DISCRETE TIME QUEUES WITH DELAYED INFORMATION

Eitan ALTMAN
INRIA, B.P. 93
2004 Route des Lucioles
06902 Sophia-Antipolis, France

Daniel KOFMAN
Département Réseaux
E.N.S.T.
75634 Paris Cedex 13 France

Uri YECHIALI
Department of Statistics and Operations Research
Tel-Aviv University
Tel-Aviv 69978, Israel

December 1994

Abstract

We study the behavior of a single-server discrete-time queue with batch arrivals, where the information on the queue length and possibly on service completions is delayed. Such a model describes situations arrising in high speed telecommunication systems, where information arrives in messages, each comprised of a variable number of fixed-length packets, and it takes one unit of time (a slot) to transmit a packet. Since it is not desirable to attempt service when the system may be empty, we study a model where we assume that service is attempted only if, given the information available to the server, it is certain that there are messages in the queue. We characterize the probability distribution of the number of messages in the queue under some general stationarity assumptions on the arrival process, when information on the queue size is delayed K slots, and derive explicit expressions of the PGF of the queue length for the case of i.i.d. batch arrivals and general independent service times. We further derive the PGF of the queue size when information on <u>both</u> the queue length and service completion is delayed $K = 1$ units of time. Finally, we extend the results to priority queues and show that when all messages are of unit length, the $c\mu$ rule remains optimal even in the case of delayed information.

Keywords: Discrete time queues, delayed information.

1 Introduction

Queues with delayed information are quite useful for modeling high-speed telecommunication networks, where propagation delay of information required for controlling the system is nonnegligible with respect to transmission time. A typical example is the following. In Asynchronous Transfer Mode (ATM), several types of customers can use the same User Network Interface (UNI) to access the network. Priority may be assigned to the various types. A field in the Protocol Data Unit (ATM-PDU), called Generic Flow Control (GFC), can be used to control the access at a single UNI. In Broadband Integrated Services Digital Networks (B-ISDN) the transmission time of a cell (the time slot in our model below) equals 0.68μsec when dealing with the 622Mb/sec normalized UNI, and 2.73 μsec when dealing with the 155Mb/sec normalized UNI. The propagation delay is about 4μsec/km. As a consequence, the propagation delay between the controller and the customers is not negligible with respect to the cell transmission time. Such type of applications motivated recent work on queues with delayed information, see [1, 2, 3, 4, 5] and references therein, which focused on control aspects. The purpose of this paper is to study performance measures of such queues.

We analyze a discrete-time queue with batch arrivals of messages and a single server. Each message is comprised of a variable number of packets. The transmission time of a single packet is one time slot, so that the number of time slots required for serving (transmitting) a message may vary from message to message. Service is initiated only at the beginning of a slot, and messages that arrive during a time slot are candidates to begin service only at the beginning of the next slot. The server has a delayed information on the number of messages in the queue, i.e., at the beginning of time slot n, the server knows how many messages were present in the queue at the beginning of time slot $n - K$. The server has no information on the length of any message, nor does it know the number of messages that arrived during any of the last K slots. We assume that it is not desirable to try to serve messages when the system may be empty, since the server may be used for performing other tasks at those times.

We consider two situations regarding information on service completions. In the first (Section 2), the server has full and non delayed information on service completions in previous time slots. For this case we obtain a general decomposition rule assuming that the arrival process is stationary and independent of the service process. We then derive explicit expressions for the Probability

Generating Function (PGF) of the queue length when the number of message-arrivals per slot is an i.i.d. sequence, and the service durations are general and i.i.d..

The second situation is when the service completion information is also delayed (Section 4). We restrict the analysis to the case of a unit-slot delay, i.e. $K = 1$. The server will attempt to serve a message at time slot n if and only if, based on all information that the server possesses at that time, it is sure that the system is non empty. Consequently, the server will attempt to serve the queue at time n if either there were at least two messages in the queue at the beginning of time slot $n-1$, or if there was one message at the beginning of time slot $n-1$ but service was not attempted (service would not be attempted at time $n-1$ if, for example, at time $n-2$ the system were empty). We obtain explicit expressions for the steady state PGF of the number of messages in the system by viewing the problem as a discrete time queue with vacations, and using a work decomposition result. Finally (Section 5), we discuss a priority queue with delayed information, where all messages are of unit length. We obtain explicit expressions for the queue size and mean waiting time of a class-m customer, and show that the $c\mu$ rule remains optimal even in the case of delayed information.

2 Delayed information on arrivals, instantaneous information on service completion

Consider a single server queue, where time is slotted, i.e., time is subdivided into some basic units of length δ; slot j begins at time $j\delta$, where $j = ..., -1, 0, 1, 2, 3, ...$ We consider below batch arrivals of messages into a single queue and we are interested in the process that describes the number of messages in the queue. Each message is composed of a sequence of packets (each with unit size). Thus, the time it takes to transmit a message may vary from message to message. The underlying process is the "late arrival" model described in Section 6.1 of Takagi [13].

More precisely, let $\{X_n\}$, $n = -K, -K+1, ..., -1, 0, 1, 2, 3, ...,$ be the process of queue lengths, where X_n denotes the number of <u>messages</u> in the queue at the <u>beginning</u> of time slot n. During time-slot n, a batch of A_n messages arrives. Let U_n be a decision variable denoting whether service is attempted at time n ($U_n = 1$), or not ($U_n = 0$). U_n depends on the information available to the server (see below) at the beginning of slot n. We define a sequence $C_n \in \{0,1\}$ which corresponds to a potential service completion; if service is attempted at the beginning of slot n (and the system is nonempty), then service completion (of a whole message) occurs at the end of the slot if and only if $C_n = 1$. The interpretation of C_n is as follows. Suppose the kth arriving message is of a given length, say $\tau_k \geq 1$ packets, and at time n the server starts to serve that message and continues serving its packets until the whole message is transmitted. Since we assume that it takes one time slot to serve one packet, then $C_m = 0$ for $m = n, n+1, ..., n+\tau_k - 2$, and $C_m = 1$ for $m = n+\tau_k - 1$. Hence, $C_m = 1$ at those time slots at the end of which the service (the transmission) of a whole message is completed.

If $U_n = 1$ and $C_n = 1$, then a message leaves the system. This yields the following dynamics:

$$X_{n+1} = X_n + A_n - U_n C_n \tag{1}$$

The service decisions are governed by the following rule: service is attempted at time n if the server is sure that the system is nonempty at that time. The available information for this decision is the queue length till time $n-K$, as well as the previous decisions U_l and the values of C_l, $l < n$. More precisely, we assume that U_n is given by

$$U_n = 1\left\{ X_{n-K} > \sum_{i=1}^{K} U_{n-i} C_{n-i} \right\}. \tag{2}$$

Note, in particular, that if service is initiated at time n, then it will continue till the current message leaves the system. Indeed, assume that $U_n = 1$, and let $m > n$ be the instant at which the message that starts being served at time n is fully transmitted and leaves the queue. Then for any l such that $n < l < m$,

$$X_{l-K} \geq X_{n-K} - \sum_{i=n-K}^{l-K-1} U_i C_i, \tag{3}$$

This follows since $\sum_{i=n-K}^{l-K-1} U_i C_i$ is the number of messages transmitted during time slots $n-K$ to $l-K-1$. Now,

$$\sum_{i=1}^{K} U_{n-i} C_{n-i} - \sum_{i=n-K}^{l-K-1} U_i C_i \;=\; \sum_{i=n-K}^{n-1} U_i C_i - \sum_{i=n-K}^{l-K-1} U_i C_i = \sum_{i=l-K}^{n-1} U_i C_i = \sum_{i=l-K}^{l-1} U_i C_i$$

$$= \sum_{i=1}^{K} U_{l-i} C_{l-i}. \tag{4}$$

The equation before the last follows since the message whose transmission has started at time n is not leaving until time m and so $U_n C_n = U_{n+1} C_{n+1} = ... = U_{l-1} C_{l-1} = 0$. Since we assumed that service was attempted at time n, it follows from (2) that $X_{n-K} > \sum_{i=1}^{K} U_{n-i} C_{n-i}$. By combining this with (3) and (4), we obtain

$$X_{l-K} \geq X_{n-K} - \sum_{i=n-K}^{l-K-1} U_i C_i > \sum_{i=1}^{K} U_{n-i} C_{n-i} - \sum_{i=n-K}^{l-K-1} U_i C_i = \sum_{i=1}^{K} U_{l-i} C_{l-i}, \qquad n < l \leq m-1$$

Hence, using (2) implies that $U_l = 1, \ n \leq l \leq m-1$. Service is thus attempted till the message whose service has started at time n, leaves the system.

Thus, since each message is continuously served until transmission, one may view the system as a discrete time queue with non preemptive server vacations. The vacations are caused by the delay in information which may idle the server even when the system is non-empty. We will use this observation in the sequel.

Next, we make some probabilistic assumptions on the exogenous sequences $\{A_n\}$, and $\{C_n\}$. We consider a probability space on which the two random processes $\{A_n\}$ and $\{C_n\}$ are defined, where $n = -K, -K+1, ..., -1, 0, 1, 2, 3, ...$. We assume that the sequence $\{A_n\}$ is stationary, and independent of the sequence $\{C_n\}$. (We make no further assumptions on the sequence $\{C_n\}$). Finally, in order for the processes X_n and U_n to be well defined, we make assumptions on the initial conditions: we assume that

$$U_{-K}, ..., U_{-1} = 0, \text{ and } X_{-K} \text{ has some initial distribution } \gamma. \tag{5}$$

Let X^0 denote the queue length process corresponding to no delay in information, i.e. $K = 0$, and an initial distribution γ (the same as in (5)).

Proposition 2.1 *(i) We have*

$$X_n =_d X_n^0 + \sum_{i=1}^{K} A_{n-i+K} = X_n^0 + \sum_{i=n}^{n+K-1} A_i. \tag{6}$$

(ii) If $\{A_n\}$ are i.i.d. then

$$X_n =_d X_n^0 + \sum_{i=1}^{K} \tilde{A}_i, \tag{7}$$

where $\tilde{A}_i$ are i.i.d., distributed like A_1, and are independent of X_n^0.

Remark 2.1 *Note that in (i), X_n^0 and $\sum_{i=1}^{K} A_{n-i+K}$ need not be independent. This follows from the fact that X_n^0 depends on the whole arrival process $\{A_l, l < n\}$, and since the sequence $\{A_n\}$ are not assumed to be i.i.d., the variables $\{A_n, ..., A_{n+K-1}\}$ also depend on $\{A_l, l < n\}$, and hence the possible dependence between X_n^0 and $\sum_{i=0}^{K-1} A_{n+i}$.*

Proof of Proposition 2.1: We define the quantity

$$Y_n = X_{n-K} - \sum_{i=n-K}^{n-1} U_i C_i \tag{8}$$

which is the number of messages in the queue at (the start of) time n if during the last K slots there were no arrivals. Combining (2) and (8) we obtain

$$U_n = 1\{Y_n > 0\}. \tag{9}$$

From (1) and (8) we get

$$X_{n-K+1} = \left(Y_n + \sum_{i=1}^{K} U_{n-i} C_{n-i} \right) + A_{n-K} - U_{n-K} C_{n-K} = Y_n + \sum_{i=1}^{K-1} U_{n-i} C_{n-i} + A_{n-K} \tag{10}$$

and using again (8) we get

$$\begin{aligned} Y_{n+1} &= X_{n-K+1} - \sum_{i=0}^{K-1} U_{n-i} C_{n-i} = Y_n + A_{n-K} - U_n C_n \\ &= Y_n + A_{n-K} - 1\{Y_n > 0\} C_n \end{aligned} \tag{11}$$

where the last equality follows from (9). Finally, by repeated use of (1),

$$X_n = X_{n-K} + \sum_{i=1}^{K} A_{n-i} - \sum_{i=1}^{K} U_{n-i} C_{n-i} = Y_n + \sum_{i=1}^{K} A_{n-i}. \tag{12}$$

Since (i) the sequence $\{A_{n-K}\}$ has the same distribution as the sequence $\{A_n\}$, (ii) equation (11) describes the evolution of the system with $K=0$ and a shifted arrival process, and (iii) the distribution of Y_0 is γ (by (5) and (8)), we have

$$\{Y_n, A_{n-K}, n=0,1,...\} =_d \{X_n^0, A_n, n=0,1,...\}. \tag{13}$$

This, together with (12), implies (6). Finally, if the arrivals are i.i.d., Y_n does not depend on $A_{n-i},\ i=1,...,K$, from which (7) follows. ∎

As a consequence of Proposition 2.1, we may obtain explicit expressions for the Probability Generating Function (PGF) of the queue length in steady state, for the case where both the number of arrivals per slot and service durations are each an i.i.d. sequence, independent of each other. (This model is denoted in Takagi [13] as a Batch Arrival $Geo/G/1$ queue). Let τ_n be the number of slots required for serving the nth message. Denote the PGF of τ_1 and A_1 by

$$\tilde{\tau}(z) = E\left[z^{\tau_1}\right], \qquad \tilde{A}(z) = E\left[z^{A_1}\right],\ |z| \le 1.$$

Let $\lambda = E[A_1]$, $\rho = \lambda E[\tau_1]$.

The queue length in a Batch Arrival $Geo/G/1$ discrete time queue with non-idling service (and no delay in information) evolves according to

$$X_{n+1}^0 = \max(X_n^0 - C_n, 0) + A_n. \tag{14}$$

X_n^0 is known to be stable if $\rho < 1$; in that case it converges in distribution to a finite random variable X^0, whose PGF is given by (see [13] page 21)

$$\tilde{X}^0(z) = E\left[z^{X^0}\right] = \frac{(z-1)\tilde{\tau}(\tilde{A}(z))}{z - \tilde{\tau}(\tilde{A}(z))}(1-\rho)\,,\ |z| \le 1, \tag{15}$$

and

$$E\left[X^0\right] = \rho + \frac{E[\tau_1(\tau_1 - 1)]\lambda^2 + E[\tau_1]E[A_1(A_1-1)]}{2(1-\rho)}. \tag{16}$$

Note that since $\rho < 1$ is the (sufficient and necessary) stability condition for the process X_n^0 (and thus of Y_n), it follows from (12) that it is the stability condition for the process X_n. Combining (15), (16) with Proposition 2.1, we obtain

Corollary 2.1 *Assume that $\rho < 1$, and let $\{A_n\}$ and $\{\tau_n\}$ be i.i.d. and independent. Then the queue length for the system with K slots delayed information is stable: the queue length converges in distribution to a finite random variable X^K, whose PGF is given by*

$$\tilde{X}^K(z) = E\left[z^{X^K}\right] = \frac{(z-1)\tilde{\tau}(\tilde{A}(z))}{z-\tilde{\tau}(\tilde{A}(z))}\,(1-\rho)\,[\tilde{A}(z)]^K, \ |z| \le 1, \tag{17}$$

$$E\left[X^K\right] = E\left[X^0\right] + K\lambda, \tag{18}$$

where $E\left[X^0\right]$ is given in (16).

For the case of stationary arrival processes which are not i.i.d., and if the sequence $\{A_n\}$ is associated (see definition below), one can still obtain some performance bounds on the moments of the queue length in term of those of the system without delayed information, by using Proposition 2.1. This could be useful for MAPs (Markov Arrival Processes) or MMPPs (Markov Modulated Poisson Processes).

Following Barlow and Proschan [7], the $\mathbf{R}$-valued RV's $\alpha = \{\alpha^1, ..., \alpha^N\}$ are said to be *associated* if the inequality

$$E[f(\alpha)g(\alpha)] \ge E[f(\alpha)]E[g(\alpha)] \tag{19}$$

holds for all *monotone non-decreasing* mappings $f, g : \mathbf{R}^{\mathrm{N}} \to \mathbf{R}$ for which the expectations $E[f(\alpha)], E[g(\alpha)]$ and $E[f(\alpha)g(\alpha)]$ exist.

Lemma 2.1 *(i) Assume that $\{X_n^0, \sum_{i=0}^{K-1} A_{n+i}\}$ are associated for some n. Then for any integer l,*

$$E\left[(X_n)^l\right] \ge E\left[\left(X_n^0 + \sum_{i=1}^{K} \tilde{A}_i\right)^l\right], \tag{20}$$

where $\{\tilde{A}_i, i = 1, ..., K\}$ are distributed like $\{A_i, i = 1, ..., K\}$ and are independent of X_n^0.

(ii) If $\{A_m, m \le n + K - 1\}$ are associated then (20) holds.

(iii) Assume that the queue is stable, i.e., X_n^0 converges in distribution to a finite random variable X_0; assume that for some integer m, $E[(X_n^0)^l]$ converges to $E[(X^0)^l]$ for all $l \le m$, and that the

first m moments of A_1 exist. Assume that $\{X_n^0, \sum_{i=0}^{K-1} A_{n+i}\}$ are associated for all n, or that $\{A_m, m = 1, 2, ...\}$ are associated. Then

$$E\left[(X)^l\right] \geq E\left[\left(X^0 + \sum_{i=1}^{K} \tilde{A}_i\right)^l\right], \forall l \leq m \tag{21}$$

where $\{\tilde{A}_i, i = 1, ..., K\}$ are distributed like $\{A_i, i = 1, ..., K\}$ and are independent of X^0.

Proof:

$$\begin{aligned} E\left[(X_n)^l\right] &= E\left[\left(X_n^0 + \sum_{i=0}^{K-1} A_{n+i}\right)^l\right] = \sum_{j=0}^{l} \binom{l}{j} E\left[\left(X_n^0\right)^j \left(\sum_{i=0}^{K-1} A_{n+i}\right)^{l-j}\right] \\ &\geq \sum_{j=0}^{l} \binom{l}{j} E\left[\left(X_n^0\right)^j\right] E\left[\left(\sum_{i=0}^{K-1} A_{n+i}\right)^{l-j}\right] = E\left[\left(X_n^0 + \sum_{i=0}^{K-1} \tilde{A}_{n+i}\right)^l\right]. \end{aligned}$$

The first equality follows from (6), and the inequality follows from (19) since $\{X_n^0, \sum_{i=0}^{K-1} A_{n+i}\}$ are associated. This implies (i). If $\{A_m, m \leq n+K-1\}$ are associated then $\{X_n^0, A_{n+i}, i = 0, ..., K-1\}$ are associated (see [8] Section III.4.2) and hence $\{X_n^0, \sum_{i=0}^{K-1} A_{n+i}\}$ are associated. Indeed, this follows from the fact that any non-decreasing function of associated RVs generate a set of associated RVs, see Lemma III.4. in [8] (in our case, $\{X_n^0, \sum_{i=0}^{K-1} A_{n+i}\}$ can be written trivially as a non-decreasing function of $\{X_n^0, A_{n+i}, i = 0, ..., K-1\}$). Claim (ii) now follows from (i). Claim (iii) follows by going to the limits. ∎

3 Applying decomposition results for vacation models

In order to be able to analyze the case of delayed information of both arrivals and service completions, we need the following result (from [10]) on work decomposition for discrete time queues with batch arrivals and server vacations. A vacation is defined as a time period in which the server is not available for service. In Section 2, a vacation corresponds to a sequence of consecutive time-slots n for which $U_n = 0$, where U_n is defined in equation (2).

Consider the following discrete time Batch Arrival $Geo/G/1$ queue, with service vacations. During time slot n, a batch of A_n messages arrives. The number of packets in the mth arriving message (and hence the time duration in slots it takes to serve the mth message) is τ_m. $\{A_n\}$ and $\{\tau_m\}$ are i.i.d. sequences, independent of each other. Occasionally, the server takes a vacation.

We shall use the results from [10] for the more general vacation structure described in [12]. We make the following additional assumptions from [12]: (i) Each service time is independent of the sequence of vacation periods that preceded that service time. (ii) All customers (messages) arriving to the system are eventually served; the system has an infinite queueing capacity, and $\rho = E[A_1]E[\tau_1] < 1$. (iii) Customers are served in an order that is independent of their service times. (iv) Service is nonpreemptive. (v) The rules that govern when the server begins and ends vacations do not anticipate future jumps of the arrival process ("anticipate" in the sense of Wolff [16]), see Assumption 5 in [12].

Lemma 3.1 *In steady state, the amount of work W at the beginning of an arbitrary slot is equal in distribution to the sum of the following independent RVs: (i) V, the amount of work in the same queue at the beginning of an arbitrary slot when there are no vacations; (ii) Y, the amount of work at the beginning of an arbitrary non-serving slot. The same decomposition holds also for the queue length: $X =_d X_{nv} + Q$, where X_{nv} is the queue length in the beginning of an arbitrary slot in the queue without vacations and Q is the number of messages at the beginning of an arbitrary non-serving slot.*

Proof: The proof for the work in the system can be found in Theorem 1 of [10], which is the discrete time version of the proof in [9]. Although the theorem is presented for cyclic polling systems with given service policies, i.e. for a special kind of vacations, it can easily be checked that the proof holds also in our case. (Indeed, let $\overline{C}$ be an arbitrary arriving message, and let $\overline{C}_A$ be his ancestor, in the sense defined in [9, 12]. Then, in order for the proof in [9, 10] to carry, one has to show that the amount of work found by $\overline{C}_A$'s batch is distributed like Y. This follows indeed from the assumptions on the vacation discipline, as follows from the argument in the first paragraph of the proof of Proposition 5 in [12]). The statement for the queue length follows from similar arguments. ∎

We show how Corollary 2.1 can be obtained using the above decomposition. We first show that Q, the queue length at the beginning of an arbitrary non-serving slot, is distributed like

$\sum_{i=1}^{K} A_i$. Assume that the server is busy till time $n-1$ and idles from time n (hence $U_n = 0$ and $U_{n-1}C_{n-1} = 1$) till time $m > n$ (at time slot m it restarts serving). It follows that all messages present at time $n-K$ have been served till the end of slot $n-1$, i.e. $X_{n-K} = \sum_{i=1}^{K} U_{n-i}C_{n-i}$. The number of messages in the queue at the beginning of the non-serving period, i.e. at time slot n, is thus $X_n = \sum_{i=n-K}^{n-1} A_i =_d \sum_{i=1}^{K} A_i$. Combining this with (15) and with Lemma 3.1, we obtain the result in Corollary 2.1.

Remark 3.1 In fact, one can show that at each idling slot, $n \le l \le m-1$, the number of messages in the queue is distributed like $\sum_{i=1}^{K} A_i$. Indeed, assume that $n+1 = l < m$. Then $A_{n-K} = 0$, for otherwise the server would not stay idle at time $l = n+1$. The number of messages in the queue at time $n+1$ is thus $X_{n+1} = X_n + A_n = \sum_{i=n-K+1}^{n} A_i =_d \sum_{i=1}^{K} A_i$. Iterating this argument we see that for each idling slot, the number of messages in the queue is indeed distributed like $\sum_{i=1}^{K} A_i$.

4 Delayed information on arrivals and on service completion

In this section we study the case where information on both the queue size and messages' service completion is delayed $K = 1$ slots. Consequently, at time n, the value of C_{n-1} is unknown.

The dynamics (1) stays unchanged, but the decision at time n whether to serve a message or not will differ from that of Section 2. We consider the following dynamics

$$U_n = 1\left(\{X_{n-1} > 1\} \cup \{X_{n-1} = 1, U_{n-1} = 0\}\right) \tag{22}$$

That is, the server attempts to serve the queue if either there were at least two messages at the beginning of time slot $n-1$, or if there was one message at that time but service was not attempted. Otherwise, the server takes a vacation of one slot, at the end of which new information is obtained.

We assume below that $\{A_n\}$ and $\{C_n\}$ are two independent i.i.d. sequences, and that C_n are Bernoulli with parameter μ (i.e. $C_n = 1$ w.p. $0 \le \mu \le 1$). We shall assume that the system operates in a steady state regime ($\rho < 1$).

We now compute the PGF of Q, the number of messages in steady state at an arbitrary non-serving period. Let n be the instant when a non-serving period begins. Hence, at the beginning of time slot $n-1$ there was still one message in the system, and service was attempted during time slot $n-1$. This means that at the beginning of time slot n, the queue contains the messages that have arrived during time slot $n-1$, plus (perhaps) the message whose service was attempted (and not necessarily completed) at time slot $n-1$.

There are two possible types of non-serving periods, I_1 and I_2 (these will also be used to denote durations of these periods).

Type 1: If the service of the single message that was present in the system at time slot $n-1$ was completed at the end of that time slot, then at the beginning of time slot n, there are A_{n-1} messages present. The duration of the non-serving period is geometrically distributed with parameter $\tilde{A}(0)$ (the probability of no arrival in a time slot), and its expected duration is $1/(1-\tilde{A}(0))$.

Type 2: The service of the single message that was present in the system at time slot $n-1$ was not completed at the end of time slot $n-1$. At the beginning of time slot n, there are $A_{n-1}+1$ messages. The server takes a vacation of only one time slot, since at time slot $n+1$ the server will be aware that the message served at time $n-1$ is still in the system regardless of the value of A_{n-1}.

It follows that the expected duration of an arbitrary non-serving period I is

$$E[I] \quad = \quad E[I_1]P(C_1=1)+E[I_2]P(C_1=0)=\frac{1}{(1-\tilde{A}(0))}\times\mu+1\times(1-\mu).$$

The probability $P(I_2)$ that an arbitrary message that arrives during a non-serving period arrives to a type-2 non-serving period is $(1-\mu)/E[I]$, and the probability $P(I_1)$ that it arrives to a type-1 non-serving period is $1-(1-\mu)/E[I]$.

Let $\tilde{Q}(z)$ denote the PGF of Q. Then

$$\begin{aligned}\tilde{Q}(z)=E\left[z^Q\right] &= E\left[z^Q\middle| I_1\right]P(I_1)+E\left[z^Q\middle| I_2\right]P(I_2) \qquad (23)\\ &= \tilde{A}(z)P(I_1)+\tilde{A}(z)zP(I_2)\\ &= \tilde{A}(z)\left(1+(z-1)(1-\mu)\left[\frac{\mu}{(1-\tilde{A}(0))}+1-\mu\right]^{-1}\right).\end{aligned}$$

Applying the decomposition from Lemma 3.1 and using (15), the PGF of the number of messages in the system at steady state is

$$\tilde{X}(z) = \frac{(z-1)\tilde{\tau}(\tilde{A}(z))}{z-\tilde{\tau}(\tilde{A}(z))}(1-\rho) \times \tilde{A}(z)\left(1+(z-1)(1-\mu)\left[\frac{\mu}{(1-\tilde{A}(0))}+1-\mu\right]^{-1}\right). \tag{24}$$

The assumption that $\{C_n\}$ is a Bernouilli sequence is equivalent to assuming that $\{\tau_m\}$ is a sequence of geometric i.i.d. random variables, i.e., $P(\tau_m = j) = (1-\mu)^{j-1}\mu,\ j = 1,2,3,....$ The PGF, $\tilde{\tau}(z)$, of τ_1 and its moments are given by

$$\tilde{\tau}(z) = \frac{\mu z}{1-(1-\mu)z}, \qquad E[\tau_1] = \mu^{-1}, \qquad E[\tau_1(\tau_1 - 1)] = \frac{2(1-\mu)}{\mu^2}. \tag{25}$$

The PGF of W, the amount of work in the system is then given by

$$\begin{aligned} E\left[z^W\right] &= E\left\{E\left[z^W \middle| X\right]\right\} = E\left\{E\left[z^{\sum_{i=1}^{X}\tau_i} \middle| X\right]\right\} \\ &= E\left\{(E[z^{\tau_1}])^X\right\} = \tilde{X}\left(\tilde{\tau}(z)\right) \\ &= \tilde{X}\left(\frac{\mu z}{1-(1-\mu)z}\right). \end{aligned}$$

5 Priority queues

Consider N queues and a single server. Time is slotted as before. Let $\{X_n = (X_n^1, ..., X_n^N)\}$, $n = -K, -K+1, ..., -1, 0, 1...$, be the process of queue lengths, where X_n^i denotes the number of messages in queue i at the beginning of time slot n. During time-slot n, a batch of A_n^i messages arrives to queue i. The arrivals to different queues at a time-slot may be correlated. U_n^i denotes whether service is attempted at time n at queue i. We consider the simpler case where all messages are of one unit length, so that if service is attempted at time n in queue i then a message leaves queue i at the end of that slot. This yields the following dynamics:

$$X_{n+1}^i = X_n^i + A_n^i - U_n^i, \qquad i = 1, ..., N \tag{26}$$

The service decisions are governed by the following rule:

(i) service is attempted at time n if the server is sure that the system is nonempty at that time. The available information for this decision is the queue lengths till time $n-K$, as well as the previous decisions $U_l = (U_l^1, ..., U_l^N)$ for $l < n$.

(ii) The queues have a (non-preemptive) priority order according to their indices, so that queue 1 has the highest priority and queue N has the lowest. This is reflected in the following evolution of U_n:

$$\begin{aligned}
U_n^1 &= 1\{X_{n-K}^1 > \sum_{i=1}^{K} U_{n-i}^1\}, \\
U_n^2 &= 1\{U_n^1 = 0, X_{n-K}^2 > \sum_{i=1}^{K} U_{n-i}^2\}, \\
U_n^m &= 1\{U_n^l = 0, l = 1, ..., m-1, X_{n-K}^m > \sum_{i=1}^{K} U_{n-i}^m\}, \quad m = 3, ..., N. \qquad (27)
\end{aligned}$$

We assume that $U_j^m = 0$ for $m = 1, ..., N$, $j = -K, -K+1, ..., -1$.

The above enables one to express the distribution of the total number of messages belonging to queues with priority higher than or equal to l, $1 \le l \le N$.

We show first that for every m, $1 \le m \le N$, and all $n \ge 0$,

$$X_{n-K}^m \ge \sum_{i=1}^{K} U_{n-i}^m. \qquad (28)$$

Indeed, this is true for $n = 0$, since $U_j^m = 0$ for $m = 1, ..., N$, $j = -K, -K+1, ..., -1$. Now assume it holds up to some n. We show that it also holds for $n+1$, thus establishing the property by induction. (i) if

$$X_{n-K}^m > \sum_{i=1}^{K} U_{n-i}^m$$

then

$$X_{n+1-K}^m \ge X_{n-K}^m - U_{n-K}^m > \sum_{i=1}^{K} U_{n-i}^m - U_{n-K}^m = \sum_{i=1}^{K-1} U_{n-i}^m \ge \sum_{i=1}^{K} U_{n+1-i}^m - 1$$

and hence, $X^m_{n+1-K} + 1 > \sum_{i=1}^K U^m_{n+1-i}$, so that

$$X^m_{n+1-K} \geq \sum_{i=1}^{K} U^m_{n+1-i}.$$

(ii) Now, if

$$X^m_{n-K} = \sum_{i=1}^{K} U^m_{n-i}.$$

Then $U^m_n = 0$ by (27), and

$$\begin{aligned} X^m_{n+1-K} &= X^m_{n-K} + A^m_{n-K} - U^m_{n-K} = \sum_{i=1}^{K} U^m_{n-i} + A^m_{n-K} - U^m_{n-K} \\ &= \sum_{i=1}^{K-1} U^m_{n-i} + A^m_{n-K} = \sum_{i=1}^{K-1} U^m_{n-i} + U^m_n + A^m_{n-K} = \sum_{i=1}^{K} U^m_{n+1-i} + A^m_{n-K} \geq \sum_{i=1}^{K} U^m_{n+1-i}. \end{aligned}$$

Define now, for each l, $1 \leq l \leq N$, $X^*_n(l) = \sum_{i=1}^l X^i_n$, $A^*_n(l) = \sum_{i=1}^l A^i_n$, $U^*_n(l) = \sum_{i=1}^l U^i_n$. Then, by (28), $X^*_{n-K}(l) > \sum_{i=1}^K U^*_{n-i}(l)$ if and only if, for some $r = 1, ..., l$, $X^r_{n-K} > \sum_{i=1}^K U^r_{n-i}$. It then follows that

$$X^*_{n+1}(l) = X^*_n(l) + A^*_n(l) - U^*_n(l), \tag{29}$$

$$U^*_n(l) = 1\left\{ X^*_{n-K}(l) > \sum_{i=1}^{K} U^*_{n-i}(l) \right\}. \tag{30}$$

We see that (29) and (30) are special case of the dynamics (1) and (2). It thus follows that we may apply Proposition 2.1, Corollary 2.1 and Lemma 2.1 of Section 2 to obtain expressions (and bounds) for the distribution of $X^*_n(l)$ and its steady state version.

Finally, in order to compute the expectation of X^m_n, the number of messages of priority m at time n, it suffices to note that $X^m_n = X^*_n(m) - X^*_n(m-1)$ and hence $E[X^m_n] = E[X^*_n(m)] - E[X^*_n(m-1)]$. A similar argument allows to compute the expected number of messages of priority m in steady state (see (18)), the expressions for which are given in the sequel.

Optimal order of priority

Let $X^{K,m}$ be the steady state queue size of priority m when the delay is K slots, and let $X^{K,*}(m) =$

$\sum_{l=1}^{m} X^{K,l}$, $X^{K,*}(0) = 0$. Then for $1 \le m \le N$, $X^{K,m} = X^{K,*}(m) - X^{K,*}(m-1)$. From (18),

$$E\left[X^{K,*}(m)\right] = E\left[X^{0,*}(m)\right] + K\left(\sum_{j=1}^{m} \lambda_j\right) \tag{31}$$

where $\lambda_j = E[A_1^j]$. Therefore,

$$E\left[X^{K,m}\right] = E\left[X^{0,m}\right] + K\lambda_m. \tag{32}$$

Let $W^{K,m}$ be the waiting time (until service) of class-m customer, where the delay is K slots. Then, since each message is of length one,

$$E\left[X^{0,m}\right] = \lambda_m \left(E\left[W^{0,m}\right] + 1\right) \tag{33}$$

and

$$E\left[W^{K,m}\right] = \frac{E\left[X^{K,m}\right]}{\lambda_m} - 1. \tag{34}$$

It follows that

$$\begin{aligned} E\left[W^{0,m}\right] \quad = \quad & \frac{\sum_{k=1}^{m} E[A_1^k(A_1^k - 1)]}{2\left(1 - \sum_{i=1}^{m-1} \lambda_i\right)\left(1 - \sum_{i=1}^{m} \lambda_i\right)} + \frac{\sum_{k=2}^{m} \sum_{j=1}^{k-1} E[A_1^k A_1^j]}{\left(1 - \sum_{i=1}^{m-1} \lambda_i\right)\left(1 - \sum_{i=1}^{m} \lambda_i\right)} \\ & + \frac{E[A_1^m(A_1^m - 1)] + 2\sum_{k=1}^{m-1} E[A_1^k A_1^m]}{2\lambda_m\left(1 - \sum_{i=1}^{m-1} \lambda_i\right)}, \end{aligned} \tag{35}$$

see Takagi [13] equation (6.39), and Takahashi and Hashida [14], [15]. Combining (32), (33) and (35) yields explicit expressions for $E\left[X^{K,m}\right]$.

We now wish to order the priority classes such that the mean weighted waiting time of an arbitrary customer is minimized. That is, our objective is to

$$\text{Minimize} \left\{\sum_{m=1}^{N} c_m \frac{\lambda_m}{\lambda} E\left[W^{K,m}\right]\right\}, \tag{36}$$

where c_m is the rate of cost of class-m customers. By (32), (33) and (34),

$$E\left[W^{K,m}\right] = \frac{E\left[X^{0,m}\right]}{\lambda_m} + K - 1 = E\left[W^{0,m}\right] + K.$$

Hence, (36) is equivalent to

$$\text{Minimize} \sum_{m=1}^{N} c_m \lambda_m \left[E\left[W^{0,m} \right] + K \right]. \tag{37}$$

Since $\sum_{m=1}^{N} c_m \lambda_m K$ is constant over all priority orders, (37) is minimized by the well known $c\mu$ rule which, in our case, assigns priorites by decreasing order of c_m. Hence, the queue with the highest c_m gets the highest priority, that is, the lowest index. This follows since the $c\mu$ rule is already optimal for the case of no-delay (see [6] and [11] p. 84-85).

Acknowledgement: We thank the Associated Editor and the referees for their instructive comments.

References

[1] E. Altman and G. Koole, "Control of a random walk with noisy delayed information", Report BS-R9316, Department of OR, Statistics and System Theory, CWI, Amsterdam, Netherlands. to appear in *Systems and Control Letters.*

[2] E. Altman and G. Koole, "Stochastic scheduling games with Markov decision arrival processes", *Journal Computers and Mathematics with Appl.*, 3rd special issue on "Differential Games", pp. 141-148, 1993.

[3] E. Altman and P. Nain, "Closed-Loop control with delayed information", *Performance Evaluation Review* **20**, pp. 193-204, 1992.

[4] E. Altman and S. Stidham, "Optimality of monotonic policies for two-action Markovian decision processes", submitted to *QUESTA*, 1994.

[5] D. Artiges, "Optimal routing into two heterogeneous service stations with delayed information", to appear in IEEE-AC.

[6] J. S. Baras, D. -J. Ma, and A. M. Makowski, "K competing queues with geometric service requirements and linear costs: the μc rule is always optimal," *Systems and Control Letters* **6**, pp. 173-180, 1985.

[7] R.E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, Reading, (MA), 1975.

[8] F. Baccelli and A.M. Makowski, "Queueing models for systems with synchronization constraints", *Proceedings of the IEEE*, Vol. 77 No. 1, Special Issue on Dynamics of Discrete Event Systems, L. Ho Ed., pp. 138-161, 1989.

[9] O. J. Boxma, "Workloads and Waiting Times in Single-Server Systems with Multiple Customer Classes", *Queuing Systems*, **5**, pp. 185-214, 1989.

[10] O. J. Boxma and W. P. Groenendijk, "Waiting times in discrete-time cyclic-server systems", *IEEE Transactions on Communications*, **36**, pp. 164-170, 1988.

[11] D. R. Cox and W. L. Smith, *Queues*, Methuen, London, 1961.

[12] S. W. Fuhrmann and R. R. Cooper, "Stochastic decomposition in the M/G/1 queue with generalized vacations", *Operations Research*, **33**, pp. 1117-1129, 1985.

[13] H. Takagi, *Queueing Analysis, Vol. 3: Discrete-Time Systems*, North-Holland, 1993.

[14] Y. Takahashi and O. Hashida, "Delay analysis of discrete-time priority queue with structured inputs", *Queueing Systems*, **8**, pp. 149-164, 1991.

[15] Y. Takahashi and O. Hashida, "Correction to our paper: Delay analysis of discrete-time priority queue with structured inputs", *Queueing Systems*, **13**, pp. 449-450, 1993.

[16] R. W. Wolff, "Poisson arrivals see time averages", *Operations Research* **30**, pp. 223-231, 1982.