Access-Control Disciplines for Multi-Access Communication Channels: Reservation and TDMA Schemes

IZHAK KUBIN, MEMBER, IEEE

Abstract—Reservation and TDMA schemes are studied for governing the access-control discipline for a network of terminals communicating through a multi-access broadcast channel. A single repeater is employed to allow a fully connected network structure. A channel can be characterized as inducing a low propagation-delay value, as for terrestrial radio or line networks, or as being associated with a higher propagation-delay value, as for a satellite communication channel. A synchronized (slotted) communication medium is considered. Messages are composed of a random number of packets, governed by an arbitrary message-length distribution. The process describing the number of reserved message arrivals within each time frame is assumed to be a sequence of i.i.d. random variables, governed by an arbitrary distribution. (A Poisson arrival stream thus becomes a special case.) The reservation access-control disciplines studied in this paper employ message-switching distributed-control procedures. The performance of each access-control scheme is evaluated according to its delay-throughput function. In particular, schemes are developed to adapt their structure, or protocol, dynamically to the underlying fluctuating network traffic flow values. A fixed-reservation access-control (FRAC) discipline is studied, employing a fixed periodic pattern of reservation and service periods. The reservation periods are used for the transmission of reservation packets as well as for the integrated service of other groups of network stations. The latter stations can access the channel during these periods, using any proper access control procedure. As a special case, message-delay distributions and moments under a TDMA scheme are obtained. Using dynamic estimates of the underlying message traffic parameters, a dynamic fixed-reservation access-control (DFRAC) scheme is obtained. An analytical technique, which employs a Markov ratio limit theorem, is presented for the derivation of the delay-throughput performance curves of dynamic demand-assignment reservation schemes. To illustrate its application, asynchronous reservation demand-assignment (ARDA) schemes are developed to adapt automatically to the underlying network traffic characteristics. Such schemes establish reservation slots dynamically according to observed network service demands and queue sizes.

I. INTRODUCTION

A MULTI-ACCESS broadcast channel is considered. The channel is employed to provide communication media between a set of geographically distributed terminals. Due to the broadcast nature of the channel, each terminal (user) is assumed to be able to listen to and receive any messages transmitted by any other terminal in the network, including itself. To achieve these communication capabilities, we assume the network to incorporate a single repeater. This repeater receives the messages transmitted (uplink) by the terminals and transmits them (downlink) back to the terminals through a second distinct channel (using a downlink frequency band different from the uplink one). The repeater can be a satellite transponder in geosynchronous orbit with the earth. The multi access broadcast satellite channel serves as the main motivation for our studies. However, such a repeater can also serve as a radio-repeater station in a radio network where terminals (stationary or mobile) communicate with each other through this station (as is the case for packet radio networks). Furthermore, our access-control studies here are applicable to many line-network situations. This is, for example, the case when an actual loop line network, required to operate in a broadcast mode, is considered. For the latter, the necessary broadcast propagation of information among the terminals is observed to be equivalent, for access-control considerations, to the corresponding flow occurring between the terminals and the repeater in the above-mentioned radio networks. In particular, one notes that each network is characterized by an appropriate round trip message propagation delay (designating the total uplink and downlink message propagation delay), and the latter will serve as an important parameter in our studies.

To utilize efficiently the bandwidth of such a channel and grant acceptable message response times to the terminals which wish to share this channel, one needs to apply an appropriate access-control discipline. The latter will govern the allocation of the channel in time and bandwidth dimensions among the network terminals. Furthermore, one generally wishes to implement a channel-scheduling procedure which will adapt appropriately to the underlying message traffic characteristics. The latter are many times not only a priori unknown, but also tend to fluctuate greatly (between regular and crisis situations, as a function of the time of day, etc.). Such demand-assignment access-control disciplines are studied in this paper.

Associated with the management of the channel access-control function, one generally distinguishes between centralized and distributed control functions. If a centralized control function is applied, a central control station needs to be established to supervise terminal
scheduling. Such an intelligent station can allow the implementation of very sophisticated and efficient access-control disciplines, but will necessitate excess propagation delays for control signals to and from the central controller and will be clearly vulnerable to failures of the latter. On the other hand, a distributed control scheme will generally require more complex local terminals. The latter increased sophistication of the local terminals is necessary to allow the local terminal to carry out its control functions and become integrated within the overall network protocol. Central supervising functions might still be required even for a distributed control scheme to provide, for example, periodic checks or synchronization updates.

Clearly, if each terminal in the network emits a steady flow of messages so that message interarrival times for each terminal have low variance, a fixed scheduling discipline such as frequency division multiple access (FDMA) or time division multiple access (TDMA) will yield an efficient utilization of the channel as well as low user response times. Using a TDMA scheme, the channel is time-shared, on a fixed basis, between the network terminals. In particular, such a scheme precludes fluctuations in the number and character of the network terminals. The same is true for an FDMA scheme. The latter observations also indicate the desirability of having an access-control procedure which allows terminals to join and leave the network efficiently.

When traffic in a terminal is characterized as bursty, low duty-cycle with a high ratio of peak-to-average traffic intensity, a fixed-channel assignment (such as TDMA or FDMA) is not efficient. Instead, a packet-switching technique can be utilized efficiently. Under a packet-switching procedure, a packet is given the whole bandwidth of the channel at the appropriate instants of time, as specified by the access-control discipline.

A number of packet-switching access-control disciplines for a multi-access broadcast channel have been studied, while only a few have been analyzed (in the open literature), using mainly simulation or numerical procedures. A great many more disciplines, which have been proposed and are being presently developed for a multitude of applications, have contained no analytical performance studies. For that purpose, useful analytical (and combined analytical-simulation) techniques are developed in this paper. The packet-switching procedures considered can be divided into two classes, designated as random-access disciplines and reservation disciplines.

When a random-access discipline is utilized, any newly arriving packet is allowed access to the channel upon its arrival. If two or more packets collide, the latter are scheduled for retransmission (at appropriate future random times) by again incorporating a random-access procedure. (For certain studies of these schemes see, for example, [6], [8], and [19].) In particular, one notes that such schemes yield low message delays for very low channel throughput values (the latter is defined as the average rate of successful transmissions). As the channel throughput increases, these schemes will become unstable, yielding excessive message-delay values and diminishing throughput values. (See [19] for the analysis and design of a stabilized group random-access scheme.) The unslotted and slotted random-access schemes (also called ALOHA schemes, with the slotted discipline allowing a terminal to transmit a packet only at the network-synchronized start of a time slot) have maximal throughput values of \( (2e)^{-1} \) and \( 1/e \) of successful packets per slot, respectively; a slot being equal to a packet transmission time. Such low throughput values, compared with the maximal value of one, are unacceptable for many applications. Using additional network observations, the delay-throughput performance of a random-access discipline can be somewhat improved, at the expense of requiring a more complex system. (See [11] for such schemes for packet radio networks.)

To improve the throughput utilization of the channel, a variety of reservation schemes have been proposed. In a reservation scheme, each terminal sends a reservation packet containing request information regarding channel time required by this terminal. These requests are received by the central controller, or by each active terminal in a distributed-control scheme, and are subsequently used to assign channel service times to the requesting terminals, following the system's access-control discipline. Channel bandwidth resources are thus appropriately divided into reservation and service components. The first component is used to process reservation request packets while the second is dedicated for message service. When this division is made on a temporal basis, so that the whole channel bandwidth is appropriately switched between these two modes, one subsequently observes the separation of the channel active time into reservation and service epochs. Different access-control disciplines will thus utilize different protocols to establish the latter epochs.

For example, in a reservation scheme presented (and analyzed, using certain approximations) in [7], terminals are allowed to send reservation packets during preassigned fixed periodic reservation time slots. Terminals contend for channel time during a reservation slot, to achieve a successful transmission of a reservation packet, following a (slotted ALOHA) random-access procedure. Messages which have successfully made their reservations are subsequently assigned dedicated service slots within the available service period. This scheme further allows use of the whole channel bandwidth for reservations as long as the channel is idle.

Another scheme, presented in [10] (with no analysis), assumes a TDMA synchronous time frame to be established, so that each terminal is associated with a certain time slot within the frame. Considering the time frame duration also to exceed the message propagation time, each terminal can make packet reservations in the protocol part if its own slot, to be acted upon during the following time frame. These reservations, which are received by all terminals, are then used by each terminal to
compute its own allocated service slot, following the system's access-control discipline. The discipline used in [10] assigns each requesting terminal its own slot within the time frame, while allocating the free slots among the requesting terminals on a round-robin basis. Another procedure presented in [9] used a similar TDMA frame structure, but implements a slot control algorithm which allows a terminal to own a slot position within the frame if it owned that slot position in the previous frame, and thus yields a more favorable response time to longer messages. Other reservation schemes assume a central controller and a separate link for reservations (see, for example, [11] when considering a packet radio channel). For satellite channels, where long propagation delays exist, such schemes can induce excessive message response times.

Thus, while random-access procedures yield low packet delays for low channel throughput values, reservation schemes when appropriately designed can yield acceptable message delays at higher throughput values. The channel traffic capacity can further be allowed to approach its maximal value. A reservation scheme will usually result in better service for multi-packet messages, while also being able to incorporate a multitude of priority service disciplines.

Queue-size distributions for TDMA schemes have been presented in [20] and [21]. However, results concerning the limiting distribution of the message delay under a TDMA procedure have apparently not been published. (A virtual message waiting-time analysis appears in [20]. See also [26], where the same virtual waiting-time approach is followed.) For a special TDMA structure (each station being assigned a single slot within each time frame), when the message-arrival process is assumed to be a Poisson stream, an expression for the limiting average message delay is obtained in [22] (using a rather complicated approach; see also [23]).

In this paper we present two analytical techniques for deriving message-delay distributions under reservation and TDMA schemes. The first technique involves the derivation of the steady state distribution and the moments of the message delay for a channel which is managed by a general TDMA scheme. We thus obtain message-delay results for a fixed reservation access control (FRAC) scheme, as well as for a TDMA scheme. Under the FRAC scheme, periodic time frames are identified. Each time frame is divided into two periods: a reservation period and a service period. Network stations transmit their reserved messages in the service periods. The reservation periods are used for the transmission of reservation packets. Furthermore, the duration of the reservation period can be chosen also to incorporate the utilization of the channel, within these periods, by other station groups (governed by any proper access-control procedure).

The delay-performance results for a FRAC scheme yield, as a special case, message-delay distributions under a TDMA scheme (where a TDMA station has the dedicated use of a certain number of time slots within each time frame). In particular, we note that these results are derived for a general message-arrival process, generated by requiring the random variables representing the number of message arrivals within each time frame to be independent identically distributed (i.i.d.). (Thus a Poisson stream becomes a special case. Statistical models for more bursty-type arrival streams can subsequently be applied.)

The second technique involves the computation of the limiting message-delay distribution and the moments for certain dynamic access-control schemes, under which the channel state process is modeled as a Markov chain. Employing a Markov ratio limit theorem, the delay-throughput performance curve is derived analytically or, sometimes more efficiently (see [12]), through combined simulation and analysis. This technique is illustrated by applying it to derive the delay-throughput performance curves for a class of asynchronous-reservation demand-assignment (ARDA) schemes. These are simple reservation schemes which dynamically adapt their structure to the underlying message traffic fluctuations. This technique can be applied to compute the delay-throughput performance curves of a multitude of general access-control schemes. It has been used in [19] to study group random-access schemes and in [25] to analyze integrated random-access/reservation schemes.

The system model is presented in Section II. The message-delay analysis for the FRAC scheme is presented in Section III, and message-delay results under a TDMA procedure are deduced. In Section IV we introduce the dynamic ARDA scheme. An analytical technique for the computation of the message delay under such dynamic schemes is presented. We note that the techniques and results given here serve as useful tools in the delay-throughput performance analysis of various access-control schemes temporally integrated with reservation and TDMA schemes.

II. THE NETWORK MODEL

The System

We consider a multi-access broadcast communication channel serving a network of terminals. The channel utilizes a repeater (such as a satellite transponder or a radio relay station) to enable each terminal in the network to communicate (through the repeater) with any other terminal (Fig. 1). Messages transmitted by the terminals are directed, through the channel uplink, to the repeater. The latter then shifts the messages into a disjoint frequency band and broadcasts them (so that each terminal can receive any signal reflected by the transmitter) through the downlink channel towards the network terminals. (Note that no scheduling for message downlink transmissions are required.) We assume a synchronized network structure. Thus time (referenced henceforth with respect to the repeater's time) is divided into fixed-length durations, of \( \tau \) s each, called slots. Terminals will thus start transmissions of messages only at times coinciding with starting times of the synchronized time slots. The
channel is characterized by a propagation delay of \( R \tau \) s or \( R \) slots. (Propagation delays are of the order of milliseconds for packet radio channels and around 0.25 s for geosynchronous satellite channels. Their effective values, however, will frequently be increased to include transceiver turn-around times.)

Terminal messages are considered to be composed of fixed-length packets. Each packet contains \( \mu^{-1} \) bits (including protocol, information, and parity-check bits). Messages are transmitted through the channel at a rate of \( C \) bit/s. Packet transmission time across the channel is thus \((\mu C)^{-1}\) s/packet. We set

\[
\tau = (\mu C)^{-1},
\]

so that a slot duration is equal to the packet transmission time. When considering multi-packet terminal messages, the number of packets associated with the \( n \)th message of any terminal is denoted by \( B_n \). We assume \( \{B_n, n \geq 1\} \) to be an i.i.d. sequence of random variables, with a discrete distribution \( \{\beta_k, k \geq 1\} \) and moment-generating function \( \beta(\theta) \), where

\[
\beta_k = P(B_n = k), \quad k = 1, 2, \cdots; \quad \beta(\theta) = E(\theta^{B_n}), \quad |\theta| < 1;
\]

and moments \( \{b_i, i \geq 1\}, \)

\[
b_i = E(B_n^i)
\]

so that \( b = b_1 < \infty \) and \( b_2 < \infty \). Any two different terminals produce statistically independent (identically distributed) message durations.

New messages arrive at the \( i \)th terminal, \( i = 1, 2, \cdots, M \), according to a stochastic stream of intensity \( \lambda_i \) messages/slot. The overall network message-arrival stream is thus a stochastic point process with an intensity of \( \lambda = \sum_{i=1}^{M} \lambda_i \) messages/slot. Upon the arrival of a new message, a terminal will immediately try to gain access into the channel for this message. If the message has to wait for its access, it is stored within the terminal buffer, possibly along with other messages that have arrived previously but have not yet been granted channel access. Terminals can also be considered to possess buffer storage for only a single message. A terminal will then be "locked" (or blocked) for new message arrivals as long as a prior message at the terminal has not been transmitted. Considering the situation, however, where a large (around ten or more) number of bursty terminals are active, the local terminal congestion is insignificant, so that the second mode of (blocked) terminal operation yields the same network delay-throughput performance as the first one (see for example [8]). The first mode (of unlocked terminals) will thus be assumed henceforth.

We present the analysis of a multitude of access-control disciplines in this paper. The delay-throughput curve will be used to assess the system’s performance as managed by the access-control discipline under consideration. The network throughput is defined as the average (steady-state) number of packets per slot transmitted successfully across the channel. Since equilibrium is achieved for all schemes under consideration and no terminal blocking is applied, the network throughput will equal \( \lambda \) for all disciplines studied here. The message-delay measure, denoted by \( D \), is defined as the time (expressed in time slots) elapsed from the message time of arrival until a successful transmission of the message over the channel is received by the network terminals. Note that, since the system is time-slotted (while messages can sometimes be assumed to arrive according to a continuous-time process), a message arriving within a slot will be considered for channel access only at the start of the following slot. Hence, if actual message arrivals are uniformly distributed over each slot, an average delay component of \( 1/2 \) slot is always associated with any message to describe the delay within its slot of arrival. This extra delay component is common to all the schemes under consideration and is therefore not included in \( D \).

Assuming a distributed-control realization of the access-control disciplines considered here, each terminal is required to store in a queueing table the appropriate state of the system. Each terminal updates its queueing table incorporating the information about channel transmissions that it receives through the broadcast channel. The same disciplines can also be realized by using a centralized controller. The mathematical analysis will then follow readily from that presented for the distributed-control schemes by appropriately incorporating additional reservation propagations to and from the central controller; this analysis will thus be omitted.

The Reservation Period

The access-control disciplines under consideration will require each terminal to make reservations for all the messages that it wishes to transmit. A terminal will then have to send a reservation packet containing information about its identity, the messages it wishes to transmit, and other parameters such as message lengths or priorities. A header containing synchronization bits and parity-check bits must also be included in such a reservation packet.
This packet, however, will generally be shorter than a regular message packet.

Assuming reservation packets to be of fixed length, each containing \( \mu_r^{-1} \) bits, we set
\[
\eta = \lceil \mu^{-1} / \mu_r^{-1} \rceil,
\]
where \( \lceil x \rceil \) denotes the largest integer which is not larger than \( x \). Thus, within a slot duration of \( \tau \) s, at most \( \eta \) reservation packets can be transmitted.

Our main purpose here is to study the system's behavior induced by contentions for channel access of messages that have already made their reservations. We subsequently distinguish between this latter process of contention and the reservation procedure in the following manner. During certain (prescribed) periods of times, called reservation periods, it is assumed that the channel is not available for the transmission of message packets. These periods of times are dedicated for the transmission of reservation packets, as well as for other network service purposes such as the service of other terminal groups sharing the channel, within these periods of times, and governed by any proper access-control procedure.

Reservation packets that employ any proper access control technique can be transmitted by the terminals during reservation periods. In particular, if every terminal is assigned a dedicated reservation minislot within each reservation period, reservations can be made on a contention-free basis. To reduce the size of the contention-free reservation periods, one can, for example, assign (on a fixed basis) reservation minislots to different stations at different duty cycles. For example, certain stations might require a number of reservation minislots in each reservation period, while others might be assigned a single dedicated reservation minislot only once every \( n \) reservation periods, with \( n \geq 2 \). To adapt to the network varying traffic-message characteristics, the assignment of reservations slots could also be dynamically readjusted. It is also possible to realize a contention-free reservation period by using proper code-division multiple-access (CDMA) techniques. Under a CDMA procedure, reservation packets are encoded so that, although their transmissions may overlap in time, each receiver can reliably extract the information directed to it. (The spread-spectrum multiple access (SSMA) technique serves as an example.) Such a procedure will require the transmission of extra code bits (and subsequently will increase network congestion) but can become efficient when short reservation packets are considered or when a joint access/error-control scheme is desired.

To utilize the reservation periods more efficiently, it is sometimes preferable to use a contention scheme to control the access of reservation packets. For example, a random-access scheme, such as the slotted ALOHA [6] or the group random-access (GRA) scheme [19] can be employed. Under the GRA procedure, stations transmit reservation packets at random times within certain reservation periods. Each reservation packet transmission occupies a minislot. If two or more reservation packets are transmitted at the same minislot, the colliding packets will be retransmitted at random times within the next reservation period. Such a scheme can yield very low delays in granting access for reservation packets, provided that the overall reservation traffic within each reservation period is kept low (up to \( 1/e \) reservation packets per minislot for a synchronized slotted channel). This can be achieved by scheduling long enough reservation periods at a high enough duty cycle. (In many actual network situations, reservation packets are very short so that the collision rate is very small as well. The underlying reservation periods can then be regarded as essentially collision-free. This is, for example, the case for the MAROTS network.)

We will not consider explicitly the behavior of the access procedure chosen for the reservation packets. Instead, we assume a statistical model for the number of messages that have made their reservations within each reservation period. For a contention-free reservation procedure, these values represent the number of reservations made by the associated stations to which reservation minislots are dedicated within the period. Under a random-access reservation scheme, they represent the number of successfully transmitted (noncolliding) reservations within the reservation period. The delay-throughput analysis associated with the service of message packets can proceed as if a contention-free reservation were employed. Of course, in assessing the overall message response time, we must include the proper reservation delay. Also, we note that the statistics of the process representing the number of reservations (successfully) made within each reservation period will depend on the specific reservation scheme. To simplify the analysis, we model this process as a sequence of independent random variables. (This is actually the case for contention-free reservation schemes.)

In this paper we present two analytical techniques. The first is used to analyze a family of fixed-reservation access-control schemes. As special cases we also obtain general message-delay results for FDMA and TDMA schemes. The results also apply to networks where a number of station groups, each governed by its own access scheme, share the channel. The second analytical technique is employed to perform the analysis of a dynamic reservation scheme. This procedure is illustrated in Section IV. The reader is referred to [12] for more details.

An arbitrary service discipline can be considered for serving messages which send reservations within a single time period for purposes of average message-delay calculations. To obtain the variance or the distribution of message delay, we assume that these messages are chosen at random for service. [As noted from the analysis, one can also incorporate any other service discipline.] Messages for which reservations were made at an earlier reservation period will however be served prior to a message for which a reservation packet is transmitted within the present reservation period.

In calculating delay-throughput performance curves for the disciplines under consideration, we choose, for illustrative purposes, three characteristic propagation delay values of \( R = 0, 1, 12 \). The first two values apply to radio (or broadcast line) networks where either very short
propagation delays are involved, inducing $R \approx 0$, or short propagation delays and transceiver turn-around times, inducing $R \approx 1$. These values are also representative of satellite communication systems that employ long slot durations (induced for example by low data rates). The value $R = 12$ is associated with a 50-kbit/s satellite channel, inducing a propagation delay of 0.27 s when packets of length 1125 bits are used (see for example [8]).

III. A FIXED-RESERVATION ACCESS-CONTROL SCHEME

Protocol

Channel time is divided into successive time periods, called time frames, each containing $N_F$ slots. Each time frame is divided into two periods. The first serves as the reservation period, and it contains $N_R$ slots. During this period, while reservations are made, other channel transmissions can also take place. The second period in a time frame is the service period. It serves for the orderly transmission of message packets. Each service period contains $N_S$ slots, so that

$$N_F = N_R + N_S.$$ 

The simplest and most basic distributed-control access-control procedure utilizing message reservations is the fixed-reservation access-control (FRAC) scheme governed by the following protocol (see Fig. 2, where we have set $N_R = 1, N_S = N$).

Protocol for FRAC Scheme

A time frame consisting of a reservation period and a service period is established. Then

1) each arriving message will cause a reservation packet to be sent in the next reservation period, or any subsequent reservation period;
2) upon reception of reservations by terminals, the corresponding messages are assigned (by each terminal, observing the state of the system as recorded in its queueing table) free service slots (within the service periods) following an agreed-upon service discipline.

Following each reservation period, each terminal updates the contents of its queueing table. The latter contains information about the presently occupied service slots, subsequently enabling each terminal to deduce the assignment of free service slots following the next reservation period.

Example 1: Consider a reservation period which accommodates all $M$ network terminals on a contention-free basis. The reservation period is then required to contain $M$ minislots. Since, by (2.4), each slot consists of $\eta$ minislots, the minimum number of reservation slots required is

$$N_R = \lceil M / \eta \rceil,$$

where $\lceil x \rceil$ denotes the smallest integer not smaller than $x$.

Example 2: Each reservation period accommodates, on a contention-free basis, $1 / (lM)$ network terminals. Then we require

$$N_R = \lceil M / l \eta \rceil.$$

A network terminal can now, for example, be given the opportunity to send a reservation packet once every $l$ reservation periods.

Example 3: Terminals transmit reservation packets within the reservation periods on a random-access (GRA) basis. Reservation packets now contend for channel access. This contention scheme must be properly designed and controlled (see [19]) to yield acceptable delay-throughput values for reservation packets.

In particular, assume that an average of $\lambda$ new messages arrive within each slot, so that an average of $\lambda N_F$ new messages wish to make reservations within each reservation period. Assume that each reservation period contains

$$K = \eta N_R$$

assigned for the transmission of reservation packets. The throughput $s$ (expressing the limiting average number of successful reservations made in each reservation minislot) is expressed by

$$s = \frac{\lambda (1 - P_R) N_F}{K} = \frac{\lambda (1 - P_R) (N_R + N_S)}{N_R \eta} \leq \frac{1}{\rho},$$

where $P_R$ denotes the probability that a reservation packet is rejected (not admitted; assuming a control scheme which, at certain instants of time, rejects reservation packets; see [19]). In general, to obtain low enough reservation packet delays, we choose $s < 0.3$ and set the control scheme to yield $P_R < 1$ (see [19]). These considerations indicate how to determine the values for $N_R$ and $N_S$ properly when a GRA scheme is employed.

For example, for $\eta = 10$, we find that we can realize a GRA scheme with throughput $s = 0.2$ and $P_R = 0$, provided that we set $N_S / N_R = 5\rho^{-1}$, where $\rho = \lambda (1 + (N_R / N_S))$. The parameter $\rho < 1$ is the underlying reservation traffic intensity. For $\rho = 0.8$, we can use the value $N_S / N_R = 6$, so that a single reservation slot will be provided for each six service slots. At these values we find that the average delay of a reservation packet (from the time it is first transmitted to the time it is first successfully transmitted) is approximately equal to $D_R \sim 0.2(N_R + N_S)$. Note that an
average of about 0.2 retransmissions per reservation packet are realized by a proper control scheme (see [19]).

Example 4: As a special case we assume now that each terminal is assigned a dedicated service period. No reservation transmissions are thus required.

Consider an arbitrary network terminal, say terminal number 1. Assume that, within each time frame, all the $N_S$ service slots are dedicated for the service of this terminal. The remaining $N_R$ reservation slots within each time frame are distributed among the remaining $M-1$ network terminals.

Thus our general FRAC procedure has now become, as a special case, a time-division multiple-access (TDMA) scheme.

Example 5: As a further special case, consider the situation where only a single network terminal is assumed to exist, $M = 1$. All slots are then taken to be dedicated to the transmission of packets generated by this terminal, and we set $N_R = 0$. We note that the FRAC scheme describes the message behavior in a single channel of a slotted frequency-division multiple-access (SFDMA) scheme. Message transmission times are now properly increased to account for the reduced bandwidth (or transmission rate) available to the terminal.

These examples illustrate well the generality of the FRAC model and its generalized TDMA structure. A fixed-reservation scheme that is structurally identical to a FRAC scheme (with $N_R = 1$) has been presented in [7]. This scheme assumes random-access (slotted ALOHA) contention within the reservation slot. The analysis of reserved-message waiting time in [7] incorporates an approximation of the system by an $M/G/l$ queueing system, to be noted and extended in the next section.

We present here an exact-message waiting-time analysis for the general FRAC scheme. Exact results for special cases of the FRAC scheme, particularly for TDMA schemes, are then deduced. We also provide simple upper and lower bounds to the message delay.

The Reservation Arrival Process

Let $R_n$ denote the number of messages making reservations in the $n$th reservation period. We assume that $(R_n, n > 1)$ is a sequence of i.i.d. random variables governed by the probabilities

$$r_k = P(R_n = k), \quad k = 0, 1, 2, \ldots, \sum_{k=0}^{\infty} r_k = 1. \quad (3.1)$$

We set

$$R = E(R_n) = \lambda N_S - \lambda (N_R + N_S), \quad (3.2)$$

so that $\lambda$ denotes the average number of reservations per slot that are carried out. Also assume that $R^2 = E(R_n^2) < \infty$, $R' = E(R_n')$. Let $R^*(z)$ be the $z$-transform (generating function) of $(r_k)$,

$$R^*(z) = \sum_{k=0}^{\infty} r_k z^k, \quad |z| < 1. \quad (3.3)$$

The number of packets for which reservations have been made during the $n$th reservation period is denoted by $\widehat{R}_n, n > 1$. Clearly, $(\widehat{R}_n, n > 1)$ is a sequence of i.i.d. random variables governed by the distribution $(\widehat{r}_k, k = 0, 1, \ldots)$ and $z$-transform $\widehat{R}^*(z)$, where

$$\widehat{r}_k = \sum_{i=k}^{\infty} \beta_i^{(n)} \widehat{r}_k, \quad k = 0, 1, 2, \ldots, \quad (3.4)$$

$$\widehat{R}^*(z) = \sum_{k=0}^{\infty} \beta_k^{(n)} \widehat{R}^*(z), \quad \widehat{R}^*(z) = R^*(\beta(z)), \quad (3.5)$$

Thus, with probability $r_n$, no reservations will be made within a reservation period. On the other hand, with probability $(1-r_n)$, a number of reservations are made within a reservation period. A set of such reservations made within a certain reservation period is called a reservation group. The process representing the arrival of (successful) reservations at the system is thus described (see [13]) as the stochastic jump process $(A_n, G_n, n > 1)$. The random variables $A_n$ and $G_n$ denote, respectively, the time (i.e., index of reservation period) of arrival and the group size (in terms of the number of messages in the group) of the $n$th group. This stochastic jump process is characterized as follows.

The group arrival sequence $(A_n, n > 1)$ is a discrete-time renewal point process with geometrically distributed inter-arrival times (being therefore a discrete-time binomial counting process). Thus, letting $\widehat{T}_n$ denote the inter-arrival time between the $n$th and the $(n+1)$st group, counting service slots alone, we conclude that $(\widehat{T}_n, n > 1)$ is a sequence of i.i.d. random variables governed by the geometric distribution

$$P(\widehat{T}_n = kN_S) = (1-r_0)^{k-1}r_0, \quad k = 1, 2, \ldots, \quad (3.6)$$

for $n > 1$. Note that $T_n = A_n - A_{n-1}$ is related to $\widehat{T}_n$ by the relationship $T_n = \widehat{T}_n + N_R N_S^{-1} \widehat{P}_n$.

The group-size process $(G_n, n > 1)$ is a sequence of i.i.d. random variables statistically independent of $(A_n, n > 1)$, with the distribution $(g_k)$ given by

$$g_k = P(G_n = k) = (1-r_0)^{-1}r_k, \quad k = 1, 2, 3, \ldots, n > 1. \quad (3.7)$$

The number of packets contained in the $n$th reservation group $G_n$ is denoted by $\widehat{G}_n$. Since message lengths are i.i.d. random variables, the process $(\widehat{G}_n, n > 1)$ is a sequence of i.i.d. random variables with the distribution

$$\widehat{g}_k = P(\widehat{G}_n = k) = (1-r_0)^{-1} \widehat{r}_k, \quad k = 1, 2, 3, \ldots, \quad (3.8)$$

The Overall Message Delay

Having characterized the sequences of reservation group sizes and arrival times, we now derive the message (steady-state) delay distribution. The delay $D_n$ of the $n$th message represents the number of slots between the slot of its arrival and the instant it is received by the network stations. It is decomposed into the sum

$$D_n = D_n^{(R)} + \hat{D}_n + D_n^{(S)}. \quad (3.9)$$
The variable $D_n^{(R)}$ denotes the total number of slots between the time of arrival of the $n$th message and the time that its (first) associated group reservation transmission is completed. When a newly arriving message transmits its reservation packet in the next reservation period, we have, for each $n \geq 1$,

$$D_n^{(R)} = U_n + N_n,$$  \hspace{1cm} (3.10)

where $U_n$ denotes the number of slots between the time of arrival of the $n$th message and the starting time of the next reservation period. The delay component $D_n^{(R)}$ consists of the sum of the delay variable $U_n$ and the reservation period delay $N_n$, as expressed by (3.10). We assume that all reservations made within a certain reservation period must be broadcast before any service slot allocation by the stations can be applied.

The distribution of $U_n$ depends on the statistics of the message arrival process. For example, if new message arrivals are uniformly distributed over any time frame, we have, for any $n > 1$,

$$P(U_n = i) = \frac{1}{N_F}, \quad i = 0, 1, \ldots, N_F - 1,$$ \hspace{1cm} (3.11)

so that

$$E(U_n) = \frac{1}{2} (N_F - 1); \quad \text{var}(U_n) = \frac{1}{12} (N_F^2 - 1).$$ \hspace{1cm} (3.12)

This is, in particular, the case when the number of new message arrivals within each slot constitute a sequence of i.i.d. random variables; or when an underlying continuous-time arrival process with stationary independent increments, such as a Poisson process, is assumed.

The variable $\hat{D}_n$ in (3.9) denotes the $n$th message reservation contention delay, describing the period of time since transmission of the first reservation packet is completed to the time when this reservation packet is successfully transmitted. The reservation propagation delay is included in the third delay component ($D_n^{(S)}$). Thus, for a GRA scheme, when $N_F > R$ we have

$$\hat{D}_n = Z_n N_F,$$ \hspace{1cm} (3.13)

where $Z_n$ denotes the number of retransmissions of the $n$th reservation packet. See [19] for the computation of the limiting distribution of $(Z_n)$. In particular, $Z_n = 0$ with probability one, if the $n$th reservation packet is successfully transmitted upon its first transmission.

Under a contention-free reservation procedure we clearly have

$$\hat{D}_n = 0$$ \hspace{1cm} (3.14)

with probability one for each $n \geq 1$. Under a contention reservation scheme the proper expression for the limiting distribution of $\hat{D}_n$ must be incorporated.

The variable $D_n^{(S)}$ in (3.9) represents the $n$th message service delay. It denotes the number of slots between the end of the reservation period containing the time when the $n$th message reservation packet is (successfully) transmitted and the time when the whole $n$th message has been broadcast and received by the network stations. In this section we derive the limiting distribution of $D_n^{(S)}$.

The limiting distributions for the above-mentioned delays, when they exist, are defined as the corresponding Cezaro-one limits. Thus we set

$$D_n^{(S)}(k) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} P(D_n^{(S)} = k), \quad k \geq 0.$$ \hspace{1cm} (3.15)

The limiting distributions $D_R(k)$, $\hat{D}(k)$, and $D(k)$ are defined similarly. The corresponding z-transforms are denoted by $D_R^*(z)$, $\hat{D}^*(z)$, and $D^*(z)$, where

$$D_R^*(z) = \sum_{k=0}^{\infty} z^k D_R(k), \quad |z| < 1.$$ \hspace{1cm} (3.16)

We set $D_R$, $D_R^{(S)}$, $\hat{D}$, and $D$ to be random variables governed by the above corresponding limiting distributions. From (3.10) and (3.11) we have

$$E(D_R) = \frac{1}{2} (N_F - 1) + N_R; \quad \text{var}(D_R) = \frac{1}{12} (N_F^2 - 1).$$ \hspace{1cm} (3.17)

For a contention-free reservation procedure, from (3.14),

$$\hat{D}^*(z) = 1, \quad E(\hat{D}) = \text{var}(\hat{D}) = 0.$$ \hspace{1cm} (3.19)

We will derive the limiting-delay transform $D_R^*(z)$ and obtain the limiting mean $E(D_R)$ and variance $\text{var}(D_R)$. The overall delay transform $D^*(z)$ is then given by (3.9). In particular, since messages which make reservations within the same period are set for service on a random ordering basis, the variable $D_n^{(S)}$ is statistically independent of the variables $D_n^{(R)}$ and $\hat{D}_n$. If the reservation contention scheme is such that $\hat{D}_n$ and $D_n^{(R)}$ are also statistically independent (as is the case under a GRA procedure), we have

$$D^*(z) = D_R^*(z) \hat{D}^*(z) D_n^{(S)}(z), \quad |z| < 1.$$ \hspace{1cm} (3.20)

In particular, under a contention-free reservation scheme (when (3.14) and (3.19) hold), if the reservation delay is described by (3.9) (so that (3.1) holds), we obtain

$$D^*(z) = D_R^*(z) Z_n^{(S)}[N_F(1-z)]^{-1}[1-z^{N_F}], \quad |z| < 1.$$ \hspace{1cm} (3.21)

The Service-Delay Distribution for Reserved Messages

The $n$th message service delay $D_n^{(S)}$ can be expressed as the sum of three random variables,

$$D_n^{(S)} = W_n^{(L)} + W_n^{(G)} + S_n, \quad n \geq 1.$$ \hspace{1cm} (3.22)

The first component $W_n^{(L)}$ is the waiting time (expressed in slots) of the leader among all messages which have made reservations in the reservation period used by the $n$th message. In each arriving group the message which will be served first is designated as the group leader. The associated waiting-time variable, which results when only service slots are counted, is denoted by $\hat{W}_n^{(L)}$. 

The second component $W^{(G)}_j$ denotes the waiting time of the $n$th message beyond that of the leader of its group. When only service slots are counted, the corresponding variable is set to be $\tilde{W}^{(G)}_n$. The third component $S_n$ represents the number of slots required for the transmission and propagation of the $n$th message. (This clearly will depend on the position of the first packet to be transmitted within the time frame and the message length $B_n$.) The number of service slots required for the $n$th message transmission is equal to $B_n$.

The $n$th message service delay in service slots, not including the additional propagation delay, is denoted by $D^{(S)}_n$. Corresponding to (3.22), we now have

$$D^{(S)}_n = \tilde{W}^{(L)}_n + \tilde{W}^{(G)}_n + B_n, \quad n \geq 1.$$  \hspace{1cm} (3.23)

Setting $\lfloor x \rfloor$ to denote the largest integer which is strictly smaller than $x$, we thus relate the delay variables by the expression

$$D^{(S)}_n = \tilde{D}^{(S)}_n + N[R_1 \tilde{D}^{(S)}_n / N_1] + R.$$  \hspace{1cm} (3.24)

The associate (Cezaro-one) limiting $z$-transforms, when they exist, are denoted by $\tilde{W}^{(S)}_L(z)$, $\tilde{W}^{(S)}_G(z)$, and $\tilde{D}^{(S)}_S(z)$. The three random variables in (3.23) are statistically independent. We obtain

$$\tilde{D}^{(S)}_S(z) = \tilde{W}^{(S)}_L(z) \tilde{W}^{(S)}_G(z) \beta(z), \quad |z| < 1.$$  \hspace{1cm} (3.25)

The random variables governed by the above corresponding limiting-delay distributions are denoted by $\tilde{W}^{(S)}_L$, $\tilde{W}^{(S)}_G$, and $\tilde{D}^{(S)}_S$. By (3.25), we note that

$$E(\tilde{D}^{(S)}_S) = E(\tilde{W}^{(S)}_L) + E(\tilde{W}^{(S)}_G) + b;$$  \hspace{1cm} (3.26a)

$$\text{var}(\tilde{D}^{(S)}_S) = \text{var}(\tilde{W}^{(S)}_L) + \text{var}(\tilde{W}^{(S)}_G) + b - b^2.$$  \hspace{1cm} (3.26b)

We will use relations (3.25) and (3.26) to compute $\tilde{D}^{(S)}_S(z)$ and the moments of $\tilde{D}^{(S)}_S$. The distribution and moments of $D_S$ subsequently follow by relation (3.24), since

$$D_S = \tilde{D}^{(S)}_S + N[R_1 \tilde{D}^{(S)}_S / N_1] + R.$$  \hspace{1cm} (3.27)

By (3.27), we note that, with probability one,

$$D^{(S)}_S - N_R < D_S < D^{(S)}_S + N^{(-1)}_S N_1 R^{(s)} + R.$$  \hspace{1cm} (3.28)

The delay $D_S$ is upper bounded by the random variable $D^{(s)}_S$ (being within $N_R$ slots of it), whose distribution is obtained readily from that of $\tilde{D}^{(S)}_S$. (Note that $D_S = D^{(s)}_S - N_R$ when $N_1 = 1$.)

In the remainder of this section we will derive the distribution of $\tilde{D}^{(S)}_S$. The limiting distribution, or $z$ transform $\tilde{W}^{(S)}_n(z)$, of the waiting-time component $\tilde{W}^{(G)}_n$ is computed as follows. Consider the i.i.d. sequence of group sizes $\{G_n, n \geq 1\}$. Associate with it the discrete-time renewal point process $\{C_n, n \geq 1\}$, where $C_n = \sum_{j=1}^{n} G_j, n > 1$. A time index for this point process represents a message, while an event (which occurs at time $C_n$ for some $n$) represents the leader of a group. We choose the $m$th unit of time at random and compute the time elapsed $\eta_m$ between this time and the preceding event occurrence. This time, called the backward recurrence time, clearly represents the number of messages preceding the $m$th message in its own group. (Note that a message is granted service at random among all messages in its group.) From renewal theory, we conclude that [see for example [2, p. 114]] the limiting distribution of $\eta_m$ always exists, and its $z$-transform which is denoted by $\eta^*(z)$, is given by

$$\eta^*(z) = \frac{1 - G^*(z)}{G^*(1 - z)} = \frac{1 - R^*(z)}{R^*(1 - z)}, \quad |z| < 1.$$  \hspace{1cm} (3.29)

Noting that the $k$th message contains $B_k$ packets, we conclude that

$$\tilde{W}^{*}_G(z) = \eta^*(\beta(z)) = \frac{1 - R^*(\beta(z))}{R(1 - \beta(z))}, \quad |z| < 1.$$  \hspace{1cm} (3.30)

The mean and variance are

$$E(\tilde{W}^*_G) = \frac{1}{2} b \left[ (R^*)^{-1} R^2 - 1 \right];$$  \hspace{1cm} (3.31)

$$\text{var}(\tilde{W}^*_G) = \frac{1}{2} (b_2 - b_1^2) \left[ (R^*)^{-1} R^2 - 1 \right]$$

$$+ b^2 \left[ \frac{1}{3} R^3 (R^*)^{-1} - \frac{1}{4} \left( \frac{R^2}{R^*} \right)^2 - \frac{1}{12} \right].$$  \hspace{1cm} (3.32)

where $R^2 = E(R_k^2) < \infty$. As a special case, if the reservation arrival process is a Poisson stream, we have

$$r_k = \lambda N_{R(k)} = e^{-\lambda N_{R(k)}} \frac{(\lambda N_{R(k)})^k}{k!}, \quad k = 0, 1, 2, \ldots.$$  \hspace{1cm} (3.33)

$$R^*(z) - \exp[-\lambda N_{R(1-z)}], \quad |z| < 1.$$  \hspace{1cm} (3.34)

From (3.30)–(3.32) we obtain

$$\tilde{W}^{*}_G(z) = \frac{1 - \exp[-\lambda N_{R(1-\beta(z))}]}{\lambda N_{R(1-\beta(z))}}, \quad |z| < 1,$$  \hspace{1cm} (3.35)

$$E(\tilde{W}^*_G) = \frac{1}{2} b N_S.$$  \hspace{1cm} (3.36)

where

$$\rho = \lambda b N_{R(1)}^{-1}$$  \hspace{1cm} (3.37)

and

$$\text{var}(\tilde{W}^*_G) = \frac{1}{2} b N_S \left[ b_2 b_1^{-1} + \frac{1}{6} \rho N_S \right].$$  \hspace{1cm} (3.38)

It remains to obtain the limiting distribution of $\tilde{W}^{(L)}_n$. The stochastic process $\{\tilde{W}^{(L)}(n), n \geq 1\}$ is a Markov chain governed by the recurrence relationship

$$\tilde{W}^{(L)}(n+1) = \left[ \tilde{W}^{(L)}(n) + G_n - \tilde{R}_{n+1} - \tilde{R} \right]^+ + \tilde{R},$$  \hspace{1cm} (3.39)

where $[x]^+ = \max(0,x)$ and $\tilde{R}$ is defined by the relation

$$\tilde{R} = \tilde{R} + N[R_1 \tilde{R} / N_1].$$  \hspace{1cm} (3.40)

so that $\tilde{R}$ denotes the propagation delay counted in service slots (starting at the beginning of a service period).

Relationship (3.39) is explained by noting that the waiting time (in service slots) $\tilde{W}^{(L)}(n+1)$ of the $(n+1)$st leader is equal to the difference between the overall waiting time and the service time dedicated to the $n$th group $\tilde{W}^{(L)}_n + G_n$.
and the interarrival time $T_{n+1}$, between the $n$th and the $(n+1)$st groups, provided that the latter difference is not smaller than $R$ service slots. It takes $R$ service slots for the information in the $(n+1)$st reservation to propagate and be broadcast to all terminals, so that slots could be assigned to the requesting terminals. Therefore, if the above-mentioned difference is smaller than $R$, we have $W^{(l+1)} = R$. This explains (3.39).

The Markov chain $\{V_n, n \geq 1\}$ is now defined by

$$V_n = W^{(l)} - R, \quad n \geq 1. \quad (3.41)$$

By (3.39), we conclude that $\{V_n, n \geq 1\}$ satisfy the recurrence relationship

$$V_{n+1} = [V_n + \bar{G}_n - \bar{T}_{n+1}]^+, \quad n \geq 1. \quad (3.42)$$

The limiting (Cezaro-one) z-transform $V^*(z)$ of the distribution of $V_n$, is obtained in Appendix A by using results from queueing theory. The variable $V_\infty$ is defined as the random variable governed by this limiting distribution.

Note that the recurrence (3.42) is identical to that satisfied by the message waiting time in a single server GI/G/1 queueing system with message interarrival times $\{T_n\}$ and service times $\{G_n\}$. Therefore, applying Lindley's theorem ([2, p. 168]) to this GI/G/1 queueing system, we conclude that the Markov chain $\{V_n\}$ has a proper limiting distribution, independent of $V_1$, if and only if $E(G_n)/E(T_{n+1}) < 1$; or, equivalently, using (3.8) and (3.6), if and only if the system traffic intensity parameter $\rho$ satisfies

$$\rho = \frac{bN_p b}{N} \rho N_p b / N < 1. \quad (3.43)$$

By the same theorem, if $\rho > 1$, we obtain that $\lim_{n \to \infty} P(V_n < x) = 0$, for every $x$. Using these observations and the results obtained in Appendix A, we obtain the following.

**Theorem 1:** A (Cezaro-one) limiting service delay distribution exists for a FRAC scheme when $\rho < 1$. Its z-transform $\tilde{D}^D(z)$ is given by

$$\tilde{D}^D(z) = \tilde{W}^*_L(z) \tilde{W}^*_R(z) \beta(z), \quad (3.44)$$

where $\tilde{W}^*_L(z)$ is expressed by (3.30), $\tilde{W}^*_R(z)$ is given by

$$\tilde{W}^*_L(z) = z^R V^*(z), \quad (3.45)$$

$$V^*(z) = \frac{N_p (1 - \rho) (1 - z)}{R^* \beta(z)} \prod_{r=1}^{N_p - 1} \frac{z - \eta_r}{1 - \eta_r}, \quad (3.46)$$

and $\eta_r$, $r = 1, 2, \cdots, N_p - 1$ are the distinct $N_p - 1$ roots with $|\eta_r| < 1$ obtained by the limit

$$\eta_r = \lim_{n \to \infty} \eta_r(\omega), \quad (3.47)$$

where $\eta_r(\omega), r = 1, 2, \cdots, N_p$ are the $N_p$ distinct roots of the functional equation

$$z^{N_p} = \omega \tilde{R}^* (\beta(z)), \quad |z| < 1, \quad |\omega| < 1. \quad (3.48)$$

(The roots are ordered so that $\eta_0 = 1$.) For $N_p = 1$ the product in (3.46) is set equal to one.

Furthermore, for $\rho < 1$, $V^*(z)$ is also the z-transform of the limiting distribution of the sequence $\{X_n, n \geq 1\}$, which satisfies the recurrence relationship

$$X_{n+1} = [X_n + \bar{R}_n - N_S]^+, \quad n \geq 1. \quad (3.49)$$

The transform $V^*(z)$ can also be expressed as

$$V^*(z) = \exp \left\{ - \sum_{n=1}^{\infty} \frac{1}{n} \left[ 1 - \sum_{j=0}^{\infty} \sum_{k} \beta_{j+nN_p} f^{(k)}(z) \right] \right\}. \quad (3.50)$$

The limiting mean value $E(V_\infty)$ of the random variable $V_\infty$, the transform of whose distribution is given by $V^*(z)$, is equal to

$$E(V_\infty) = \tilde{V}_u - \frac{1}{2} \rho N_p b \left[ 1 + \frac{1}{2} \sum_{r=1}^{N_p - 1} \frac{1}{1 - \eta_r} \right]. \quad (3.51)$$

and $\rho = \lambda bN_p N_{r-1} < 1$. Furthermore,

$$\tilde{V}_u = \frac{\rho}{2(1-\rho)} \left[ b \bar{b} - b \right]. \quad (3.52)$$

The limiting variance $\text{var}(V_\infty)$ is given by (A-11). For $N_p = 1$, the summations in (3.51) and (A-11) are set equal to zero. For $\rho > 1$, $V_n$ and $\tilde{D}^D_n$ become arbitrarily large as $n \to \infty$.

**Theorem 2:** For a FRAC scheme with $\rho < 1$, the limiting delay transform $D^*(z)$ is given by (3.20). Equation (3.27) is used to obtain $D^D_{\infty}(z)$ from $D^D_{\infty}(z)$. In particular, from (3.10) and (3.11), when arrivals are uniformly distributed over the time frame and when $D^D_{\infty}$ of (3.28) is used to replace (and upper bound) $D_{\infty}$, the resulting upper bound $D^u$ on the limiting delay is given by

$$D^u = \tilde{D}_u N_p N_{r-1} + R + \tilde{D} + U_u + N_R. \quad (3.54)$$

For $\rho < 1$, the limiting mean delay $E(D^u)$ is

$$E(D^u) = N_p N_{r-1} E(V_\infty) + 2R + b N_p N_{r-1} \left[ \frac{1}{2} \sum_{r=1}^{N_p - 1} \frac{1}{1 - \eta_r} \right]$$

$$+ \frac{1}{2} N_p N_{r-1} b \left[ R^{-1} \bar{R}^2 - 1 \right] + \frac{1}{2} \left( N_r - 1 \right) + N_R + E(\tilde{D}). \quad (3.55)$$

The limiting variance $\text{var}(D^u)$ is

$$\text{var}(D^u) = F \left( N_p N_{r-1} \right) \text{var}(V_\infty) + \left( N_p N_{r-1} \right)^2 \left( b \bar{b} - b^2 \right)$$

$$+ \left( N_p N_{r-1} \right)^2 \left( \frac{1}{2} (b \bar{b} - b^2) \left[ \frac{R^*}{R} \right] \right)$$

$$+ \left( b^2 - b \right) \left[ \frac{1}{3} R^2 \frac{R^*}{R} - \frac{1}{12} \right]$$

$$+ \frac{1}{12} \left( N^2 - 1 \right) + \text{var}(\tilde{D}). \quad (3.56)$$

In particular, when the upper bound (3.53) is applied to (3.55), the resulting upper bound to the mean delay,
denoted by \( E(D_u) \), is given by

\[
E(D_u) = E(\hat{D}) + 2R + bN_FN_s^{-1} + N_R
\]

\[
+ \frac{1}{2} (N_F - 1) + \frac{1}{2} N_FN_s^{-1} b \left[ (\lambda N_F)^{-1} R^2 + 1 \right]
\]

\[
+ \frac{\rho N_F}{2N_s(1-\rho)} \left\{ b_2 b_1^{-1} + b \left[ \frac{R^2 - (\lambda N_F)^2}{\lambda N_F} \right] - 1 \right\},
\]

\[
(3.57a)
\]

when \( \rho - \lambda b N_FN_s^{-1} < 1 \). We have

\[
E(D_u) - \frac{1}{2} \rho N_s - N_R \leq E(D^u) - N_R
\]

\[
\leq E(D) \leq E(D^u) \leq E(D_u).
\]

\[
(3.57b)
\]

We note that \( E(V_\infty) \) and \( \text{var}(V_\infty) \) are efficiently computed as the limiting sample moments when (3.49) is simulated. Their values are then used in (3.55), (3.56) to compute the mean and variance message delay functions. Equations (3.57) also present simple computable bounds to the mean message delay. (An upper bound to \( \text{var}(D) \) is similarly obtained by using (A-1).)

In the special case when the reservation arrival process is a Poisson stream, relations (3.33)-(3.38) are incorporated into the above expressions to yield the message delay distribution and moments. (See also (A-13)-(A-15) in Appendix A.) In particular, we obtain \( E(R) = \text{var}(R) \), so that the expression (3.50) for \( V_u \) becomes

\[
V_u = \frac{\rho b_2 b_1^{-1}}{2(1-\rho)}.
\]

Then \( E(D_u) \) for a Poisson arrival stream is given by

\[
E(D_u) = E(\hat{D}) + 2R + bN_FN_s^{-1} + N_R
\]

\[
+ \frac{1}{2} (N_F - 1) + \frac{1}{2} N_FN_s^{-1} b \left[ (\lambda N_F)^{-1} R^2 + 1 \right]
\]

\[
+ \frac{\rho N_F}{2N_s(1-\rho)} \left\{ b_2 b_1^{-1} + b \left[ \frac{R^2 - (\lambda N_F)^2}{\lambda N_F} \right] - 1 \right\},
\]

\[
(3.59)
\]

when \( \rho < 1 \).

We further note that \( V_u \) provides a tight bound to \( E(V_\infty) \) under heavy traffic conditions (\( \rho \approx 1 \)). Under such conditions the distribution of \( V_\infty \) can also be approximated by the geometric distribution

\[
P(V_\infty > k) = \left[ 1 - (\bar{V}_u)^{-1} \right]^k.
\]

An approximation to the limiting mean message delay, to be denoted by \( E(D_A) \), which is frequently used (see for example [7]) when a FRAC scheme is considered and the arrival process is a Poisson stream, is derived as follows. For purposes of analysis, the channel is assumed not to be blocked for access for a reservation period every \( N_F \) slots (and available for service for the rest of the time at a rate of one packet/slot). Instead, the channel is assumed to be available continuously to the users with a reduced channel service rate of \( N_FN_s^{-1} \) packets/slot. Under this assumption, the message waiting time is given by the waiting time of a message in a single-server \( M/G/1 \) queuing system, with message arrival rate \( \lambda \) [messages/slot] and message lengths (service times) equal to \( \{N_FN_s^{-1} b_n : n \geq 1 \} \).

From the characteristics of this queuing system, we find that \( E(D_A) < \infty \) if and only if \( \rho < 1 \), so that the channel traffic capacity is equal to that obtained under the above precise analysis. The steady-state average waiting time for this \( M/G/1 \) system follows from the Pollaczek–Khintchine formula (see for example [2], p. 256) when the above-mentioned modified service times are incorporated, thus yielding the following expression for \( E(D_A) \):

\[
E(D_A) = E(\hat{D}) + 2R + N_R + bN_FN_s^{-1}
\]

\[
+ \frac{1}{2} (N_F - 1) + \frac{N_R b_2 b_1^{-1}}{2N_s(1-\rho)}, \quad \text{for } \rho < 1.
\]

(3.60)

We note that \( E(D_A) - E(D_u) = \frac{1}{2} N_FN_s^{-1} \) slots. Consequently, by (3.57b), both \( E(D_A) - E(D) \) and \( E(D) - E(D_A) \) are upper bounded by \( \lfloor N_R \rfloor \frac{1}{2} N_FN_s^{-1} \). Note however that \( \text{var}(D_A) \) will not generally serve as a good approximation.

If a FRAC scheme with fixed \((N_R, N_s)\) values is used, the latter values should be chosen to yield an acceptable message delay over the predicted range of the network traffic-message characteristic fluctuations and, in particular, the range of the network flow rate \( \lambda \). If, however, the latter range is very wide, involving large fluctuations among low, medium, and high throughput values, an adaptive scheme must be realized. Such a scheme, to be called dynamic fixed-reservation access-control (DFRAC), can be implemented as follows.

We incorporate an estimator (on a centralized or distributed basis) into the FRAC scheme to estimate the present underlying flow rate \( \lambda \) (and any other relevant traffic-message statistics). The estimator computes the values of \( (N_R, N_s) \), to be denoted by \( (N_R^*(\lambda), N_s^*(\lambda)) \), which yield the minimal delay value of \( E(D) = D(\lambda, N_R, N_s) \), denoted by \( D^*(\lambda) \). Thus, for each fixed \( \lambda \) (and other traffic-message statistics), \( \rho < 1 \), the mean delay versus throughput curve \( D^*(\lambda) \) of a DFRAC scheme is

\[
D^*(\lambda) = D(\lambda, N_R^*, N_s^*) = \min_{(N_R, N_s)} D(\lambda, N_R, N_s).
\]

(3.61a)

The DFRAC scheme attains the maximal traffic capacity of one [packet/slot].

For the special case where \( N_R = 1, N_s = N \), a single reservation slot is assumed to allow contention-free reservations \( \hat{D} = 0 \), and the arrival-message process is a Poisson stream with intensity \( \lambda \), we use the exact results presented in Theorems 1 and 2, and show in Fig. 3 the delay-throughput curves \( D_\lambda(\lambda) = E(D^*) \). The service period duration \( N \) is used as a parameter. We assume \( R = 12 \) and single-packet messages \( (b_2 = b = 1, z = 1) \). Note, by (3.55), that \( D - 2R \) is independent of \( R \), so that the corresponding delay-throughput curves for any \( R \) readily follow from Fig. 3. The associated DFRAC scheme will be governed by the delay-throughput curve \( D^*(\lambda) \) where, for each fixed \( \lambda \),

\[
D^*(\lambda) = D(\lambda, N_R^*(\lambda), N_s^*(\lambda)) = \min_{(N_R, N_s)} D(\lambda, N_R, N_s).
\]

(3.61b)

The curve \( D^*(\lambda) \) thus forms the lower envelope of the
curves \( D_s(\lambda), N > 1 \) (see Fig. 3). We have also found that (3.57a) yields a close upper bound for \( E(D^*) \) (for the parameters used in Fig. 3, see [12]).

A different adaptive reservation scheme, whose structure changes dynamically and automatically to accommodate fluctuations in \( \lambda \), is presented in Section IV.

**Message Delays for TDMA Schemes**

It is of particular importance to derive message delays for a TDMA scheme, which is a special case of the FRAC scheme (see Example 4 in Section II). Every station is assigned a dedicated service period within the time frame, so that no reservations are required and \( D_s = 0 \). The message-delay distribution is thus specified as follows.

**Theorem 3:** Consider a TDMA scheme where the station under consideration (called station 1) is assigned \( N_S \) slots within each time frame of \( N_F \) slots, while the other network stations use the remaining \( N_R = N_F - N_S \) slots within each time frame. Assume that \( R_n \), station 1 messages arrive within the nth time frame, \( \{ R_n, n > 1 \} \) is a sequence of i.i.d. random variables, and that the nth message contains \( B_n \) packets where \( \{ B_n, n > 1 \} \) are i.i.d. random variables. When we set \( D = 0 \), the station 1 limiting message-delay distribution is given by the results stated in Theorems 1 and 2.

Explicit message-delay results for a TDMA scheme, where station 1 is assigned a single service slot within a time frame \( (N_S = 1) \), are summarized in Theorem 4. (Now \( D_s = D_S N_F + \tilde{R} - N_R \) and \( D = D_S + U \), so that the service slot is located at the start of the time frame.)

**Theorem 4:** Consider a TDMA scheme as described in Theorem 3. Assume now that station 1 is assigned a single service slot (so that \( N_S = 1 \)) every time frame (every \( N_F \) slots). Then, for \( \rho - \lambda b N_F < 1 \), the (Cezaro-one) limiting message delay distribution exists, and its \( z \)-transform \( D^*(z) \) is given by

\[
D^*(z) = z^{R-N} U^*(z) V^*_\eta(z) W^*_{\eta}(z^N),
\]

where

\[
U^*(z) = \sum_{k=0}^{N_F-1} P(U=k) z^k
\]

is the \( z \)-transform of the variable \( U \) representing the number of slots between the message arrival slot and station 1’s next service slot, \( V^*_\eta(z) \) is given by (3.30), and \( W^*_{\eta}(z) \) is given by (3.46) with \( N_S = 1 \), so that

\[
V^*_\eta(z) = \frac{1 - R^* (\beta(z))}{R [1 - \beta(z)]},
\]

\[
W^*_{\eta}(z) = \frac{1 - R^* (\beta(z))}{R [1 - \beta(z)]}.
\]

The mean \( E(D) \) and variance \( \text{var}(D) \) of the station 1 message-delay variables are given by

\[
\begin{align*}
E(D) &= R - N_R + E(U) + N_F b + \frac{1}{2} N_F b \left[ \frac{\rho N_F}{2(1-\rho)} \left\{ b_2 b^{-1} + b \left[ \text{var}(R_s)(\lambda N_F)^{-1} - 1 \right] \right\} - \frac{1}{2} \rho N_F \right] \\
\text{var}(D) &= \text{var}(U) + N_F^2 (b_2 - b^2) + \frac{N_F^2 \text{var}(\tilde{W}_\eta)}{12} + N_F^2 \frac{E(\tilde{R}_n^2 - 1)}{3(1-\rho)} + \left[ N_F \frac{\text{var}(\tilde{R}_n) - (1-\rho^2)}{2(1-\rho)} \right]^2,
\end{align*}
\]

where \( E(U) \) and \( \text{var}(U) \) denote the mean and the variance of \( U \), and \( \text{var}(W_\eta) \) is given by (3.32). In particular, when \( U \) is uniformly distributed over \( (0, N_F - 1) \), we have \( E(U) = \frac{1}{2} (N_F - 1), \text{var}(U) = (N_F^2 - 1)/12 \). For \( \rho > 1 \), the message delay becomes arbitrarily large.

These results can be used to compute the message-delay distribution and moments for TDMA schemes. We note that, in the special case where Poisson arrivals are assumed, the expression (3.66) for the mean message delay is identical to the expression derived in [22]. The expression for the variance of the message delay is given by (3.67) when relations (3.33)-(3.38) are incorporated. The moments of \( \tilde{R}_n \) are computed from (3.5). The same procedure is used if any other service ordering discipline is employed.

Note that, if a TDMA station is allocated a number of slots within each frame so that these slots are uniformly distributed over the frame (rather than assigned contiguously), the message-delay distribution formulas are given by those presented in Theorem 4 when the frame duration \( (N_F) \) is taken to be equal to the number of slots between any two service slots of this station. (See [24] for details.) These message-delay distribution (or moment) formulas can also serve as good simple approximations to the corresponding delay distributions attained when nonuniform frame-slot allocations are invoked.
IV. AN ASYNCHRONOUS-RESERVATION DEMAND-ASSIGNMENT (ARDA) SCHEME

An access-control scheme using reservations that is required to yield low message response-time values over a wide range of fluctuating traffic values must adapt its structure (or protocol) dynamically to the actual network traffic characteristics. We have presented such an adaptive scheme, DFRAC, in Section III. DFRAC uses state observations to estimate the network message traffic parameters and to change dynamically the structure of a fixed reservation scheme. In this section we present a family of reservation schemes called asynchronous-reservation demand-assignment (ARDA), the protocols of which automatically adapt to network traffic values. This is achieved, assuming as before a decentralized broadcast control procedure, by declaring a slot to be a reservation slot only when it becomes necessary, rather than in a fixed periodic fashion as in the FRAC scheme. Many different ARDA schemes can be developed, and the choice of a specific scheme will depend on various network characteristics and implementation constraints. The following scheme, denoted by ARDA I, will prove to be particularly efficient at low propagation-delay values, while a modification noted below will yield lower message delays for higher propagation-delay values. Furthermore, the following analysis of the ARDA I scheme will demonstrate the use of an analytical technique introduced for deriving the system delay-throughput performance curves. This technique will be used in future studies to analyze other dynamic access-control disciplines.

To simplify the analysis, we assume here that each reservation slot can be used, on a contention-free basis (or on a GRA basis with a very small probability of reservation packet retransmission), by all network terminals. We also assume that messages arrive according to a Poisson stream with intensity $\lambda$ [messages/slot]. We illustrate the computation of the limiting mean message waiting time $w$ (or delay $\mu = E(D)$). This procedure may also be applied to compute the limiting message-delay distribution under more general arrival streams and contention or non-contention reservation procedures.

In fact, in a typical operation of the ARDA I scheme to be described in this section, we will find that a reservation period of $R+1$ slots is generally established. We can therefore assume that an ARDA I scheme (later called gapless) establishes, when required, a reservation period consisting of $R+1$ slots. This period is available for the contention-free transmission of reservation packets by all network terminals which wish to make reservations in this period.

The Analytical Technique

The analytical technique is based on the following procedure. The state of the system is described by a discrete-time vector (of $K$ dimensions, $K \geq 1$) Markov chain $\{X_n, n \geq 1\}$, $X_n = (X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(K)})$ over the state space $S = X_{n+1}^{(1)}I_j$, where $I_j = \{0, 1, 2, \ldots, K\}$, is the set of nonnegative integers. The state $X_n$ describes the evolution of the process over its $n$th time period (defined appropriately, as induced by the access-control discipline). This Markov chain has a transition probability function $P(X_n, X_{n+1})$ and is irreducible and aperiodic. Furthermore, for the cases of interest, this Markov chain is also positive-recurrent. It thus has a stationary distribution $(u(i), i \in S)$ which is the unique (distribution) solution to the set of linear equations

$$u(i) = \sum_{j \in S} u(j)P(i,j), \quad j \in S.$$  (4.1)

Letting $W_n$ denote the waiting time of the $n$th message in the system, we wish to evaluate the steady state average waiting-time function

$$\bar{W} = \lim_{N \to \infty} N^{-1}E\left(\sum_{n=1}^{N} W_n\right).$$  (4.2)

when it exists.

The procedure uses the following ratio-limit theorem for Markov chains (see [1, p. 91, theorem 1]).

**Theorem 5:** Consider the irreducible aperiodic positive-recurrent Markov chain $\{X_n, n \geq 1\}$ with stationary distribution $(u(i))$. Then, for any two functions $f$ and $g$ from $S$ to $(-\infty, \infty)$ for which the two sums

$$\sum_{i \in S} u(i)f(i), \quad \sum_{i \in S} u(i)g(i),$$

converge absolutely and at least one is not zero, we have

$$\lim_{n \to \infty} \frac{\sum_{i=0}^{n} f(X_i)}{\sum_{i=0}^{n} g(X_i)} = \frac{Uf}{Ug},$$

with probability one. Also,

$$\lim_{n \to \infty} \frac{\sum_{i=0}^{n} E[ f(X_i) ]}{\sum_{i=0}^{n} E[ g(X_i) ]} = \frac{Uf}{Ug},$$

independent of the initial states assumed when taking the expectations in (4.4). \qed

Theorem 5 is now used to evaluate $\bar{W}$. We set $N(X_n, X_{n+1})$ and $W(X_n, X_{n+1})$ to be the mean number of messages served and the mean sum of the waiting times of these messages during the $(n+1)$st time period associated with the state $X_{n+1}$, given states $(X_n, X_{n+1})$. For our applications the latter two functions are time-homogeneous, for each $n > 0$. Also, as $M \to \infty$,

$$\sum_{n=0}^{N} N(X_n, X_{n+1}) \to \infty, \quad \text{with probability one.}$$  (4.5)

Hence

$$\lim_{N \to \infty} \frac{\sum_{n=1}^{N} W_n}{N} = \lim_{M \to \infty} \frac{M}{\sum_{n=0}^{N} W(X_n, X_{n+1})},$$  (4.6)
with probability one. To apply the ratio-limit Theorem 5 to (4.6), we consider the vector Markov chain \( \{ Y_n, n \geq 1 \} \), where \( Y_n = (X_n, X_{n+1}) \). This is an irreducible aperiodic positive-recurrent Markov chain with the stationary distribution \( \pi(i,j), i \in S, j \in S \) given by
\[
\pi(i,j) = u(i)P(i,j).
\]
(4.7)
We can therefore apply Theorem 5, (4.3), to (4.6), with the functions \( W(\cdot) \) and \( N(\cdot) \) replacing the functions \( f \) and \( g \), to conclude that
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i,j} W_n(i,j)u(i)P(i,j) = \frac{1}{N} \sum_{i,j} N(i,j)u(i)P(i,j),
\]
(4.8)
with probability one, provided that the two summations on the right side of (4.8) converge. For \( \lambda > 0 \), the summation in the denominator of (4.8) is positive. Finally, when (4.8) holds, it follows by the ergodicity or asymptotic ergodicity of \( \{ W_n, n \geq 1 \} \) (see [17] and the following waiting time analysis) that
\[
\frac{1}{N} \sum_{i,j} E(W_n) = \lim_{n \to \infty} \frac{1}{N} \sum_{i,j} W_n(i,j)u(i)P(i,j) = \lim_{n \to \infty} \frac{1}{N} \sum_{i,j} N(i,j)u(i)P(i,j),
\]
(4.9)
with probability one. Hence (4.8) and (4.9) yield a useful procedure for the evaluation of \( \bar{W} \). This is summarized in the following theorem.

**Theorem 6:** For the access-control discipline under consideration, assume the system state to be described by an irreducible aperiodic positive-recurrent Markov chain \( \{ X_n, n > 0 \} \) with transition probability function \( P(i,j) \) and stationary distribution \( \{ u(i) \} \). Let \( N(X_n, X_{n+1}) \) and \( W(X_n, X_{n+1}) \) denote the mean number of messages served and the mean sum of the waiting times of these messages during the \( (n+1) \)st time period associated with state \( X_{n+1} \), given the states \( (X_n, X_{n+1}) \). Then the steady-state average waiting time \( \bar{W} \), defined by (4.2), is given by (4.9) and is equal to
\[
\bar{W} = \frac{E[W(X_n, X_{n+1})]}{E[N(X_n, X_{n+1})]},
\]
(4.10)
where
\[
E[W(X_n, X_{n+1})] = \sum_{i,j} W(i,j)u(i)P(i,j)
\]
(4.11a)
\[
E[N(X_n, X_{n+1})] = \sum_{i,j} N(i,j)u(i)P(i,j),
\]
(4.11b)
provided that the summations in (4.11a) and (4.11b) converge and at least one of them is positive.

We note that, to evaluate \( \bar{W} \) by (4.10) and (4.11), we must obtain the functions \( W(\cdot) \) and \( N(\cdot) \) and derive the stationary distribution \( \{ u(i) \} \). However, we will observe that for the access-control discipline under consideration, as well as many others, the stationary probabilities of only a limited number of states are required. Note also that the same procedure is readily extended to evaluate any limiting moment of the message waiting time and to general discrete-time or continuous-time Markov state processes.

**An ARDA Scheme and its Analysis**

The asynchronous-reservation demand-assignment scheme (ARDA I) presented in this section assumes the following protocol.

**Protocol (ARDA I Scheme):** At any time, the first slot not allocated to serve any packet is declared to be a reservation slot. Messages are served, following a reservation, according to the system service discipline.

The ARDA I scheme thus operates in the following manner. When new messages arrive at a terminal, the terminal waits for the first slot that it knows to have been declared a (noncontention) reservation slot and sends a reservation packet (for all its presently unreserved messages) within this slot. Its reservation packet is received, through the broadcast channel, by all the network terminals following a round trip propagation delay of \( R \) slots. At that time the reserving terminal is assigned service slots during which it will transmit its corresponding messages, according to the network agreed-upon service discipline.

A specific or random ordering can be used for messages sending reservations within the same reservation slot, while first-come first-served ordering is used for serving messages reserving in different reservation slots. (See the previous discussion concerning service orderings.) Each terminal, or a central controller when centralized control is implemented, records in its own queueing table the channel reservations that have been made. Consequently, when a terminal wishes at time \( t \) to send a reservation packet, it can identify the first slot following time \( t \) that has not been allocated to serve a message. This slot is subsequently declared to be a reservation slot.

In Fig. 4 sample functions for the processes describing the evolution of channel reservation and service periods are shown for \( R = 0, 1, 2 \). Assuming single packet messages, the processes describing the number of reservations made in reservation slots are denoted by \( \{ N_n^0 \} \), and the service period durations are denoted by \( \{ X_n \} \). We note that for \( R = 0, 1 \), the channel state process is composed of an alternating sequence of \( (R+1) \) reservation slots followed by a service period. For \( R \geq 2 \), the corresponding processes behave similarly, except that this pattern can be broken occasionally by a gap in the service period (see Fig. 4(d) for example). This happens mainly for low traffic values and short messages, when the services required by previous reservations occupy a number of slots that extend less than \( R \) slots following the present reservation slot. We will observe that this particular phenomenon is insignificant in determining the basic evolution of the underlying state sequence. Hence we will assume, for each \( R \geq 2 \), a gapless channel process for the purpose of deriving the transition probability functions of the underlying state sequence. In obtaining the average
message delay for low traffic values, we will however derive an expression that will incorporate the above phenomenon. Note that the analysis remains precise for \( R = 0 \) or \( R = 1 \), while for \( R > 2 \) the delay-throughput expressions derived under the above gapless assumption are in virtually complete agreement with simulation results.

Having a channel process that is composed of alternate reservation and service periods, the state sequence is defined as follows. We call a channel period, or simply a period, the time interval composed of \( R + 1 \) reservation slots followed by the service period allocated to serve these reservations, provided that at least a single reservation is made in the first reservation slot. Otherwise, the first slot under consideration is an empty slot which we regard as constituting the whole corresponding channel period. We let \( X_n \) denote the state vector associated with the \( n \)th channel period, so that \( \{X_n, n \geq 1\} \) is the underlying state sequence. We let \( N_n^{(i)} \) denote the number of messages making reservations during the \( i \)th reservation slot of the \( n \)th channel period, \( i = 1, 2, \ldots, R + 1 \) (setting \( N_n^{(i)} = 0 \) for \( j > 2 \), when \( N_n^{(i)} = 0 \)). The number of messages served during the \( n \)th period is denoted by \( Y_n \), while the message lengths of the messages reserving and served during this period are denoted by \( \{B_k(n), 1 \leq k < Y_n\} \) for \( Y_n > 1 \). The service duration of the \( n \)th period, denoted by \( Y_n \) (and measured in slots), is given by

\[
Y_n = \sum_{i=1}^{Y_n} B_k(n). \tag{4.12}
\]

provided that \( Y_n > 1 \), while \( Y_n = 0 \) when \( X_n = 0 \). The \( n \)th state vector \( X_n \) is thus defined as

\[
X_n = \{N_n^{(1)}, N_n^{(2)}, \ldots, N_n^{(R+1)}, Y_n\}, \quad n \geq 1 \tag{4.13}
\]

with \( X_n = 0 \) when \( N_n^{(i)} = 0 \). We note that \( X_n \) is obtained from \( X_n \) by

\[
X_n = I(N_n^{(1)} > 0) \cdot \sum_{i=1}^{R+1} N_n^{(i)}, \quad n \geq 1 \tag{4.14}
\]

where \( I(A) \) is the indicator function associated with event \( A \), so that

\[
I(A) = \begin{cases} 1, & \text{if } A \text{ holds} \\ 0, & \text{if } A \text{ does not hold.} \end{cases} \tag{4.15}
\]

The state process \( \{X_n, n \geq 1\} \) is an \((R+2)\)-dimensional Markov chain, with a transition probability function \( P(i,j) \) readily obtained from (4.12)-(4.14) (see [27]). It is noted that \( \{X_n, n \geq 0\} \) is an irreducible aperiodic Markov chain and that \( X_{n+1} \) depends on \( X_n \) only through \( Y_n \). Sequences \( \{X_n, n \geq 0\} \) describing the number of messages sent and \( \{Y_n, n \geq 0\} \) describing the number of packets served during the consecutive service periods are also Markov chains. The transition probability function for \( \{Y_n, n \geq 0\} \) denoted by \( \{P_y(i,j)\} \) is given by

\[
P_y(i,j) = \begin{cases} 1, & \text{if } A \text{ holds} \\ 0, & \text{if } A \text{ does not hold.} \end{cases} \tag{4.16}
\]

for \( i > 0, j > 1 \), while for \( i > 0, j = 0 \) we have

\[
P_y(i,0) = \exp \{ -\lambda(i+1) \}. \tag{4.16b}
\]

The transition probability function for \( \{X_n, n \geq 0\} \) is written similarly. In particular, we note that \( X_{n+1} \) is related to \( X_n \) by the following recurrence relationship:

\[
X_{n+1} = I(N_n^{(1)} > 0) \cdot \left[ N_n^{(1)} + \sum_{i=2}^{R+1} N_n^{(i)} \right], \tag{4.17a}
\]

where for \( k > 0, i > 0, \)

\[
P \{ N_n^{(i)} = k \mid X_n = i \} = \sum_{j=i}^{\infty} e^{-\lambda(j+1)} \frac{[\lambda(j+1)]^k}{k!} \beta^{(i,k)}, \tag{4.17b}
\]

with \( \beta^{(0)} = \delta_0 \) and, for \( i = 2, 3, \ldots, R+1, k > 0, i > 0, \)

\[
P \{ N_n^{(i)} = k \mid X_n = i \} = e^{-\lambda \frac{k}{k}}. \tag{4.17c}
\]

It is further observed that the Markov chains \( \{Y_n, n \geq 0\} \) and \( \{X_n, n \geq 0\} \) are irreducible aperiodic and that one is positive-recurrent if and only if the other is, and if and only if \( \{X_n, n \geq 0\} \) is positive-recurrent. In the latter case, the stationary distribution \( \{u(j)\} \) of \( \{X_n, n \geq 0\} \) is obtained from (4.12)-(4.17),
and \( \{ u_i, i \geq 0 \} \) denotes the stationary distribution of \( \{ Y_n, n \geq 0 \} \),
\[
u_i = \lim_{n \to \infty} P \{ Y_n = i \}. \quad (4.19)
\]
Thus, to obtain the stationary distribution of \( \{ X_n, n \geq 0 \} \), it is necessary only to derive the distribution \( \{ u_i, i \geq 0 \} \) or its \( z \)-transform
\[
U^*(z) = \sum_{i=0}^{\infty} u_i z^i, \quad |z| < 1. \quad (4.20)
\]
The latter distribution is obtained by solving the set of linear equations
\[
u_j = \sum_{i=0}^{\infty} u_i P_j(i,j), \quad j > 0, \quad (4.21)
\]
with \( P_j(i,j) \) given by (4.16). In particular, we will find that only
\[
u_0 = \lim_{n \to \infty} P \{ Y_n = 0 \} = \lim_{n \to \infty} P \{ X_n = 0 \}
\]
is needed for computing the message limiting average waiting time. The condition for the positive-recurrence of the underlying Markov chains, and expressions for the transform \( U^*(z) \) and \( u_0 \) are presented in the following lemma. These results are derived by solving (4.21). (For proofs the reader is referred to [12] or [27].)

**Lemma 1:** The Markov chains \( \{ X_n, n \geq 0 \} \), \( \{ Y_n, n \geq 0 \} \), and \( \{ X_n, n \geq 0 \} \) are positive-recurrent if and only if the network traffic intensity \( \rho \) satisfies
\[
\rho < 1. \quad (4.22)
\]
For \( \rho < 1 \), the steady-state distribution \( \{ u_i, i \geq 0 \} \) exists as the unique distribution satisfying (4.21), and its transform \( U^*(z) \) is given by
\[
U^*(z) = \nu_0 B(z) + A(z), \quad \text{for } |z| < 1, \quad (4.23a)
\]
where
\[
A(z) = \prod_{i=1}^{R+1} (gh_i)^{i} \quad (4.23b)
\]
\[
B(z) = 1 - g R + \sum_{n=0}^{\infty} g^{R+1} (gh_i)^{R+1} (gh_j)^{R+1}
\]
\[
\cdots \cdot (gh_n)^{R+1} [1 - (gh_{n+1})^R], \quad (4.23c)
\]
and
\[
g = g(z) = \exp( - \lambda (1 - z))
\]
\[
h = h(z) = \beta( g(z))
\]
\[
h_{i+1} = h_{i+1}(z) = h_i(z), \quad i > 0, h_0(z) = z
\]
\[
gh_i = g(h_i(z)), \quad i > 0. \quad (4.23d)
\]
Furthermore, we have
\[
u_0 = A(0) [1 - B(0)]^{-1} \quad (4.24)
\]
where \( A(0) \) and \( B(0) \) are given by (4.24b) and (4.24c), respectively, with \( g = g(0) = \exp( - \lambda) \) and \( h = h(0) = \beta(\exp( - \lambda)) \).

Theorem 6 is now used to obtain the average message waiting time \( \bar{W} \). Assuming a gapless channel process, the function \( W_1(X_n, X_{n+1}) \), denoting the sum of waiting times of messages served during the \( (n+1) \)st channel period, is given by
\[
W_1(X_n, X_{n+1}) = I( N_{n+1}^{(1)} > 0) \left( \frac{1}{2} N_{n+1}^{(1)} Y_n + X_{n+1} + \right. \right.
\]
\[
\left. \left[ R N_{n+1}^{(1)} + (1 + R) N_{n+1}^{(2)} + \cdots + N_{n+1}^{(R)} \right]
\]
\[
+ \left. \left[ B^{(n+1)}(X_{n+1} - 1) + B^{(n+1)}(X_{n+1} - 2) + \cdots + B^{(n+1)}(X_{n+1} - 1) \right] \right) \quad (4.25)
\]
where \( B^{(n+1)} \) denotes the length of the \( i \)th message served during the \( (n+1) \)st period. In (4.25) we assume that \( X_{n+1} \) has been extended to include the message lengths \( B^{(n+1)} \) with no change in the previous conclusions since they are a sequence of i.i.d. random variables. Expression (4.25) contains the following message waiting-time components. If \( N_{n+1}^{(1)} = 0 \), no messages arrive during the \( (n+1) \)st channel period and therefore \( W(\cdot) = 0 \). Otherwise, \( N_{n+1}^{(1)} \) messages make reservations during the first slot of the latter period. These messages have arrived at their terminals during the previous service period (which is \( Y_n \) slots long), and have thus experienced (with respect to their Poisson arrivals) a total average waiting time of \( \frac{1}{2} N_{n+1}^{(1)} Y_n \) slots, yielding the first term on the right side of (4.25). The second term accounts for the delay of a reservation slot, the third term incorporates the delay experienced by a message waiting solely for the service of the other messages which have made reservations in the same reservation slot used by this message.
The number of messages \( N(X_n, X_{n+1}) \) served during the 
\((n+1)\)st channel period is given in both cases by
\[
N(X_n, X_{n+1}) = X_{n+1}. \quad (4.27)
\]

From (4.10) and (4.11) we obtain the average message 
waiting-time functions \( \bar{W}_1 \) and \( \bar{W}_2 \) by using respectively 
(4.25) and (4.26). Clearly, \( \bar{W}_1 > \bar{W}_2 \) for medium and high 
throughput values, when \( \bar{W} \approx \bar{W}_1 \), while \( \bar{W}_1 < \bar{W}_2 \) for low 
throughput values, when \( \bar{W} \approx \bar{W}_2 \). Therefore, the average 
message waiting-time estimate
\[
\bar{W} = \max(\bar{W}_1, \bar{W}_2) \quad (4.28)
\]
will yield a close approximation to \( \bar{W} \) over the whole 
range of traffic intensities. In fact, simulation results to be 
presented later show \( \bar{W} \) to yield delay values identical to 
those obtained by simulation over the whole throughput 
range, for single packet messages. (For multi-packet 
messages, \( \bar{W}_1 \) is clearly an excellent approximation.) In 
particular, we note that for \( R = 0 \) and \( R = 1 \) no service 
gaps can arise, and \( \bar{W}_1 \) yields the exact delay-throughput 
curves for these important cases. Using Theorem 3, 
Lemma 1, (4.10), (4.11), (4.25), (4.26) and the characteriza-
tions of the underlying Markov chains, we obtain the 
system delay-throughput curves. The reader is referred to 
[12] or [27] for detailed derivations and results. In those 
papers it is noted that, for \( \rho = \lambda b < 1 \), the mean waiting-
time expressions \( (\bar{W}_1 \text{ and } \bar{W}_2) \) depend only on the parame-
ters \( u_0, \lambda, b, b_2 \), and \( R \). (For \( \rho > 1 \), arbitrarily high 
limiting delays result.)

In particular, for single-packet messages, when \( \rho = \lambda < 1 \), \( R > 0 \), the following delay-throughput formulas are obtained:
\[
D_1(\lambda) = (1-\lambda)^{-1}[1.5(1+R) - 0.5\lambda R] + \frac{1}{2} + \frac{1+R}{1+R(1-u_0)}, \quad (4.29)
\]
\[
D_2(\lambda) = 2R + 2 \cdot \frac{1.5\lambda}{1-\lambda} + \frac{\lambda}{2} \cdot \frac{1}{1-\lambda} + \frac{2R}{1-R} \cdot \frac{(R-1)}{(1-u_0)(R^2 + \lambda^2)} \cdot \frac{(R+1)}{R(1-u_0)}, \quad (4.30)
\]
where \( u_0 \) is given by (4.24) with \( b(z) = z \). For single-packet 
messages, with \( \rho = \lambda < 1 \), since \( 0 < u_i < 1 \), we obtain from 
(4.29) the bounds
\[
\tilde{D}(\lambda) < D_1(\lambda) < \tilde{D}(\lambda) + \frac{1}{2} R, \quad R > 0, \quad (4.31a)
\]
by setting \( u_0 = 0,1 \), where
\[
\tilde{D}(\lambda) = (1-\lambda)^{-1}[2+1.5R - 0.5\lambda(1+R)]. \quad (4.31b)
\]
The lower and upper bounds in (4.31a) differ by only 
0.5R slots and are closely followed at high and low traffic 
intensity values, respectively. Furthermore, we note that 
\( \tilde{D}(\lambda) \) is a simple expression which does not require the 
computation of \( u_0 \). For \( R = 0 \), we obtain \( D(\lambda) = D_1(\lambda) = D_2(\lambda) = \tilde{D}(\lambda) \), while for \( R = 1 \), \( D(\lambda) = D_2(\lambda) \) is within 0.5 
slot of \( \tilde{D}(\lambda) \). Similar observations can be made for \( D_2(\lambda) \) 
and for multi-packet messages to yield simple approxi-
mate formulas (or close bounds) for evaluating the system 
delay-throughput curves. We also note that, as \( \lambda \to 0 \),
\[
\lim_{\lambda \to 0} D(\lambda) = \lim_{\lambda \to 0} D_1(\lambda) = \lim_{\lambda \to 0} D_2(\lambda) = b + 1 + 2R. \quad (4.32)
\]
Relation (4.32) indicates that as \( \lambda \to 0 \) an arriving message 
experiences no waiting time but an average time delay 
consisting of \( b \) slots for message transmission time, one 
slot for reservation and \( 2R \) slots propagation delay while 
broadcasting the reservation packet and the message itself.

**Performance Evaluation**

Fig. 5 shows delay-throughput curves for \( R = 0 \) and 
\( R = 1 \) (for applications to ground radio or line networks, 
as well as to satellite channels with long slots) as given by 
\( D(\lambda) \) of (4.29), which yields the exact performance curves. 
In this figure, we have also drawn the corresponding 
curves for \( R = 0,1 \) under a DFRAC scheme \( (D^*(\lambda)) \). We 
note that, for both \( R = 0 \) and \( R = 1 \), access-control scheme 
ARDA I yields delay-throughput curves uniformly lower 
over the whole throughput range than those obtained by 
the corresponding DFRAC scheme. Thus while the latter 
scheme uses estimates of the underlying traffic intensity 
values to adjust a FRAC scheme optimally, an ARDA I 
scheme requires reservation slots to be assigned dynami-
cally and to adapt automatically to traffic fluctuations 
and demonstrates better performance over the whole 
throughput range for \( R = 0,1 \). Recall that this comparison 
applies to the case where all network terminals can 
use a single contention-free reservation slot. (The gapless 
ARDA scheme in effect uses a reservation period of \( R + 1 \) 
slots.)

For a higher value of round-trip propagation delay, 
\( R = 12 \) slots, and single-packet messages, we compare the 
formulas \( D_1(\lambda), D_2(\lambda) \), with the delay-throughput curve 
\( \bar{D}(\lambda) \) obtained by simulation of a channel under an 
ARDA I scheme (Fig. 6). We note that, as indicated 
above, \( D(\lambda) = D_1(\lambda) \) and \( D(\lambda) > D_2(\lambda) \) for higher through-
put (\( \lambda \)) values, while \( D(\lambda) > D_2(\lambda) \), \( D(\lambda) = D_2(\lambda) \) 
for lower traffic intensity values. Furthermore, \( D(\lambda) = 
\max(D_1(\lambda), D_2(\lambda)) \) is observed to yield \( D(\lambda) \) to great 
precision. Furthermore, \( D_1(\lambda) \) is also a very close approxima-
tion to \( D(\lambda) \) over the whole throughput region.

In Fig. 7 we compare the throughput-delay perfor-
mance curves of a DFRAC scheme and an ARDA I 
scheme for \( R = 12 \) and single-packet messages. These two 
schemes yield close average message-delay values for the 
throughput region \( 0 < \lambda < 0.5 \). However, for higher 
throughput values the ARDA I scheme is not as efficient, 
yielding much higher message-delay values. The reasons 
for the degraded performance of an ARDA I scheme for 
high propagation-delay values are explained in the next 
section. There we also modify the ARDA I scheme to 
obtain an ARDA II scheme which is as efficient as a 
DFRAC scheme for high values of \( R \), over the whole 
throughput range.
An ARDA II Access-Control Scheme

The inefficient operation of an ARDA I scheme for high values of propagation delay and throughput stems mainly from the following two points, both associated with the reservation process. First, new messages arriving at the network have to wait to the end of the present (long) service periods, then transmit a reservation packet and subsequently wait another $R$ slots (for the reservation packet to be broadcast to the network users) before a channel service period can start. Second, for high traffic intensity values, the long reservation period (of $R+1$ slots) preceding any service period constitutes an inefficient use of channel time. This is because the average number of reservations made during the first reservation slot in a reservation period is large (since it contains the reservations made by all messages arriving during the preceding long service period), while during the following $R$ reservation slots only a small number of reservations are made (since in each such reservation slot only messages arriving during the preceding reservation slot will make their reservations).

Assuming that all network terminals can make their reservations within a single contention-free slot, we correct the above-mentioned inefficiencies by modifying the ARDA I scheme at high values of $R$ to obtain the scheme ARDA II which has the following protocol. Assume henceforth that $R > 1$.

Protocol (ARDA II): $R$ slots after a reservation slot, or after the last in a group of reservations slots, the remaining service time in the present channel service period is observed. Depending on the duration of this remaining service period, one of the following two methods is used to establish the location of the next reservation slot.

1) If the above remaining service period is not longer than $R$ slots, the group of $R+1$ slots following the present service period is declared to be a group of reservation slots, as for an ARDA I scheme.

2) If the remaining service period is longer than $R$ slots, the slot which precedes by $R$ slots the end of this service period is established as a reservation slot. Arriving messages then make reservations during the first reservation slot established following their arrival and, $R$ slots later, are assigned service slots (following an appropriate service discipline) and served.

The channel process under scheme ARDA II behaves as follows. When the reservations made in a reservation slot, or the last reservation slot in a reservation group, are received by all the network users ($R$ slots following this reservation slot), the remaining service period is observed. If the latter is not longer than $R$, the group of $R+1$ slots following the present service period is declared to be a group of reservation slots, as for an ARDA I scheme. On the other hand, if the remaining service time is longer than $R$, as might be the case under high throughput conditions, a reservation slot is established $R$ slots prior to the end of this service period. Messages which made reservations during the latter service period can be served within the service period that starts immediately following it. In this manner both causes of the inefficiency of scheme ARDA I for high $R$ and throughput values are corrected, since a single reservation slot (rather than a group of $R+1$ reservation slots) is set up $K$ slots prior to the end of the present service period, so that the additional delay of $R$
slots experienced by a message in an ARDA I scheme following the present service period is saved. Furthermore, the observation of the remaining service period is made at the first possible slot following a reservation slot (being \( R \) slots following the latter), while the establishment of the next reservation slot, provided the latter period is longer than \( R \), is at the latest slot enabling succession of service periods, thus causing minimal delay for messages served within the present service period.

The waiting-time analysis for the ARDA II scheme is similar to that of the ARDA I scheme, using the technique presented by Theorem 6, and is therefore not presented here in detail. (The reader is referred to [12] for details.)

For single-packet messages, \( R = 12 \), the delay-throughput curve for an ARDA II scheme, is shown in Fig. 7. For low traffic intensity values (around \( 0 < \lambda < 0.5 \)), ARDA I and II yield virtually identical delay-throughput values, which are also very close to those obtained under a DFRAC scheme. For higher traffic intensity values, the ARDA II scheme demonstrates a significant improvement in performance over the ARDA I scheme. The average message-delay values obtained under an ARDA II scheme are only slightly greater than those obtained under a DRAC scheme in the throughput region \( 0.5 < \lambda < 0.9 \) and are smaller than the DFRAC values when \( \lambda > 0.9 \).

In comparing the delay-throughput performance of an ARDA II scheme with that of a DFRAC scheme, one should note however that the latter requires a perfect estimate of the underlying network traffic intensity value. Such an estimate is readily acquired through a long enough sequence of observations of the channel process, when long-term traffic fluctuations are considered. However, when short-term traffic fluctuations need to be incorporated, the ARDA II scheme is more appropriate since its structure is modified dynamically, based upon the most recent size of the service period, and thus the most recent magnitude of the network traffic intensity.

V. CONCLUSION

We have studied reservation and TDMA schemes for the message-switching access-control of a multi-access broadcast channel. A single repeater is used to yield a fully connected broadcast network structure. An arbitrary propagation-delay value is associated with the channel so that the results apply to channels with low propagation-delay values such as terrestrial radio-relay or line networks, as well as to channels characterized by long propagation delays, such as satellite channels. Network users have been assumed to be synchronized. Channel slots for packet transmissions are then defined.

Message-delay distribution and moments are derived when considering a multi-access channel which employs a fixed-reservation access-control procedure. Under such a scheme, time frames are identified and each time frame is divided into two periods: the reservation period and the service period. The reservation periods are used by network terminals to transmit reservation packets. They can also be partially used by certain network terminals for message information transmission, governed by any proper access-control discipline. The service periods are used by the network terminals under consideration to transmit their reserved messages.

We have computed the limiting message-service delay distribution under the FRAC scheme, assuming the number of reservations made in each reservation period to be described as an i.i.d. sequence of random variables governed by arbitrary distribution. The message reservation delay, describing the time elapsed between message arrival and the proper transmission of its reservation packet, is then added to the message service delay.

The FRAC scheme assumes a TDMA structure in that service periods are temporally distributed. As a special case we obtain expressions for the limiting distribution and the moments of the delay of messages which access the channel on a TDMA basis. A TDMA network station is assigned a set of slots within each time frame. This station thus requires no reservations. It assigns its messages locally to its dedicated slots. The delay distributions of these messages are derived. The underlying message-arrival process is characterized statistically by describing the number of messages arriving within each time frame as a sequence of i.i.d. random variables. (A Poisson arrival process thus becomes a special case.)

We have presented a technique for the analytical and joint analytical-simulation computation of the limiting message-delay distribution and moments for a class of dynamic access control schemes, properly described by a Markov state process. We have illustrated the use of this technique by deriving the mean delay-throughput functions for a simple asynchronous-reservation demand-assignment (ARDA I) scheme. This scheme adapts its structure dynamically in accordance with the underlying message traffic fluctuations. The scheme is modified (yielding the ARDA II procedure) to give better performance characteristics at higher propagation delay values.

The techniques and results presented serve as basic tools for message-delay computations for access-control schemes which involve TDMA and reservation procedures, as well as the integration of these schemes with other access control procedures. For example, the TDMA results have been used in [24] to study and compare FDMA and TDMA schemes. The analytical technique used to study the ARDA schemes, which employs a Markov ratio limit theorem, can be applied to the analysis of various adaptive reservation schemes, such as those presented in [9] and [10]. It has been used in [19] to analyze group random-access schemes and, in [25], to study integrated random-access/reservation access control techniques. Under such an integrated scheme, newly arriving packets are first transmitted on a random-access or reservation basis. A packet that is transmitted on a random-access basis and subsequently collides with other packets is retransmitted by employing a reservation procedure, rather than in a random-access fashion. Such schemes are shown in [25] to yield good delay-throughput
characteristics over the entire range of throughput values when short propagation delays are involved. They also perform well under high propagation delay values for low-to-medium throughput values. For higher throughputs the reservation schemes presented in this paper exhibit superior delay-throughput characteristics.

Appendix A

Proof of Theorem 1

We assume that \( p < 1 \). To evaluate the steady-state distribution of the queue size \( V_n \) in the GI/G/1 queueing system, described by (3.42), we proceed as follows. For this queueing system, let \( X_n \) denote the queue size prior to the \( n \)th reservation period. The Markov chain \( \{ X_n, n \geq 1 \} \) is governed by the recurrence

\[
X_{n+1} = [X_n + \bar{R} - N_S]^+, \quad n \geq 1.
\]  

(A-1)

Equation (A-1) follows by noting that the number of messages queueing in the system prior to the \((n+1)\)st reservation period is determined by the corresponding number prior to the \(n\)th reservation slot, the number of reservations made in the \(n\)th reservation slot \( (\bar{R}_n) \), and the number of service slots \( (N_S) \) between any two reservation periods. Then \( \{ V_n, n \geq 1 \} \) is obtained by sampling \( \{ X_n, n \geq 1 \} \) at those reservation slots which correspond to reservation group arrivals. However, the latter are described by a discrete-time renewal point process with geometric inter-arrival times following distribution (3.6), and this constitutes a memorylessness of a random sequence of events. Therefore, it is easily shown (and is also well known from queueing theory, see [2]) that \( V_n \) and \( X_n \) have identical steady state distributions. In particular,

\[
V\ast(z) = X\ast(z),
\]  

(A-2)

where \( X\ast(z) \) is the \( z \)-transform of the limiting distribution of \( X_n \) when the latter exists.

The recurrence (A-1) is identical to that describing \( X_n \) as the \( n \)th message waiting time, or as the embedded (packet) queue size at time \( r_n \), if \( n \geq 1 \), in a single server queueing system \( D/G/1 \) with constant inter-arrival times (equal to \( N_S \)) between message groups (requiring service duration) of size \( \bar{R}_n \) slots. Note that the latter may require zero service, corresponding to no arrivals within the recent group of service slots. For the analysis of this system, we define the queue-size sequence \( \{ Z_n, n \geq 1 \} \), where \( Z_n \) denotes the queue size (i.e., number of packets in the system) at time \( r_n \), for system (A-1). Since at time \( r_n \) a group of size \( \bar{R}_n \) packets (or slots) arrives, we have

\[
Z_n = X_n + \bar{R}_n.
\]  

(A-3)

Substituting (A-3) into (A-1), we find

\[
Z_{n+1} = [Z_n - N_S]^+ + \bar{R}_{n+1}, \quad n \geq 1.
\]  

(A-4)

We use (A-4) to derive the steady-state distribution of \( Z_n \) from which the corresponding functions for \( X_n \) and \( V_n \) follow by (A-3). In particular,

\[
V\ast(z) = Z\ast(z)[\bar{R}\ast(z)]^{-1},
\]  

(A-5)

where \( Z\ast(z) \) is the \( z \)-transform of the limiting distribution of \( Z_n \).

Using (A-4), or directly from (A-1), we obtain the limiting transform \( Z\ast(z) \) by solving the associated steady-state equations for the underlying Markov chains. We find that (see, for example, [3, p. 160] or [20]), if

\[
\rho = (N_S)^{-1} E(\bar{R}) = b\bar{R}N_S^{-1} < 1,
\]  

(A-6)

then \( V_n \) has a limiting distribution with an associated \( z \)-transform \( V\ast(z) \) given by

\[
V\ast(z) = \frac{N_S(1 - \rho)(1 - z)}{R\ast(z) - zN_S} \sum_{r=1}^{n-1} \frac{z - r}{1 - \eta_r},
\]  

(A-7)

where \( \eta_r, r = 1, 2, \ldots, N_S - 1 \), are the distinct \( N_S - 1 \) roots with \( |\eta_r| < 1 \) obtained by setting

\[
\eta_r = \lim_{\omega \to 1} \eta_r(\omega),
\]  

(A-8)

where \( \eta_r(\omega), r = 1, 2, \ldots, N_S \) are the \( N_S \) distinct roots of

\[
z^{N_S} = \omega R\ast(z), \quad |z| < 1, \quad |\omega| < 1.
\]  

(A-9)

Only one of the roots of (A-9) tends to unity as \( \omega \to 1 \). For \( N_S = 1 \), (A-7) holds if we set the product term equal to one. The steady-state moments of \( V_n \) are obtained by differentiating (A-7) and letting \( z \to 1 \). We obtain for \( p < 1 \)

\[
E(V_\omega) = \frac{1}{N_S(1 - \rho)} - \frac{1}{2} E(\bar{R}) + \frac{1}{2} \sum_{r=1}^{N_S - 1} \frac{1 + \eta_r}{1 - \eta_r} + \frac{1}{2} \sum_{r=1}^{N_S - 1} \frac{\eta_r}{(1 - \eta_r)^2}.
\]  

(A-10)

The summations in (A-10) and (A-11) are set equal to zero when \( N_S = 1 \) (see also [4, p. 185]). The roots \( \{ \eta_r(\omega) \} \) can be obtained by solving the set of equations

\[
z = \omega[\bar{R}\ast(z)]^{1/N_S} \exp\left\{ -2\pi i(k - 1)N_S^{-1} \right\}, \quad k = 1, 2, \ldots, N_S
\]  

(A-12)

Each equation in (A-12) yields a distinct root, so that (A-12) results in the desired \( N_S \) roots \( \{ \eta_r(\omega) \} \). For example, assume the reservations arrive according to a Poisson stream with intensity \( \lambda [\text{reservations}/\text{slot}] \), so that (3.33) and (3.34) hold, the message lengths are geometrically distributed with parameter \( q \), so that

\[
&z^{-1} = \exp \left[ \lambda N_S z^{-1} \right].
\]  

(A-13)

Then, using a power series expansion, we obtain

\[
\eta_r(\omega) = \sum_{n=1}^{\infty} \frac{\omega^{nN_S}}{n} \left( \sum_{j=1}^{n-1} \frac{q^{j-1} - 1}{j!} \left[ m(1 - q)N_S^{-1} \right]^j \right), \quad k = 1, 2, \ldots, N_S
\]  

(A-14)

where

\[
\phi_k = \lambda N_S^{-1} + 2\pi i(k - 1)N_S^{-1}.
\]  

(A-15)

Considering the \( D/G/1 \) queueing system governed by the waiting-time process of (A-1), the following expressions for the limiting transform \( V\ast(z) \) and the mean \( E(V_\omega) \) hold (see for example [2, pp. 280–282], and [14]):

\[
V\ast(z) = \exp \left\{ - \sum_{n=1}^{\infty} \frac{1}{n} \left[ 1 - \sum_{j=0}^{\infty} z^j P(C_n = j) \right] \right\},
\]  

(A-16)

\[
E(V_\omega) = \frac{1}{n} \sum_{j=0}^{\infty} j P(C_n = j).
\]  

(A-17)
where $C_n$ is a random variable following a distribution which is equal to the $n$th convolution of the distribution of $\tilde{R}_n - N$. Hence

$$P(C_n = j) = \rho^{(n)}j \cdot \sum_k \rho^{(n)}k \cdot \frac{1}{N^k}, \quad j = 0, 1, \ldots \quad (A-18)$$

Substituting (A-18) into (A-16) and (A-17), we obtain (3.50) and (3.51). An upper bound on $E(V_\infty)$, denoted by $\bar{V}_n$, is obtained by using the result derived in [15] for a GI/G/I queueing system. Considering system (A-1), we have

$$E(V_\infty) \leq \bar{V}_n = \frac{\text{var}(\tilde{R}_n)}{2N^2(1-\rho)}, \quad (A-19)$$

for $\rho < 1$. Hence

$$\bar{V}_n = \frac{\rho}{2(1-\rho)} \left[ b_2 b^{-1} + b(\text{var}(\tilde{R}_n)[E(R_n)]^{-1} - 1) \right], \quad (A-20)$$

yielding (3.52) and (3.53). The lower bounds in (3.53) are obtained by dropping the last term in (3.51) and noting that $0 < \rho < 1$.

REFERENCES