Message Delay and Queue-Size Analysis for Circuit-Switched TDMA Systems

Izhak Rubin, Fellow, IEEE, and Zhensheng Zhang, Member, IEEE

Abstract—We consider a multiple-access communications channel which is shared among network stations using a circuit-switched TDMA (CS-TDMA) scheme. Each station is allocated a fixed number of slots (N) during each frame. A station provides access to the channel for its sessions, by assigning (on a queued FCFS basis) to a ready session one of its dedicated circuits (each circuit consisting of a single slot per frame), for the total duration of the session. Such CS-TDMA schemes, implemented on a fixed-assignment or demand-assignment basis, are commonly used by local, metropolitan, and satellite networks as effective procedures for providing multiple-access support to real-time services, such as those involving voice, periodic telemetry, video, and high-speed data transmissions. In this paper, we carry out queue-size message delay analysis for CS-TDMA systems. We first derive the generating function of the queue size and of the waiting time distribution for a discrete-time Geom//Geom/N queuing system. This result is used to obtain the generating function of the system size for the circuit-switched TDMA scheme. The associated computation requires the solution of (N + 1)² linear equations. To derive a more computationally effective procedure, tight lower and upper bounds are obtained, requiring the solution of at most 3N linear equations. We prove that a slot allocation scheme which distributes station slots uniformly over the frame yields a message-delay lower bound. We also discuss the application of our results to the analysis of demand-assigned CS-TDMA systems.

I. INTRODUCTION

Time-division multiple-access (TDMA) schemes are widely used to provide for the sharing of a multiaccess communications channel by network stations. Under a TDMA procedure, each station is allocated a fixed number of slots during each time frame. A station is allowed to transmit its messages across the shared channel only during its allocated time-slots.

We differentiate between packet-switched TDMA and circuit-switched TDMA schemes. Consider a time frame structure under which during each time frame, which consists of M = N + L slots, the tagged station is allocated N contiguous slots. Under a packet-switched TDMA access-control procedure, a station transmits its queued messages during its allocated N time-slots, during each frame. In this manner, a station’s packet transmissions take place during its contiguously allocated time-slots. Queued messages can be selected for transmission by the station in accordance with an FCFS or a priority-based service discipline. Such a procedure is commonly used to provide a transport mechanism for store-and-forward packet-switched based services.

In turn, many real-time services are accommodated through the use of a circuit-switched TDMA procedure. Under the latter, the station sets up a session connection for the duration of a session, through the dedication of a number of its allocated time-slots. The number of time-slots per frame allocated to each session depends on the session’s information rate. We assume here that all sessions operate at the same information rate, so that each supported session is assumed to require a dedication of a single station slot during each frame. For example, a voice or video session is supported by allowing the transmission of a single voice or video packet during each frame; the circuit rate is selected to minimize buffering requirements and provide real-time service support. Thus, under the circuit-switched TDMA scheme, the tagged station can accommodate up to N active sessions at a time. Requests arriving at a station for provision of session channel access are assumed here to be queued at the station request buffer (rather than be blocked when all the station’s circuits are occupied), and then served on a first-come–first-served basis.

The queue-size and message delay performance of packet-switched TDMA systems have been investigated by many researchers. Chu and Konheim [3], Kobayashi and Konheim [7], Rubin [11], and Lam [9] have derived the generating function of the queue length for a TDMA scheme under which a single slot per frame is allocated to each station. The message delay distribution has been obtained for a single slot per frame TDMA model by Rubin [11] and [12]. For the multiple slot per frame TDMA scheme, an approximate analysis was carried out by Rubin [12] and Bruncl [2] and an exact analysis has been recently performed by Rubin and Zhang [14]. For schemes that involve a general allocation of station slots across the frame, the following results are available. For such TDMA systems with Poisson arrival streams and single packet messages, generating functions for the packet-size distribution are obtained in [1]. In [5], using a similar model, it is shown that a uniform distribution yields a mean packet delay lower bound. In [10], this model is analyzed with the assumption of multipacket messages to yield expressions for the generating functions of the system packet-size and of message delays. De Moraes and Rubin [4] and Rubin and Tsai [13] analyzed message delays for a TDMA scheme under priority service.
disciplines. A related contiguous-slot allocation model has been studied by Konheim and Meister [8].

We have found no published works involving queue-size and message delay performance analysis for circuit-switched TDMA schemes. Such schemes are modeled and analyzed in this paper. A mathematically precise method for analyzing these schemes is presented. We first analyze related discrete-time queueing systems of the Geom[X]/Geom/N type. We then use these results to carry out exact message delay analysis for circuit-switched TDMA systems. Such an analysis requires the solution of \((N+1)2^N\) linear equations. To reduce the computational requirements, we also derive close upper and lower bounds to the message delay, which we show to be computationally efficient, requiring the solution of at most \(3N\) linear equations. We note that the derived results for these queueing system models are of prime importance for their own sake, in providing for the analysis of many other buffering, queueing, multiplexing, switching, and multiple-access system models. We also examine the dependence of system performance upon the allocation of the station’s slots across the frame. We prove that a uniform slot allocation pattern yields lower message delay levels than those resulting from a contiguous slot allocation scheme.

In Section II, we present the analysis of the discrete-time Geom[X]/Geom/N queueing system. Exact analysis for the circuit-switched TDMA scheme is performed in Section III, while in Section IV, upper and lower bounds for the mean message delay are given. In Section V, we present a numerical example, as well as demonstrate the application of our results to the analysis of demand-assigned (reservation) circuit-switched TDMA systems. Our analysis is also noted to be of importance for application to the performance evaluation of integrated packet-switched and circuit-switched TDMA systems.

II. The Geom[X]/Geom/N QUEUING SYSTEM

We start by carrying out the analysis of a Geom[X]/Geom/N queueing system which provides service to messages generated by a single station. The study of such a discrete-time queueing model is of importance by itself. These results are used in Section III as part of the analysis of the circuit-switched TDMA system.

A. The System Size Distribution

Consider an \(N\)-server Geom[X]/Geom/N queueing system, defined as follows. Time is divided into equal slots, each of duration \(\tau\) seconds. Time-slots are set to start at times \(t = n\tau, n = 0, 1, \ldots\), and slot \(k\) spans the period \((k-1)\tau, k\tau\), \(k \geq 1\). Each message consists of a random number of packets. The transmission time of a packet is set equal to the slot duration \(\tau\). Let \(N_b\) be the number of messages arriving during the \(n\)th slot and \(B_n\) be the number of packets contained in the \(n\)th message. Assume that \(\{N_b, n \geq 1\}\) and \(\{B_n, n \geq 1\}\) are independent identically distributed (i.i.d) sequences of random variables, independent of each other, and that message lengths are geometrically distributed, \(P(B_n = j) = q(1-q)^{j-1}, j \geq 1, 0 < q \leq 1\).

To characterize the message arrival process, we denote the arrival probabilities as \(e_j = P(N_b = j), j \geq 0\), the mean arrival rate as \(\sum_{j=1}^{\infty} j e_j = E(N_b)\), and let the generating function of the number of arrivals per slot be denoted as \(E(z) = \sum_{i=0}^{\infty} e_i z^i, |z| \leq 1\). Message arrivals are recorded at the end of each slot (following any departures that may occur at this time). The arrival process is a geometric batch process, so that a batch (or group) of arrivals occurs at a slot with probability \(1 - e_0\), and the batch distribution is governed by the probabilities \(e_j/(1 - e_0), j \geq 1\). Define \(X_n = X_{n+}\) as the number of messages in the system just after the termination of the \(n\)th slot (or at the start of the \((n+1)\)th slot), including all departures and arrivals that have occurred during the \(n\)th slot. Assume steady-state probabilities to exist and be denoted as \(P_i = \lim_{n \to \infty} P(X_n = i)\), \(P(z) = \sum_{j=0}^{\infty} P_j z^j, |z| \leq 1\). Arriving messages are queued at this station; when any one of the \(N\) servers becomes available, the message at the head of the queue is selected for service. Since \(X = \{X_n, n \geq 1\}\) is a Markov chain, the limiting probabilities \(\{P_i, j \geq 0\}\) are noted to satisfy the following equations:

\[
P_i = \sum_{k=0}^{N+i} P_k P_{ki} \quad i \geq 0, \quad (2.1)
\]

where \(P_{ki} = P(X_{n+1} = i | X_n = k)\), for \(0 \leq k \leq N + i\), is given by

\[
P_{ki} = \sum_{j=(k+i)^+}^{kN} e_{i+j-k}\left(\frac{kN}{j}\right)q^j(1-q)^{kN-j} \quad (2.2)
\]

with \(kN = \min(k, N), (x)^+ = \max(0, x)\). In writing (2.2), one notes that each one of the \(kN\) messages in service at the start of the slot will depart at the end of this slot with probability \(q\) and will stay in the system for at least another slot with probability \(1 - q\). Multiplying by \(z^i\) on both sides of (2.1) and summing over \(i\), we obtain after some algebraic manipulations the following expression as shown below for \(P(z)\).

In order to determine \(P(z)\), we need to find \(P_k\), \(k = 0, 1, \ldots, N - 1\). To this end, we derive \(N\) linear equations for these \(N\) unknowns. Using Takacs' Lemma [15, p. 83], it can be shown that if \(\alpha < N\) (we assume henceforth that this stability condition holds), the denominator of the right-hand side of (2.3) has exactly \(N - 1\) roots in the domain \(|z| < 1\). We denote these roots by \(w_j, j = 1, 2, \ldots, N - 1\). Since \(P(z)\)
is analytic in \(|z| < 1\), the numerator of the right-hand side of (2.3) must be zero at these points. That is

$$
\sum_{k=0}^{N-1} ((1 - q)w_j + q^{j-N} - w_j^{j-N})P_k = 0,
$$

\(j = 1, 2, \ldots, N - 1\).

(2.4)

By the normalizing condition \(P(1) = 1\), we have from (2.3) that

$$
P(z) = \frac{E(z)(N - \alpha)(z - 1)}{z^N - E(z)} \prod_{j=1}^{N-1} \frac{z - w_j}{1 - w_j},
$$

(2.6)

where \(w_j, j = 1, 2, \ldots, N - 1\) are the \(N - 1\) roots of \(z^N - E(z)\) in \(|z| < 1\). We further note that this equation also yields the transform of the system packet-size distribution for any Geom\([X]/G/N\) system with \(E(z)\) replaced by \(E(B(z))\) where \(B(z)\) is the generating function of the message-length distribution.

The roots \(w_j, j = 1, 2, \ldots, N - 1\) of the denominator in (2.3) can be obtained by solving the equation

$$
z = \exp(2\pi ij/N)E(z)^{1/N}(1 - q)z + q,
$$

(2.7)

using a repeated substitution algorithm, separately for each \(j = 1, 2, \ldots, N - 1\). The condition \(\alpha < Nq\) guarantees convergence by the fixed point theorem [6, p.45]. In this way, the roots are computed through the solution of \(N - 1\) separate equations, rather than by solving a single functional equation for \(N - 1\) roots. Noting that if \(w_j\) is a complex root, then its complex-conjugate \(w_j^*\) is also a root. This property is used to reduce the numerical computational requirements.

4) For the Geom\([X]/Geom/\infty\) queuing system, we let \(N \rightarrow \infty\) in (2.1) and (2.2), and obtain the generating function of the system size \(P_\infty(z)\) to satisfy the functional equation:

$$
P_\infty(z) = E(z)P_\infty(q + (1 - q)z).
$$

(2.8)

Taking derivatives on both sides of (2.8) and setting \(z = 1\), we obtain the mean system size \(E(X_\infty)\) to be

$$
E(X_\infty) = \frac{q}{\alpha}.
$$

From (2.8), using power series expansion around \(z = 1\), we obtain

$$
P_\infty(z) = \sum_{n=0}^{\infty} g_n (z - 1)^n
$$

(2.9)

where \(g_n\) is given by the following recursive formula:

$$
g_n = \sum_{l=0}^{n} \binom{n}{l}(1 - q)g_l E^{(n-l)}(1)/(1 - (1 - q)^n),
$$

(2.10)

where \(E^{(n)}(1) = \frac{d^n}{dz^n} E(z)|_{z=1}\). The above recursive formula is obtained by computing the \(n\)th \((n = 1, 2, \ldots)\) derivatives on both sides of (2.8) and setting \(z = 1\).

For the Geom\([X]/Geom/N\) system, to simplify the calculation of the mean system-size by avoiding the computation of the above-mentioned \(N - 1\) roots, we can write the following upper and lower bounds.

**Lemma 1:** For the Geom\([X]/Geom/N\) system, at steady state, for \(\alpha < Nq\), we have

$$
Q_L \leq E(X) \leq Q_L + \frac{N - 1}{2}(2 - q).
$$

(2.11)

where

$$
Q_L = \alpha + \frac{2}{Nq - \alpha}(2 - q).
$$

(2.12)

**Proof:** Equations (2.11) and (2.12) are obtained by taking the derivative of \(P(z)\) as given by (2.3), setting \(z = 1\), and obtaining the relationship

$$
E(X) = Q_L + \frac{q(2 - q)}{2(Nq - \alpha)} \sum_{k=0}^{N-1} k(N - k)P_k.
$$

(2.13)

We then use the inequality

$$
0 \leq \sum_{k=0}^{N-1} k(N - k)P_k
$$

\(\leq (N - 1)q \sum_{k=0}^{N-1} (N - k)P_k = (N - 1)(Nq - \alpha)
$$

whereby the last equality is derived using relation (2.5). □

Since the mean system size \(E(X)\) decreases as \(N\) increases, we conclude that \(E(X) \geq E(X_\infty) = \frac{q}{\alpha}\). Combining the latter lower bound with the lower bound given by (2.11), we have

$$
E(X) \geq \max\left(\frac{Q_L}{q}, \frac{\alpha}{q}\right).
$$

(2.14)

Note from (2.11) that the upper bound is within \((N - 1)\) slots of the lower bound, independent of the system's traffic intensity.

**B. The Waiting Time Distribution**

Messages arriving in separate groups (or batches) are served in accordance with an FCFS policy, while messages arriving in the same group are selected randomly for service. Let \(W_1\) present the message steady-state waiting time. The message
waiting time is defined as the time duration elapsed between the message arrival time and the time that this message transmission starts. Assume that \( \lambda < Nq \). In this section, we derive an expression for the steady-state message waiting time distribution \( P(W = k) \), \( k \geq 0 \).

Denote \( E_i \) as the event that a message chosen randomly from its arriving group is in the \( i \)th position in the system queue (including itself and counting also messages which are currently being served). Let \( \hat{P}_i \) be the probability that event \( E_i \) occurs. Then

\[
\hat{P}_i = P(E_i) = \sum_{t=0}^{i-1} \hat{P}_t r_{i-t} \quad i \geq 1 \tag{2.15}
\]

where \( r_i = \sum_{m=1}^{\infty} \frac{e_m}{1 - \alpha} \frac{1}{1 - \alpha} = \sum_{m=1}^{\infty} e_m / \alpha, \ i \geq 1 \) is the probability that a randomly selected message is in the \( i \)th position in its group; \( \hat{P}_i \) is the probability that there are \( i \) messages in the system prior to a batch arrival, and is equal to \( P(X_n = i) \), where \( X_n \) represents the number of messages in the system at time \( n \) — including any departure occurring at this time, but excluding arrivals occurring during the \( n \)th slot. Since \( X_n = X_n + N, \hat{P}_i \) is determined by \( \hat{P}_i = \hat{P}_i^* e_i = \sum_{i=0}^{i} \hat{P}_t e_{i-t} \). Hence, given \( \{\hat{P}_i, i \geq 0\} \), we can compute \( \{\hat{P}_i, i \geq 0\} \) recursively with the relationship

\[
\hat{P}_i = \frac{1}{e_0} \left( \hat{P}_i - \sum_{t=0}^{i-1} \hat{P}_t e_{i-t} \right) \quad i \geq 0. \tag{2.16}
\]

To compute the waiting time distribution, we condition on the event \( E_i \) to obtain

\[
P(W = k) = \sum_{i=1}^{\infty} P(W = k | E_i) \hat{P}_i. \tag{2.17}
\]

For \( k = 0 \), we have

\[
P(W = 0) = \sum_{i=1}^{N} \hat{P}_i. \tag{2.18}
\]

For \( k \geq 1, P(W = k | E_i) = 0 \), when \( i \leq N \), and

\[
P(W = k | E_{N+i}) = \sum_{j=(d-N)^+}^{d-1} \binom{N(k-1)}{j} \cdot q^j (1-q)^{N(k-1)-j} \sum_{l=d-j}^{N} \binom{N}{l} q^l (1-q)^{N-l} \tag{2.19}
\]

for \( d \geq 1 \), noting that \( \binom{N(k-1)}{j} q^j (1-q)^{N(k-1)-j} \) is the probability that \( j \) messages depart during \( (k-1) \) slots, and \( \sum_{l=d-j}^{N} \binom{N}{l} (1-q)^{N-l} \) is the probability that at least \( d - j \) messages depart in the \( k \)th slot. Hence, from (2.17), (2.19), we obtain the probability that an arbitrary message must wait \( k \) slots prior to the start of its transmission, to be given for \( k \geq 1 \) as

\[
P(W = k) = \sum_{d=1}^{Nk} \sum_{j=(d-N)^+}^{d-1} \binom{N(k-1)}{j} q^j (1-q)^{N(k-1)-j} \tag{2.20}
\]

The following observations are made.

1) The steps involved in computing \( P(W = k) \) are as follows.
   a) Compute \( \{\hat{P}_i, i \geq 0\} \) using recursive relation (2.16).
   b) Compute \( \hat{P}_i, i \geq 1 \) using (2.15).
   c) Calculate \( P(W = 0) = \hat{P}_i = \hat{P}_0 r_1 = \hat{P}_0 r_1 / e_0 \) \( \tag{2.21} \)
   d) Calculate \( P(W = k) \) for \( k \geq 1 \) using recursive relation (2.16).

2) For \( N = 1 \), we have

\[
P(W = k) = \frac{\hat{P}(k-1) q^{k-1} (1-q)^{d-k+1}}{d-1} \tag{2.22}
\]

For this case \( (N = 1) \), we have a Geo[\( X \)]/Geo/1 system, and the generating function of the waiting time defined by

\[
W(z) = \sum_{j=0}^{\infty} P(W = j) z^j
\]

is obtained from (2.21) and (2.22) and is given by

\[
W(z) = \hat{P}(k-1) \frac{1 - E(B(z))}{\alpha (1 - B(z))} = \frac{P(B(z))}{E(B(z))} \frac{1 - E(B(z))}{\alpha (1 - B(z))} \tag{2.23}
\]

where \( B(z) = \frac{q z}{1 - (1-q)z} \) and \( P(z) \) is given now by (2.3) with \( N = 1 \).

Define the message delay \( D \) as the time duration elapsed between the message arrival time and the time that this message departs (at the termination of its transmission) from the system. Note that the message delay \( D \) is equal to the sum of the message waiting \( W \) and the message transmission time. The mean steady-state message delay \( E(D) \), for \( N = 1 \), is then computed from (2.23) to be given by

\[
E(D) = E(W) + \frac{1}{q} - \frac{1}{q - \alpha} \left( 1 - \alpha + E'(1) \right) \tag{2.24}
\]

which checks with Little’s formula \( E(D) = \frac{1}{\alpha} E(X) \). The latter formula can also be used to compute \( E(D) \), for \( N \geq 1 \), with \( E(X) \) calculated by using the exact expression (2.13), or the bounds (2.11).
III. EXACT ANALYSIS OF THE CIRCUIT-SWITCHING TDMA MODEL

We assume a circuit-switched TDMA scheme under which a session requires a single dedicated station slot during each frame. Time is divided into slots as defined in Section II. In the following analysis we consider the session queue-size and delay behavior of a tagged station. This station is allocated (the first) \( N \) contiguous slots during each time frame. A frame consists of \( M = N + L \) slots. The tagged station's \( N \) slots in a frame are identified as slot index numbers \( 1, 2, \ldots, N \). Session arrivals are assumed to be recorded at the end of each slot. Sessions arriving at a station within a frame are queued for transmission that can start in the next slot belonging to the next frame, so that a session is not allocated a circuit until after its frame of arrival. We assume that available slots are allocated to a newly arriving session in the following manner. The new session is allocated the slot which has the lowest index, when considering all available slots within the frame. Once a slot position is allocated to a session, this position is reserved for this session within each subsequent frame, until this session's transmission terminates. The session holding time (measured in frames) is assumed to be geometrically distributed. Let \( A_n \) be the number of sessions arriving during the \( n \)th slot and \( B_n \) be the holding time of the \( n \)th session (measured in frames). Note, however, that the session's holding time in frames is assumed to be geometrically distributed.

We define, at steady state, the Variables \( Y_k \) and \( X_k \) in the form:

\[
Y_k = \{A_m, m \geq n\} \quad \text{and} \quad X_k = \{B_m, m \geq k\}
\]

where

\[
Y_k = \{A_m, m \geq n\} \quad \text{and} \quad X_k = \{B_m, m \geq k\}
\]

and

\[
P_{kj} = P(X_k = j) = \sum_{x|k| = j} P_k(x) \quad 0 \leq j \leq N
\]

where

\[
P_k(x) = P(X_{1k} = x_1, X_{2k} = x_2, \ldots, X_{Nk} = x_N), \quad x_k = [x]\}
\]

Clearly, \( \{Y_k, k \geq 1\} \) is a discrete-time Markov chain, for which the following equilibrium (steady-state) equations are written. For \( 0 \leq X_k = [x] = j \leq N \), and \( k = 1 \), we have

\[
P_k(x_1, \ldots, x_N) = \sum_{y=0}^{[x]} \alpha_F^{y} \sum_{\{x|y| = [x] - 1, x \geq y\}} P_{N+L}(y)
\]

while for \( 2 \leq k \leq N + 1 \)

\[
P_k(x) = (1 - x_{k-1})(P_{k-1}(x) + q)
\]

\[
P_{k-1}(x_1, \ldots, x_{k-2}, 1, x_k, \ldots, x_N) + (1 - q)x_{k-1}P_{k-1}(x)
\]

and for \( N + 2 \leq k \leq N + L \), we have

\[
P_k(x_1, \ldots, x_N) = P_{k-1}(x_1, \ldots, x_N)
\]

where \( x \geq y \) represents that \( x_i \geq y_i, \quad i = 1, \ldots, N \). For \( X_k = j \geq N \),

\[
P_{kj} = \sum_{m=0}^{j} \alpha_F^{m} P_{N+1+j-m}
\]

\[
P_k(x) = (1 - q)P_{k-1}(x) + qP_{k-1+j+1}
\]

\[
P_k(x) = P_{k-1}(x) + 2 \leq k \leq N + 1
\]

where \( \alpha_F^m = P \) (number of arrivals during a frame = \( m \)) = \( \alpha_F^m \), denoting the \( (N + L) \)th order convolution of \( \{a_n, m \geq 0\} \).

We set \( Z_k \) to denote the steady-state number of sessions in the system at the start of the \( k \)th slot within a frame, including some frame session arrivals (and departures) which have occurred prior to the start of the \( k \)th slot. At steady state, the variable \( Z_k \) has the same distribution as the sum of \( X_k \) and the number of arrivals occurring during the first \( (k - 1) \) slots of the frame, i.e., \( Z_k = X_k + \sum_{i=1}^{k-1} A_i \). Let

\[
P_j = \frac{1}{N+L} \sum_{k=1}^{N+L} P(Z_k = j, \quad j \geq 0
\]

Express the steady-state probability for \( \{Z_k, 1 \leq k \leq N + L\} \) averaged (positionwise) over the frame. Define, for \( |z| \leq 1 \), the generating functions

\[
P(z) = \sum_{j=0}^{\infty} P_j z^j
\]

\[
P_k(z) = \sum_{j=0}^{\infty} P_{kj} z^j
\]

For the \( z \)-transform of \( \{P_j, j \geq 0\} \) we then obtain

\[
P(z) = \frac{1}{N+L} \sum_{k=1}^{N+L} A(z)^{k-1}P_k(z)
\]

It is readily noted that \( P_k(z) \), which involves the queue size at the starting slots of the frames, is identical to the corresponding generating function obtained for a Geo/Geo system, so that \( P_k(z) \) is given by (2.3), with \( E(z) = A(z)^{N+L} \), assuming that \( \lambda < Nq/(N+L) \). Therefore, the probabilities \( \{P_j, 0 \leq j \leq N - 1\} \) are computed by solving (2.4) and (2.5) with

\[
E(z) = A(z)^{N+L}
\]

For \( P_j, 0 \leq j \leq N - 1 \). We then set

\[
P_j = P_0, 0 \leq j \leq N - 1
\]

From (3.1d)-(3.1f), we obtain the
expression for $P_k(z)$, $1 \leq k \leq N + L$ to be given as

$$P_1(z) = A(z)^{N+L}P_{N+L}(z) + A(z)^{N+L} \sum_{j=0}^{N-1} P_{N+1+j}z^j - \sum_{j=0}^{N-1} P_{1+j}z^j$$  (3.3)

$$P_k(z) = (1 - q + qz^{-1})P_{k-1}(z) - qz^{N-1}P_{k-N} \quad 2 \leq k \leq N + 1$$  (3.4)

$$P_k(z) = P_{k-1}(z) \quad N + 2 \leq k \leq N + L$$  (3.5)

and for $P_k(z)$, we have

$$P_k(z) = \sum_{j=0}^{N-1} P_{kj}z^j + P_k(z) \quad 1 \leq k \leq N + L.$$  (3.6)

From (3.2)-(3.6), we note that, to determine $P(z)$, we need to find $P_{kj}$, $1 \leq k \leq N + 1, 0 \leq j \leq N$, which requires the computation of $P_k(z)$, for $0 \leq |z| \leq N, 1 \leq k \leq N + 1$. The latter probabilities are obtained as the solutions to (3.1a), (3.1b), noting that by (3.1c), $P_{N+1}(y) = P_{N+1}(y)$, and using $\sum_{|z|=j} P_1(z) = P_{1+j}, 0 \leq j \leq N - 1$. Since for each $k$, the number of unknowns, $P_k(z), 0 \leq |z| \leq N$, is equal to $2N$, the total number of unknowns is then equal to $(N + 1)2N$.

Another expression for $P_1(z)$, to be used in Section IV for deriving performance bounds, is obtained by solving (3.3)-(3.6), to yield (3.7) below.

The probabilities $P_{N+1+j}, 0 \leq j \leq N - 1$ and $P_{kN}, 1 \leq k \leq N$ can be obtained by solving the sets of the following linear equations. Noting that the roots $\{w_j, 1 \leq j \leq N - 1\}$ in $|z| < 1$ of the denominator must also be roots of the numerator, since $P_1(z)$ is analytic in $|z| < 1$, we have, for $1 \leq j \leq N - 1$

$$-q \sum_{k=1}^{N} ((1 - q)w_jN^{-k}N_{kj}P_{kN} + \sum_{l=0}^{N-1} w_j^l(P_{N+1+j} - ((1 - q)w_jq)^NP_{1+j}) = 0 \quad (3.8)$$

and for $0 \leq j \leq N - 1$ by (3.1d) we have

$$P_{1+j} = \sum_{m=0}^{N} a_m P_{N+1+j-m}.$$  (3.9)

Recall that the probabilities $\{P_{1+j}, 0 \leq j \leq N - 1\}$ have been computed by solving (2.4) and (2.5) as mentioned above. Equations (3.8) and (3.9) form then a set of $2N - 1$ linear equations for the $2N$ unknowns $\{P_{kN}, 1 \leq k \leq N; P_{N+1+j}, 0 \leq j \leq N - 1\}$. We derive another equation by using the normalization condition $P_1(1) = 1$ to obtain

$$-q \sum_{k=1}^{N} ((N - k)(1 - q) + k - 1)P_{kN} + \sum_{l=0}^{N-1} z^l(P_{N+1+l} - ((1 - q)z + q)^NP_{1+l}) = 0. \quad (3.10)$$

Note that when $q = 1$, the system model reduces to that of a "please wait" packet-switched TDMA scheme with single packet messages, where arriving packets can be assigned slots starting from the frame following their frame of arrival. An exact analysis for the system size and message delay for such a system can also be carried out in a manner similar to that used in [14], where it is assumed that messages can be served starting from their frame of arrival.

IV. UPPER AND LOWER BOUNDS FOR THE PERFORMANCE OF A CIRCUIT-SWITCHED TDMA SYSTEM

From the previous section, we notice that it is impractical (especially when $N$ is large) to obtain numerically exact values for system performance measures, such as mean system size $E(X)$ or mean message delay $E(D)$. In this section, we derive upper and lower bounds for these performance functions. For that purpose, the following result is first obtained.

**Proposition 1:** If $(N + L)\lambda < Nq$, then

$$C \leq E(X) = P'(1) \leq C + (N - 1)d \quad (4.1)$$

where

$$C = \frac{N + L - 1}{2}\lambda + 1 \quad \frac{N + L}{N(L + 1)} \left( \sum_{k=1}^{N+L} P_k(1) + \sum_{j=1}^{N-1} jP_{1+j} + \sum_{j=0}^{N-1} jP_{N+1+j} \right)$$  (4.2)

and $d = \frac{1}{N + L} \left( (N - 1) \sum_{j=1}^{N-1} P_{1+j} + q \sum_{j=1}^{N-1} (N - j)P_{1+j} \right)$  (4.3)

and $P'_{1}(1) = \frac{d}{dz}P_k(z)|_{z=1}$.

**Proof:** From (3.2)-(3.6), we have

$$P'(1) = C + \frac{1}{N + L} \sum_{k=2}^{N} \sum_{j=1}^{N-1} jP_{kj}.$$  (4.4)

From (3.2)-(3.6), we have

$$P_1(z) = \frac{z^N A(z)^{N+L} \left\{-q \sum_{k=1}^{N} ((1 - q)z + q)^{N-k}k^{-1}P_{kN} + \sum_{l=0}^{N-1} z^l(P_{N+1+l} - ((1 - q)z + q)^NP_{1+l}) \right\}}{z^N - A(z)^{N+L}((1 - q)z + q)^N}.$$  (3.7)
It remains to only calculate \[ \sum_{k=2}^{N} \sum_{j=1}^{N-1} jP_{kj} \] which requires the solution of the \((N + 1)2^N\) equations \((3.1a)\) and \((3.1b)\). To simplify the computation of the latter summation, we derive upper and lower bounds for it. Setting \(z = 1\) in \((3.4)\), noting that \(P_k(1) = 1 - \sum_{j=0}^{N-1} P_{kj}, 2 \leq k \leq N + 1\), and applying recursively the resulting expression, we obtain
\[
\sum_{j=0}^{N-1} P_{kj} = \sum_{j=0}^{N-1} P_{kj} + q \sum_{j=0}^{N-1} P_{jN}, \quad 2 \leq k \leq N. \tag{4.5}
\]

By \((3.1b)\), we observe that \(P_{k0} \geq P_{k-10}, 2 \leq k \leq N + 1\), noting that \(P_{00} = P(0)\). Hence, from \((4.5)\),
\[
\sum_{j=1}^{N} P_{kj} \leq \sum_{j=1}^{N-1} P_{j1} + q \sum_{j=1}^{N-1} P_{jN}, \quad 2 \leq k \leq N.
\]

Thus, using these bounds we obtain
\[
0 \leq \sum_{j=2}^{N} \sum_{j=1}^{N-1} jP_{kj} \leq (N - 1) \sum_{j=1}^{N} \sum_{j=2}^{N-1} P_{kj}
\leq (N - 1) \left[ \sum_{j=1}^{N-1} \sum_{j=1}^{N-1} P_{j1} + q \sum_{j=1}^{N-1} P_{jN} \right]
\leq (N - 1) \left[ (N - 1) \sum_{j=1}^{N-1} P_{j1} + q \sum_{j=1}^{N-1} (N - l)P_{jN} \right] .
\]

Substituting \((4.5a)\) in \((4.4)\) yields \((4.1)\).

The parameters \(C\) and \(d\) in \((4.1)\) are computed as follows. As noted before, the probabilities \(P_{kj}, 0 \leq j \leq N - 1\), are first obtained by solving \((2.4)\) and \((2.5)\) for \(\{P_j, 0 \leq j \leq N - 1\}\) with \(E(z) = A(z)^{N+1}\) and setting \(P_{0j} = P_j, 0 \leq j \leq N - 1\). The probabilities \(P_{kJ}, 1 \leq j \leq N\) and \(P_{N+1j}, 0 \leq j \leq N - 1\), are obtained by solving \((3.8)-(3.10)\). These computations require solving two sets of linear equations; the first consists of the set \((2.4)\) and \((2.5)\) of \(N\) linear equations and the second involves the set of \(2N\) linear equations \((3.8)-(3.10)\). The parameters \(P_{k0}(1), 1 \leq k \leq N + L\), needed to compute \(C\), are then obtained recursively as follows:

\[
P_{k0}(1) = E(X_d) - \sum_{j=1}^{N-1} jP_{kj}

P_{k0}(1) = P_{k-10}(1) - q(N - 1)P_{k-1N}

- q \left( 1 - \sum_{j=0}^{N-1} P_{kj} - \sum_{j=1}^{N-1} P_{jN} \right) \quad 2 \leq k \leq N + 1

P_{k0}(1) = P_{(k-1)0}(1) \quad N + 2 \leq k \leq N + L
\]

and \(E(X_d)\) is given in \((2.13)\) as \(E(X)\) by setting \(E(z) = A(z)^{N+L}\). By \((4.3)\) and \((4.5)\), we observe that \(d < (N + L)^{-1}(N - 1) \sum_{j=0}^{N-1} P_{nj} \leq (N - 1)/(N + L)\); since generally, \(L > N\), we conclude that \(d \ll 1\), so that \((4.1)\) provides tight upper and lower bounds to \(E(X)\).

For the session delay \(D\), which is defined as the time period elapsed between the arrival time of the request for the session setup and the time that this session terminates its transmission across the channel, we conclude the following result.

**Proposition 2:** If \((N + L)\lambda < Nq\), then
\[
C_1 \leq E(D) \leq C_1 + (N - 1) \tag{4.6}
\]
with
\[
C_1 = \frac{N + L - 1}{2} + E(D_{N,N+L})(N + L) - (N + L - 1)
\]
where \(E(D_{N,M})\) (measured in frames; each frame being equal to \(M\) slots) is the message delay in the corresponding Geom\([X]/\text{Geom}/N\) system with \(E(z) = A(z)^M\).

**Proof:** We decompose the session delay \(D\) into the sum \(D = U + W_1 + W_2 + T\). The component \(U\) denotes the frame latency, representing the number of slots between the time of arrival of the request for setting up the tagged session and the starting time of the next frame. The component \(W_1\) is the waiting time component representing the number of slots elapsed from the start of the frame following the sessions arrival and the start of the frame in which the session is allocated its first slot. The variable \(W_2\) is governed by the same distribution as that obtained for the waiting time of an arbitrary message in a Geom\([X]/\text{Geom}/N\) system with arrivals governed by the generating function \(E(z) = A(z)^{N+L}\), and a message service time which is set to equal the session’s holding time and is expressed in frames, noting that each slot in this discrete-time queueing system corresponds to a frame in the corresponding TDMA system. The wait component \(W_2\) expresses the time delay between the start of the frame (in which the first transmission slot is allocated) to the time position of that allocated slot within this frame. Thus, \(W_2 = i - 1, 1 \leq i \leq N, \) if the session’s first slot is the \(i\)th slot within the frame. The variable \(T\) expresses the effective session holding time, which is equal to the period elapsed between the start of the first slot allocated to the session and the termination of the last slot provided for this session. It is easy to see that \(E(U) = N\lambda L + 1\) and the average of the waiting-time component \(W_1\) (measured in slots) is \(E(W_1) = E(D_{N,N+L}) - \frac{1}{q}(N + L)\).

The derivation of an exact expression for \(W_2\) is a complex task. Hence, we use the obvious bounds \(0 \leq W_2 \leq N - 1\).

The effective sessions holding time component \(T\) is \(T = (\lambda - 1)(N + L)\).

Combining the above four terms, we conclude result \((4.6)\).

In \((4.7)\), the delay \(E(D_{N,M})\) is computed as \(E(D_{N,M}) = (M\lambda)^{-1}E(X_d)\), where \(E(X_d)\) is given by \((2.13)\) with \(E(z) = A(z)^{N+L}\) and requires the solution of \(N\) linear equations.

We observe that Proposition 2 also holds for a more general arrival process, under which \(\{N_n, n \geq 1\}\) is a sequence of i.i.d. random variables where \(N_n\) represents the number of messages arriving during the \(n\)th frame, and the distribution of \(N_n\) is governed by the mgf \(E(z)\), provided the appropriate expression is substituted for the average frame latency component \(E(U)\).

Note that \((4.6)\) yields tight upper and lower bounds on the mean message delay. These bounds differ by less than a single frame duration. Numerical results show that when \(\lambda = (N + L)\lambda/Nq\) is close to 1, computing \(E(D) = \lambda^{-1}E(X)\)
using (4.1) yields tighter bounds than those computed using (4.6). For low $\rho$ values, the bounds in (4.6) are observed to be tighter than those obtained by using (4.1) (see Section V, Table I). Combining the two propositions and using Little's formula, we obtain the following useful upper and lower bounds for the message delay.

**Proposition 3:** For $(N + L)\lambda < Nq$,

$$\max(C/\lambda, C_1) \leq E(D) \leq \min(C/\lambda + (N - 1)d/\lambda, C_1 + N - 1) \quad (4.8)$$

where $C$ and $C_1$ are given by (4.2) and (4.7), respectively.

It is also interesting to compare the mean system-size (or message delay) values induced under uniform and contiguous slot allocations across the circuit-switched TDMA frame. As in the case of packet-switched TDMA [14], we conclude the following result.

**Proposition 4:** For the case in which $(N + L)/N$ is a positive integer, the steady-state mean message delay under a contiguous slot allocation scheme is higher than or equal to the mean message delay achieved when slots are allocated uniformly over the frame.

**Proof:** Assume that $\frac{N + L}{N} = R$, where $R$ is a positive integer. Consider the TDMA scheme which employs a uniform slot allocation pattern, so that a station is provided a single slot every $R$ slots. Clearly, this scheme is identical to a TDMA scheme which provides for a frame duration of $R$ slots, and for which the tagged station is allocated a single slot per frame ($N = 1$). For the latter scheme, denote $E(D_u)$ as the mean message delay. Let $E(D_C)$ be the mean message delay under a contiguous slot allocation scheme. From (2.11) and (2.12), for the corresponding Geom[1]/Geom/N system, using Little’s formula to calculate $E(D_{N,M}) = E(X)(M\lambda)^{-1}$, we obtain (4.9) and (4.10) shown below,

$$E(D_{1,R}) = 1 + \frac{(R - 1)\lambda^2 + A'(1) + 2\lambda(1 - q)}{2\lambda(q - R\lambda)} \quad (4.9)$$

Using (4.6), (4.7), and substituting (4.9), (4.10) for $E(D_{N,N+L})$ in (4.7), we conclude (see bottom equation) noting that for $N = 1, E(D_u) = C_1$, which completes the proof.

$$E(D_{N,N+L}) \geq 1 + \frac{N\lambda(NR - 1)}{2} + N\lambda A'(1) + 2N\lambda(1 - q) \quad (4.10)$$

From the proof of Proposition 4, we observe the mean delay $E(D_u)$ induced by a uniform slot allocation serves as a lower bound to the mean message delay $E(D_C)$ under contiguous slot allocation, but it is a weaker lower bound than that provided by $C_1$, noting that $C_1 \geq E(D_u)$. Note further that, under a contiguous slot allocation, the mean message delay $E(D)$ can be reduced, if one can reorder the station transmissions every frame, so that they always occupy the first portion of their dedicated period.

**V. NUMERICAL EXAMPLE AND CONCLUSIONS**

Consider a geometric message arrival process, so that a single message arrives at a slot with probability $p$, while no arrivals occur with probability $1 - p$. For $p \leq 0.2$, the upper and lower bounds for the mean message delay versus throughput function, using (4.1), (4.6), and (4.8), respectively, for $L = 9, N = 3$, and $q = 0.4$. We notice that the derived lower and upper bounds are very tight.

The results presented in this paper can be directly applied to the analysis of a demand-assigned (reservation) circuit-switched TDMA (DA/CS-TDMA) system. To illustrate this application, consider such a system employing an in-band signaling/reservation channel. Every frame is now divided into two fixed-length periods: 1) a reservation period which consists of $L$ slots; 2) a service period which contains $N$ slots. A ready station makes first a reservation for a circuit (which consists of a single slot per frame), by transmitting a reservation packet during the reservation period. Assume the mean reservation delay to be separately computed (in accordance with the underlying reservation access control protocol) as $D_{res}$. Reservation requests for circuits are then queued (at a central station, or by all stations) and served on an FCFS basis. The mean value of the service request wait-time ($W_{ser}$) can then be computed by using the results derived here, noting that the underlying station to which $N$ slots are allocated every $N + L$ slots now represents the totality of network stations, so that its arrival process is given as the superposition of the arrival processes of the individual stations, with the corresponding mgf of the number of arrivals per slot being given (assuming arrivals to constitute a sequence of
packet-switched multiple-access procedures. TDMA systems which use integrated circuit-switched and time for a circuit allocation is then computed as a scheme, and may be used as an approximation, otherwise) by reservation periods, the reservation delay under which each station is dedicated reservation slots within D\textsubscript{ren} and \( A_i(z) \) is the corresponding mgf for arrivals at the 4th station, \( i = 1, 2, \ldots, M \).

Note that for the special, but commonly implemented case, or, is also obtained is the corresponding mgf for arrivals at the 4th station, \( i = 1, 2, \ldots, M \).

A similar analysis was carried out in [12] for commercial and military telecommunications systems and networks, computer communications networks; local, metropolitan, and satellite communications networks; queuing and multiple-access schemes; information processing and stochastic process-based models; and C\textsubscript{3} systems. From 1979 to 1980, he served as acting Chief Scientists of the Xerox Telecommunications Network (XTEN). At UCLA, he is currently leading a large research group. He also serves as President of IRI corporation, a leading team of telecommunications, computer communications, and C\textsubscript{3} experts that provides consulting, analysis, design, and seminar services.

Dr. Rubin is a member ofEta Kappa Nu. He served as Co-chairman of the 1981 IEEE International Symposium on Information Theory, a Program Chairman of the 1984 NSF–UCLA workshop on Personal Communications, and as Program Chairman of the 1987 IEEE INFOCOM Conference. He also serves as an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS.
Zhensheng Zhang (M'90) received the B.Sc. degree in applied mathematics from Shandong University, China in 1982, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, in 1984 and 1989, respectively.

He has been with the Center for Telecommunications Research at Columbia University, New York, NY, as an Associate Research Scientist, since 1989. His research interests include performance evaluation and modeling of computer communication networks, queueing theory, ISDN switching systems and optimization, wireless networks, and multihop lightwave networks.

Dr. Zhang is a recipient of the 1988 Phi Beta Kappa Scholarship Award.