Queueing Systems, 5 (1989) 185–214

WORKLOADS AND WAITING TIMES IN SINGLE-SERVER SYSTEMS WITH MULTIPLE CUSTOMER CLASSES

O.J. BOXMA

Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands; Faculty of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Received 8 August 1988; revised 31 March 1989

Abstract

One of the most fundamental properties that single-server multi-class service systems may possess is the property of work conservation. Under certain restrictions, the work conservation property gives rise to a conservation law for mean waiting times, i.e., a linear relation between the mean waiting times of the various classes of customers. This paper is devoted to single-server multi-class service systems in which work conservation is violated in the sense that the server's activities may be interrupted although work is still present. For a large class of such systems with interruptions, a decomposition of the amount of work into two independent components is obtained; one of these components is the amount of work in the corresponding system *without* interruptions. The work decomposition gives rise to a (pseudo)conservation law for mean waiting times, just as work conservation did for the system without interruptions.

Keywords: Single-server multi-class service system, service interruptions, work decomposition, conservation law.

1. Introduction

One of the most fundamental properties that single-server multi-class service systems may possess is the property of *work conservation*. Suppose that the server serves at constant rate, and that he serves if and only if at least one customer is present. Also suppose that the scheduling discipline, the procedure for deciding which customer(s) should be in service at any time, has the following property: it does not affect the amount of service given to a customer, or the arrival time of any customer. Then a sample path consideration shows that the amount of work in the system is the same, whatever scheduling discipline with the above-mentioned property is chosen. A pleasant consequence is that the analysis of the workload process in some system with complex priority structure can be reduced to the analysis of the workload process in a system with a more convenient scheduling discipline, like FCFS or LCFS.

Heyman and Sobel [27, p. 383] use the term 'system properties' for such properties as work conservation, Little's theorem, and 'Poisson arrivals see time averages' (PASTA): they are global properties, shared by a large number of specific models. These system properties are mostly based on sample path observations. They can be used in structured models to obtain more specific conclusions. For example, under certain assumptions the mean workload of a particular class of customers can be expressed in the mean number of those customers, and then, via Little's theorem, in their mean sojourn time. Thus the principle of work conservation may lead to the so-called *conservation law*, which states a certain linear relation between the mean waiting (or sojourn) times of customers of all classes in a single-server, multi-class service system:

$$\sum_{n=1}^{N} \rho_n EW_n = C; \tag{1.1}$$

here ρ_n and EW_n are the traffic load and mean waiting time of class n customers, and C is a function of the traffic characteristics of the system but not of the scheduling discipline. The implication of the conservation law is that, if a change in the scheduling discipline causes one of the mean waiting times to decrease, this must happen at the expense of other mean waiting times.

It should be noted that Little's theorem $L = \lambda W$, PASTA and the conservation law have in common that they relate a time average and a customer average. For thorough discussions of these system properties and their interrelations we refer to Chapter 11 of Heyman and Sobel [27], and to the fundamental papers of Brumelle [10] and Wolff [49]. In particular, Heyman and Sobel [27, p. 432] present a proof of the conservation law that is based on a generalization of $L = \lambda W$, viz. Brumelle's [10] formula $H = \lambda G$; here H and G are respectively time and customer averages of quantities which bear a certain relationship to each other but are otherwise unspecified.

The present paper is mainly devoted to single-server multi-class service systems in which the principle of work conservation is *violated* in the sense that the service process may be interrupted although work is still present. A prime example is the 'polling' system in which the server visits the classes in cyclic order, requiring switchover times (interruptions) between classes. For such cyclic-service systems, it has recently been shown [3] that, under the additional assumption of Poisson arrivals, a simple work decomposition result is valid: the amount of work in the system is distributed as the sum of two independent quantities, viz. (i) the amount of work in the corresponding system with identical traffic characteristics but *without* switchover times (hence *with* work conservation), and (ii) the amount of work in the original system at some epoch covered by a switching interval.

The main purposes of the present paper are (i) to extend the validity of the decomposition result beyond cyclic-service systems with switchover times, and (ii) to explore the possibilities to derive a conservation law for mean waiting times in

single-server multi-class systems with interruptions, hence without work conservation. Indeed, such a (pseudo)conservation law is shown to hold, under rather restrictive assumptions regarding the scheduling discipline and the interruption process (e.g., none of them should preempt a service in progress). We use the affix 'pseudo', because the resulting expression for $\sum_{n=1}^{N} \rho_n E W_n$ (cf. (1.1)) now generally *does* depend on the scheduling discipline.

The paper is organized in the following way. Section 2 is devoted to the concept of work. After a brief discussion of work conservation (§ 2.1), the above-mentioned work decomposition result is shown to hold for a rather general single-server multi-class system with interruptions of the service process (§ 2.2). (Pseudo)conservation laws for mean waiting times form the main topic of section 3. First the classical conservation laws for mean sojourn and waiting times are reviewed (§ 3.1). Subsequently the extension to systems with interruptions is made (§ 3.2), after which some special cases are considered for which the pseudoconservation law can be worked out in more detail: the cyclic-service system with switchover times (§ 3.3), a polling system with more general (not strictly cyclic) service order of the classes (§ 3.4), a polling system in which the server visits the classes according to a Markov routing chain (§ 3.5), and a network with a single server in which both server and customers move from queue to queue (§ 3.6).

Conservation laws for mean waiting times serve several useful purposes. In many complex systems they are the only meaningful exact results that can be obtained. Thus they provide important qualitative insight into the behavior of such systems. They can also serve as a test for approximations, and be instrumental in constructing approximations for individual mean waiting times. Section 4 illustrates the latter point. Section 5 presents some conclusions and a list of a few challenging open problems in this area of queueing theory.

The paper partly has the character of a survey. Subsections 2.1 and 3.1, which respectively discuss the principle of work conservation and the conservation law for systems without interruptions, contain hardly any new material; for more fundamental discussions the reader is referred to the books of Gelenbe and Mitrani [22] and Heyman and Sobel [27]. These subsections mainly serve as introduction to the subsections 2.2 and 3.2, where the extension to systems with interruptions is made. Much of the material in the latter subsections is a generalization of results recently obtained for polling systems with cyclic service. A survey of the analysis of polling systems with cyclic service (without the particular emphasis on conservation laws) is given by Takagi [45]. We also refer to Takagi [44,45] for examples of polling systems from a wide range of computer-, communication- and production networks.

Recently several decomposition results for queue lengths and for waiting times have been obtained for single-server queues with vacations of the server. A server vacation is also a form of interruption, and waiting time decomposition is clearly related to workload decomposition, in particular in the case of Poisson arrivals

-who see time averages. The paper of Doshi [11] is an extensive survey of decomposition results for queueing systems with vacations.

2. Work

This section is devoted to a discussion of the amount of work in a single-server service system with multiple classes of customers. The speed of the server is supposed to be constant. Assume, without loss of generality, that the speed of the server is 1. The amount of work in the system at time t is defined to be the sum of the remaining required service times of all customers who are present at that time.

In subsection 2.1 we consider the case in which no work is created or destroyed in the system, i.e., the server works as long as there is work and customers do not leave the system before their service has been completed. Next we turn to the case where work may be created in the sense that the service process may be interrupted although work is still present. In subsection 2.2 it will be shown that, under mild assumptions, the work in system can be decomposed into the work in the corresponding system *without* such interruptions, plus an additional term.

2.1. WORK CONSERVATION

A scheduling discipline is a procedure for deciding which customer(s), if any, should be in service at any moment of time [22]. In single-server multi-class service systems there is a wide range of possible scheduling disciplines. The server, S, may serve all customers according to a global discipline like FCFS, LCFS, Processor Sharing or Shortest Remaining Processing Time First; or he may visit the classes in some order (fixed, or random, or following a static or dynamic priority rule) and serve customers within each class according to a global discipline–and this does not yet exhaust all possibilities.

Following Heyman and Sobel [27, p. 418] we introduce, for multi-server multi-class service systems:

DEFINITION 2.1

A scheduling discipline is called *work-conserving* if

(i) no server is free when at least one customer is waiting, and

(ii) the discipline does not affect the amount of service time given to a customer or the arrival time of any customer.

Definition 2.1 excludes the creation and destruction of work. In a single-server system, the work in the system obviously follows the same sample path for any work-conserving discipline. This is not true in multi-server systems, even if all servers have the same speed, unless assumption (i) in definition 2.1 is changed

into 'no server is free when at least one customer is present'. In the following we restrict ourself to single-server systems.

Let $V^{SD}(t)$ denote the amount of work in a single-server multi-class system at time t for a scheduling discipline SD. Assume that the stochastic process $\{V^{SD}(t),\ t \geqslant 0\}$ has an equilibrium distribution and let V^{SD} denote a s.v. with distribution this equilibrium distribution. The above observation, that all work-conserving disciplines applied to a certain realization of the arrival and service demand processes lead to exactly the same realization of the work process, implies the following weaker statement which suffices for most purposes:

$$V^{SD} \overset{D}{=} V^{FCFS}, \tag{2.1}$$

where $\overset{D}{=}$ stands for equality in distribution.

Gelenbe and Mitrani [22, p. 174] present the work-conserving principle in terms of means, using the following formulation:
For any single-server queueing system in equilibrium there exists a constant EV, determined only by the parameters of the arrival and service demand processes, such that

$$EV^{SD} = EV, \tag{2.2}$$

for all work-conserving scheduling disciplines SD.

Rewrite (2.2) as

$$\sum_{n=1}^{N} EV_n^{SD} = EV, \tag{2.3}$$

where EV_n^{SD} is the expected steady-state amount of work due to customers of class n (the sum of the expected remaining service times of all class n customers in the system at a random epoch in the steady state). The implication is [22] that the vector $(EV_1^{SD}, \ldots, EV_N^{SD})$ always varies with the scheduling discipline in such a way that the sum of its elements remains constant.

Under certain assumptions concerning the scheduling discipline and the arrival and service demand processes, the mean amount of work due to class n can be expressed in the mean number of class n customers in the system and hence, via Little's theorem, in the mean sojourn time of class n customers. Therefore (2.3) might lead to a relation between the various sojourn times. We turn to this topic in Section 3. First we investigate, in subsection 2.2, the extent to which the fundamental property (2.1) can be generalized when the work-conserving property is violated by allowing a specific form of work creation.

2.2. WORK DECOMPOSITION

Again consider the single-server multi-class service system, but extend the set of states in which server S can be from $\{free,\ serving\}$ to $\{free,\ interrupted,$

serving}. S is in the state 'interrupted' when he is not serving customers although at least one customer is in the queue; he is in the state 'free' iff there are no customers present. Generally, we shall lump the states 'free' and 'interrupted' into the state 'non-serving'. Interruptions may occur in various forms:
- the server takes a vacation;
- the server requires switchover times between classes, or between customers, or even between service intervals of one and the same customer;
- the server experiences a breakdown.

Accordingly, the process of service interruptions is a stochastic process which may be intricately interwoven with the arrival and service processes and the scheduling discipline.

Interruptions destroy the work-conserving property of the system; in Kleinrock's terminology [32], work is created when interruptions take place. To be still able to make general and useful statements about the work in the system, we restrict the generality of the arrival process: in the following we consider a batch Poisson arrival process with a correlation structure, as introduced in Levy and Sidi [36] in their recent study of cyclic polling systems. This arrival process is defined below.

DEFINITION 2.2

Arrival epochs occur according to a Poisson process with rate λ. At each arrival epoch, batches of size $\boldsymbol{K} = (\boldsymbol{K}_1, \ldots, \boldsymbol{K}_N)$ of customers of the classes $1, \ldots, N$ arrive with some arbitrary joint batch size distribution. The elements of the vector $\boldsymbol{K}$ are assumed to have the same joint distribution at each arrival epoch, and this distribution is independent of previous or future arrival epochs. The arrival rate of customers of class n is denoted by $\lambda_n := \lambda E\boldsymbol{K}_n$. Finally

$$K_{n,n} := E\boldsymbol{K}_n^2 - E\boldsymbol{K}_n, \quad K_{m,n} := E\boldsymbol{K}_m\boldsymbol{K}_n, \quad m \neq n. \tag{2.4}$$

Note that this arrival process offers the possibility to model the synchronization of several arrival streams.

Let $\boldsymbol{V}^{SD,I}(t)$ denote the amount of work in the system at time t for a scheduling discipline SD and interruption process I. We introduce the following

ASSUMPTION 2.1

1. The stochastic process $\{\boldsymbol{V}^{SD,I}(t),\ t \geqslant 0\}$ possesses an equilibrium distribution.
2. The scheduling discipline SD is work-conserving.
3. The interruption process does not affect the amount of service time given to a customer or the arrival time of any customer.
4. The arrival process is the Poisson process introduced in definition 2.2.

It should be noted that the third assumption does not exclude the possibility that lengths of service interruptions depend on the class of customer whose

service was interrupted, or the class of customer to be served next, or on numbers of customers being present. It is not accidental that the second and third assumptions put similar restrictions on the scheduling discipline and the interruption process: an interruption could also be viewed as the service of class $N+1$ customers, which at the start of such an interruption have higher priority than all other customer classes, but whose work is not counted in $V^{SD,I}(t)$.

From now on, we restrict ourself to the consideration of *steady-state* distributions (see also the first part of assumption 2.1). $V_{SD,I}$ denotes a s.v. with distribution the equilibrium distribution of $\{V_{SD,I}(t),\ t \geqslant 0\}$. In the sequel, the 'corresponding' M/G/1 system indicates a single-server multi-class system with exactly the same arrival and service demand process and scheduling discipline as the system under consideration, but without service interruptions. According to (2.1), the amount of work in that corresponding M/G/1 system is the same for all work-conserving scheduling disciplines. We denote the steady-state amount of work in that system by V. The main result of this section is the following *work decomposition* result:

THEOREM 2.1

Consider a single-server multi-class service system under assumption 2.1. The steady-state amount of work in the system, $V^{SD,I}$, is distributed as the sum of the steady-state amount of work in the corresponding M/G/1 system, V, and the steady-state amount of work, Y, present in the original system at a nonserving interval:

$$V^{SD,I} \overset{D}{=} V + Y. \tag{2.5}$$

Furthermore, V and Y are independent.

Proof

In [3] we have formulated and proved the same decomposition result for the special case of a cyclic polling system with single Poisson arrivals and switchover times (i.e., interruptions) of the server in moving from one class (queue) to the next on the cycle. That proof can almost literally be used in the present more general setting. To make the paper self-contained, we repeat the main line of the argument below.

In the proof we need the concepts of 'ancestral line' and 'offspring' of a customer (cf. Fuhrmann and Cooper [20]). Let C_A be a customer who arrives during a non-serving interval. The customers who arrive during the service of C_A are called the first generation offspring of C_A. The customers who arrive during the service of customers of the first generation offspring are called the second generation offspring of C_A, etc. The set of all customers who belong to the offspring of C_A, including C_A, is called the ancestral line of C_A, and C_A is called the ancestor of all customers in this ancestral line.

Adapting an idea of Fuhrmann and Cooper [20], 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 system with interruptions, in which the server takes vacations *exactly* during the non-serving periods of the system with interruptions. The LCFS discipline is assumed to be nonpreemptive, with one exception: if a service is interrupted by a vacation, forced upon the LCFS system by the system with interruptions, 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 system with interruptions at the arrival epoch of an arbitrary customer, say C. Obviously, the amounts of work in the system with interruptions and in the corresponding LCFS system with vacations are identical at any time, so we can concentrate on the amount of work in the LCFS system at a batch arrival epoch. Because of the 'Poisson arrivals see time averages' property [50], this amount of work has the same distribution as the steady-state amount of work.

C's ancestor is called C_A. Note that C could be C_A himself. By definition, C_A has arrived during a non-serving period (or, in this LCFS case: a vacation). Another application of the PASTA property implies that the amount of work found by C_A's batch upon arrival, Y_{C_A}, is distributed like Y. Note that, because of the LCFS service discipline, Y_{C_A} will still be present when C arrives. Also note that it is possible that other customers have arrived after C_A's batch, in the same non-serving period (vacation). They do not belong to his ancestral line, they are served before C_A and so are their offspring - so they are of no interest to us.

The rest of the work, present at C's arrival epoch, is distributed as the amount of work in the corresponding M/G/1 system with batch arrivals, at an arrival (or arbitrary) epoch. Consider the epoch at which the service of C_A's batch starts (see fig. 2.1). Apart from Y_{C_A} no further work is present; and we ignore Y_{C_A}. The residual amount of work now evolves just as in the corresponding M/G/1

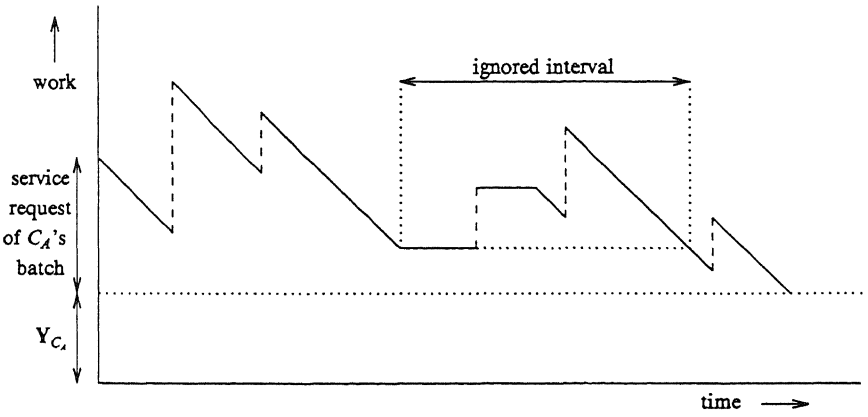

Fig. 2.1. Amount of work in the LCFS system during service of the ancestral line of C_A's batch.

system, with one exception: during the vacation periods, forced upon the LCFS system by the system with interruptions, 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 C_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 $\boldsymbol{Y}_{C_A}$ and work required by the offspring of C_A's batch is present.

This reasoning shows that, at the arrival epoch of C's batch, the amount of work present is composed of two independent parts: an amount of work $\boldsymbol{Y}_{C_A}$ that is distributed like $\boldsymbol{Y}$, and an amount of work that is distributed like the amount of work in the corresponding M/G/1 queue with batch arrivals. As observed above, the PASTA property implies that the amount of work present at the arrival epoch of C's batch has the same distribution as the steady-state amount of work. This proves the theorem.

REMARK 2.1

For the case of the cyclic polling system with single Poisson arrivals and switchover times [3], B.T. Doshi kindly showed us a different proof of the decomposition result. That proof is based on a level crossing argument. We present it below; its extension to the present model is straightforward.

Let λ denote the rate of the Poisson arrival process. Let $B(.)$ denote the service time distribution of an arbitrary customer (averaged over the classes), with mean β and Laplace-Stieltjes transform $\beta(.)$. The traffic intensity equals $\rho := \lambda\beta$. Let $V(.)$ and $Y(.)$ denote the distributions of $\boldsymbol{V}^{SD,I}$ and $\boldsymbol{Y}$ in the cyclic system with switchover times. Assume for simplicity that their densities exist; denote them by $v(.)$ and $y(.)$, and denote the Laplace transforms of these densities by $\phi(.)$ and $\eta(.)$. Equating the downcrossing and upcrossing rates of level $x > 0$ yields:

$$v(x) - (1-\rho)y(x) = \lambda \int_{0-}^{x} (1 - B(x-y))v(y)\,\mathrm{d}v.$$

Combining this relation with

$$v(0) = (1-\rho)y(0),$$

and taking Laplace transforms leads to:

$$\phi(s) = (1-\rho)\eta(s) + \lambda\phi(s)\frac{1-\beta(s)}{s}.$$

Hence

$$\phi(s) = \frac{(1-\rho)s}{s - \lambda + \lambda\beta(s)}\eta(s),$$

which proves the decomposition into two independent components. The same argument has been used by Doshi in [11], p. 58, to give a new proof of another work decomposition result: a result of Ott [39] for a model with a single server

and two customer classes, with class 1 customers arriving according to a Poisson process and class 2 customers arriving according to a very general process.

REMARK 2.2

The paper of Doshi [11] mentioned above is a survey on queueing systems with vacations. It presents a beautiful methodological overview of decomposition results for queueing systems in which the server works on primary and secondary customers (vacations). The paper concentrates on (decompositions for) waiting time distributions. Doshi [12] considers the decomposition of the steady-state amount of *work* in a single-server single-class system with vacations. The arrival process is allowed to be a semi-Markov process. The form of the work decomposition in [12] differs from ours in the sense that the vacations in [12] are considered as additional work. Another recent paper devoted to decompositions for the M/G/1 queue with vacations is Fuhrmann and Cooper [20]. Their study concentrates on *queue length* distributions (at departure epochs). The proof of Theorem 2.1 is based on an idea of [20]; but work decomposition appears to be more natural than queue length decomposition, and indeed our assumptions are less restrictive than those needed in [20]. In particular, when amounts of work are considered instead of queue lengths, Assumptions 3 and 4 of [20] may be replaced by the assumption that the service discipline is work-conserving.

REMARK 2.3

In [4] we have formulated and proved a decomposition result for a cyclic polling system with switchover times in a discrete-time setting. In this setting, time is divided into slots, and numbers of arrivals in successive slots are independent, identically distributed s.v. Letting the slot size tend to zero leads to continuous-time results, with batches arising in a natural way. One of the few subtleties required in proving Theorem 2.1 in discrete time is the replacement of the PASTA property by the BASTA property, 'Bernoulli Arrivals See Time Averages'; cf. [4] and [26].

3. Conservation laws for mean sojourn and waiting times

As remarked at the end of § 2.1, under certain restrictions the mean amount of work due to customers of class n can be related to the mean number of class n customers in the system, and hence also to the mean sojourn time of class n customers. Thus (2.3) leads to a relation between the various mean sojourn times, the prime performance measures in most service systems. Such a relation is sometimes referred to as a conservation law (Kleinrock [30-32]). In § 3.1 a conservation law is presented for various examples of the single-server multi-class system without interruptions. In § 3.2 the same is done for the case *with*

interruptions. In §§ 3.3-3.6 particular attention is paid to polling systems with either a fixed (e.g., cyclic) or random service order, and with switchover times.

3.1. NO INTERRUPTIONS–A CONSERVATION LAW FOR MEAN SOJOURN AND WAITING TIMES

Starting point is relation (2.3):

$$\sum_{n=1}^{N} EV_n^{SD} = EV.$$

This relation for mean amounts of work is generally valid for single-server multi-class systems in equilibrium, with a work-conserving scheduling discipline. In order to go from here to mean sojourn times, and arrive at useful relations between them, one has to impose several restrictions. The discussion below is mainly based on Gelenbe and Mitrani [22]. Following [22, p. 175], we first introduce

ASSUMPTION 3.1

Only information about the current state and the past of the queueing process is used in making scheduling decisions; thus, it is possible to discriminate among customers on the basis of their expected remaining service times (since their classes and attained service are known), but not on the basis of exact remaining service times.

The purpose of the restriction is to exclude scheduling disciplines, like Shortest Remaining Processing Time First, for which the mean service time of a customer, who is still present, differs from an arbitrary mean service time. Further we introduce the following

ASSUMPTION 3.2

1. Successive interarrival times of class n customers are independent, identically distributed s.v. with mean $1/\lambda_n$.
2. Successive required service times of class n customers are independent, identically distributed s.v. with distribution $B_n(.)$, with mean β_n and second moment $\beta_n^{(2)}$, $n = 1, \ldots, N$.
3. The arrival processes and the service demand processes are independent stochastic processes.

It will be seen later that this assumption is unnecessarily restrictive (cf. [41]).

Denote the traffic intensity of class n by $\rho_n = \lambda_n \beta_n$, $n = 1, \ldots, N$, and the total traffic intensity by $\rho = \rho_1 + \ldots + \rho_N < 1$. Under the above assumptions, mean amount of work can be easily related to mean numbers of customers, in the

following two cases: (i) all required service times are exponentially distributed, and (ii) the scheduling discipline is nonpreemptive. In case (i),

$$EV_n^{SD} = \beta_n EX_n^{SD}, \quad n = 1, \ldots, N, \tag{3.1}$$

with EX_n^{SD} the mean number of class n customers in the system under scheduling discipline SD. In case (ii),

$$EV_n^{SD} = \beta_n E\bar{X}_n^{SD} + \rho_n \frac{\beta_n^{(2)}}{2\beta_n}, \tag{3.2}$$

with $E\bar{X}_n^{SD}$ the mean number of waiting class n customers in the system under scheduling discipline SD (because SD is nonpreemptive, service of those customers has not yet been started), and with $\beta_n^{(2)}/2\beta_n$ the mean residual service time of a class n service in progress.

In both cases, application of Little's formula leads to a *conservation law*. The results are formulated in the following theorem. In the sequel, $ES_n^{SD}(EW_n^{SD})$ denotes the mean sojourn (waiting) time of a class n customer, $n = 1, \ldots, N$, under scheduling discipline SD.

THEOREM 3.1

Consider a single-server multi-class service system with work-conserving scheduling discipline SD, under the assumptions 3.1 and 3.2.

(i) When the required service times are exponentially distributed there exists a constant EV, determined only by the interarrival time distributions and the mean service times, such that

$$\sum_{n=1}^{N} \rho_n ES_n^{SD} = EV. \tag{3.3}$$

(ii) When SD is nonpreemptive there exists a constant EV, determined only by the interarrival and service time distributions, such that

$$\sum_{n=1}^{N} \rho_n ES_n^{SD} = EV + \sum_{n=1}^{N} \rho_n \left(\beta_n - \frac{\beta_n^{(2)}}{2\beta_n} \right), \tag{3.4}$$

and

$$\sum_{n=1}^{N} \rho_n EW_n^{SD} = EV - \tfrac{1}{2} \sum_{n=1}^{N} \lambda_n \beta_n^{(2)}. \tag{3.5}$$

EV is generally unknown; in case (ii), it equals the – unknown – mean amount of work in the GI/G/1 queue. When all arrival processes are independent Poisson processes, EV can be determined by considering an M/G/1 queue with $SD = FCFS$, in which all customer classes are lumped together into one customer

class with arrival rate $\Lambda = \lambda_1 + \ldots + \lambda_N$ and service time distribution $\Sigma(\lambda_n/\Lambda)B_n(.)$. The Pollaczek-Khintchine formula then yields in case (i):

$$EV = \sum_{n=1}^{N} \rho_n \beta_n/(1-\rho),$$

and in case (ii):

$$EV = \sum_{n=1}^{N} \lambda_n \beta_n^{(2)}/(2(1-\rho)).$$

(3.3)–(3.5) now reduce to simple expressions for a weighted sum of mean sojourn (waiting) times. These expressions were first obtained by Kleinrock [30,31]. As observed by him, their implication is that, if a change in the scheduling discipline causes one of the mean sojourn (waiting) times to decrease, this must happen at the expense of other mean sojourn (waiting) times. This justifies the use of the word conservation. Formula (3.5) for a general arrival process is due to Schrage [41]; Schrage in fact made no assumptions about independence of the interarrival times and of the service times, see also [27, p. 432].

After having established that the mean sojourn times must satisfy a certain linear relation, Gelenbe and Mitrani [22] proceed to narrow the possibilities further by showing that a certain inequality constraint holds for $\Sigma_{n \in g} \rho_n E\boldsymbol{S}_n^{SD}$, for all subsets g of $\{1, \ldots, N\}$. To illustrate the concept we state their theorem 6.5 for case (ii):

THEOREM 3.2

In any N-class M/G/1 system in equilibrium, for every non-empty subset g of customer class indices, and for any work-conserving nonpreemptive scheduling discipline SD for which assumption 3.1 holds, the mean sojourn times satisfy the inequality

$$\sum_{n \in g} \rho_n E\boldsymbol{S}_n^{SD} \geqslant \tfrac{1}{2} \sum_{k=1}^{N} \lambda_k \beta_k^{(2)} \sum_{n \in g} \rho_n \Big/ \Big(1 - \sum_{n \in g} \rho_n\Big) + \sum_{n \in g} \lambda_n \beta_n^2. \tag{3.6}$$

Moreover, (3.6) becomes an equality if SD gives nonpreemptive priority to g-customers.

The proof is based on the following considerations. In order to minimize $\Sigma_{n \in g} \rho_n E\boldsymbol{S}_n^{SD}$, the customers from g should receive nonpreemptive priority over the non-g customers. Now the sojourn time of a customer from g can only be influenced by non-g customers if he finds one of those in service. With Poisson arrivals, the probability of this event is not influenced by the scheduling strategy. This reasoning implies that the minimal value of $\Sigma_{n \in g} \rho_n E\boldsymbol{S}_n^{SD}$ can be obtained by lumping all customers from g into one class, all other customers into a second class, and giving head-of-the-line priority to the customers from g. The theorem now follows.

Subsequently Gelenbe and Mitrani [22] present so-called characterization results, which state that all mean sojourn time vectors which satisfy the constraints, can indeed be realized by choosing a specific scheduling discipline. We omit discussion of this topic, but refer the reader to chapter 6 of [22] for some interesting results and further references.

REMARK 3.1

Heyman and Sobel [27, p. 432] extend (3.5) to a multi-server multi-class queue. For this extension they require that all customer classes have the same service time distribution. See also lemma 1 of Federgruen and Groenevelt [16]. The latter authors subsequently extend theorem 3.2 above to a multi-server queue. They show that the performance space, the set of mean waiting time vectors which are achievable under some nonpreemptive work-conserving scheduling discipline, is a polyhedron described by $2^N - 1$ inequalities. The special structure of this polyhedron allows for efficient ($O(N^2 \log N)$) procedures to minimize any convex (separable) function of the vector of mean waiting times.

REMARK 3.2

A minor but interesting extension of case (ii) of theorem 3.1 is the following. Consider a network of service stations $Q_1, \ldots, Q_N$. Customers of class n arrive at Q_n; after having been served in Q_n they move to some queue Q_m with transition probability p_{nm}, $n, m = 1, \ldots, N$, becoming class m customers, etc. With probability $1 - \sum_{m=1}^{N} p_{nm}$, a class n customer leaves the system. There is one server in the network, who moves from queue to queue. For the moment, we assume that switchover times of the server between queues are negligible (and we assume the same for customer switchover times). Each service is assumed to be nonpreemptive. The conditions given in case (ii) of theorem 3.1 are satisfied; the extension lies in the fact that customers may change class. Formula (3.2) must be replaced by:

$$EV_n^{SD} = E\bar{X}_n^{SD}\left[\beta_n + \sum_{m=1}^{N}\sum_{k=1}^{\infty} p_{nm}^{(k)}\beta_m\right] + \rho_n\left[\frac{\beta_n^{(2)}}{2\beta_n} + \sum_{m=1}^{N}\sum_{k=1}^{\infty} p_{nm}^{(k)}\beta_m\right], \qquad (3.7)$$

with $p_{nm}^{(k)}$ the k-step transition probability from Q_n to Q_m, and ρ_n the total traffic intensity of class n customers.

Networks of queues with one single server arise in various models of computer-communication systems. Klimov [34] has studied the problem of moving the server in such a way as to minimize some objective function. Foss [18] relaxes Klimov's assumption of Poisson arrivals. Several papers have been devoted to an exact queue-length analysis for the special case of tandem configurations with one moving server, cf. Nair [38], Taube-Netto [46] and the recent study of Katayama [28], to which we also refer for further references.

3.2. INTERRUPTIONS–A PSEUDOCONSERVATION LAW FOR MEAN WAITING TIMES

In this subsection we try to extend the conservation law results of the preceding subsection to the case with interruptions. Consider a single-server multi-class system under the assumption 2.1. Theorem 2.1 implies that

$$\sum_{n=1}^{N} EV_n^{SD,I} = EV + EY. \tag{3.8}$$

EV is the mean amount of work in the corresponding M/G/1 system with batch arrival process as defined in definition 2.2; viewing the batches as super-customers, EV is also the mean amount of work in an M/G/1 system with single arrivals (with arrival rate λ) and service time distribution the distribution of the total service time of a batch. In line with our earlier notation, individual service times of class n customers have distribution $B_n(.)$ with mean β_n and second moment $\beta_n^{(2)}$; the arrival rate of class n customers is $\lambda_n = \lambda EK_n$, with EK_n the mean batch size of class n arrivals; and $\rho_n = \lambda_n \beta_n$, $\rho = \rho_1 + \ldots + \rho_N$. Denoting the second moment of the service time of a supercustomer by $b^{(2)}$, we can write [36]:

$$b^{(2)} = \sum_{m=1}^{N} \sum_{n=1}^{N} \beta_m \beta_n K_{m,n} + \sum_{n=1}^{N} \beta_n^{(2)} EK_n, \tag{3.9}$$

and

$$EV = \frac{\lambda b^{(2)}}{2(1-\rho)}. \tag{3.10}$$

From (3.8) and (3.10),

$$\sum_{n=1}^{N} EV_n^{SD,I} = \frac{\lambda b^{(2)}}{2(1-\rho)} + EY. \tag{3.11}$$

We are left with two problems. We have to relate $EV_n^{SD,I}$ to $ES_n^{SD,I}$ ($EW_n^{SD,I}$), the mean sojourn (waiting) time of a class n customer; and we have to determine EY, the mean amount of work present in some epoch covered by a non-serving interval. Solution of the first problem requires similar restrictive assumptions as were made in § 3.1 for the case without interruptions. In particular, along with assumption 3.1, we also impose:

ASSUMPTION 3.3

The interruption process uses only information about the current state and the past of the queueing process; thus, no information about exact remaining service times is used.

Thus we exclude, e.g., the case that there is always an interruption when the residual service time of the customer in service equals d.

One can now prove the following pseudoconservation laws, which are the counterparts of the conservation laws in theorem 3.1 (note the differences in the arrival processes):

THEOREM 3.3

Consider a single-server multi-class M/G/1 service system with scheduling discipline SD and interruption process I under the assumptions 2.1, 3.1 and 3.3.

(i) If the required service times are exponentially distributed, then

$$\sum_{n=1}^{N} \rho_n E S_n^{SD,I} = \frac{\lambda b^{(2)}}{2(1-\rho)} + EY. \tag{3.12}$$

(ii) If the scheduling discipline and the interruption process are such that services are not preempted, then

$$\sum_{n=1}^{N} \rho_n E S_n^{SD,I} = \frac{\lambda b^{(2)}}{2(1-\rho)} + \sum_{n=1}^{N} \rho_n \left(\beta_n - \frac{\beta_n^{(2)}}{2\beta_n} \right) + EY, \tag{3.13}$$

and

$$\sum_{n=1}^{N} \rho_n E W_n^{SD,I} = \frac{\lambda b^{(2)}}{2(1-\rho)} - \tfrac{1}{2} \sum_{n=1}^{N} \lambda_n \beta_n^{(2)} + EY. \tag{3.14}$$

REMARK 3.3

The reason for not allowing interruptions during a service time in case (ii) is the same as the one for not allowing preemptions by the scheduling discipline: we have to exclude that an arbitrary waiting customer already has received some service time. In case (i) the non-anticipating assumption 3.3 suffices.

REMARK 3.4

With similar modifications as above, case (ii) can be extended to the Klimov network with a single server that was discussed in remark 3.2. See also § 3.6 below.

We now turn to the second problem that eq. (3.11) gave rise to, viz., the determination of EY. Since EV is completely independent of the scheduling discipline and interruption process, their whole influence on $EV^{SD,I}$ is concentrated in EY. Therefore we can hardly expect to make meaningful statements about EY without specifying the scheduling discipline and interruption process in detail.

In the rest of this section we concentrate on case (ii) of theorem 3.3, while the only interruptions are switches of S between classes. § 3.3 is devoted to the case of cyclic service, i.e., S successively visits classes $1, 2, \ldots, N, 1, 2, \ldots$, requiring switchover times between classes; in § 3.4 the cyclic order is generalized to an

arbitrary fixed class visit order, and in § 3.5 to a random (in fact, Markovian) polling order; finally in § 3.6 the Klimov network with a single server and switchover times of the server between classes (queues) is briefly considered.

3.3. A PSEUDOCONSERVATION LAW FOR CYCLIC-SERVICE SYSTEMS

Single-server multi-class service systems with cyclic service of the classes and switchover times frequently arise in the performance analysis of computer-communication networks. An important example is provided by local area networks with a token ring protocol. This example, along with several others, is discussed in more detail by Takagi [44]; his study also contains a detailed analysis and extensive survey of cyclic-service models.

In this subsection we are going to derive a pseudoconservation law for mean waiting times in a cyclic-service system, by working out an expression for EY in (3.14). The resulting pseudoconservation law is an extension of the one in [3]. For some of the service strategies to be considered below, exact expressions for individual mean waiting times have been derived in the literature, usually as the solution of a large set of linear equations; for further details and references the reader is referred to the surveys of Takagi [44,45].

We still have to specify the switchover process. As we started from a very general model in section 2 and gradually restricted ourself more and more, it seems appropriate to give a concise

Model description

A single server S serves N classes of customers, or rather N queues $Q_1, \ldots, Q_N$ with infinite waiting rooms, in cyclic order: $Q_1, Q_2, \ldots, Q_N, Q_1, Q_2, \ldots$. The switchover times of S between the nth and $(n+1)$th queue are independent, identically distributed s.v. with first moment s_n and second moment $s_n^{(2)}$. The first moment of the total switchover time during a cycle of the server, s, is given by $s = \sum_{n=1}^{N} s_n$; its second moment is denoted by $s^{(2)}$. When S finds a queue empty, he immediately begins to switch to the next queue.

The arrival process of customers is the correlated Poisson process introduced in definition 2.2. The service times of class n customers are independent, identically distributed s.v. with distribution $B_n(.)$, with mean β_n and second moment $\beta_n^{(2)}$. As before, the traffic intensities are denoted by ρ_n, $n = 1, \ldots, N$, and $\rho = \sum_{n=1}^{N} \rho_n$.

The interarrival, service demand and switchover time processes are mutually independent, apart from the correlation between the sizes of simultaneously arriving batches (in fact, one might also relax the independence of successive switchover times).

For the service strategies at the queues there are various possibilities, which differ in the numbers of customers who may be served in a queue during a visit of

S to that queue. Before specifying a number of such strategies, let us see how far we can get from theorem 3.3 with the above specification of the interruption (switchover) process.

Assume that the queueing system under consideration is in equilibrium (the ergodicity conditions depend on the service strategies at the queues; obviously $\rho < 1$ is a necessary condition). The conditions of case (ii) of theorem 3.3 are fulfilled, so (3.14) holds. To determine EY, we follow the approach in [3]. Denote by EY_n the mean amount of work in the cyclic-service system at some epoch covered by a switchover from Q_n to Q_{n+1}. So

$$EY = \sum_{n=1}^{N} \frac{s_n}{s} EY_n. \tag{3.15}$$

EY_n is composed of three terms:

1. $EM_n^{(1)}$: the mean amount of work in Q_n at a departure epoch of S from Q_n;
2. $EM_n^{(2)}$: the mean amount of work in the rest of the system at a departure epoch of S from Q_n;
3. $\rho(s_n^{(2)}/2s_n)$: the mean amount of work that arrived in the system during the past part of the switching interval under consideration.

To calculate $EM_n^{(1)}$ and $EM_n^{(2)}$, we need the following two globally valid results for cyclic-service systems (cf. Takagi [44]):

The mean *cycle time*, i.e., the mean time between two successive visits of S to, say, Q_n, is independent of n; it is given by

$$EC = \frac{s}{1-\rho}. \tag{3.16}$$

The mean *visit time* of S at Q_n, i.e., the mean time between the arrival and subsequent departure of S at Q_n, is given by

$$EVI_n = \rho_n EC = \rho_n \frac{s}{1-\rho}, \quad n = 1, \ldots, N. \tag{3.17}$$

(3.16) and (3.17) follow from general traffic balance arguments. Repeated use of (3.17) yields:

$$\begin{aligned} EM_n^{(2)} = {} & \rho_{n-1}\left(s_{n-1} + \frac{\rho_n s}{1-\rho}\right) + \rho_{n-2}\left(s_{n-2} + \frac{\rho_{n-1}s}{1-\rho} + s_{n-1} + \frac{\rho_n s}{1-\rho}\right) \\ & + \ldots + \rho_{n+1}\left(s_{n+1} + \frac{\rho_{n+2}s}{1-\rho} + s_{n+2} + \frac{\rho_{n+3}s}{1-\rho} + \ldots + s_{n-1} + \frac{\rho_n s}{1-\rho}\right) \\ & + \sum_{j \neq n} EM_j^{(1)}; \end{aligned} \tag{3.18}$$

and we find

$$\sum_{n=1}^{N} \frac{s_n}{s} EM_n^{(2)} = \frac{\rho}{s} \sum\sum_{h<k} s_h s_k + \frac{s}{1-\rho} \sum\sum_{h<k} \rho_h \rho_k + \sum_{n=1}^{N} \frac{s_n}{s} \sum_{j \neq n} EM_j^{(1)}. \tag{3.19}$$

From (3.15), (3.18) and (3.19),

$$EY = \sum_{j=1}^{N} EM_j^{(1)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)} \left[\rho^2 - \sum_{n=1}^{N} \rho_n^2 \right]. \tag{3.20}$$

Finally, from (3.9), (3.14) and (3.20), suppressing the superscript SD,I:

$$\sum_{n=1}^{N} \rho_n EW_n = \rho \frac{\sum_{n=1}^{N} \lambda_n \beta_n^{(2)}}{2(1-\rho)} + \lambda \frac{\sum_{m=1}^{N} \sum_{n=1}^{N} \beta_m \beta_n K_{m,n}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s}$$
$$+ \frac{s}{2(1-\rho)} \left[\rho^2 - \sum_{n=1}^{N} \rho_n^2 \right] + \sum_{j=1}^{N} EM_j^{(1)}. \tag{3.21}$$

The last three terms, together constituting EY, reflect the influence of the presence of switchover times. The term $\rho s^{(2)}/2s$ 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 $s^{(2)}/2s$ equals 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 switchover times. The next 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. Finally $\sum_{j=1}^{N} EM_j^{(1)}$ represents the mean total amount of work left behind by S at the various queues in one cycle. This is the *only* term that cannot be determined without specifying the service strategies at the various queues. A pleasing consequence of the global validity - irrespective of the service strategies - of the expressions (3.16) and (3.17) for, respectively, mean cycle time and mean visit times, is that $EM_j^{(1)}$ only depends on the service strategy at Q_j, and not on the service strategies at the other queues. Another consequence is that the correlation between batch sizes also has no effect on $EM_j^{(1)}$.
We now turn to the

Determination of $EM_j^{(1)}$ for various service strategies

1. Exhaustive: S serves class j customers until Q_j is empty.

$$EM_j^{(1)} = 0. \tag{3.22}$$

2. Gated: S serves exactly those class j customers that were present upon his arrival at Q_j.

$$EM_j^{(1)} = \rho_j EVI_j = \rho_j^2 \frac{s}{1-\rho}. \tag{3.23}$$

3. Reserved gated (also called fully gated [2, section 3.5.2]): S serves exactly those class j customers that were present upon his departure from Q_{j-1}.

Similarly as for gated service one obtains:

$$EM_j^{(1)} = \rho_j EVI_j + \rho_j s_{j-1} = \rho_j^2 \frac{s}{1-\rho} + \rho_j s_{j-1}. \tag{3.24}$$

4. Binomial-gated: when S finds N_j customers present upon his arrival at Q_j, he serves a number of customers that is binomially distributed with parameters N_j and p_j, $0 < p_j \leqslant 1$. Note that $p_j = 1$ corresponds to gated service. The binomial-gated strategy, which has been introduced and analyzed by Levy [35], allows assigning priorities to the queues of a cyclic service system by choosing the probabilities p_j. It is easily seen [35] that

$$EM_j^{(1)} = \rho_j \left(\rho_j + \frac{1-p_j}{p_j} \right) \frac{s}{1-\rho}. \tag{3.25}$$

5. Binomial-exhaustive: when S finds N_j customers present upon his arrival at Q_j, he sets aside a number of customers that is binomially distributed with parameters N_j and p_j, $0 \leqslant p_j < 1$, and he serves the other customers and those arriving during their service, etc. Note that $p_j = 0$ corresponds to exhaustive service. This service strategy was suggested by W.P. Groenendijk [personal communication]. A simple calculation yields:

$$EM_j^{(1)} = \frac{p_j}{1-p_j} \rho_j (1-\rho_j) \frac{s}{1-\rho}. \tag{3.26}$$

6. 1-limited: S serves exactly one customer at Q_j.

$$EM_j^{(1)} = \rho_j \frac{\lambda_j s}{1-\rho} EW_j + \rho_j^2 \frac{s}{1-\rho} + \frac{\lambda s}{2(1-\rho)} \beta_j K_{j,j}. \tag{3.27}$$

This formula can be derived from (4.15) and (5.8) of [4]. The latter study only considers uncorrelated batch arrivals; but as has been observed above, the correlation of batch sizes has no effect on $EM_j^{(1)}$. Formula (3.27) can be written in the following way:

$$EM_j^{(1)} = \left(1 - \frac{\lambda_j s}{1-\rho} \right) 0 + \frac{\lambda_j s}{1-\rho} \left[\rho_j \{ EW_j + \beta_j \} + \beta_j \frac{K_{j,j}}{2EK_j} \right],$$

with as interpretation: $\lambda_j s/(1-\rho)$ equals the fraction of visits of S to Q_j that result in a service; $\rho_j \{ EW_j + \beta_j \}$ equals the mean amount of work that has arrived during the sojourn time of the departing customer; and $\beta_j K_{j,j}/(2EK_j)$ equals the mean amount of work of the customers who arrived in the same batch as the departing customer but are served after him.

7. Bernoulli: after each service which does not leave Q_j empty, S serves another customer with probability $1-p_j$ and moves to the next queue with probability p_j. This discipline has been introduced by Keilson and Servi [29]. The expression for $EM_j^{(1)}$ is strongly related to its counterpart in the 1-limited case

(an explanation will be given at the end of § 3.5):

$$EM_j^{(1)} = p_j\left[\rho_j \frac{\lambda_j s}{1-\rho} EW_j + \rho_j^2 \frac{s}{1-\rho} + \frac{\lambda s}{2(1-\rho)} \beta_j K_{j,j}\right]. \tag{3.28}$$

Tedijanto [47] has derived (3.28) for the case of single arrivals.

8. Semi-exhaustive: S continues serving class j customers until the number present is one less than the number present upon his arrival. From (4.25) and (5.8) of [4]:

$$EM_j^{(1)} = \rho_j \frac{\lambda_j s(1-\rho_j)}{1-\rho} EW_j - \frac{\lambda_j^2 s}{2(1-\rho)} \rho_j \beta_j^{(2)} - \frac{\lambda s}{2(1-\rho)} \beta_j \rho_j K_{j,j}. \tag{3.29}$$

It follows from (3.21) and the subsequent discussion that, in a cyclic-service system with a mixture of the above listed service strategies (e.g., exhaustive at one queue, 1-limited at the next, etc.) one can easily determine an exact expression for a weighted sum of mean waiting times. Note that for the 1-limited, Bernoulli and semi-exhaustive strategies the weight factor is not equal to the traffic intensity at the queue.

If one of the queues has yet another service strategy, one only has to determine its corresponding $EM_j^{(1)}$. However, this may be a very difficult problem. Consider the G-limited and E-limited service strategies: S serves a queue according to the gated or the exhaustive service strategy, with the restriction that he serves at most, say, k customers. $k = 1$ reduces to 1-limited service, whereas $k = \infty$ reduces to gated respectively exhaustive service. Everitt [14,15] has derived a pseudo-conservation law for G-limited respectively E-limited service, but his formulas still contain the unknown second factorial moment of the number of customers served in the queue at a visit of S. An exact expression for this term is probably hard to come by. Replacing the second factorial moment by zero immediately yields Fuhrmann's bound [19] for the weighted sum of mean waiting times.

REMARK 3.5

In none of the above cases it is necessary to specify the order of service within a class. It suffices to make the restriction to work-conserving nonpreemptive service disciplines for which assumption 3.1 holds.

REMARK 3.6

For cyclic-service systems with switchover times and uncorrelated arrivals, a pseudoconservation law has first been obtained by Ferguson and Aminetzah [17] for the cases of exhaustive service at all queues and gated service at all queues, and by Watson [48] for the same two cases and also for 1-limited service at all queues. The last result is particularly noteworthy, because the mean waiting times at the individual queues are not known apart from the two-queue case [5], for which singular integral expressions are obtained. In [3] these pseudoconservation

laws have been unified and generalized for a mixture of exhaustive, gated, 1-limited and semi-exhaustive strategies at the queues. The probabilistic proof in [3] explains the validity of the pseudoconservation law, and allows an interpretation of the various terms. A discrete-time version of this pseudoconservation law, and an extension to batch arrivals, are presented in [4]. A discrete-time version for the case of gated service at all queues has also been obtained in [40]. Levy and Sidi [36] have further generalized the results of [3] to the case of the correlated batch Poisson arrival process of definition 2.2.

REMARK 3.7

For $N = 1$ queue the above calculations yield some, mostly known, expressions for mean waiting times in M/G/1 queues with various kinds of vacations.

REMARK 3.8

For all listed strategies, apart from the reserved gated one, $EM_j^{(1)}$ is linear in s; and so is EY, if $s^{(2)}/2s$ is linear in s. The pseudoconservation law thus gives an interesting insight into the influence of the total mean switchover time, s, on workload and – to some extent – on the (weighted sum of the) mean waiting times.

3.4. POLLING SYSTEMS WITH A GENERAL SERVICE ORDER TABLE

A generalization of single-server multi-class systems with cyclic service is obtained by allowing the server to visit the queues according to a fixed – but not necessarily cyclic – pattern, like the star pattern: $Q_1, Q_2, Q_1, Q_3, \ldots, Q_1, Q_N, Q_1, Q_2, \ldots$. Polling systems with a general service order table arise naturally in many computer-communication networks; some examples are the token bus local area network, and a computer with multi-drop terminals in a star configuration. The possibility of using general service order tables is also interesting from the viewpoint of optimization; it gives one the opportunity to assign stations higher priority by listing them more often in the table.

Baker and Rubin [1] have presented an exact analysis of waiting times for polling systems with a general service order table, with exhaustive service at all queues. In [6] a pseudoconservation law has been derived for such polling systems but with a mixture of various service disciplines (see also [23] for a related result). Following an idea of [1], the system with a service order table is reduced to a cyclic-service system by introducing *pseudostations*. We illustrate the concept using the star pattern example. The star pattern repeats itself after $2N - 2$ queue visits. The introduction of $2N - 2$ pseudostations $PS_1, \ldots, PS_{2N-2}$ leads to a cyclic-service system, with one complication: $PS_1, PS_3, \ldots, PS_{2N-3}$ all refer to Q_1, and arrivals at these pseudostations really are arrivals at Q_1. Determination of the mean visit times of these pseudostations is not as trivial as before. However, the pseudoconservation law (3.14) holds, and evaluation of an expres-

sion for EY is only slightly more complicated than for the strictly cyclic model. Formula (3.15) remains valid, with $2N-2$ switchover times; the decomposition of EY_n into three terms $EM_n^{(1)}$, $EM_n^{(2)}$ and $\rho(s_n^{(2)}/2s_n)$ also goes through, but the first two terms now refer to pseudostations, and their determination requires a careful bookkeeping of earlier visits to other pseudostations that refer to the same queue.

3.5. POLLING SYSTEMS WITH MARKOVIAN ROUTING OF THE SERVER

Another generalization of single-server multi-class systems with cyclic service is obtained by allowing the server to visit the queues according to a probabilistic routing scheme. Kleinrock and Levy [33] have introduced the *random polling* scheme in which, after a server visit period to a queue, the next queue to be served is Q_j with probability p_j. In [8] the more general *Markovian polling* scheme is considered where a visit to Q_i is with probability p_{ij} followed by a visit to Q_j: S visits the queues according to a Markov chain.

Again a pseudoconservation law can be formulated. Determination of $EM_j^{(1)}$ proceeds as in § 3.3. Determination of $EM_j^{(2)}$, the mean amount of work in the rest of the system at a departure epoch of S from Q_j, provides some difficulties; it requires a careful study of the mean time between a departure of S from Q_j and the last previous departure from, say, Q_i.

The flexibility of the Markovian polling scheme is illustrated by the following example. Consider the case of 1-limited service at all queues, with the following server routing probabilities:

$$p_{ij} = 1 - p_i \quad \text{if } j = i,$$
$$p_{ij} = p_i \quad \text{if } j = i+1,$$
$$p_{ij} = 0 \quad \text{else;}$$

and with the following mean switchover times:

$$s_{ii} = 0,$$
$$s_{i,i+1} = s_i;$$

it is easily seen that this leads to cyclic service with a Bernoulli service strategy at all queues. This observation has been exploited in [8] to derive the pseudo-conservation law (3.28) for the case of single arrivals. In fact, for the Markovian polling scheme, it is seen that the probability that a server visit to Q_j results in a service equals $\lambda_j p_j s/(1-\rho)$, after which the reasoning below (3.27) can again be applied to determine $EM_j^{(1)}$.

3.6. A QUEUEING NETWORK WITH A SINGLE SERVER

Let us return to the network of service stations $Q_1, \ldots, Q_N$ with one single server, that was introduced in remark 3.2. Assume that the conditions of case (ii)

of theorem 3.3 are fulfilled. Then we have (cf. (3.11) and (3.7)):

$$\sum_{n=1}^{N}\left(\rho_n+\bar{\lambda}_n\left[\sum_{m=1}^{N}\sum_{k=1}^{\infty}p_{nm}^{(k)}\beta_m\right]\right)EW_n^{SD,I}$$
$$=\frac{\lambda b^{(2)}}{2(1-\rho)}-\tfrac{1}{2}\sum_{n=1}^{N}\lambda_n\beta_n^{(2)}-\sum_{n=1}^{N}\rho_n\left[\sum_{m=1}^{N}\sum_{k=1}^{\infty}p_{nm}^{(k)}\beta_m\right]+EY, \qquad (3.30)$$

with $p_{nm}^{(k)}$ the k-step transition probability from Q_n to Q_m, $\bar{\lambda}_n$ the total arrival rate at Q_n, and $\rho_n=\bar{\lambda}_n\beta_n$ the total traffic intensity of class n customers.

EY still has to be determined. We consider a similar special case as in § 3.3. The interruption process is specified by assuming that the server, S, visits the queues in a cyclic order; the only interruptions are those caused by the switches of S between queues. The model description is identical to the model description in § 3.3, apart from the fact that customers may move from queue to queue (without switchover times), and change class accordingly. We claim that determination of EY proceeds similarly as in § 3.3. In particular, (3.15)–(3.17) remain valid; (3.18) requires some adaptation because customers can reach a queue from another queue only during particular periods. As before, determination of $EM_j^{(1)}$, the mean amount of work in Q_j at a departure epoch of S from Q_j, depends on the service strategy at Q_j.

Independent of the present study, Sidi and Levy [42] have analysed a network with one cyclically moving server and either exhaustive or gated service at all queues; for this case they have also obtained (3.30).

4. Conservation-law based mean waiting time approximations

Exact expressions for mean waiting times in single-server multi-class service systems are known only in exceptional cases (see Takagi [44,45] for most of the references regarding cyclic-service models). In view of this, (pseudo)conservation laws for mean waiting times are extremely useful, if only as a tool to test approximations or to base approximations upon. As an illustration, we briefly discuss conservation-law based mean waiting time approximations for cyclic-service systems with switchover times (the model of § 3.3). We restrict ourself to single Poisson arrivals, and to FCFS service at all queues.

Recently, several mean waiting time approximations for cyclic-service systems have been suggested in literature. Most of these approximations are based on the following idea, that has independently been developed in [13], for the two cases of exhaustive service and gated service at all queues, and in [7] for 1-limited service at all queues. First obtain a linear relation between the mean waiting time EW_n at queue Q_n and the mean residual cycle time Erc_n for a cycle starting with the arrival of server S at Q_n. The first cycle time moment is the same for all queues; the residual cycle time moments generally differ, but these differences are usually

quite small (cf. the exact analysis of a special case in [5]). Now assume that the N mean residual cycle time moments are exactly the same: $\boldsymbol{Erc}_n = \boldsymbol{Erc}$. Finally substitute the obtained N linear relations between $\boldsymbol{EW}_n$ and $\boldsymbol{Erc}$ in the pseudo-conservation law and solve for the one unknown, $\boldsymbol{Erc}$. Groenendijk [24] has shown that this approach can be applied to cyclic-service systems with a *mixture* of exhaustive, gated and 1-limited service strategies. We briefly indicate the main steps of his approximation.

(i) Q_n has gated service. Then

$$\boldsymbol{EW}_n = (1 + \rho_n)\boldsymbol{Erc}_n. \tag{4.1}$$

Indeed (Groenendijk [personal communication]), the mean waiting time of a tagged class n customer consists of two components. Firstly a mean residual cycle time $\boldsymbol{Erc}_n$, because with gated service a customer is never served in the cycle in which he arrives. Secondly, the mean time from the instant the server arrives at Q_n until the service completion of all class n customers who arrived before the tagged customer, in the same cycle: $(\lambda_n \boldsymbol{Erc}_n)\beta_n$.

(ii) Q_n has exhaustive service. Then one can prove [24] that

$$\boldsymbol{EW}_n = (1 - \rho_n)\boldsymbol{E\tilde{r}c}_n, \tag{4.2}$$

where $\boldsymbol{E\tilde{r}c}_n$ is the mean residual cycle time at Q_n with a cycle starting at a *departure* epoch of S from Q_n. The following simple argument of Doshi [personal communication] immediately leads to (4.2): $\boldsymbol{E\tilde{r}c}_n$ consists of two components; firstly $\boldsymbol{EW}_n$, the mean waiting time of the hypothetical customer whose arrival marks the beginning of the residual cycle, and secondly $\rho_n \boldsymbol{E\tilde{r}c}_n$, the mean work that arrives at Q_n during the residual cycle. A minor variant of this argument is to write

$$\boldsymbol{Erc}_n = \boldsymbol{EW}_n + (\lambda_n \boldsymbol{EW}_n)(\beta_n/(1 - \rho_n)),$$

the last term in the righthand side denoting the mean number of arrivals at Q_n during $\boldsymbol{EW}_n$ times the mean length of the busy period at Q_n generated by such an arrival (note that the hypothetical customer himself should not contribute to $\boldsymbol{Erc}_n$!). Formula (4.2) gives rise to the approximation

$$\boldsymbol{EW}_n \approx (1 - \rho_n)\boldsymbol{Erc}_n. \tag{4.3}$$

(iii) Q_n has 1-limited service. No exact relation between $\boldsymbol{EW}_n$ and $\boldsymbol{Erc}_n$ is known in this case. Groenendijk [24] applies the following idea of [7]. Denoting the number of waiting customers at Q_n found by an arriving class n customer by $\tilde{X}_n$, and the length of a cycle at Q_n which contains a class n service by $\boldsymbol{C}_{b,n}$, one has:

$$\boldsymbol{EW}_n \approx \boldsymbol{Erc}_n + \boldsymbol{E\tilde{X}}_n \boldsymbol{EC}_{b,n} = \boldsymbol{Erc}_n + \lambda_n \boldsymbol{EW}_n \boldsymbol{EC}_{b,n},$$

leading to

$$\boldsymbol{EW}_n \approx \frac{\boldsymbol{Erc}_n}{1 - \lambda_n \boldsymbol{EC}_{b,n}}. \tag{4.4}$$

$EC_{b,n}$ is not exactly known, but for this term an accurate and simple approximation can be obtained. Finally, substitution of (4.1), (4.3) and (4.4), with $Erc_n \equiv Erc$, into the pseudoconservation law (3.14) yields an expression for Erc, and hence an approximation for the individual mean waiting times. The resulting approximation has the following features.

- It is an explicit formula for EW_n, which gives much qualitative insight into the behavior of cyclic-service systems;
- it is exact in the completely symmetric case (same traffic characteristics, switchover time distributions and service strategies at all queues);
- it is an excellent approximation for low and medium traffic;
- it is not very accurate when traffic is high and asymmetric, in particular when the system contains 1-limited service queues.

The main source of the just mentioned inaccuracy is approximation (4.4) for queues with 1-limited service. A more detailed study of cycle times, taking into account information about previous cycles, led Groenendijk [25] to replace (4.4) by

$$EW_n \approx \frac{Erc_n}{1-\lambda_n EC_{b,n}} + H_n. \tag{4.5}$$

Here H_n is a correction term that must be calculated iteratively. The resulting approximation is more accurate but less transparent than the one using (4.4).

The ideas in [25] are partly based on those of Srinivasan [43]. For the case of 1-limited service at all queues, Srinivasan had also improved upon [7] by taking a closer look at (conditional) cycle times before eventually applying the pseudoconservation law.

Finally we mention two more studies which present mean waiting time approximations based on a pseudoconservation law. Fuhrmann and Wang [21] consider the difficult and important cases of G-limited and E-limited service, discussed below (3.29). They derive heuristic mean waiting time approximations, based on tight bounds [19] for the pseudoconservation law. These bounds reduce to the exact pseudoconservation law for exhaustive, gated and 1-limited service in the corresponding limiting cases.

Pang and Donaldson [40] suggest a very accurate mean waiting time approximation for discrete-time cyclic-service systems with gated service at all queues. They express the mean waiting time at Q_n in the second moment $v_{n,n}$ of the sum of Q_n's visit time and the subsequent switchover time; next they obtain a linear relation between $v_{n+1,n+1}$ and $v_{n,n}$ for all n; and finally they solve for the $v_{n,n}$ by deriving an extra linear relation between $v_{1,1}, \ldots, v_{N,N}$. At this last stage the conservation law is elegantly brought into the picture.

5. Conclusions

This paper has been devoted to single-server multi-class service systems with interruptions. The main results are:

- Under rather weak restrictions, the steady-state amount of work in the system, $V^{SD,I}$, is distributed as the sum of two independent quantities, viz. (i) V, the steady-state amount of work in the corresponding system with identical characteristics but without interruptions, and (ii) Y, the steady-state amount of work in the original system at an epoch at which the server is not serving. V does not depend on the scheduling discipline, nor on the interruption process; all the information provided by those two system entities is contained in Y.
- Under stronger restrictions, a pseudoconservation law holds for the mean waiting times of the classes of customers.

The pseudoconservation law has already proved its usefulness in cyclic-service systems. In section 4 it has been shown how one can employ the pseudoconservation law to derive approximations for individual mean waiting times.

The righthand side of the pseudoconservation law contains EY. Evaluation of EY has been discussed in many special cases, mainly derived from polling systems. In view of the importance of such systems in computer-communication networks, a further study of EY for various service strategies at the queues is of interest. However, in relation to conservation principles in service systems there are several more challenging and fundamental problems to be solved. We end this paper with a list of some of those problems.

1. Relaxation of the Poisson assumption

The assumption of Poisson arrival processes is not always realistic. Extension of the work decomposition in theorem 2.1, and of the pseudoconservation law in theorem 3.3, to more general arrival processes would be of considerable interest. It should be noted that some of the waiting time decompositions for queues with vacations hold for general interarrival time distributions (cf. Doshi [11,12]). Of course, when a work decomposition for such a case is obtained, there is still the problem that the mean amount of work in a G/G/1 queue is not known.

2. Multi-server queues

As observed in remark 3.1, extension of the conservation law to multi-server queues without interruptions has only been accomplished under the severe restriction of equal service time distributions for all classes. It would be interesting to study the concepts of work conservation, work decomposition and (pseudo)conservation law for multi-server multi-class systems. Thus new insight might be obtained into the behavior of cyclic-service systems with multiple servers, a subject which has received relatively little attention but for which

several applications exist. Takagi [45] contains the references to the few papers that have appeared on this subject.

3. Optimization

Although the extensive research on cyclic-service system has been useful for performance evaluation, it has not yet led to a clear ability to control the systems under consideration and to affect their design. Modern developments in computer and communication technology enable the use of more sophisticated scheduling disciplines, while the need to control complex networks makes the use of such disciplines imperative. Recently a few studies have appeared which open up possibilities for optimization; much more research is needed here. Levy's [35] binomial-gated strategy (cf. § 3.3) leads to a tractable mathematical model in which the choice of binomial probabilities of numbers of customers served at the queues allows prioritization. Levy et al. [37] compare several service disciplines w.r.t. the total amount of work in the system. Using a sample path analysis they show that some policies dominate other policies in the sense that, *at any time*, the total amount of unfinished work in the system under one policy is at most as large as under another policy. The analysis can be used to construct a hierarchy of several common service disciplines.

Browne and Yechiali [9] present a semi-dynamic polling policy in which the server, at the beginning of a cycle, determines a visiting order of the queues for this cycle so as to minimize some objective function. Finally, the use of a fixed service order table (cf. [1,6] and § 3.4) enables the assignment of priorities by listing a queue more often in the table, and a similar remark holds for polling systems in which the server visits the classes according to a Markov routing chain (cf. [33,8] and § 3.5).

Acknowledgment

The author is indebted to J.W. Cohen, B.T. Doshi, S.W. Fuhrmann, W.P. Groenendijk, Y.T. Wang and several other colleagues for interesting discussions and for useful comments on an earlier draft of this paper.

References

[1] J.E. Baker and I. Rubin, Polling with a general-service order table, IEEE Trans. Commun., COM-35 (1987) 283–288.

[2] D.P. Bertsekas and R.G. Gallager, *Data Networks* (Prentice-Hall, Inc., Englewood Cliffs, 1987).

[3] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, J. Appl. Prob. **24** (1987) 949–964.

[4] O.J. Boxma and W.P. Groenendijk, Waiting times in discrete-time cyclic-service systems, IEEE Trans. Commun. COM-**36** (1988) 164–170.
[5] O.J. Boxma and W.P. Groenendijk, Two queues with alternating service and switching times, In: *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen*, eds. O.J. Boxma and R. Syski (North-Holland, Amsterdam, 1988) 261–282.
[6] O.J. Boxma, W.P. Groenendijk and J.A. Weststrate, A pseudoconservation law for service systems with a polling table, Report Centre for Mathematics and Computer Science, Amsterdam, 1988; to appear in IEEE Trans. Commun.
[7] O.J. Boxma and B. Meister, Waiting-time approximations for cyclic-service systems with switch-over times, Performance Evaluation Review **14** (1986) 254–262.
[8] O.J. Boxma and J.A. Weststrate, Waiting times in polling systems with Markovian server routing, Report Centre for Mathematics and Computer Science, Amsterdam, 1989; to appear in: *Messung, Modellierung und Bewertung von Rechensystemen und Netzen*, eds. G. Stiege and J.S. Lie (Springer, Berlin, 1989).
[9] S. Browne and U. Yechiali, Dynamic priority rules for cyclic-type queues, Adv. Appl. Prob. **21** (1989) 432–450.
[10] S.L. Brumelle, On the relation between customer and time averages in queues, J. Appl. Prob. **8** (1971) 508–520.
[11] B.T. Doshi, Queueing systems with vacations–a survey, Queueing Systems **1** (1986) 29–66.
[12] B.T. Doshi, Generalizations of the stochastic decomposition results for single server queues with vacations, Report AT&T Bell Labs, Holmdel (NJ), 1988.
[13] D.E. Everitt, Simple approximations for token rings, IEEE Trans. Commun. COM-**34** (1986) 719–721.
[14] D.E. Everitt, A conservation-type law for the token ring with limited service, Br. Telecom Technol. *J.* **4** (1986) 51–61.
[15] D.E. Everitt, A note on the pseudoconservation laws for cyclic service systems with limited service disciplines, IEEE Trans. Commun. COM-**37** (1989) 781–783.
[16] A. Federgruen and H. Groenevelt, M/G/*c* queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules, Management Sci. **34** (1988) 1121–1138.
[17] M.J. Ferguson and Y.J. Aminetzah, Exact results for nonsymmetric token ring systems, IEEE Trans. Commun. COM-**33** (1985) 223–231.
[18] S.G. Foss, Queues with customers of several types, In: *Advances in Probability Theory: Limit Theorems & Related Problems*, ed. A.A. Borovkov (Optimization Software Inc., New York, 1984).
[19] S.W. Fuhrmann, Inequalities for cyclic service systems with limited service, In: *Proc. GLOBECOM '87.*
[20] S.W. Fuhrmann and R.B. Cooper, Stochastic decompositions in the M/G/1 queue with generalized vacations, Oper. Res. **33** (1985) 1117–1129.
[21] S.W. Fuhrmann and Y.T. Wang, Mean waiting time approximations of cyclic service systems with limited service, In: *Performance '87*, eds. P.-J. Courtois and G. Latouche (North-Holland Publ. Cy., Amsterdam, 1988) 253–265.
[22] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems* (Academic Press, New York, 1980).
[23] N.P. Giannakouros and A. Laloux, A general conservation law for a priority polling system, Report Telecommunications Laboratory, Leuven University, 1988.
[24] W.P. Groenendijk, Waiting-time approximations for cyclic-service systems with mixed service strategies, In: *Proc. 12th ITC* (North-Holland Publ. Co., Amsterdam, 1989).
[25] W.P. Groenendijk, A conservation-law based approximation algorithm for waiting times in polling systems, Report Centre for Mathematics and Computer Science, Amsterdam, 1988.

[26] S. Halfin, Batch delays versus customer delays, Bell System Tech. J. **62** (1983) 2011–2015.
[27] D.P. Heyman and M.J. Sobel, *Stochastic Models in Operations Research*, Vol. I. (McGraw-Hill Book Company, New York, 1982).
[28] T. Katayama, A cyclic service tandem queueing model with semi-exhaustive service, Report NTT Communication Switching Laboratories, Tokyo, 1987; to appear in the *Proc. Int. Seminar on Performance of Distributed and Parallel Systems, Kyoto, Japan.*
[29] J. Keilson and L.D. Servi, Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules, J. Appl. Prob. **23** (1986) 790–802.
[30] L. Kleinrock, *Communication Nets – Stochastic Message Flow and Delay* (Dover, New York, 1964).
[31] L. Kleinrock, A conservation law for a wide class of queueing disciplines, Naval Res. Logist. Quart. **12** (1965) 181–192.
[32] L. Kleinrock, *Queueing Systems*, Vol. II (Wiley, New York, 1976).
[33] L. Kleinrock and H. Levy, The analysis of random polling systems, Oper. Res. **36** (1988), 716–732.
[34] G.P. Klimov, Time-sharing service systems, I, Theory of Prob. and its Appl. **19** (1974) 532–551.
[35] H. Levy, Optimization of polling systems via binomial service, Report Department of Computer Science, Tel-Aviv University, 1988.
[36] H. Levy and M. Sidi, Correlated arrivals in polling systems, Report Department of Computer Science, Tel-Aviv University, 1988.
[37] H. Levy, M. Sidi and O.J. Boxma, Dominance relations in polling systems, Report Department of Computer Science, Tel-Aviv University, 1988; to appear in Queueing Systems.
[38] S.S. Nair, A single server tandem queue, J. Appl. Prob. **8** (1971) 95–109.
[39] T.J. Ott, On the M/G/1 queue with additional inputs, J. Appl. Prob. **21** (1984) 129–142.
[40] J.W.M. Pang and R.W. Donaldson, Approximate delay analysis and results for asymmetric token-passing and polling networks, IEEE J. Sel. Areas in Commun. SAC-**4** (1986) 783–793.
[41] L. Schrage, An alternative proof of a conservation law for the queue G/G/1, Oper. Res. **18** (1970) 185–187.
[42] M. Sidi and H. Levy, A queueing network with a single cyclically roving server, Report Electrical Engineering Department, Technion, 1988.
[43] M.M. Srinivasan, An approximation for mean waiting times in cyclic server systems with nonexhaustive service, Performance Evaluation **9** (1988) 17–33.
[44] H. Takagi, *Analysis of Polling Systems* (The MIT Press, Cambridge, Massachusetts, 1986).
[45] H. Takagi, Queuing analysis of polling models, ACM Comp. Surveys **20** (1988) 5–28.
[46] M. Taube-Netto, Two queues in tandem attended by a single server, Oper. Res. **25** (1977) 140–147.
[47] Tedijanto, Exact results for the cyclic-service queue with a Bernoulli schedule, Report Electrical Engineering Department and Systems Research Center, University of Maryland, 1988.
[48] K.S. Watson, Performance evaluation of cyclic service strategies–a survey, In: *Performance '84*, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam, 1988) 521–533.
[49] R.W. Wolff, Work-conserving priorities, J. Appl. Prob **7** (1969) 327–337.
[50] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. **30** (1982) 223–231.