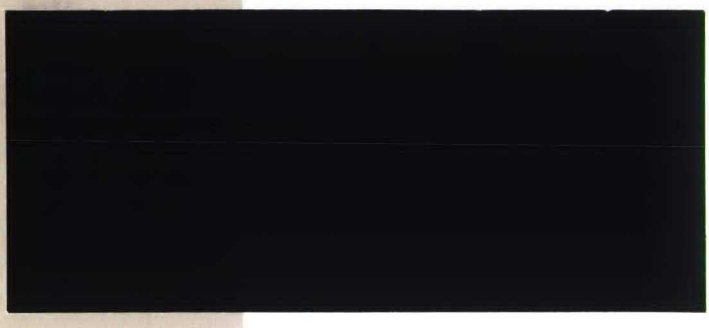
entER
for
omic Research

Discussion paper



TILBURG UNIVERSITY

Center
for
Economic Research

No. 9514

POLLING SYSTEMS WITH MARKOVIAN SERVER ROUTING

By R.D. van der Mei

February 1995

ISSN 0924-7815

Polling systems with Markovian server routing

R.D. van der Mei

Abstract

In this paper we study the performance of polling systems in which the server is routed along the queues according to some probabilistic routing mechanism. It is shown how the performance of the system can be analyzed by means of the so-called power-series algorithm (PSA), a tool for the numerical analysis of a broad class of multiple-queue models. We compare the performance of polling systems with probabilistic server routing with the performance of identical systems in which the server visits the queues in a fixed order. Numerical experiments with the PSA suggest that the mean amount of work in the system is structurally smaller in the case of fixed visit orders. In addition, it is shown that a similar dominance relation is not generally valid for the individual mean waiting times at the queues. Subsequently, we consider the problem of finding optimal combinations of server routing probabilities. We observe a tendency of the optimal probabilistic routing towards deterministic routing. The influence of system parameters on the optimal matrices of routing probabilities is examined. These investigations point out that the optimal routing matrices can be classified into a limited number of types of solutions, each having specific characteristics that can be interpreted rather easily. Finally, we give some guidelines for constructing optimal routing matrices.

1 Introduction

A polling system basically consists of a number of queues attended to by a single server. The server visits the queues in some order to render service to the customers present at the queues, typically incurring a non-negligible switch-over time while moving from one queue to another. Polling models are widely applicable for the modeling of systems in which several types of user compete for access to a common service facility. Applications of polling models can be found in the areas of communication systems, computer networks, maintenance, manufacturing and production environments (cf. Levy and Sidi [19] and Takagi [26] for extensive surveys on the applicability of polling models).

In many cases the server has no global information about the queue lengths. Therefore, in most polling models it is assumed that the server visits the queues in a cyclic order. However, in some cases it is desirable to visit particular queues more frequently than others, e.g. when the queues are not equally loaded. Therefore, a number of generalizations of the cyclic visit order has been considered in the literature. The most common generalization of purely cyclic polling is *periodic* polling, in which the server visits the queues periodically according to a fixed service order table (cf. [17], [2]). In this way, queues can be given higher priority by listing them more often on the (polling) table. Alternatively, the server can be routed along the

queues according to some *probabilistic* routing mechanism. In this paper, we investigate the performance of polling models with so-called Markovian server routing. Under this routing mechanism, with routing probability $p_{i,j}$ the server is routed to queue j after a departure from queue i, independent of the actual state of the system. In this way, the customers at the different queues can be assigned relative priorities by varying specific routing probabilities. In this perspective, Markovian polling can be viewed as the *stochastic counterpart* of periodic polling.

Motivation
This study is motivated by a number of reasons. First, there is a number of specific applications of polling systems with probabilistic server routing. Evidently, for these systems performance analysis and optimization is very useful. Second, only very little is known about the performance of polling systems under probabilistic server routing, whereas polling models with periodic server routing have received much attention in the literature. For this reason, we believe it to be interesting to provide an insight into the performance of polling models with probabilistic server routing, and to investigate how the performance of polling models under periodic and probabilistic server routing is related. Third, we believe it to be interesting to see how the PSA can be applied to determine detailed performance measures of the model, and how the specific structure of the present model can be explored to make the implementation of the PSA more efficient.

Applications
Polling models with probabilistic server routing find a number of specific applications. For instance, they may be used to model distributed systems, such as a shared broadcast channel where from time to time a decision has to be made as to who gets the right for transmission. These decisions are usually based on some probabilistic algorithms, rather than on a fixed order (cf. [15]). Alternatively, polling models with Markovian server routing may also be used to predict the expected delay in an exhaustive slotted ALOHA system. In such a system, a station is granted the exclusive right to transmit during some time period. When a transmitting station no longer reserves the channel, some or all stations start contending to seize the channel. Both the length of the contention period and the next station that will seize the channel are random (cf. [18]). Markovian polling is also useful for the modeling of the so-called Orwell slotted-ring protocol. In this protocol, a number of unit-buffer slots of equal length rotate around a ring, and a packet in a slot filled by a station is addressed to some other station with a certain probability, where it is emptied and passed on empty to the next downstream station (cf. [21], [28]). This is a major difference from other slotted-ring protocols, where a slot can be released only by the station that filled it. As another alternative, polling models with Markovian server routing can be used to model material handling systems such as an Automated Guided Vehicle (AGV) system in which a single vehicle serves a manufacturing cell by moving loads from one machining center to another. When the AGV delivers a load to the center, it inspects the output buffer of that center to determine if there are any loads waiting to be transported. If so, the AGV takes some amount of time to pick up load from this output buffer, and a certain amount of time to transport the load and deliver it at its destination, and the AGV polls the output buffer of the center which receives the load. Otherwise, the AGV switches to poll the next center in some order (cf. [9]).

Literature
In the literature, only a few papers have been devoted to the analysis of polling models with

probabilistic server routing, and detailed results are restricted to special cases. Kleinrock and Levy [15] analyze the behavior of so-called random polling models in which after a departure from an arbitrary queue, the server is routed to queue j with some given probability p_j, irrespective of the queue it has just departed from. It should be noted that random polling occurs as a special case of Markovian polling by taking the routing probabilities $p_{i,j} = p_j$ for all i. For infinite-buffer models in which either all queues are served according to the gated service discipline or in which all queues are served exhaustively, Kleinrock and Levy give the mean waiting times at the queues as the solution of a system of linear equations. For symmetrical models with 1-limited service, they determine a closed-form expression for the mean waiting time. For polling models with Markovian server routing with mixtures of exhaustive, gated and 1-limited service, Boxma and Weststrate [8] derive a pseudo-conservation law (PCL), i.e. an exact expression for a specific weighted sum of the mean waiting times at the queues. For models in which either all queues are served exhaustively or all queues are served according to the gated service discipline, Weststrate [31] derives a set of linear equations to obtain the mean waiting times at the queues. However, the number of linear equations increases cubically in the number of queues, so that this approach is restricted to rather small systems. Srinivasan [24] derives a PCL for polling models with Markovian server routing, in which the routing probabilities may depend on whether customers have been served during the last visit of the server to a queue. Chung et al. [11] analyze Markovian polling models with unit buffers. They derive exact expressions for the generating function of the joint queue length at polling instants, the Laplace-Stieltjes Transforms (LSTs) of the waiting times and the LST of the cycle-time distribution of each queue.
In addition, they derive a set of linear equations to determine the mean waiting times. The number of equations however increases exponentially in the number of queues.
For polling models with probabilistic server routing that are not covered in these references, to the best of the author's knowledge, no alternative algorithms are available to compute performance measures concerning queue-length and waiting-time distributions.

The power-series algorithm
The power-series algorithm (PSA) is a device for the numerical analysis for a broad class of multiple-queue models, requiring a continuous-time Markov chain representation of the process. The basic idea of the PSA is the transformation of the non-recursively (infinite) set of global balance equations into, an principle, recursively solvable set of equations by adding one dimension to the state space. This transformation is realized by expressing the state probabilities as power series in the offered load to the system in light traffic. The basic idea of the PSA stems from Hooghiemstra et al. [13], who applied the PSA to the coupled-processor model. The algorithm has been further developed by Blanc, which has led to more efficient implementations of the algorithm. The PSA has been applied to a number of models, such as the shortest-queue model, a variety of polling models, models with correlated arrivals and Markovian queueing networks (cf. [4] for a survey on the applicability of the PSA). Blanc and Van der Mei [5] have extended the PSA to the computation of derivatives of the performance measures with respect to a broad class of system parameters. This extension is very useful for performing sensitivity analysis and for optimization purposes. Recently, Koole [16] has shown that the PSA is, formally, applicable to general Markov processes.
The PSA can be used to compute numerical values for general performance measures which are functions of the state probabilities. We emphasize that the PSA can not only be applied to determine global performance measures like mean waiting times and queue lengths, but

can also be used to compute more detailed performance measures like tail probabilities and individual state probabilities.
The main limitations of the PSA are the available amounts of storage capacity and computation time, restricting the use to fairly small and moderately-sized models. We refer to [4] for a fairly complete survey of various aspects of the PSA, including useful ideas about efficient memory management and improvements of the convergence of the power series, which have strongly improved the performance and the applicability of the PSA.

Overview of the results
Section 2 contains a detailed model description. In section 3 we show how the PSA can be applied to analyze polling models with Markovian server routing. It is also shown how derivatives of the performance measures with respect to the routing probabilities can be determined by means of the PSA, opening the possibility of performing sensitivity analysis and optimization of performance measures with respect to the routing probabilities.
In section 4, the PSA is used for comparing the performance of polling models with probabilistic and periodic server routing. Numerical experiments with the PSA indicate that the mean total amount of work in the system is structurally *larger* under probabilistic polling. However, it is shown that a similar dominance relation is not generally valid for the individual mean waiting times at the queues. We give an intuitive argument for these observations in terms of the spacing of the visits in time.
In section 5 we consider the problem of characterizing combinations of routing probabilities that minimize the mean amount of waiting work in the system. However, the dimension of this *optimization* problem grows quadratically in the number of queues, making numerical procedures based on standard techniques for non-linear optimization very time consuming when the number of queues becomes large. Therefore, we focus on finding *qualitative*, instead of quantitative, properties of optimal routing matrices. For symmetrical models, Liu et al. [20] have shown that each cyclic server routing (which occurs as a special case of Markovian server routing by taking $p_{i,j} = 1$ if $j = i + 1$ and 0 otherwise) is optimal. Numerical experiments with the PSA suggest that the cyclic service order is a *stable* optimum in the sense that it remains optimal for slight perturbations of the system parameters. The validity of this statement is confirmed by a number of numerical examples. When the model becomes even more asymmetrical, the cyclic visit order may become suboptimal. We observe the *tendency* of the optimal probabilistic server routing towards (partially) *deterministic routing*. That is, for surprisingly many queues i there exists a specific queue k_i such that the optimal routing probabilities are equal to $p_{i,j} = 1$ if $j = k_i$ and 0 otherwise. In addition, we examine the influence of system parameters on optimal routing matrices. These investigations point out that the optimal routing matrices can be roughly classified into a limited number of types of solutions, each having specific characteristics that can be interpreted fairly easily. On the basis of the insights obtained from numerical experience, we propose a number of rough guidelines for constructing optimal routing matrices. The validity of these guidelines is illustrated by a number of examples.
Finally, in Section 6 we discuss some topics for further research.

2 Model description

Consider a polling model with s infinite-buffer queues $Q_1,\ldots,Q_s$. Customers arrive at Q_i according to Poisson arrival process with rate λ_i, $i=1,\ldots,s$. The total arrival rate is denoted by $\Lambda := \Sigma_{i=1}^{s}\lambda_i$. The service times of customers at Q_i are Coxian distributed with parameters $\pi_i^{1,\xi},\mu_i^{1,\xi},\Psi_i^1$, $\xi=1,\ldots,\Psi_i^1$; that is, with probability $\pi_i^{1,\xi}$ a service at Q_i is composed of subsequent phases $\xi,\xi-1,\ldots,1,\xi=1,\ldots,\Psi_i^1$, $i=1,\ldots,s$. Denote by $\beta^{(1)}=(\beta_1^{(k)},\ldots,\beta_s^{(k)})$ the vector of k-th moments of the service times at the various queues, $k=1,2$. Denote by $\beta_k := (1/\Lambda)\Sigma_{i=1}^{s}\lambda_i\beta_i^{(k)}$ the k-th moment of an arbitrary service time, $k=1,2$. Let $\rho := \Sigma_{i=1}^{s}\lambda_i\beta_i^{(1)}$ denote the total offered load to the system. Because the offered load ρ will be used as a variable in the PSA, we define

$$a_i := \lambda_i/\rho, \tag{1}$$

referred to as the relative arrival rate to Q_i, $i=1,\ldots,s$. Let $\boldsymbol{a} := (a_1,\ldots,a_s)$. Note that it follows from the definition of the relative arrival rates (cf. (1)) that $\Sigma_{i=1}^{s}a_i = 1/\beta_1$.

The service discipline at Q_i is the so-called Bernoulli service strategy with parameter q_i ($0 \le q_i \le 1$), which works as follows. When the server arrives at Q_i finding that queue non-empty, at least one customer at Q_i is served; otherwise, the server moves to the next queue. Moreover, if after a service completion at Q_i the queue is still non-empty, with probability q_i another customer at Q_i is served; otherwise, the server proceeds to the next queue. It should be noted that the class of Bernoulli service disciplines contains the classical 1-limited and exhaustive service strategies at special cases for $q_i=0$ and $q_i=1$, respectively. The vector of Bernoulli parameters $\boldsymbol{q}=(q_1,\ldots,q_s)$ is referred to as a *Bernoulli schedule*.

The server visits the queues according to a Markovian polling scheme with routing matrix $\mathbf{P}=(p_{i,j})$; that is, after a departure of the server from Q_i the server starts to move to Q_j with probability $p_{i,j}$, $i,j=1,\ldots,s$. In this way, the process of successive visits of the server to the various queues can be described as a discrete-time Markov chain $D=\{d_k,\ k=0,1,\ldots\}$ with state space $\{1,\ldots,s\}$, where $\{d_k=i\}$ denotes the event that the k-th visited queue is Q_i, $i=1,\ldots,s$, $k=0,1,\ldots$. Throughout it is assumed that D is irreducible. The times needed by the server to move from Q_i to Q_j are Coxian distributed with parameters $\pi_{i,j}^{0,\xi},\mu_{i,j}^{0,\xi},\Psi_{i,j}^{0,\xi}$, $\xi=1,\ldots,\Psi_{i,j}^0$, $i,j=1,\ldots,s$, which are used in a similar way as for the service times. Denote by $\sigma_{i,j}^{(k)}$ the k-th moment of the switch-over times to move from Q_i to Q_j, $i,j=1,\ldots,s$, $k=1,2$. Because D is an irreducible Markov chain on a finite state space, it possesses a stationary distribution $\{\omega_i,\ i=1,\ldots,s\}$, which is uniquely determined by the following set of equations (cf. [23]):

$$\omega_i = \sum_{j=1}^{s}\omega_j p_{j,i}\ (i=1,\ldots,s); \qquad \sum_{j=1}^{s}\omega_j = 1. \tag{2}$$

Necessary and sufficient conditions for the stability of the system have been derived in [12]. For the present model with Markovian server routing these conditions read:

$$\rho\left[1+\frac{a_i\sigma(1-q_i)}{\omega_i}\right] < 1, \tag{3}$$

where

$$\sigma := \sum_{i=1}^{s}\omega_i\sum_{j=1}^{s}p_{i,j}\sigma_{i,j}^{(1)}, \tag{4}$$

i.e. the mean of an arbitrary switch-over time. Throughout, it is assumed that these conditions are satisfied and that the system is in steady state.
Finally, we introduce some national conventions. For an event E, the expression $I\{E\}$ will stand for the indicator function on E. The vector $\boldsymbol{e}_j \in \mathbb{N}^s$ will stand for the j-th unit vector, i.e. the vector for which the j-th component is equal to 1 and all other components are 0. For a vector $\boldsymbol{v} \in \mathbb{N}^s$, the symbol $|\boldsymbol{v}|$ will stand for $v_1 + \cdots + v_s$. For a set $\mathcal{A}$, the symbol $|\mathcal{A}|$ will stand for the cardinality of $\mathcal{A}$.

3 The power-series algorithm

In this section we show how the present model can be analyzed by means of the PSA. To apply the PSA, we first describe the present model as a continuous-time Markov chain representation of the model. Therefore, we first We define the state probabilities and formulate the global balance equations. Then the state probabilities, and their derivatives with respect to the routing probabilities, are expressed as power series in the offered load to the system. Finally, we derive a computational scheme to compute the coefficients of these power series.

3.1 Balance equations

Let $\{\boldsymbol{N}(t) = (N_1(t), \ldots, N_s(t)), t \geq 0\}$ be the joint queue-length process. Evidently, this process is not a Markov process, e.g. because the departure rate depends on whether the server is switching or serving. To transform the process $\{\boldsymbol{N}(t), t \geq 0\}$ into a Markov process, we introduce a triple $(H(t), G(t), \Xi(t))$ of supplementary variables. Let $\{H(t) = h;\ G(t) = 1;\ \Xi(t) = \xi\}$ denote the event that at time t the server is serving at Q_h and that ξ is the current phase number of this service, $h = 1, \ldots, s$, $\xi = 1, \ldots, \Psi_h^1$, $t \geq 0$. Moreover, let the event $\{H(t) = h;\ G(t) = -g;\ \Xi(t) = \xi\}$ indicate that at time t the server is switching from Q_g towards Q_h, and that ξ is the current phase number of this switch-over, $g, h = 1, \ldots, s$, $\xi = 1, \ldots, \Psi_{g,h}^0$, $t \geq 0$. For ease of the discussion, it is assumed that the supplementary space is the same for all $\boldsymbol{n} \in \mathbb{N}^s$, and is given by

$$\mathcal{S} := \{1, \ldots, s\} \times \{-s, \ldots, -1, 1\} \times \{1, \ldots, K\}, \text{ where } K = \max_{i,j} \left\{\Psi_{i,j}^0, \Psi_i^1\right\}. \tag{5}$$

Denote by $(\boldsymbol{N}, H, G, \Xi)$ random variables with as joint distribution the stationary distribution of $(\boldsymbol{N}(t), H(t), G(t), \Xi(t))$.

Define the state probabilities as follows: for $(\boldsymbol{n}, h, -g, \xi) \in \mathbb{N}^s \times \mathcal{S}$,

$$p(\boldsymbol{n}, h, -g, \xi) := \Pr\{(\boldsymbol{N}, H, G, \Xi) = (\boldsymbol{n}, h, -g, \xi)\}. \tag{6}$$

Because of the stability of the system the rate into each state is equal to the rate out of that state. The state probabilities satisfy the following balance equations for the states in which the server is switching (from Q_g to Q_h): for $\boldsymbol{n} \in \mathbb{N}^s$, $g, h = 1, \ldots, s$, $\xi = 1, \ldots, \Psi_{g,h}^0$,

$$\begin{aligned}
&\left[\rho \sum_{j=1}^{s} a_j + \mu_{g,h}^{0,\xi}\right] p(\boldsymbol{n}, h, -g, \xi) = \mu_{g,h}^{0,\xi+1} p(\boldsymbol{n}, h, -g, \xi + 1) I\left\{\xi < \Psi_{g,h}^0\right\} \\
&+\rho \sum_{j=1}^{s} a_j p(\boldsymbol{n} - \boldsymbol{e}_j, h, -g, \xi) I\{n_j > 0\} + \pi_{g,h}^{0,\xi} p_{g,h} \sum_{f=1}^{s} \mu_{f,g}^{0,1} p(\boldsymbol{n}, g, -f, 1) I\{n_g = 0\} \\
&+\mu_g^{1,1} \pi_{g,h}^{1,\xi} p_{g,h} p(\boldsymbol{n} + \boldsymbol{e}_g, g, 0, 1)[1 - q_g I\{n_g > 0\}].
\end{aligned} \tag{7}$$

The first term at the right-hand side indicates a phase transition in a switch-over time from Q_g to Q_h. The second term corresponds to an arrival while the server is switching from Q_g to Q_h. The third term describes that the server finds Q_g empty upon arrival and immediately starts to move to Q_h. Finally, the fourth term indicates that the server departs from Q_g after service completion of a customer at that queue and proceeds to Q_h.

The global balance equations for the states in which the server is serving (at Q_h) read as follows: for $\boldsymbol{n} \in \mathbb{N}^s$, $h = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_h^1$, $n_h > 0$,

$$\begin{aligned}&\left[\rho \sum_{j=1}^{s} a_j + \mu_h^{1,\xi}\right] p(\boldsymbol{n},h,1,\xi) = \mu_h^{1,\xi+1} p(\boldsymbol{n},h,1,\xi+1) I\{\xi < \Psi_h^1\} \\ &+\rho \sum_{j=1}^{s} a_j p(\boldsymbol{n}-\boldsymbol{e}_j,h,1,\xi) I\{n_j > 0\} + \pi_h^{1,\xi} \sum_{g=1}^{s} \mu_{g,h}^{0,1} p(\boldsymbol{n},h,-g,1) \\ &+q_h \mu_h^{1,1} \pi_h^{1,\xi} p(\boldsymbol{n}+\boldsymbol{e}_h,h,1,1).\end{aligned} \tag{8}$$

The first term indicates a phase transition in a service of a customer at Q_h. The second term corresponds to an arrival during the service of a customer at Q_h. The third term describes that the server arrives at Q_h and immediately starts to serve a customer at that queue. The fourth term indicates that after a service completion at Q_h the server immediately starts to serve the next customer at that queue.

Because the server can not be serving at an empty queue, we have: for $\boldsymbol{n} \in \mathbb{N}^s$, $h = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_h^1$,

$$p(\boldsymbol{n},h,1,\xi) = 0 \quad \text{if} \quad n_h = 0, \tag{9}$$

and according to the law of total probability, we have

$$\sum_{\boldsymbol{n}\in\mathbb{N}^s} \sum_{h=1}^{s} \left\{ \sum_{g=1}^{s} \sum_{\xi=1}^{\Psi_{g,h}^0} p(\boldsymbol{n},h,-g,\xi) + \sum_{\xi=1}^{\Psi_h^1} p(\boldsymbol{n},h,1,\xi) \right\} = 1. \tag{10}$$

The set of balance equations (7), (8), together with the law of total probability (10), forms an infinite set of linear equations between the state probabilities. However, this set of equations is not recursively solvable. In the next section we will show how the PSA can be used to transform this set of equations into a (mainly) recursively solvable set of equations.

3.2 Computational scheme

The basic idea of the PSA is to transform a non-recursively solvable set of balance equations into a recursively solvable set of equations by expressing the state probabilities (6) as power series in the offered load to the system in light traffic. By substituting these expressions into the balance equations, one may obtain a complete recursive computational scheme to calculate the coefficients of these power series. In this section we will show how such a computational scheme can be obtained for the present model.

The PSA relies on the following light-traffic property: for $(\boldsymbol{n},h,-g,\xi) \in \mathbb{N}^s \times \mathcal{S}$,

$$p(\boldsymbol{n},h,-g,\xi) = O(\rho^{|\boldsymbol{n}|}), \quad \rho \downarrow 0. \tag{11}$$

Here, the limits are taken in such a way that the relative arrival rates remain fixed (cf. (1)). We refer to Van den Hout and Blanc [29] for conditions under which this property is valid. For the present model these properties are satisfied. Based on property (11), we express the state probabilities as power series in ρ as follows: for $(\boldsymbol{n}, h, -g, \xi) \in \mathbb{N}^s \times \mathcal{S}$,

$$p(\boldsymbol{n}, h, -g, \xi) = \rho^{|\boldsymbol{n}|} \sum_{k=0}^{\infty} \rho^k b_0(k; \boldsymbol{n}, h, -g, \xi). \tag{12}$$

We refer to Van den Hout and Blanc [29], [30] for conditions on the convergence of the power-series expansions.

There are various ways to define the derivatives of the routing probabilities with respect to other routing probabilities, $\frac{\partial p_{g,h}}{\partial p_{i,j}}$. One way to do so is to consider routing probability $p_{g,h}$ as function of underlying variables $t_{g,h} \geq 0$ as follows: for $g, h = 1, \ldots, s$,

$$p_{g,h} = \frac{t_{g,h}}{\sum_{k=1}^{s} t_{g,k}}, \tag{13}$$

evaluated at $\Sigma_{k=1}^{s} t_{g,k} = 1$. We define the derivatives of the routing probabilities as follows: for $g, h, i, j = 1, \ldots, s$,

$$\frac{\partial p_{g,h}}{\partial p_{i,j}} := \left[\frac{\partial p_{g,h}}{\partial t_{i,j}} \right]_{\sum_{k=1}^{s} t_{g,k} = 1}. \tag{14}$$

It is readily verified by applying standard rules for differentiation that: for $g, h, i, j = 1, \ldots, s$,

$$\frac{\partial p_{g,h}}{\partial p_{i,j}} = I\{i = g\} \left[I\{j = h\} - p_{g,h} \right]. \tag{15}$$

Using this definition (14), the derivatives of the state probabilities with respect to the routing probabilities are well-defined. For notational convenience, we define the following linear ordering of the derivatives: for $r = 1, \ldots, s^2$, $(\boldsymbol{n}, h, -g, \xi) \in \mathbb{N}^s \times \mathcal{S}$,

$$p_r(\boldsymbol{n}, h, -g, \xi) := \frac{\partial}{\partial p_{i,j}} p(\boldsymbol{n}, h, -g, \xi), \tag{16}$$

where i, j and r are related through

$$r = (i - 1)s + j, \quad i, j = 1, \ldots, s. \tag{17}$$

The derivatives of the state probabilities (16) can be expressed as power series in ρ as follows: for $r = 1, \ldots, s^2$, $(\boldsymbol{n}, h, -g, \xi) \in \mathbb{N}^s \times \mathcal{S}$,

$$p_r(\boldsymbol{n}, h, -g, \xi) = \rho^{|\boldsymbol{n}|} \sum_{k=0}^{\infty} \rho^k b_r(k; \boldsymbol{n}, h, -g, \xi). \tag{18}$$

Because ρ does not depend on the routing probabilities, the coefficients $b_r(k; \boldsymbol{n}, h, -g, \xi)$ can be obtained by termwise differentiation of the coefficients $b_0(k; \boldsymbol{n}, h, -g, \xi)$: for $r = 1, \ldots, s^2$, $(k; \boldsymbol{n}, h, -g, \xi) \in \mathbb{N}^{1+s} \times \mathcal{S}$,

$$b_r(k;\boldsymbol{n},h,-g,\xi) := \frac{\partial}{\partial p_{i,j}} b_0(k;\boldsymbol{n},h,-g,\xi), \tag{19}$$

where i,j and r are related through (17). Substituting the power-series expansions (12) into the balance equations (7) and (8), and equating corresponding powers of ρ leads to the following sets of linear relations between the coefficients of the power series in (12) and (16): for $r = 0,1,\ldots,s^2$, $\boldsymbol{n} \in \mathbb{N}^s$, $h,g = 1,\ldots,s$, $\xi = 1,\ldots,\Psi^0_{g,h}$, $k = 0,1,\ldots$,

$$\begin{aligned}
&\mu^{0,\xi}_{g,h} b_r(k;\boldsymbol{n},h,-g,\xi) = \mu^{0,\xi+1}_{g,h} b_r(k;\boldsymbol{n},h,-g,\xi+1) I\left\{\xi < \Psi^0_{g,h}\right\} \\
&+ \sum_{j=1}^{s} a_j \left[b_r(k;\boldsymbol{n}-\boldsymbol{e}_j,h,-g,\xi) I\{n_j > 0\} - b_r(k-1;\boldsymbol{n},h,-g,\xi) I\{k > 0\}\right] \\
&+\pi^{0,\xi}_{g,h} p_{g,h} \sum_{f=1}^{s} \mu^{0,1}_{f,g} b_r(k;\boldsymbol{n},g,-f,1) I\{n_g = 0\} \\
&+\pi^{0,\xi}_{g,h} \left[\frac{\partial p_{g,h}}{\partial p_{i,j}}\right] \sum_{f=1}^{s} \mu^{0,1}_{f,g} b_0(k;\boldsymbol{n},g,-f,1) I\{r > 0\} I\{n_g = 0\} \\
&+\mu^{1,1}_{g} \pi^{0,\xi}_{g,h} p_{g,h} b_r(k;\boldsymbol{n}+\boldsymbol{e}_g,g,1,1)\left[1 - q_g I\{n_g > 0\}\right] I\{k > 0\} \\
&+\mu^{1,1}_{g} \pi^{0,\xi}_{g,h} \left[\frac{\partial p_{g,h}}{\partial p_{i,j}}\right] b_0(k;\boldsymbol{n}+\boldsymbol{e}_g,g,1,1)\left[1 - q_g I\{n_g > 0\}\right] I\{k > 0\} I\{r > 0\};
\end{aligned} \tag{20}$$

and for the coefficients corresponding to the states in which the server is serving: for $r = 0,1,\ldots,s^2$, $\boldsymbol{n} \in \mathbb{N}^s$, $h = 1,\ldots,s$, $\xi = 1,\ldots,\Psi^1_h$, $k = 0,1,\ldots$,

$$\begin{aligned}
&\mu^{1,\xi}_{h} b_r(k;\boldsymbol{n},h,1,\xi) = \mu^{1,\xi+1}_{h} b_r(k;\boldsymbol{n},h,1,\xi+1) I\{\xi < \Psi^1_h\} \\
&+ \sum_{j=1}^{s} a_j \left[b_r(k;\boldsymbol{n}-\boldsymbol{e}_j,h,1,\xi) I\{n_j > 0\} - b_r(k-1;\boldsymbol{n},h,1,\xi) I\{k > 0\}\right] \\
&+\pi^{1,\xi}_{h} \sum_{g=1}^{s} \mu^{0,1}_{g,h} b_r(k;\boldsymbol{n},h,-g,1) + q_h \mu^{1,1}_{h} \pi^{1,\xi}_{h} b_r(k-1;\boldsymbol{n}+\boldsymbol{e}_h,h,1,1) I\{k > 0\}.
\end{aligned} \tag{21}$$

For convenience, we rewrite the set of equations (20) as follows: for $r = 0,1,\ldots,s^2$, $\boldsymbol{n} \in \mathbb{N}^s$, $h,g = 1,\ldots,s$, $\xi = 1,\ldots,\Psi^0_{g,h}$, $k = 0,1,\ldots$,

$$\begin{aligned}
\mu^{0,\xi}_{g,h} b_r(k;\boldsymbol{n},h,-g,\xi) &= \pi^{0,\xi}_{g,h} p_{g,h} \sum_{f=1}^{s} \mu^{0,1}_{f,g} b_r(k;\boldsymbol{n},g,-f,1) I\{n_g = 0\} \\
&\quad + y_r(k;\boldsymbol{n},h,-g,\xi),
\end{aligned} \tag{22}$$

where

$$\begin{aligned}
&y_r(k;\boldsymbol{n},h,-g,\xi) := \mu^{0,\xi+1}_{g,h} b_r(k;\boldsymbol{n},h,-g,\xi+1) I\left\{\xi < \Psi^0_{g,h}\right\} \\
&+ \sum_{j=1}^{s} a_j \left[b_r(k;\boldsymbol{n}-\boldsymbol{e}_j,h,-g,\xi) I\{n_j > 0\} - b_r(k-1;\boldsymbol{n},h,-g,\xi) I\{k > 0\}\right] \\
&+\pi^{0,\xi}_{g,h} \left[\frac{\partial p_{g,h}}{\partial p_{i,j}}\right] \sum_{f=1}^{s} \mu^{0,1}_{f,g} b_0(k;\boldsymbol{n},g,-f,1) I\{r > 0\} I\{n_g = 0\} \\
&+\mu^{1,1}_{g} \pi^{0,\xi}_{g,h} p_{g,h} b_r(k-1;\boldsymbol{n}+\boldsymbol{e}_g,g,1,1)\left[1 - q_g I\{n_g > 0\}\right] I\{k > 0\} \\
&+\mu^{1,1}_{g} \pi^{0,\xi}_{g,h} \left[\frac{\partial p_{g,h}}{\partial p_{i,j}}\right] b_0(k-1;\boldsymbol{n}+\boldsymbol{e}_g,g,1,1)\left[1 - q_g I\{n_g > 0\}\right] I\{k > 0\} I\{r > 0\}.
\end{aligned} \tag{23}$$

To derive a computation order for the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$, we need to explore the structure of the set of equations (22). To this end, we assign to each $\boldsymbol{n} \in \mathbb{N}^s$ the *null-set* corresponding to $\boldsymbol{n}$ as follows: for $\boldsymbol{n} \in \mathbb{N}^s$,

$$\mathcal{N}^{(0)}_{\boldsymbol{n}} := \{1 \le g \le s \mid n_g = 0\}, \tag{24}$$

i.e. the set of empty queues when the joint queue-length vector is $\boldsymbol{n}$. In addition, we define: for $(k;\boldsymbol{n}) \in \mathbb{N}^{1+s}$, $g \in \mathcal{N}_{\boldsymbol{n}}^{(0)}$, $r = 0,1,\ldots,s^2$,

$$C_r(k;\boldsymbol{n},g) := \sum_{f=1}^{s} \mu_{f,g}^{0,1} b_r(k;\boldsymbol{n},g,-f,1). \tag{25}$$

Then, by summing both sides of the equations (22) over $g = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_{g,h}^0$, we obtain the following set of equations: for $r = 0,1,\ldots,s^2$, $(k;\boldsymbol{n}) \in \mathbb{N}^{1+s}$, $h \in \mathcal{N}_{\boldsymbol{n}}^{(0)}$,

$$C_r(k;\boldsymbol{n},h) = \sum_{g \in \mathcal{N}_{\boldsymbol{n}}^{(0)}} C_r(k;\boldsymbol{n},g) p_{g,h} + \bar{y}_r(k;\boldsymbol{n},h), \tag{26}$$

where

$$\bar{y}_r(k;\boldsymbol{n},h) := \sum_{g=1}^{s} \sum_{\xi=1}^{\Psi_{g,h}^0} y_r(k;\boldsymbol{n},h,-g,\xi). \tag{27}$$

Once for given triple $(r,k;\boldsymbol{n})$ the quantities $C_r(k;\boldsymbol{n},g)$, $g \in \mathcal{N}_{\boldsymbol{n}}^{(0)}$, are known, the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$, $(h,-g,\xi) \in \mathcal{S}$, can be obtained from the following relation: for $r = 0,1,\ldots,s^2$, $(k;\boldsymbol{n}) \in \mathbb{N}^{1+s}$, $h,g = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_{g,h}^0$,

$$\mu_{g,h}^{0,\xi} b_r(k;\boldsymbol{n},h,-g,\xi) = \pi_{g,h}^{0,\xi} p_{g,h} C_r(k;\boldsymbol{n},g) + y_r(k;\boldsymbol{n},h,-g,\xi), \tag{28}$$

with the convention that $C_r(k;\boldsymbol{n},g) := 0$ for $g \notin \mathcal{N}_{\boldsymbol{n}}^{(0)}$.

We are now ready to define an ordering of the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$ such that they can be determined recursively. Let us first define an ordering for the states with $r = 0$. To this end, we define the following ordering $\prec$ over the $(k;\boldsymbol{n})$-combinations: for $(k;\boldsymbol{n},h,-g,\xi) \in \mathbb{N}^{1+s} \times \mathcal{S}$,

$$\begin{aligned} &(k;\boldsymbol{n},h,-g,\xi) \prec (\hat{k};\hat{\boldsymbol{n}},\hat{h},-\hat{g},\hat{\xi}) \\ &\text{if } \left[k + |\boldsymbol{n}| < \hat{k} + |\hat{\boldsymbol{n}}|\right] \vee \left[k + |\boldsymbol{n}| = \hat{k} + |\hat{\boldsymbol{n}}| \wedge k < \hat{k}\right]. \end{aligned} \tag{29}$$

For given $(k;\boldsymbol{n})$ *and* h, we define the following ordering over the couples $(-g,\xi)$, $g = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_{g,h}^0$, $\hat{\xi} = 1,\ldots,\Psi_h^1$,

$$(k;\boldsymbol{n},h,-g,\xi) \prec (k;\boldsymbol{n},h,1,\hat{\xi}); \tag{30}$$

thus, for given $(k;\boldsymbol{n},h)$, the coefficients corresponding to the states in which the server is serving are of higher order than those corresponding to states in which the server is switching. In addition, for given $(k;\boldsymbol{n},h)$, the states in which the server is switching are (partially) ordered as follows:

$$(k;\boldsymbol{n},h,-g,\xi) \prec (k;\boldsymbol{n},h,-\hat{g},\hat{\xi}) \text{ if } \left[g \notin \mathcal{N}_{\boldsymbol{n}}^{(0)} \wedge \hat{g} \in \mathcal{N}_{\boldsymbol{n}}^{(0)}\right]; \tag{31}$$

thus, the coefficients corresponding to states in which the server is switching after a departure from a non-empty queue are of lower order than those for the states in which the server is moving just after a departure from an empty queue. For given $(k;\boldsymbol{n},h)$ *and* $g = 1,\ldots,s$, the states $(k;\boldsymbol{n},h,-g,\xi)$ are ordered as follows: for $(k;\boldsymbol{n}) \in \mathbb{N}^{1+s}$, $h,g = 1,\ldots,s$, $\xi,\tilde{\xi} = 1,\ldots,\Psi_{g,h}^0$,

$$(k;\boldsymbol{n},h,-g,\xi) \prec (k;\boldsymbol{n},h,-g,\hat{\xi}) \text{ if } \xi > \hat{\xi}; \tag{32}$$

thus, for the states in which the server is moving from a given queue Q_g towards Q_h are ranked in increasing order with respect to the component ξ as $\Psi^0_{g,h}, \Psi^0_{g,h} - 1, \ldots, 1$. We emphasize that this ordering, defined by (29)-(32) is only partial, so that not all couples of vectors $(k;\boldsymbol{n},h,-g,\xi) \in \mathbb{N}^{1+s} \times \mathcal{S}$ are mutually ordered.

Thus far, we have only considered the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$ with $r = 0$. To derive an ordering for the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$, $r = 0,1,\ldots,s^2$, we extend the partial ordering $\prec$, defined in (29)-(32), to the states $(r,k;\boldsymbol{n},h,-g,\xi)$ as follows: for $(r,k;\boldsymbol{n},h,-g,\xi)$, $(\hat{r},\hat{k},\hat{\boldsymbol{n}};\hat{h},-\hat{g},\hat{\xi}) \in \{0,1,\ldots,s^2\} \times \mathbb{N}^{1+s} \times \mathcal{S}$,

$$\begin{aligned}&(r,k;\boldsymbol{n},h,-g,\xi) \tilde{\prec} (\hat{r},\hat{k};\hat{\boldsymbol{n}},\hat{h},-\hat{g},\hat{\xi})\\ &\text{if } [r = 0 \wedge \hat{r} > 0] \vee \left[r = \hat{r} \wedge (k;\boldsymbol{n},h,-g,\xi) \prec (\hat{k},\hat{\boldsymbol{n}};\hat{h},-\hat{g},\hat{\xi})\right].\end{aligned} \tag{33}$$

One may verify that under the partial ordering $\tilde{\prec}$ all coefficients in (23) are of lower order than $b_r(k;\boldsymbol{n},h,-g,\xi)$, and that all terms at the right-hand side of (21) are of lower order with respect to $\tilde{\prec}$ than $b_r(k;\boldsymbol{n},h,1,\xi)$.

Hence, it remains to consider the solvability of the set of equations (22) for given $(r,k;\boldsymbol{n})$. To this end, note that for given $(r,k;\boldsymbol{n})$ the set of equations (22) is uniquely solvable if and only if the set (26) is uniquely solvable. To consider the solvability of (26), a distinction has to be made between the empty and non-empty states.
For $\boldsymbol{n} \neq \boldsymbol{0}$, the reduced routing matrix $\mathbf{P} = (p_{i,j})$, $i,j \in \mathcal{N}^{(0)}_{\boldsymbol{n}}$ is substochastic (cf. [23]), which guarantees that the set (26) indeed possesses a unique solution $C_r(k;\boldsymbol{n},g)$, $g \in \mathcal{N}^{(0)}_{\boldsymbol{n}}$. To consider the solvability of the set of equations (26) for the states with $\boldsymbol{n} = \boldsymbol{0}$, one may verify, by summing both sides over $h \in \mathcal{N}^{(0)}_{\boldsymbol{0}} = \{1,\ldots,s\}$, that for given $(r,k;\boldsymbol{n})$ relations (26) form a *dependent* set of equations. One may verify that this set of equations is not contradictive because of a necessary balance between the empty states and states with exactly one customer in the system, implying: for $r = 0,1,\ldots,s^2$, $k = 0,1,\ldots$,

$$\sum_{h=1}^{s} \bar{y}_r(k;\boldsymbol{0},h) = \sum_{g=1}^{s}\sum_{h=1}^{s}\sum_{\xi=1}^{\Psi^0_{g,h}} y_r(k;\boldsymbol{0},h,-g,\xi) = 0. \tag{34}$$

An additional equation follows directly from the law of total probability (10), which implies: for $r = 0,1,\ldots,s^2$, $k = 0,1,\ldots$,

$$\sum_{g=1}^{s}\sum_{h=1}^{s}\sum_{\xi=1}^{\Psi^0_{g,h}} b_r(k;\boldsymbol{0},h,-g,\xi) = Y_r(k), \tag{35}$$

where for $r = 0,1,\ldots,s^2$, $Y_r(0) := I\{r = 0\}$ and for $k = 1,2,\ldots$,

$$\begin{aligned}&Y_r(k) :=\\ &-\sum_{0<|\boldsymbol{n}|\leq k}\sum_{h=1}^{s}\left\{\sum_{g=1}^{s}\sum_{\xi=1}^{\Psi^0_{g,h}} b_r(k-|\boldsymbol{n}|;\boldsymbol{n},h,-g,\xi) + \sum_{\xi=1}^{\Psi^1_h} b_r(k-|\boldsymbol{n}|;\boldsymbol{n},h,1,\xi)\right\}.\end{aligned} \tag{36}$$

Equation (35) can be rewritten in terms of the variables $C_r(k;\boldsymbol{0},g)$ in the following way: for $r = 0,1,\ldots,s^2$, $k = 0,1,\ldots$,

$$\sum_{g=1}^{s}\left(\sum_{h=1}^{s}\sum_{\xi=1}^{\psi_{g,h}^0}\frac{p_{g,h}}{\mu_{g,h}^{0,\xi}}\sum_{\psi=\xi}^{\psi_{g,h}^0}\pi_{g,h}^{0,\psi}\right)C_r(k;\mathbf{0},g) = Y_r(k) - \sum_{g=1}^{s}\sum_{h=1}^{s}\sum_{\xi=1}^{\psi_{g,h}^0}\frac{1}{\mu_{g,h}^{0,\xi}}\sum_{\psi=\xi}^{\psi_{g,h}^0}y_r(k;\mathbf{0},h,-g,\psi). \tag{37}$$

Now, it suffices to show that the set of equations (26), (37) or equivalently, the set (22), (35) is uniquely solvable. To this end, it should be noted that the coefficients at the left-hand side of these sets of equations are independent of r and k, so that it is sufficient to consider the solvability of these sets of equations for $r = 0$ and $k = 0$. Then the solvability is readily established by observing that the continuous-time Markov process $\{(\boldsymbol{N}(t), H(t), G(t), \Xi(t)),\ t \geq 0\}$, conditioned on the event $\boldsymbol{N}(t) = \mathbf{0}$, is irreducible on the state space $\{\mathbf{0}\} \times \mathcal{S}$. Alternatively, it is rather tedious, but straightforward, to verify that the determinant of all but one of the equations (22), together with (35), is equal to $\Delta = \sigma\Pi_{g=1}^{s}\Pi_{h=1}^{s}\Pi_{\xi=1}^{\psi_{g,h}^0}\mu_{g,h}^{0,\xi} > 0$, implying that this set of equations indeed possesses a unique solution.

In practice, one is usually only interested in a limited number, say L, of performance measures, instead of in all individual state probabilities. Let $g^{(l)}(\boldsymbol{n}, h, -g, \xi)$ be a real-valued function of the state space. Then general performance measures of the form $\mathrm{E}\{g^{(l)}(\boldsymbol{N}, H, G, \Xi)\}$, $l = 1, \ldots, L$, can be expressed in terms of the coefficients of the power series as: for $l = 1, \ldots, L$,

$$\mathrm{E}\{g^{(l)}(\boldsymbol{N}, H, G, \Xi)\} = \sum_{k=0}^{\infty}\rho^k f^{(l)}(k), \tag{38}$$

where for $k = 0, 1, \ldots$,

$$f^{(l)}(k) := \sum_{0\leq|\boldsymbol{n}|\leq k}\sum_{(h,-g,\xi)\in\mathcal{S}} g^{(l)}(\boldsymbol{n}, h, -g, \xi)b_0(k - |\boldsymbol{n}|; \boldsymbol{n}, h, -g, \xi). \tag{39}$$

We assume that the function $g^{(l)}(\boldsymbol{n}, h, -g, \xi)$ does not depend on the routing probabilities and that the performance measures in (38) are (partially) differentiable with respect with respect to the routing probabilities. Then the performance measures (38) and their derivatives (40) can be expressed as power series in ρ as follows: for $i, j = 1, \ldots, s$, $r = 1, \ldots, s^2$, $l = 1, \ldots, L$,

$$\frac{\partial}{\partial p_{i,j}}\mathrm{E}\{g^{(l)}(\boldsymbol{N}, H, G, \Xi)\} = \sum_{k=0}^{\infty}\rho^k f_r^{(l)}(k), \tag{40}$$

where for $k = 0, 1, \ldots$,

$$f_r^{(l)}(k) := \sum_{0\leq|\boldsymbol{n}|\leq k}\sum_{(h,-g,\xi)\in\mathcal{S}} g^{(l)}(\boldsymbol{n}, h, -g, \xi)b_r(k - |\boldsymbol{n}|; \boldsymbol{n}, h, -g, \xi), \tag{41}$$

and where i, j and r are related through (17).

Because of limitations on the available amounts of computation time and storage capacity, only a limited number of coefficients can be computed. Let M be the number of terms that one wants or has to compute. The following computational scheme shows how the performance measures $\mathrm{E}\{g^{(l)}(\boldsymbol{N}, H, G, \Xi)\}$, $l = 1, \ldots, L$, and their derivatives with respect to the routing probabilities, can be computed:

step 1 : $m := 0$; for $l = 1,\ldots,L$, $f^{(l)}(0) := 0$ and $f_r^{(l)}(0) := 0$, $r = 1,\ldots,s^2$;

step 2 : for all $(k;\boldsymbol{n})$ with $\boldsymbol{n} \neq \boldsymbol{0}$ and $k + |\boldsymbol{n}| = m$,

(1) for $g \notin \mathcal{N}_{\boldsymbol{n}}^{(0)}$, determine $b_r(k;\boldsymbol{n},h,-g,\xi)$, $h = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_{g,h}^0$, from (22), $r = 0,1,\ldots,s^2$; update $f^{(l)}(m)$ and $f_r^{(l)}(m)$, $l = 1,\ldots,L$, $r = 1,\ldots,s^2$, according to (39) and (41), respectively;

(2) determine $C_r(k;\boldsymbol{n},g)$, $g \in \mathcal{N}_{\boldsymbol{n}}^{(0)}$, by solving the set of equations (26) and determine $b_r(k;\boldsymbol{n},h,-g,\xi)$, $h = 1,\ldots,s$, $g \in \mathcal{N}_{\boldsymbol{n}}^{(0)}$, $\xi = 1,\ldots,\Psi_{g,h}^0$, from (28), in increasing order with respect to $\prec$ (33), $r = 0,1,\ldots,s^2$; update $f^{(l)}(m)$ and $f_r^{(l)}(m)$, $l = 1,\ldots,L$, $r = 1,\ldots,s^2$, according to (39) and (41), respectively;

(3) determine $b_r(k;\boldsymbol{n},h,1,\xi)$, $h = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_h^0$, according to (21), $r = 0,1,\ldots,s^2$; update $f^{(l)}(m)$ and $f_r^{(l)}(m)$, $l = 1,\ldots,L$, $r = 1,\ldots,s^2$, according to (39) and (41), respectively;

step 3 : determine $C_r(m;\boldsymbol{0},g)$, $g = 1,\ldots,s$, according to (26), (37), and determine $b_r(m;\boldsymbol{0},h,-g,\xi)$, $g,h = 1,\ldots,s$, $\xi = 1,\ldots,\Psi_{g,h}^0$, according to (28), $r = 0,1,\ldots,s^2$; update $f^{(l)}(m)$ and $f_r^{(l)}(m)$, $l = 1,\ldots,L$, $r = 1,\ldots,s^2$, according to (39) and (41), respectively;

step 4 : $m := m + 1$; if $m \leq M$ then return to *step 2*; otherwise STOP.

The mean queue lengths (including the customer in service) can be obtained by taking $g^{(l)}(\boldsymbol{n},h,-g,\xi) := n_l$, $l = 1,\ldots,s$. The mean waiting times can then be obtained by Little's formula.

The reader is referred to [5] for a detailed discussion of practical aspects of the implementation of derivatives into the PSA.

The PSA is most effective when the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$ can be computed recursively. Therefore, Coxian distributed are generally preferred to general phase-type distributions for the service times and switch-over times. It should be noted that the class of Coxian distributions lies dense in the class of general probability distributions with non-negative support (cf. [1]).

The present model is contained in the class of so-called quasi birth-and-death (QBD) processes, in which state transitions from state $(\boldsymbol{n},\varphi) := (\boldsymbol{n},h,-g,\xi) \in \mathbb{N}^s \times \mathcal{S}$ can only occur to states of the form $(\boldsymbol{n},\psi)$ or $(\boldsymbol{n}+\boldsymbol{e}_j,\psi)$ or $(\boldsymbol{n}-\boldsymbol{e}_j,\psi) \in \mathbb{N}^s \times \mathcal{S}$). We refer to Blanc [3] for an extensive discussion of the use of the PSA for models with a general QBD structure and to Blanc and Van der Mei [5] for the extension of this general model to the computation of derivatives. In this general setting, for *each* triple $(r,k;\boldsymbol{n})$ a set of $|\mathcal{S}|$ linear equations has to be solved to determine the coefficients $b_r(k;\boldsymbol{n},\varphi)$, $\varphi \in \mathcal{S}$.
However, by introducing the quantities $C_r(k;\boldsymbol{n},g)$ in (25), for a given triple $(r,k;\boldsymbol{n})$ the set of $|\mathcal{S}|$ linear equations for computing the coefficients $b_r(k;\boldsymbol{n},\varphi)$, $\varphi = (h,-g,\xi) \in \mathcal{S}$, has been *reduced* to the set of only $|\mathcal{N}_{\boldsymbol{n}}^{(0)}|$ linear equations for the quantities $C_r(k;\boldsymbol{n},g)$, $g \in \mathcal{N}_{\boldsymbol{n}}^{(0)}$ (cf. (26)). Apparently, the specific structure of the presented model has been explored to obtain a more effective implementation of the PSA.

To characterize whether or not, for given $(r,k;\boldsymbol{n})$, the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$, $h = 1,\ldots,s$, $g=-1,1,\ldots,s$, $\xi = 1,\ldots,\Psi^0_{g,h}$, are fully recursively solvable, let us reconsider the set of equations (22). The states $(r,k;\boldsymbol{n},h,-g,\xi)$ for which the first term at the right-hand side does not vanish are exactly those for which the server has just skipped Q_g which was empty (i.e. $n_g = 0$) upon arrival of the server at that queue. Thus, *for given* $(r,k;\boldsymbol{n})$, the set of states $(r,k;\boldsymbol{n},h,-g,\xi)$ which can not be completely ordered are those in which the server is switching between the empty queues. Hence, as long as no arrivals occur at one of the empty queues, the server can keep on switching between these empty queues, provided the routing probabilities corresponding to these switches are strictly positive. For the states $(r,k;\boldsymbol{n},h,-g,\xi)$ with $n_g = 0$, the indicator function in (22) does not vanish, so that the coefficients $b_r(k;\boldsymbol{n},h,-g,\xi)$ can not be solved recursively according to (22).
This also explains why in the special case of cyclic server routing the states can be computed fully recursively for $\boldsymbol{n} \neq \boldsymbol{0}$. To see this, for $\boldsymbol{n} \neq \boldsymbol{0}$, there exists some index i such that $n_i > 0$, so that the server can not skip Q_i and hence, can not be moving around as long as no arrivals occur. From (22) it follows that under cyclic polling with $p_{g,g+1} = 1$, $g = 1,\ldots,s$, the coefficients $b_r(k;\boldsymbol{n},g+1,-g,\xi)$ can be determined recursively in the order $i,i+1,\ldots,s,1,\ldots,i-1$ with respect to g.

For the case in which some or all switch-over times are 0 a.s., some straightforward modifications of the balance equations and of the computational scheme have to be made.

It is not easy to give bounds for the accuracy of the computations with the PSA. However, for the present polling model with Markovian server routing a rough indication of the accuracy can be obtained from the PCL, i.e. an exact expression for a specific weighted sum of the mean waiting times at the queues. The accuracy of the computations with the PSA can be rougly estimated by computing this specific weighted sum on the basis of the computed mean waiting times and comparing this value to the exact value of the right-hand side of the PCL. For polling models with Markovian server routing with mixtures of 1-limited ($q_i = 0$) and exhaustive ($q_i = 1$) (and gated) service a PCL has been derived in [8]. This PCL can be readily extended to Markovian polling models with Bernoulli service with general parameters q_i ($0 \leq q_i \leq 1$) at Q_i, $i = 1,\ldots,s$, leading to the following expression:

$$\begin{aligned}\sum_{i=1}^{s} \rho_i \mathrm{E}W_i &= \frac{\rho^2}{2(1-\rho)} \sum_{i=1}^{s} a_i \beta_i^{(2)} + \frac{\rho}{2\sigma} \sum_{i=1}^{s} \omega_i \sum_{j=1}^{s} p_{i,j} \sigma_{i,j}^{(2)} \\ &\quad + \frac{1}{\sigma} \sum_{i=1}^{s} \omega_i \sum_{j=1}^{s} p_{i,j} \sigma_{i,j}^{(1)} \sum_{k \neq i} \rho_k \mathrm{E}T_{k,i} + \sum_{i=1}^{s} \mathrm{E}M_i, \end{aligned} \tag{42}$$

where $T_{i,j}$ is defined as the time elapsed between a departure of the server from Q_j and its last previous departure from Q_i, $i,j = 1,\ldots,s$, and where M_i stands for the amount of work at Q_i at a departure epoch of the server from Q_i. One may verify that $\mathrm{E}M_i$ is related to $\mathrm{E}W_i$ by the following relation (cf. [27]): for $i = 1,\ldots,s$,

$$\mathrm{E}M_i = (1-q_i)\left[\rho_i \lambda_i \frac{1}{\omega_i} \frac{\sigma}{1-\rho} \mathrm{E}W_i + \rho_i^2 \frac{1}{\omega_i} \frac{\sigma}{1-\rho}\right], \tag{43}$$

so that the PCL for the present model with Bernoulli service disciplines reads as follows:

$$\sum_{i=1}^{s} \rho_i \left[1 - \frac{a_i(1-q_i)}{\omega_i} \frac{\rho\sigma}{1-\rho}\right] EW_i = \frac{\rho^2}{2(1-\rho)} \sum_{i=1}^{s} a_i \beta_i^{(2)} + \frac{\rho}{2\sigma} \sum_{i=1}^{s} \omega_i \sum_{j=1}^{s} p_{i,j} \sigma_{i,j}^{(2)} + \frac{1}{\sigma} \sum_{i=1}^{s} \omega_i \sum_{j=1}^{s} p_{i,j} \sigma_{i,j}^{(1)} \sum_{k \neq i} \rho_k ET_{k,i} + \frac{\sigma}{1-\rho} \sum_{i=1}^{s} \frac{\rho_i^2(1-q_i)}{\omega_i}. \quad (44)$$

The unknown quantities $ET_{k,i}$ can be obtained by solving a set of linear equations. Thus, the PCL for Markovian polling is *not* a closed-form expression, and can only be evaluated by solving a set of linear equations, as opposed to the case of cyclic polling, in which the PCL gives a closed-form expression for a weighted sum of the mean waiting times at the queues. Still, the PCL (44) is very useful for getting an indication of the accuracy of the calculations with the PSA.

4 Markovian versus periodic polling

In this section we make a *comparison* between the performance of polling models under periodic and probabilistic server routing. To this end, we have implemented the PSA for both polling models with Markovian server routing and for polling models with periodic server routing, along the lines discussed in the previous section in Blanc [3], respectively.

Due to the considerable number of degrees of freedom in specifying the relative arrival rates, the service times, the system load, the switch-over times, the service disciplines and the visit orders, we have restricted ourselves to the analysis of a number of specific models, which we believe cover the main characteristics of the variety of models. Moreover, because of the computational complexity of the PSA we have restriced ourselves to models with a rather small number of queues. The characteristics observed for these models contribute to the understanding of the behavior of polling systems. We believe that these insights are also useful for understanding the behavior of models with a large number of queues. The following models have been taken under consideration, covering fully symmetrical models, models with asymmetrical arrival rates and models with asymmetrical switch-over times. For each of these models the offered load has been varied to cover models under light and heavy traffic, and the service discipline has been varied to cover 1-limited and exhaustive service, and moreover, the asymmetry in the arrival rates and switch-over times has been varied to cover a fairly broad class of models.

Model I represents *symmetrical* models, and is specified by the following set of system parameters: $s = 3$; $\beta^{(1)}$=(1.00,1.00,1.00); $\sigma_{i,j}^{(1)}$=0.05, $i,j = 1,\ldots,s$; all service times and switch-over times are exponentially distributed; a=(1.00,1.00,1.00); $q = (q,q,q)$. The quantities ρ and q are still variable.

Model II represents models in which the *arrival rates* are *asymmetrical*, and is specified by the following set of parameters: $s = 3$; $\beta^{(1)}$=(1.00,1.00,1.00); $\sigma_{i,j}^{(1)}$=0.05, $i,j = 1,\ldots,s$; all service times and switch-over times are exponentially distributed; $q = (q,q,q)$; the relative arrival rates are given by $a=(\alpha/(\alpha+2), 1/(\alpha+2), 1/(\alpha+2))$, so that the ratios between the arrival rates are α:1:1. The quantities ρ, α and q are still variable.

Model III represents models in which the *switch-over times* between the queues are *asymmetrical*, and the system parameters are: $s = 3$; $\beta^{(1)}$=(1.00,1.00,1.00); all service times and switch-over times are exponentially distributed; α=(1/3,1/3,1/3); $q = (q,q,q)$; the mean switch-over times are given by $\sigma^{(1)}_{1,1}=\sigma^{(1)}_{2,2}=\sigma^{(1)}_{3,3}=0.005$; $\sigma^{(1)}_{2,3}=\sigma^{(1)}_{3,2}=0.25$; $\sigma^{(1)}_{1,2}=\sigma^{(1)}_{2,1}=\sigma^{(1)}_{1,3}=\sigma^{(1)}_{3,1}=\alpha$, so that α (for values $\alpha \geq 0.125$) can be viewed as the mean 'distance' between Q_1 on the one hand and Q_2 and Q_3 on the other hand. The quantities ρ, α and q are still variable.

To make a reasonable comparison between the performance of Markovian and periodic polling models, we associate with each periodic service order table $\pi = (\pi(1),\ldots,\pi(L))$ a unique Markovian counterpart in which the matrix of routing probabilities $\mathbf{P} = (p_{i,j})$ is defined by: for $i,j = 1,\ldots,s$,

$$p_{i,j} := \frac{\sum_{k=1}^{L} I\left\{\pi_k = i;\ \pi_{(k \bmod L)+1} = j\right\}}{\sum_{k=1}^{L} I\{\pi_k = i\}}, \tag{45}$$

i.e. the fraction of times the server moves to Q_j after a departure from Q_i under polling table π. For instance, if the periodic service order table is given by π =(1,2,1,3), then the probabilistic version has routing probabilities $p_{1,2} = p_{1,3} = 0.50$, $p_{2,1} = p_{3,1} = 1.00$. Throughout, the Markovian polling model that is related to a periodic polling model through (45) is referred to as the *Markovian counterpart* of the periodic polling model.

In the remainder of this section we show some of the numerical results that we have gathered to compare the performance of polling systems in which the service order is guided by a polling table π and their Markovian counterpart. For given routing matrix $\mathbf{P}$, the performance measure considered here is

$$C(\mathbf{P}) := \sum_{i=1}^{s} \rho_i E W_i, \tag{46}$$

i.e. the mean total amount of waiting work in the system.
In the numerical examples considered here, the offered load to the system is either $\rho = 0.3$ (representing lightly-loaded models) or $\rho = 0.8$ (representing heavily-loaded models). The number of terms of the power series that has been computed is equal to $M = 40$, and the estimated error in the computations is typically less than 0.001.

Let us first consider symmetrical models under symmetrical visit orders, i.e. in which the routing is statistically the same for all queues. To this end, we have computed the system performance (46) for model I under a number of symmetrical routing orders with periodic polling (indicated by *I'*) and with their Markovian counterparts (indicated by *M*). Table 1 below shows the results for $q = 0.00$ (1-limited service) and $q = 1.00$ (exhaustive service), and for $\rho = 0.3$ and $\rho = 0.8$.
Table 1 suggests that in symmetrical models the mean total amount of waiting work in the system and hence, the mean waiting times (which are the same for all queues), are smaller in the case of periodic polling than under the corresponding Markovian server routing in all considered cases. However, one may also observe that the differences are rather small.
A comparison of the system performance for the various routing orders considered here shows

		$\rho = 0.3$		$\rho = 0.8$	
routing	P/M	$q = 0.00$	$q = 1.00$	$q = 0.00$	$q = 1.00$
123	P, M	0.18	0.17	4.50	3.44
112233	P	0.19	0.18	4.55	3.62
	M	0.20	0.19	4.75	3.64
111222333	P	0.20	0.20	4.63	3.80
	M	0.22	0.21	5.00	3.84
111122223333	P	0.22	0.21	4.74	3.99
	M	0.24	0.23	5.25	4.04

Table 1: Performance under symmetrical routing mechanisms; Model I.

that the performance of the system is closely related to the *spacing* of the visits. When the visits are 'better spaced' in time, the system performance seems to be improved and vice versa. This observation is supported by the following intuitive arguments. Under a periodic visit order the visits seem to be more homogeneously spaced than under the corresponding Markovian visit order. As a consequence, the cycle times C_i of Q_i, defined as the time interval between two successive departures of the server from Q_i, seem to be more 'regular' under periodic polling than under Markovian server routing. Under Bernoulli service EW_i (approximately) relates to the first two moments of C_i according to the relation (cf. [27]): for $i = 1, \ldots, s$,

$$EW_i \approx \frac{(1 - \rho + \rho_i) - q_i\rho_i(2 - \rho)}{1 - \rho[1 + \sigma a_i(1 - q_i)/\omega_i]} \frac{EC_i^2}{2EC_i}, \tag{47}$$

where $EC_i = \sigma/(\omega_i(1 - \rho))$, $i = 1, \ldots, s$, so that for a given set of system parameters and relative visit frequencies, the mean waiting time at Q_i increases with increasing 'irregularity' of the cycle times, represented by EC_i^2. These intuitive arguments support the observation in Table 4.1 that the system performance under periodic polling is better than under Markovian server routing.

To investigate whether a similar dominance relation also holds for symmetrical models under asymmetrical server routing, we have computed the mean waiting times for model I for a number of asymmetrical service orders, specified by $\pi = (1, 2, 1, 3)$, commonly referred to as star-polling, and $\pi = (1, 2, 3, 1, 3, 2)$. Table 2 shows the results for $q = 0.00$ and $q = 1.00$ and for $\rho = 0.3$ and $\rho = 0.8$.

Table 2 suggests that in the case of periodic polling the mean amount of work is still smaller than under Markovian polling. Yet, a similar stochastic dominance relation is *not* generally valid for the *individual* mean waiting times. This observation is supported by the following intuitive arguments. Let us reconsider the model with $\pi = (1, 2, 3, 1, 3, 2)$ with $\rho = 0.8$ in Table 2. In that case the polling order suggests that the visits to Q_1 are more homogeneously spaced than the visits to Q_2 and Q_3. Accordingly, EW_1 can be expected to be smaller than EW_2 and EW_3, which indeed turns out to be the case. Moreover, one may observe that the stochastic counterpart of this model, having routing probabilities $p_{1,2} = p_{1,3} = p_{2,1} = p_{2,3} = p_{3,1} = p_{3,2} = 0.50$, is symmetric, leading to the same mean waiting times at the queues. One would expect EW_1 to be smaller under periodic polling here, because the visits to Q_1

			$\rho = 0.3$		$\rho = 0.8$	
routing	q	P/M	(EW_1, EW_2, EW_3)	$C(\mathbf{P})$	(EW_1, EW_2, EW_3)	$C(\mathbf{P})$
1213	0.00	P	(0.48,0.66,0.66)	0.18	(2.35,7.92,7.92)	4.85
	0.00	M	(0.48,0.74,0.74)	0.20	(2.31,8.29,8.29)	5.04
	1.00	P	(0.48,0.60,0.60)	0.17	(3.03,4.98,4.98)	3.46
	1.00	M	(0.48,0.67,0.67)	0.18	(3.09,5.20,5.20)	3.60
123132	0.00	P	(0.58,0.60,0.60)	0.18	(5.60,5.66,5.66)	4.51
	0.00	M	(0.61,0.61,0.61)	0.18	(5.73,5.73,5.73)	4.58
	1.00	P	(0.54,0.57,0.57)	0.17	(3.92,4.55,4.55)	3.47
	1.00	M	(0.57,0.57,0.57)	0.17	(4.38,4.38,4.38)	3.50

Table 2: Performance under asymmetrical routing mechanisms; Model I.

seem to be better spaced than under Markovian polling, in which the uncertainty leads to less well-spaced visits to Q_1. As for the mean waiting times at Q_2 and Q_3, there is a trade-off between the irregularity in the cycle times caused by the use of probabilistic polling on the one hand, and the irregularity of the cycle times caused by a rather bad spacing of the visits under periodic polling. Apparently, the former irregularity is *dominated* by the latter one. This intuitive argument supports the observation that in this example EW_2 and EW_3 are larger under periodic polling than under the corresponding Markovian polling mechanism.

To investigate whether the observations made for symmetrical models also persist for asymmetrical models, we consider the performance of the system for both periodic and Markovian service order for a model with varying asymmetry in the arrival rates. We have computed the mean waiting times in model II for a number of values of relative arrival rates. Table 3 shows the results where the ratios between the arrival rates are 1:10:10 and 10:1:1, for $q = 1.00$ and $\rho = 0.3$ and $\rho = 0.8$.

			$\rho = 0.3$		$\rho = 0.8$	
routing	ratios	P/M	(EW_1, EW_2, EW_3)	$C(\mathbf{P})$	(EW_1, EW_2, EW_3)	$C(\mathbf{P})$
1213	1:10:10	P	(0.51,0.58,0.58)	0.17	(4.02,4.35,4.35)	3.39
	1:10:10	M	(0.52,0.65,0.65)	0.18	(4.14,4.59,4.59)	3.55
	10:1:1	P	(0.48,0.72,0.72)	0.19	(2.85,10.78,10.78)	6.51
	10:1:1	M	(0.48,0.79,0.79)	0.21	(2.86,11.00,11.00)	6.63
123132	1:10:10	P	(0.59,0.56,0.56)	0.17	(4.66,4.37,4.37)	3.57
	1:10:10	M	(0.66,0.56,0.56)	0.18	(7.59,4.17,4.17)	4.25
	10:1:1	P	(0.51,0.68,0.68)	0.19	(3.14,9.43,9.43)	5.87
	10:1:1	M	(0.53,0.70,0.70)	0.19	(3.08,10.01,10.01)	6.16

Table 3: Mean waiting times for asymmetrical routing mechanisms; Model II.

The results in Table 3 confirm the observation that the mean amount of work in the system is smaller for periodic polling in all considered cases, but that in a number of cases some of the individual mean waiting times are smaller under Markovian polling.

5 Optimization

In this section we consider the following optimization problem:

$$\min_{\mathbf{P}\in\mathcal{M}_s} C(\mathbf{P}), \tag{48}$$

where $C(\mathbf{P})$ is defined in (46) and $\mathcal{M}_s$ is defined as the set of irreducible stochastic $s \times s$ matrices. In words, the problem is to find combinations of routing probabilities which minimize the mean amount of waiting work in the system. Optimal routing matrices are denoted by $\mathbf{P}^*$.
In a general parameter setting no explicit expressions for the cost function are available and hence, the optimization problem is not exactly solvable.

Boxma et al. [7] consider a similar problem of heuristically obtaining periodic polling tables which minimize the mean amount of work in the system. They propose to combine explicit square-root formulas for optimal relative visit frequencies in random polling models with the Golden Ratio procedure (cf. [14]) for the spacing of the visits. However, this approach relies on the assumption that the switch-over times depend *only* on the queue which is being switched *to*, and is independent of the queue that has just been visited. This assumption is quite restrictive, e.g. when switch-over times represent physical movement from one place to another. Yet, when this assumption is dropped, the problem of finding an optimal visit order for *given* relative visit frequencies can be formulated as a Travelling Salesman Problem (TSP), which is known to be NP-hard.

The optimization problem (48) can, in principle, be solved numerically by combining a numerical algorithm for the evaluation of the cost function (46) with some standard procedure for non-linear (constrained) optimization. However, the dimension of the optimization problem grows quadratically in the number of queues, so that in practice this approach is restricted to models with a rather small number of queues. It should be noted that in the special case in which all queues are served exhaustively, the cost function (46) can be directly obtained via the PCL (44), requiring the solution of a relatively small set of linear equations. However, in case at least one queue is served non-exhaustively, the PCL is no longer applicable to evaluate the cost function (46). Moreover, the PCL can not be used to determine more detailed performance measures like the individual mean waiting times at the queues. In those situations, the computations may be based on the use of the PSA, requiring considerably more computational effort. To find optimal routing matrices, we have computed the cost function (46), plus its derivatives with respect to the routing probabilities, in combination with the conjugate gradient method for non-linear optimization with linear constraints (cf. [22]).
We reemphasize the enormous complexity of the TSP-like optimization problem in a general parameter setting. Therefore, we restrict ourselves to obtain some qualitative, instead of quantitative, properties of optimal combinations of routing probabilities. The results presented here should be viewed in this perspective.

The remainder of this section is organized as follows. In section 5.1 we discuss properties of optimal routing matrices in the case of fully symmetrical models, and in section 5.2 we investigate optimal combinations of routing probabilities in the case of some asymmetrical models.

5.1 Symmetrical systems

For fully symmetrical models it is shown in [20] that each cyclic service order, which is contained in the class of Markovian service orders, solves the optimization problem (48). Thus, for such models with (symmetrical) Bernoulli schedule $\boldsymbol{q} = (q, \ldots, q)$, $0 \leq q \leq 1$,

$$\mathbf{P}^* = \hat{\mathbf{P}}, \tag{49}$$

where $\hat{\mathbf{P}} = (\hat{p}_{i,j}) \in \mathcal{M}_s$ with $\hat{p}_{i,j} \in \{0,1\}$, $i,j = 1,\ldots,s$. Note that there are $(s-1)!$ alternative *local optima*, each of which uniquely corresponds to a specific cyclic visit order.

Let us consider the question whether these optima are stable, i.e. whether the optimal cyclic server routing orders remain optimal when the system parameters are slightly perturbed. To this end, it should be noted that the derivatives of the cost function (48) with respect to each of the routing probabilities may provide useful information about the character of the optimal routing probabilities. Namely, when all derivatives are equal to 0 in the optimum, the optimal schedule will be an 'interior optimum' which may become suboptimal for slight changes in one of the system parameters. On the other hand, 'boundary optima' with non-zero derivatives at the optimum remain (locally) optimal for slight modifications of the parameters. To study the character of the optima, we have applied the PSA to compute the cost function (46) and the derivatives with respect to the routing probabilities for a set of symmetrical models, each of which giving similar outcomes. For a typical example, consider Model I (introduced in section 4). Table 4 shows the derivatives of the cost function (48) with respect to the routing probabilities at $\mathbf{P}^* = \left(p^*_{i,j}\right)$, with $p^*_{1,2} = p^*_{2,3} = p^*_{3,1} = 1.00$, for $\rho = 0.3$ and 0.8 and for $q = 0.00$, 0.50 and 1.00.

	q=0.0	q=0.5	q=1.0
$\rho = 0.3$	0.007 0.000 0.007	0.007 0.000 0.007	0.007 0.000 0.007
	0.007 0.007 0.000	0.007 0.007 0.000	0.007 0.007 0.000
	0.000 0.007 0.007	0.000 0.007 0.007	0.000 0.007 0.007
$\rho = 0.8$	0.030 0.000 0.030	0.028 0.000 0.028	0.025 0.000 0.025
	0.030 0.030 0.000	0.028 0.028 0.000	0.025 0.025 0.000
	0.000 0.030 0.030	0.000 0.028 0.028	0.000 0.025 0.025

Table 4: Derivatives $\partial C(\mathbf{P})/\partial p_{i,j}$ at $\mathbf{P} = \mathbf{P}^*$ in a symmetrical model.

Table 4 suggests that the optimal cyclic visit orders are *stable optima*. To illustrate this, consider the case q=0.5, $\rho = 0.8$, and consider the routing probabilities after departing from Q_1, which are under the present cyclic schedule equal to $p^*_{1,2} = 1.00$, $p^*_{1,1} = p^*_{1,3} = 0.00$. To obtain an alternative triple of routing probabilities, either $p_{1,1}$ or $p_{1,3}$, or both, must be increased, and $p_{1,2}$ will have to be decreased. Now, because the derivatives of the cost function with respect to $p_{1,1}, p_{1,2}$ and $p_{1,3}$ at $\mathbf{P}^*$ are given by 0.028, 0.000 and 0.028, respectively, a (small) increase in either $p_{1,1}$ or $p_{1,3}$ together with a (small) decrease of $p_{1,2}$ will lead to an increase of the cost function. Under the assumption that the derivatives of the cost function with respect to continuous system parameters at the cyclic optimum are continuous, slight modifications of these system parameters do not cause the cyclic optimum to become suboptimal. The fact that in Table 4 the derivatives with respect to $p_{1,2}$, $p_{2,3}$ and $p_{3,1}$ are 0.000 is due to the

definition of the derivatives of the routing probabilities in (13)-(15), which even implies that these derivatives are *exactly* equal to zero.

The observation that the cyclic optimum is stable suggests the existence of a *attraction region* 'around' the cyclic optimum for nearly-symmetrical models. That is, there is a set of 'nearly-symmetrical' models for which the optimal Markovian server routing is cyclic. In the next subsection we will present some numerical experiments which support this observation. Yet, exact expressions for this region are unknown, so that the observation is only useful for providing a qualitative, rather than a quantitative, insight into optimal combinations of routing probabilities.

5.2 Asymmetrical systems

When the model is asymmetrical, it is clear that the cyclic server routing is no longer generally optimal within the class of Markovian server routings. However, for asymmetrical models the optimization problem is not exactly solvable, and numerical procedures are needed to find optimal combinations of routing probabilities. In a number of examples discussed here, it is assumed that all queues are served exhaustively. This assumption is based on results obtained by Liu et al. [20], who have shown that in order to minimize the cost function (46), all queues should be served exhaustively. Recall that in those cases the cost function (46), and the optimal routing matrices, can be obtained relatively quickly via (44). Yet, in some applications exhaustive service may not be implemented or technically infeasible. In those situations, the cost function (46), and the optimal routing matrices, have been obtained on the basis of the PSA.

As for the accuracy of the computations, in the case of exhaustive service at all queues the cost function (46) has been accurately calculated from the PCL (44), with typical errors less than 10^{-12}. In the cases with non-exhaustive service disciplines the cost function has been evaluated by means of the PSA, where typically 40 or 50 terms of the power series have been computed and the estimated errors are typically less than 0.001. The optimization procedure is based on a grid size 0.001.

Numerical experience has taught us that the cost function (46) as function of the routing matrices generally has a number of *local optima*, similar to the case of fully symmetrical models. This difficulty, which is very common in non-linear optimization, is tackled by running the optimization procedure with a number of different initial routing matrices.

To study characteristics of optimal routing matrices for a broad class of Markovian polling models, we have computed the optimal routing probabilities for a wide variety of the parameter settings for models II and III (model I occurring as a special case), covering a diversity of models. In this section we present some of the numerical results. We emphasize that this numerical study does not aim to give a full characterization of optimal schedules, but is meant to give some useful insights, which contribute to the understanding of the characteristics of optimal routing matrices.

Influence of asymmetry in the arrival rates
To investigate the influence of the asymmetry in the arrival process on $\mathbf{P}^*$, we have computed

an optimal routing matrix for model II (cf. section 4) for $q = 1.00$ and in which the ratios between the arrival rates are given by α:1:1. Tables 5 and 6 show an optimal routing matrix $\mathbf{P}^*$ for various values of α for the system under light traffic ($\rho = 0.3$) and heavy traffic ($\rho = 0.8$), respectively. It should be noted that because in this model Q_2 and Q_3 are stochastically identical, the corresponding routing probabilities are exchangeable.

	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.50$
$\mathbf{P}^*$	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00
	0.04 0.00 0.96	0.14 0.00 0.86	0.48 0.00 0.52	0.81 0.00 0.19	1.00 0.00 0.00
	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00
$C(\mathbf{P}^*)$	0.155	0.157	0.162	0.164	0.165

	$\alpha = 1.00$	$\alpha = 4.00$	$\alpha = 10.00$	$\alpha = 100.00$	$\alpha = 1000.00$
$\mathbf{P}^*$	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.44 0.56	0.00 0.50 0.50	0.92 0.04 0.04
	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00
	0.00 1.00 0.00	0.00 1.00 0.00	0.80 0.20 0.00	1.00 0.00 0.00	1.00 0.00 0.00
$C(\mathbf{P}^*)$	0.165	0.163	0.160	0.152	0.146

Table 5: Optimal routing probabilities for model II; $\rho = 0.3$.

	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.50$
$\mathbf{P}^*$	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00
	0.05 0.00 0.95	0.17 0.00 0.83	0.58 0.00 0.42	0.94 0.00 0.06	1.00 0.00 0.00
	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00
$C(\mathbf{P}^*)$	3.347	3.363	3.402	3.422	3.434

	$\alpha = 1.00$	$\alpha = 10.00$	$\alpha = 25.00$	$\alpha = 50.00$	$\alpha = 100.00$
$\mathbf{P}^*$	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.40 0.60	0.00 0.50 0.50	0.22 0.39 0.39
	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00
	0.00 1.00 0.00	0.00 1.00 0.00	0.66 0.34 0.00	1.00 0.00 0.00	1.00 0.00 0.00
$C(\mathbf{P}^*)$	3.440	3.350	3.311	3.287	3.274

Table 6: Optimal routing probabilities for model II; $\rho = 0.8$.

The results displayed in Tables 5 and 6 reveal some characteristics of optimal routing matrices. First, we observe the surprisingly large fraction of the routing probabilities that are equal to 0.00 or 1.00, indicating that the *optimal routing decisions have a tendency towards deterministic routing*. This observation is supported by the observation in section 4 that the cycle times under deterministic routing are more regular than under probabilistic routing decisions, generally leading to a better system performance (cf. (47)).
Second, for lightly- and heavily-loaded systems the results suggest that the optimal matrices for different values of α can be divided into a small number of classes, each of which has specific characteristics that can be interpreted easily, providing an insight into the behavior

of optimal routing matrices. Each class corresponds to a specific interval of values of α. We will now discuss characteristics of these classes, letting α increase from 0 to infinity.
Let us first consider the limiting case $\alpha \downarrow 0$. In that case the arrival rate at Q_1 is negligible compared with the arrival rates at Q_2 and Q_3, so that in the optimum Q_1 will be visited only very seldom. The results in Tables 5 and 6 suggest that $\mathbf{P}^*$ tends to a limiting routing matrix with $p^*_{1,3} = p^*_{2,3} = p^*_{3,2} = 1.00$. Under this routing matrix, state 1 (corresponding to visits to Q_1) is only a transient state, and the states 2 and 3 form an absorbing set of states. When α is somewhat increased, starting from 0.00, Q_1 will be visited more frequently, but still less frequently than Q_2 and Q_3. This situation appears to give optimal routing matrices of the form $p^*_{1,3} = p^*_{3,2} = 1.00$, $p^*_{2,1} = r$, and $p^*_{2,3} = 1 - r$, $0 \leq r < 1$, where the value of r increases with increasing value of α. That is, after a departure from Q_1 the server always moves to Q_3 and subsequently, to Q_2. The only random routing decisions are made after departures from Q_2. So, the server visits the queues in cyclic order, typically interceded by a number of switches back and forth between Q_2 and Q_3. When α is further increased to approach 1.00 the arrival rates become of the same order of magnitude. For $\alpha = 1.00$ the system is symmetric, and it is known that in that case the optimal routing is cyclic (cf. [20]). Moreover, Tables 5 and 6 suggest that the *cyclic* routing is still optimal when α is varied within some interval around $\alpha = 1.00$. This observation supports the conjecture that there is some region 'around' the cyclic optimum $\mathbf{P}^*$ in which $\mathbf{P}^*$ is still optimal (cf. section 5.1). When the value of α is further increased, Q_1 becomes considerably more heavily loaded than the other queues, so that above some threshold value for α the cyclic visit order is no longer optimal. We then typically observe optimal routing matrices $\mathbf{P}^*$ of the form $p^*_{2,1} = 1.00$, $p^*_{1,2} = r_1$, $p^*_{1,3} = 1 - r_1$, $p^*_{3,1} = r_2$, $p^*_{3,2} = 1 - r_2$ $(0 < r_1, r_2 < 1)$, and zeros elsewhere. Under this type of service orders, Q_1 is implicitly given higher priority than the other queues. This is because after most visits to either Q_2 or Q_3 the next queue to be served is Q_1. The only exception is when Q_2 is visited after a visit to Q_3. In those cases Q_1 will be immediately visited afterwards. This type of routing matrix may be seen as an *intermediate* between the cyclic server routing (for smaller values of α) and another type of polling order which occurs when α is increased further. In the latter case Q_1 dominates the system in such a strong way that Q_1 is always visited immediately after a visit to one of the other queues, so that the optimum $\mathbf{P}^*$ is typically of the form $p^*_{1,2} = p^*_{1,3} = 0.50$, $p^*_{2,1} = p^*_{3,1} = 1.00$. This type of routing matrix may be seen as a stochastic counterpart of the periodic *star-type* polling. Finally, when α is increased even further Q_1 becomes so relatively heavily loaded that switches from Q_1 to itself (throughout referred to as *self transitions*) become optimal, while Q_1 is always visited immediately after a visit to one of the other queues. When α approaches infinity, the optimal routing matrix tends to the routing matrix $\mathbf{P}^*$ with $p^*_{1,1} = p^*_{2,1} = p^*_{3,1} = 1.00$ and zeros elsewhere.

Influence of the asymmetry in the switch-over times
To investigate the influence of the switch-over times on the optimal routing matrix, we have computed optimal probabilities for a variety of models which are contained in the class described in model III (cf. section 4). For these models the ratios between the arrival rates are equal, and mean switch-over times are given by $\sigma^{(1)}_{1,1} = \sigma^{(1)}_{2,2} = \sigma^{(1)}_{3,3} = 0.005$; $\sigma^{(1)}_{1,2} = \sigma^{(1)}_{2,1} = \sigma^{(1)}_{1,3} = \sigma^{(1)}_{3,1} = \alpha$; $\sigma^{(1)}_{2,3} = \sigma^{(1)}_{3,2} = 0.25$. Note that, for $\alpha \geq 0.125$, the parameter α can basically be viewed as the mean 'distance' between Q_1 on the one hand and Q_2 and Q_3 on the other hand. Tables 7 and 8 show the results for various values of α, $q = 1$, and for ρ=0.3 and 0.8,

respectively.

	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.25$
$\mathbf{P}^*$	0.00 0.50 0.50	0.00 0.50 0.50	0.00 0.50 0.50	0.00 0.00 1.00
	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00
	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	0.00 1.00 0.00
$C(\mathbf{P}^*)$	0.130	0.139	0.236	0.311

	$\alpha = 0.50$	$\alpha = 1.00$	$\alpha = 2.50$	$\alpha = 10.00$
$\mathbf{P}^*$	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00
	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00
	0.95 0.05 0.00	0.54 0.46 0.00	0.25 0.75 0.00	0.07 0.93 0.00
$C(\mathbf{P}^*)$	0.437	0.686	1.412	5.019

Table 7: Optimal routing probabilities for model III; $\rho = 0.3$.

	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.25$
$\mathbf{P}^*$	0.00 0.50 0.50	0.00 0.50 0.50	0.00 0.00 1.00	0.00 0.00 1.00
	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00	1.00 0.00 0.00
	1.00 0.00 0.00	1.00 0.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00
$C(\mathbf{P}^*)$	3.208	3.280	3.933	4.400

	$\alpha = 0.50$	$\alpha = 1.00$	$\alpha = 2.50$	$\alpha = 10.00$
$\mathbf{P}^*$	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00	0.00 1.00 0.00
	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00
	1.00 0.00 0.00	0.70 0.30 0.00	0.41 0.59 0.00	0.19 0.81 0.00
$C(\mathbf{P}^*)$	5.213	6.805	11.231	32.129

Table 8: Optimal routing probabilities for model III; $\rho = 0.8$.

The results in Tables 7 and 8 reveal some properties of the character of optimal routing matrices. Similar to the case of varying the relative arrival rates discussed above, we observe again that the optimal routing decisions have a *tendency towards deterministic routing* (cf. the discussion of the results in Tables 5 and 6).

Moreover, Tables 7 and 8 indicate that the optimal routing matrices for varying values of α can be divided into a number of types of routing matrices. We will briefly discuss characteristics of each of these classes. When α is small the optimal routing matrix routes the server to Q_1 (with probability 1.00) after a visit to one of the other queues. In this way, the relatively long 'journey' between Q_2 and Q_3 is avoided, and Q_1 serves as a 'bridge' between these queues. When α approaches 0.25 the system becomes symmetrical and the cyclic visit order becomes optimal. Again it is observed that this cyclic optimum remains optimal for slight perturbations in the switch-over times. When α becomes considerably larger than the switch-over times between Q_2 and Q_3, Q_1 basically becomes relatively 'isolated' from Q_2 and

Q_3, or equivalently, Q_2 and Q_3 may be viewed as relatively 'clustered'. The optimal routing matrices $\mathbf{P}^*$ appear to have a specific structure of the form $p^*_{1,2} = p^*_{2,3} = 1.00$, and $p^*_{3,1} = r$, $p^*_{3,2} = 1 - r$ $(0 < r < 1)$, where r decreases with increasing α. This specific structure can be interpreted as follows. After having emptied Q_1 the server always moves towards Q_3, and after a visit to Q_3 the server moves to Q_2 with probability 1.00. Then the server keeps on alternating between Q_2 and Q_3 before making the relatively long trip to Q_1. The latter implies that in this way one avoids making two successive relatively long journeys without having visited both queues in the cluster of Q_2 and Q_3.

Influence of the service disciplines

In the cases considered so far it is assumed that the queues are served exhaustively. We will now study the influence of the service discipline on optimal routing matrices. To this end, consider the case $\alpha = 1.00$ for Model III (cf. also Tables 7 and 8). Recall that in this case Q_1 is relatively 'isolated' from Q_2 and Q_3. We study the influence of the service discipline at Q_1 on the optimal routing matrices. To this end, we have computed the optimal routing matrices for various Bernoulli service policies with parameter $q_1 = q$ $(0 \leq q \leq 1)$. The service discipline at Q_2 and Q_3 is assumed to be exhaustive. Tables 9 and 10 below show optimal routing matrices for q=0.00, 0.50 and 1.00, and for $\rho = 0.3$ and ρ=0.8, respectively.

	$q = 0.00$	$q = 0.50$	$q = 1.00$
$\mathbf{P}^*$	0.96 0.04 0.00	0.95 0.05 0.00	0.00 1.00 0.00
	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00
	0.56 0.44 0.00	0.55 0.45 0.00	0.54 0.46 0.00
$C(\mathbf{P}^*)$	0.696	0.692	0.686

Table 9: Optimal routing probabilities for different service disciplines; $\rho = 0.3$.

	$q = 0.00$	$q = 0.50$	$q = 1.00$
$\mathbf{P}^*$	0.97 0.03 0.00	0.95 0.05 0.00	0.00 1.00 0.00
	0.00 0.00 1.00	0.00 0.00 1.00	0.00 0.00 1.00
	0.77 0.23 0.00	0.76 0.24 0.00	0.70 0.30 0.00
$C(\mathbf{P}^*)$	7.301	7.163	6.805

Table 10: Optimal routing probabilities for different service disciplines; $\rho = 0.8$.

The results in Table 9 indicate that the service discipline may have a considerable impact on the optimal routing matrices. In particular, we observe a striking difference in the optimal routing probabilities between exhaustive service on the one hand and non-exhaustive service on the other hand. We observe that in the case of non-exhaustive service (i.e. $q < 1$) self transitions occur frequently here, whereas for exhaustive service similar self transitions occur with probability 0. To give an intuitive argument for this observation, recall that it is shown in [20] that all queues should be served exhaustively to minimize the cost function (46). Consider the case $q < 1$, i.e. Q_1 is served non-exhaustively, so that after a visit of the server at Q_1 there may be customers present at Q_1. Note that the switch-over times needed by the

server for a self transition (with mean 0.005) are negligible compared with the switch-over times between different queues (with means 0.25 or 1.00). Hence, the server can almost immediately return to Q_1 to check whether there is another customer waiting at that queue. If so, the next customer at Q_1 will be served, and if not so, the 'cost' of this 'unnecessary' travel from Q_1 to itself is very small. This argument intuitively explains why for non-exhaustive service at Q_1 self transitions occur with large probability (typically ≥ 0.9). In this way, Q_1 is served 'nearly exhaustively'. Obviously, in the case of exhaustive service, self transition from Q_1 would probably not make much sense, because no customers are present at a departure instant of the server at Q_1.

The above-mentioned considerations indicate that when self transitions can be made instantaneously (i.e. $\sigma_{i,i}^{(1)} = 0$, $i = 1, \ldots, s$), then under exhaustive service at Q_i, the cost function does not depend on $p_{i,i}$, provided $p_{i,i} < 1$. Moreover, for systems with $\sigma_{i,i}^{(1)} = 0$, $q_i = 0$, $i = 1, \ldots, s$, the service discipline at Q_i can basically be viewed as a *Bernoulli* service discipline with parameter $\tilde{q}_i = p_{i,i}$, $i = 1, \ldots, s$ (cf. also [8]). In this way, the problem of finding optimal Bernoulli parameters $\tilde{q} = (\tilde{q}_1, \ldots, \tilde{q}_s)$ in cyclic polling models occurs as a special case by putting the additional restriction $p_{i,i} + p_{i,i+1} = 1$ (cf. [6]). The only difference here is that $\tilde{q}_i$ should be strictly smaller than 1 (to guarantee the irreducibility of the Markov chain D, cf. section 2), while $\tilde{q}_i = 1$ is also allowed in the optimization problem discussed in [6].

Guidelines for constructing optimal routing matrices

Based on the results presented in Tables 5 to 10, we will now give some general ideas that may be useful for heuristically constructing routing matrices for larger systems. We reemphasize that these ideas only aim to give some insight into the qualitative, rather than the quantitative, behavior of optimal routing matrices, and should be viewed in that perspective.

Suppose the distance structure is such that the queues $Q_1, \ldots, Q_s$ can somehow be partitioned into a relatively small number of *clusters* of queues, $C_1, \ldots, C_m$, $m < s$, in such a way that the mean switch-over times between queues within the same cluster are considerably smaller than the distances between queues in different clusters. In this perspective, each of these clusters can be viewed as *super queues*. The numerical results in Tables 7 and 8 suggest that in each cluster C_k there is a 'front door' queue C_k^F such that the server can 'enter' cluster C_k only through a visit at queue C_k^F and not through a visit at another queue in C_k. This suggestion implies that for (nearly) optimal routing matrices we have $p_{i,j}^* = 0$ if $Q_i \notin C_k$, $Q_j \in C_k$ and $Q_j \neq C_k^F$, $k = 1, \ldots, m$. Similarly, each cluster C_k seems to have a 'back door' queue C_k^B such that the server can only depart from cluster C_k through C_k^B, $k = 1, \ldots, m$, i.e. $p_{i,j}^* = 0$ if $Q_i \in C_k$, $Q_j \notin C_k$ and $Q_i \neq C_k^B$, $k = 1, \ldots, m$. Moreover, one may expect that the optimal routing probabilities within each cluster C_k will be such that after the server has entered C_k (through an arrival at C_k^F) *all* queues within that cluster are certainly visited at least once during the visit of the server to that cluster. The problem of determining optimal routing probabilities between the different clusters (super queues), i.e. $p_{i,j}^*$, $Q_i = C_k^B$, $Q_j = C_k^F$, $k = 1, \ldots, m$, is roughly similar to the problem of determining optimal routing matrices for systems with $m < s$ (super) queues $\tilde{Q}_1, \ldots, \tilde{Q}_m$, where the parameters of super queue $\tilde{C}_k$ can be determined by *aggregating* over the parameters of the queues in C_k in a straightforward manner. This observation suggests a *hierarchical* procedure for obtaining optimal routing matrices for larger systems.

As an illustration of the validity of the above-mentioned guidelines, we consider the model with the following combination of system parameters: $s = 4$; $\boldsymbol{a} = (1.00, 1.00, 1.00, 1.00)$; $\boldsymbol{\beta}^{(1)} = (1.00, 1.00, 1.00, 1.00)$; all service times and switch-over times are exponentially distributed; $\boldsymbol{q} = (1.00, 1.00, 1.00, 1.00)$; $\sigma_{1,j}^{(1)} = \sigma_{j,1}^{(1)} = \alpha$, for $j = 2,3,4$; $\sigma_{i,j}^{(1)} = 0.05$ in all other cases. Note that for values of α large enough, the queues can be basically partitioned into clusters $C_1 = \{Q_1\}$ and $C_2 = \{Q_2, Q_3, Q_4\}$. Table 11 shows optimal routing matrices for α=0.10, 0.25 and 5.00, and for ρ=0.3, and Table 12 shows the results for $\rho = 0.8$.

	$\alpha = 0.10$	α=0.25	α=5.00
$\mathbf{P}^*$	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00
	0.00 0.00 1.00 0.00	0.00 0.00 1.00 0.00	0.00 0.00 1.00 0.00
	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00	0.00 0.32 0.00 0.68
	1.00 0.00 0.00 0.00	0.13 0.87 0.00 0.00	0.04 0.96 0.00 0.00
$C(\mathbf{P}^*)$	0.201	0.608	2.404

Table 11: Optimal routing probabilities; $\rho = 0.3$.

	$\alpha = 0.10$	α=0.25	$\alpha = 5.00$
$\mathbf{P}^*$	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00
	0.00 0.00 1.00 0.00	0.00 0.00 1.00 0.00	0.00 0.00 1.00 0.00
	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00
	1.00 0.00 0.00 0.00	0.24 0.76 0.00 0.00	0.10 0.90 0.00 0.00
$C(\mathbf{P}^*)$	3.713	6.222	16.347

Table 12: Optimal routing probabilities; $\rho = 0.8$.

The results in Tables 11 and 12 confirm the characteristics discussed above. Obviously, the queues can be clustered as $C_1 = \{Q_1\}$ and $C_2 = \{Q_2, Q_3, Q_4\}$. We observe that C_2 is only entered through Q_2 (i.e. $C_2^F = Q_2$) and is only departed from at Q_4 (i.e. $C_2^B = Q_4$). In all cases considered here the server moves to C_2 after departing from Q_1, and visits the queues in C_2 a number of times (geometrically distributed with parameter $1 - p_{1,4}^*$) before returning to Q_1. We also observe that once the server has entered C_2 through a visit at Q_2, all queues in C_2 are served at least once during that visit.

As an alternative, consider the model with the same system parameters as the above-discussed model, but with the following mean switch-over times: $\sigma_{i,i}^{(1)} = 0.05$, $i = 1, \ldots, 4$; $\sigma_{i,j}^{(1)}$=1.00 if $i, j \in \{1,2\}$ or $i, j \in \{3,4\}$; $\sigma_{i,j}^{(1)} = \alpha$ in all other cases. Note that for $\alpha > 1.00$, the queues can basically be clustered into clusters $C_1 = \{Q_1, Q_2\}$ and $C_2 = \{Q_3, Q_4\}$. Table 13 shows the optimal routing matrices for α=1.00, 10.00 and 50.00 for $\rho = 0.3$, and Table 14 shows the results for $\rho = 0.8$.
Tables 13 and 14 support the characteristics of the optimal routing matrices discussed in this section. In all cases considered here we have $C_1^F = Q_1$, $C_1^B = Q_2$, $C_2^F = Q_3$ and $C_2^B = Q_4$.

	$\alpha = 1.00$	$\alpha = 10.00$	$\alpha = 50.00$
$\mathbf{P}^*$	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00
	0.00 0.00 1.00 0.00	0.48 0.00 0.52 0.00	0.84 0.00 0.16 0.00
	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00
	1.00 0.00 0.00 0.00	0.52 0.00 0.48 0.00	0.16 0.00 0.84 0.00
$C(\mathbf{P}^*)$	1.071	5.753	26.144

Table 13: Optimal routing probabilities; $\rho = 0.3$.

	$\alpha = 1.00$	$\alpha = 10.00$	$\alpha = 50.00$
$\mathbf{P}^*$	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00	0.00 1.00 0.00 0.00
	0.00 0.00 1.00 0.00	0.36 0.00 0.64 0.00	0.72 0.00 0.28 0.00
	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00	0.00 0.00 0.00 1.00
	1.00 0.00 0.00 0.00	0.64 0.00 0.36 0.00	0.28 0.00 0.72 0.00
$C(\mathbf{P}^*)$	10.000	41.090	164.522

Table 14: Optimal routing probabilities; $\rho = 0.8$.

The server typically moves from one cluster to the other, interceded by a number of switches back and forth between the queues within the respective clusters.

6 Topics for further research

The guidelines for constructing optimal routing matrices in the previous section are based on the insights obtained by the numerical study presented in sections 4 and 5. However, the guidelines are qualitative, and do not yield a heuristic approach to obtain optimal routing matrices. The following idea may be worthwhile to consider for obtaining a quantitative heuristic approach. For a given planar distance structure between the queues (due to the switch-over times), there are various algorithms available for partitioning the set of queues ('points') into a number of subsets ('clusters') of queues (e.g. single-link clustering, complete-link clustering, furthest neighbor method, cf. e.g. [10]). Each of these algorithms provides a means to define a clustering structure depending on whether the planar distances between certain combinations of queues exceed some threshold value d. For a given clustering algorithm, one may build a *tree structure* of clusters by successively decreasing the threshold distance d, starting with $d = \infty$ (in which all queues form one cluster) until $d = 0$ (in which each forms a cluster by itself). Such a tree structure suggests an *iterative* approach for heuristically obtaining optimal routing matrices for large systems, with decreasing threshold value d.

By definition of 'iteration', at each step of the iteration at least one couple of clusters, say C_1 and C_2 is united to one cluster $C_{12} := C_1 \cup C_2$. In this way, one should construct (i) a simple heuristic approach to define the 'front door' C_{12}^F and the 'back door' C_{12}^B of cluster C_{12} (defined in section 5), and (ii) a simple heuristic to 'merge' the 'local' routing probabilities for C_1 and C_2 to routing probabilities for C_{12}. As for the first problem, one should probably select C_{12}^B either C_1^B or C_2^B, and a similar approach may be used to determine C_{12}^F. The

second problem may be handled by adopting the intra-cluster ('local') routing probabilities. As for the inter-cluster routing probabilities, one may set $p_{i,j} = 0$, $i \in C_1$, $j \in C_2$, *unless* $Q_i = C_1^B$ *and* $Q_j = C_2^F$. The routing probabilities between C_1^B, C_1^F, C_2^B, C_2^F can be determined numerically by considering the optimization problem discussed in section 5 for small (two-queue) polling models with Markovian server routing, which can be solved in a similar way as was done in section 5. Note that the observed tendency towards deterministic routing (cf. section 5) may be used here to set certain routing probabilities equal to 1.00.
This iterative algorithm converges when each queue forms a cluster by itself (for small values of the threshold disctance d). It should be noted that the algorithm converges after *at most* s iterations, because (by definition) at each stage at least two queues are united.

We reemphasize the enormous mathematical and numerical complexity of the optimization problem considered here and the idea should be viewed in that perspective. Although the idea of hierarchical clustering is rather intuitive and may hide some unforseen complications, we believe it is interesting to pursue this idea further in the future.

Acknowledgement
The author wishes to thank thank J.P.C. Blanc and O.J. Boxma for their useful comments.

References

[1] S. Asmussen (1987). *Applied Probability and Queues* (Wiley, Chichester).

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

[3] J.P.C. Blanc (1992). Performance evaluation of polling systems by means of the power-series algorithm. *Ann. Oper. Res.* **35**, 155–186.

[4] J.P.C. Blanc (1993). Performance analysis and optimization with the power-series algorithm. In: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson (North-Holland, Amsterdam), 53–80.

[5] J.P.C. Blanc and R.D. van der Mei (1994). Computation of derivatives by means of the power-series algorithm. Submitted.

[6] J.P.C. Blanc and R.D. van der Mei (1992). Optimization of polling systems with Bernoulli schedules. To appear in *Perf. Eval.*.

[7] O.J. Boxma, H. Levy and J.A. Weststrate (1993). Efficient visit orders for polling systems. *Perf. Eval.* **18**, 103–123.

[8] O.J. Boxma and J.A. Weststrate (1989). Waiting times in polling systems with Markovian server routing. In: *Messung, Modellierung und Bewertung von Rechensystemen und Netze*, eds. G Stiege and J.S. Lie (Springer, Berlin), 89–104.

[9] Y.A. Bozer and M.M. Srinivasan (1991). Tandem configurations for automated guided vehicle systems and the analysis of single loops. *IIE Trans.* **23**, 72–82.

[10] C. Chatfield and A.J. Collins (1980). *Introduction to Multivariate Analysis* (Chapman and Hall, London).

[11] H. Chung, C.K. Un and W.Y. Jung (1994). Performance analysis of Markovian polling systems with single buffers. *Perf. Eval.* **19**, 303–315.

[12] C. Fricker and M.R. Jaïbi (1994). Stability of a polling model with a Markovian scheme. INRIA Technical Report 2278, France.

[13] G. Hooghiemstra, M.S. Keane and S. van de Ree (1988). Power series for stationary distributions of coupled processor models. *SIAM J. Appl. Math.* **48**, 1159–1166.

[14] A. Itai and Z. Rosberg (1984). A Golden Ratio control policy for a multiple-access channel. *IEEE Trans. Autom. Control* **29**, 712–718.

[15] L. Kleinrock and H. Levy (1988). The analysis of random poling systems. *Oper. Res.* **36**, 716–732.

[16] G. Koole (1994). On the power-series algorithm. In: *Evaluation of Parallel and Distributed Systems-Solution Methods*, eds. O.J. Boxma and G. Koole, CWI Tract 105 & 106 (CWI, Amsterdam), 139–155.

[17] J.B. Kruskal (1969). Work-scheduling algorithms: a non-probabilistic queueing study (with applications to No 1 ESS). *Bell Syst. Techn. J.* **48**, 2963–2974.

[18] H. Levy (1984). *Non-Uniform Structures and Synchronization Patterns in Shared-Channel Communication Networks.* Ph.D. Thesis, UCLA.

[19] H. Levy and M. Sidi (1990). Polling systems: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750–1760.

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

[21] I. Mitrani, J.L. Adams and R.M. Falconer (1986). A modelling study of the Orwell ring protocol. In: *Teletraffic Analysis and Computer Performance Evaluation*, eds. O.J. Boxma, J.W. Cohen and H.C. Tijms (North-Holland, Amsterdam), 429–438.

[22] S.S. Rao (1984). *Optimization Theory and Applications* (Wiley Eastern Limited, New Delhi, 2nd ed.).

[23] E. Seneta (1981). *Non-Negative Matrices and Markov Chains* (Springer-Verlag, Berlin).

[24] M.M. Srinivasan (1988). Non-deterministic polling systems. *Mgmt. Sc.* **37**, 667–681.

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

[26] H. Takagi (1991). Applications of polling models to computer networks. *Comp. Netw. and ISDN Syst.* **22**, 193–211.

[27] T.E. Tedijanto (1990). *Non-Exhaustive Policies in Polling Systems and Vacation Models.* Ph.D. Thesis, University of Maryland.

[28] B. van Arem (1990). *Queueing Network Models for Slotted Transmission Systems.* Ph.D. Thesis, Twente University, Enschede, The Netherlands.

[29] W.B. van den Hout and J.P.C. Blanc (1994). The power-series algorithm for a wide class of Markov processes. Center discussion paper 9487, Tilburg University, The Netherlands.

[30] W.B. van den Hout and J.P.C. Blanc (1994). The power-series algorithm extended to the BMAP/PH/1 queue. Center discussion paper 9360, Tilburg University, The Netherlands.

[31] J.A. Weststrate (1992). *Analysis and Optimization of Polling Systems.* Ph.D. Thesis, Tilburg University, The Netherlands.

Discussion Paper Series, CentER, Tilburg University, The Netherlands:

(For previous papers please consult previous discussion papers.)

No.	Author(s)	Title
9447	G. Koop, J. Osiewalski and M.F.J. Steel	Hospital Efficiency Analysis Through Individual Effects: A Bayesian Approach
9448	H. Hamers, J. Suijs, S. Tijs and P. Borm	The Split Core for Sequencing Games
9449	G.-J. Otten, H. Peters, and O. Volij	Two Characterizations of the Uniform Rule for Division Problems with Single-Peaked Preferences
9450	A.L. Bovenberg and S.A. Smulders	Transitional Impacts of Environmental Policy in an Endogenous Growth Model
9451	F. Verboven	International Price Discrimination in the European Car Market: An Econometric Model of Oligopoly Behavior with Product Differentiation
9452	P.J.-J. Herings	A Globally and Universally Stable Price Adjustment Process
9453	D. Diamantaras, R.P. Gilles and S. Scotchmer	A Note on the Decentralization of Pareto Optima in Economies with Public Projects and Nonessential Private Goods
9454	F. de Jong, T. Nijman and A. Röell	Price Effects of Trading and Components of the Bid-ask Spread on the Paris Bourse
9455	F. Vella and M. Verbeek	Two-Step Estimation of Simultaneous Equation Panel Data Models with Censored Endogenous Variables
9456	H.A. Keuzenkamp and M. McAleer	Simplicity, Scientific Inference and Econometric Modelling
9457	K. Chatterjee and B. Dutta	Rubinstein Auctions: On Competition for Bargaining Partners
9458	A. van den Nouweland, B. Peleg and S. Tijs	Axiomatic Characterizations of the Walras Correspondence for Generalized Economies
9459	T. ten Raa and E.N. Wolff	Outsourcing of Services and Productivity Growth in Goods Industries
9460	G.J. Almekinders	A Positive Theory of Central Bank Intervention
9461	J.P. Choi	Standardization and Experimentation: Ex Ante Versus Ex Post Standardization

No.	Author(s)	Title
9462	J.P. Choi	Herd Behavior, the "Penguin Effect", and the Suppression of Informational Diffusion: An Analysis of Informational Externalities and Payoff Interdependency
9463	R.H. Gordon and A.L. Bovenberg	Why is Capital so Immobile Internationally?: Possible Explanations and Implications for Capital Income Taxation
9464	E. van Damme and S. Hurkens	Games with Imperfectly Observable Commitment
9465	W. Güth and E. van Damme	Information, Strategic Behavior and Fairness in Ultimatum Bargaining - An Experimental Study -
9466	S.C.W. Eijffinger and J.J.G. Lemmen	The Catching Up of European Money Markets: The Degree Versus the Speed of Integration
9467	W.B. van den Hout and J.P.C. Blanc	The Power-Series Algorithm for Markovian Queueing Networks
9468	H. Webers	The Location Model with Two Periods of Price Competition
9469	P.W.J. De Bijl	Delegation of Responsibility in Organizations
9470	T. van de Klundert and S. Smulders	North-South Knowledge Spillovers and Competition. Convergence Versus Divergence
9471	A. Mountford	Trade Dynamics and Endogenous Growth - An Overlapping Generations Model
9472	A. Mountford	Growth, History and International Capital Flows
9473	L. Meijdam and M. Verhoeven	Comparative Dynamics in Perfect-Foresight Models
9474	L. Meijdam and M. Verhoeven	Constraints in Perfect-Foresight Models: The Case of Old-Age Savings and Public Pension
9475	Z. Yang	A Simplicial Algorithm for Testing the Integral Property of a Polytope
9476	H. Hamers, P. Borm, R. van de Leensel and S. Tijs	The Chinese Postman and Delivery Games
9477	R.M.W.J. Beetsma	Servicing the Public Debt: Comment
9478	R.M.W.J. Beetsma	Inflation Versus Taxation: Representative Democracy and Party Nominations
9479	J.-J. Herings and D. Talman	Intersection Theorems with a Continuum of Intersection Points
9480	K. Aardal	Capacitated Facility Location: Separation Algorithms and Computational Experience

No.	Author(s)	Title
9481	G.W.P. Charlier	A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model with Individual Fixed Effects and Application to Labour Force Participation
9482	J. Bouckaert and H. Degryse	Phonebanking
9483	B. Allen, R. Deneckere, T. Faith and D. Kovenock	Capacity Precommitment as a Barrier to Entry: A Bertrand -Edgeworth Approach
9484	J.-J. Herings, G. van der Laan, D. Talman, and R. Venniker	Equilibrium Adjustment of Disequilibrium Prices
9485	V. Bhaskar	Informational Constraints and the Overlapping Generations Model: Folk and Anti-Folk Theorems
9486	K. Aardal, M. Labbé, J. Leung, and M. Queyranne	On the Two-level Uncapacitated Facility Location Problem
9487	W.B. van den Hout and J.P.C. Blanc	The Power-Series Algorithm for a Wide Class of Markov Processes
9488	F.C. Drost, C.A.J. Klaassen and B.J.M. Werker	Adaptive Estimation in Time-Series Models
9489	Z. Yang	A Simplicial Algorithm for Testing the Integral Property of Polytopes: A Revision
9490	H. Huizinga	Real Exchange Rate Misalignment and Redistribution
9491	A. Blume, D.V. DeJong, Y.-G. Kim, and G.B. Sprinkle	Evolution of the Meaning of Messages in Sender-Receiver Games: An Experiment
9492	R.-A. Dana, C. Le Van, and F. Magnien	General Equilibrium in Asset Markets with or without Short-Selling
9493	S. Eijffinger, M. van Rooij, and E. Schaling	Central Bank Independence: A Paneldata Approach
9494	S. Eijffinger and M. van Keulen	Central Bank Independence in Another Eleven Countries
9495	H. Huizinga	The Incidence of Interest Withholding Taxes: Evidence from the LDC Loan Market
9496	V. Feltkamp, S. Tijs and S. Muto	Minimum Cost Spanning Extension Problems: The Proportional Rule and the Decentralized Rule
9497	J.P.J.F. Scheepens	Financial Intermediation, Bank Failure and Official Assistance

No.	Author(s)	Title
9498	A.L. Bovenberg and R.A. de Mooij	Environmental Tax Reform and Endogenous Growth
9499	J. Ashayeri, R. Heuts, A. Jansen and B. Szczerba	Inventory Management of Repairable Service Parts for Personal Computers: A Case Study
94100	A. Cukierman and S. Webb	Political Influence on the Central Bank - International Evidence
94101	G.J. Almekinders and S.C.W. Eijffinger	The Ineffectiveness of Central Bank Intervention
94102	R. Aalbers	Extinction of the Human Race: Doom-Mongering or Reality?
94103	H. Bester and W. Güth	Is Altruism Evolutionarily Stable?
94104	H. Huizinga	Migration and Income Transfers in the Presence of Labor Quality Externalities
94105	F.C. Drost, T.E. Nijman, and B.J.M. Werker	Estimation and Testing in Models Containing both Jumps and Conditional Heteroskedasticity
94106	V. Feltkamp, S. Tijs, and S. Muto	On the Irreducible Core and the Equal Remaining Obligations Rule of Minimum Cost Spanning Extension Problems
94107	D. Diamantaras, R.P. Gilles and P.H.M. Ruys	Efficiency and Separability in Economies with a Trade Center
94108	R. Ray	The Reform and Design of Commodity Taxes in the Presence of Tax Evasion with Illustrative Evidence from India
94109	F.H. Page	Optimal Auction Design with Risk Aversion and Correlated Information
94110	F. de Roon and C. Veld	An Empirical Investigation of the Factors that Determine the Pricing of Dutch Index Warrants
94111	P.J.-J. Herings	A Globally and Universally Stable Quantity Adjustment Process for an Exchange Economy with Price Rigidities
94112	V. Bhaskar	Noisy Communication and the Fast Evolution of Cooperation
94113	R.C. Douven and J.E.J. Plasmans	S.L.I.M. - A Small Linear Interdependent Model of Eight EU-Member States, the USA and Japan
94114	B. Bettonvil and J.P.C. Kleijnen	Identifying the Important Factors in Simulation Models with Many Factors
94115	H. Uhlig and N. Yanagawa	Increasing the Capital Income Tax Leads to Faster Growth
9501	B. van Aarle, A.L. Bovenberg and M. Raith	Monetary and Fiscal Policy Interaction and Government Debt Stabilization

No.	Author(s)	Title
9502	B. van Aarle and N. Budina	Currency Substitution in Eastern Europe
9503	Z. Yang	A Constructive Proof of a Unimodular Transformation Theorem for Simplices
9504	J.P.C. Kleijnen	Sensitivity Analysis and Optimization of System Dynamics Models: Regression Analysis and Statistical Design of Experiments
9505	S. Eijffinger and E. Schaling	The Ultimate Determinants of Central Bank Independence
9506	J. Ashayeri, A. Teelen and W. Selen	A Production and Maintenance Planning Model for the Process Industry
9507	J. Ashayeri, A. Teelen and W. Selen	Computer Integrated Manufacturing in the Chemical Industry: Theory & Practice
9508	A. Mountford	Can a Brain Drain be Good for Growth?
9509	F. de Roon and C. Veld	Announcement Effects of Convertible Bond Loans Versus Warrant-Bond Loans: An Empirical Analysis for the Dutch Market
9510	P.H. Franses and M. McAleer	Testing Nested and Non-Nested Periodically Integrated Autoregressive Models
9511	R.M.W.J. Beetsma	The Political Economy of a Changing Population
9512	V. Kriman and R.Y. Rubinstein	Polynomial Time Algorithms for Estimation of Rare Events in Queueing Models
9513	J.P.C. Kleijnen, W.J.H. van Groenendaal and R.Y. Rubinstein	Optimization and Sensitivity Analysis of Computer Simulation Models by the Score Function Method
9514	R.D. van der Mei	Polling Systems with Markovian Server Routing