Load-Dependent Service Queues With Application to Congestion Control in Broadband Networks

Kin K. Leung
AT&T Labs, Room 4-120
100 Schulz Drive
Red Bank, NJ 07701
Email: kkleung@research.att.com

Abstract: We analyze $D/G/1$ and $M/G/1$ queues where the service time for an arrival depends on the amount of work in the system upon arrival. The models are motivated by the bit dropping methods (e.g., [12]) as a congestion control in broadband networks. The idea there is to reduce packet (cell) transmission time in case of congestion, while maintaining satisfactory quality of service.

To study the load-dependent-service queues, we first derive a functional equation that characterizes the response time for an arbitrary arrival. Then, we employ and extend a technique based on Laguerre functions in [8] to solve the equation. A new recursion is developed in this paper, which significantly simplifies and enhances the applicability of the technique, and makes it easy to program on computers. Numerical examples for packetized voice in a broadband network are included. The enhanced technique is also applicable to other communication and computer systems that can be characterized by similar functional equations.

1. Introduction

In this paper, we analyze a $D/G/1$ and an $M/G/1$ queue with load-dependent service time. Specifically, the service time for an arrival can be selected from a fixed number of probability distribution functions (PDFs) and the choice of PDF for an arrival depends on the amount of unfinished work in the system upon arrival. For example, when the system has a large amount of work backloged, new arrivals are forced to receive low-grade service (i.e., short service time). To our knowledge, the queueing models with such load-dependent service have not been analyzed previously. Let these models be referred to as the $D/\tilde{G}/1$ and $M/\tilde{G}/1$ queue in the following.

These models are motivated by bit dropping methods (e.g., [6], [11], [12]) used as a congestion control for continuous-bit-rate (CBR) traffic in broadband networks. For CBR services such as packetized voice, voice samples of a conversation are taken periodically and transmitted over the communication channel. In case of congestion, the idea of the bit dropping methods is to discard certain portion of the traffic such as the least significant bits of voice samples as a way to "reduce" the transmission time, while maintaining satisfactory quality of service as perceived by end users. An alternative congestion control is to discard certain cells entirely; see e.g., [14], [3] and [2]. Cell discard is to drop all bits of a cell, and thus can be viewed as a special case of the bit dropping methods.

Bit dropping methods can be classified as *input* bit dropping (IBD) and *output* bit dropping (OBD), respectively. In the IBD, bits may be dropped when the packets are placed in the queue waiting for transmission. In contrast, bits are possibly discarded in the OBD only from a packet being transmitted over the channel. Depending on actual implementation of the bit dropping algorithm, the IBD may be more efficient than the OBD or vice versa. In terms of performance analysis, it is natural to use an $M/\tilde{G}/1/K$ model with state-dependent service time to study the queue-length distribution and other performance measures for the OBD (e.g., [11]). This is so because possible bit dropping occurs only on the packet being transmitted. On the other hand, the same model is not appropriate for analyzing the IBD since the queue length of packets with bits possibly dropped does not adequately represent the actual amount of workload for the communication channel. The models under study in this paper can be used to address such shortcoming.

As multimedia traffic including voice, data and video shares common resources in broadband networks, providing low-grade service by discarding less critical portion of traffic as a congestion-control strategy is increasingly useful. The models studied in this paper will help us understand the performance characteristics of such control strategies.

Queueing models with periodical arrivals are applicable to many communications and computer systems [4]. However, except for well-structured service time distributions, previous analysis techniques for the standard $D/G/1$ queue often resorted to approximations; see e.g., [13]. More recently, an efficient technique has been proposed in [1] for computing the waiting-time distribution for $GI/G/1$ queue by numerical transform inversion. However, it is not clear how the inversion technique can be extended to study the $D/\tilde{G}/1$ and $M/\tilde{G}/1$ queue.

To analyze the queueing models with load-dependent service time, we first derive a functional equation that characterizes the response (waiting plus service) time for an arbitrary arrival. Then, we employ and enhance a numerical technique in [8] to solve the equation. The technique closely approximates the *complementary cumulative function (CCF)* for the response time for an arbitrary arrival by a weighted sum of Laguerre functions. The functional equation is then transformed into a set of linear equations from which the coefficients of the Laguerre functions are computed. A new recursion is developed in this paper, which significantly simplifies and enhances the applicability of the technique, and makes it easy to program on computers. Numerical examples for packetized voice in a broadband network are included to demonstrate the validity of the enhanced technique.

It is noteworthy that the functional equations for the $D/\tilde{G}/1$ and $M/\tilde{G}/1$ queue are similar to the vacation models with time-limited service in [8] and [9]. In addition, with some modeling restrictions, similar functional equations can be obtained for other communication and computer systems where timers are used to allocate service among multiple types of customers. One of such examples is token-passing networks with a token-holding timer to limit packet transmission at each station. Furthermore, one can derive a similar functional equation, and apply the enhanced technique to perform numerical inversion of Laplace transforms (LT). Because of many applications of this analysis approach, we use the generic terminology in the following discussion.

The organization of the rest of this paper is as follows. In Section 2, we first derive a functional equation that characterizes the response time for an arbitrary arrival in the $D/\tilde{G}/1$ queue. Then, we extend the result to obtain a similar functional equation for the $M/\tilde{G}/1$ queue in Section 3. By obtaining a new recursion in Section 4, we employ and enhance the Laguerre-function technique to transform the functional equation into a set of of linear equations, thus obtaining the response-time distribution. In Section 5, we also derive the probabilities of different grades of service. Section 6 gives several numerical examples for packetized voice in a broadband network. Finally, Section 7 presents our conclusions and future work on the subject.

2. A Functional Equation for the $D/\tilde{G}/1$ Queue

Consider the $D/\tilde{G}/1$ queue with load-dependent service time. Let T be the fixed time period between two successive arrivals. Let Y_i for $i=0$ to K be real, non-negative thresholds. We also define $Y_0 \equiv 0$ and $Y_K \rightarrow \infty$. Further, let the system have K different grades of service and $G_i(t)$ be the PDFs for the grade-i service time for $i=1$ to K. The service time for an arrival is selected from one of $G_i(t)$'s upon arrival according to the amount of unfinished work in the system (denoted by W) immediately prior to the arrival instant as follows:

$$X(t) \;=\; \begin{cases} G_1(t) & \text{if } \; W \le Y_1 \\ G_2(t) & \text{if } \; Y_1 \;<\; W \le Y_1+Y_2 \\ G_3(t) & \text{if } \; Y_1+Y_2 \;<\; W \le Y_1+Y_2+Y_3 \\ \ldots & \ldots \\ G_K(t) & \text{if } \; Y_1+Y_2+\cdots+Y_{K-1} \;<\; W \end{cases} \qquad (1)$$

where $X(t)$ is the service-time PDF for the arrival. For example, an arrival receives grade-1 service with the PDF $G_1(t)$ if the amount of work in the system found by the arrival is less that Y_1.

We make the following assumptions: 1) the system has infinite waiting room, 2) the system is in steady state and by standard arguments, such equilibrium exists if T is greater than the average service time for the grade-K service, and 3) arrivals are served on a first-come-first-served basis.

Let U_n be the response time (waiting plus service time) for the n^{th} arrival. By the Lindley recursion, we have

$$U_n \;=\; W_n \;+\; X_n \;=\; [\; U_{n-1} \;-\; T \;]^+ \;+\; X_n \qquad (2)$$

where $[Z]^+ \;=\; 0$ if $Z \le 0$ and Z otherwise, and W_n and X_n are the waiting and service time (which are correlated because of the load-dependent service) for the n^{th} arrival, respectively. At steady state, let $F_u(t)$, $f_u(t)$ and $U^*(s)$ denote the PDF, the probability density function (pdf) and its Laplace transform (L.T.) of the response time U for an arbitrary arrival. Thus, dropping the subscript n at steady state and taking the L.T. on the left and right-hand sides of (2) yield

$$U^*(s) \;=\; \int_0^{\infty} E[\;\exp(-s[t-T]^+ - sX)\;]\; f_u(t)\,dt. \qquad (3)$$

By dividing the range of integration in (3) according to (1) for the selection of service-time distribution, we obtain

$$\begin{aligned} U^*(s) = & \int_0^T E[\exp(-sX(1))]\, f_u(t)\,dt \\ & + \int_T^{T+Y_1} E[\exp(-s(t-T)^+ - sX(1))]\, f_u(t)\,dt \\ & + \int_{T+Y_1}^{T+Y_1+Y_2} E[\exp(-s(t-T)^+ - sX(2))]\, f_u(t)\,dt \\ & + \int_{T+Y_1+Y_2}^{T+Y_1+Y_2+Y_3} E[\exp(-s(t-T)^+ - sX(3))]\, f_u(t)\,dt + \cdots \\ & + \int_{T+Y_1+\ldots+Y_{K-1}}^{\infty} E[\exp(-s(t-T)^+ - sX(K))]\, f_u(t)\,dt \end{aligned}$$

where $X(i)$ denotes the service time with the PDF $G_i(t)$. To see the physical meaning of the integrals, let us consider the third term on the right-hand side. The response time for an arbitrary arrival is between $T+Y_1$ and $T+Y_1+Y_2$. Thus, the next arrival after T time units finds the amount of work in the system lying between Y_1 and Y_1+Y_2. According to (1), the service time for the new (also arbitrary) arrival is chosen from $G_2(t)$. Similarly interpretation applies to other terms. After some algebraic manipulation, this equation becomes

$$\begin{aligned} U^*(s) = & F_u(T)\, G_1^*(s) + e^{sT} G_1^*(s) \int_T^{T+Y_1} e^{-st} f_u(t)\,dt \\ & + e^{sT} G_2^*(s) \int_{T+Y_1}^{T+Y_1+Y_2} e^{-st} f_u(t)\,dt + \cdots \\ & + e^{sT} G_K^*(s) \int_{T+Y_1+\cdots+Y_{K-1}}^{\infty} e^{-st} f_u(t)\,dt. \end{aligned} \qquad (4)$$

By definition of L.T., $U^*(s) = \int_0^{\infty} e^{-st} f_u(t)\,dt$. Thus, (4) is the functional equation for $f_u(t)$. Since $f_u(t)$ is a pdf, we have

$$\int_0^{\infty} f_u(t)\;dt \;=\; 1. \qquad (5)$$

Therefore, we need to solve (4) for $f_u(t)$ subject to the condition in (5). Equations (4) and (5) have a unique solution for $f_u(t)$ as the system has a steady state.

The regular $D/G/1$ queue where the service time is independent of the workload in the system can be viewed as a special of the $D/\tilde{G}/1$ queue with $Y_1 \rightarrow \infty$. Using the definition of L.T., it is clear from (4) that the functional equation for the regular $D/G/1$ queue is

$$\int_0^{\infty} e^{-st} f_u(t)\,dt = G^*(s) \int_0^T f_u(t)\,dt \;+\; e^{sT} G^*(s) \int_T^{\infty} e^{-st} f_u(t)\,dt \qquad (6)$$

where $G^*(s)$ is the L.T. for the service time.

3. A Functional Equation for the $M/\tilde{G}/1$ Queue

Consider an $M/\tilde{G}/1$ queue that is identical to the $D/\tilde{G}/1$ queue described above except that arrivals arrive according to a Poisson process with rate λ in the former model. For convenience, let us use the same notation. We notice that (2) to (4) remain valid for the $M/\tilde{G}/1$ queue, although T now is a random variable with an exponential distribution. Thus, the functional equation for the $M/\tilde{G}/1$ queue can be readily obtained by unconditioning the variable T in (4) with the distribution. That is,

$$\begin{aligned} U^*(s) = {} & G_1^*(s) \int_{v=0}^{\infty} \lambda e^{-\lambda v} \int_{t=0}^{v} f_u(t)\, dtdv \\ & + G_1^*(s) \int_{v=0}^{\infty} \lambda e^{-\lambda v} e^{sv} \int_{t=v}^{v+Y_1} e^{-st} f_u(t)\, dtdv \\ & + G_2^*(s) \int_{v=0}^{\infty} \lambda e^{-\lambda v} e^{sv} \int_{t=v+Y_1}^{v+Y_1+Y_2} e^{-st} f_u(t)\, dtdv + \cdots \\ & + G_K^*(s) \int_{v=0}^{\infty} \lambda e^{-\lambda v}\; e^{sv} \int_{t=v+Y_1+\cdots+Y_{K-1}}^{\infty} e^{-st} f_u(t)\, dtdv. \end{aligned} \tag{7}$$

One can verify that

$$G_1^*(s) \int_{v=0}^{\infty} \lambda e^{-\lambda v} \int_{t=0}^{v} f_u(t)\, dt\; dv \;=\; G_1^*(s)\; U^*(\lambda). \tag{8}$$

To consider the rest of integrals on the RHS of (7), we define

$$H(s,a,b) \;\equiv\; \int_{v=0}^{\infty} \lambda e^{(s-\lambda)v} \int_{t=v+a}^{v+b} e^{-st}\; f_u(t)\, dt\; dv \tag{9}$$

where $b \geq a$ and both are real. We change the order of integration in (9) by observing that the ranges of integration can be split as follows:

$$\int_{v=0}^{\infty} \int_{t=v+a}^{v+b} \quad <=> \quad \int_{t=a}^{b} \int_{v=0}^{t-a} \quad + \quad \int_{t=b}^{\infty} \int_{v=t-b}^{t-a}$$

Applying this to (9) and after some algebra, we obtain

$$\begin{aligned} H(s,a,b) = \frac{\lambda}{s-\lambda}\; \{ & \; e^{-a(s-\lambda)} \int_a^{\infty} e^{-\lambda t} f_u(t)\, dt - \int_a^b e^{-st} f_u(t)\, dt \\ & - e^{-b(s-\lambda)} \int_b^{\infty} e^{-\lambda t} f_u(t)\, dt \;\}. \end{aligned} \tag{10}$$

By putting (8) into (7) and expressing the integrals involving $G_i^*(s)$ in form of (10), we obtain

$$\begin{aligned} U^*(s) = U^*(\lambda)\, G_1^*(s) + \sum_{k=1}^{K} \frac{\lambda G_k^*(s)}{s-\lambda} \; \{ & \; \exp[-(s-\lambda) \sum_{i=0}^{k-1} Y_i] \\ & \int_{Y_{k-1}}^{\infty} \exp(-\lambda t)\, f_u(t)\, dt - \int_{\sum_{i=0}^{k-1} Y_i}^{\sum_{i=0}^{k} Y_i} \exp(-st)\, f_u(t)\, dt \\ & - \exp[-(s-\lambda) \sum_{i=0}^{k} Y_i] \int_{\sum_{i=0}^{k} Y_i}^{\infty} \exp(-\lambda t)\, f_u(t)\, dt \;\} \end{aligned} \tag{11}$$

where $Y_0 = 0$ and $Y_K \to \infty$ as defined previously. By definition, replacing $U^*(s)$ by $\int_0^{\infty} e^{-st} f_u(t)\, dt$ in (11) yields the functional equation in $f_u(t)$ for the $M/\tilde{G}/1$ queue.

By standard arguments, this functional equation subject to the condition in (5) has a unique solution for $f_u(t)$ if the average service time for the grade-K service is less than $1/\lambda$.

4. An Enhanced Solution Technique

To make this paper concise, we focus only on a solution technique to solve (4) for the $D/\tilde{G}/1$ queue in this section. Due to the similarity between (4) and (11), the same approach is also applicable to solve the equation for the $M/\tilde{G}/1$ queue.

To solve for $f_u(t)$ in (4), we employ and extend the Laguerre-function technique in [8]. We first closely approximate the CCF, $F_u^c(t)$, by a weighted sum of Laguerre functions with unknown weighting coefficients. Then, the approximations for both $F_u(t)$ and $U^*(s)$ are substituted into (4) to produce a set of linear equations. Solving these equations gives the unknown coefficients.

4.1 Close Approximations by Laguerre Functions

$F_u^c(t)$ is approximated by a weighted sum of Laguerre functions:

$$F_u^c(t) \;=\; 1 \;-\; F_u(t) \;\approx\; \sum_{n=0}^{N} a_n\; e^{-t/2C} L_n(t/C) \tag{12}$$

where $C > 0$ is a time-scaling factor, the a_n's are unknown coefficients and $L_n(.)$ is the Laguerre polynomial of degree n. Since $f_u(t) = d[1 - F_u^c(t)]/dt$, differentiating (12) yields

$$f_u(t) \;\approx\; \left[1 - \sum_{n=0}^{N} a_n \right] \delta(t) \;-\; \sum_{n=0}^{N} a_n \frac{d}{dt} \left[e^{-t/2C} L_n(t/C) \right] \tag{13}$$

where $\delta(t)$ is the unit impulse. Taking the L.T. on both sides of (12) and using the results for integrals involving Laguerre polynomials (e.g., Item 6 of 7.414 on p.844 of [5]) yield

$$U^*(s) \;\approx\; 1 \;-\; \sum_{n=0}^{N} a_n \; \frac{s\; (s \;-\; 1/2C)^n}{(s \;+\; 1/2C)^{n+1}}\; . \tag{14}$$

It is important to note that putting $s = 0$ in (14) yields $U^*(0) = 1$, which is equivalent to the condition in (5). This is so because $F_u^c(t)$ as approximated by (12) tends to zero as $t \to \infty$ and this is equivalent to the pdf having unit area.

Substituting $F_u(t)$ in (12) into (4), we get

$$\begin{aligned} U^*(s) = {} & G_1^*(s) [1 - \sum_{n=0}^{N} a_n e^{-t/2C} L_n(t/C)] \\ & - e^{sT} G_1^*(s) \sum_{n=0}^{N} a_n \int_T^{T+Y_1} e^{-st} d[e^{-t/2C} L_n(t/C)] \\ & - e^{sT} G_2^*(s) \sum_{n=0}^{N} a_n \int_{T+Y_1}^{T+Y_1+Y_2} e^{-st} d[e^{-t/2C} L_n(t/C)] - \cdots \\ & - e^{sT} G_K^*(s) \sum_{n=0}^{N} a_n \int_{T+Y_1+\ldots+Y_{K-1}}^{\infty} e^{-st} d[e^{-t/2C} L_n(t/C)]. \end{aligned} \tag{15}$$

To transform (15) into a set of linear equations, one could use the original approach in [8] to carry out the integrations by expanding the Laguerre polynomials $L_n(.)$'s. However, such an approach is very tedious and it turns out that an efficient recursion exists for such integrations as follows.

4.2 A New Recursion for the Laguerre Integrations

Let us first define

$$I_n(a,x,y) \;\equiv\; \int_x^y e^{-at}\; L_n(t)\, dt \tag{16}$$

for $n = 1,2,3,\cdots$ By integration by parts and using the property of derivatives of Laguerre polynomials (i.e., Item 1 of 8.971 on p.1037 in [5]), we obtain from (16) the following recursion:

$$I_n(a,x,y) = \frac{1}{a}\left\{e^{-ax}[L_n(x) - L_{n-1}(x)] - e^{-ay}[L_n(y) - L_{n-1}(y)]\right\}$$

$$+ (1 - \frac{1}{a}) I_{n-1}(a,x,y). \qquad (17)$$

For $n=0$, it is easy to find

$$I_0(a,x,y) = \frac{1}{a}\,(e^{-ax} - e^{-ay}). \qquad (18)$$

Using (17) and (18) together with the recurrence relation for Laguerre polynomials:

$$\begin{aligned} L_0(x) &= 1 \\ L_1(x) &= 1 - x \\ nL_n(x) &= (2n - 1 - x)L_{n-1}(x) - (n - 1)L_{n-2}(x) \qquad (19) \end{aligned}$$

for $n \geq 2$, one can efficiently compute $I_n(a,x,y)$ for all $n \geq 0$ and given values of a, x and y.

4.3 Conversion of Functional Equation to Linear Equations

Let us consider the integrals in (15). One can verify that

$$J_n(x,y) \equiv \int_x^y e^{-st} d[e^{-t/2C} L_n(\frac{t}{C})]$$

$$= \exp[-(s + \frac{1}{2C})y]\; L_n(\frac{y}{C}) - \exp[-(s + \frac{1}{2C})x] L_n(\frac{x}{C})$$

$$+\; sC\; I_n(sC + \frac{1}{2}, \frac{x}{C}, \frac{y}{C}) \qquad (20)$$

where $I_n(.)$'s and $L_n(.)$'s can be obtained recursively from (17) to (19). Expressing all integrals in (15) in the form of (20) and replacing $U^*(s)$ by (14), we finally get

$$\sum_{n=0}^{N} a_n \{\exp(sT) \sum_{k=1}^{K-1} \exp[-(s + \frac{1}{2C})(T + \sum_{i=1}^{k} Y_i)]$$

$$L_n(\frac{T + \sum_{i=1}^{k} Y_i}{C})[G^*_{k+1}(s) - G^*_k(s)]$$

$$+ \exp(sT) \sum_{k=1}^{K} G^*_k(s)\, sC\; I_n(sC + \frac{1}{2}, \frac{T + \sum_{i=1}^{k-1} Y_i}{C}, \frac{T + \sum_{i=1}^{k} Y_i}{C})$$

$$- \frac{s(s - 1/2C)^n}{(s + 1/2C)^{n+1}}\} \;=\; G^*_1(s) - 1. \qquad (21)$$

Clearly, (21) is a linear equation for the unknown a_n's for any s with $\mathrm{Re}(s) \geq 0$. Given the values for s and the model parameters, the coefficients in this equation can thus be efficiently computed by the recursion formulas (17) to (19) for $I_n(.)$'s and $L_n(.)$'s.

It is important to note that the new recursion in (17) and (18) significantly enhances the Laguerre-function technique in terms of its applicability to various systems, numerical stability and computer programming. This is so because the original approach in [8] requires a great deal of algebraic manipulation, which may cause difficulty in writing computer programs for the technique. In fact, the new recursion can readily be used to simplify the analysis of the vacation models with time-limited service in [8] and [9]. Furthermore, besides the load-dependent-service queues studied in this paper, we expect the enhanced technique is also applicable to analyze other communication and computer systems that can be characterized by functional equations similar to (4) or (11).

To solve for the a_n's from (21), we force both sides of the equation to be equal for $N+1$ appropriately chosen values of s. Specifically, we consider s to be pure imaginary, i.e., $s = i\omega$. Since both sides of (21) are analytic functions, it can be reduced to one that equates only the real part of each side. As $N+1$ unknowns are involved, substituting $N+1$ appropriately chosen values of ω into the equation yields $N+1$ linear equations.

To take advantage of a trigonometric property, we change ω to a new variable θ where $\omega = cot(\theta/2)/(2T)$. We choose the same values of θ as used in [8] and [9] to generate the linear equations. That is,

$$\theta_j = \frac{2j + 1}{N + 1} \cdot \frac{\pi}{2} \qquad (22)$$

for $j = 0,1,...,N$. Accordingly, $\omega_j = cot(\theta_j/2)/(2T)$. For each $j = 0,1,...,N$, substituting ω_j into (21) generates one linear equation. As a result, a_n for $n = 0,1,...,N$, can be solved from the set of $N+1$ linear equations.

5. Response Time & Probability of Grades of Service

For both $D/\tilde{G}/1$ and $M/\tilde{G}/1$ queue, once the a_n's are computed, the response-time distribution $F_u(t)$ can be obtained from (12). We remark that by use of recursion in (19), $F_u(t)$ as expressed in terms of Laguerre functions in (12) is easy to compute for any given argument. Differentiating (14) at $s = 0$ yields the average response time, $\bar{u}$. That is,

$$\bar{u} = \sum_{n=0}^{N} (-1)^n (2C) a_n. \qquad (23)$$

Now, let us consider the probabilities of different grades of service in the $D/\tilde{G}/1$ queue. Let P_i be the probability for an arbitrary arrival to receive a service based on the PDF $G_i(t)$ (i.e., the grade-i service) for $i = 1$ to K. We have

$$P_1 = F_u(T + Y_1) \qquad (24)$$

because $F_u(T + Y_1)$ is probability that an arrival sees, including its service time, the amount of work in the system less than or equal to $T + Y_1$ immediately after the arrival epoch. According to (1), this guarantees that the next arrival arriving T time units later is to receive service based on the PDF $G_1(t)$. Similarly, we have

$$P_k = F_u(T + \textstyle\sum_{i=1}^{k} Y_i) - F_u(T + \textstyle\sum_{i=1}^{k-1} Y_i) \qquad (25)$$

for $k = 2$ to K.

Following the same reasoning in Section 3, the P_k's for the $M/\tilde{G}/1$ queue can be obtained by unconditioning T in (24) and (25) by the exponential distribution. Doing so yields

$$P_1 = F_u(Y_1) + \int_{Y_1}^{\infty} e^{-\lambda(t - Y_1)}\; f_u(t)\; dt.$$

Substituting (13) into the above, we obtain

$$P_1 = F_u(Y_1) - \sum_{n=0}^{N} a_n\; e^{\lambda Y_1}\; J_n(\lambda, Y_1, \infty) \qquad (26)$$

where the $J_n(.)$'s in (20) can be computed recursively from (17) to (19). For $2 \leq k \leq K$, unconditioning T in (25) yields

$$P_k = \int_{v=0}^{\infty} \int_{t=v+Y_1+\cdots+Y_{k-1}}^{v+Y_1+\cdots+Y_k} \lambda e^{-\lambda v} f_u(t)\ dt\ dv.$$

This leads us to an observation that $P_k = H(0, \sum_{i=1}^{k-1} Y_i, \sum_{i=1}^{k} Y_i)$ defined by (9). Hence, using (10), one can verify that

$$P_k = F_u(\sum_{i=1}^{k} Y_i) - F_u(\sum_{i=1}^{k-1} Y_i)$$

$$+ \sum_{n=0}^{N} a_n e^{\lambda \sum_{i=1}^{k-1} Y_i} \left[J_n(\lambda, \sum_{i=1}^{k-1} Y_i, \infty) - e^{\lambda Y_k} J_n(\lambda, \sum_{i=1}^{k} Y_i, \infty) \right]. \quad (27)$$

6. Numerical Results for Packetized Voice

Before presenting our numerical examples, let us discuss several implementation considerations. The general approaches to the selections of the time-scaling factor, *C*, and the highest degree of Laguerre polynomial, *N*, as discussed in [8], are applicable to the models studied in this paper. From our numerical experience, *N* chosen between 30 to 80 will be adequate for most of parameter settings. As for the value of *C*, we only need to get an approximation of the average response time (denoted by T_a) for the $D/\tilde{G}/1$ or $M/\tilde{G}/1$ queue. Then, *C* can be selected to be $\alpha T_a / N$ where α ranges between 1 to 12, depending on the shape of the pdf for the service time. If the initial selection of *N* and *C* does not produce satisfactory results as discussed below, one can repeat the calculation with a revised choice of *N* and *C*.

To ensure computation correctness, one can check whether or not the magnitudes of the coefficients a_n's (relatively monotonically) reduce to very small numbers (e.g., $< 10^{-7}$) as *n* approaches *N*. If so, the computation is performed correctly. In general, the magnitude of the last few coefficients provides a good indication of the amount of error in the results.

To validate the proposed approach for the $D/\tilde{G}/1$ queue, we first compare our numerical results to existing results. In particular, we consider several instances of the regular $D/M/1$ queue [7]. We find that the average response times obtained from our computation match with the exact results for at least 5 or 6 decimal digits.

Now, let us demonstrate the potential use of the $D/\tilde{G}/1$ to study the bit dropping method for packetized voice in a broadband network. Following the spirit of the study in [11], we assume voice samples from a number of voice sources are generated to form a packet periodically, which is waiting for transmission by a communication link. Each sample has four bits and, to enable efficient bit dropping, the sample bits from all sources are grouped into four blocks in the packet, one block for each of the four bits, such that the whole blocks with less significant bits can be dropped when needed. Prior to the possible bit dropping, the packet transmission time is proportional to the number of active voice sources (i.e., conversations in progress). We assume the lengths of successive packets are independent of each other. (This assumption can be justified for certain scenarios.)

Let the input bit dropping scheme be used in the system and there be one level of bit dropping. That is, the block of the least significant bits in a packet is dropped if it finds upon arrival (excluding itself) that the amount of work backloged for the link exceeds a fixed threshold *Y*. As an illustrative example, we assume that packet transmission time with full 4-bit samples is represented by an Erlang-4 distribution with an average of one millisecond. When the bit dropping occurs, the packet transmission time is characterized by an Erlang-3 distribution with an average of 0.75 millisecond.

For several selected threshold *Y* (in millisecond) for bit dropping, Fig.1 shows the relationship between the average packet response time (from the arrival of the packet until its transmission is completed) and the packet interarrival time. As intuitively expected, when the threshold is low, the average packet response time is also low because bit dropping often occurs. Since the number of bits in a voice sample as transmitted by the link significantly affects the voice quality as perceived by users, it is important to quantify the probability of bit dropping. Fig.2 presents such probabilities for the illustrative example. It is interesting to observe the spread of the probabilities for different thresholds when the packet interarrival time is above 0.9 millisecond. This is so because if the interarrival time is further lower than 0.9 millisecond, the link becomes so saturated that bit dropping is very likely to take place regardless of the actual threshold value.

Each curve in Figure 1 and 2 contains 50 data points. For each data point, the enhanced Laguerre-function technique consumes less than one second of CPU time on a SUN SPARC 5 workstation. To ensure correctness of our results, we examine the magnitudes of the last few coefficients a_n's. They are on the order of 10^{-6} or less in all cases. So, we are confident the results are correct.

7. Conclusions

We have analyzed the $D/\tilde{G}/1$ and $M/\tilde{G}/1$ queues where the service time for an arrival can be chosen from a number of distribution functions, depending on the amount of work in the system upon arrival. The models are motivated by the bit dropping methods used as a congestion control algorithm for packetized voice in broadband networks. Functional equations for the queueing models are derived. To solve the equations, we employ and enhance a Laguerre-function technique in [8] to compute the performance measures of interest such as response-time distribution, average response time and probability of different grades of service. In particular, a new recursion for the integrations involving Laguerre functions has been derived, which helps us avoid tedious algebraic manipulation required in the original approach, thus significantly simplifying computer implementation of the technique. Numerical examples have been presented to demonstrate the potential use of the $D/\tilde{G}/1$ model to study the average packet delay and the probability of bit dropping for packetized voice service in a broadband network. (Certainly, the same approach can also be applied to study effects of bit dropping for video traffic.)

With the new recursion for integrations involving Laguerre functions, the enhanced approach is much amenable to computer programming. Thus, in addition to the $D/\tilde{G}/1$and $M/\tilde{G}/1$ models, we expect that the technique can be applied to analyze other communication and computer systems that can be characterized by functional equations similar to those derived in this paper. For example, we are exploring to use the technique to analyze the superposition of multiple deterministic arrivals for packetized voice [10]. Furthermore, it is also possible to extend the technique for bursty arrival processes such as the Markov modulated Poisson process (MMPP), which has been commonly used to study traffic and performance issues for B-ISDN. In this case, one distribution function as approximated by a weighted sum of Laguerre functions may be needed for each state of the MMPP. As a step further, since jitter delay for CBR traffic in ATM networks is often studied by a $D+MMPP/D/1$ queue, the Laguerre approach may also be applicable to provide exact characterization of the jitter.

Acknowledgment. Thanks are due to Gagan Choudhury, Martin Eisenberg, K. Sriram and Ward Whitt for their helpful discussion.

References

[1] J.A. Abate, G.L. Choudhury and W. Whitt, "Calculation of the GI/G/1 Waiting-Time Distribution and its Cumulants from Pollaczek's Formulas," *AEÜ*, Vol.47, pp.311-321, 1993.

[2] H.J. Chao and H. Cheng, "A New QoS-Guaranteed Cell Discarding Strategy: Self-Calibrating Pushout," *Proc. IEEE Globecom'94*, San Francisco, Nov. 1994, pp.929-934.

[3] A.E. Eckberg, B.T. Doshi and R. Zoccolillo, "Controlling Congestion in B-ISDN/ATM: Issues and Strategies," *IEEE Commun. Mag.*, pp.64-70, Sept. 1991.

[4] A.A. Fredericks, B.L. Farrell and D.F. DeMaio, "Approximate Analysis of a Generalized Clocked Schedule," *AT&T Tech. J.*, Vol.64, pp.597-615, 1985.

[5] I.S. Gradshteyn and I.M. Ryzhik, Table of Integrals, Series and Products, Academic Press, 1980.

[6] J.M. Holtzman, "The Interaction Between Queueing and Voice Quality in Variable Bit Rate Packet Voice Systems," *Proc. ITC 11*, Kyoto, Japan, Sept 1985, paper 2.2A4.

[7] L. Kleinrock, *Queueing Systems, Vol. I: Theory*, New York: Wiley, 1975.

[8] K.K. Leung and M. Eisenberg, "A Single-Server Queue With Vacations and Gated Time-Limited Service," *IEEE Trans. on Commun.*, Vol.38, pp.1454-1462, Sept. 1990.

[9] K.K. Leung and M. Eisenberg, "A Single-Server Queue With Vacations and Non-Gated Time-Limited Service," *Perf. Eval.*, Vol.12, pp.115-125, 1991.

[10] B. Sengupta, "A Queue with Superposition of Arrival Streams with an Application to Packet Voice Technology," *Performace'90*, P.J.B. King and R.J. Pooley (Ed.), pp.53-59, Elsevier Science Publishers B.V. (North-Halland), 1990.

[11] K. Sriram and D.M. Lucantoni, "Traffic Smoothing Effects of Bit Dropping in a Packet Voice Multiplexer," *IEEE Tran. on Commun.*, Vol.37, pp.703-712, July 1989.

[12] K. Sriram, R.S. McKinney and M.H. Sherif, "Voice Packetizaton and Compression in Broadband ATM Networks," *IEEE J. Select. Areas Commun.*, Vol.9, pp.294-304, April 1991.

[13] H.C. Tijms, *Stochastic Modelling and Analysis: a Computational Approach*, New York: Wiley, 1986.

[14] N. Yin, S.Q. Li and T.E. Stern, "Congestion Control for Packet Voice by Selective Packet Discarding," *Proc. IEEE Globecom'87*, Tokyo, Japan, Nov. 1987, pp.1782-1786.

Figure 1. Avg. Response Time Vs. Arrival Period

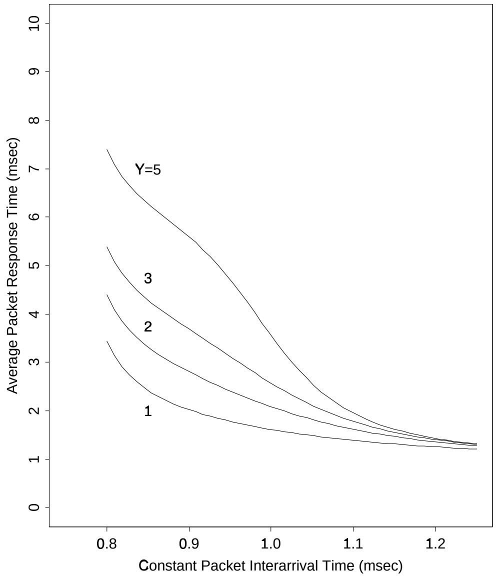
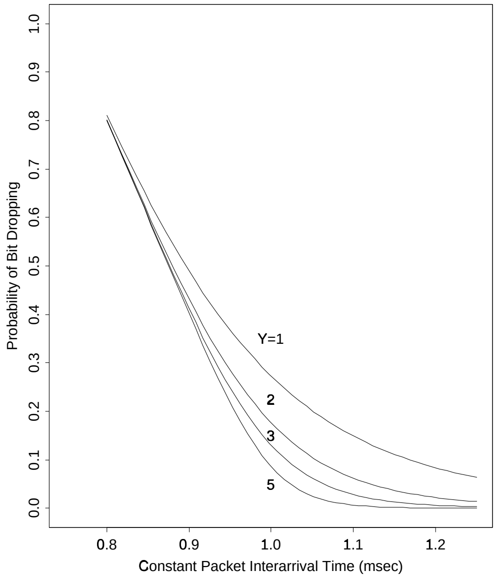


Figure 2. Prob. of Bit Dropping Vs. Arrival Period


Probability of Bit Dropping
Y=1
Constant Packet Interarrival Time (msec)