Periodic versus exhaustive service in a multi-product production center

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 95-01

Periodic versus exhaustive service in a multi-product production center

M.J.A. van Eenige
I.J.B.F. Adan
J.A.C. Resing
J. van der Wal

Eindhoven, January 1995
The Netherlands

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariat: Main building 9.10
Telephone: 040-47 3130

ISSN 0926 4493

Periodic versus exhaustive service in a multi-product production center

M.J.A. van Eenige, I.J.B.F. Adan, J.A.C. Resing, and J. van der Wal
Eindhoven University of Technology
Department of Mathematics and Computing Science
P.O. Box 513
5600 MB Eindhoven
The Netherlands

Abstract

In this paper, we consider a single machine at which two types of orders with deterministic processing times arrive according to independent Bernoulli processes. When the machine switches from processing one type of order to another, fixed switch-over times are incurred. For both types of orders, we determine the delivery-time distribution for a periodic switching strategy and for an exhaustive service policy. Using these delivery-time distributions, we compare for these switching strategies both the average delivery time and the fraction of orders that are processed in time. Numerical examples show that a periodic switching strategy performs considerably worse than an exhaustive service policy. For the periodic switching strategy, however, delivery times of orders can be guaranteed, which is not the case for the exhaustive service policy. Hence, for this latter policy, we also determine the expected delivery times of orders if for each state upon arrival of an order the promised delivery time has to be met with a certain probability. When this probability is high, the periodic switching strategy might outperform the exhaustive service policy. Finally, we compute the switching strategy that minimizes the number of orders at the machine, and compare it with the aforementioned switching strategies.

1 Introduction

In production centers, machines usually have to (partly) manufacture different types of products. Whenever such a machine switches from producing one type of product to another, switch-over times are incurred, for instance, to clean and tune up the machine, or to replace tools. For these production centers, important performance measures are the average delivery time of an order and the fraction of orders that is handled in time. These performance measures obviously depend on the switching strategy of the machine. In this paper, we consider two kinds of switching strategies, which occur in a natural way in practice: a periodic and an exhaustive strategy. Of course, other switching strategies are also possible, e.g., a strategy that optimizes some objective function like the number of orders or the total amount of work at the machine. In general, however, such optimal strategies are quite complex and therefore not suitable for implementation in practice.

In a periodic switching strategy, the machine switches at predetermined and periodic time epochs. So, for each type of order, time can be divided into a periodic pattern of intervals of fixed length during which the machine is either available to this type of order (*on-periods*) or not (*off-periods*). One such fixed periodic sequence of on- and off-periods will be called a cycle. When implementing a periodic switching strategy, the production center has to define for each type of order the number of on-periods in a cycle (of course at least one), the length of each of these on-periods, and their order in the cycle.

In an exhaustive service policy, orders for each type of product are handled exhaustively before the machine switches to handling orders for another type of product. For this kind of switching strategy, the production center has to decide in which order the machine is available to the different types of orders, and it has to define a rule which governs the behaviour of the machine when there are no orders at all at the machine.

A periodic switching strategy has several important practical advantages with respect to an exhaustive service policy. For instance, it is easier to control the inventory level of materials and it is possible to give accurate estimates of the delivery times of orders. It may be expected, however, that a periodic switching strategy performs worse with respect to the aforementioned performance measures than an exhaustive service policy. The aim of this paper is to get some insight in the change in performance when a periodic and more practical switching strategy is preferred to an exhaustive service policy.

To simplify the analysis of these switching strategies, we consider a discrete-time queueing system with two types of jobs or products, which are manufactured by a common machine. For each type of job, orders arrive according to a Bernoulli process and have deterministic service times. Whenever the machine switches from handling orders for one type of job to the other, fixed switch-over times are incurred. It is natural to model these queueing systems as systems in which each type of order has its own queue. Therefore, we will henceforth say that the server switches between queues rather than between job types. For the periodic switching strategy, we assume that a cycle consists of exactly one on-period for each type of job. For the exhaustive service policy, we use the natural rule that, when there are no orders at the machine, the production center waits for the next order before switching. For both switching strategies, we determine the delivery-time distribution for each type of order, so that we are able to compute and compare the aforementioned performance measures. Numerical results show that the average delivery time of an order and the probability that an order is not served in time are in general considerably larger for a periodic switching strategy than for the exhaustive service policy. Only when the production center has to give delivery times based on the number of orders at the machine upon arrival and these delivery times have to be met with a very high probability, it turns out that a periodic switching strategy might outperform the exhaustive service policy. Finally, we compare these two switching strategies with the switching policy that minimizes the number of orders at the machine. Our numerical examples show that when the service times of the different type of jobs do not differ much, the performance of the exhaustive service policy and the optimal one are almost the same. In a forthcoming paper, we will focus on the rate at which the performance of a periodic switching strategy can be improved when introducing opportunities to work in overtime and to produce on stock.

Discrete-time queueing systems with a periodic switching strategy have been studied by several authors. By using a generating function technique, Newell (cf. [11]) obtained approximations for the expected number of orders in a queue just before the machine switches to the other queue. Darroch (cf. [3]) derived bounds on the same quantity. Van Eenige, Resing and Van der Wal (cf. [4]) applied the matrix-geometric approach of Neuts [10] to this kind of queueing system, and derived, for a special case, the sojourn-time distribution of an arbitrary arriving order.

The queueing system with the exhaustive service policy considered here is usually called a polling system with a dormant (also called stopping or patient) server. Most of the literature on polling systems, however, is focussed on the case that the machine continuously switches when there are no orders at this machine (see, e.g., Takagi [13] for a review). Continuous-time polling systems with a dormant server were first analysed in Eisenberg [5] for the two queue case. For each type of order, Eisenberg derived the Laplace-Stieltjes transform for the waiting-time distribution. Only recently, this polling system again received some attention. Liu, Nain and Towsley (cf. [9]) showed for the symmetric case that this exhaustive service policy is optimal in the sense that for this policy the total amount of unfinished work in the system and the number of orders in the system are minimal. Eisenberg [6], and Srinivasan and Gupta [12] extended, among other things, the former results of [5] to the case of an arbitrary number of queues and to other switching rules when the system is empty. Borst (cf. [2]), and Srinivasan and Gupta (cf. [12]) derived a pseudo-conservation law for these polling systems. Finally, Blanc and Van der Mei (cf. [1]) used the power-series algorithm for a numerical analysis of these queueing systems.

The outline of the paper is as follows. In Section 2, the two queueing systems will be described in detail. The analysis of the queue-length processes of these models will be the topic of Section 3, together with some remarks with respect to the determination of the switching strategy that minimizes the number of orders at the machine. Numerical examples will be given in Section 4, and the paper will be concluded with some final remarks in Section 5.

2 The models

In this section, we give a detailed description of the models. We first describe the common characteristics of both models, and then for each of the models the specific ones.

We consider a discrete-time queueing system with two queues, which are attended by a common server. Whenever the server switches from one queue to the other, switch-over times are incurred. Time is divided into intervals of equal length, henceforth called slots. During each slot, the server is either serving a customer, or idling at a queue, or switching from one queue to the other.

At queue i, $i = 1, 2$, customers arrive according to a Bernoulli process with parameter p_i ($0 < p_i < 1$) with deterministic service times equal to b_i slots. For each queue, customers are served in order of arrival and the arrival processes are assumed to be mutually independent. The switch-over times from queue 1 to queue 2 (from queue 2 to queue 1) are also deterministic, and equal to s_{12} slots (s_{21} slots). Hence, for these queueing systems, only the arrival processes are random.

We assume that all events, i.e., customer arrivals and service completions, occur at slot boundaries. Since we want to determine the sojourn-time distribution of an arbitrary customer, we are interested in the queue lengths at arrival instants. For convenience, we assume that service completions occur *just before* a slot boundary, and customer arrivals *just after* a slot boundary. In the sequel, we will say that a customer arriving (departing) at the slot boundary between slot $k-1$ and slot k (i.e., the k-th slot boundary) is arriving in slot k (departing in slot $k-1$). Also decisions whether to start service, to idle, or to switch are made at slot boundaries. We assume that such decisions are taken at the instants immediately after a possible arrival. For example, if a customer arrives at an empty queue which the server is attending, his service starts in the same slot in which he arrived.

In the sequel, we denote the queueing model with a periodic switching strategy as the PER-model, and the queueing model with the exhaustive service policy as the EXH-model.

2.1 The PER-model

In the PER-model, the server is assigned to the queues according to a predetermined (fixed) periodic schedule. More specifically, the server is first assigned to queue 1 for a_1 consecutive slots. Then, it switches from queue 1 to queue 2 (which takes s_{12} slots), after which the server is available to customers of queue 2 for a_2 consecutive slots. Finally, it switches back to queue 1 (which takes s_{21} slots), after which the periodic switching pattern starts over again. If, due to the periodic pattern, the server has to switch to the other queue during the service of a customer, upon return of the server this service is resumed where it was interrupted.

The time interval between two successive departures of the server from queue 1 is called a cycle. Clearly, the length of a cycle is equal to $a_1 + s_{12} + a_2 + s_{21}$ slots, and will be denoted by N. The slots in a cycle are numbered $1, 2, \ldots, N$.

2.2 The EXH-model

In the EXH-model, the server keeps serving customers of the attended queue as long as this queue is not empty at instants immediately after possible customer arrivals (recall that customer arrivals occur just after slot boundaries). If the attended queue is empty at this instant, then the server only switches if the other queue is not empty. Otherwise, the server idles at the queue it is attending until the next arrival at one of the queues.

3 The queue-length processes

In this section, we first analyse the queue-length processes of the PER-model and the EXH-model. The analysis of these two models are completely different, since in the PER-model the queue-length processes can be analysed separately, while in the EXH-model they have to be analysed simultaneously. We conclude this section with describing a way to approximate the switching policy that minimizes the number of customers in the system.

3.1 The PER-model

Since the queue-length processes do not affect each other, they can be analysed separately. In the remaining part of this subsection, we will consider only one of the queues. For notational convenience, we will simply write b for the service times of customers of the queue to be analysed, a for the length of the on-period in a cycle for customers of this queue, and p for the arrival rate of these customers, respectively. Furthermore, we will assume without loss of generality that these a slots are the last a slots in the cycle.

Van Eenige, Resing and Van der Wal (cf. [4]) considered a more general model, and applied the matrix-geometric approach of Neuts to the analysis of the queue-length process. This approach involves the determination of the rate matrix $\boldsymbol{R}$ (see [10]). For the present problem, the dimension of $\boldsymbol{R}$ is Nb and can be very large, so that the matrix-geometric technique is of limited value for numerical purposes. For this reason, we have sought an alternative (i.e., more efficient) way to analyse the queue-length process.

3.1.1 Analysing the queue embedded at the beginning of the cycle

Consider the queueing system embedded at the beginning of the cycle (i.e., at the beginning of the *off-period*), and let Y_k denote the total number of slots work in the queue at the beginning of the k-th cycle. Then, the process $\{Y_k\}_{k\in\mathbb{N}_0}$ is a discrete-time parameter Markov chain with state space consisting of the nonnegative integers $\{j|j = 0,1,2,\ldots\}$. We assume that this chain is ergodic (i.e., $Npb < a$), and hence, that the stationary distribution $\{\pi_j, j = 0,1,2,\ldots\}$ exists. This stationary distribution is the unique normalized solution of the one-cycle equilibrium equations

$$\pi_j = q_{0,j}\pi_0 + q_{1,j}\pi_1 + \cdots + q_{j+a,j}\pi_{j+a}, \qquad j = 0,1,2,\ldots, \tag{1}$$

where $q_{i,j}$ denotes the one-cycle transition probability from state i to state j, i.e.,

$$q_{i,j} := \Pr\{Y_1 = j|Y_0 = i\}.$$

We will first show that for $j \geq Nb - a$ the transition probabilities $q_{i,j}$ satisfy

$$q_{i,j} = \begin{cases} \beta_m & \text{if } i = j - mb + a \text{ for some } m = 0,1,2,\ldots,N, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where β_m denotes the probability of m customers arriving in a cycle consisting of N slots, i.e.,

$$\beta_m := \binom{N}{m} p^m(1-p)^{N-m}, \quad m = 0,1,2,\ldots,N.$$

Hence, for $j \geq Nb - a$, the equilibrium equations (1) reduce to

$$\pi_j = \beta_N\pi_{j-(Nb-a)} + \beta_{N-1}\pi_{j-((N-1)b-a)} + \cdots + \beta_0\pi_{j+a}.$$

Later on, we will use the structure of these equations to determine the stationary distribution.

To prove (2), let $j \geq Nb - a$. If the server is idle in a slot between $N - a$ and N, then, at the end of the cycle there would be no more than $(a-1)(b-1) < N(b-1) \leq Nb - a$ slots of work. So, if at the end of the cycle the number of slots work equals j, the server cannot have been idle in the cycle. Hence, $q_{i,j} = 0$ for $i \neq j - mb + a$ for any $m = 0, 1, \ldots, N$. If $i = j - mb + a$ for some $m = 0, 1, \ldots, N$, then there should have been exactly m arrivals in the cycle. Even if these m arrivals take place in the last m slots, the server cannot become idle as there are $i = j - mb + a \geq (N - m)b \geq N - m$ slots of work at the beginning of the cycle. Thus, any pattern of m arrivals prevents the server from becoming idle, and lead to j slots of work at the end of the cycle. Hence, $q_{i,j} = \beta_m$ if $i = j - mb + a$ for some $m = 0, 1, \ldots, N$.

So, the equilibrium equations (1) can be partitioned as follows

$$\pi_j = q_{0,j}\pi_0 + q_{1,j}\pi_1 + \cdots + q_{j+a,j}\pi_{j+a}, \qquad j = 0, 1, 2, \ldots, Nb - a - 1, \quad (3)$$
$$\pi_j = \beta_N \pi_{j-(Nb-a)} + \beta_{N-1}\pi_{j-((N-1)b-a)} + \cdots + \beta_0 \pi_{j+a}, \qquad j = Nb - a, Nb - a + 1, \ldots, \quad (4)$$

Equations (4) are called inner conditions and equations (3) are called boundary conditions. In the boundary conditions, for $i \geq a$, the transition probabilities $q_{i,j}$ are also given by (2). For $i < a$, the transition probabilities $q_{i,j}$ can iteratively be obtained from the one-slot transition probabilities. These one-slot transition probabilities can easily be determined, since in each slot at most one customer can arrive at the queue and at most one slot of work can be handled.

As already mentioned, the stationary distribution is the unique normalized solution of the equilibrium equations. We will seek solutions of the inner conditions of the form $\pi_j = z^j$, and use these solutions to construct a linear combination that also satisfies the boundary conditions and the normalization, such that we finally obtain

$$\pi_j = \sum_i c_i z_i^j, \quad j = 0, 1, 2, \ldots,$$

where the c_i's denote the coefficients of the linear combination.

Inserting $\pi_j = z^j$ in the equations (4) and dividing by $z^{j-(Nb-a)}$ yields

$$z^{Nb-a} = \beta_0 z^{Nb} + \beta_1 z^{(N-1)b} + \cdots + \beta_N. \tag{5}$$

Clearly, we are only interested in solutions z with $|z| < 1$. The following lemmas state that there are $Nb - a$ solutions with $|z| < 1$, and that exactly one of these solutions is positive and real, which is the largest solution in absolute value. The proofs of these lemmas can be found in Appendix A.

Lemma 1 *Equation (5) has exactly $Nb - a$ roots inside the unit circle.*

These $Nb - a$ roots will be denoted by $z_1, z_2, \ldots, z_{Nb-a}$.

Lemma 2 *There is exactly one root in the interval* $(0, 1)$.

This root will be denoted by z_1.

Lemma 3 *If $Nb - a$ is not a multiple of b, then $|z_i| < |z_1|$ for $i = 2, 3, \ldots, Nb - a$. If $Nb - a$ is a multiple of b, there are $b - 1$ roots, $z_2, z_3, \ldots, z_b$, say, with $|z_i| = |z_1|$, and $|z_i| < |z_1|$ for $i = b + 1, b + 2, \ldots, Nb - a$.*

For convenience, we assume that the $Nb-a$ roots of equation (5) are distinct, so that we have $Nb-a$ basic solutions z_i^j satisfying the inner conditions. Now, we will try to represent π_j as a linear combinations of these basic solutions

$$\pi_j = \sum_{i=1}^{Nb-a} c_i z_i^j, \qquad j = 0, 1, 2 \ldots \tag{6}$$

The assumption that the roots of equation (5) are distinct is not restrictive, because if, for instance, roots z_i and z_{i+1} are identical, z_{i+1}^j should be replaced in (6) by jz_i^j, so that the number of basic solutions is sufficient to represent the stationary distribution $\{\pi_j, j = 0, 1, 2, \ldots\}$. Similarly, if there are more roots which are equal to z_i, then higher powers of j should be used. It remains to determine the $Nb-a$ coefficients c_i such that the boundary conditions and the normalization equation are satisfied.

Since the equilibrium equations are dependent, one of the boundary conditions in (3) may be omitted, the boundary condition for state $j = 0$, say. Substituting the form (6) with the unknown coefficients c_i into the reduced set of boundary conditions and the normalization equation leads to a set of linear equations. The solution of this set of equations is unique and it yields the desired coefficients c_i.

To use solution (6), we first have to find the $Nb-a$ roots of equation (5), and then we have to solve the boundary conditions and the normalization equation for the coefficients c_i. The single positive root, i.e., z_1, can easily be determined numerically, e.g., by using bisection. Finding all roots numerically, however, can be difficult. These difficulties especially occur when $Nb-a$ is large or when (some of) the roots are closely clustered inside the unit circle. Moreover, even if we are able to compute all roots accurately, the set of equations consisting of the boundary conditions (after omitting the equation for state $j = 0$) and the normalization equation is close to singular if the roots are closely clustered (see, e.g., Section 1.6 in Neuts [10]). Therefore, we will not use the exact solution (6) for numerical purposes. Instead, we will approximate the stationary distribution numerically by exploiting the asymptotic behaviour of the stationary probabilities. In the next subsection, it will be shown that this behaviour is determined by the single positive root z_1.

3.1.2 Numerical approach

The asymptotic behaviour of the stationary distribution (6) is obviously determined by the root(s) with largest absolute value. By Lemma 3, this root (z_1) is unique if $Nb-a$ is not a multiple of b, and there are b largest roots $(z_1, z_2, \ldots, z_b)$ if $Nb-a$ is a multiple of b. In the former case, we have

$$\lim_{j\to\infty} \frac{\pi_{j+1}}{\pi_j} = z_1, \tag{7}$$

and hence, the tail behaviour of the stationary distribution is geometric. Since this limit does not exist in the latter case, we consider in this case the quotient π_{j+b}/π_j, and, by using $z_i^b = z_1^b$ for $i = 2, 3, \ldots, b$, we obtain

$$\lim_{j\to\infty} \frac{\pi_{j+b}}{\pi_j} = \lim_{j\to\infty} \frac{\sum_{i=1}^{Nb-a} c_i z_i^{j+b}}{\sum_{i=1}^{Nb-a} c_i z_i^j} = \lim_{j\to\infty} \frac{z_1^b \sum_{i=1}^{b} c_i z_i^j + \sum_{i=b+1}^{Nb-a} c_i z_i^{j+b}}{\sum_{i=1}^{b} c_i z_i^j + \sum_{i=b+1}^{Nb-a} c_i z_i^j} = z_1^b. \tag{8}$$

So, the single positive root characterizes the asymptotic behaviour of the stationary distribution completely. This behaviour will be exploited by using the numerical approach of Tijms and Van de Coevering (cf. [15]).

In [15], the starting point is the assumption of the existence of limit (7). Since this limit does not exist if b is a divisor of $Nb-a$, their approach is slightly adapted for the present problem by exploiting the asymptotic behaviour of π_{j+b}/π_j instead of π_{j+1}/π_j. We choose a positive integer J, and set

$$\pi_j = z_1^b \pi_{j-b}, \qquad j \geq J+1. \tag{9}$$

So, we implicitly assume that z_1^b is a (fairly) good approximation for the quotient π_{j+b}/π_j for $j > J$. The probabilities $\pi_0, \pi_1, \ldots, \pi_J$ are then obtained by solving the equilibrium equations for $j = 0, 1, 2 \ldots, J$ (in which approximation (9) is substituted), and the normalization equation, which reads as follows

$$\sum_{j=0}^{J-b} \pi_j + \sum_{i=1}^{b} \frac{1}{1-z_1^b} \pi_{J-b+i} = 1.$$

The value of J (which obviously depends on the required accuracy) has to be determined experimentally.

As suggested in [15] and confirmed by our numerical examples, the asymptotic behaviour appears rather quickly, i.e., for already small values of J. In particular, for our examples, J is much smaller than $Nb-a$, which is the number of boundary conditions needed to be solved for the exact analysis, and consequently, J is also much smaller than the dimension of the rate matrix Nb which has to be determined when using the matrix-geometric approach.

From this approximated stationary distribution, we can determine the sojourn-time distribution of an arbitrary arriving customer as follows. Since in each slot at most one customer can arrive and at most one slot work can be handled, the stationary distribution of the number of slots work at the n-th slot boundary can easily be computed once the stationary distribution at the $(n-1)$-st slot boundary is known. So, from the stationary distribution $\{\pi_j, j = 0, 1, 2 \ldots\}$, we can iteratively compute the stationary distribution of the number of slots work at the n-th slot boundary, $n = 2, 3, \ldots, N$, in the cycle. Due to the strictly periodic switching pattern, the number of slots work upon arrival, uniquely determines the slots in the cycle in which the service of this customer is started, interrupted, and completed. Hence, given the slot of arrival and (by using the Bernoulli-arrivals-see-time-average property (cf. Halfin [7])) the stationary distribution of the number of slots work upon arrival, we easily obtain the sojourn-time distribution of this customer.

Finally, we remark that the above analysis does not depend on the number of job types (or the number of queues). For this analysis, it is only important that for each queue a cycle contains exactly one on-period of fixed length.

3.2 The EXH-model

In contrast to the PER-model, the queue-length processes of the EXH-model cannot be studied separately. In order to fully characterize the latter model at slot boundaries, we need the number of customers in queue 1 and queue 2, the position of the server (i.e., the server is at one of the queues or is switching), and the residual number of slots service or switching at these slot boundaries. Let X_k^i denote the number of customers in queue i, $i = 1, 2$, at the k-th slot boundary. Let Y_k denote the position of the server during slot $k-1$. More specifically, if the server is at queue 1 (at queue 2) during slot $k-1$, we set $Y_k := 1$ ($Y_k := 2$), and if the server is switching from queue 1 to queue 2 (from queue 2 to queue 1) during slot $k-1$, we set $Y_k := 3$ ($Y_k := 4$). Let Z_k denote the residual number of slots service of the customer in service (i.e, if $Y_k \in \{1,2\}$), or the residual number of slots switching (i.e., if $Y_k \in \{3,4\}$), at the k-th slot boundary. If the queue, which is attended by the server at the k-th slot boundary, is empty, or if a customer departure occurred in slot $k-1$, we set $Z_k := 0$. The process $\{(X_k^1, X_k^2, Y_k, Z_k)\}_{k\in\mathbb{N}}$ is an aperiodic and irreducible discrete-time parameter Markov chain with state space

$$\begin{aligned}\mathcal{S} \;=\;& \{(0,l_2,1,0)|l_2\in\mathbb{N}_0\} \cup \{(l_1,l_2,1,n)|l_1\in\mathbb{N}, l_2\in\mathbb{N}_0, n=0,1,2\ldots,b_1-1\} \cup \\ & \{(l_1,0,2,0)|l_1\in\mathbb{N}_0\} \cup \{(l_1,l_2,2,n)|l_1\in\mathbb{N}_0, l_2\in\mathbb{N}, n=0,1,2\ldots,b_2-1\} \cup \\ & \{(l_1,l_2,3,n)|l_2\in\mathbb{N}, n=0,1,\ldots,s_{12}-1, l_1=0,1,\ldots,s_{12}-1-n\} \cup \\ & \{(l_1,l_2,4,n)|l_1\in\mathbb{N}, n=0,1,\ldots,s_{21}-1, l_2=0,1,\ldots,s_{21}-1-n\}.\end{aligned}$$

A necessary and sufficient condition for positive recurrence is that the average number of slots work brought to the system per slot is less than 1, i.e.,

$$p_1b_1 + p_2b_2 < 1.$$

Henceforth, we will assume that this condition is satisfied, so that this Markov chain has a stationary distribution.

To our knowledge, so far only expressions for the generating function of the queue-length process and the Laplace-Stieltjes transform of the waiting-time and sojourn-time distributions are known for the continuous-time version of the EXH-model (see, e.g., Eisenberg [5] and [6], and Srinivasan and Gupta [12]). Since we are mainly interested in tail probabilities of the sojourn-time distribution, we use a numerical approach for approximating the stationary distribution of the aforementioned Markov chain. Once this (approximated) stationary distribution is known, we can give explicit expressions for the sojourn-time distribution of an arbitrary customer.

For the numerical approach, we assume that the buffers at both queues are finite. The size of these buffers is chosen sufficiently large. Customers arriving at a queue with a full buffer will be rejected. For the Markov chain, which corresponds to this model with finite buffers, we use the iteration method of Van der Wal and Schweitzer (cf. [16]) to solve the finite set of equilibrium equations and the normalization equation, so that we obtain an approximation for the stationary distribution of the original model.

From this approximated stationary distribution, we can obtain performance measures like the sojourn-time or waiting-time distribution of an arbitrary arriving customer. Here, we briefly outline the basic idea of the derivation of the sojourn-time distribution. For more details, we refer to Appendix B. We condition on the state of the system upon arrival, and determine the conditional sojourn-time distribution. If the server is at or switching to the queue at which the customer arrived, the sojourn time of this customer is completely known. Otherwise, the sojourn time of this customer is the sum of the fixed time to switch and to serve customers of his own queue, plus the random amount of time to serve customers of the other queue. The latter amount of time is basically the distribution of a busy period starting with a known initial number of slots work.

3.3 The optimal switching strategy

As mentioned in the Introduction, we also determine the switching strategy that minimizes the number of customers in the system for comparing it with the periodic and exhaustive strategy. To our knowledge, so far the structure of this policy is known for the (continuous-time) case that the service-time distributions of both type of customers are identical. In this case Liu, Nain and Towsley (cf. [9]), and Hofri and Ross (cf. [8]) show that each queue should be served exhaustively, and furthermore, Hofri and Ross conjecture that the server should only switch from an empty queue to the other queue if the number of customers in the latter queue exceeds a certain number. The queueing system with the optimal switching policy will be denoted by OPT-model. To find this optimal policy, we first model our problem as a Markov decision problem.

The decision epochs are the instants immediately after possible customer arrivals. The state of the system at these instants is characterized similarly as for the EXH-model. For all decision epochs k, the set of all possible states is

$$\begin{aligned}\hat{\mathcal{S}} \;=\; & \{(0,l_2,1,0)|l_2 \in \mathbb{N}_0\} \cup \{(l_1,l_2,1,n)|l_1 \in \mathbb{N}, l_2 \in \mathbb{N}_0, n = 0,1\ldots,b_1-1\} \cup \\ & \{(l_1,0,2,0)|l_1 \in \mathbb{N}_0\} \cup \{(l_1,l_2,2,n)|l_1 \in \mathbb{N}_0, l_2 \in \mathbb{N}, n = 0,1\ldots,b_2-1\} \cup \\ & \{(l_1,l_2,3,n)|l_1,l_2 \in \mathbb{N}_0, n = 0,1,\ldots,s_{12}-1\} \cup \\ & \{(l_1,l_2,4,n)|l_1,l_2 \in \mathbb{N}_0, n = 0,1,\ldots,s_{21}-1\}.\end{aligned}$$

Note that $\hat{\mathcal{S}}$ differs slightly from $\mathcal{S}$, because in the EXH-model, during a switch the number of customers at the queue from which the server switches is smaller than the elapsed time of the switch. If at a decision epoch the server is already switching, there is only one possible decision: the server keeps switching. Otherwise, there are two possible actions: the server switches to the other queue or the server stays at the queue. If the server stays at the queue, it keeps serving customers (if any) of this queue. For ease of notation, we denote a state in $\hat{\mathcal{S}}$ by y or z instead of (l_1,l_2,m,n). The set of all possible decisions when the state of the system is y will be denoted by $\mathcal{A}(y)$. Since in each slot at most one customer can arrive at each queue and at most one customer can leave the system, the transition probabilities $p_{y,z}(a)$ can easily be deduced for all states $y,z \in \hat{\mathcal{S}}$ and actions $a \in \mathcal{A}(y)$. Finally, the direct costs for the forthcoming slot are the (expected) number of customers in that slot, i.e.,

$$c_y(a) := l_1 + l_2, \quad y = (l_1,l_2,m,n) \in \hat{\mathcal{S}},$$

when the system is in state y upon a decision epoch and action a is taken.

Since the state space $\hat{\mathcal{S}}$ is infinitely large, we will compute approximations for the optimal switching strategy and the corresponding minimal costs, by solving the system with finite buffers exactly. Here the capacity of the buffers is chosen sufficiently large to obtain good approximations for the minimal costs and the optimal switching strategy (at least, for states close to the origin, i.e., states which are not too close to the states with one of the buffers full). Customers who arrive at a full buffer are rejected. Let $\hat{\mathcal{S}}'$ be the (finite) state space of the finite buffer model, and $V_k(y)$ be the total expected costs with k slots left to the time horizon given initial state y. The value vectors $V_k(y)$ can be computed recursively by using the value-iteration method (see, e.g., Tijms [14]) for all $y \in \hat{\mathcal{S}}'$:

$$V_{k+1}(y) = \min_{a\in\mathcal{A}(y)} \{a : c_y(a) + \sum_{z\in\hat{\mathcal{S}}'} p_{y,z}(a)V_k(z),\} \tag{10}$$

where $V_0(z) = 0$ for all z. Let f_{k+1} denote the minimizing switching strategy obtained in the $(k+1)$-st iteration, and $g(f_{k+1})$ denote the average costs of the stationary strategy f_{k+1}. In each iteration, the following bounds on $g(f_{k+1})$ and the minimal average costs g^* are obtained

$$\min_{z\in\mathcal{S}'}\{V_{k+1}(z) - V_k(z)\} \leq g^* \leq g(f_{k+1}) \leq \max_{z\in\mathcal{S}'}\{V_{k+1}(z) - V_k(z)\}.$$

Under certain conditions, the difference

$$V_{k+1}(y) - V_k(y)$$

converges for all $y \in \hat{\mathcal{S}}'$ to the minimal average number of customers in the system when k goes to infinity. A sufficient condition for convergence is that each stationary policy satisfies the unichain property and the corresponding Markov chain is aperiodic (see, e.g., Tijms [14]). When the service times are larger than one, this condition is satisfied. As simple examples show, for service times equal to one, this condition may not hold. We iterate relation (10) until for all $y \in \hat{\mathcal{S}}'$

$$\max_{z\in\hat{\mathcal{S}}'}\{V_{k+1}(z) - V_k(z)\} - \min_{z\in\hat{\mathcal{S}}'}\{V_{k+1}(z) - V_k(z)\} \leq \epsilon \cdot \min_{z\in\hat{\mathcal{S}}'}\{V_{k+1}(z) - V_k(z)\},$$

with ϵ a prespecified accuracy level, so that the difference $g(f_{k+1}) - g^*$ is at most an ϵ-percentage of $g(f_{k+1})$.

4 Numerical results

In the Introduction, we already mentioned that delivery times of orders are an important performance measure for a production center. In this section, we give numerical results for the delivery times of orders of the different switching strategies, using the following basic example. In a production center, a machine is used to produce two types of products. This production center can run 8 hours a day for 5 days a week. In each of these hours (i.e., in each slot) at most one order arrives for a product of type 1 and at most one order for a product of type 2. The probability of an arrival of these orders is p_1 and p_2, respectively. The time for producing one product is equal to 2 hours for products of type 1, and 3 hours for products of type 2. When the production center decides to change from handling orders of type 1 to orders of type 2 or vice versa, the machine has to be cleaned thoroughly. The cleaning of the machine takes 6 hours. Finally, we define $\rho = 2p_1 + 3p_2$, i.e., the average number of hours work arriving at the machine per hour.

Before giving numerical results concerning the delivery time of orders, we first make some remarks about the length of the on-periods and how quick the geometric behaviour of the stationary queue-length distribution appears in the PER-model. Recall that we have not specified yet the length of the on-periods (i.e., a_1 and a_2). For the numerical examples, these lengths are chosen such that the average number of orders at the machine is minimized, so that we can make a fair comparison with the OPT-model. For several examples, the optimal values for a_1 and a_2 are given in Table 1.

We see that balancing the load of the queues ρ_i is often optimal and that the lengths of the on-periods are mostly multiples of the service times. The corresponding cycle length N turns out (of course) to depend on the load of the system, but seems rather insensitive for values of the arrival rate of orders which give this load. In fact, N seems to be roughly $\mathcal{O}(1/(1-\rho))$.

ρ	p_1	p_2	N	Type 1 a_1	Type 1 ρ_1	Type 2 a_2	Type 2 ρ_2
0.50	$\frac{1}{10}$	$\frac{1}{10}$	47	14	0.67	21	0.67
	$\frac{1}{8}$	$\frac{1}{12}$	48	18	0.67	18	0.67
	$\frac{1}{16}$	$\frac{1}{8}$	49	10	0.61	27	0.68
0.75	$\frac{3}{20}$	$\frac{3}{20}$	92	32	0.86	48	0.86
	$\frac{3}{16}$	$\frac{1}{8}$	96	42	0.86	42	0.86
	$\frac{3}{32}$	$\frac{3}{16}$	102	24	0.80	66	0.87
0.90	$\frac{9}{50}$	$\frac{9}{50}$	226	86	0.95	128	0.95
	$\frac{9}{40}$	$\frac{3}{20}$	228	108	0.95	108	0.95
	$\frac{9}{80}$	$\frac{9}{40}$	238	58	0.92	168	0.96

Table 1: The cycle length and the length of the on-periods in hours, and the effective load of the queues.

To obtain the numerical results of the PER-model, we make use of the asymptotic behaviour of the stationary queue-length distribution. For very accurate results, this behaviour (cf. (8)) appears for type 1 orders for values of J which are about 60, 90 and 120, when ρ equals 0.50, 0.75 and 0.90, respectively. For orders of type 2, this behaviour appears for values of J which are about 100, 130 and 200, when ρ is 0.50, 0.70 and 0.90, respectively, with an exceptional 300 for the first example of the case 0.90. However, results with an accuracy of about three or four digits can already be obtained for much smaller values of J. Note that these values of J are much smaller than Nb_i, which is the dimension of the rate matrix $\boldsymbol{R}$ when using the matrix-geometric approach, and consequently they are much smaller than the number of boundary conditions to be solved when using the analytical approach in Section 3.1.1.

We now return to our performance measure of interest: the delivery times of orders. In Table 2, we give the average delivery time of both types of orders in the OPT-model, EXH-model, and PER-model. Numerical results suggest that the optimal switching policy is a so-called threshold

ρ	p_1	p_2	Type 1 product OPT	Type 1 product EXH	Type 1 product PER	Type 2 product OPT	Type 2 product EXH	Type 2 product PER
0.50	$\frac{1}{10}$	$\frac{1}{10}$	11.60	11.60	19.97	11.23	11.23	17.55
	$\frac{1}{8}$	$\frac{1}{12}$	10.72	10.72	17.02	11.93	11.93	20.54
	$\frac{1}{16}$	$\frac{1}{8}$	12.89	12.66	23.95	9.68	9.86	14.47
0.75	$\frac{3}{20}$	$\frac{3}{20}$	22.05	22.05	39.19	18.23	18.23	32.39
	$\frac{3}{16}$	$\frac{1}{8}$	19.51	19.51	32.39	20.27	20.27	39.26
	$\frac{3}{32}$	$\frac{3}{16}$	26.65	26.24	46.76	14.95	15.23	25.90
0.90	$\frac{9}{50}$	$\frac{9}{50}$	51.62	51.62	95.42	37.87	37.87	82.59
	$\frac{9}{40}$	$\frac{3}{20}$	43.69	43.69	80.00	44.29	44.29	98.57
	$\frac{9}{80}$	$\frac{9}{40}$	66.05	66.07	118.33	28.67	28.69	61.48

Table 2: Expected values for the delivery times of orders in hours.

policy, i.e., the machine switches only if the number of orders at the other queue is at least equal to a specific threshold. In most of these examples the thresholds for both queues are equal to 1, which implies that the optimal policy and exhaustive service policy coincide. Only for the latter examples of $\rho = 0.50, 0, 75, 0.90$ the thresholds are not equal to 1. When the machine is at queue 2, the threshold is 2 for each of these latter examples, and when the machine is at queue 1, the threshold is 1 for $\rho = 0.50$ and zero for $\rho = 0.75$ and $\rho = 0.90$. However, the differences between the average number of orders at the machine in the OPT-model and EXH-model are quite small. One may expect, however, that when the difference in service times increases, the optimal switching policy and the exhaustive service policy will differ more, and consequently, the differences between the average number of orders at the machine for these policies.

				Type 1 product				Type 2 product			
ρ	p_1	p_2	Model	α: 0.70	0.80	0.90	0.95	α: 0.70	0.80	0.90	0.95
0.50	$\frac{1}{10}$	$\frac{1}{10}$	EXH	15	18	22	26	14	16	19	22
			PER	28	31	35	39	24	28	30	35
0.50	$\frac{1}{8}$	$\frac{1}{12}$	EXH	14	17	20	24	15	17	21	24
			PER	24	28	31	33	28	32	35	42
0.50	$\frac{1}{16}$	$\frac{1}{8}$	EXH	16	19	24	29	13	15	17	19
			PER	33	37	41	47	20	23	26	29
0.75	$\frac{3}{20}$	$\frac{3}{20}$	EXH	27	34	43	53	23	27	33	39
			PER	52	59	68	80	42	47	58	70
0.75	$\frac{3}{16}$	$\frac{1}{8}$	EXH	24	30	38	45	25	30	37	44
			PER	45	51	56	65	51	57	70	84
0.75	$\frac{3}{32}$	$\frac{3}{16}$	EXH	32	40	53	66	19	22	26	31
			PER	63	72	80	92	35	39	47	57
0.90	$\frac{9}{50}$	$\frac{9}{50}$	EXH	65	81	106	130	47	58	74	90
			PER	124	139	165	192	101	119	151	182
0.90	$\frac{9}{40}$	$\frac{3}{20}$	EXH	55	68	89	108	55	68	88	106
			PER	105	118	138	161	121	141	178	214
0.90	$\frac{9}{80}$	$\frac{9}{40}$	EXH	82	104	140	175	35	42	55	66
			PER	155	174	203	235	75	89	115	141

Table 3: Percentiles for the delivery-time distribution of orders in hours.

Table 3 shows α-percentiles for the delivery times of both types of orders in the EXH-model and PER-model. The α-percentile is defined as the smallest integer s satisfying $\Pr\{S_i \le s\} \ge \alpha$, where S_i denotes the delivery time of an arbitrary order for a product of type i, $i = 1, 2$. As expected, Table 2 and Table 3 show that the performance of the EXH-model is considerably better than the performance of the PER-model. More precisely, the average delivery times in the PER-model are roughly twice as large as in the EXH-model. For the α-percentiles of the delivery-time distribution a similar remark can be made.

An advantage of the PER-model above the EXH-model is that, given the state of the system upon arrival, the production center can guarantee delivery times (and the average guaranteed delivery time is equal to $E\{S_i\}$ for orders of type i). In the EXH-model, this center generally cannot guarantee delivery times at all. Suppose that in the EXH-model, the production center promises delivery times based on the state of the system upon arrival of an order, and this delivery time has to be met at least with probability α. Let $\mathrm{E}\{D_{i,\alpha}\}$ denote the expectation of this promised delivery time for an arbitrary order of type i, that is,

$$\mathrm{E}\{D_{i,\alpha}\} = \sum_{y=(l_1,l_2,m,n)\in\mathcal{S}} s_{i,y}\tilde{\pi}_y,$$

with

$$s_{i,y} = \min\{s \mid \Pr\{S_i \leq s|\text{the state upon arrival is } y\} \geq \alpha\},$$

S_i the delivery time of a type i order, and $\tilde{\pi}_y$ the stationary probability of state y. Table 4 shows this expectation for several values of α. Comparing Table 2 and Table 4 shows that the differences between the promised delivery time in the EXH-model and the guaranteed delivery time in the PER-model decreases when α increases. In fact, when the promised delivery times have to be met with a very high probability the PER-model might outperform the EXH-model.

			Type 1 product				Type 2 product		
ρ	p_1	p_2	α : 0.90	0.95	0.99		α : 0.90	0.95	0.99
0.50	$\frac{1}{10}$	$\frac{1}{10}$	13.62	15.25	18.30		12.11	12.70	14.11
	$\frac{1}{8}$	$\frac{1}{12}$	12.20	13.25	15.62		13.42	13.98	15.64
	$\frac{1}{16}$	$\frac{1}{8}$	16.00	17.80	22.93		10.35	10.70	11.24
0.75	$\frac{3}{20}$	$\frac{3}{20}$	27.78	30.96	37.90		20.11	21.00	23.00
	$\frac{3}{16}$	$\frac{1}{8}$	23.22	25.19	29.75		23.25	24.69	27.73
	$\frac{3}{32}$	$\frac{3}{16}$	36.75	42.33	56.28		16.01	16.47	17.34
0.90	$\frac{9}{50}$	$\frac{9}{50}$	63.97	69.67	82.74		41.31	42.70	45.72
	$\frac{9}{40}$	$\frac{3}{20}$	51.19	54.61	61.98		50.08	52.51	57.59
	$\frac{9}{80}$	$\frac{9}{40}$	92.67	105.79	136.94		30.00	30.54	31.66

Table 4: The promised delivery times in hours in the EXH-model for different values of α.

5 Conclusions and final remarks

In this paper, we considered a discrete-time queueing system with two types of jobs served by a common server. Orders for these jobs arrive at their queue according to a Bernoulli process with deterministic service times. Whenever the server changes from serving one type of order to another, fixed switch-over times are incurred. For two switching strategies, we determined the sojourn-time distribution of both types of orders. For the first strategy, the server switches to a queue at predetermined and fixed periodic time epochs. For the second strategy, the server serves each queue exhaustively. The periodic switching strategies has several advantages which may be quite relevant in practice. However, the sojourn times of orders in the periodic switching strategy are considerably larger than in the exhaustive service policy. Only, when the promised delivery times have to be met with a very high probability, the periodic switching strategy might outperform the exhaustive service policy. In a forthcoming paper, we will investigate the rate at which the performance of the periodic switching strategy is improved when incorporating opportunities for working in overtime or producing on stock.

Further, numerical examples show that the exhaustive service policy practically minimizes the average number of orders in the system when the service times do not differ much. However, when the differences in service time between the types of orders become larger, one expects that the difference between the exhaustive service policy and the policy that minimizes the number of orders in the system (and hence, the difference in the average delivery time of an order for these switching strategies) also increases.

Appendix A

In this appendix, we prove Lemma 1, Lemma 2, and Lemma 3. Define

$$f(z) := z^{Nb-a} \quad \text{and} \quad g(z) := \beta_0 z^{Nb} + \beta_1 z^{(N-1)b} + \cdots + \beta_N.$$

A.1 Proof of Lemma 1

Lemma 1 *Equation $f(z) = g(z)$, i.e., equation (5), has exactly $Nb - a$ roots inside the unit circle.*

Proof. Since the Markov chain is ergodic, the average number of slots work arriving at the queue per cycle (i.e., pNb) is less than the average service capacity of the server for this queue per cycle (i.e., a), and thus

$$g'(1) = \sum_{i=0}^{N}(N-i)b\beta_i = Nb - pNb > Nb - a = f'(1).$$

However since $f(1) = g(1)$, there exists a sufficiently small $\epsilon > 0$ such that

$$f(1-\epsilon) > g(1-\epsilon). \tag{11}$$

For all z with $|z| = 1 - \epsilon$, we have

$$\begin{aligned} |g(z)| &\leq \beta_0|z^{Nb}| + \beta_1|z^{(N-1)b}| + \cdots + \beta_N \\ &= \beta_0(1-\epsilon)^{Nb} + \beta_1(1-\epsilon)^{(N-1)b} + \cdots + \beta_N = g(1-\epsilon), \\ |f(z)| &= |z^{Nb-a}| = (1-\epsilon)^{Nb-a} = f(1-\epsilon). \end{aligned}$$

Hence, together with inequality (11) we have

$$|g(z)| < |f(z)| \quad \text{for } |z| = 1-\epsilon.$$

Applying Rouchés Theorem to the circle $|z| = 1 - \epsilon$ and letting ϵ tend to zero, completes the proof. □

A.2 Proof of Lemma 2

Lemma 2 *There is exactly one root in the interval* $(0,1)$.

Proof. In order to prove this lemma, we divide equation (5) by z^{Nb-a}, to get

$$1 = \beta_0 z^a + \beta_1 z^{a-1} + \cdots + \beta_N z^{-(Nb-a)} =: P(z). \tag{12}$$

Since $P''(z) > 0$ for all $z > 0$, the polynomial $P(z)$ is strictly convex on $(0,\infty)$. Hence, equation (12) has at most two positive solutions. Since

$$\lim_{z\downarrow 0} P(z) = \infty \text{ (because } \beta_N > 0), \quad P(1) = 1, \quad \text{and} \quad P'(1) > 0,$$

the result follows immediately. □

A.3 Proof of Lemma 3

Lemma 3 *If $Nb-a$ is not a multiple of b, then $|z_i| < |z_1|$ for $i = 2,3,\ldots,Nb-a$. If $Nb-a$ is a multiple of b, there are $b-1$ roots, $z_2, z_3, \ldots, z_b$, say, with $|z_i| = |z_1|$, and $|z_i| < |z_1|$ for $i = b+1, b+2, \ldots, Nb-a$.*

Proof. First, we consider the case that $Nb-a$ is not a multiple of b or that $b = 1$. From $f(0) = 0 < \beta_N = g(0)$ and Lemma 2, we have

$$\begin{aligned} f(z) &< g(z) \quad \text{for } 0 \le z < z_1, \\ f(z) &> g(z) \quad \text{for } z_1 < z < 1. \end{aligned} \tag{13}$$

For z_i, $i = 2,3,\ldots,Nb-a$, we have

$$f(|z_i|) = |f(z_i)| = |g(z_i)| = |\sum_{j=0}^{N} \beta_j z_i^{(N-j)b}| \le \sum_{j=0}^{N} \beta_j |z_i^{(N-j)b}| = g(|z_i|), \tag{14}$$

and thus by (13)

$$|z_i| \le |z_1| \quad \text{for } i = 2,3,\ldots,Nb-a.$$

If equality holds in (14), then from (13) it follows that $|z_i| = z_1$. Since $\beta_N > 0$ and $\beta_{N-1} > 0$, equality in (14) holds if and only if z_i^b is real and $z_i^b > 0$. Suppose that $z_i^b > 0$ and so $|z_i| = z_1$. Then, we have that $z_i = \varphi z_1$ for some φ with $\varphi^b = 1$. However, since $Nb-a$ is not a multiple of b, φz_1 is not a root of the equation $f(z) = g(z)$, and so the equality of (14) does not hold, and hence we have proved that

$$|z_i| < z_1 \quad \text{for } i = 2,3,\ldots,Nb-a.$$

In case that $Nb-a$ is a multiple of b ($Nb-a = rb$, say, for some positive integer r) but b is not equal to 1, we can rewrite equation (5) by substituting $y = z^b$ as

$$y^r = \beta_0 y^N + \beta_1 y^{N-1} + \cdots + \beta_N. \tag{15}$$

From Lemma 1 and Lemma 2, and the first part of Lemma 3,

$$|y_i| < y_1 \quad \text{for } i = 2,3,\ldots,r.$$

The proof is completed by noting that (since $z = \sqrt[b]{y}$) with each root y of equation (15), b roots z of equation (5) correspond, and these b roots have the same absolute value. □

Appendix B

In this appendix, we derive the sojourn-time distribution of an arbitrary arriving customer at one of the queues, queue 1, say. We condition the sojourn time of this customer on the state upon arrival (we know the probability of finding a certain state upon arrival by using the Bernoulli-arrivals-see-time-average (cf. Halfin [7])). If the server is at or switching to queue 1 upon arrival of the customer, then the sojourn time of this customer is completely known. If, however, the server is at or switching to queue 2, then the sojourn time of this customer consists of a fixed and a random part. The fixed part is the sum of the switch-over time (including a possible residual switch-over time to queue 2) and the time for serving customers in queue 1. The random part is the number of slots to customers of queue 2. Clearly, this latter part is the length of a busy period starting with a certain initial number of slots work. If the server was already at queue 2 upon arrival, this initial number of slots work is equal to the number of slots work in queue 2 upon arrival. Otherwise, the initial number of slots work is equal to the number of slots work in queue 2 upon arrival, plus the number of slots work arriving at queue 2 during the residual switch-over to queue 2.

Let V denote this initial number of slots work. We now determine the conditional probability $\Pr\{B = s|V = v\}$ that the busy period, which starts at a slot boundary with v slots of work, ends after s slots. In the sequel, we call the first slot the server is serving queue 2 after the arrival of our customer slot 1. The consecutive slots are numbered consequently $2, 3, \ldots$.

Since the service times of customers of queue 2 are equal to b_2, we have $\Pr\{B = s|V = v\} = 0$ if $s - v \neq mb_2$ for any $m = 0, 1, 2, \ldots$. Suppose $s - v = mb_2$ for some nonnegative integer m, i.e., m customers arrive at queue 2 (since slot 1) before the ending of the busy period. We first consider the case that $v > 0$ and then the case that $v = 0$.

Since m customers arrive at queue 2 before the end of the busy period, the busy period is ended at the end of slot $v + mb_2$. Suppose, for the sake of the argument, that these m customers arrive in the first m consecutive slots $1, 2, \ldots, m$. Then, no customer will arrive at queue 2 in slots $m + 1, m + 2, \ldots, v + mb_2$. Furthermore, no customer may arrive in slot $v + mb_2 + 1$, since the busy period would otherwise not be completed. Hence, the probability that m customers arrive at queue 2 in the first m slots, and that the busy period ends at the end of slots $v + mb_2$ is equal to

$$p_2^m(1 - p_2)^{v+mb_2-m+1}. \tag{16}$$

There are, however, also other ways in which these m customers may arrive such that the busy period lasts $v + mb_2$ slots. The probability of each of these feasible possibilities is given by (16). So, it remains to determine the number of feasible possibilities.

Let $\mathrm{comb}(v, m)$ denote the number of possibilities that m customers arrive at queue 2 such that the busy period, which starts with v slots of work, ends at the end of slot $v + mb_2$. This number can be determined recursively by conditioning on the number of the slot (using the aforementioned numbering of slots) in which the first of these m customers arrive.

1. For all v, we obviously have $\mathrm{comb}(v, 0) := 1$, and set $k := 1$.

2. Compute for all v

   $$\mathrm{comb}(v, k) = \sum_{i=1}^{v+1} \mathrm{comb}(v + b_2 - i, k - 1),$$

   stop if $k = m$, and otherwise set $k := k + 1$ and repeat this step.

If, on the other hand, $v = 0$, then there are two possibilities. First, there is no arrival at queue 2 in slot 1, so that the server switches immediately to queue 1. So, the conditional busy period is equal to zero with probability $1 - p_2$. Second, there is an arrival at queue 2 in slot 1, which occurs with probability p_2. Then, we consider the busy period which starts at the beginning of slot 2 with $b_2 - 1$ number of slots work. Applying the same arguments as for the case $v > 0$, we obtain the conditional probability distribution of the busy period which starts in slot 2 with $b_2 - 1$ number of slots work.

References

[1] Blanc, J.P.C., Mei, R.D. van (1993). The power-series algorithm applied to polling systems with a dormant server, Report 9346, CentER for Economic Research, Tilburg University, Tilburg, The Netherlands.

[2] Borst, S.C. (1993). A polling system with a dormant server, Report BS-R9313, Center for Mathematics and Computer Science, Amsterdam, The Netherlands.

[3] Darroch, J.N. (1964). On the traffic-light queue, *Annals of Mathematical Statistics* **35**, 380-388.

[4] Eenige, M.J.A. van, Resing, J.A.C., Wal, J. van der (1993). A matrix-geometric analysis of queueing systems with periodic service interruptions, Memorandum COSOR 9332, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, The Netherlands.

[5] Eisenberg, M. (1971). Two queues with changeover times, *Operations Research* **19**, 386-401.

[6] Eisenberg, M. (1994). The polling system with a stopping server, *Queueing Systems* **18**, 387-431.

[7] Halfin, S. (1983). Batch delays versus customer delays, *The Bell Systems Technical Journal* **62**, 2011-2015.

[8] Hofri, M., Ross, K.W. (1987). On the optimal control of two queues with server setup times and its analysis, *SIAM Journal on Computing* **16**, 399-420.

[9] Liu, Z., Nain, P., Towsley, D. (1992). On optimal polling policies, *Queueing Systems* **11**, 59-83.

[10] Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models (An Algorithmic Approach)*, The Johns Hopkins University Press, Baltimore.

[11] Newell, G.F. (1960). Queues for a fixed-cycle traffic light, *Annals of Mathematical Statistics* **31**, 589-597.

[12] Srinivasan, M.M., Gupta, D. (1994). When should a roving server be patient, *to appear.*

[13] Takagi, H. (1990). Queueing analysis of polling models: an update, in: *Stochastic Analysis of Computer and Communications Systems*, ed. H. Takagi, Elsevier Science Publishers, North-Holland, Amsterdam.

[14] Tijms, H.C. (1986). *Stochastic Modelling and Analysis: a Computational Approach*, Wiley, Chester.

[15] Tijms, H.C., Coevering, M.C.T. van de (1991). A simple numerical approach for infinite-state Markov chains, *Probability in the Engineering and Informational Sciences* **5**, 285-295.

[16] Wal, J. van der, Schweitzer, P.J. (1987). Iterative bounds on the equilibrium distribution of a finite Markov chain, *Probability in the Engineering and Informational Sciences* **1**, 117-131.

List of COSOR-memoranda - 1995

Number	Month	Author	Title
95-01	January	M.J.A. van Eenige I.J.B.F. Adan J.A.C. Resing J. van der Wal	Periodic versus exhaustive service in a multi-product production center