



**Centrum voor Wiskunde en Informatica**  
Centre for Mathematics and Computer Science

---

O.J. Boxma, W.P. Groenendijk

Waiting times in discrete-time cyclic-service systems

Department of Operations Research and System Theory

Report OS-R8707

March

---

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

# Waiting Times in Discrete-Time Cyclic-Service Systems

O.J. Boxma, W.P. Groenendijk

Centre for Mathematics and Computer Science  
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

This paper considers single-server, multi-queue systems with cyclic service in discrete time. Non-zero 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 & Groenendijk [1] where the continuous-time case is treated.

1980 Mathematics Subject Classification: 60K25, 68M20.

Key Words & Phrases: discrete-time queueing system, cyclic service, (pseudo-) conservation law.

Note: This report will be submitted for publication elsewhere.

## 1. 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, multi-queue 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, multi-queue 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 *pseudo-conservation laws* (Watson [12], Ferguson & 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 pseudo-conservation laws have been generalized by allowing a mixture of different service strategies at different queues. The proof of the resulting unified pseudo-conservation 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 [4] to models with switch-over times. The decomposition also allows a simple probabilistic interpretation of the various terms of the unified pseudo-conservation 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 [11]. 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 & Meister [6], Swartz [9], Rubin & DeMoraes [8] and Takagi [10]; but the bulk of the literature in this area is devoted to continuous-time systems. See Takagi [10] for an extensive survey of cyclic-service systems, and Takagi [11] 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 2 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 3. The main result of the paper, the *discrete-time pseudo-conservation law for mean waiting times*, is proved in Section 4. In Section 5 the relation between the obtained discrete-time results and results for the continuous-time case is presented. Section 6 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)  $Q_1, \dots, Q_N$ , 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  $j$ -th slot.

#### Arrival process

Let,

$x_i(j) :=$  number of messages arriving at station  $i$  in the  $j$ -th 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  $x_i(j)$ ,  $j=1,2,\dots$  are assumed to be independent, identically distributed random variables with  $z$ -transform, first and second moment:

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

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

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

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

Let

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

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

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

Further introduce:

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

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

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

The total offered traffic,  $\rho$ , 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 indices  $i = 1, 2, \dots, N$  ("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  $Q_i$ :

- I. Exhaustive service (E):  $S$  serves type- $i$  messages until  $Q_i$  is empty.
- II. Gated service (G):  $S$  serves exactly those type- $i$  messages present upon his arrival at  $Q_i$  (a gate closes upon his arrival).
- III. 1-Limited service (1L):  $S$  serves at most one type- $i$  message (the term nonexhaustive has often been used for this strategy).
- IV. Semi-exhaustive 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., semi-exhaustive at  $Q_1$ , exhaustive at  $Q_2$  and  $Q_4$ , 1-limited at  $Q_3$  and gated at  $Q_5, \dots, Q_N$ ). 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  $i$ -th 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 of the total switch-over time during a cycle of the server,  $s$ , 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:

- $X_i$ : 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 queuing 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  $i$ , and from a balancing argument it follows that the mean cycle time equals  $EC$  with

$$EC = \frac{s}{1-\rho}. \quad (1.8)$$

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,

$$\lambda_i EC = \frac{EV_i}{\beta_i}, \quad (1.9)$$

and hence, from (1.8),

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

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 semi-exhaustive service, we have the following additional condition:

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

This reflects the fact that, for semi-exhaustive 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 an 1L or SE strategy.

## 2. 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  $\prod_1^N A_i(0)$  and a train does arrive with probability  $1 - \prod_1^N A_i(0)$ . An arbitrarily chosen message in this train poses a service request whose  $z$ -transform is the mixture  $\sum_1^N \frac{\lambda_i}{\lambda} B_i(z)$ .

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:

$$V_c \stackrel{D}{=} V_{M/G/1}. \quad (2.1)$$

According to Kobayashi & Konheim [5], the mean number of messages in the corresponding system at an arbitrary epoch is given by:

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

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  $\rho$ ; the residual service time of the message in service is  $\frac{\beta^{(2)}}{2\beta} + \frac{1}{2}$ . Hence

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

### REMARK 3

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

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

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

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:

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

We propose to call Equation (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.

### 3. 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 and 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 1. Suppose the system is ergodic and stationary. Then the amount of work in this system at an arbitrary epoch,  $V_c$ , is distributed as the sum of the amount of work in the "corresponding" M/G/1 system at an arbitrary epoch,  $V_{M/G/1}$ , and the amount of work,  $Y$ , in the cyclic-service system at an arbitrary epoch in a switching interval. In other words:

$$V_c \stackrel{D}{=} V_{M/G/1} + Y, \quad (3.1)$$

where  $\stackrel{D}{=}$  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. 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) The LCFS discipline has the following property: if a train arrives during a switch-over period, then the work found upon his arrival is untouched by the server until this train, and trains arriving during its service, etc., have been fully served. This allows the decomposition  $V_c^{LCFS} = V_{M/G/1}^{LCFS} + Y$ .

We refer to [1] for details; the only changes required for the discrete-time case are that,

- (a) instead of considering customers we consider *trains*, consisting of the batch of messages that arrive in one slot;
- (b) instead of using the PASTA-property [13], we now use its discrete-time equivalent, 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.

#### REMARK 4

In Levy & Kleinrock [7],  $Y$  represents "the additional delay due to the presence of the starter".

### 4. THE PSEUDO-CONSERVATION LAW

As a consequence of Theorem 1:

$$EV_c = EV_{M/G/1} + EY, \quad (4.1)$$

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

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



We now derive an expression for  $EY$ , thus obtaining a very general pseudo-conservation 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,  $EY = \sum_{i=1}^N \frac{s_i}{s} EY_i$ . 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.  $\rho\{\frac{s_i^{(2)}}{2s_i} - \frac{1}{2}\}$ : 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:

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

and hence for  $EY$ :

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

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  $\frac{s^{(2)}}{2s} - \frac{1}{2}$  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))

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

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):

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

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:

$$EM_i^{(1)} = 0; \quad (4.6)$$

$Q_i$  gated (cf. (1.10)):

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

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  $\frac{\lambda_i s}{1-\rho}$ , and no service with probability  $1 - \frac{\lambda_i s}{1-\rho}$ .

Hence, with  $ET_i$  the amount of work left behind at a departure epoch of a type- $i$  message,

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

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:

$$Q_i(z) = W_i(A_i(z))B_i(A_i(z)); \quad (4.9)$$

$\tilde{Q}_i(z)$ , 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:

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

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

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

where

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

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

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

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

From (4.8) and (4.14) we obtain:

For  $Q_i$  1-limited:

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

Finally, we consider semi-exhaustive 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

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

(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, see also (2.2). Subsequently express  $EM_i^{(1)}$  in the first term in the right-hand side of (4.16):

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

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  $\lambda_i$  and mean number of packets per message  $\beta_i$ ), we have

$$EV_i = \frac{\rho_i s}{1 - \rho} = Pr\{U_i \geq 1\} \frac{\beta_i}{1 - \rho_i}, \quad (4.18)$$

so

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

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

$$\rho_i EW_i + \rho_i \beta_i + \frac{\lambda_i^{(2)} - \lambda_i}{2\lambda_i} \beta_i = \frac{EM_i^{(1)}}{\lambda_i s \frac{1 - \rho_i}{1 - \rho}} + \left[ \frac{\lambda_i^2 \beta_i^{(2)}}{2(1 - \rho_i)} + \frac{\lambda_i^{(2)} - \lambda_i^2 - \lambda_i}{2\lambda_i(1 - \rho_i)} \rho_i + \rho_i \right] \beta_i, \quad (4.20)$$

and so we have

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

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 1. Denote by

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

Then

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

## REMARK 5

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) (resp. (5.23),(6.60)) of [10].

## REMARK 6

If we assume Poisson arrivals in (4.22) (which corresponds to taking  $\lambda_i^{(2)} = \lambda_i^2 + \lambda_i$ ) we obtain the following relation for the weighted sum of the mean waiting times.

$$\begin{aligned} & \sum_{i \in e} \rho_i E W_i + \sum_{i \in g} \rho_i E W_i + \sum_{i \in ll} \rho_i [1 - \frac{\lambda_i s}{1-\rho}] E W_i + \sum_{i \in se} \rho_i [1 - \frac{\lambda_i s (1-\rho_i)}{1-\rho}] E W_i \\ &= \frac{\lambda \beta^{(2)}}{2(1-\rho)} \rho + \rho \frac{s^{(2)}}{2s} - \frac{1}{2} \rho + \frac{s}{2(1-\rho)} [\rho^2 - \sum_{vi} \rho_i^2] + \frac{s}{(1-\rho)} \sum_{i \in g, ll} \rho_i^2 \\ & - \frac{s}{2(1-\rho)} \sum_{i \in se} \lambda_i^2 \beta_i^{(2)} \rho_i + \frac{s}{2(1-\rho)} \sum_{i \in ll} \lambda_i \rho_i + \frac{s}{(1-\rho)} \sum_{i \in se} \lambda_i \rho_i (1-\rho_i). \end{aligned} \quad (4.23)$$

## 5. 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  $\Delta$  we are able, by taking the limit  $\Delta \rightarrow 0$ , 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  $\lambda_i$  and  $\lambda_i^{(2)}$  respectively. Denote by  $\tilde{A}_i(z)$  the number of message arrivals at  $Q_i$  per time unit. Then

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

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

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

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,

$$\begin{aligned} \tilde{\beta}_i &= \beta_i \Delta, & \tilde{\beta}_i^{(2)} &= \beta_i^{(2)} \Delta^2; \\ \tilde{s}_i &= s_i \Delta, & \tilde{s}_i^{(2)} &= s_i^{(2)} \Delta^2. \end{aligned} \quad (5.3)$$

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

$$E\tilde{W}_i = E W_i \Delta. \quad (5.4)$$

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  $\Delta$  we obtain from (4.22) and (5.2)-(5.4):

$$\begin{aligned} & \sum_{i \in e} \rho_i E \tilde{W}_i \frac{1}{\Delta} + \sum_{i \in g} \rho_i E \tilde{W}_i \frac{1}{\Delta} + \sum_{i \in ll} \rho_i [1 - \frac{\tilde{\lambda}_i \tilde{s}}{1-\rho}] E \tilde{W}_i \frac{1}{\Delta} + \sum_{i \in se} \rho_i [1 - \frac{\tilde{\lambda}_i \tilde{s} (1-\rho_i)}{1-\rho}] E \tilde{W}_i \frac{1}{\Delta} \\ &= \frac{\tilde{\lambda} \tilde{\beta}^{(2)}}{2(1-\rho)} \rho \frac{1}{\Delta} + \frac{(\tilde{\lambda}^{(2)} - \tilde{\lambda}^2 - \tilde{\lambda}) \tilde{\beta}}{2\tilde{\lambda}(1-\rho)} \rho \frac{1}{\Delta} + \rho \frac{\tilde{s}^{(2)}}{2\tilde{s}} \frac{1}{\Delta} - \frac{1}{2} \rho + \frac{\tilde{s}}{2(1-\rho)} [\rho^2 - \sum_{vi} \rho_i^2] \frac{1}{\Delta} + \end{aligned}$$

$$\begin{aligned}
& + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in g, ll} \rho_i^2 \frac{1}{\Delta} - \frac{\tilde{s}}{2(1-\rho)} \sum_{i \in se} \tilde{\lambda}_i^2 \tilde{\beta}_i^{(2)} \rho_i \frac{1}{\Delta} + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in ll} \frac{\tilde{\lambda}_i^{(2)} - (1-\Delta)\tilde{\lambda}_i^2 - \tilde{\lambda}_i}{2\tilde{\lambda}_i} \rho_i \frac{1}{\Delta} \\
& + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in se} \frac{\tilde{\lambda}_i^{(2)} - (1-\Delta)\tilde{\lambda}_i^2 - \tilde{\lambda}_i}{2\tilde{\lambda}_i} \rho_i (1-\rho_i) \frac{1}{\Delta} - \frac{\tilde{s}}{2(1-\rho)} \sum_{i \in se} (\tilde{\lambda}_i^{(2)} - \tilde{\lambda}_i^2 - \tilde{\lambda}_i) \tilde{\beta}_i \rho_i \frac{1}{\Delta}. \quad (5.5)
\end{aligned}$$

In (5.5) we can take the limit for  $\Delta \rightarrow 0$  by multiplying the left- and right-hand side with  $\Delta$  and substituting  $\Delta = 0$ . If we do so, we obtain

$$\begin{aligned}
& \sum_{i \in e} \rho_i E\tilde{W}_i + \sum_{i \in g} \rho_i E\tilde{W}_i + \sum_{i \in ll} \rho_i [1 - \frac{\tilde{\lambda}_i \tilde{s}}{1-\rho}] E\tilde{W}_i + \sum_{i \in se} \rho_i [1 - \frac{\tilde{\lambda}_i \tilde{s} (1-\rho_i)}{1-\rho}] E\tilde{W}_i \\
& = \frac{\tilde{\lambda} \tilde{\beta}^{(2)}}{2(1-\rho)} \rho + \frac{(\tilde{\lambda}^{(2)} - \tilde{\lambda}^2 - \tilde{\lambda}) \tilde{\beta}}{2\tilde{\lambda}(1-\rho)} \rho + \rho \frac{\tilde{s}^{(2)}}{2\tilde{s}} + \frac{\tilde{s}}{2(1-\rho)} [\rho^2 - \sum_{i \in ll} \rho_i^2] \\
& + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in g, ll} \rho_i^2 - \frac{\tilde{s}}{2(1-\rho)} \sum_{i \in se} \tilde{\lambda}_i^2 \tilde{\beta}_i^{(2)} \rho + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in ll} \frac{\tilde{\lambda}_i^{(2)} - \tilde{\lambda}_i^2 - \tilde{\lambda}_i}{2\tilde{\lambda}_i} \rho_i \\
& + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in se} \frac{\tilde{\lambda}_i^{(2)} - \tilde{\lambda}_i^2 - \tilde{\lambda}_i}{2\tilde{\lambda}_i} \rho_i (1-2\rho_i). \quad (5.6)
\end{aligned}$$

At this point some remarks are in order. To obtain Formula (5.6) it is not necessary to specify precisely how the above limit  $\Delta \rightarrow 0$  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 1, 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:

$$A_i(z) = A_i(0) + [1 - A_i(0)]G_i(z), \quad (5.7)$$

and hence, with (5.1):

$$\begin{aligned}
\tilde{A}_i(z) &= (A_i(0) + [1 - A_i(0)]G_i(z))^{1/\Delta} \\
&= (1 - [\frac{1 - A_i(0)}{\Delta} - \frac{1 - A_i(0)}{\Delta} G_i(z)]\Delta)^{1/\Delta}.
\end{aligned} \quad (5.8)$$

Let,

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

$\gamma_i$  denotes the arrival intensity of type- $i$  batches. Note that  $\gamma_i$  is also equal to  $\tilde{\lambda}_i / G_i'(1)$ . Now if in (5.8) we let  $\Delta \rightarrow 0$  in such a way that  $\gamma_i$  remains constant, the  $z$ -transform for the number of message arrivals per time unit at  $Q_i$  becomes:

$$\tilde{A}_i(z) = e^{\gamma_i(G_i(z)-1)}, \quad (5.9)$$

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 (5.6) reduces to the pseudo-conservation law in continuous time derived in [1]:

$$\sum_{i \in e} \rho_i E\tilde{W}_i + \sum_{i \in g} \rho_i E\tilde{W}_i + \sum_{i \in ll} \rho_i [1 - \frac{\tilde{\lambda}_i \tilde{s}}{1-\rho}] E\tilde{W}_i + \sum_{i \in se} \rho_i [1 - \frac{\tilde{\lambda}_i \tilde{s} (1-\rho_i)}{1-\rho}] E\tilde{W}_i =$$

$$\frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho + \rho\frac{\tilde{s}^{(2)}}{2s} + \frac{\tilde{s}}{2(1-\rho)}[\rho^2 - \sum_{vi} \rho_i^2] + \frac{\tilde{s}}{(1-\rho)} \sum_{i \in g, ll} \rho_i^2 - \frac{\tilde{s}}{2(1-\rho)} \sum_{i \in se} \tilde{\lambda}_i^2 \tilde{\beta}_i^{(2)} \rho_i. \quad (5.10)$$

Formula (5.6) 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.

## 6. 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 "pseudo-conservation law". This pseudo-conservation 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 pseudo-conservation 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] BOXMA, O. J., GROENENDIJK, W.P. (1986) Pseudo-conservation laws in cyclic-service systems. Report OS-R8606, Centre for Mathematics and Computer Science, Amsterdam. To appear in: *J. Appl. Prob.*, Vol. 24, 1987.
- [2] BOXMA, O. J., MEISTER, B. (1986) Waiting-time approximations for cyclic-service systems with switch-over times. *Performance Evaluation Review* 14, 254-262.
- [3] FERGUSON, M.J., AMINETZAH, Y.J. (1985) Exact results for nonsymmetric token ring systems. *IEEE Trans. Commun.*, Vol. COM-33, 223-231.
- [4] KLEINROCK, L. (1976) *Queueing Systems*, Vol. 2. Wiley, New York.
- [5] KOBAYASHI, H., KONHEIM, A.G. (1977) Queueing models for computer communications systems analysis. *IEEE Trans. Commun.*, Vol. COM-25, 2-29.
- [6] KONHEIM, A.G., MEISTER, B. (1974) Waiting lines and times in a system with polling. *J. Ass. Comput. Mach.*, 21, 470-490.
- [7] LEVY, H., KLEINROCK, L. (1986) A queue with starter and a queue with vacations: delay analysis by decomposition. *Oper. Res.* 34, 426-436.
- [8] RUBIN, I., DEMORAES, L.F. (1983) Message delay analysis for polling and token multiple-access schemes for local communication networks. *IEEE J. Select. Areas Commun.*, Vol. SAC-1, 935-947.
- [9] SWARTZ, G.B. (1980) Polling in a loop system. *J. Ass. Comput. Mach.*, 27, 42-59.
- [10] TAKAGI, H. (1986) *Analysis of Polling Systems*. The MIT Press, Cambridge, Massachusetts.
- [11] TAKAGI, H. (1987) A survey of queueing analysis of polling models. To appear in: *Proceedings of the Third International Conference on Data Communication Systems and their Performance*, Rio de Janeiro.
- [12] WATSON, K. S. (1984) Performance evaluation of cyclic service strategies - a survey. In: *Performance '84*, ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 521-533.
- [13] WOLFF, R. W. (1982) Poisson arrivals see time averages, *Oper. Res.* 30, 223-231.

ONTVANGEN 19 MAART 1987