the consistent gated/limited priority policy (c-G/L). It will be assumed that the limited service provided to $\mathcal{L}$-$\mathcal{Q}$ does not distinguish among simultaneously arrived low priority customers. In a slotted, discrete-time environment the previous implies that all low priority customer arrivals within the same time unit (slot) are served, before the server leaves $\mathcal{L}$-$\mathcal{Q}$. As it is outlined in the conclusions of this work, the latter assumption may be modified to some extent and the developed approach still be applicable. The proposed queueing system is a discrete-time version of that formulated in [13] as a queueing model for a station in the DQDB (IEEE 802.6) Metropolitan Area Network [14]. Exact analysis of the continuous-time system in [13] and the one proposed here have not been carried out in the past. Details regarding this application are presented in the next section.

Paper approved by Hideaki Takagi, the Editor for Queueing and Networking Performance of the IEEE Communications Society. Manuscript received May 30, 1991; revised January 21, 1992, April 27, 1992 and September 8, 1992. This work was supported in part by the National Science Foundation under Grant NCR-9011962. This paper was presented in part at the IEEE INFOCOM'92 Conference, May 4-8, 1992, Florence, Italy.

The author is with the Department of Computer Science and Electrical Engineering of the University of Vermont, Burlington, Vermont 05405-0156, USA.

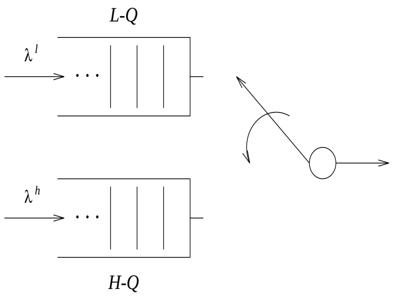

Fig. 1. The queueing system.

Finally, the proposed priority service policy can be applied to establish fairness (consistency) to a gated/limited type of service discipline and improve the performance of the high priority class. Thus, it may serve as a compromise between the HoL and the straightforward gated/limited servicy priority policies.

II. Description of the Service Policy

The proposed c-G/L policy is described in this section. Time is assumed to be slotted and customer service time deterministic and equal to one slot. Customers will be referred to as packets (of information). Let $\mathcal{H}$ ($\mathcal{L}$) denote the high (low) priority class of packets. To facilitate the description of the adopted service policy, it is assumed that new arrivals from a certain class join the corresponding queue (buffer) assigned to that class. As a result, two queues are formed (Fig. 1); let $\mathcal{H}$-$\mathcal{Q}$ and $\mathcal{L}$-$\mathcal{Q}$ denote the high and the low priority queues, respectively; infinite queue capacities are assumed.

The packet arrival processes associated with the two classes are assumed to be mutually independent discrete-time arrival processes. For each priority class, the number of packet arrivals over a slot follows a general distribution. Packet arrivals over consecutive slots are assumed to be independent. A group of $\mathcal{L}$-packets ($\mathcal{H}$-packets) is defined to be the set of all $\mathcal{L}$-packets ($\mathcal{H}$-packets) arriving over the same slot. Events (packet arrivals and service completions) are assumed to occur at the slot boundaries. The server switching time between the queues is assumed to be zero. The following rules define the proposed service policy (c-G/L) and determine the operation of the queueing system.

The Service Policy

The system is work conserving (*WC*). That is, the server is never idle in the presence of a packet in the system and the service policy does not affect the amount of service time or the arrival time of any customer. When the system is empty the server is considered to be in a neutral position. If an $\mathcal{H}$-packet ($\mathcal{L}$-packet) arrives to a previously empty system, the server visits and starts serving $\mathcal{H}$-$\mathcal{Q}$ ($\mathcal{L}$-$\mathcal{Q}$) at the beginning of the next slot. If both $\mathcal{H}$-packets and $\mathcal{L}$-packets arrive to a previously empty system, the server starts serving $\mathcal{H}$-$\mathcal{Q}$; then, it operates as the policy indicates. Packets within each queue are served according to a FIFO (First in-First Out) service policy. The server switches from $\mathcal{L}$-$\mathcal{Q}$ to $\mathcal{H}$-$\mathcal{Q}$ after serving the group of packets which contains the packet at the head of the $\mathcal{L}$-$\mathcal{Q}$ (limited one-group service). If upon switching to $\mathcal{H}$-$\mathcal{Q}$, $\mathcal{L}$-$\mathcal{Q}$ is left non-empty, then the server serves all $\mathcal{H}$-packets present in $\mathcal{H}$-$\mathcal{Q}$ at the switching instant. Then it switches back to $\mathcal{L}$-$\mathcal{Q}$. If upon switching to $\mathcal{H}$-$\mathcal{Q}$, $\mathcal{L}$-$\mathcal{Q}$ is left empty, then the server remains at $\mathcal{H}$-$\mathcal{Q}$ and serves all the $\mathcal{H}$-packets which arrived prior to or over the same slot with the next $\mathcal{L}$-packet. Then, the server switches back to $\mathcal{L}$-$\mathcal{Q}$. Note that the priority service policy described above is consistent and a compromise between the (consistent) HoL and the (inconsistent) corresponding s-G/L policies.

A continuous time version of the consistent G/L priority policy described above has been originally introduced in [13] as a queueing model for the DQDB MAN. A version of this policy has been adopted for the development of the simple DQDB simulator in [15]. A description of the DQDB MAN may be found in [14; 15; 16] or in the references in [17]. The consistent G/L priority policy captures basic functions of the queueing systems formulated at a DQDB station, as explained below. The limited one-group service policy may be easily modified to a one-packet limited service one (section V). Additional approximations may be required, though, to accommodate various dependencies present in the complex DQDB network. Nevertheless, the simulator in [15] seems to perform satisfactorily.

Let $\mathcal{L}$-$\mathcal{Q}$ model the queue of a tagged DQDB station. After the service (transmission) of a group of $\mathcal{L}$-packets is completed (say at t), the next (if any) group of $\mathcal{L}$-packets is forwarded to the head of $\mathcal{L}$-$\mathcal{Q}$. Before this group of $\mathcal{L}$-packets is served, all pending requests for service—that is, requests coming from the downstream users which have been registered and passed by the tagged station by t but not served yet—will have to be served (by allowing for empty slots to come by the tagged station), before the group of $\mathcal{L}$-packets is served. Additional requests, which may pass by the tagged station while waiting for the service of the pending (at t) requests, will not be considered before the service of the group of $\mathcal{L}$-packets waiting at the head of $\mathcal{L}$-$\mathcal{Q}$. The previous establishes the gated nature of the queueing model for the DQDB station. The interferring upstream traffic (busy slots passing by the tagged station) may be seen as additional pending requests which should be served before the next group of $\mathcal{H}$-packets. Thus, the interfering busy slots and the pending requests may be associated with a high priority traffic feeding $\mathcal{H}$-$\mathcal{Q}$. The input traffic to this queue may be modeled as consisted of two Bernoulli streams with rates equal to those of the cumulative upstream and downstream user traffic. The consistency of the queueing behavior may be established by noting that all requests passing by the tagged station by t (defined as the arrival time of a group of $\mathcal{L}$-packets to the head of $\mathcal{L}$-$\mathcal{Q}$) must be served before that group of $\mathcal{L}$-packets. Although the consistent G/L priority policy analyzed in this paper may be capable of capturing the basic functions of the queueing behavior of a DQDB station, the fine tuning of the model and the achievable accuracy are beyond the scope of this paper.

III. Analysis of the Consistent G/L Priority Policy

III.A. The Proposed Methodology for the Study of Priority Policies

The analysis of the proposed c-G/L policy will be based on arguments from renewal theory, the theory of infinite dimensional linear systems of equations and a work-conservation law. Renewal arguments and solutions of infinite dimensional equations have been considered in the past for the study of distributed queueing systems, as formulated in random-access multi-user communication networks [18; 19; 20]. This is the first time that such a methodology is applied for the analysis of priority queueing systems. The most difficult part in applying this methodology to the analysis of random-access multi-user protocols is related to the establishment of the system stability region and the conditions for the existence of a non-negative and finite solution to an infinite dimensional system of linear equations. This is trivially carried out in the case of priority queueing systems, as long as the service policy is non-preemptive and work-conserving (*NP-WC*). The supporting theory and the general methodology are presented in this sub-section.

Let the *equivalent* FIFO system be defined as the infinite capacity, FIFO queueing system whose arrival process is identical to the cumulative arrival process of the priority queueing system under consideration. Quantities associated with the equivalent FIFO system will be marked with the superscript FIFO. Greek letters denote real constants; letters i, j, and k (with or without subscripts) denote non-negative integer numbers. Random variables are denoted by a lower case letter and their expected values by the corresponding upper case letter.

Basic theory for *WC* queueing systems [8; 9; 21] asserts that the busy and idle period processes of a *WC* system and its *equivalent* FIFO system are identical for all realizations. Let x be the random variable which describes the length (in slots) of the time interval between two consecutive instants when the *WC* system is empty. Then, the following lemma is obvious.

Lemma 1: For a *WC* system $x = x^{FIFO}$ for all realizations and thus $X = X^{FIFO}$.

Let λ denote the arrival rate to the *equivalent* FIFO system; let the random variable s denote the service time of an arbitrary arrival; let $\rho = \lambda E\{s\}$ denote the utilization of the system; Then, under the condition $\rho < 1$ and $E\{s^2\} < \infty$ and assuming finite second moment for the batch size of the arrival process (stability conditions) [8], the queue is stable, the induced delay in the system is finite and ρ is equal to the fraction of time that the server is busy; $1-\rho$ is equal to the fraction of time that the server is idle or the system is empty. The latter is also given by $1/X^{FIFO}$. Since the stability of a *WC* queueing system is not affected by the order of service, the following lemma can be easily proven, in view of the above.

Lemma 2: For a stable *WC* queueing system, X is given by

$$X = \frac{1}{1-\rho} < \infty \qquad (1)$$

and the operation of the system induces renewal points with finite mean cycle length.

Lemma 2 will be used for the establishment of the existence and the actual calculation of lower bounds on the delay associated with each of the priority classes of a *NP-WC* priority system. Then, upper bounds will be derived by using the corollary to the next theorem.

Theorem 1: Consider a *NP-WC* priority queueing system supporting K priority classes. Let λ^i (D^i) denote the arrival rate (mean delay) of the ith priority customers. Under the assumption that the customer service requirements do not depend on their priority class [8; 9; 21],

$$\lambda D^{FIFO} = \sum_{i=1}^{K} \lambda^i D^i \ , \qquad (2a)$$

$$\lambda = \sum_{i=1}^{K} \lambda^i \qquad (2b)$$

Let D^i_{lo} , $1 \leq i \leq K$, denote a lower bound on the mean delay of the ith priority customers in a priority queueing system, as described in Theorem 1. The following corollary provides for an upper bound on D^i, $1 \leq i \leq K$; its proof is evident in view of Theorem 1.

Corollary 1: An upper bound, D^i_{up}, on D^i, $1 \leq i \leq K$, for the priority queueing system described in Theorem1, is given by

$$D^i_{up} = \frac{1}{\lambda^i}\left[\lambda D^{FIFO} - \sum_{k=1, k\neq i}^{K} \lambda^k D^k_{lo}\right] \qquad (3)$$

Note that the above theory is valid for a *NP-WC* queueing system supporting fixed length packets with different priorities. From now on, the customers will be considered to be fixed length packets whose service time is equal to one slot, independently of their priority class.

Let $\{z_j\}_{j\geq 1}$ denote the sequence of slot boundaries at which the queueing system is empty; $\{z_j\}_{j\geq 1}$ is a renewal sequence with mean cycle length given by (1), under the stability conditions (1). Let $\{x_j\}_{j\geq 1}$ denote the sequence of the lengths of these cycles. Let c^i_j denote the cumulative delay of the ith priority packets which arrived (were transmitted) over the jth cycle. Under stability conditions, $\{c^i_j\}_{j\geq 0}$ is a regenerative process with respect to the renewal process $\{z_j\}_{j\geq 0}$ with $E\{c^i_j\} = C^i < \infty$. Note that under the stability conditions (see above Lemma 2), the second moment of the renewal cycle is finite. Notice that the packet delay cannot exceed the length of a cycle. More precisely, the delay of a packet cannot exceed the time interval between the packet arrival instant and the end of the current cycle (excess time). Since the mean excess time is finite if the first two moments of the cycle length are finite [8], the mean packet delay—and thus the cumulative

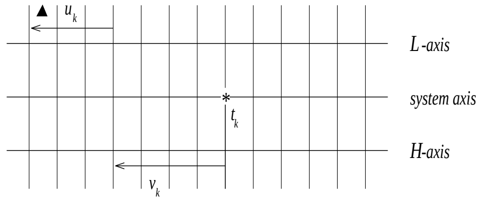

Fig. 2. Time axes for the definition of the state of the system.

packet delay as well—will be finite. The mean delay of an ith priority packet can then be obtained from [19; 20],

$$D^i = \frac{C^i}{\lambda^i X} = \frac{1-\rho}{\lambda^i} C^i \tag{4}$$

Notice that $\lambda^i X$ is equal to the average number of ith priority packet arrivals over a cycle.

To compute the expected value of the cumulative delay of the ith priority packets, the specific priority discipline has to be taken into consideration. This is carried out in the sequel for the queueing system under the c-G/L policy. The approach is potentially applicable to queueing systems under other priority policies. It turns out that the computation of C^i requires the solution of an infinite number of linear equations. A lower bound on C^i, C^i_{lo}, is obtained by solving a truncated, finite set of these equations; then, a lower bound on D^i, D^i_{lo}, is obtained by substituting C^i_{lo} in (4). Finally, Corollary 1 is invoked for the computation of an upper bound on D^i, D^i_{up}, by utilizing the lower bounds on D^i, for $0 \le i \le K$. By considering a sufficiently large number of equations in the truncated version, arbitrarily tight bounds may be obtained. The approach is illustrated in the next subsection, where lower bounds on the mean $\mathcal{L}$-packet and $\mathcal{H}$-packet delays are obtained for the queueing system under the c-G/L policy.

III.B. Mean Delay Bounds for the Consistent G/L Priority Policy

Consider the two-priority *NP-WC* queueing system described in Section II. The slot boundaries determine a discrete-time axis, referred to as the system axis, which is the time reference for all processes involved in the analysis of the system. To facilitate the description of the operation of the system—through the introduction of axes on which single category of events are marked—arrivals of $\mathcal{H}$-packets ($\mathcal{L}$-packets) are marked on a fictitious time axis referred to as $\mathcal{H}$-axis ($\mathcal{L}$-axis); these axes are otherwise identical to the system axis. For reasons which become clear later, the mean cycle length X, given by (1), is computed first.

Let $\{t_k\}_{k \ge 0}$ denote the sequence of time instants at which the service of the $\mathcal{H}$-packets found upon the kth visit to the $\mathcal{H}$-$\mathcal{Q}$ is completed. At a time instant t_k, let j_k, $0 \le j_k < \infty$, be a random variable denoting the length (in slots) of the unexamined interval on the $\mathcal{H}$-axis; that is, $\mathcal{H}$-packets which arrived over the interval j_k have not been considered for service by t_k (Fig. 2). Let $i_k + j_k$, $0 \le i_k < \infty$, be a random variable describing the distance from t_k of the group of the $\mathcal{L}$-packets which contains the packet at the head of the $\mathcal{L}$-$\mathcal{Q}$ at time t_k, that is, the oldest group of $\mathcal{L}$-packets in the $\mathcal{L}$-$\mathcal{Q}$. Let $\{r_k\}_{k \ge 0}$ be a stochastic process embedded at $\{t_k\}_{k \ge O}$ with state space $S = \{(i,j)$: $0 \le i < \infty$, $0 \le j < \infty\}$, where i and j are the values of i_k and j_k at the current instant $t_k \epsilon \{t_k\}_{k \ge 1}$. Since packet arrivals over consecutive slots are independent, it is easily established that $\{r_k\}_{k \ge 0}$ is a Markov chain embedded at $\{t_k\}_{k \ge 0}$. The following quantities are used in the analysis.

Let $y(i,j)$ be a random variable (r.v.) describing the length of the time interval (in slots) between some time instant t_k (as defined above) when the system is in state (i,j), and the first time in the future (including t_k) when the system becomes empty, $i \ge 0$, $j \ge 0$, and $i + j \ne 0$. Let $y(0,0)$ be a r.v. describing the length of the time interval between two consecutive instants when the system is empty. Notice that $y(0,0)$ is the same as x, defined earlier, and not equal to $y(i,j)$, as defined above, evaluated at $i = 0$, $j = 0$; the latter is equal to 0 since the system is empty, while $y(0,0) = x$ is always greater than zero since the interval between consecutive slots at which the system is empty can not be less than one. Let l be a r.v. describing the number of $\mathcal{L}$-packets arrived over a slot; let $g^l(k)$, $0 \le k \le M^l < \infty$, and λ^l denote its probability mass function and its expected value, respectively. Let h_1 be the same as l applied to $\mathcal{H}$-packets with corresponding parameters $g^h(k)$, $0 \le k \le M^h < \infty$, and λ^h. Let l_c be a r.v. describing the number of $\mathcal{L}$-packets arrived over a slot, given a group of $\mathcal{L}$-packets has arrived over this slot; let $g^l_c(k)$, $1 \le k \le M^l$, and μ^l denote its probability mass function and its expected value, respectively. Let h_c be the same as l_c applied to $\mathcal{H}$-packets with corresponding parameters $g^h_c(k)$, $1 \le k \le M^h$, and μ^h. Let h_k be a r.v. describing the number of $\mathcal{H}$-packets arrived over k slots. Let $h(k,j)$, $0 \le j \le kM^h$, denote its probability mass function which is given by the k-fold convolution of $g^h()$. Let a_k $[b_k]$ be a r.v. indicating the location—time instant t—of the first $\mathcal{H}$-packet [$\mathcal{L}$-packet] arrived over an unexamined interval of length k; the value of a_k $[b_k]$ is equal to the number of slots between t and the current time. Let $c^l(i,j)$ $[c^h(i,j)]$ be a r.v. describing the cumulative delay of all $\mathcal{L}$-packets [$\mathcal{H}$-packets] which are transmitted (arrived) over $y(i,j)$, $i \ge 0$, $j \ge 0$. Let $C^l(i,j)$, $C^h(i,j)$ and $Y(i,j)$ denote expected values of the quantities denoted by the corresponding lower case letter.

At this point a procedure is developed for the computation of $Y(0,0) = X$. Although the latter quantity may be computed from (1), an alternative computation approach is followed for two reasons. First, the bounds to be computed through this approach are required for the derivation of tight bounds on the mean packet delay, as explained in section IV. Second, it is conceptually easier to present this approach by applying it for the computation of the quantities $y(i,j)$, $i \ge 0$, $j \ge 0$. Based on this approach and some of the derived results, C^l and C^h will then be computed in a straightforward manner. The latter quantities are required for the mean delay calculation in (4).

It is easy to establish that $y(0,0)$ is given by (see

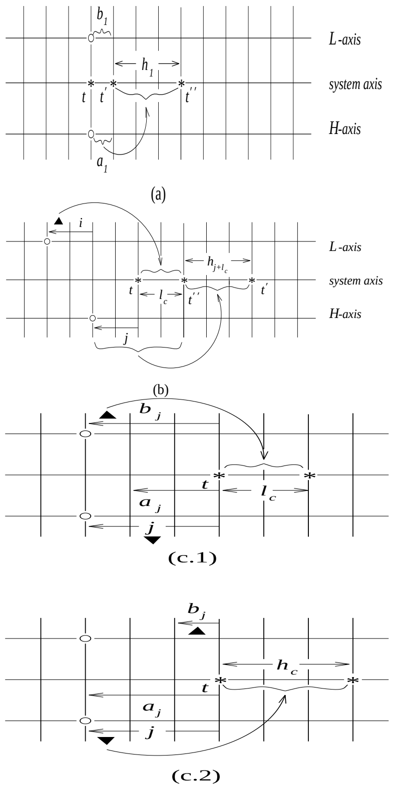

Fig. 3. Illustration of the derivation of the equations in (6); mark * indicates a potential time instant in $\{t_k\}_{k\geq 0}$; mark o indicates the boundary between the examined (on the left) and unexaamined (on the right) intervals on the corresponding axes; mark Δ indicates packet arrivals; system axis is the real time axis.

Fig. 3(a)).

$$y(0,0) = x = \begin{cases} 1 & \text{if } a_1 + b_1 = 0 \\ 1 + h_1 + y(b_1, h_1) & \text{if } a_1 + b_1 \neq 0 \end{cases} \tag{5a}$$

where (see Fig. 3(b)) for $i \geq 1$, $j \geq 0$,

$$y(i,j) = \begin{cases} l_c & \text{if } b_{i-1+j+l_c} + h_{j+l_c} = 0 \\ l_c + h_{j+l_c} + & \\ y(b_{i-1+j+l_c}, h_{j+l_c}) & \text{otherwise} \end{cases} \tag{5b}$$

and (see Fig. 3(c)) for $i = 0$, $j \geq 1$,

$$y(i,j) = \begin{cases} 0 & \text{if } a_j + b_j = 0 \\ y(1, b_j - 1) & \text{if } b_j > a_j,\, b_j > 0 \\ h_c + y(b_1, a_j - 1 + h_c) & \text{if } a_j \geq b_j,\, a_j > 0 \end{cases} \tag{5c}$$

Equation (5a) is easily explained by considering Fig. 3(a). Let t be a discrete-time instant (slot boundary) at which the buffer is empty. The next such time instant will be t' (the next slot) if no arrivals take place over the slot (t, t'), that is, if $a_1 + b_1 = 0$; in this case $x = 1$. If h_1 ($h_1 \geq 0$) $\mathcal{H}$-packets arrive over (t, t') then these packets will be transmitted over the next h_1 time slots. The completion time, t'', of these transmissions (which coincides with t' if $h_1 = 0$) corresponds to a time instant from the sequence $\{t_k\}_{k\geq 0}$, on which the Markov chain $\{r_k\}_{k\geq 1}$ has been defined. At time t'' the unexamined interval on the $\mathcal{H}$-axis is h_1. If a group of $\mathcal{L}$-packets have arrived over the examined interval of the $\mathcal{H}$-axis (t, t'), then $b_1 = 1$; $b_1 = 0$ if no such packets have arrived over that interval. Thus, the state of the system can be defined to be (b_1, h_1). By definition, $y(b_1, h_1)$ slots are required for the system to become empty after t''. Thus, if $h_1 + b_1 \neq 0$ (or equivalently $a_1 + b_1 \neq 0$), the time required for the system to reach an empty buffer state, starting from an empty buffer state at t, is given by $1 + h_1 + y(b_1, h_1)$. The values of $y(i,j)$, $i \geq 0$, $j \geq 0$, and $i + j \neq 0$ are given by the linear equations (5b) and (5c).

When $i > 0$, there is always some $\mathcal{L}$-packet to be served after instant $t \in \{t_k\}_{k\geq 0}$ (Fig. 3(b)). Let $l_c > 1$ denote the number of these packets. If no $\mathcal{H}$-packet arrived over the unexamined interval of length $j + l_c (h_{j+l_c} = 0)$ and no $\mathcal{L}$-packet arrived over the unexamined interval $i - 1 + j + l_c (b_{i-1+j+l_c} = 0)$ then the system is empty at t'' and thus, $y(i,j) = l_c$. If, on the other hand, $h_{j+l_c} \neq 0$, then the $\mathcal{H}$-packets which arrived over the interval of length $j + l_c$ are served. The time instant of the completion of this service, t', corresponds to a point in $\{t_k\}_{k\geq 1}$. Thus, in this case, $y(i,j)$ equals $l_c + h_{j+l_c} + y(b_{i-1+j+l_c}, h_{j+l_c})$. When $h_{j+l_c} = 0$ but $b_{i-1+j+l_c} \neq 0$ then $t'' \in \{t_k\}_{k\geq 1}$ and $y(i,j)$ is equal to $l_c + y(b_{i-1+j+l_c}, 0)$.

When $i = 0$ and $j \geq 1$, then no $\mathcal{L}$-packet which arrived over the examined interval on the $\mathcal{H}$-axis is in $\mathcal{L}$-$\mathcal{Q}$. Since $t \in \{t_k\}$, the server will serve the oldest group of packets independently of the class in which they belong. If no packet arrived over the interval of length j, then t corresponds to an empty buffer time instant and thus, $y(0,j) = 0$. If a group of $\mathcal{L}$-packets have arrived first over that interval, that is, if $b_j > a_j$, $b_j > 0$ (Fig. 3(c.1)), then the server moves to $\mathcal{L}$-$\mathcal{Q}$ at t and the state at that instant can be described as $(1, b_j - 1)$. Thus, $y(0,j) = y(1, b_j - 1)$. If a group consisting of h_c $\mathcal{H}$-packets have arrived first (or simultaneously with some group of $\mathcal{L}$-packets) over the interval of length j, that is, if $a_j \geq b_j$, $a_j > 0$ (Fig. 3(c.2)), then the server serves the group of the $\mathcal{H}$-packets and then is ready to move to $\mathcal{L}$-$\mathcal{Q}$. Thus, $y(0,j) = h_c + y(b_1, a_j - 1 + h_c)$, where $b_1 = 1$ if a group of $\mathcal{L}$-packets arrived simultaneously with that group of $\mathcal{H}$-packets and $b_1 = 0$ otherwise.

By applying the expectation operator to (5), the following equations are obtained.

$$Y(i,j) = e(i,j) + \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} b(i,j,i_1,i_2) Y(i_1,i_2), \tag{6}$$

where $0 \leq i \leq \infty$, $0 \leq j \leq \infty$. The constants $e(i,j) \geq 0$ and $b(i,j,i_1,i_2) \geq 0$ for $i,j,i_1,i_2 \geq 0$, may be found in [22]. The following lemma provides for the existence of a nonnegative and finite solution to the system in (6); Its proof is evident in view of the fact that the solutions of (6) are nonnegative and finite, under stability conditions [23]. Then Theorem 2 provides for a lower bound on $X = Y(0,0)$.

Lemma 3: If $\lambda^l + \lambda^h < 1$, then the system in (6) has a unique nonnegative and finite solution.

Theorem 2: For $\lambda^h + \lambda^l < 1$, a lower bound on $Y = X = Y(0,0)$ is given by $Y_{lo} = Y_{lo}(0,0)$, where $Y_{lo}(0,0)$ is the solution for $Y(0,0)$ of the finite system of linear equations

$$Y(i,j) = e(i,j) + \sum_{i_1=0}^{N_1} \sum_{i_2=0}^{N_2} b(i,j,i_1,i_2) Y(i_1,i_2) \quad (7)$$

for $0 \le i \le N_1 < \infty$, $0 \le j \le N_2 < \infty$, where $Y_{lo}(0,0)$ increases monotonically to $Y(0,0)$ as $N_1, N_2 \to \infty$.

Proof: For $\lambda^h + \lambda^l < 1$, the infinite dimensional system in (6) has a unique nonnegative finite solution (Lemma 3). Thus, the truncated version of (6) shown in (7) has solutions, $Y_{lo}(i,j)$, which satisfy

$$Y_{lo}(i,j) \le Y(i,j) \; and \lim_{N_1, N_2 \to \infty} Y_{lo}(i,j) = Y(i,j),$$

for $0 \le i \le N_1$, $0 \le j \le N_2$ [23].

Numerical results verify that for sufficiently large N_1, N_2, $Y_{lo}(0,0)$ is very close to X given by (1).

ooooooo

The previous approach can be applied directly for the calculation of lower bounds on C^h and C^l, which are unknown. These quantities are computed by formulating similar equations, as in (5), with respect to c^l and c^h. These equations, may be found in [22]. By applying the expectation operator to these equations, two sets of infinite dimensional systems of linear equations are obtained. These systems of equations are of the form of that shown in (6). In fact, the coefficients of the unknowns are identical to those in (6). It is only the constants which are different from $e(i,j)$, $i, j \ge 0$; these constants are denoted by $e^l(i,j)$ or $e^h(i,j)$ and they may be found in [22].

Tight lower bounds on $C^l = C^l(0,0)$ and $C^h = C^h(0,0)$, denoted by C^l_{lo} and C^h_{lo}, respectively, can be obtained by solving truncated versions of the corresponding infinite dimensional systems of linear equations (Theorem 2). Under the stability conditions for the queue, $C^l(i,j)$ and $C^h(i,j)$ are finite since $Y(i,j)$ are finite, for all finite i and j. Then, lower bounds on the mean packet delay can be obtained for each priority class from (see (4)):

$$D^h_{lo} = \frac{1-\rho}{\lambda^h} C^h_{lo} \quad \text{and}$$
$$D^l_{lo} = \frac{1-\rho}{\lambda^l} C^l_{lo} \quad . \qquad (8)$$

Finally, upper bounds on the mean delay for each class can be obtained from (4), provided that the mean delay for the *equivalent* FIFO queueing system is known. The latter quantity is given by,

$$D^{FIFO} = 1 + \frac{\sigma - \lambda}{2\lambda(1-\lambda)} \quad , \qquad (9)$$

where σ denotes the second moment of the cumulative number of packet arrivals per slot. Equation (9) is a known result which can be obtained, for instance, by applying the analysis in [19].

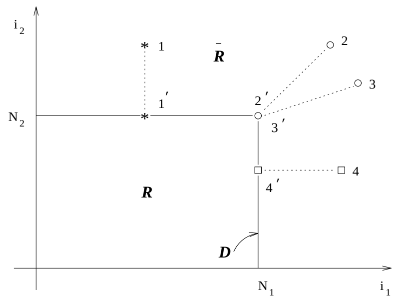

Fig. 4. The solution region $\mathbf{R}$ and its boundary $\mathbf{D}$.

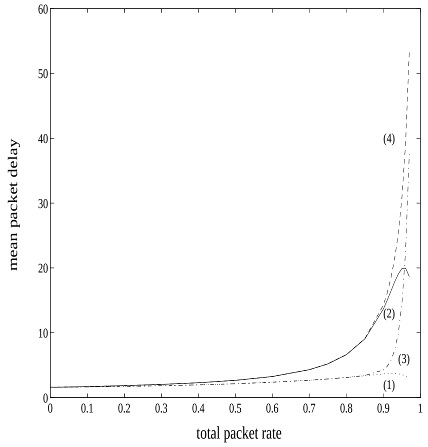

Fig. 5. Upper (1) and lower (2) bounds on the low priority packets; Upeer (3) and lower (4) bounds on the high priority packets.

IV. Numerical Results

In this section the performance of the proposed priority policy is evaluated, in terms of the induced mean packet delay for each priority class. Since the derivation of exact results requires the solution of infinite dimensional systems of linear equations, upper and lower bounds on the induced mean packet delay are computed. To improve the accuracy of the computed results, some techniques are developed for the derivation of tight bounds. The new developments will be presented with respect to quantities associated with the $\mathcal{L}$-packets, but they hold for the corresponding quantities associated with the $\mathcal{H}$-packets, as well. The following definitions are useful for the discussion of this section (see Fig. 4). Let $\mathbf{R} = \{(i_1, i_2) : 0 \le i_1 \le N_1, 0 \le i_2 \le N_2\}$ denote the solution region for the system of linear equations in (7); let $\overline{\mathbf{R}}$ denotes its complement. Let $\mathbf{D} = \{(i_1, i_2) : i_1 = N_1 \text{ and/or } i_2 = N_2\}$ denote the boundary of the solution region $\mathbf{R}$.

When the traffic load $\lambda = \lambda^h + \lambda^l$ is small or moderate (e.g., $\lambda < .6$), then the computed lower bounds C^h_{lo} and C^l_{lo} are very close to the true values of $C^h = C^h(0,0)$ and $C^l = C^l(0,0)$. This is justified in view of the observed tightness of the resulting lower and upper bounds on D^h and D^l, computed from (8) and (2). C^h_{lo} and C^l_{lo} are computed by solving (7) for some finite N_1 and N_2, where $e(i,j)$ is replaced by $e^h(i,j)$ and $e^l(i,j)$, $0 \le i \le N_1$ and $0 \le j \le$

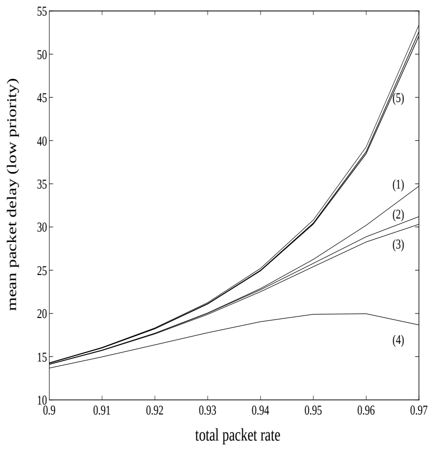

Fig. 6. Upper (5) and lower ((1),(2),(3),(4)) bounds on the low priority packets (see section IV).

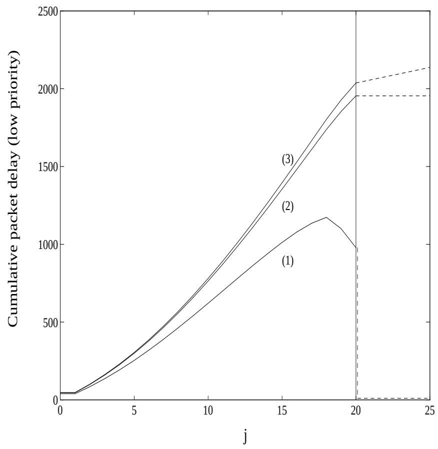

Fig. 8. Lower bounds on $C^l(0,j)$, $0 \leq j \leq 20$. The three curves are explained in section IV.

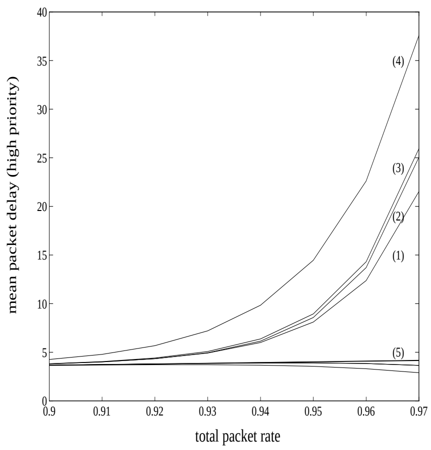

Fig. 7. Upper ((1),(2),(3),(4)) and lower (5) bounds on the high priority packets (see section IV).

N_2, respectively. The results for $N_1 = 100$, $N_2 = 15$ and symmetric packet traffic, defined by $g_c^k(1) = .5$, $g_c^k(2) = .25$, $g_c^k(3) = .25$ ($k = l$ or $k = h$), are shown in Fig. 5. Notice that the bounds become loose for total packet rate greater than .95. The latter is shown in more detail on Fig. 6 for D^l and Fig. 7 for D^h (curve (4)). The observed (erroneous) decrement of the lower bounds, as the load increases beyond .95, is due to the truncation effect on the infinite dimensional system in (7). This behavior of D^l_{lo} is explained in the next paragraph. The behavior of the bounds on D^h can be similarly explained.

From the definition of $C^l(i,j)$ it is evident that the following monotone behavior is expected: for some fixed i_0 and j_0, $C^l(i_0,j_1) \leq C^l(i_0,j_2)$ for $j_1 \leq j_2$ and $C^l(i_1,j_0) \leq C^l(i_2,j_0)$ for $i_1 \leq i_2$. This behavior is clearly not present in the results computed by (7). For instance, a typical behavior of $C^l_{lo}(i_0,j)$ for $0 \leq j \leq N_2$ is shown in Fig. 8 (curve (1)), for $i_0 = 0$, $N_1 = 60$ and $N_2 = 20$. The values of $C^l_{lo}(i_0,j)$ for j close to the truncation boundary $\mathbf{D}(j = N_2)$ decrease, as j increases. This is easily explained in view of the fact that the values of $C^l(i_0,j)$ for $(i_0,j) \in \overline{\mathbf{R}}(j > N_2)$, which affect the computation of the values of $C^l_{lo}(i_0,j)$ close to the boundary $\mathbf{D}$, are set equal to zero in (7). When the traffic load is low or moderate, and N_1 and N_2 are sufficiently large, then the states (i,j) which are close to the boundary $\mathbf{D}$ are visited very rarely. As a result, their (inaccurate) contribution to the computed value $C^l_{lo}(0,0)$ is insignificant and the resulting bounds on D^l are tight.

For large traffic loads (e.g., $\lambda > .9$) the boundary effect becomes significant and the bounds become very loose, unless N_1 and N_2 are very large. Under such traffic conditions, the values of $C^l(i,j)$ for $(i,j) \in \overline{\mathbf{R}}$ are set to be equal to the value $C^l(\hat{\imath},\hat{\jmath})$, instead of zero; $(\hat{\imath},\hat{\jmath})$ denotes the state on $\mathbf{D}$ with the minimum distance from (i,j). The latter is implemented by increasing the values of $b(i,j,i_1,i_2)$ for $(i_1,i_2) \in \mathbf{D}$, to include the weight of the coefficients $b(i,j,i_1,i_2)$ for $(i_1,i_2) \in \overline{\mathbf{R}}$. The approach is briefly described in Appendix A; the resulting new coefficients $\tilde{b}(i,j,i_1,i_2)$ may be found in [22]. The values of $\tilde{C}^l_{lo}(i,j)$ computed from (7), by incorporating the new coefficients $\tilde{b}(i,j,i_1,i_2)$, present the monotone behavior of the exact values $C^l(i,j)$, obtained from the solution of (6). A typical behavior is shown in Fig. 8 (curve (2)). The smaller the solution region $\mathbf{R}$ of (7), the larger the improvement on the tightness of the bounds on D^l achieved by utilizing the increased values of the coefficients on the boundary $\mathbf{D}$. The values $\tilde{C}^l_{lo}(i,j)$ computed from this approach are lower bounds on the true values $C^l(i,j)$, since lower bounds on $C^l(i_1,i_2)$ have been used for $(i_1,i_2) \in \overline{\mathbf{R}}$. This is formally proven in Appendix A. The resulting improvement on the bounds on the delay is shown in Fig. 6 (curve (3)).

Further improvement on $\tilde{C}^l_{lo}(i,j)$ can be achieved by boosting the values of $C^l(i,j)$ for $(i,j) \in \overline{\mathbf{R}}$. This is achieved by setting the values of $C^l(i,j)$ for $(i,j) \in \overline{\mathbf{R}}$ to be equal to the value of $C^l(\hat{\imath},\hat{\jmath})$, plus a term which increases with i and/or j; this term is a lower bound on the difference between the true value $C^l(i,j)$ and the representing value $C^l(\hat{\imath},\hat{\jmath})$. This approach is briefly described in Appendix B; the values of the resulting increased con-

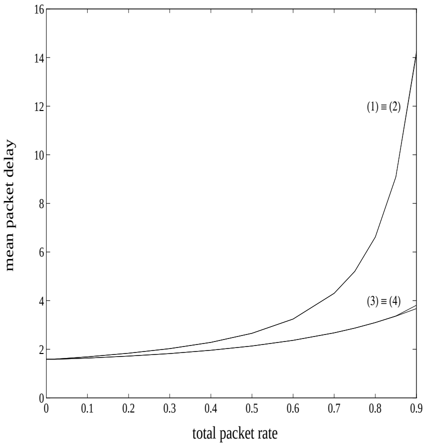

Fig. 9. Results obtained by incorporating the dominant system (section IV). Upper (1) and lower (2) bounds on the low priority packets; Upper (3) and lower (4) bounds on the high priority packets.

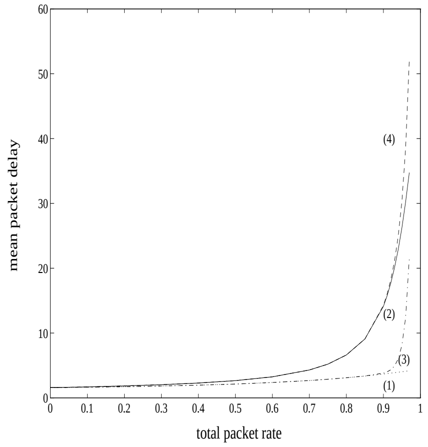

Fig. 10. Results as those in Fig. 9 over larger range of λ.

stants $\hat{e}^l(i,j)$, $0 \leq i \leq N_1$, $0 \leq j \leq N_2$, may be found in [22]. Again, the values $\hat{C}^l_{lo}(i,j)$ computed from this approach are lower bounds on $C^l(i,j)$, since lower bounds on $C^l(i_1,i_2)$ for $(i_1,i_2) \in \overline{\mathbf{R}}$ are used. Curve (3) in Fig. 8 shows the improved values of $C^l(0,j)$ for $0 \leq j \leq 20$. Curve (2) in Fig. 6 shows the resulting lower bound on D^l.

Finally, tight bound on D^l can be obtained from $D^l_{dom-} = \tilde{C}^l_{lo}/(\lambda^l \tilde{Y}_{lo})$ where $\tilde{C}^l_{lo}$ and $\tilde{Y}_{lo}$ are computed by using constant (nonzero) values for $C^l(i,j)$ and $Y(i,j)$ for $(i,j) \in \overline{\mathbf{R}}$, as explained above. In the following it is shown that D^l_{dom-} is indeed a lower bound on D^l which is better than D^l_{lo}. Let D^l_{dom} denote the mean delay of the $\mathcal{L}$-packets induced by a stochastically dominant system which rejects all packets appearing over unexamined intervals outside the region $\mathbf{R}$. Assuming otherwise identical systems, it is easily established that $\lambda^l_{dom} \leq \lambda^l$ and $D^l_{dom} \leq D^l$. Furthermore,

$$D^l_{dom} = \frac{\tilde{C}^l_{lo}}{\lambda^l_{dom} \tilde{Y}_{lo}} \geq \frac{\tilde{C}^l_{lo}}{\lambda^l \tilde{Y}_{lo}} = D^l_{dom-} \geq \frac{\tilde{C}^l_{lo}}{\lambda^l Y} = D^l_{lo}$$

Thus, $D^l_{lo} \leq D^l_{dom-} \leq D^l$. Similarly, a tight upper bound on D^h is given by $D^h_{dom-} = \frac{\tilde{C}^h_{lo}}{\lambda^h \tilde{Y}_{lo}}$, where $\tilde{C}^h_{lo}$ is the equivalent to $\tilde{C}^l_{lo}$ for $\mathcal{H}$-packets.

Curve (1) in Figs. 6 and 7 presents the obtained results for D^l_{dom-} and D^l_{dom-}, respectively. These bounds on D^l and D^h turned out to be the tightest obtained by any of the bounding approaches applied for the case considered here. These bounds on the mean delay results for each priority class are shown in Fig. 9, for $0 \leq \lambda \leq .90$, and Fig. 10 for $0 \leq \lambda \leq 1$. Notice that the upper and lower bounds coincide for $\lambda < .90$, in this case.

V. Concluding Remarks

The main contributions of this work may be summarized as follows:

1. A general methodology has been developed for the analysis of queueing systems with priorities, based on renewal arguments, the theory of infinite dimensional linear equations and a work conservation law.
2. The consistency property associated with a queueing system has been introduced and the first consistent gated/limited service priority policy (c-G/L) has been studied analytically.
3. The potential applicability of the studied priority policy to approximately model the queueing behavior of a DQDB station has been discussed.

Some comments regarding the comparison of c-G/L policy with some other relevant policies follow. Let Q^l_p (Q^h_p) denote the mean occupancy (queue length) of $\mathcal{L}$-$\mathcal{Q}$ ($\mathcal{H}$-$\mathcal{Q}$) when some policy p is in effect. The following inequalities are easily shown to hold regarding c-G/L, the Head of the Line (HoL) and the Exhaustive/Limited one-group (E/Lg) priority policies:

$$\begin{aligned} &Q^h_{\mathrm{HoL}} \leq Q^h_{\mathrm{E/Lg}} \leq Q^h_{\mathrm{c-G/L}} \quad \text{and} \\ &Q^l_{\mathrm{HoL}} \geq Q^l_{\mathrm{E/Lg}} \geq Q^l_{\mathrm{c-G/L}} \end{aligned} \tag{10}$$

The above inequalies may be shown by noting that they hold for all realizations of the buffer occupancy processes (under identical arrivals for all policies) and thus they hold for the expected values as indicated in (10). From (10) the following may be easily established.

$$\begin{aligned} &D^h_{\mathrm{HoL}} \leq D^h_{\mathrm{E/Lg}} \leq D^h_{\mathrm{c-G/L}} \quad \text{and} \\ &D^l_{\mathrm{HoL}} \geq D^l_{\mathrm{E/Lg}} \geq D^l_{\mathrm{c-G/L}} \end{aligned} \tag{11}$$

Notice that all the policies considered above are consistent. The exhaustive/limited policy is consistent only if limited one-group service is provided to $\mathcal{L}$-$\mathcal{Q}$; otherwise, the same inconsistency problem arises as for the straightforward gated/limited priority policy (Introduction).

Regarding the performance of the developed c-G/L policy relatively to the comparable non-consistent s-G/L one-group (s-G/Lg) policy, the following may be established as before.

$$Q^h_{\mathrm{c-G/L}} \leq Q^h_{\mathrm{s-G/Lg}}, \quad Q^l_{\mathrm{c-G/L}} \geq Q^l_{\mathrm{s-G/Lg}} \quad \text{and thus}$$

$$D^h_{\text{c-G/L}} \le D^h_{\text{s-G/Lg}}, \quad D^l_{\text{c-G/L}} \le D^l_{\text{s-G/Lg}} \tag{12}$$

Notice that priority policies c-G/L and s-G/Lg become identical if the completion of the service of a group of $\mathcal{L}$-packets leaves $\mathcal{L}$-$\mathcal{Q}$ always non-empty. Thus, $D^k_{\text{c-G/L}} \to D^k_{\text{s-G/Lg}}$, $k \in \{l, h\}$, as the total traffic load, λ, increases to one, since $\mathcal{L}$-$\mathcal{Q}$ will always be left non-empty in the limit.

Finally, it should be noted that the developed analysis technique is directly applicable to a queueing system under a service policy which provides limited one-packet service to $\mathcal{L}$-$\mathcal{Q}$ and which is otherwise identical to c-G/L policy. An additional assumption in this case will be that the number of $\mathcal{L}$-packet arrivals per slot is geometrically distributed. In this case, similar equations may be written. For instance, equation (5a) will still be valid, while equations (5b) and (5c) will need to be slightly modified to account for the fact that the arrival slot of the $\mathcal{L}$-packets which has just been served will have to be reexamined for possible additional packet arrivals, since one packet and not one group is served at each visit to $\mathcal{L}$-$\mathcal{Q}$. Due to the geometric distribution of the number of $\mathcal{L}$-packets arrivals per slot, no information will be required regarding the number of $\mathcal{L}$-packets from the same arrival slot (group) already served.

Finally it should be noted that the developed methodology for the study of priority policies (section III.A) is valid when the service time distributions of the customers are identical without being necessarily constant and equal to one. The latter is true since Theorem 1 is valid as long as the customer service requirements do not depend on their priority class. On the other hand, the mean delay bound calculation procedure (section III.B) is based on the fact that the service time is constant and equal to one. This service requirement determines systems which may serve as models of queueing systems appearing in slotted packetized communication networks. When the service requirement is other than one time unit, some other approach will need to be followed for the mean bound calculation. The analysis presented in section III.A will still be applicable provided that the service time distributions of all types of customers be identical.

A. Constant Boundary

The increased coefficients $\tilde{b}(i, j, i_1, i_2)$:

Let $(\hat{\imath}_1, \hat{\imath}_2) \in \mathbf{D}$ denote the point on $\mathbf{D}$ which has the minimum distance from (i_1, i_2) (Fig. 4). Since $\hat{\imath}_1 \le i_1$ and $\hat{\imath}_2 \le i_2$, the monotonicity of $C^l(i, j)$ implies that $\epsilon(i_1, i_2) = C^l(i, j) - C^l(\hat{\imath}, \hat{\jmath})$, for $(i, j) \in \overline{\mathbf{R}}$, is non-negative. The system in (7) associated with $C^l(i, j)$ can be written in the following way.

$$\begin{aligned} C^l(i,j) = e^l(i,j) &+ \sum_{i_1=0}^{N_1}\sum_{i_2=0}^{N_2} b(i,j,i_1,i_2) C^l(i_1,i_2) \\ &+ \sum\sum_{(i_1,i_2)\in\overline{\mathbf{R}}} b(i,j,i_1,i_2) \\ &\cdot \left[C^l(\hat{\imath}_1, \hat{\imath}_2 + \epsilon(i_1,i_2)\right] \\ = e_l(i,j) &+ \sum\sum_{(i_1,i_2)\in\overline{\mathbf{R}}} b(i,j,i_1,i_2)\epsilon(i_1,i_2) \\ &+ \sum_{i_1=0}^{N_1}\sum_{i_2=0}^{N_2} \tilde{b}(i,j,i_1,i_2) C^l(i_1,i_2) \\ = \tilde{e}_l &+ \sum_{i_1=0}^{N_1}\sum_{i_2=0}^{N_2} \tilde{b}(i,j,i_1,i_2) C^l(i_1,i_2) \end{aligned} \tag{13}$$

where

$$\tilde{e}^l(i,j) = e^l(i,j) + \sum\sum_{(i_1,i_2)\in\overline{\mathbf{R}}} b(i,j,i_1,i_2)\epsilon(i_1,i_2).$$

The coefficients $\tilde{b}(i, j, i_1, i_2)$ within the solution region $\mathbf{R}$ (excluding the boundary $\mathbf{D}$) are given by

$$\tilde{b}(i,j,i_1,i_2) = b(i,j,i_1,i_2)\,,$$

where $0 \le i \le N_1$, $0 \le j \le N_2$, and $(i_1, i_2) \in \mathbf{R} - \mathbf{D}$. The coefficients $\tilde{b}(i, j, i_1, i_2)$ for $(i_1, i_2) \in \mathbf{D}$ may be found in [22]. Note that the above system of equations is equivalent to system (6) with respect to the unknowns $C^l(i, j)$ for $(i, j) \in \mathbf{R}$. Thus, system (13 has a non-negative solution if system (6) has a nonnegative solution; the latter is the case under stability conditions. Since $\tilde{e}^l(i, j) \ge e^l(i, j)$, system (13) is a majorant for the one with constants equal to $\tilde{e}^l(i, j) = e^l(i, j)$ (system (14)). Thus, under stability conditions, system (14) has a non-negative solution $\tilde{C}^l_{lo}$, which is upper bounded by the solution C^l that of system (13) [23].

B. Boosting $C^l(i, j)$ for $(i, j) \in \overline{\mathbf{R}}$

The increased constants $\hat{e}(i, j)$:

Further improvement on $\tilde{C}^l_{lo}(i, j)$ can be achieved by boosting the values of $C^l(i, j)$ for $(i, j) \in \overline{\mathbf{R}}$. This is achieved by setting the values of $C^l(i, j)$ for $(i, j) \in \overline{\mathbf{R}}$ to be equal to the value of $C^l(\hat{\imath}, \hat{\jmath})$, plus $\delta^l(\hat{\imath}_1, \hat{\imath}_2)$; $(\hat{\imath}, \hat{\jmath})$ denotes the state on $\mathbf{D}$ with the minimum distance from (i, j); the additional term $\delta^l(\hat{\imath}_1, \hat{\imath}_2)$ is a lower bound on the difference between the true value $C^l(i, j)$ and the representing value $C^l(\hat{\imath}, \hat{\jmath})$. That is,

$$C^l(i_1, i_2) \ge C^l(\hat{\imath}_1, \hat{\imath}_2) + \delta^l(\hat{\imath}_1, \hat{\imath}_2) \tag{15}$$

where the nonzero values of $\delta^l(\hat{\imath}_1, \hat{\imath}_2)$ may be found in [22]. The exact value of $C^l(i_1, i_2)$ is greater than the right hand side of (15), since the additional cumulative delay $\delta^l(\hat{\imath}_1, \hat{\imath}_2)$, in excess of $C^l(i_1, i_2)$, is a lower bound on the difference between $C^l(i_1, i_2)$ and $C^l(\hat{\imath}_1, \hat{\imath}_2)$. This bound represents, in general, the delay contribution of the $\mathcal{L}$-packet arrivals over the initial unexamined intervals in $Y(i_1, i_2)$ which are not present in $Y(\hat{\imath}_1, \hat{\imath}_2)$. The cumulative delay of the additional packets due to the longer length of $Y(i_1, i_2)$ compared to that of $Y(\hat{\imath}_1, \hat{\imath}_2)$ is not represented by $\delta^l(\hat{\imath}_1, \hat{\imath}_2)$.

By substituting the above lower bounds on $C^l(i_1, i_2)$ for $(i_1, i_2) \in \overline{\mathbf{R}}$ in the infinite dimensional system in (6)—with $e(i, j)$ replaced by $e^l(i, j)$—an $(N_1+1)(N_2+1)$-dimensional system of linear equations is obtained. Its coefficients

$\tilde{b}(i,j,i_1,i_2)$ are identical to those derived in Appendix (A); its constants are increased due to the additional terms $\delta^l(\hat{\imath}_1,\hat{\imath}_2)$. The new constants $\hat{e}^l(i,j)$ are given by

$$\hat{e}^l(i,j) = e^l(i,j) + \sum_{i_1=0}^{i-1+j+M^l} \sum_{i_2=0}^{(j+M^l)M^h} b(i,j,i_1,i_2)\delta^l(\hat{\imath}_1,\hat{\imath}_2), \tag{16}$$

where $0 \le i \le N_1$, $0 \le j \le N_2$.

Finally, it is easily justified that the solutions of the system in (7) with constants as derived above and coefficients $\tilde{b}(i,j,i_1,i_2)$, as computed in Appendix A, are upper bounded by the exact solutions computed from (6). The latter can be shown by writing

$$C^l(i,j) = C^l(\hat{\imath},\hat{\jmath}) + \delta^l(\hat{\imath},\hat{\jmath}) + \epsilon(i,j)$$

for $(i,j) \in \overline{\mathbf{R}}$ and applying the approach presented in Appendix A. A similar approach may be followed for the derivation of the increased constants $\hat{e}^h$ [22].

C. Comments on equations (9) and (10) and the associated constants

XXX Does this belong here? It was found after the bibliography.

The derivation of the equations in systems (9) and (10) is similar to that of the equations in system (6). Some of these equations are explained in this section.

The system in (9) can be explained in the following way (see Fig. 3(b)). The cumulative delay of the $\mathcal{L}$-packets which arrived (transmitted) over the time interval $y(i,j)$, for some $i \ge 1$ and $j \ge 0$, is equal to the contributions of the l_c $\mathcal{L}$-packets which arrived $j+i$ slots before the current time t, if no other packet has arrived by time t''. This quantity is equal to the delay experienced by t (which is equal to $(i+j-1)l_c$), plus the cumulative delay experienced between the beginning and the completion of the transmission of that group $(1+2+\ldots+l_c = (1/2)l_c(l_c+1))$. If some other packets arrived by t'', their cumulative delay will be that associated with an interval $y(b_{i-1+j+l_c}, h_{j+l_c})$, which is denoted by $c^l(b_{i-1+j+l_c}, h_{j+l_c})$.

The equations in system (10) can be explained in the following way. The cumulative delay of $\mathcal{H}$-packets which arrived over a period $y(i,j)$, for some $i \ge 1$ and $j \ge 0$, is equal to zero if no other packet has arrived by t''. If only some $\mathcal{L}$-packets have arrived by t'', then no contribution to $c^h(i,j)$ has taken place by t''; thus, $c^h(i,j)$ will be equal to $c^h(b_{i-1+j+l_c},0)$, which describes the corresponding quantity over the interval $y(b_{i-1+j+l_c},0)$ which follows t''. If some $\mathcal{H}$-packets have arrived by t'', that is, if $h_{j+l_c} > 0$, their contribution to $c^h(i,j)$ will be equal to $(1/2)h_{j+l_c}(j+l_c-1)$ by time t'' and $1+2+\ldots+h_{j+l_c} = (1/2)h_{j+l_c}(h_{j+l_c}+1)$ by time t'. The rest of the contribution to $c^h(i,j)$ corresponds to that associated with a period $y(b_{i-1+j+l_c}, h_{j+l_c})$ that follows t' and it is denoted by $c^h(b_{i-1+j+l_c}, h_{j+l_c})$.

By considering the expectations of the systems (9) and (10), a system of linear equations is formed. The coefficients of the unknown are identical to those in (7). The non-zero constants $e^l(i,j)$ and $e^h(i,j)$, associated with the systems (9) and (10), are given by

$$e^l(i,j) = (i+j-\frac{1}{2})\mu^l + \frac{1}{2}E\{l_c^2\}, \quad i \ge 1,\ j \ge 0,$$

$$e^h(i,j) = \frac{1}{2}\lambda^h\mu^h + \frac{1}{2}E\{h_1^2\}, \quad i = 0,\ j = 0$$

$$e^h(i,j) = \sum_{k=1}^{M^h} g_c^l(k) \cdot \left\{ \frac{k+j}{2}E\{h_{j+k}\} + \frac{1}{2}E\{h_{j+k}^2\} \right\}, \quad i \ge 1,\ j \ge 0$$

$$e^h(i,j) = s_{12}(j)\left[\frac{1}{2}E\{h_c^2\} - \frac{1}{2}\mu^h\right] + mu^h \sum_{i_1=1}^{j} s(i_1,j)i_1, \quad i = 0,\ j \ge 1$$

References

[1] M. Reiser, "Performance evaluation of data networks," *Proc. IIII*, vol. 70, no. 2, 1982.

[2] H. Kobayashi and A. Konheim, "Queueing models for computer communications system analysis," *IEEE Trans. Commun.*, vol. 25, Jan. 1977.

[3] H. Kobayashi, *Discrete-time Queueing Systems*, pp. 53–85. In Louchard and Latouche [24], 1983.

[4] Dafermos and M. Neuts, "A single server queue in discrete-time," *Cahier du Centre d'Etudes de Recherche Operationelle*, vol. 13, pp. 23–40, 1971.

[5] B. Kim, "A single server discrete-time queueing system: with and without priorities," in *Proc. IEEE Globecom'88*, pp. 522–526, IEEE Communications Society, 1988.

[6] D. Bertsekas and R. Gallagher, *Data Networks*. Prentice Hall, 1987.

[7] L. Kleinrock, *Queueing Systems*, vol. I and II. Wiley, 1976.

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

[9] G. Barberis, "A useful tool in the theory of priority queueing," *IEEE Trans. Commun.*, vol. 28, Sept. 1980.

[10] I. Rubin, "Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems," *IEEE Trans. Inform. Theory*, vol. 35, May 1989.

[11] I. Stavrakakis, "Statistical multiplexing under non-i.i.d. packet arrival processes and different priority policies," *Performance Evaluation Journal*, no. 12, pp. 181–189, 1991.

[12] T. Ozaka, "Alternative service queues with mixed exhaustive and k-limited service," *Performance Evaluation Journal*, no. 12, pp. 165–175, 1990.

[13] C. Bisdikian, "A queueing model with applications to bridges and the DQDB (IEEE 802.6) MAN," Research Report RC 15218, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., Dec. 1989. Accepted, *Computer Networks and ISDN Systems Journal*.

[14] R. Newman, Z. Budrikis, and J. Hullett, “The QPSX man,” *IEEE Communications Magazine*, vol. 26, Apr. 1988.

[15] C. Bisdikian, “A quasi-gated service discipline model for a DQDB data station,” Research Report RC 15587, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., Mar. 1990. Presented at the 2nd ORSA Telecommunications Workshop, March 1992, Boca Raton, Fla.

[16] C. Bisdikian, “Waiting time analysis in a single buffer DQDB (802.6) network,” *IEEE Journal on Selected Areas in Communications*, vol. 8, Oct. 1990.

[17] B. Mukherjee and C. Bisdikian, “A journet through the DQDB literature,” Research Report RC 17016, IBM Thomas J. Watson Research Center, Yorktown Heights N.Y., Mar. 1991. To appear in *Performance Evaluation Journal*, Special Issue on High Speed Telecommunications Systems, 1992.

[18] L. Georgiades, L. Merakos, and P. Papantoni-Kazakos, “A method for the delay analysis of random access algorithms whose delay process is regenerative,” *IEEE Journal on Selected Areas in Communications*, vol. 5, July 1987.

[19] I. Stavrakakis and D. Kazakos, “A limited sensing protocol for multi-user packet radio systems,” *IEEE Trans. Commun.*, Apr. 1989.

[20] L. Georgiades and P. Papantoni-Kazakos, “Limited sensing algorithms for the packet broadcast channel,” *IEEE Trans. Inform. Theory*, vol. 31, pp. 280–294, Mar. 1985.

[21] L. Schrage, “An alternative proof of a conservation law for the queue G/G/1,” *Operations Research*, vol. 18, pp. 185–187, 1970.

[22] I. Stavrakakis, “Delay bounds on a queueing system with consistent priorities,” Research Report CSEE/92/04-01, CS/EE Dept., University of Vermont, Burlington, Vt., Apr. 1992.

[23] L. Kantorovich and V. Krylov, *Approximate Methods of Higher Analysis.* New York: Interscience, 1958.

[24] G. Louchard and G. Latouche, eds., *Probability Theory and Computer Science.* London: Academic Press, 1983.

[25] I. Stavrakakis, “Analysis of a statistical multiplexer under a general input traffic model,” in *Proc. IEEE INFOCOM'90*, (San Francisco), IEEE Communications Society, 5–7June 1990.