Waiting Times in Discrete-Time Cyclic-Service Systems

ONNO J. BOXMA AND WIM P. GROENENDIJK

Abstract—**This paper considers single-server, multiqueue systems with cyclic service in discrete time. Nonzero switch-over times between consecutive queues are assumed; the service strategies at the various queues may differ. A decomposition for the amount of work in such systems is obtained, leading to an exact expression for a weighted sum of the mean waiting times at the various queues. The present paper is the companion paper of Boxma and Groenendijk [1] where the continuous-time case is treated.**

I. INTRODUCTION

IN local communication networks, a number of stations often compete for the use of a common transmission medium. Various polling schemes are employed to coordinate and control the access to the communication channel. The performance of such polling schemes can be analyzed by studying single-server, multiqueue queueing systems. For example, in a token ring local area network, the common transmission channel may be represented by the single server, and the workstations attached to the ring by the queues. The circulation of the token along the ring implies that the stations are polled in a cyclic order. The resulting single-server, multiqueue system with cyclic service and switch-over times between queues is the subject of the present paper.

The main performance measure of interest in polling systems is the waiting time of messages at the stations. Unfortunately, explicit analytical results for even mean waiting times in cyclically served queueing systems are only available in some exceptional cases. The recent discovery of so-called *pseudoconservation laws* (Watson [13], Ferguson and Aminetzah [3]) is an important step forward. These laws are exact expressions for weighted sums of the mean waiting times. They can be readily used to obtain and/or test approximations for the mean waiting times at the various queues (cf. [2]). In [1], those pseudoconservation laws have been generalized by allowing a mixture of different service strategies at different queues. The proof of the resulting unified pseudoconservation law is based on a stochastic decomposition of the amount of work in the cyclic-service system. This decomposition provides a generalization of Kleinrock's work conservation principle [5] to models with switch-over times. The decomposition also allows a simple probabilistic interpretation of the various terms of the unified pseudoconservation law.

All above-mentioned results are for continuous-time systems. The main goal of the present paper is to obtain discrete-time analogs of the results of [1]—thus solving a problem posed by Takagi [12]. Our motivation is that discrete-time arrival and service processes naturally fit the generally time-synchronized configuration of practical communication networks (while continuous-time cyclic-service results can be easily obtained from their discrete-time counterparts). Discrete-time polling systems have been studied before, cf. Konheim and Meister [7], Swartz [10], Rubin and DeMoraes [9], and Takagi [11], but the bulk of the literature in this area is devoted to continuous-time systems. See Takagi [11] for an extensive survey of cyclic-service systems, and Takagi [12] for an update reflecting the rapid development and strong interest in this area of research.

The organization of the rest of the paper is as follows. In Section II, we consider cyclic-service systems without switch-over times. For such systems the principle of work conservation clearly holds. This principle naturally leads to a discrete-time version of Kleinrock's conservation law for mean waiting times. The extension of the work conservation principle to the case *with* switch-over times is made in Section III. The main result of the paper, the *discrete-time pseudoconservation law for mean waiting times*, is proved in Section IV. In Section V the relation between the obtained discrete-time results and results for the continuous-time case is presented. Section VI contains some concluding remarks and topics for further research. We close this introductory section by presenting a more detailed model description and some basic results of general validity.

Model Description

We consider a discrete-time queueing system with N stations (queues) $Q_1, \cdots, Q_N$ where each station has an infinite buffer capacity to store waiting messages (customers). Each message consists of a number of packets, which are assumed to be of fixed length. Time is slotted with slot size equal to the transmission time of the data contained in a packet (the service time of a packet). We shall call the time interval $[j, j+1]$ the jth slot.

Arrival Process

Let

$x_i(j) :=$ number of messages arriving at station i in the jth slot,

$b_i :=$ number of packets included in a message at station i.

The message arrival process at each station is assumed to be independent of those at other stations. The stochastic processes $\{x_i(j)\}$ and $\{b_i\}$ are assumed to be mutually independent. The $x_i(j)$, $j = 1, 2, \cdots$ are assumed to be independent, identically distributed random variables with z transform, first and second moment

$$A_i(z) := E[z^{x_i(j)}],\ \lambda_i := E[x_i(j)],\ \lambda_i^{(2)} := E[x_i^2(j)]. \tag{1.1}$$

Note that we can view the arrival process at Q_i as a Bernoulli arrival process with batch arrivals

$$A_i(z) = A_i(0) + [1 - A_i(0)]\, G_i(z)$$

Paper approved by the Editor for Queueing Networks and Performance of the IEEE Communications Society. Manuscript received March 17, 1987; revised July 24, 1987.

The authors are with the Centre for Mathematics and Computer Science, Amsterdam, The Netherlands.

IEEE Log Number 8718267.

with $G_i(z)$ denoting the z transform of the size of a type i batch.

Let

$$\lambda := \sum_{i=1}^{N} \lambda_i, \ \lambda^{(2)} := E\left[\left(\sum_{i=1}^{N} x_i(j)\right)^2\right]. \quad (1.2)$$

The z transform, first and second moment of the number of packets, b_i, in a message at Q_i are given by

$$B_i(z) := E[z^{b_i}], \ \beta_i := E[b_i], \ \beta_i^{(2)} := E[b_i^2]. \quad (1.3)$$

Further, introduce

$$\beta := \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \beta_i, \ \beta^{(2)} := \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \beta_i^{(2)}. \quad (1.4)$$

Note that $B_i(0) = \Pr\{b_i = 0\} = 0$ by definition. The offered traffic at the ith station, ρ_i, is defined as

$$\rho_i := \lambda_i \beta_i, \ i = 1, 2, \cdots, N. \quad (1.5)$$

The total offered traffic ρ is defined as

$$\rho := \sum_{i=1}^{N} \rho_i. \quad (1.6)$$

Service Strategy

We assume that a single server S visits the N stations in the order of their indexes $i = 1, 2, \cdots, N$ ("cyclic service"). For the service strategies at the queues there are various possibilities, which differ in the number of messages which may be served in a queue during a visit of server S to that queue. Assume that S visits Q_i. When Q_i is empty, S immediately begins to switch to Q_{i+1} (we disregard variants in which S does not switch if none of the queues contains messages). Otherwise, S acts as follows, depending on the service strategy at Q_i.

1) Exhaustive service (E): S serves type i messages until Q_i is empty.
2) Gated service (G): S serves exactly those type i messages present upon his arrival at Q_i (a gate closes upon his arrival).
3) 1-limited service (1-L): S serves one type i message (the term nonexhaustive has often been used for this strategy; in [1] we have accordingly used NE instead of 1L).
4) Semiexhaustive service (SE): S continues serving type i messages until the number present is one less than the number present upon his arrival.

In this paper, we will allow mixed cyclic-service strategies (e.g., semiexhaustive at Q_1, exhaustive at Q_2 and Q_4, 1-limited at Q_3, and gated at $Q_5, \cdots, Q_N$). The order of service within each queue is first-come-first-served (FCFS). This assumption is not essential. In the sequel, the system is assumed to be in equilibrium.

Remark 1

Consideration of mixed service strategies will enable us to prove results for various cyclic-service systems in a unified manner. However, it is also of practical interest to study mixed strategies. For example, according to the draft IEEE 802.6 recommendation of the committee on metropolitan area networks, two or more token ring local area networks are to be interconnected by a backbone ring through bridges. It is often natural to assign a higher priority to the queues which represent the bridges than to the other queues at the ring. The service discipline at the ordinary queues usually is 1-limited, but at the "bridge queues" one may consider another service discipline to model the preferential treatment received by these queues.

Switching Process

A switch-over time is needed to switch from one station to the next. The switch-over times of the server between the ith and the $(i + 1)$th station (measured in slots) are independent, identically distributed random variables with first moment s_i and second moment $s_i^{(2)}$. The first moment s of the total switch-over time during a cycle of the server is given by

$$s := \sum_{i=1}^{N} s_i \quad (1.7)$$

its second moment is given by $s^{(2)}$.

Some additional notation we shall be needing is the following:

- X_i: the number of type i messages in the system at an arbitrary epoch;
- X_i^w: the number of *waiting* type i messages in the system at an arbitrary epoch;
- W_i: waiting time of a type i message; the waiting time is counted from the beginning of the slot following the one in which the message arrived.

Remark 2

It should be noted that, as customary in discrete-time queueing literature, an *arbitrary epoch* is supposed to be the instant just after the beginning of a slot.

Below, we state a few general results for future reference. For any strictly cyclic-service system, we can define the cycle time C_i for Q_i as the time between two successive arrivals of S at Q_i. It is easily seen that the *mean* cycle time for Q_i, EC_i, is independent of i; we will denote it by EC. The visit time V_i of S for Q_i is the time between the arrival of S at Q_i and his subsequent departure from that queue. Balancing the flow of type i messages in and out of the system during a cycle shows that

$$\rho_i EC = EV_i. \quad (1.8)$$

Summing over i, we obtain

$$\rho EC = \sum_{i=1}^{N} EV_i = EC - s.$$

This yields

$$EC = \frac{s}{1-\rho} \quad (1.9)$$

and hence, from (1.8) and (1.9)

$$EV_i = \frac{\rho_i s}{1-\rho}. \quad (1.10)$$

The intervisit time I_i for Q_i is defined as

$$I_i := C_i - V_i. \quad (1.11)$$

Now some remarks about the conditions for ergodicity of these cyclic-service systems are in order. Clearly, $\rho < 1$ is a necessary condition. For exhaustive and gated service, this condition is also sufficient. For a queue Q_i with 1-limited service, it can be seen that

$$\frac{\lambda_i s}{1-\rho} < 1 \quad (1.12)$$

is an additional condition for the ergodicity of the cyclic-

service system; indeed, the mean number of type i arrivals during a cycle should be less than one. Note that it is possible that, even if Q_i is unstable, some of the other queues are stable.

Similarly, for a queue Q_i with semiexhaustive service, we have the following additional condition:

$$\lambda_i EI_i = \frac{\lambda_i s(1-\rho_i)}{1-\rho} < 1. \tag{1.13}$$

This reflects the fact that, for semiexhaustive service, the mean number of type i arrivals during the intervisit time I_i should be less than one, for during visit times the number of type i messages is at most reduced by one.

For the mixed strategies that we allow, the conditions (1.12) and (1.13) should be added to the stability condition $\rho < 1$ for those queues at which we have a 1L or SE strategy.

II. Conservation Law for the Discrete-Time $M/G/1$ Model

In this section, the switch-over times are taken to be zero; hence, the server works whenever there is work in the system, and is idle when there is no work in the system. Therefore, the principle of work conservation holds: the total amount of work V_c in the cyclic-service system does not depend on the order of service, and should hence equal the amount of work in a "corresponding" FCFS $M/G/1$ queueing system. This observation will allow us to derive a conservation law for mean waiting times in the cyclic-service system without switch-over times. We first introduce the notion of the "corresponding" $M/G/1$ queueing model. This is a discrete-time queueing model, consisting of one queue and one server with a Bernoulli (Memoryless = M) arrival process with batch arrivals. The arrival process is constructed as follows: the arrival streams at all N queues of the cyclic-service model are aggregated into a single arrival stream. The batch of all the messages arriving in a slot is called a *train*. In any slot, no train arrives with probability $\Pi_1^N A_i(0)$ and a train does arrive with probability $1 - \Pi_1^N A_i(0)$. An arbitrarily chosen message in this train poses a service request whose z transform is the mixture $\Sigma_1^N (\lambda_i/\lambda)B_i(z)$.

The principle of work conservation now states that V_c equals the amount of work in the corresponding $M/G/1$ system, $V_{M/G/1}$. Therefore, V_c also equals $V_{M/G/1}$ in distribution

$$V_c \overset{D}{=} V_{M/G/1}. \tag{2.1}$$

According to Kobayashi and Konheim [6], the mean number of messages in the corresponding system at an arbitrary epoch is given by

$$EX_{M/G/1} = \frac{\lambda^2\beta^{(2)}}{2(1-\rho)} + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2(1-\rho)} + \rho. \tag{2.2}$$

Note that the second term in the right-hand side disappears when the arrival process is Poisson. The mean number of messages in service is ρ; the residual service time of the message in service is $\beta^{(2)}/2\beta + 1/2$. Hence,

$$EV_{M/G/1} = \left[\frac{\lambda^2\beta^{(2)}}{2(1-\rho)} + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2(1-\rho)}\right]\beta + \rho\left[\frac{\beta^{(2)}}{2\beta} + \frac{1}{2}\right]. \tag{2.3}$$

Remark 3

It should be observed that in the renewal process in discrete time with interevent-time distribution with first moment β and second moment $\beta^{(2)}$, $\beta^{(2)}/2\beta + 1/2$ is the mean residual life time and $\beta^{(2)}/2\beta - 1/2$ is the mean past life time.

On the other hand, we can write EV_c as (cf. the definitions above Remark 2)

$$EV_c = \sum_{i=1}^{N} \beta_i EX_i^w + \sum_{i=1}^{N} \rho_i \left[\frac{\beta_i^{(2)}}{2\beta_i} + \frac{1}{2}\right] = \sum_{i=1}^{N} \rho_i EW_i + \sum_{i=1}^{N} \rho_i \left[\frac{\beta_i^{(2)}}{2\beta_i} + \frac{1}{2}\right]. \tag{2.4}$$

The second equality is based on Little's formula.

From (2.1), (2.3), and (2.4), we obtain the following expression for a weighted sum of the mean message waiting times

$$\sum_{i=1}^{N} \rho_i EW_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho. \tag{2.5}$$

We propose to call (2.5) the $M/G/1$ conservation law in discrete time. We have found no references to this relation in the literature, although it seems highly likely that it has been derived before.

III. A Stochastic Decomposition Result

In the sequel, switch-over times are incorporated in the systems under consideration. Because now the server may be idle (switching) although there is work in the system, Kleinrock's principle of work conservation is no longer valid. However, Theorem 1 below presents a natural modification of this work conservation principle. In the theorem, an arbitrary epoch is considered to be "in" a switching interval if it marks the beginning of a switching slot; the "corresponding" $M/G/1$ system is the system (without switch-over times) introduced in the preceding section.

Theorem 1

Consider a single-server cyclic-service system with mixed service strategies as described in Section I. Suppose the system is ergodic and stationary. Then the amount of work V_c in this system at an arbitrary epoch is distributed as the sum of the amount of work $V_{M/G/1}$ in the "corresponding" $M/G/1$ system at an arbitrary epoch and the amount of work Y in the cyclic-service system at an arbitrary epoch in a switching interval. In other words,

$$V_c \overset{D}{=} V_{M/G/1} + Y \tag{3.1}$$

where $\overset{D}{=}$ stands for equality in distribution. Furthermore, $V_{M/G/1}$ and Y are independent.

Proof: The proof is similar to that of Theorem 1 in [1] for the continuous-time case, apart from the fact that *trains* are considered instead of customers. It is based on the following observations:

1) $V_{M/G/1}$ is not affected when the service discipline is LCFS nonpreemptive instead of FCFS.

2) V_c is also not affected when, instead of cyclic service, the following service strategy is enforced: all arriving trains are served LCFS, but service is interrupted precisely during the switch-over periods of the cyclic-service system.

3) It now suffices to prove that, in distribution, $V_c^{\text{LCFS}} = V_{M/G/1}^{\text{LCFS}} + Y$. The validity of this decomposition is a consequence of the LCFS discipline. Consider a train T that arrives during a switch-over period. It has to wait until trains that arrived after T, in the same switch-over period, have been served (and also trains arriving during their service, etc.). When, finally, T is taken into service, the only work present is the work that T found upon his arrival. This latter quantity is distributed like Y. Here, we use a discrete-time equivalent of the PASTA property [14], which we should like to call the BASTA property (Bernoulli arrivals see time averages); because the input of trains to the system is Bernoulli (and due

to the memoryless property of the underlying geometric distribution), the distribution of the amount of work at an arbitrary epoch is equal to the distribution of the amount of work immediately before an arrival epoch of a train. T initiates a busy period, which evolves exactly like a busy period in the "corresponding" $M/G/1$ system. So during the busy period initiated by T, the amount of work present in the system is distributed as the sum of Y and the amount of work during a busy period of the $M/G/1$ system. We refer to [1] for details.

Remark 4

Theorem 1 is the discrete-time analog of Theorem 1 of [1]. The latter theorem was motivated by, and its proof uses arguments suggested by, Fuhrmann and Cooper [4].

However, the reasoning in [4] is held for *customers* at *departure epochs* instead of *work* at *arbitrary epochs*. In [4], this leads to a similar decomposition as [1, Theorem 1] and (3.1), for *queue lengths*, for a class of so-called vacation systems. For our purposes, the amount of unfinished work is the natural quantity. Decomposition (3.1) holds for this quantity under very general assumptions (the restriction to cyclic service can in fact be relaxed). In the next section, decomposition (3.1) will be exploited to obtain a relation between the mean waiting times at the various queues of the cyclic-service system.

Remark 5

In Levy and Kleinrock [8], Y represents "the additional delay due to the presence of the starter."

IV. The Pseudoconservation Law

As a consequence of Theorem 1

$$EV_c = EV_{M/G/1} + EY \tag{4.1}$$

and hence, cf. (2.3) and (2.4),

$$\sum_{i=1}^{N} \rho_i EW_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho + EY. \tag{4.2}$$

We now derive an expression for EY, thus obtaining a very general pseudoconservation law for the weighted sum of the mean waiting times at the various queues. Let EY_i denote the amount of work in the cyclic-service system at an arbitrary switching epoch during a switch-over from Q_i to Q_{i+1}. Obviously, $EY = \sum_{i=1}^{N} (s_i/s)EY_i$. As in the continuous-time case, EY_i is composed of three terms:

1) $EM_i^{(1)}$: the mean amount of work in Q_i at a departure epoch of the server from Q_i.

2) $EM_i^{(2)}$: the mean amount of work in the rest of the system at a departure epoch of S from Q_i.

3) $\rho\{s_i^{(2)}/2s_i - 1/2\}$: the amount of work that arrived in the system during the past part of the switching interval under consideration (cf. also Remark 3).

Again, as in the continuous-time case, we have

$$\sum_{i=1}^{N} \frac{s_i}{s} EM_i^{(2)} = \frac{\rho}{s}\sum\sum_{h<k} s_h s_k + \frac{s}{1-\rho}\sum\sum_{h<k} \rho_h\rho_k + \sum_{i=1}^{N}\frac{s_i}{s}\sum_{j\neq i} EM_j^{(1)} \tag{4.3}$$

and hence for EY

$$EY = \rho\left(\frac{s^{(2)}}{2s} - \frac{1}{2}\right) + \frac{s}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{N}\rho_i^2\right) + \sum_{j=1}^{N} EM_j^{(1)}. \tag{4.4}$$

The first term in the right-hand side of (4.4) represents the mean amount of work that arrived at all queues *during the switching intervals* after the last visit of the server to those queues. Note that $s^{(2)}/2s - 1/2$ represents the mean total past switching time from the departure of the server from an arbitrary queue to the present random switching epoch. This interpretation explains why only s and $s^{(2)}$ occur, and no moments of individual switch-over times. The second term reflects the interaction between queues; it represents the mean amount of work that arrived at queues, after the last visit of S, during the subsequent service periods of other queues. Its most natural representation is perhaps [cf.(1.10)]

$$\frac{1}{2}\sum\sum_{h\neq k} \rho_k EV_h.$$

Finally, $\sum_{j=1}^{N} EM_j^{(1)}$ represents the mean amount of work that arrived at queues during the last service periods of those queues, but that was not handled by S at those service periods. From (4.2) and (4.4)

$$\sum_{i=1}^{N} \rho_i EW_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho + \rho\frac{s^{(2)}}{2s} - \frac{1}{2}\rho + \frac{s}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{N}\rho_i^2\right) + \sum_{i=1}^{N} EM_i^{(1)}. \tag{4.5}$$

Note that the form of formula (4.5) is still independent of the service strategies at the various queues; only the $EM_i^{(1)}$ depend on the choice of service strategies.

The $EM_i^{(1)}$ are readily found for an exhaustive or gated strategy at Q_i:

Q_i exhaustive:

$$EM_i^{(1)} = 0. \tag{4.6}$$

Q_i gated [cf.(1.10)]:

$$EM_i^{(1)} = \rho_i EV_i = \rho_i^2 \frac{s}{1-\rho}. \tag{4.7}$$

For the 1-limited strategy somewhat more work is required. At a departure epoch of S from Q_i, S has just completed one service of a message with probability $\lambda_i s/(1-\rho)$, and no service with probability $1 - \lambda_i s/(1-\rho)$. Hence, with ET_i the amount of work left behind at a departure epoch of a type i message

$$EM_i^{(1)} = \frac{\lambda_i s}{1-\rho} ET_i. \tag{4.8}$$

To determine ET_i, we calculate the mean number of packets left behind by a departing type i message. Let $W_i(z)$ be the z transform for the waiting time of an arbitrarily chosen type i message (the tagged message); $EW_i = W_i^{(1)}(1)$. Note that the messages left behind at station Q_i when the service of the tagged message has been completed are those which arrived during the sojourn time of the tagged message, and those which arrived in the same slot as the tagged message but were placed behind the tagged message (the sojourn time is counted from the beginning of the slot next to the one in which the arrival took place). The z transform $Q_i(z)$ for the number of messages who arrived during the sojourn time of the tagged message is given by

$$Q_i(z) = W_i(A_i(z))B_i(A_i(z)). \tag{4.9}$$

$\tilde{Q}_i(z)$, the z transform for the number of messages which arrived in the same slot as the tagged message, but were placed behind the tagged message, is given by the backward

recurrence time transform

$$\tilde{Q}_i(z)=\frac{1-A_i(z)}{\lambda_i(1-z)}. \tag{4.10}$$

These numbers of messages are not independent, but we can still determine the first moment of the sum, i.e.,

$$Q_i^{(1)}(1)+\tilde{Q}_i^{(1)}(1) \tag{4.11}$$

where

$$Q_i^{(1)}(1)=\lambda_i(EW_i+\beta_i) \tag{4.12}$$

$$\tilde{Q}_i^{(1)}(1)=\frac{\lambda_i^{(2)}-\lambda_i}{2\lambda_i}. \tag{4.13}$$

And so ET_i, the mean amount of work left behind in Q_i at a departure epoch of a type i message, equals

$$ET_i=\rho_iEW_i+\rho_i\beta_i+\frac{\lambda_i^{(2)}-\lambda_i}{2\lambda_i}\beta_i. \tag{4.14}$$

From (4.8) and (4.14), we obtain the following.

For Q_i 1-limited,

$$EM_i^{(1)}=\frac{\lambda_i s}{1-\rho}\rho_iEW_i+\rho_i^2\frac{s}{1-\rho}+\frac{\rho_i s}{1-\rho}\frac{\lambda_i^{(2)}-\lambda_i}{2\lambda_i}. \tag{4.15}$$

Finally, we consider semiexhaustive service. With the above definition of ET_i, (4.14) again holds. Denote by U_i the number of messages in Q_i at an arrival epoch of S at Q_i. Due to the structure of the SE strategy, we can also write

$$ET_i=\beta_iE[U_i-1\,|\,U_i\geq 1]+\left[\frac{\lambda_i^2\beta_i^{(2)}}{2(1-\rho_i)}+\frac{(\lambda_i^{(2)}-\lambda_i^2-\lambda_i)\beta_i}{2(1-\rho_i)}+\rho_i+\frac{\lambda_i^{(2)}-\lambda_i}{2\lambda_i}\right]\beta_i. \tag{4.16}$$

Note that the second term in the right-hand side represents the amount of work left behind by a departing message in a discrete-time $M/G/1$ queue with $A_i(z)$ and $B_i(z)$, respectively, the z transform of the number of message arrivals per slot and the number of packets per message; the first three terms between square brackets represent the mean number of messages that have arrived during the sojourn time of the departing message (cf. (2.2) and Little's formula), and the fourth term is the mean number of messages that have arrived in the same slot as this message, but were placed behind it, cf. (4.13). Subsequently, express $EM_i^{(1)}$ in the first term in the right-hand side of (4.16)

$$EM_i^{(1)}=\beta_iE[\max(0,\,U_i-1)]=\beta_iE[U_i-1\,|\,U_i\geq 1]\Pr\{U_i\geq 1\}. \tag{4.17}$$

Because the mean visit time of S at Q_i during a cycle, when positive, equals $\beta_i/(1-\rho_i)$ (the mean busy period of a discrete-time $M/G/1$ system with mean number of arrivals per slot λ_i and mean number of packets per message β_i), we have

$$EV_i=\frac{\rho_i s}{1-\rho}=\Pr\{U_i\geq 1\}\frac{\beta_i}{1-\rho_i} \tag{4.18}$$

so

$$\Pr\{U_i\geq 1\}=\frac{\lambda_i s(1-\rho_i)}{1-\rho}. \tag{4.19}$$

Combining (4.14), (4.16), (4.17), and (4.19),

$$\rho_iEW_i+\rho_i\beta_i=\frac{EM_i^{(1)}}{\lambda_i s\dfrac{1-\rho_i}{1-\rho}}+\left[\frac{\lambda_i^2\beta_i^{(2)}}{2(1-\rho_i)}+\frac{\lambda_i^{(2)}-\lambda_i^2-\lambda_i}{2\lambda_i(1-\rho_i)}\rho_i+\rho_i\right]\beta_i. \tag{4.20}$$

And so we have

$$EM_i^{(1)}=\rho_i\frac{\lambda_i s(1-\rho_i)}{1-\rho}EW_i-\frac{\lambda_i s(1-\rho_i)}{1-\rho}\cdot\left[\frac{\lambda_i\beta_i^{(2)}}{2(1-\rho_i)}\rho_i+\frac{(\lambda_i^{(2)}-\lambda_i^2-\lambda_i)}{2\lambda_i(1-\rho_i)}\rho_i\beta_i\right]. \tag{4.21}$$

Combining (4.5) and the four expressions for $EM_i^{(1)}$ in the cases of E, G, $1L$ and SE service strategy at Q_i, respectively, we have proved our main result.

Theorem 2

Consider an ergodic cyclic-service system with one server and mixed service strategies as described in Section I. Denote by

e: the group of E(xhaustive) queues,
g: the group of G(ated) queues,
$1l$: the group of $1L$(imited) queues, and
se: the group of S(emi) E(xhaustive) queues.

Then,

$$\begin{aligned}&\sum_{i\in e}\rho_iEW_i+\sum_{i\in g}\rho_iEW_i+\sum_{i\in 1l}\rho_i\left[1-\frac{\lambda_i s}{1-\rho}\right]EW_i+\sum_{i\in se}\rho_i\left[1-\frac{\lambda_i s(1-\rho_i)}{1-\rho}\right]EW_i\\&=\frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho+\frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho+\rho\frac{s^{(2)}}{2s}-\frac{1}{2}\rho+\frac{s}{2(1-\rho)}\left[\rho^2-\sum_{\forall i}\rho_i^2\right]\\&\quad+\frac{s}{(1-\rho)}\sum_{i\in g,1l}\rho_i^2-\frac{s}{2(1-\rho)}\sum_{i\in se}\lambda_i^2\beta_i^{(2)}\rho_i+\frac{s}{(1-\rho)}\sum_{i\in 1l}\frac{\lambda_i^{(2)}-\lambda_i}{2\lambda_i}\rho_i\\&\quad-\frac{s}{2(1-\rho)}\sum_{i\in se}(\lambda_i^{(2)}-\lambda_i^2-\lambda_i)\beta_i\rho_i.\end{aligned} \tag{4.22}$$

Remark 6

The case of $N=1$ queue yields expressions for mean waiting times in discrete-time $M/G/1$ queues with some form of server vacations. In the completely symmetric case with all queues having identical characteristics and the same exhaustive (gated, 1-limited) service strategy, formula (4.22) reduces to formula (3.63b) [respectively, (5.23), (6.60)] of [11].

Remark 7

If we assume Poisson arrivals in (4.22) (and hence take $\lambda_i^{(2)}$

$= \lambda_i^2 + \lambda_i$), we obtain the following relation for the weighted sum of the mean waiting times.

$$\sum_{i\in e}\rho_i EW_i+\sum_{i\in g}\rho_i EW_i+\sum_{i\in 1/}\rho_i\left[1-\frac{\lambda_i s}{1-\rho}\right]EW_i + \sum_{i\in se}\rho_i\left[1-\frac{\lambda_i s(1-\rho_i)}{1-\rho}\right]EW_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho+\rho\frac{s^{(2)}}{2s}-\frac{1}{2}\rho+\frac{s}{2(1-\rho)} \cdot\left[\rho^2-\sum_{\forall i}\rho_i^2\right]+\frac{s}{(1-\rho)}\sum_{i\in g,1/}\rho_i^2 -\frac{s}{2(1-\rho)}\sum_{i\in se}\lambda_i^2\beta_i^{(2)}\rho_i+\frac{s}{2(1-\rho)}\sum_{i\in 1/}\lambda_i\rho_i. \tag{4.23}$$

V. Relation to the Continuous-Time Case

In the present paper, we have expressed all quantities involved, including waiting times, in slots with the slot length equal to the time unit. If, instead, we assume a slot to be of length Δ we are able, by taking the limit $\Delta \to 0$, to pass the results over to continuous time.

First, we express the arrival process in messages per time unit. Recall that the z transform of the number of message arrivals at Q_i in a slot is given by $A_i(z)$, with first and second moment λ_i and $\lambda_i^{(2)}$, respectively. Denote by $\tilde{A}_i(z)$ the number of message arrivals at Q_i per time unit. Then

$$\tilde{A}_i(z)=[A_i(z)]^{1/\Delta} \tag{5.1}$$

($1/\Delta$ is the number of slots per time unit). From (5.1), we find

$$\tilde{\lambda}_i=\frac{\lambda_i}{\Delta}, \qquad \tilde{\lambda}_i^{(2)}=\frac{\lambda_i^{(2)}}{\Delta}+\frac{1}{\Delta}\left(\frac{1}{\Delta}-1\right)\lambda_i^2. \tag{5.2}$$

For the service (switching) process let $\tilde{\beta}_i$, $\tilde{\beta}_i^{(2)}(\tilde{s}_i, \tilde{s}_i^{(2)})$ denote the first and second moment, respectively, of the service (switching) time expressed in time units. It may be easily seen that

$$\tilde{\beta}_i=\beta_i\Delta, \qquad \tilde{\beta}_i^{(2)}=\beta_i^{(2)}\Delta^2;$$
$$\tilde{s}_i=s_i\Delta, \qquad \tilde{s}_i^{(2)}=s_i^{(2)}\Delta^2. \tag{5.3}$$

Similarly, cf. (1.2), (1.4),

$$\tilde{\lambda} := \sum_{i=1}^{N}\tilde{\lambda}_i=\frac{\lambda}{\Delta}, \qquad \tilde{\lambda}^{(2)} := \tilde{\lambda}^2+\sum_{i=1}^{N}(\tilde{\lambda}_i^{(2)}-\tilde{\lambda}_i^2) \tag{5.4}$$

hence,

$$\tilde{\lambda}^{(2)}-\tilde{\lambda}^2-\tilde{\lambda}=\frac{1}{\Delta}(\lambda^{(2)}-\lambda^2-\lambda);$$

furthermore,

$$\tilde{\beta} := \sum_{i=1}^{N}\frac{\tilde{\lambda}_i}{\tilde{\lambda}}\tilde{\beta}_i=\beta\Delta, \qquad \tilde{\beta}^{(2)} := \sum_{i=1}^{N}\frac{\tilde{\lambda}_i}{\tilde{\lambda}}\tilde{\beta}_i^{(2)}=\beta^{(2)}\Delta^2. \tag{5.5}$$

For the mean waiting time in time units $E\tilde{W}_i$, we have

$$E\tilde{W}_i=EW_i\Delta. \tag{5.6}$$

Of course $\rho_i = \lambda_i\beta_i = \tilde{\lambda}_i\tilde{\beta}_i$. We can now express (4.22) in time units. With the slot length equal to Δ, we obtain from (4.22) and (5.2)–(5.6):

$$\sum_{i\in e}\rho_i E\tilde{W}_i\frac{1}{\Delta}+\sum_{i\in g}\rho_i E\tilde{W}_i\frac{1}{\Delta}+\sum_{i\in 1/}\rho_i\left[1-\frac{\tilde{\lambda}_i\tilde{s}}{1-\rho}\right] \cdot E\tilde{W}_i\frac{1}{\Delta}+\sum_{i\in se}\rho_i\left[1-\frac{\tilde{\lambda}_i\tilde{s}(1-\rho_i)}{1-\rho}\right]E\tilde{W}_i\frac{1}{\Delta} =\frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho\frac{1}{\Delta}+\frac{(\tilde{\lambda}^{(2)}-\tilde{\lambda}^2-\tilde{\lambda})\tilde{\beta}}{2\tilde{\lambda}(1-\rho)}\rho\frac{1}{\Delta} +\rho\frac{\tilde{s}^{(2)}}{2\tilde{s}}\frac{1}{\Delta}-\frac{1}{2}\rho+\frac{\tilde{s}}{2(1-\rho)}\left[\rho^2-\sum_{\forall i}\rho_i^2\right]\frac{1}{\Delta} +\frac{\tilde{s}}{(1-\rho)}\sum_{i\in g,1/}\rho_i^2\frac{1}{\Delta}-\frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}\tilde{\lambda}_i^2\tilde{\beta}_i^{(2)}\rho_i\frac{1}{\Delta} +\frac{\tilde{s}}{(1-\rho)}\sum_{i\in 1/}\frac{\tilde{\lambda}_i^{(2)}-(1-\Delta)\tilde{\lambda}_i^2-\tilde{\lambda}_i}{2\tilde{\lambda}_i}\rho_i\frac{1}{\Delta} -\frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}(\tilde{\lambda}_i^{(2)}-\tilde{\lambda}_i^2-\tilde{\lambda}_i)\tilde{\beta}_i\rho_i\frac{1}{\Delta}. \tag{5.7}$$

In (5.7), we can take the limit for $\Delta \to 0$ by multiplying the left- and right-hand side with Δ and substituting $\Delta = 0$. If we do so, we obtain

$$\sum_{i\in e}\rho_i E\tilde{W}_i+\sum_{i\in g}\rho_i E\tilde{W}_i+\sum_{i\in 1/}\rho_i\left[1-\frac{\tilde{\lambda}_i\tilde{s}}{1-\rho}\right]E\tilde{W}_i + \sum_{i\in se}\rho_i\left[1-\frac{\tilde{\lambda}_i\tilde{s}(1-\rho_i)}{1-\rho}\right]E\tilde{W}_i =\frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho+\frac{(\tilde{\lambda}^{(2)}-\tilde{\lambda}^2-\tilde{\lambda})\tilde{\beta}}{2\tilde{\lambda}(1-\rho)}\rho+\rho\frac{\tilde{s}^{(2)}}{2\tilde{s}} +\frac{\tilde{s}}{2(1-\rho)}\left[\rho^2-\sum_{\forall i}\rho_i^2\right] +\frac{\tilde{s}}{(1-\rho)}\sum_{i\in g,1/}\rho_i^2-\frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}\tilde{\lambda}_i^2\tilde{\beta}_i^{(2)}\rho_i +\frac{\tilde{s}}{(1-\rho)}\sum_{i\in 1/}\frac{\tilde{\lambda}_i^{(2)}-\tilde{\lambda}_i^2-\tilde{\lambda}_i}{2\tilde{\lambda}_i}\rho_i -\frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}(\tilde{\lambda}_i^{(2)}-\tilde{\lambda}_i^2-\tilde{\lambda}_i)\beta_i\rho_i. \tag{5.8}$$

At this point, some remarks are in order. To obtain formula (5.8), it is not necessary to specify precisely how the above limit $\Delta \to 0$ is taken. However, the structure of the resulting arrival process does depend on it. Let us take a closer look at the arrival process. As has been noted in Section I, the message arrival process at Q_i is a Bernoulli process with batch arrivals. We have a Bernoulli arrival process in the sense that

$$\Pr\{\text{ type } i \text{ batch arrives in a slot }\} = 1 - A_i(0)$$

$$\Pr\{\text{ type } i \text{ batch does not arrive in a slot }\} = A_i(0).$$

With respect to the batch arrivals, let $G_i(z)$ denote the z transform of the size of a type i batch. Then, we can write

$A_i(z)$ as

$$A_i(z) = A_i(0) + [1 - A_i(0)]G_i(z) \tag{5.9}$$

and hence, with (5.1):

$$\tilde{A}_i(z) = (A_i(0) + [1 - A_i(0)]G_i(z))^{1/\Delta} = \left(1 - \left[\frac{1 - A_i(0)}{\Delta} - \frac{1 - A_i(0)}{\Delta} G_i(z)\right]\Delta\right)^{1/\Delta}. \tag{5.10}$$

Let

$$\gamma_i := \frac{1 - A_i(0)}{\Delta}.$$

γ_i denotes the arrival intensity of type i batches. Note that γ_i is also equal to $\tilde{\lambda}_i/G_i^{(1)}(1)$. Now, if in (5.10) we let $\Delta \to 0$ in such a way that γ_i remains constant, the z transform for the number of message arrivals per time unit at Q_i becomes

$$\tilde{A}_i(z) = e^{\gamma_i(G_i(z)-1)} \tag{5.11}$$

which is the z transform of a compound Poisson process. If we take $G_i(z) = z$ (single arrivals), we obtain the z transform of the "ordinary" Poisson process; in this case $\tilde{\lambda}_i^{(2)} = \tilde{\lambda}_i^2 + \tilde{\lambda}_i$, and (5.8) reduces to the pseudoconservation law in continuous time, formula (3.22), derived in [1]:

$$\sum_{i\in e} \rho_i E\tilde{W}_i + \sum_{i\in g} \rho_i E\tilde{W}_i + \sum_{i\in lf} \rho_i \left[1 - \frac{\tilde{\lambda}_i \tilde{s}}{1-\rho}\right] E\tilde{W}_i + \sum_{i\in se} \rho_i \left[1 - \frac{\tilde{\lambda}_i \tilde{s}(1-\rho_i)}{1-\rho}\right] E\tilde{W}_i = \frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho + \rho\frac{\tilde{s}^{(2)}}{2\tilde{s}} + \frac{\tilde{s}}{2(1-\rho)}\left[\rho^2 - \sum_{\forall i} \rho_i^2\right] + \frac{\tilde{s}}{(1-\rho)} \sum_{i\in g,lf} \rho_i^2 - \frac{\tilde{s}}{2(1-\rho)} \sum_{i\in se} \tilde{\lambda}_i^2 \tilde{\beta}_i^{(2)} \rho_i. \tag{5.12}$$

Formula (5.8) presents a slight extension to this result, in that the message arrival process at Q_i is allowed to be a Poisson process with *batch* arrivals.

VI. Discussion

In this paper, we have derived a stochastic decomposition for the amount of work in discrete-time cyclic-service systems with mixed service strategies. This decomposition is analogous to one that has recently been proved in [1] for the continuous-time case. The work decomposition result is used to derive an exact expression for a weighted sum of mean waiting times—a so-called "pseudoconservation law." This pseudoconservation law, stated in Theorem 2, forms a natural extension of the $M/G/1$ conservation law in discrete time as stated in formula (2.5). Its derivation clearly exposes the meaning of all terms. Theorem 2 presents a remarkably simple result, in view of the fact that expressions for the individual mean waiting times (in continuous- or discrete-time) are in general either not known or very complicated.

In [2], it has been shown, for the 1-limited case, how pseudoconservation laws can be used to obtain simple, yet quite accurate, approximations for individual mean waiting times. In a future report, this approximation will be extended to more general cyclic-service models with mixed service strategies.

Finally, we should like to stress the fact that Theorem 1 and decomposition (4.5) can be proved for more general models than the one under consideration. In particular, other service strategies may also be included—and for each extension the challenge is to determine $\sum_{i=1}^{N} EM_i^{(1)}$, the sum of the mean amounts of work left behind by the server in the queues.

References

[1] O. J. Boxma and W. P. Groenendijk, "Pseudo-conservation laws in cyclic-service systems," Rep. OS-R8606, 1986, Centre for Mathematics and Computer Science, Amsterdam; also, *J. Appl. Prob.*, vol. 24, 1987.

[2] O. J. Boxma and B. Meister, "Waiting-time approximations for cyclic-service systems with switch-over times," *Perform. Eval. Rev.*, vol. 14, pp. 254–262, 1986.

[3] M. J. Ferguson and Y. J. Aminetzah, "Exact results for nonsymmetric token ring systems," *IEEE Trans. Commun.*, vol. COM-33, pp. 223–231, 1985.

[4] S. W. Fuhrmann and R. B. Cooper, "Stochastic decompositions in the $M/G/1$ queue with generalized vacations," *Oper. Res.*, vol. 33, pp. 1117–1129, 1984.

[5] L. Kleinrock, *Queueing Systems, Vol. 2.* New York: Wiley, 1976.

[6] H. Kobayashi and A. G. Konheim, "Queueing models for computer communications systems analysis," *IEEE Trans. Commun.*, vol. COM-25, pp. 2–29, 1977.

[7] A. G. Konheim and B. Meister, "Waiting lines and times in a system with polling," *J. Ass. Comput. Mach.*, vol. 21, pp. 470–490, 1974.

[8] H. Levy and L. Kleinrock, "A queue with starter and a queue with vacations: Delay analysis by decomposition," *Oper. Res.*, vol. 34, pp. 426–436, 1986.

[9] I. Rubin and L. F. DeMoraes, "Message delay analysis for polling and token multiple-access schemes for local communication networks," *IEEE J. Select. Areas Commun.*, vol. SAC-1, pp. 935–947, 1983.

[10] G. B. Swartz, "Polling in a loop system," *J. Ass. Comput. Mach.*, vol. 27, pp. 42–59, 1980.

[11] H. Takagi, *Analysis of Polling Systems.* Cambridge, MA: M.I.T. Press, 1986.

[12] ——, "A survey of queueing analysis of polling models," in *Proc. Third Int. Conf. Data Commun. Syst. Perform.*, Rio de Janeiro, Brazil, 1987.

[13] K. S. Watson, "Performance evaluation of cyclic service strategies—a survey," in *Performance '84*, E. Gelenbe, Ed. Amsterdam: North-Holland, 1984, pp. 521–533.

[14] R. W. Wolff, "Poisson arrivals see time averages," *Oper. Res.*, vol. 30, pp. 223–231, 1982.

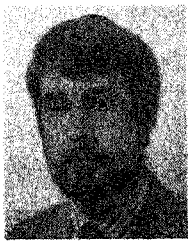
★

Onno J. Boxma received the Master's degree from Delft Technological University, The Netherlands, in 1974, and the Ph.D. from the University of Utrecht, The Netherlands, in 1977, both in mathematics.

During 1978–1979, he was an IBM Postdoctoral Fellow in Yorktown Heights, NY. In 1984, he spent three months at the IBM Zürich Research Laboratory. Since August 1985, he has been with the Centre for Mathematics and Computer Science (CWI) where he leads a small research group in queueing theory and performance evaluation. Since August 1987 he has been Professor of Operations Research at Tilburg University, with a 0.2 assignment. His research interests include queueing theory, computer performance, and stochastic scheduling.

He is a member of IFIP W.G.7.3 and serves on the Editorial Board of the journals *Queueing Systems: Theory and Applications* and *Performance Evaluation*.

★



Wim P. Groenendijk was born in Utrecht, The Netherlands, in December 1961. He received the Master's degree from the University of Utrecht, The Netherlands, in January 1986. Since February 1986, he has been working towards his Ph.D. degree at the Centre for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands.

His research interests include queueing theory and applications, with specific emphasis on performance evaluation of computer-communication systems.