J. Appl. Prob. **24**, 949–964 (1987)
Printed in Israel

PSEUDO-CONSERVATION LAWS IN CYCLIC-SERVICE SYSTEMS

O. J. BOXMA AND
W. P. GROENENDIJK,* *Centre for Mathematics and Computer Science, Amsterdam*

Abstract

This paper considers single-server, multi-queue systems with cyclic service. Non-zero switch-over times of the server between consecutive queues are assumed. A stochastic decomposition for the amount of work in such systems is obtained. This decomposition allows a short derivation of a 'pseudo-conservation law' for a weighted sum of the mean waiting times at the various queues. Thus several recently proved conservation laws are generalised and explained.

QUEUEING SYSTEM; SWITCH-OVER TIMES; MEAN WAITING TIME; CONSERVATION LAW

1. Introduction

The principle of work conservation has in the past proven to be very useful in the analysis of queueing systems with a non-FCFS service discipline. When no work is created or destroyed within the system, the amount of work present should not depend on the order of service — and hence should equal the amount of work in the 'corresponding' system with FCFS service discipline. If, moreover, the queueing discipline selects customers in a way that is independent of (any measure of) the service time, then the distribution of the number of customers in the system is also independent of the order of service [9]. But even if this is not the case, as in priority systems with different service requirements for different classes of customers, the principle of work conservation yields a useful expression for a weighted sum of the mean queue lengths; hence (by using Little's formula) a weighted sum of mean waiting times can be obtained.

Priority systems with switch-over times between different classes do not possess the work-conserving property, because the server is forced to be idle although work is present (introducing switch-over times can be interpreted as creating additional work within the system). A prime example of such systems

Received 26 June 1986; revision received 26 September 1986.
* Postal address: Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

is the single-server multi-queue model with cyclic service and switch-over times; one server visits a set of queues in a fixed cyclic order, taking some non-negligible time to switch between queues. This model has played an important role in the analysis of polling schemes; presently it is finding a new application in local area networks with a ring or bus topology, employing a medium access control protocol based on token passing. It is also useful for analysing, for example, congestion at traffic lights.

Because of the importance of this cyclic-service model, and the complexity of its mathematical analysis, the recent discovery of 'pseudo-conservation laws', expressions for a weighted sum of the mean waiting times at the various queues of the cyclic system [6], [14], has attracted considerable attention. Unfortunately, the derivation of these conservation laws was lengthy and cumbersome, and no satisfactory explanation for the occurrence of these laws was provided. The goal of the present paper is to generalise and unify the known conservation laws, and to explain why they should hold.

Let us first present a more detailed *model description*. The model under consideration consists of N queues $Q_1, \cdots, Q_N$; each queue has infinite capacity. Customers arrive at all queues according to independent Poisson processes with arrival intensities $\lambda_1, \cdots, \lambda_N$; the total arrival rate is given by

$$\Lambda := \sum_{i=1}^{N} \lambda_i.$$

Customers who arrive at Q_i are called type-i customers.

The queues are attended by a single server S who visits the queues in a fixed cyclic order: $Q_1, Q_2, \cdots, Q_N, Q_1, Q_2, \cdots$. The switch-over times of the server between the ith and $(i+1)$th queue are independent, identically distributed stochastic 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:

$$s := \sum_{i=1}^{N} s_i;$$

its second moment is denoted by $s^{(2)}$. The service times of type-i customers are independent, identically distributed stochastic variables; their distribution $B_i(\cdot)$ has first moment β_i and second moment $\beta_i^{(2)}$. We assume that the interarrival process, the service process and the switch-over process are mutually independent. The offered traffic at Q_i, ρ_i, is defined as

$$\rho_i := \lambda_i \beta_i, \qquad i = 1, \cdots, N.$$

The total offered traffic, ρ, is defined as

$$\rho := \sum_{i=1}^{N} \rho_i.$$

For the service strategies at the queues there are various possibilities, which differ in the number of customers who may be served in a queue during a visit of 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 customers). Otherwise, S acts as follows, depending on the service strategy at Q_i:

I. Exhaustive service (E): S serves type-i customers until Q_i is empty.

II. Gated service (G): S serves exactly those type-i customers present upon his arrival at Q_i (a gate closes upon his arrival).

III. Non-exhaustive service (NE): S serves only one type-i customer (the generalisation to 'service of at most k customers' has hardly been analysed, and will also not be considered here).

IV. Semi-exhaustive service (SE): S continues serving type-i customers until the number present is one less than the number present upon his arrival.

For detailed references and an extensive discussion of the E, G and NE strategies see for instance Takagi [13]. The SE discipline has recently been introduced by Takagi [12], who studies it in the case where all arrival rates, service-time and switch-over time distributions are the same for all queues. Boxma [2] contains a concise survey, with special emphasis on detailed mathematical studies of two-queue models.

In this paper we will allow mixed cyclic-service strategies (e.g., semi-exhaustive at Q_1, exhaustive at Q_2 and Q_4, non-exhaustive 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, as will be discussed in Section 4. In what follows the cyclic-service system under consideration will be assumed to be in equilibrium.

Below we state a few general, known, results for future reference. For any strictly cyclic service system we can define the cycle time $\mathbf{C}_i$ for Q_i as the time between two successive arrivals of S at Q_i. It is easily seen that $E\mathbf{C}_i$ is independent of i, and from a balancing argument it follows that the mean cycle time equals $E\mathbf{C}$ with

$$E\mathbf{C} = \frac{s}{1-\rho}. \tag{1.1}$$

Furthermore we can define the visit time $\mathbf{V}_i$ of S for Q_i as the time between the arrival of S at Q_i and his subsequent departure from that queue. Balancing the flow of type-i customers in and out of the system during a cycle shows that

$$\lambda_i E\mathbf{C} = \frac{E\mathbf{V}_i}{\beta_i}, \tag{1.2}$$

and hence, from (1.1),

$$E\mathbf{V}_i = \frac{\rho_i S}{1-\rho}. \tag{1.3}$$

The intervisit time, $\mathbf{I}_i$, for Q_i is defined as

$$\mathbf{I}_i := \mathbf{C}_i - \mathbf{V}_i. \tag{1.4}$$

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 non-exhaustive service, it can be seen that

$$\frac{\lambda_i S}{1-\rho} < 1 \tag{1.5}$$

is an additional condition for the stability of Q_i, $i = 1, \cdots, N$; indeed, for every $i = 1, \cdots, N$ the mean number of type-i arrivals during a cycle should be less than 1. Note that it is possible that, even if Q_i is unstable, some of the other queues are stable.

Similarly, for the SE case we have the following additional conditions:

$$\lambda_i E\mathbf{I}_i = \frac{\lambda_i S(1-\rho_i)}{1-\rho} < 1, \qquad i = 1, \cdots, N. \tag{1.6}$$

This reflects the fact that, for semi-exhaustive service, the mean number of type-i arrivals during the intervisit time $\mathbf{I}_i$ should be less than 1, for during visit times the number of type-i customers is at most reduced by 1.

For the mixed strategies that we allow, the conditions (1.5) and (1.6) should be added to the stability condition $\rho < 1$ for those queues at which we have an NE or SE strategy.

In the case of zero switch-over times, it is well-known that a *conservation law* holds for the total amount of work in the system. This amount should not depend on the order of service, and should hence equal the amount of work in an $M/G/1$ queue with arrival rate Λ and service time distribution the mixture $\Sigma(\lambda_i/\Lambda)B_i(\cdot)$ (this system is subsequently denoted as the 'corresponding' $M/G/1$ queue). Let $E\mathbf{X}_i$ denote the mean number of type-i customers waiting at an arbitrary epoch, and $E\mathbf{W}_i$ the mean waiting time of type-i customers (until their start of service). The foregoing implies [11] that, regardless of the service discipline, the amount of work required by the waiting customers equals:

$$\sum_{i=1}^{N} \beta_i E\mathbf{X}_i = \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} - \sum_{i=1}^{N} \rho_i \frac{\beta_i^{(2)}}{2\beta_i} = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)}.$$

Application of Little's formula allows one to translate this work-conserving property into an (again invariant) expression for a weighted sum of the mean waiting times. Thus the following conservation law is obtained (cf. Schrage [11], Kleinrock [9]):

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)}. \tag{1.7}$$

Recently this conservation law has been generalised by Watson [14] to the cases E, G and NE *with* switch-over times (see also [6], for the cases E and G) and by Boxma [2] to the SE case. Below we state all four pseudo-conservation laws, in a form slightly different from Watson's. The reason for speaking of *pseudo*-conservation laws is the following: in models *with* switch-over times the amount of work (or the weighted sum of the waiting times) is no longer independent of the service strategy.

$$\mathrm{E}: \sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{N} \rho_i^2\right]. \tag{1.8}$$

$$\mathrm{G}: \sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left[\rho^2 + \sum_{i=1}^{N} \rho_i^2\right]. \tag{1.9}$$

$$\mathrm{NE}: \sum_{i=1}^{N} \rho_i \left[1 - \frac{\lambda_i s}{1-\rho}\right] E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left[\rho^2 + \sum_{i=1}^{N} \rho_i^2\right]. \tag{1.10}$$

$$\mathrm{SE}: \sum_{i=1}^{N} \rho_i \left[1 - \frac{\lambda_i s(1-\rho_i)}{1-\rho}\right] E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}(1-\lambda_i s \rho_i/\rho)}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{N} \rho_i^2\right]. \tag{1.11}$$

The known derivations of these conservation laws are not quite satisfactory in the following respects:

— they involve very lengthy calculations,

— they do not explain why such simple expressions exist for weighted sums of mean waiting times (which themselves are sometimes very complicated).

— they do not fully explain the meaning of the various terms on the right-hand sides.

The present paper provides a derivation of a new pseudo-conservation law for cyclic-service systems with mixed service strategies, which contains (1.8)–(1.11) as special cases. This derivation involves few algebraic manipulations and yields an interpretation for each of the terms in the right-hand sides of (1.8)–(1.11). Finally, it answers the question why such relations as (1.8)–(1.11) may be expected to hold.

The derivation is motivated by results in two very interesting recent papers of Fuhrmann and Cooper [8] and Fuhrmann [7]. Fuhrmann [7] gives a simple proof of the pseudo-conservation laws for E, G and NE in the special, symmetric, case where all queues have identical characteristics. By suitably modifying his argument we are able to handle the general case.

Remark 1. The main reason for allowing mixed service strategies is to give a unifying proof for recently obtained mean waiting-time results of four cyclic-service systems with the same service discipline (E, G, NE or SE) at all queues. However, mixed strategies may also be of practical interest. For example, in local area networks where several rings are connected to each other by bridges, the queues which represent the bridges should have higher priority than the other queues at the ring. The service discipline at the ordinary queues usually is non-exhaustive, but at the 'bridge queue' one may consider another service discipline to model the preferential treatment received by these queues.

Remark 2. For E and G, the exact mean waiting times can be numerically calculated by solving $O(N^2)$ linear equations [6]. For NE and SE the mean waiting times are only known when $N = 1$ or $N = 2$ (see [1] for NE and [5] for SE); this fact obviously stresses the importance of the above-mentioned conservation laws.

The remaining part of this paper is divided into three sections. In Section 2 a stochastic decomposition, analogous to a decomposition result in Fuhrmann and Cooper [8], is proven. Our result states that the following relation holds in the cyclic-service systems with mixed service strategies that we have described:

$$\mathbf{V}_c \overset{D}{=} \mathbf{V} + \mathbf{Y}, \tag{1.12}$$

with $\overset{D}{=}$ denoting equality in distribution and

$\mathbf{V}_c :=$ amount of work in a cyclic-service system at an arbitrary epoch,

$\mathbf{V} :=$ amount of work in the 'corresponding' $M/G/1$ system at an arbitrary epoch,

$\mathbf{Y} :=$ amount of work in a cyclic-service system at an arbitrary epoch in a switching period;

$\mathbf{V}$ and $\mathbf{Y}$ are independent.

In proving this result we adapt an approach of Fuhrmann [7]; but where he considers the distribution of the total number of customers and therefore has to restrict himself to the case of N symmetric queues, we consider the total amount of work in the system and do not impose that restriction.

As a by-result of (1.12) we obtain the relation

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + E\mathbf{Y}. \tag{1.13}$$

In Section 3 we shall specify $E\mathbf{Y}$ further, in this way deriving a very general pseudo-conservation law for the weighted sum of the mean waiting times at the various queues in a single-server cyclic-service system with mixed service strategies. Section 4 contains some concluding remarks and topics for further research.

2. A stochastic decomposition result

This section is devoted to a proof of the following decomposition result:

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*, $\mathbf{V}_c$, *is distributed as the sum of the amount of work in the 'corresponding' M/G/*1 *system at an arbitrary epoch*, $\mathbf{V}$, *and the amount of work*, $\mathbf{Y}$, *in the cyclic-service system at an arbitrary epoch in a switching interval. In other words,*

$$\mathbf{V}_c \overset{\mathrm{D}}{=} \mathbf{V} + \mathbf{Y}, \tag{2.1}$$

where $\overset{\mathrm{D}}{=}$ *stands for equality in distribution. Furthermore,* $\mathbf{V}$ *and* $\mathbf{Y}$ *are independent.*

Proof. In the cyclic-service system, the server S is in one of two possible states: S is either serving or switching. As the system is ergodic and stationary, and an amount of work ρ per time unit is offered to the server, we have

$$\Pr\{S \text{ is serving}\} = \rho,$$

$$\Pr\{S \text{ is switching}\} = 1 - \rho.$$

Hence we obtain (with (A) denoting the indicator function of the event A),

$$
\begin{aligned}
&E[\exp(-\omega \mathbf{V}_c)] \\
&\quad = E[\exp(-\omega \mathbf{V}_c)(S \text{ is serving})] + E[\exp(-\omega \mathbf{V}_c)(S \text{ is switching})] \\
&\quad = \rho E[\exp(-\omega \mathbf{V}_c) \mid S \text{is serving}] + (1-\rho) E[\exp(-\omega \mathbf{V}_c) \mid S \text{ is switching}] \\
&\quad = \rho E[\exp(-\omega \mathbf{V}_c) \mid S \text{ is serving}] + (1-\rho) E[\exp(-\omega \mathbf{Y})], \quad \operatorname{Re} \omega \geqq 0.
\end{aligned} \tag{2.2}
$$

We now need the following lemma.

Lemma 1. *The amount of work in the cyclic-service system at an arbitrary epoch in a service interval is distributed as the sum of two independent quantities, viz., the amount of work in the 'corresponding' $M/G/1$ queue at an arbitrary epoch in a service interval and the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval. In other words*:

$$
\begin{aligned}
&E[\exp(-\omega \mathbf{V}_c) \mid S \text{ is serving}] \\
&\quad = E[\exp(-\omega \mathbf{V}) \mid \text{server in } M/G/1 \text{ is serving}] E[\exp(-\omega \mathbf{Y})], \\
&\qquad\qquad \operatorname{Re} \omega \geqq 0.
\end{aligned} \tag{2.3}
$$

Note that the distribution of $\mathbf{V}$ does not depend on the service discipline in the $M/G/1$ queue, as long as no work is created or destroyed within the system; this is the principle of work conservation. From (2.2) and (2.3):

$$
\begin{aligned}
&E[\exp(-\omega \mathbf{V}_c)] \\
&\quad = E[\exp(-\omega \mathbf{Y})][1-\rho+\rho E[\exp(-\omega \mathbf{V}) \mid \text{server in } M/G/1 \text{ is serving}]] \\
&\quad = E[\exp(-\omega \mathbf{Y})] E[\exp(-\omega \mathbf{V})], \quad \operatorname{Re} \omega \geqq 0.
\end{aligned}
$$

Hence we have proved Theorem 1 once we have proved Lemma 1.

In the proof of Lemma 1 we shall need the concepts of 'ancestral line' and 'offspring' of a customer (cf. Fuhrmann and Cooper [8]). Let K_A be a customer who arrives during a switching interval. The customers who arrive during the service of K_A are called the first-generation offspring of K_A. The customers who arrive during the service of customers of the first-generation offspring are called the second-generation offspring of K_A, etc. The set of all customers who belong to the offspring of K_A, including K_A, is called the ancestral line of K_A, and K_A is called the ancestor of all customers in this ancestral line.

Proof of Lemma 1. Adapting an idea of Fuhrmann and Cooper [8], we consider an $M/G/1$ system with a last-come-first-served (LCFS) service discipline and with identically the same traffic process offered as the cyclic-service system, in which the server takes vacations *exactly* during the switching periods of the cyclic-service system (a switching *period* may consist of several consecutive switching *intervals*, e.g., switching intervals from Q_i to Q_{i+1} and from Q_{i+1} to Q_{i+2}). The LCFS discipline is assumed to be non-preemptive,

with one exception: if a service is interrupted by a vacation, forced upon the LCFS system by the cyclic-service system, and if during this vacation new customers arrive, then the interrupted service is resumed when all new customers (and offspring of these customers) have left.

Now consider the cyclic-service system at an arbitrary service epoch. Obviously, the amount of work in the cyclic-service system and in the corresponding LCFS system with vacations are identical at any time, and therefore we can (and we shall) from now on concentrate on the amount of work in the LCFS system at an arbitrary service epoch.

Let K denote the customer who is presently in service in the LCFS system. His ancestor is called K_A. Note that K could be K_A himself. By definition, K_A has arrived during a switching period (or, in this case: a vacation). Because of the 'Poisson arrivals see time averages' property [15], the amount of work found by K_A upon arrival, $\mathbf{Y}_{K_A}$, is distributed like $\mathbf{Y}$. Note that, because of the LCFS service discipline, $\mathbf{Y}_{K_A}$ will still be present when K is in service.

We claim that the rest of the work, present at an arbitrary epoch at which K is being served, is distributed as the amount of work in an ordinary $M/G/1$ system at a service epoch (the service discipline in this $M/G/1$ system may be FCFS or LCFS; or any other work-conserving discipline). Note that it is possible that other customers have arrived after K_A, in the same switching period (vacation). They do not belong to his ancestral line, they are served before K_A and so are their offspring — so they are of no interest to us.

Now consider the epoch at which the service of K_A starts (see Figure 1).

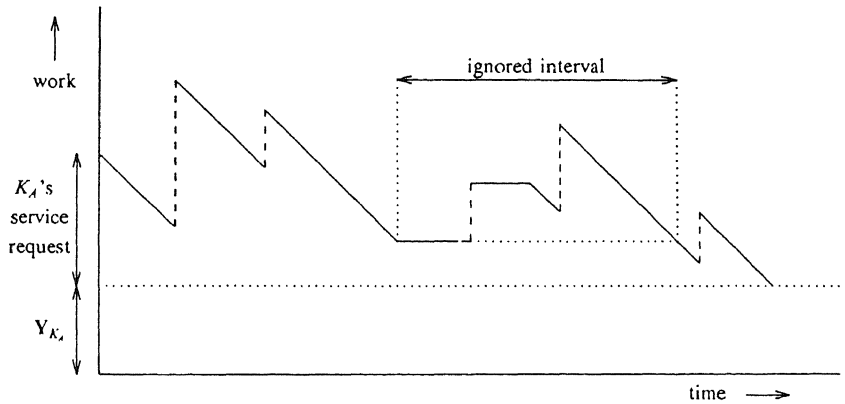

Figure 1. Amount of work in the LCFS system during service of K_A's ancestral line

Apart from $\mathbf{Y}_{K_A}$ no further work is present; and we ignore $\mathbf{Y}_{K_A}$. The residual amount of work now evolves just as in an ordinary $M/G/1$ system with non-preemptive LCFS (or any other work-conserving service discipline) with one exception: during the vacation periods, forced upon the LCFS system by the cyclic-service system, the work remains constant or may increase because of new arrivals. But these new arrivals, and their offspring, are served first (and

do not belong to the ancestral line of K_A), and finally the work level is back again at the level immediately before the vacation started. Note that, due to the memoryless property, the arrival process also starts afresh and that, once more, only $\mathbf{Y}_{K_A}$ and work required by the offspring of K_A is present.

The reasoning shows that, at an arbitrary service epoch of K, the amount of work present is composed of two independent parts: an amount of work $\mathbf{Y}_{K_A}$ that is distributed like $\mathbf{Y}$, and an amount of work that is distributed like the amount of work in an $M/G/1$ queue at an arbitrary service epoch. This proves Lemma 1 and hence Theorem 1 is proven.

Remark 3. In the proof of Lemma 1 the same line of reasoning is used as in the proof of Proposition 5 of Fuhrmann and Cooper [8]; but the reasoning in [8] is held for *customers* at *departure epochs* instead of *work* at *arbitrary epochs*. In [8] this leads to a similar relation as (2.1) for *queue lengths*, for a class of so-called vacation systems. Our cyclic-service model does not fall into this class, because Assumption 3 of [8] is not fulfilled. It is easy to see that, when amounts of work are considered instead of queue lengths, in [8] Assumptions 3 and 4 may be replaced by the assumption that the service discipline is work conserving.

Remark 4. Fuhrmann [7] uses the results of [8] to prove the pseudo-conservation laws (1.8), (1.9) and (1.10) for E, G and NE in the special case of N symmetric queues (identical arrival rates, service-time distributions and switch-over time distributions). His proof is based on the above-mentioned Proposition 5 of [8]. By considering workloads instead of queue lengths, in the next section we prove these pseudo-conservation laws in the more general setting of non-identical queues. In this respect note that work conservation is a more general property than customer conservation.

Remark 5. Ott [10] considers a single-server queueing system with two *independent* input streams, one being of M/G type and the other being a much more general process which need not be Markov. For this system he proves a similar decomposition result (Theorem 2.1) as our Theorem 1.

3. A pseudo-conservation law for weighted waiting times

In this section we use Theorem 1 to derive a pseudo-conservation law for a cyclic-service system with mixed service strategies (e.g., semi-exhaustive at Q_1, exhaustive at Q_2 and Q_4, non-exhaustive at Q_3 and gated at $Q_5, \cdots, Q_N$). This pseudo-conservation law contains (1.8)–(1.11) as special cases. From (2.1),

$$E\mathbf{V}_c = E\mathbf{V} + E\mathbf{Y}, \tag{3.1}$$

and hence,

$$EV_c = \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + EY. \tag{3.2}$$

On the other hand,

$$\begin{aligned} EV_c &= \sum_{i=1}^{N} \beta_i EX_i + \sum_{i=1}^{N} \rho_i \frac{\beta_i^{(2)}}{2\beta_i} \\ &= \sum_{i=1}^{N} \rho_i EW_i + \frac{1}{2} \sum_{i=1}^{N} \lambda_i \beta_i^{(2)}. \end{aligned} \tag{3.3}$$

The first equality follows by noting that, at an arbitrary epoch, a type-i customer is being served with probability ρ_i, while his residual service time has mean $\beta_i^{(2)}/2\beta_i$. From (3.2) and (3.3),

$$\sum_{i=1}^{N} \rho_i EW_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + EY. \tag{3.4}$$

To obtain an expression for this weighted sum of mean waiting times, it remains to determine EY, the mean amount of work in the cyclic-service system at an arbitrary epoch in a switching interval. Denote by Y_i the amount of work in the cyclic-service system at an arbitrary switching epoch during a switchover from Q_i to Q_{i+1}; then it is easily seen that

$$EY = \sum_{i=1}^{N} \frac{s_i}{S} EY_i. \tag{3.5}$$

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 (S) 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. $\rho(s_i^{(2)}/2s_i)$: the mean amount of work that arrived in the system during the past part of the switching interval under consideration.

Hence we have

$$EY_i = EM_i^{(1)} + EM_i^{(2)} + \rho \frac{s_i^2}{2s_i}. \tag{3.6}$$

It will turn out that $EM_i^{(1)}$ is the only term in the right-hand side of (3.6) which depends on the service strategy at Q_i; it can only be determined when the service strategy at Q_i is specified. Hence we shall first consider $EM_i^{(2)}$, the total amount of work in $Q_{i+1}, \cdots, Q_N, Q_1, \cdots, Q_{i-1}$ at a departure epoch of S from

Q_i. By noting that the mean visit time at Q_h is given by $\rho_h s/(1-\rho)$ (cf. (1.3)), we obtain the following relation:

$$
\begin{aligned}
E\mathbf{M}_i^{(2)} &= \rho_{i-1}\left(s_{i-1}+\frac{\rho_i s}{1-\rho}\right)+\rho_{i-2}\left(s_{i-2}+\frac{\rho_{i-1}s}{1-\rho}+s_{i-1}+\frac{\rho_i s}{1-\rho}\right)\\
&\quad+\cdots+\rho_{i+1}\left(s_{i+1}+\frac{\rho_{i+2}s}{1-\rho}+s_{i+2}+\frac{\rho_{i+3}s}{1-\rho}+\cdots+s_{i-1}+\frac{\rho_i s}{1-\rho}\right)\\
&\quad+\sum_{j\neq i} E\mathbf{M}_j^{(1)},
\end{aligned}
\tag{3.7}
$$

and

$$
\sum_{i=1}^{N}\frac{s_i}{s}E\mathbf{M}_i^{(2)}=\frac{\rho}{s}\sum\sum_{h<k}s_h s_k+\frac{s}{1-\rho}\sum\sum_{h<k}\rho_h\rho_k+\sum_{i=1}^{N}\frac{s_i}{s}\sum_{j\neq i}E\mathbf{M}_j^{(1)}. \tag{3.8}
$$

Hence

$$
\begin{aligned}
E\mathbf{Y}=\sum_{i=1}^{N}\frac{s_i}{s}E\mathbf{Y}_i&=\sum_{j=1}^{N}E\mathbf{M}_j^{(1)}+\frac{\rho}{s}\left[\sum\sum_{h<k}s_h s_k+\frac{1}{2}\sum_{i=1}^{N}s_i^{(2)}\right]+\frac{s}{1-\rho}\sum\sum_{h<k}\rho_h\rho_k\\
&=\sum_{j=1}^{N}E\mathbf{M}_j^{(1)}+\rho\frac{s^{(2)}}{2s}+\frac{s}{2(1-\rho)}\left[\rho^2-\sum_{i=1}^{N}\rho_i^2\right].
\end{aligned}
\tag{3.9}
$$

Finally, from (3.4) and (3.9):

$$
\begin{aligned}
\sum_{i=1}^{N}\rho_i E\mathbf{W}_i&=\rho\frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)}+\rho\frac{s^{(2)}}{2s}+\frac{s}{2(1-\rho)}\left[\rho^2-\sum_{i=1}^{N}\rho_i^2\right]\\
&\quad+\sum_{j=1}^{N}E\mathbf{M}_j^{(1)}.
\end{aligned}
\tag{3.10}
$$

A word about the meaning of the terms in the right-hand side of (3.10) is in order. The first term is the mean amount of work waiting in the corresponding cyclic-service system *without* switch-over times (cf. (1.7)). The second, third and fourth terms reflect the influence of the presence of switch-over times. In fact they constitute the mean amount of work present at a switching epoch. The second term represents the mean amount of work that arrived at all queues *during the switching intervals* after the last visit of S to those queues. Note that $s^{(2)}/2s$ represents the mean total past switching time from the departure of S 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 third 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

$$\frac{1}{2}\sum\sum_{h\neq k}\rho_k E\mathbf{V}_h.$$

Finally $\sum_{j=1}^{N} E\mathbf{M}_j^{(1)}$ 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. $E\mathbf{M}_j^{(1)}$ depends on the service strategy at Q_j; hence it can only be determined when the service strategy at Q_j is specified. We now turn to the determination of $E\mathbf{M}_j^{(1)}$ in the four cases of E, G, NE and SE strategy at Q_j.

1. E(xhaustive):

$$E\mathbf{M}_j^{(1)} = 0. \tag{3.11}$$

2. G(ated):

$$E\mathbf{M}_j^{(1)} = \rho_j E\mathbf{V}_j = \rho_j \frac{\rho_j s}{1-\rho} = \rho_j^2 \frac{s}{1-\rho}. \tag{3.12}$$

3. N(on) E(xhaustive):

This requires a bit more work. At a departure epoch of S from Q_j, S has just completed one service with probability $\lambda_j s/(1-\rho)$ and no service with probability $1-\lambda_j s/(1-\rho)$. Hence, with $\mathbf{T}_j$ the amount of work left behind in Q_j at the departure epoch of a *customer* from Q_j,

$$E\mathbf{M}_j^{(1)} = \frac{\lambda_j s}{1-\rho} E\mathbf{T}_j. \tag{3.13}$$

Using an up-and-down-crossings argument and the well-known PASTA-property [15], it follows that the mean queue length at Q_j at a departure epoch of a customer from Q_j and at an arbitrary epoch are equal, and hence, with Little's formula,

$$E\mathbf{T}_j = \beta_j(E\mathbf{X}_j + \rho_j) = \rho_j E\mathbf{W}_j + \rho_j\beta_j. \tag{3.14}$$

From (3.13) and (3.14):

$$E\mathbf{M}_j^{(1)} = \rho_j \frac{\lambda_j s}{1-\rho} E\mathbf{W}_j + \rho_j^2 \frac{s}{1-\rho}. \tag{3.15}$$

4. S(emi) E(xhaustive):

Again, with the above definition of $\mathbf{T}_j$,

$$E\mathbf{T}_j = \rho_j E\mathbf{W}_j + \rho_j\beta_j.$$

Denote by $\mathbf{U}_j$ the number of customers in Q_j at an arrival epoch of S at Q_j. Due to the structure of the SE strategy we can also write

$$E\mathbf{T}_j = \beta_j E[\mathbf{U}_j - 1 \mid \mathbf{U}_j \geqq 1] + \beta_j \left[\frac{\lambda_j^2 \beta_j^{(2)}}{2(1-\rho_j)} + \rho_j \right], \tag{3.16}$$

(note that the second term in the right-hand side represents the amount of work left behind by a departing customer in an $M/G/1$ queue with arrival rate λ_j and service time distribution $B_j(\cdot)$). Subsequently express $E\mathbf{M}_j^{(1)}$ in the first term in the right-hand side of (3.16):

$$E\mathbf{M}_j^{(1)} = \beta_j E[\max(0, \mathbf{U}_j - 1)] = \beta_j E[\mathbf{U}_j - 1 \mid \mathbf{U}_j \geqq 1] \Pr\{\mathbf{U}_j \geqq 1\}. \tag{3.17}$$

Because the mean visit time of S at Q_j during a cycle, when positive, equals the mean busy period of an $M/G/1$ system with arrival rate λ_j and service time distribution $B_j(\cdot)$, we have

$$E\mathbf{V}_j = \frac{\rho_j s}{1-\rho} = \Pr\{\mathbf{U}_j \geqq 1\} \frac{\beta_j}{1-\rho_j}, \tag{3.18}$$

so

$$\Pr\{\mathbf{U}_j \geqq 1\} = \frac{\lambda_j s (1-\rho_j)}{1-\rho}. \tag{3.19}$$

Combining (3.14), (3.16), (3.17) and (3.19),

$$\rho_j E\mathbf{W}_j + \rho_j \beta_j = \frac{E\mathbf{M}_j^{(1)}}{\lambda_j s \dfrac{1-\rho_j}{1-\rho}} + \beta_j \left[\frac{\lambda_j^2 \beta_j^{(2)}}{2(1-\rho_j)} + \rho_j \right]; \tag{3.20}$$

and so we have

$$E\mathbf{M}_j^{(1)} = \rho_j \frac{\lambda_j s (1-\rho_j)}{1-\rho} E\mathbf{W}_j - \frac{\lambda_j s}{2(1-\rho)} \lambda_j \rho_j \beta_j^{(2)}. \tag{3.21}$$

Combining (3.10) and the four expressions for $E\mathbf{M}_j^{(1)}$ in the cases of E, G, NE and SE service strategy at Q_j, 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*,
ne: *the group of N(on) E(xhaustive) queues*,
se: *the group of S(emi) E(xhaustive) queues*.

Then

$$\sum_{i\in e}\rho_i E\mathbf{W}_i+\sum_{i\in g}\rho_i E\mathbf{W}_i+\sum_{i\in ne}\rho_i\left[1-\frac{\lambda_i s}{1-\rho}\right]E\mathbf{W}_i+\sum_{i\in se}\rho_i\left[1-\frac{\lambda_i s(1-\rho_i)}{1-\rho}\right]E\mathbf{W}_i$$
$$=\rho\sum_{\forall i}\frac{\lambda_i\beta_i^{(2)}}{2(1-\rho)}-\sum_{i\in se}\frac{\lambda_i^2\beta_i^{(2)}\rho_i s}{2(1-\rho)}+\rho\frac{s^{(2)}}{2s}+\frac{s}{2(1-\rho)}\left[\rho^2-\sum_{\forall i}\rho_i^2\right]$$
$$+\frac{s}{(1-\rho)}\sum_{i\in g,ne}\rho_i^2. \tag{3.22}$$

Remark 6. The case of $N=1$ queue yields some (mostly well-known) expressions for mean waiting times in $M/G/1$ queues with some form of server vacations.

4. Conclusion and final remarks

In this paper we have derived a stochastic decomposition for the amount of work in cyclic-service systems with mixed service strategies, and we have used this decomposition result to obtain a pseudo-conservation law for such systems. These results form a natural extension of Kleinrock's conservation law [9]; the amount of work is the essential quantity, the relation for the waiting times is a by-product.

Pseudo-conservation laws like (3.22) seem to be very useful in several respects. Firstly, they are useful for obtaining (or testing) approximations for individual mean waiting times (e.g., an approximation of Bux and Truong [4] for E satisfies (1.8), and an approximation in [3] for NE was specifically constructed to satisfy (1.10); in a future study we shall use Theorem 2 for obtaining mean waiting-time approximations in cyclic-service systems with mixed service strategies). Such approximations are badly needed in analytically untractable cases (as in the case of non-exhaustive service) but also in analytically tractable cases; the latter because, when the number of queues is large, the numerical computation of the exact formulas can become very cumbersome. Secondly, pseudo-conservation laws can also be used to study asymptotics, yielding information about what happens when the number of queues becomes very large or when the offered traffic at a particular queue approaches its stability limit (cf. Watson [14]), etc.

In the model description at the beginning of this paper we have assumed a first-come–first-served (FCFS) service discipline at the various queues. The reasoning in the preceding sections reveals that, instead of FCFS, one may allow any work-conserving service discipline (as long as it fits in with the global service strategy).

Finally, some topics for further research. It is worthwhile to investigate whether the assumption that the server visits the queues in a fixed cyclic order can be weakened. Furthermore, (many) more service strategies than just E, G,

NE, SE and mixtures of these four can be considered, provided they fit in the model description given previously. One could for instance think of a generalisation of the NE-strategy, in which the server S serves at most k (instead of one) customers during his visit to a queue (see [7] for a partial result). Another interesting variant might be that the server spends at most T time units at a queue.

References

[1] BOXMA, O. J. (1984) Two symmetric queues with alternating service and switching times. In *Performance '84*, ed. E. Gelenbe, North-Holland, Amsterdam, 409–431.

[2] BOXMA, O. J. (1986) Models of two queues: a few new views. In *Teletraffic Analysis and Computer Performance Evaluation*, eds. O. J. Boxma, J. W. Cohen and H. C. Tijms, North-Holland, Amsterdam, 75–98.

[3] BOXMA, O. J. AND MEISTER, B. (1986) Waiting-time approximations for cyclic-service systems with switch-over times. *Performance Eval. Rev.* **14**, 254–262.

[4] BUX, W. AND TRUONG, H. L. (1983) Mean-delay approximations for cyclic-service queueing systems. *Performance Eval.* **3**, 187–196.

[5] COHEN, J. W. (1987) A two-queue model with semi-exhaustive alternating service. *Performance '87*. To appear.

[6] FERGUSON, M. J. AND AMINETZAH, Y. J. (1985) Exact results for nonsymmetric token ring systems. *I.E.E.E. Trans. Communications* **33**, 223–231.

[7] FUHRMANN, S. W. (1985) Symmetric queues served in cyclic order. *Operat. Res. Letters* **4**, 139–144.

[8] FUHRMANN, S. W. AND COOPER, R. B. (1985) Stochastic decompositions in the $M/G/1$ queue with generalised vacations. *Operat. Res.* **33**, 1117–1129.

[9] KLEINROCK, L. (1976) *Queueing Systems*, Vol. 2. Wiley, New York.

[10] OTT, T. J. (1984) On the $M/G/1$ queue with additional inputs. *J. Appl. Prob.* **21**, 129–142.

[11] SCHRAGE, L. (1970) An alternative proof of a conservation law for the queue $G/G/1$. *Operat. Res.* **18**, 185–187.

[12] TAKAGI, H. (1984) Mean message waiting time in a symmetric polling system. In *Performance '84*, ed. E. Gelenbe, North-Holland, Amsterdam, 293–302.

[13] TAKAGI, H. (1986) *Analysis of Polling Systems*. The MIT Press, Cambridge, Mass.

[14] WATSON, K. S. (1984) Performance evaluation of cyclic service strategies — a survey. In *Performance '84*, ed. E. Gelenbe, North-Holland, Amsterdam, 521–533.

[15] WOLFF, R. W. (1982) Poisson arrivals see time averages. *Operat. Res.* **30**, 223–231.