Optimization of polling systems with Bernoulli schedules

Blanc, J.P.C.; van der Mei, R.D.

Publication date:
1992

TILBURG UNIVERSITY
KATHOLIEKE UNIVERSITEIT BRABANT
POSTBOX 90153
5000 LE TILBURG
THE NETHERLANDS

[redacted]

DEPARTMENT OF ECONOMICS

RESEARCH MEMORANDUM

OPTIMIZATION OF POLLING SYSTEMS WITH BERNOULLI SCHEDULES

R35

J.P.C. Blanc
R.D. van der Mei

Queuing Theory
Polling Systems
Optimization

FEW 563

Communicated by Prof.dr.ir. O.J. Boxma

OPTIMIZATION OF POLLING SYSTEMS WITH BERNOULLI SCHEDULES

J.P.C. Blanc
R.D. van der Mei
Tilburg University, Faculty of Economics
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

In this paper optimization of cyclic polling systems with Bernoulli schedules is considered. An arbitrary weighted sum of the steady-state mean waiting times at the queues is to be minimized with respect to the Bernoulli parameters. Light- and heavy-traffic asymptotes of the optimal schedule are presented. A partial solution of the optimization problem is obtained by means of monotonicity of the mean waiting times as function of the Bernoulli parameters. A numerical approach to compute the optimal schedule, based on the use of the power-series algorithm, is discussed. The influence of system parameters on the optimal Bernoulli schedule is examined. Finally, a fast approach to approximate the optimal schedule is presented and tested.

Keywords: polling systems, optimization, Bernoulli schedules, power-series algorithm

1. Introduction

A polling system is a multi-queue model, attended to by a single server. Polling models arise naturally in the modeling of many computer-communication and production networks, where several users compete for access to a common resource, e.g., a central computer or a transmission channel. The reader is referred to [20] and [28] for overviews of the large variety of applications.

In this paper optimization of continuous-time cyclic polling systems with a so-called Bernoulli service strategy at all queues will be considered. A Bernoulli service strategy, with parameter q_i $(0 \leq q_i \leq 1)$ for queue i, works as follows: if there are one or more customers present at queue i upon a service completion epoch at queue i, with probability q_i another customer at queue i is served; otherwise the server proceeds to the next queue, and so on. The optimization problem considered in this paper is to find a combination of the Bernoulli parameters which minimizes an arbitrary weighted sum of the steady-state mean waiting times at the various queues. In general, the optimization problem is not analytically solvable. Therefore, we shall discuss some properties of the optimal Bernoulli schedule and moreover, propose a numerical approach to compute the optimal schedule accurately, based on the use of the power-series algorithm (p.s.a.), cf. [2, 3, 4, 5, 6]. As this approach may be rather time and memory consuming, we finally present an approximation method for finding the optimal Bernoulli schedule, for which the time and memory requirements are negligible and which yields fairly accurate results over a wide range of admissible parameter values.

Motivation

The use of Bernoulli service strategies is motivated by situations in which the server in a polling system is to give higher priority to some queues; in that case, the Bernoulli parameters may serve as control parameters. The Bernoulli service discipline stochastically limits the number of customers served during one visit of the server to a particular queue through a series of random decisions. The optimization problem comes down to the assignment of priorities to each of the queues.

The Markovian character of the Bernoulli strategies simplifies the analysis of the model. Getting an insight into such a system may serve as a set-up for the analysis of models with more complex (non-Markovian) service strategies, such as the limited service strategies; under a limited service strategy, the number of customers served during one visit of the server to a particular queue has a fixed upper bound.

Classification

The order in which the server visits the queues is based on some routing mechanism; such a mechanism may be static or dynamic. In static routing mechanisms the server may visit the queues according to some probabilistic routing mechanism (probabilistic polling) or in a fixed order according to a so-called polling table (periodic polling). In dynamic routing the order in which the server visits the queues is changing dynamically, i.e., the routing depends on the actual state of the system. In the present model the server moves along the queues in a cyclic order; this is a special case of periodic polling. The proposed optimization method is readily modified to other routing mechanisms.

The number of customers served during a visit of the server to a particular queue is determined by the service strategy at that queue, also referred to as the service discipline. The large variety of service disciplines can be classified into the exhaustive-type and the gated-type service disciplines, cf. [21]. If a queue has an exhaustive-type service discipline, all customers present upon the arrival of the server at the queue *plus* all those arriving during the same visit period of the server are *candidates* for service; under a gated-type service strategy, only those customers present upon the arrival of the server at that queue are candidates for service. The Bernoulli service discipline is of the exhaustive-type. Fuhrmann [14] provides another classification of the service disciplines. He classifies the service strategies into two classes, depending on whether or not the service strategies satisfy the following property, which will be referred to as the randomization property: all customers present at a queue upon the server arrival epoch to that queue can be effectively replaced by a random population, the sizes of which are independent, identically distributed (i.i.d.) random variables. Moreover, Fuhrmann shows that models in which all service strategies satisfy the randomization property are relatively easy to analyze. In particular, for such models Resing [26] gives exact expressions for the generating function of the joint queue length at polling instants, using the observation that the joint queue length process in those models constitutes a multi-type branching process with immigration. In general, the Bernoulli service discipline does *not* satisfy the randomization property. As a consequence, there is a need for numerical techniques to analyze queueing systems with Bernoulli schedules.

Related literature

The variety in routing mechanisms and service strategies opens possibilities for efficient operation and optimization. Nevertheless, optimization of polling systems has received little attention in the literature.

For models with zero switch-over times in which there is full information about the queue lengths and one is allowed to choose the next queue to be visited after each service completion, the rule which minimizes an arbitrary weighted sum of the mean waiting times is the well-known μc-rule, cf., e.g., [23].

For models with non-negligible switch-over times, Hofri and Ross [16] suggest that in a system with full information about the queue lengths the optimal switch-over rule is of threshold-type, i.e., there exist thresholds that determine when the server should switch from one queue to the other. Boxma et al. [10] propose, for mixtures of gated, 1-limited and exhaustive service strategies, simple square-root rules for the 'optimal' relative visit frequencies in a polling table so as to minimize an arbitrary weighted sum of the mean waiting times, cf. [8] for an overview of their results. Browne and Yechiali [11] consider optimization for a model with so-called semi-dynamic server routing, in which the server, at the beginning of a cycle, chooses a visiting order of the queues for this cycle that minimizes the mean duration of the cycle. The reader is referred to [31] for an overview of the results on semi-dynamic server routing. Levy et al. [21] prove that, for each given visit order, the service strategy which minimizes the amount of unfinished work in the system, is to serve during each visit as many customers as possible at each queue. Liu et al. [22] discuss the problem of finding polling strategies that stochastically minimize the amount of unfinished work and the number of customers in the system at all time. Their results imply that the strategy that routes the server to the longest queue is optimal when all queue lengths are known, and that the cyclic server routing is optimal in case the only available information is the set of previous decisions.

Bernoulli strategies were introduced by Keilson and Servi [18]. For the G/G/1-Bernoulli vacation model they show that the steady-state waiting time is the convolution of two components, being: (i) the waiting time in the corresponding G/G/1 model *without* vacation times, using so-called modified service times, i.e., service times which are modified by the convolution of a service time *plus* a vacation time with probability q, where q is the Bernoulli parameter in the model, and (ii) the forward recurrence time of the length of the vacation period. For systems with Poisson arrivals, a closed-form expression for the Laplace-Stieltjes Transform (L.S.T.) of the waiting time distribution in an M/G/1-Bernoulli vacation model is given in [24].
The most general exact results obtained for polling systems are the formulations of so-called pseudo-conservation laws; a pseudo-conservation law is a closed form expression for a *specific* weighted sum of the mean waiting times. A general framework to derive pseudo-conservation laws is presented in [9].
However, more detailed results for polling systems with service disciplines which do not satisfy the randomization property, are restricted to two-queue models. In these models, exact detailed analysis is only possible in a few special cases. In particular, Weststrate [30] gives a closed-form expression for the L.S.T. of the waiting time distribution in a two-queue model in which one queue is served according to a Bernoulli service strategy and the other is served exhaustively. Moreover, he presents an iterative algorithm to compute the mean

waiting times at both queues.

As a consequence, there is a need for numerical algorithms or approximation methods to optimize polling systems with Bernoulli service strategies. For these models, the Markovian character of Bernoulli strategies admits the use of the power-series algorithm, introduced in [17], and developed in [2, 3, 4, 5, 6]. The p.s.a. can be used for the numerical analysis of polling models in which the joint queue length process has the structure of a multi-dimensional quasi birth-death process. The p.s.a. is based on power-series expansions of the state-probabilities and the moments of the joint queue length distribution as function of the load of the system in light-traffic. Leung [19] proposes another numerical technique based on discrete Fourier Transforms to solve the waiting time distributions in polling models with a so-called probabilistically-limited service strategy. The main disadvantage of both algorithms is that the time and memory requirements increase exponentially with the number of queues, so that their use is restricted to fairly small systems.

Overview of the paper

The rest of the paper is organized as follows. In Section 2 a detailed description of the model will be given and the optimization problem will be formulated. In Section 3 the p.s.a. will be used to derive algebraïc expressions for the light-traffic asymptotes of the individual mean waiting times, which yield light-traffic asymptotes of the optimal Bernoulli schedule. For systems with non-negligible switch-over times, the stability condition puts a lower bound on the range of possible values of the components of the optimal Bernoulli parameter. Because for increasing load of the system this lower bound increases to one for each queue, it follows that the optimal service strategy tends to the exhaustive strategy for all queues when the load of the system tends to the boundary of the stability region. Subsequently, we shall discuss a monotonicity property for mean waiting times, which states that increasing the Bernoulli parameter for a particular queue implies a decrease of the mean waiting time at that queue, and an increase of the mean waiting times at all other queues. Using this property, one can find a partial solution for the optimization problem. That is, the queues for which the relative cost for the waiting time, divided by the relative offered load, is maximal over all queues, are to be served exhaustively; for systems with negligible switch-over times, the queues for which this ratio is minimal over all queues are to be served 1-limited. In Section 4, a numerical approach to compute the optimal Bernoulli schedule accurately will be discussed; the approach combines the use of the p.s.a. with some optimization procedure. Subsequently, in Section 5 the influence of system parameters on the optimal Bernoulli schedule will be discussed. As the numerical approach discussed in Section 4 may be time and memory consuming, in Section 6 we shall propose and test an approximation method for finding the optimal Bernoulli schedule, for which the time and memory requirements are negligible.

2. Model description and problem formulation

Consider a queueing system consisting of s queues, attended to by a single server. Customers arrive at queue i according to a Poisson process with rate λ_i , $i=1,...,s$. Each queue may contain an unbounded number of customers. The queues are visited in a cyclic order. The number of customers which is served during a visit to a certain queue depends on the service discipline at that queue. As service disciplines Bernoulli schedules will be taken. A Bernoulli schedule is a vector of s probabilities $\mathbf{q}=(q_1,...,q_s)$, which is used as follows. When the server arrives at a queue, at least one customer is served, provided the queue is non-empty. If the queue is empty at the server arrival epoch, the server directly proceeds to the next queue. After completion of a service at queue i, the server starts to serve another customer at queue i with probability q_i , if queue i has not yet been emptied; otherwise, the server proceeds to the next queue. In the case $q_i=0$, the service discipline is referred to as 1-limited; if $q_i=1$, queue i is said to be served exhaustively. At each queue the customers are served on a first-in-first-out basis.

The service times at queue i are assumed to be i.i.d. random variables with finite first and second moments, denoted by $\beta_i^{(1)}$ and $\beta_i^{(2)}$ respectively, $i=1,...,s$. The switch-over times from queue i-1 to queue i are assumed to be i.i.d. random variables with finite first two moments $\sigma_i^{(1)}$ and $\sigma_i^{(2)}$, $i=2,...,s$; the first two moments of the switch-over times from queue s to queue 1 are denoted by $\sigma_1^{(1)}$ and $\sigma_1^{(2)}$.

All service times, switch-over times and interarrival times are assumed to be mutually independent and independent of the state of the system.

The sum of the arrival processes at the various queues is a Poisson process with rate

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

The first two moments β_1 and β_2 of the service time distribution of an arbitrary customer are given by

$$\beta_1:=\frac{1}{\Lambda}\sum_{i=1}^{s}\lambda_i\beta_i^{(1)},\qquad \beta_2:=\frac{1}{\Lambda}\sum_{i=1}^{s}\lambda_i\beta_i^{(2)}. \tag{2}$$

The *offered load* to queue i, ρ_i , and the *total offered load* to the system, ρ, are defined by

$$\rho_i := \lambda_i \beta_i^{(1)},\ i=1,\dots,s, \quad \rho := \sum_{i=1}^{s} \rho_i. \tag{3}$$

Because ρ will be used as a variable in the p.s.a., define the following quantities

$$\eta_i := \rho_i/\rho, \quad a_i := \lambda_i/\rho,\ i=1,\dots,s, \tag{4}$$

referred to as the *relative load* and the *relative arrival rate* of queue i, i = 1,...,s.

The first two moments of the total switch-over time during one cycle of the server along the queues are given by

$$\sigma_1 := \sum_{i=1}^{s} \sigma_i^{(1)}, \quad \sigma_2 := \sum_{i=1}^{s} \sigma_i^{(2)} + 2\sum_{i=1}^{s}\sum_{j=1}^{i-1} \sigma_i^{(1)} \sigma_j^{(1)}. \tag{5}$$

Necessary and sufficient conditions for the stability of polling systems have been derived in [13]. For the present model with Bernoulli schedules these conditions read:

$$\chi := \rho + \sigma_1 \max_{i=1,\dots,s} \{\lambda_i(1-q_i)\} < 1; \tag{6}$$

χ is called the *occupancy* of the system. In the sequel it is assumed that the stability condition (6) is satisfied and that the system is in steady-state.

The reader is referred to [3, 4, 5, 6] for a description of the p.s.a. for this model.

Optimization problem

The optimization problem is to minimize the cost function, defined by

$$C(\mathbf{q}) := \sum_{i=1}^{s} c_i EW_i, \tag{7}$$

over all $\mathbf{q}$ for which (6) holds; here, c_i, i = 1,...,s, are arbitrary strictly positive cost coefficients and EW_i, i = 1,...,s, are the steady-state mean waiting times at the various queues, which depend on $\mathbf{q}$. The optimal Bernoulli schedule will be denoted by $\mathbf{q}^*$. Without loss of generality, it is assumed that the cost coefficients are normalized such that

$$\sum_{i=1}^{s} c_i = 1. \tag{8}$$

In general, the optimization problem is not analytically solvable. Therefore, in Section 3 we shall discuss some properties of the optimal Bernoulli schedule. In particular, light- and heavy-traffic asymptotes of $\mathbf{q}^*$ will be discussed. Moreover, a partial solution to the optimization problem will be given; that is, explicit values of q_i^* for some particular queues i will be given.

3. Light-traffic and heavy-traffic properties; partial solution

The most general exact results obtained for polling systems are the formulations of the so-called *pseudo-conservation laws*, cf. [9]; a pseudo-conservation law is a closed-form expression for a *specific* weighted sum of the mean waiting times at the various queues. For a cyclic polling system with Bernoulli schedules at all queues the pseudo-conservation law reads (cf. [7]):

$$\sum_{i=1}^{s} \rho_i \left[1 - a_i(1-q_i)\frac{\sigma_1 \rho}{1-\rho}\right] EW_i =$$
$$\frac{\rho^2}{1-\rho}\frac{\beta_2}{2\beta_1} + \rho\frac{\sigma_2}{2\sigma_1} + \frac{\sigma_1}{1-\rho}\sum_{i=1}^{s}\rho_i^2(1-q_i) + \frac{\sigma_1}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{s}\rho_i^2\right]. \tag{9}$$

Light-traffic properties

An elegant tool to get an insight into the light-traffic behaviour of the system is the use of the p.s.a. The algorithm relies on the property that the mean waiting time at an arbitrary queue i, $i = 1,\ldots,s$, can be expressed as power-series in ρ, say

$$EW_i = \sum_{n=0}^{\infty} \gamma_n^{(i)} \rho^n, \tag{10}$$

for all values of ρ for which the stability condition (6) is satisfied; the coefficients $\{\gamma_n^{(i)}\}$, $i = 1,\ldots,s$, $n = 0,1,\ldots$, can be computed by means of an iteration scheme for the coefficients of the power-series expansions of the state-probabilities, cf. [3, 6]. Algebraïcally calculating the first few terms of the power-series expansions (10) gives an insight into the light-traffic behaviour of the system and, moreover, provides useful information about the optimization of Bernoulli schedules in light-traffic systems. In order to derive light-traffic asymptotes, we first determine the coefficients of the power-series expansions of the state-probabilities of the queue lengths according to the recursion relations of the p.s.a. Then, the coefficients of the power-series expansions of the mean waiting times follow with the aid of Little's formula, yielding the following results.

If $\sigma_1 = 0$, *then for* $i = 1,...,s$,

$$EW_i = \rho \frac{\beta_2}{2\beta_1} + \rho^2[\eta_i \frac{\beta_2}{2\beta_1} + \Xi(i) + \Psi(i,q)] + O(\rho^3), \quad (\rho \downarrow 0), \tag{11}$$

with

$$\begin{aligned} \Xi(i) &:= \sum_{j=1}^{s-1} \eta_{i+j} \sum_{k=0}^{j-1} \eta_{i+k} \frac{\beta_{i+k}^{(2)}}{\beta_{i+k}^{(1)}}, \\ \Psi(i,q) &:= \sum_{j=1}^{s} q_j \eta_j^2 \frac{\beta_j^{(2)}}{\beta_j^{(1)}} - q_i \eta_i \frac{\beta_i^{(2)}}{\beta_i^{(1)}}; \end{aligned} \tag{12}$$

if $\sigma_1 > 0$, *then for* $i = 1,...,s$,

$$EW_i = \frac{\sigma_2}{2\sigma_1} + \rho[\frac{\beta_2}{2\beta_1} + \Phi(i) + \Omega(i,q)] + O(\rho^2), \quad (\rho \downarrow 0), \tag{13}$$

with

$$\begin{aligned} \Phi(i) &:= (\sigma_1 - \frac{\sigma_2}{2\sigma_1})(1 - \eta_i) + \sum_{\substack{j=1 \\ j \neq i}}^{s} \sum_{k=1}^{j} \eta_j \frac{\sigma_k^{(2)} - \sigma_k^{(1)2}}{\sigma_1}, \\ \Omega(i,q) &:= (1 - q_i)(\eta_i \sigma_1 + \frac{1}{2} a_i \sigma_2) . \end{aligned} \tag{14}$$

Here, the indices, say i, should be replaced by i *mod* s if i>s.
The coefficient of ρ^2 in (11) consists of three terms. The first term is independent of the order in which the queues are polled and of the service discipline. The term $\Xi(.)$ reflects the influence of the order in which the server visits the queues and the term $\Psi(. , .)$ depends on the service discipline at the queues. The coefficient of ρ in (13) also consists of three terms. The first term only depends on the service time distribution and is independent of the switch-over times and of the service discipline. The term $\Phi(.)$ depends on the switch-over time distributions and the third term $\Omega(. , .)$ reflects the influence of the service disciplines at the queues.

The computation of (11) and (13) according to the recursive schemes is tedious, but straightforward. We used an algebraïc formula manipulation computer program. It should be noted that the p.s.a. is only applicable to quasi birth-death processes. General probability distributions for the service times and the switch-over times are approximated by Coxian distributions. However, the memory requirements have restricted the computation to systems with 3 queues, using 2-phase Coxian distributions, for the case of non-zero switch-over times. The so-obtained coefficients in (11) and (13) have been checked numerically for numerous examples and were found to be valid in all considered cases.

Using (11) and (13), one is now able to derive light-traffic asymptotes of $\mathbf{q}^*$ both for zero and non-zero switch-over times.
For $\sigma_1 = 0$, omitting the terms which do not depend on $\mathbf{q}$ in (11), it follows that the coefficient of ρ^2 of the cost function (7) is

$$\sum_{i=1}^{s} q_i \eta_i (\eta_i - c_i) \frac{\beta_i^{(2)}}{\beta_i^{(1)}} . \tag{15}$$

Consequently, using (15) we have

if $\sigma_1 = 0$, *then for* $i = 1,\ldots,s$,
(i) if $c_i > \eta_i$, *then*

$$\lim_{\rho \downarrow 0} q_i^* = 1, \tag{16}$$

(ii) if $c_i < \eta_i$, *then*

$$\lim_{\rho \downarrow 0} q_i^* = 0. \tag{17}$$

The case $c_i = \eta_i$ is not covered by (16) and (17); apparently, the light-traffic limit of q_i^* is determined by higher order terms in that case. Note that if $c_i = \eta_i$ for all i, $i = 1,\ldots,s$, (9) implies that $C(\mathbf{q})$ is independent of $\mathbf{q}$.
For non-zero switch-over times, (7), (13) and (14) yield the following light-traffic limit of $\mathbf{q}^*$:

if $\sigma_1 > 0$, *then for* $i = 1,\ldots,s$,

$$\lim_{\rho \downarrow 0} q_i^* = 1. \tag{18}$$

Remarks:

The limits are defined such that the total arrival rate tends to zero, while the ratios between the individual arrival rates remain fixed.

Numerical experience suggests that the individual mean waiting times are smooth functions of $\mathbf{q}$. Therefore, one would expect that (16), (17) and (18) not only hold in the limiting case $\rho\downarrow 0$, but remain valid when $0 \le \rho < \varepsilon$, for some ε small enough. In fact, cases were found with $\sigma_1 > 0$ small and $c_i > \eta_i$, where there exist positive numbers $\rho^{(0)}$, $\rho^{(1)}$ and $\rho^{(2)}$, satisfying $0 < \rho^{(1)} < \rho^{(2)} < \rho^{(0)}$, such that $q_i^* = 1$ for $0 \le \rho \le \rho^{(1)}$, q_i^* decreases from one to zero as ρ increases from $\rho^{(1)}$ to $\rho^{(2)}$ and $q_i^* = 0$ for $\rho^{(2)} \le \rho \le \rho^{(0)}$, cf. e.g. Figure 5.1 in Section 5.

The fact that $\mathbf{q}$ appears in the ρ-term in (13) and does *not* appear in the ρ-term in (11) can be explained as follows. As described in [4], the ρ^k-term in (10) corresponds to states of the system in which at most k customers are present in the system, and using PASTA, to situations in which an arriving customer finds at most k customers present in the system upon arrival, $k=0,1,\ldots$. Now, it is easy to verify that for zero switch-over times the waiting time of an arriving customer, who finds at most one customer present in the system, while there are no arrivals during the arriving customer's waiting time (higher order effect), does not depend on $\mathbf{q}$, whereas in a system with non-zero switch-over times the waiting time of such a customer *does* depend on $\mathbf{q}$.

Heavy-traffic properties

Denote the *heavy-traffic residue* of the mean waiting time at queue i by

$$\omega_i := \lim_{\chi\uparrow 1} (1-\chi) EW_i, \quad i=1,\ldots,s. \tag{19}$$

These limits are defined in such a way that the total arrival rate to the system increases to a value at which one or more queues become instable, while the proportions between the arrival rates remain fixed.

When the number of queues is not too large and the parameters of a system are not too asymmetrical, it is possible to obtain accurate data for performance measures even for high occupancy of the system, i.e., χ close to one, by means of the p.s.a. From those results the heavy-traffic residues in (19) can be estimated. An important general property, which is also supported by results in [13], is the following.

For $i=1,\ldots,s$, it holds that the residue ω_i is positive if

$$a_i(1-q_i) = \max_{j=1,\ldots,s} \{a_j(1-q_j)\}; \tag{20}$$

otherwise, $\omega_i=0$ and EW_i possesses a finite limit as $\chi\uparrow 1$.

Note that (20) implies that $C(\mathbf{q})$ tends to infinity as $\chi\uparrow 1$.
In general, there is no closed-form expression for ω_i , except for a few exceptional cases. As an example of such a special case, consider a system with exhaustive service at all queues, i.e., $q_i=1$, $i=1,\ldots,s$. For the system one can obtain explicit expressions for heavy-traffic residues of the mean waiting times from the set of equations in [27] (Chapter 4):

If $q_i=1$, $i=1,\ldots,s$, then the heavy-traffic residue of the mean waiting time at queue i is given by

$$\omega_i = \frac{1-\eta_i}{\sum_{j=1}^{s} \eta_j(1-\eta_j)} \frac{\beta_2}{2\beta_1} + \tfrac{1}{2}\sigma_1(1-\eta_i), \quad i=1,\ldots,s. \tag{21}$$

Moreover, if $\omega_i=0$, the finite limit for the mean waiting time at queue i does not admit a closed-form expression, except for a few special cases. For instance, suppose (20) is satisfied for all queues, except for queue i. Then it follows that the queues $j=1,\ldots,s$, $j\neq i$, saturate as $\chi\uparrow 1$ (note that $q_i<1$ under the assumption). Hence, the number of customers served during a visit of the server to queue j, $j\neq i$, is geometrically distributed, *independent* of the states of the other queues. Therefore, the behaviour of queue i can be interpreted as an M/G/1-Bernoulli vacation model with parameter q_i ; the vacation time distribution is the *convolution* of the switchover-time distributions plus the visit times at the other queues in a cycle. A closed-form expression for the mean waiting time in an M/G/1-Bernoulli vacation model is derived in [24].

If $\sigma_1>0$, the stability condition (6) can be reformulated as

$$q_i > 1 - \frac{1-\rho}{\sigma_1 a_i \rho}, \; i=1,\ldots,s. \tag{22}$$

Hence, (22) puts a lower bound on the set of possible values of q_i^* , $i=1,\ldots,s$. Moreover, as the right hand side of (22) tends to one as $\rho\uparrow 1$, we have the following heavy-traffic asymptote for q_i^*:

if $\sigma_1 > 0$, then

$$\lim_{\rho \uparrow 1} q_i^* = 1. \tag{23}$$

If $\sigma_1 = 0$, the stability condition reads $\rho < 1$; hence, $\mathbf{q}$ has no influence on the stability of the system. Thus, unlike in the case $\sigma_1 > 0$, the ergodicity condition (6) does *not* imply that q_i^* tends to 1, for $i = 1,\ldots,s$, as $\rho \uparrow 1$.

Partial solution

In order to get an insight into the behaviour of $\mathbf{q}$ for varying system parameters, knowing some properties of the cost function, $C(\mathbf{q})$, such as monotonicity, continuïty, differentiability, etc., may be helpful.

Monotonicity properties in polling systems with more than one queue are scarcely discussed in the literature. Levy et al. [21] studied monotonicity properties of the mean amount of unfinished work in the system. They prove, under rather weak assumptions, that the total amount of unfinished work in the system is stochastically minimized when all queues are served exhaustively. Altman et al. [22] show that the station times, the intervisit times and cycle times are stochastically increasing in arrival rates, service times and switch-over times.

However, little is known about monotonicity properties of the individual mean waiting times with respect to the system parameters. Nevertheless, numerical experience suggests that the following *monotonicity* property with respect to $\mathbf{q}$ is valid.

For increasing q_i,

(i) EW_i is strictly decreasing, (24)

(ii) EW_j is strictly increasing for all $j \neq i$. (25)

Remarks:

Property (24) is confirmed by Tedijanto [29] (Chapter 4), who proves the stochastic monotonicity of the waiting times with respect to the Bernoulli parameter in an M/G/1-Bernoulli vacation model.

One would conjecture that the monotonicity property is not restricted to the case of Bernoulli service disciplines at all queues; in fact, we believe that (24) and (25) remain valid for increasing 'exhaustiveness' of the service discipline at queue i, within a broader class of service disciplines.

The monotonicity properties (24) and (25) imply the following two statements, which partly solve the optimization problem.

(i) *if* $c_i/\eta_i = \max_{j=1,\ldots,s} \{c_j/\eta_j\}$, *then* $q_i^* = 1$, (26)

(ii) *if* $\sigma_1 = 0$, *then*

if $c_i/\eta_i = \min_{j=1,\ldots,s} \{c_j/\eta_j\}$, *then* $q_i^* = 0$. (27)

Properties (26) and (27) are based on a conjecture in [21], which says that, under rather weak assumptions, the total amount of unfinished work stochastically decreases with increasing 'exhaustiveness' of the service discipline or, equivalently,

$$\frac{\partial}{\partial q_i}\sum_{j=1}^{s} \eta_j EW_j \leq 0, \quad i=1,\ldots,s. \tag{28}$$

Consequently,

$$\begin{aligned}\frac{\partial}{\partial q_i}\sum_{j=1}^{s} c_j EW_j &= \frac{\partial}{\partial q_i}\sum_{\substack{j=1\\ j\neq i}}^{s} \frac{c_j}{\eta_j}\eta_j EW_j + \frac{\partial}{\partial q_i}\frac{c_i}{\eta_i}\eta_i EW_i \\ &\leq \sum_{\substack{j=1\\ j\neq i}}^{s} \eta_j\left[\frac{c_j}{\eta_j} - \frac{c_i}{\eta_i}\right]\frac{\partial}{\partial q_i}EW_j .\end{aligned} \tag{29}$$

Hence, using the monotonicity property, (29) immediately implies (26). Moreover, if $\sigma_1 = 0$, the principle of work conservation implies equality in (28) and (29); (27) follows then immediately.

Remarks:

The validity of (26) and (27) is supported by the μc-rule for systems with zero switch-over times, in which there is full information about the queue lengths and in which the server is allowed to choose the next queue after each service completion, cf. [23]. In order to minimize

$$\sum_{j=1}^{s} \tilde{c}_j \lambda_j EW_j , \tag{30}$$

the μc-rule gives priorities to the queues in increasing order of the values of $\tilde{c}_j/\beta_j^{(1)} = [c_j/\eta_j]/\rho$. If queue i, for which $c_i/\eta_i = \max_{j=1,\ldots,s} \{c_j/\eta_j\}$, is non-empty upon a service completion epoch at queue i, then according to the μc-rule, there is no reason for the server to proceed to

another queue. Moreover, if $\sigma_1 = 0$, using the μc-rule, there is no reason for the server to serve more than one customer at that queue k, for which $c_k/\eta_k = \min_{j=1,\ldots,s} \{c_j/\eta_j\}$.
In the case $s=2$ and $\sigma_1 > 0$, (26) implies that the iterative algorithm in [30] for a two-queue model in which one queue is served exhaustively and the other is served according to a Bernoulli service strategy, can be used to solve the optimization problem.
As noted before, exact expressions for ω_i have not been found, except for a few exceptional cases. As a consequence, if $\sigma_1 = 0$, the heavy-traffic behaviour of $\mathbf{q}^*$ for values of i, which are not covered by (26) or (27), is still an open problem. In fact, we have found cases in which the asymptote of q_i^* as $\rho \uparrow 1$ lies in the interior of the interval [0,1], cf. e.g. Figure 5.2 in Section 5.

4. Numerical optimization

In the previous section properties of optimal Bernoulli schedules were presented. In general, however, the optimization problem is not analytically solvable. Therefore, in this section a numerical approach to compute the optimal Bernoulli schedule accurately will be discussed. The approach is based on the use of the p.s.a., which, in principle, can compute $C(\mathbf{q})$ within any level of accuracy for any feasible value of $\mathbf{q}$. Combining the p.s.a. with some (local) optimization procedure may lead to an accurate determination of $\mathbf{q}^*$.
In optimization theory numerous algorithms to find (local) optima are available, using information about the characteristics of the function to be optimized. The optimization techniques can be classified into two classes, depending on whether or not they use derivatives: the direct search methods, which do not use derivatives, and the gradient methods, which are generally more efficient. The reader is referred to [25] (Chapter 6) for an overview of the various optimization techniques.
As the time requirements of the use of the p.s.a. for evaluation purposes may be considerable, cf. [6] for a discussion of the complexity of the p.s.a., efficiency of the optimization procedure is of main importance. The memory requirements restrict the use of the p.s.a. to queueing systems with a fairly small number of queues and moreover, limit the number of terms in the power-series expansions that can be computed. Hence, restrictions on the available computation time and memory space may affect the accuracy of the mean waiting time approximations which, in turn, may affect the accuracy of the p.s.a.-based optimal Bernoulli schedule. Because little is known about the cost function $C(\mathbf{q})$ as a function of $\mathbf{q}$, one is forced to rely on numerical experience, which *suggests* that $C(\mathbf{q})$ is (i) continuous and (ii) differentiable, for all $\mathbf{q}$; this justifies the use of gradient methods. Because of efficiency and memory considerations, the authors propose to use the so-called conjugate gradient

method, cf. [25] (Section 6.3), with the obvious modification that the Bernoulli parameters are restricted to the interval [0,1]. Here, the partial derivatives of the cost function (7) with respect to the Bernoulli parameters are estimated on the basis of finite differences. These finite differences and the step sizes in the line searches are adapted to the accuracy of the computed cost. The latter can only be estimated on the basis of differences in the last few computed terms of the power-series, cf. [5, 6] for a discussion of the accuracy of the p.s.a.

In general, $C(\mathbf{q})$ is not a convex function of $\mathbf{q}$, so that local optimality does not guarantee global optimality of an optimum found by means of a local optimization procedure. Nevertheless, it has been supported by numerous numerical experiments with different initial schedules that $C(\mathbf{q})$ has no alternative local minima.

5. Influence of system parameters on the optimal schedule

In this section, the influence of system parameters on the optimal Bernoulli schedule is discussed. The optimal schedules have been obtained by means of the numerical optimization technique discussed in the previous section.
For notational convenience, define $\boldsymbol{\beta}^{(k)}=(\beta_1^{(k)},\dots,\beta_s^{(k)})$, $k=1,2$, $\sigma^{(k)}=(\sigma_1^{(k)},\dots,\sigma_s^{(k)})$, $k=1,2$, $c=(c_1,\dots,c_s)$ and $a=(a_1,\dots,a_s)$.

Offered load

Optimal Bernoulli schedules have been computed for various values of the offered load; here, the load is varied in such a way that the ratio's between the arrival rates remain fixed. Consider the model with the following set of parameters: $s=3$; $\boldsymbol{\beta}^{(1)}=(0.5,1.0,1.5)$; $\sigma^{(1)}=(\alpha,\alpha,\alpha)$; all service times and switch-over times are exponentially distributed; $a=(1/3,1/3,1/3)$; $c=(10/55,15/55,30/55)$, so that $q_1^*=q_3^*=1$, cf. (26). Figure 5.1 shows the values of q_2^* as function of ρ for varying values of α.

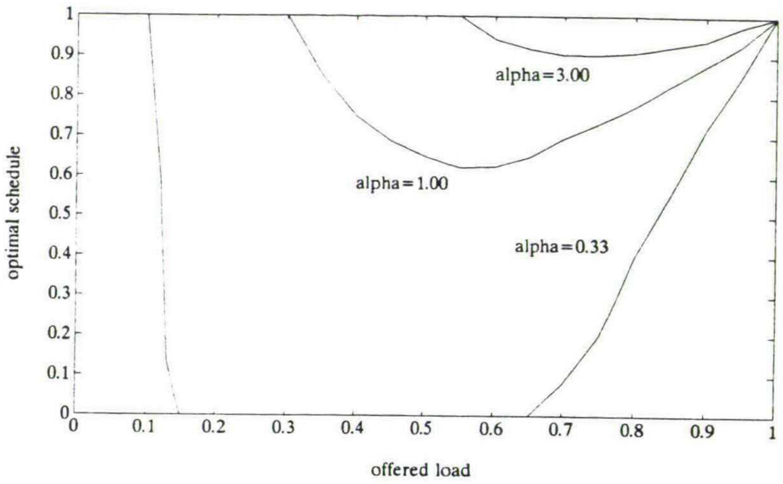

Figure 5.1 The optimal schedule as function of the offered load; $\sigma_i > 0$.

As discussed in Section 3, the characteristics of the optimal schedule as function of ρ in the case $\sigma_1 = 0$ may differ from those in the case $\sigma_1 > 0$. To illustrate this, optimal schedules for varying values of the offered load have been computed for the model with the following set of parameters: $s = 3$; $\boldsymbol{\beta}^{(1)} = (0.5, 1.0, 1.5)$; all service times at the queues 1 and 3 are exponentially distributed and the service times at queue 2 are 2-phase Coxian with squared coefficient of variation 4; $\sigma_1 = 0$; $a = (1/3, 1/3, 1/3)$. Two different cost functions are considered: $c_1 = (0.6, 0.3, 0.1)$ and $c_2 = (0.50, 0.35, 0.15)$, so that in both cases $q_1^* = 1$ and $q_3^* = 0$, cf. (26) and (27). Figure 5.2 shows q_2^* as function of the offered load for both cost functions.

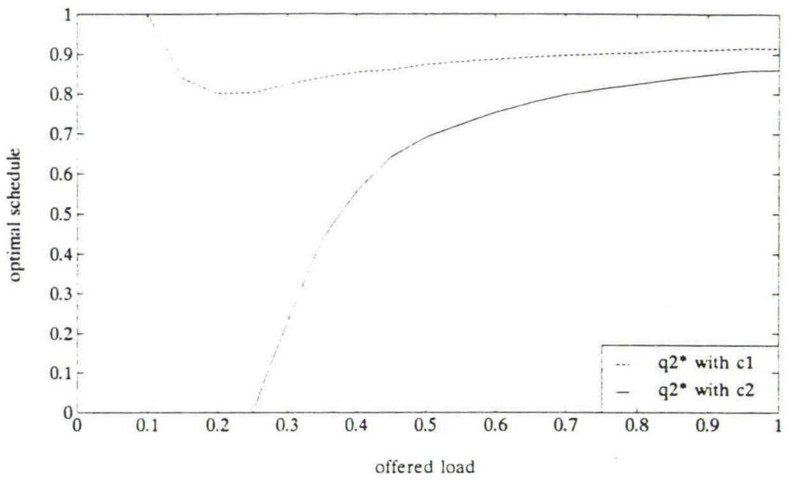

Figure 5.2 The optimal schedule as function of the offered load; $\sigma_l = 0$.

Remarks:

Figures 5.1 and 5.2 confirm the validity of the light- and heavy-traffic asymptotes, cf. (16), (17), (18) and (23).

Figure 5.2 illustrates the fact that in the case $\sigma_1 = 0$ some of the components of the heavy-traffic asymptote of $\mathbf{q}^*$ may have values in the interior of the interval [0,1], as remarked at the end of Section 3.

As illustrated in Figure 5.1, some of the components of the optimal schedule may be equal to zero for sufficiently small values of σ_1 ($\sigma_1 > 0$). For those cases, numerical experiments have confirmed that the total cost can be decreased by modifying the Bernoulli service discipline as follows: when the server arrives at a non-empty queue, a coin is flipped so as to decide whether or not the first customer is served at all. If no customer is served, the server proceeds to the next queue; otherwise, the strategy proceeds according to the 'usual' Bernoulli strategy. Note that this class of modified Bernoulli strategies does not contain the 1-limited service discipline.

We shall now discuss the influence of system parameters which are independent of the offered load: the service time distributions (higher moments) and the switch-over time distributions.

Service time distributions

Numerical experience has indicated that the optimal schedules and in particular, the optimal cost, seem to depend primarily on the second moments of the service time distributions through β_2 , rather than on the second moments of the individual service time distributions, $\beta^{(2)}$. To illustrate this, the value of $\mathbf{q}^*$ has been computed for combinations of the individual second moments of the service time distributions, $\beta^{(2)}$, where the second moment of the service time distribution of an arbitrary customer, β_2 , is kept fixed. Table 5.1 shows the value of $\mathbf{q}^*$ for different combinations $\beta^{(2)}$, which are constructed in such a way that $\beta_2 = 10$ in all considered cases. The other parameters are taken to be: $s=3$; $\beta^{(1)} = (1.0,2.0,3.0)$; all service times are 2-phase Coxian distributed; $\sigma^{(1)} = (0.05,0.05,0.05)$; all switch-over times are exponentially distributed; $a=(1/3,1/3,1/3)$; $\rho=0.8$; $c=(0.40,0.25,0.35)$, so that $q_1^*=1$, cf. (26).

$\beta^{(2)}$	$\mathbf{q}^*$	$C(\mathbf{q}^*)$
(2.0, 8.0,20.0)	(1.00,0.59,0.22)	9.06
(2.0,10.0,18.0)	(1.00,0.57,0.21)	9.08
(4.0, 8.0,18.0)	(1.00,0.60,0.22)	9.06
(1.5, 6.0,22.5)	(1.00,0.62,0.23)	9.05
(1.5,15.0,13.5)	(1.00,0.54,0.21)	9.09

Table 5.1 The optimal schedule for different values of $\beta^{(2)}$ for β_2 fixed.

The influence of the second moment of the service time distribution of an arbitrary customer, β_2 , seems to be rather unpredictable; in fact, in some cases components of the optimal schedule decrease for increasing values of β_2 , whereas in other cases components increase for increasing β_2 . To illustrate this, the optimal schedules have been computed for different values of β_2 for two different models. The first model is determined by the following set of parameters: $s=3$; $\beta^{(1)} = (1.0,2.0,3.0)$; all service times at queues 1 and 2 are exponentially distributed, and the service times at queue 3 are 2-phase Coxian distributed with squared coefficient of variation α; $\sigma^{(1)} = (0.05,0.05,0.05)$; all switch-over times are exponentially distributed; $a=(1/3,1/3,1/3)$; $\rho=0.8$; $c=(0.40,0.25,0.35)$, so that $q_1^*=1$, cf. (26). Table 5.2a shows the optimal schedules for $\alpha=0.25$, 0.50, 1.00, 2.00 and 4.00.

β_2	$\mathbf{q}^*$	$C(\mathbf{q}^*)$
7.08	(1.00,0.52,0.15)	6.62
7.83	(1.00,0.54,0.17)	7.25
9.33	(1.00,0.58,0.20)	8.51
12.33	(1.00,0.64,0.27)	11.01
18.33	(1.00,0.74,0.31)	15.95

Table 5.2a Increasing effect of β_2 on the optimal schedule.

For the second model the parameters are: $s=3$; $\boldsymbol{\beta}^{(1)}=(0.5,1.0,1.5)$; all service times at queues 1 and 3 are exponentially distributed, and the service times at queue 2 are 2-phase Coxian distributed with squared coefficient of variation α; $\sigma^{(1)}=(0.01,0.01,0.01)$; all switch-over times are exponentially distributed; $a=(1/3,1/3,1/3)$; $\rho=0.8$; $c=(0.60,0.30,0.10)$, so that $q_1^*=1$, cf. (26). Table 5.2b shows the optimal schedules for $\alpha=1$, 2, 3 and 4.

β_2	$\mathbf{q}^*$	$C(\mathbf{q}^*)$
2.33	(1.00,0.99,0.00)	2.45
2.67	(1.00,0.91,0.00)	2.80
3.00	(1.00,0.86,0.00)	3.13
3.33	(1.00,0.83,0.00)	3.45

Table 5.2b Decreasing effect of β_2 on the optimal schedule.

Remark:
As illustrated in Tables 5.2a and 5.2b the optimal cost increases when $\boldsymbol{\beta}_2$ increases.

Switch-over time distributions
As illustrated in Figure 5.1, the influence of the value of σ_1 on the optimal schedule may be considerable and, in general, seems to have an increasing effect on all components of $\mathbf{q}^*$. However, as far as the individual mean switch-over times are concerned, numerical experience has taught us that the optimal schedule depends on the individual mean switch-over times $\sigma^{(1)}$ mainly through the mean total switch-over time, σ_1 . To illustrate this, Table 5.3 shows the optimal Bernoulli schedules for varying combinations $\sigma^{(1)}$ of the individual mean switch-over times; the total switch-over times consist of three i.i.d. exponential phases, each with mean 0.05 and hence, are Erlangian-3 distributed with mean $\sigma_1=0.15$ in all considered cases.

The other model parameters are: $s=3$; $\beta^{(1)}=(1.0,2.0,3.0)$; all service times are exponentially distributed; $a=(1/3,1/3,1/3)$; $\rho=0.8$; $c=(0.40,0.25,0.35)$, so that $q_1^*=1$, cf. (26).

$\sigma^{(1)}$	q^*	$C(q^*)$
(0.05,0.05,0.05)	(1.00,0.58,0.21)	8.507
(0.15,0.00,0.00)	(1.00,0.58,0.20)	8.506
(0.00,0.15,0.00)	(1.00,0.58,0.20)	8.508
(0.00,0.00,0.15)	(1.00,0.58,0.20)	8.507

Table 5.3 The optimal schedule for different values of $\sigma^{(1)}$, with σ_1 (and σ_2) fixed.

As far as the second moments of the switch-over time distributions are concerned, we found out that their effect on the optimal schedule is generally negligible. To illustrate this, consider the same model as in Table 5.3 with $\sigma^{(1)}=(0.15,0.00,0.00)$. The optimal schedules have been computed for vaying values of α, the squared coefficients of variation of the switch-over times. Table 5.4 shows the results for $\alpha=0.25$, 0.50, 1.00, 2.00 and 4.00.

σ_2/σ_1^2	q^*	$C(q^*)$
1.25	(1.00,0.58,0.20)	8.500
1.50	(1.00,0.58,0.20)	8.516
2.00	(1.00,0.58,0.21)	8.548
3.00	(1.00,0.58,0.21)	8.612
5.00	(1.00,0.58,0.21)	8.740

Table 5.4 The optimal schedule for different values of σ_2 with σ_1 fixed.

Remarks:

Table 5.4 illustrates the fact that the optimal schedule is fairly insensitive to the second moment of the total switch-over time, wheareas the optimal cost increases for increasing values of σ_2.

In general, the optimal schedule depends on higher moments of the service time and switch-over time distributions. However, as illustrated in [6] (Table 1), these dependencies are fairly weak.

6. Approximation

In Section 4 we discussed a numerical approach to achieve an accurate approximation for the optimal Bernoulli schedule, based on the use of the p.s.a. The main disadvantage of this approach is the fact that the time and memory requirements increase exponentially with increasing number of queues and hence, its use is restricted to rather small systems. For this reason, in this section we will propose a simple and fast approach to approximate the optimal Bernoulli schedule, which uses negligible computation time and memory space and is therefore applicable to fairly large systems.

The approach is based on a simple mean waiting time approximation (instead of on the use of the p.s.a.); combined with the optimization procedure discussed in Section 4 it yields an approximation for the optimal Bernoulli schedule.
Consider the following mean waiting time approximation, cf. [29] (Section 6.7):

$$EW_i \approx \frac{(1-\rho+\rho_i) - q_i\rho_i(2-\rho)}{1-\rho-\lambda_i(1-q_i)\sigma_1}\, x; \tag{31}$$

the quantity x is determined by substituting (31) into the pseudo-conservation law (9). The approximation is a trivial extension of the pseudo-conservation law-based mean waiting time approximation proposed in [15] for mixtures of 1-limited and exhaustive (and gated) service disciplines. The latter approach relies on the observation of Everitt [12], who states that the cycle time of queue i, i.e., the length of the time interval between two successive arrivals of the server to queue i, has approximately the same second moment for all $i=1,\dots,s$.
Although the approximation is generally not very accurate for evaluation purposes, cf. numerical results in [29] (Section 6.7), combined with the optimization procedure as discussed in Section 4 it turns out to give satisfying results for optimization purposes.

Remarks:
The mean waiting time approximation (31) depends on the individual service time distributions only through β_1 and β_2, and similarly, depends on the switch-over time distributions only through σ_1 and σ_2. As discussed in Section 5, the optimal schedules and, in particular, the cost of the optimal schedule, are fairly insensitive to the individual service time and switch-over time distributions.
One may verify that the mean waiting time approximation (31) satisfies properties (24), (25) and (28). Consequently, as (31) also satisfies the pseudo-conservation law (9), properties (26) and (27) are satisfied.

As noted in Section 3, the heavy-traffic residue of the mean waiting times is unknown, except for a few exceptional cases. As a consequence, in the case $\sigma_1=0$, which is not covered by (23), the heavy-traffic asymptote of $\mathbf{q}^*$ is generally unknown. Therefore, the heavy-traffic asymptote of approximated optimum based on (31) may differ from the exact heavy-traffic asymptote. Hence, the approximated optimum and in particular, the cost belonging to the approximated optimum may be inaccurate. As a consequence, the use of the approximation method is restricted to systems in which the switch-over times are not too small.

Figure 6.1 illustrates the behaviour of the approximated optimum, $\mathbf{q}^*$(app), as function of the offered load ρ, compared with the optimum $\mathbf{q}^*$, obtained by means of the numerical optimization technique discussed in Section 4. The parameters are: $s=3$; $\boldsymbol{\beta}^{(1)}=(0.5,1.0,1.5)$; $\sigma^{(1)}=(0.05,0.05,0.05)$; all service times and switch-over times are exponentially distributed; $a=(1/3,1/3,1/3)$; $c=(0.40,0.25,0.35)$, so that $q_1^*=1$, cf. (26).

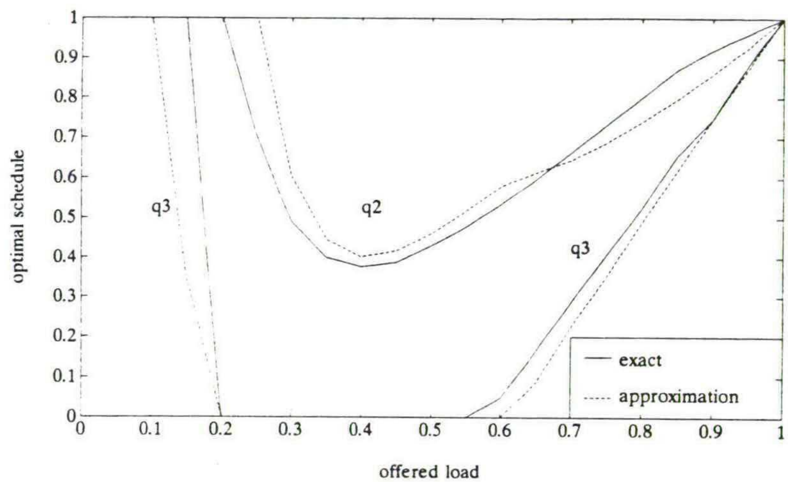

Figure 6.1 Behaviour of $\mathbf{q}^*$(app) and $\mathbf{q}^*$ as function of ρ.

The quality of the approximation can best be measured by the relative difference between the exact *cost* belonging to $\mathbf{q}^*$(app), C($\mathbf{q}^*$(app)), and the *cost* of the optimum, C($\mathbf{q}^*$), rather than by the relative accuracy of $\mathbf{q}^*$(app), compared with $\mathbf{q}^*$. Table 6.1 gives for the same set of parameters the relative error in the cost function, defined by

$$\frac{C(q^*(app)) - C(q^*)}{C(q^*)} \times 100\,\% . \tag{32}$$

ρ	C(**q***(app))	C(**q***)	err%
0.50	1.272	1.272	0.0
0.60	1.836	1.835	0.0
0.70	2.772	2.771	0.0
0.80	4.638	4.635	0.1
0.90	10.231	10.205	0.3
0.95	21.468	21.259	1.0

Table 6.1 The accuracy of the cost belonging to the approximated optimum.

In order to check the quality of the approximation for larger systems, consider a 6-queue model with the following set of parameters: $s=6$; $\beta^{(1)}=(0.60,0.80,1.00,1.20,1.40,1.60)$; $\sigma^{(1)}=(0.05,0.05,0.05,0.05,0.05,0.05)$; all service times and switch-over time are exponentially ditsributed; $a_i=1/(6\times1.1)$ for $i=1,...,6$; $c=(0.375,0.125,0.125,0.125,0.125,0.125)$, so that $q_1^*=1$, cf. (26). Table 6.2 shows the approximated optimum $\mathbf{q}^*$(app) and the exact optimum $\mathbf{q}^*$ for varying values of the offered load ρ; the relative error in the cost function (err%) is computed according to (32).

ρ	**q***(app)	C(**q***(app))	**q***	C(**q***)	err%
0.50	(1.00,1.00,1.00, 0.55,0.00,0.00)	1.45	(1.00,1.00,1.00, 0.35,0.00,0.00)	1.45	0.0
0.60	(1.00,1.00,1.00, 0.58,0.00,0.00)	2.04	(1.00,1.00,1.00, 0.50,0.00,0.00)	2.04	0.0
0.70	(1.00,1.00,1.00, 0.69,0.00,0.00)	2.99	(1.00,1.00,1.00, 0.66,0.04,0.00)	2.99	0.0
0.80	(1.00,1.00,1.00, 0.79,0.34,0.00)	4.87	(1.00,1.00,1.00, 0.91,0.34,0.00)	4.86	0.1

Table 6.2 The approximated optima compared with exact optima for a 6-queue model.

In general, the optimal schedule depends on the order in which the queues are visited, whereas the approximated optimum based on (31) does *not* depend on the order in which the queues are placed. However, the differences in the optimal schedules and in particular, in the cost belonging to the optimal schedules, have turned out to be fairly small. To illustrate this, consider the model with the following set of parameters: $s=6$; $\boldsymbol{\beta}^{(1)}=(0.5,1.5,1.5,1.5,1.5,1.5)$; $\sigma^{(1)}=(0.1,0.1,0.1,0.1,0.1,0.1)$; all service times and switch-over times are exponentially distributed; $a_i=1/(6\times 0.75)$, $i=1,\ldots,6$; $c=(0.375,0.125,0.125,0.125,0.125,0.125)$, so that $q_1^*=1$, cf. (26). Table 6.3 shows the approximated optimum $\mathbf{q}^*(\text{app})$ and the exact optimum $\mathbf{q}^*$ for varying values of the offered load ρ; the relative error in the cost function (err%) is computed according to (32). Note that $q_2^*(\text{app})=q_3^*(\text{app})=q_4^*(\text{app})=q_5^*(\text{app})=q_6^*(\text{app})$ for all values of ρ; hence, the approximated optimum $\mathbf{q}^*(\text{app})$ is denoted by $(q_1^*(\text{app}),q_{2\text{-}6}^*(\text{app}))$.

ρ	$(q_1^*(\text{app}),q_{2\text{-}6}^*(\text{app}))$	$C(\mathbf{q}^*(\text{app}))$	$\mathbf{q}^*$	$C(\mathbf{q}^*)$	err%
0.30	(1.00,1.00)	1.09	(1.00,1.00,1.00, 1.00,1.00,1.00)	1.09	0.0
0.40	(1.00,1.00)	1.50	(1.00,0.79,0.77, 0.76,0.76,0.76)	1.50	0.0
0.50	(1.00,0.95)	2.08	(1.00,0.66,0.62, 0.59,0.57,0.56)	2.06	0.1
0.60	(1.00,0.82)	2.92	(1.00,0.66,0.60, 0.57,0.55,0.53)	2.90	0.7
0.70	(1.00,0.80)	4.30	(1.00,0.72,0.67, 0.64,0.62,0.60)	4.28	0.5
0.75	(1.00,0.81)	5.40	(1.00,0.75,0.71, 0.69,0.67,0.66)	5.37	0.6
0.80	(1.00,0.84)	7.04	(1.00,0.86,0.78, 0.73,0.72,0.70)	7.01	0.4

Table 6.3 The approximated optima compared with exact optima for a 6-queue model.

An alternative approach which is more generally applicable, also for polling models for which no such approximation as (31) is available, is to use the p.s.a. with a small number of terms to find the neighbourhood of the optimal schedule with reduced computational effort, and then proceed with the p.s.a. with more terms to locally improve the optimal schedule. Further, the partial solution (26), (27), may be used to decrease the dimension of the optimization problem.

Acknowledgements: The authors wish to thank Onno Boxma and Jan Weststrate for interesting discussions and useful comments.

References

[1] E. Altman, P. Konstantopoulos, and Z. Liu, Some qualitative properties in polling systems, *Report, INRIA Centre Sophia Antipolis, France (1991)*. To appear in *Queueing Systems* **11** (1992).

[2] J.P.C. Blanc, On a numerical method for calculating state probabilities for queueing systems with more than one waiting line, *Journal of Computational and Applied Mathematics* **20** (1987) 119-125.

[3] J.P.C. Blanc, A numerical approach to cyclic-service queueing models, *Queueing Systems* **6** (1990) 173-188.

[4] J.P.C. Blanc, Cyclic polling systems: limited versus Bernoulli schedules, *Report FEW 422, Department of Economics, Tilburg University, Tilburg* (1990).

[5] J.P.C. Blanc, The power-series algorithm applied to cyclic polling systems, *Stochastic Models* **7** (1991) 527-545.

[6] J.P.C. Blanc, Performance evaluation of polling systems by means of the power-series algorithm, *Annals of Operations Research* **35** (1992) 155-186.

[7] O.J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, *Queueing Systems* **5** (1989) 185-214.

[8] O.J. Boxma, Analysis and optimization of polling systems. In: *Proceedings 13th International Teletraffic Congress, Workshop: Queueing, Performance and Control in ATM, eds. J.W.Cohen and C.D. Pack* (North-Holland, Amsterdam) (1991) 173-183.

[9] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, *Journal of Applied Probability* **24** (1987) 949-964.

[10] O.J. Boxma, H. Levy and J.A. Weststrate, Efficient visit frequencies for polling tables: minimization of waiting cost, *Queueing Systems* **9** (1991) 133-162.

[11] S. Browne and U. Yechiali, Dynamic priority rules for cyclic-type queues, *Advances in Applied Probability* **21** (1989) 432-450.

[12] D. Everitt, Simple approximations for token rings, *IEEE Transactions on Communications* **34** (1986) 719-721.

[13] C. Fricker and M.R. Jaïbi, Monotonicity and stability of periodic polling models, *Report FEW 559, Department of Economics, Tilburg University, Tilburg* (1992).

[14] S.W. Fuhrmann, A decomposition result for a class of polling models. *Report IBM*

Research Division, Zurich Research Laboratory, Switzerland (1991).

[15] W.P. Groenendijk, Waiting-time approximations for cyclic-service systems with mixed service strategies. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC 12, ed. M.Bonatti* (North-Holland, Amsterdam) (1989) 1434-1441.

[16] M. Hofri and K.W. Ross, On the optimal control of two queues with server set-up times and its analysis, *SIAM Journal on Computing* **16** (1987) 399-419.

[17] G. Hooghiemstra, S. Keane and S. van der Ree, Power series for stationary distributions of coupled processor models, *SIAM Journal of Applied Mathematics* **48** (1988) 1159-1166.

[18] J. Keilson and L.D. Servi, Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules, *Journal of Applied Probability* **23** (1986) 790-802.

[19] K.K. Leung, Cyclic-service systems with probabilistically-limited service, *IEEE Selected Areas in Communications* **9** (1991) 185-193.

[20] H. Levy and M. Sidi, Polling systems: applications, modeling and optimization, *IEEE Transactions on Communications* **38** (1990) 1750-1760.

[21] H. Levy, M. Sidi and O.J. Boxma, Dominance relations in polling systems, *Queueing Systems* **6** (1990) 155-171.

[22] Z. Liu, P. Nain and D. Towsley, On optimal polling policies, *Report, INRIA Centre Sophia Antipolis, France* (1991). To appear in *Queueing Systems* **11** (1992).

[23] I. Meilijson and U. Yechaili, On optimal right-of-way policies at a single-server station when insertion of idle times is permitted, *Stochastic Processes and their Applications* **6** (1977) 23-52.

[24] R. Ramaswamy and L. Servi, The busy period of the M/G/1 vacation model with Bernoulli schedules, *Stochastic Models* **4** (1988) 507-521.

[25] S.S. Rao, Optimization theory and applications (Wiley Eastern Limited) (1978).

[26] J.A.C. Resing, Polling systems and multi-type branching processes, *Report BS-R9128, C.W.I., Amsterdam* (1991).

[27] H. Takagi, Analysis of polling systems (The MIT Press, Cambridge, Massachusetts) (1986).

[28] H. Takagi, Queueing analysis of polling models. In: *Stochastic Analysis of Computer Communication Systems, ed. H.Takagi* (North-Holland, Amsterdam) (1990) 267-318.

[29] T.E. Tedijanto, Nonexhaustive policies in polling systems and vacation models: qualitative and approximate approach, *Ph.D. thesis, University of Maryland* (1990).

[30] J.A. Weststrate, Waiting times in a two-queue model with exhaustive and Bernoulli service, *Report FEW 437, Department of Economics, Tilburg University, Tilburg* (1990). To appear in: *ZOR Methods and Models of Operations Research.*

[31] U. Yechiali, Optimal dynamic control of polling systems. In: *Proceedings 13th International Teletraffic Congress, Workshop: Queueing, Performance and Control in ATM, eds. J.W.*

Cohen and C.D. Pack (North-Holland, Amsterdam) (1991) 195-208.

IN 1991 REEDS VERSCHENEN

466 Prof.Dr. Th.C.M.J. van de Klundert - Prof.Dr. A.B.T.M. van Schaik
Economische groei in Nederland in een internationaal perspectief

467 Dr. Sylvester C.W. Eijffinger
The convergence of monetary policy - Germany and France as an example

468 E. Nijssen
Strategisch gedrag, planning en prestatie. Een inductieve studie binnen de computerbranche

469 Anne van den Nouweland, Peter Borm, Guillermo Owen and Stef Tijs
Cost allocation and communication

470 Drs. J. Grazell en Drs. C.H. Veld
Motieven voor de uitgifte van converteerbare obligatieleningen en warrant-obligatieleningen: een agency-theoretische benadering

471 P.C. van Batenburg, J. Kriens, W.M. Lammerts van Bueren and R.H. Veenstra
Audit Assurance Model and Bayesian Discovery Sampling

472 Marcel Kerkhofs
Identification and Estimation of Household Production Models

473 Robert P. Gilles, Guillermo Owen, René van den Brink
Games with Permission Structures: The Conjunctive Approach

474 Jack P.C. Kleijnen
Sensitivity Analysis of Simulation Experiments: Tutorial on Regression Analysis and Statistical Design

475 C.P.M. van Hoesel
An $O(nlogn)$ algorithm for the two-machine flow shop problem with controllable machine speeds

476 Stephan G. Vanneste
A Markov Model for Opportunity Maintenance

477 F.A. van der Duyn Schouten, M.J.G. van Eijs, R.M.J. Heuts
Coordinated replenishment systems with discount opportunities

478 A. van den Nouweland, J. Potters, S. Tijs and J. Zarzuelo
Cores and related solution concepts for multi-choice games

479 Drs. C.H. Veld
Warrant pricing: a review of theoretical and empirical research

480 E. Nijssen
De Miles and Snow-typologie: Een explorative studie in de meubelbranche

481 Harry G. Barkema
Are managers indeed motivated by their bonuses?

513 Drs. J. Dagevos, Drs. L. Oerlemans, Dr. F. Boekema
Regional economic policy, economic technological innovation and networks

514 Erwin van der Krabben
Het functioneren van stedelijke onroerend-goed-markten in Nederland - een theoretisch kader

515 Drs. E. Schaling
European central bank independence and inflation persistence

516 Peter M. Kort
Optimal abatement policies within a stochastic dynamic model of the firm

517 Pim Adang
Expenditure versus consumption in the multi-good life cycle consumption model

518 Pim Adang
Large, infrequent consumption in the multi-good life cycle consumption model

519 Raymond Gradus, Sjak Smulders
Pollution and Endogenous Growth

520 Raymond Gradus en Hugo Keuzenkamp
Arbeidsongeschiktheid, subjectief ziektegevoel en collectief belang

521 A.G. de Kok
Basics of inventory management: Part 2
The (R,S)-model

522 A.G. de Kok
Basics of inventory management: Part 3
The (b,Q)-model

523 A.G. de Kok
Basics of inventory management: Part 4
The (s,S)-model

524 A.G. de Kok
Basics of inventory management: Part 5
The (R,b,Q)-model

525 A.G. de Kok
Basics of inventory management: Part 6
The (R,s,S)-model

526 Rob de Groof and Martin van Tuijl
Financial integration and fiscal policy in interdependent two-sector economies with real and nominal wage rigidity

527 A.G.M. van Eijs, M.J.G. van Eijs, R.M.J. Heuts
Gecoördineerde bestelsystemen
een management-georiënteerde benadering

528 M.J.G. van Eijs
Multi-item inventory systems with joint ordering and transportation decisions

529 Stephan G. Vanneste
Maintenance optimization of a production system with buffercapacity

530 Michel R.R. van Bremen, Jeroen C.G. Zijlstra
Het stochastische variantie optiewaarderingsmodel

531 Willy Spanjers
Arbitrage and Walrasian Equilibrium in Economies with Limited Information

IN 1992 REEDS VERSCHENEN

532 F.G. van den Heuvel en M.R.M. Turlings
Privatisering van arbeidsongeschiktheidsregelingen
Refereed by Prof.Dr. H. Verbon

533 J.C. Engwerda, L.G. van Willigenburg
LQ-control of sampled continuous-time systems
Refereed by Prof.dr. J.M. Schumacher

534 J.C. Engwerda, A.C.M. Ran & A.L. Rijkeboer
Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X + A^{*}X^{-1}A = Q$.
Refereed by Prof.dr. J.M. Schumacher

535 Jacob C. Engwerda
The indefinite LQ-problem: the finite planning horizon case
Refereed by Prof.dr. J.M. Schumacher

536 Gert-Jan Otten, Peter Borm, Ton Storcken, Stef Tijs
Effectivity functions and associated claim game correspondences
Refereed by Prof.dr. P.H.M. Ruys

537 Jack P.C. Kleijnen, Gustav A. Alink
Validation of simulation models: mine-hunting case-study
Refereed by Prof.dr.ir. C.A.T. Takkenberg

538 V. Feltkamp and A. van den Nouweland
Controlled Communication Networks
Refereed by Prof.dr. S.H. Tijs

539 A. van Schaik
Productivity, Labour Force Participation and the Solow Growth Model
Refereed by Prof.dr. Th.C.M.J. van de Klundert

540 J.J.G. Lemmen and S.C.W. Eijffinger
The Degree of Financial Integration in the European Community
Refereed by Prof.dr. A.B.T.M. van Schaik

541 J. Bell, P.K. Jagersma
Internationale Joint Ventures
Refereed by Prof.dr. H.G. Barkema

542 Jack P.C. Kleijnen
Verification and validation of simulation models
Refereed by Prof.dr.ir. C.A.T. Takkenberg

543 Gert Nieuwenhuis
Uniform Approximations of the Stationary and Palm Distributions of Marked Point Processes
Refereed by Prof.dr. B.B. van der Genugten

544 R. Heuts, P. Nederstigt, W. Roebroek, W. Selen
Multi-Product Cycling with Packaging in the Process Industry
Refereed by Prof.dr. F.A. van der Duyn Schouten

545 J.C. Engwerda
Calculation of an approximate solution of the infinite time-varying LQ-problem
Refereed by Prof.dr. J.M. Schumacher

546 Raymond H.J.M. Gradus and Peter M. Kort
On time-inconsistency and pollution control: a macroeconomic approach
Refereed by Prof.dr. A.J. de Zeeuw

547 Drs. Dolph Cantrijn en Dr. Rezaul Kabir
De Invloed van de Invoering van Preferente Beschermingsaandelen op Aandelenkoersen van Nederlandse Beursgenoteerde Ondernemingen
Refereed by Prof.dr. P.W. Moerland

548 Sylvester Eijffinger and Eric Schaling
Central bank independence: criteria and indices
Refereed by Prof.dr. J.J. Sijben

549 Drs. A. Schmeits
Geïntegreerde investerings- en financieringsbeslissingen; Implicaties voor Capital Budgeting
Refereed by Prof.dr. P.W. Moerland

550 Peter M. Kort
Standards versus standards: the effects of different pollution restrictions on the firm's dynamic investment policy
Refereed by Prof.dr. F.A. van der Duyn Schouten

551 Niels G. Noorderhaven, Bart Nooteboom and Johannes Berger
Temporal, cognitive and behavioral dimensions of transaction costs; to an understanding of hybrid vertical inter-firm relations
Refereed by Prof.dr. S.W. Douma

552 Ton Storcken and Harrie de Swart
Towards an axiomatization of orderings
Refereed by Prof.dr. P.H.M. Ruys

553 J.H.J. Roemen
The derivation of a long term milk supply model from an optimization model
Refereed by Prof.dr. F.A. van der Duyn Schouten

554 Geert J. Almekinders and Sylvester C.W. Eijffinger
Daily Bundesbank and Federal Reserve Intervention and the Conditional Variance Tale in DM/$-Returns
Refereed by Prof.dr. A.B.T.M. van Schaik

555 Dr. M. Hetebrij, Drs. B.F.L. Jonker, Prof.dr. W.H.J. de Freytas
"Tussen achterstand en voorsprong" de scholings- en personeelsvoorzieningsproblematiek van bedrijven in de procesindustrie
Refereed by Prof.dr. Th.M.M. Verhallen

556 Ton Geerts
Regularity and singularity in linear-quadratic control subject to implicit continuous-time systems
Communicated by Prof.dr. J. Schumacher

557 Ton Geerts
Invariant subspaces and invertibility properties for singular systems: the general case
Communicated by Prof.dr. J. Schumacher

558 Ton Geerts
Solvability conditions, consistency and weak consistency for linear differential-algebraic equations and time-invariant singular systems: the general case
Communicated by Prof.dr. J. Schumacher

559 C. Fricker and M.R. Jaïbi
Monotonicity and stability of periodic polling models
Communicated by Prof.dr.ir. O.J. Boxma

560 Ton Geerts
Free end-point linear-quadratic control subject to implicit continuous-time systems: necessary and sufficient conditions for solvability
Communicated by Prof.dr. J. Schumacher

561 Paul G.H. Mulder and Anton L. Hempenius
Expected Utility of Life Time in the Presence of a Chronic Noncommunicable Disease State
Communicated by Prof.dr. B.B. van der Genugten

562 Jan van der Leeuw
The covariance matrix of ARMA-errors in closed form
Communicated by Dr. H.H. Tigelaar