Waiting-time approximations for multiple-server polling systems

S.C. Borst, R.D. van der Mei

Department of Operations Research, Statistics, and System Theory

Report BS-R9428 September 1994

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

Waiting-Time Approximations for Multiple-Server Polling Systems

S.C. Borst
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

R.D. van der Mei
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

We consider multiple-server polling systems in which each of the servers visits the queues according to its own cyclic schedule. Such systems appear to completely defy the derivation of exact analytical waiting-time results, which motivates the search for accurate approximations. In the present paper we derive waiting-time approximations for asymmetric systems with the exhaustive and gated service discipline. The approximations are tested for a wide range of parameter combinations.

AMS Subject Classification (1991): 60K25, 68M20.
Keywords & Phrases: cyclic service, polling system, multiple servers, waiting-time approximation.

1 INTRODUCTION

A multiple-server polling system is a multiple-queue system attended by multiple servers, which visit the queues according to some routing mechanism. Like the single-server versions, multiple-server polling systems find many applications in computer systems, communication networks, and manufacturing environments. So far there are hardly any exact results known for these systems, apart from some mean-value results for global performance measures like cycle times. Motivated by the mathematical intractability, we derive in the present paper waiting-time approximations for systems with the exhaustive and gated service discipline in which each of the servers visits the queues according to its own cyclic schedule. In the remainder of the introduction we successively discuss some applications, present an overview of related literature, and describe the organization of the paper.

Applications

An example of a multiple-server polling system is a distributed system consisting of a number of computers, interconnected by a communication medium, that cooperate as follows in sharing the total load of the system, cf. [26]. The jobs entering the 'front-end' systems (corresponding to the queues) are picked up in batches by the 'back-end' systems (corresponding to the servers) according to some cyclic schedule. As soon as a back-end system completes

Report BS-R9428
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

the service of a batch, it picks up the jobs from the next front-end system. When the communication delay is negligible, it may be preferable to use a more sophisticated strategy which keeps track of the evolution of the system, like "serve the longest queue". However, when the communication overhead is significant, it makes sense to use a cyclic service schedule which requires only very little processing information.
Examples also arise in communication networks, like the underlying communication medium in the above-mentioned distributed system. Consider e.g. a Local Area Network (LAN), consisting of a number of stations, interconnected by a transmission ring. There are various protocols known for the medium access control in a LAN with a ring architecture. One variant is the multiple slotted ring, i.e., the ring is subdivided into time slots of the size of a single packet, circulating at constant speed. Occupying a slot corresponds to utilizing a server. Another medium access variant that may lead to multiple-server polling is the multiple token ring, i.e., there are multiple rings, each with a token circulating on it, representing the right of transmission on that particular ring. Holding the token corresponds to utilizing the server.
Other examples arise in manufacturing environments. In flexible manufacturing systems, e.g., one finds a group of machines that periodically change over from one type of operations to another. In some factories internal transport is provided by a chain of vehicles that follow a track along loading/unloading points, which conceptually comes very close to a slotted ring with destination release. A multiple-elevator facility is yet another example, although some features, like the acceleration effect when a floor is skipped, are not incorporated in the classical description of a polling model, cf. [17].

Related literature
Multiple-server polling systems have received remarkably little attention in the vast literature on polling systems. One of the first studies is Morris & Wang [26] in which the servers are assumed to be independent, i.e., to visit the queues independently of each other, each server according to some cyclic schedule. They obtain the mean cycle time of each server and the mean intervisit time to a queue, and derive approximate expressions for the mean sojourn time for both a gated-type and a limited-type service discipline. A very interesting phenomenon observed by Morris & Wang is the tendency for the servers to cluster if they follow identical routes, especially in heavy traffic, cf. also [24]. The phenomenon may be visualized as follows. A trailing server will tend to move fast, as it only encounters recently served queues, whereas a leading server will tend to be slowed down by queues that have not been served for a while, so that the servers tend to form bunches while constantly leapfrogging over one another. Numerical experiments indicate that the bunching of servers is likely to deteriorate the system performance. Obviously the bunching of servers is alleviated if they follow different routes. Therefore Morris & Wang advocate the use of 'dispersive' schedules to improve the system performance.
Levy & Yechiali [22] and Kao & Narayanan [20] study the joint distribution of the queue length and the number of busy servers for a Markovian multiple-server queue, where the servers individually go on vacation when there are no waiting customers left. Mitrany & Avi-Itzhak [25] and Neuts & Lucantoni [27] analyze the joint distribution of the queue length and the number of busy servers for a Markovian multiple-server queue where servers break down at exponential intervals and then get repaired.

In references [6], [19], [21], [28], [29] mean response time approximations are developed to analyze the performance of LAN's with multiple token rings. Mean response time approximations oriented to LAN's with a multiple slotted ring are contained in references [5], [6], [23], [29], [30]. Ajmone Marsan et al. [2], [3], [4] derive the mean cycle time and bounds for the mean waiting times in symmetric systems for the exhaustive, gated, and 1-limited service discipline. In [1] they illustrate how Petri-net techniques may be used to study Markovian multiple-server polling systems.
Browne & Weiss [12] is one of the few studies in which the servers are assumed to be coupled, i.e., to visit the queues together. They obtain index-type rules for determining the visit order that minimizes the mean length of individual cycles for both the exhaustive and the gated service discipline. Browne et al. [10] derive the mean waiting time for a completely symmetric two-queue system with an infinite number of coupled servers and deterministic service times. Browne & Kella [11] obtain the busy period distribution for a two-queue system with an infinite number of coupled servers, exhaustive service, and deterministic service times at one queue and general service times at the other queue. Borst [7] explores the class of systems that allow an exact analysis in the case of coupled servers. Van der Mei & Borst [24] show how a broad class of multiple-server polling systems may be analyzed numerically by means of the power-series algorithm (PSA).
The above-mentioned studies unanimously point out that multiple-server polling systems are extraordinarily hard to analyze. In fact almost none of the studies (except [10], [11], [12], [7] for the case of coupled servers, and the single-queue studies [20], [22], [25], [27]) presents any exact results, apart from some mean-value results for global performance measures like cycle times. In the case that the servers are not coupled, to the best of the authors' knowledge, there is not a single non-trivial multiple-queue system for which any exact expressions for the waiting times are known. Motivated by the mathematical intractability, we derive in the present paper waiting-time approximations for asymmetric systems with the exhaustive and gated service discipline in which each of the servers visits the queues according to its own cyclic schedule. Most of the existing approximations are only developed for systems with the 1-limited service discipline or are restricted to completely symmetric systems. Moreover, most of the approximations completely ignore the considerable influence of the visit order on the waiting times or simply assume the visit order to be dispersive.

Organization of the paper
The remainder of the paper is organized as follows. We present a detailed model description in section 2. In section 3 some preliminary results are obtained for the mean interarrival times of the various servers at the various queues, which will repeatedly be used throughout the next sections. In sections 4, 5, and 6 we derive waiting-time approximations for asymmetric systems with the exhaustive and gated service discipline. Considering the merits and drawbacks of existing approximations, we intend to (i) use pseudo-conservation-like concepts which have proven to be a very useful instrument in the single-server case, and (ii) take into account the visit orders of the servers which in the multiple-server case, through the clustering effects, appear to have a major impact on the waiting times. In section 7 the approximations are tested for a wide range of parameter combinations by comparison with either simulation results or exact numerical results obtained from the PSA, as elaborated upon in [24]. In section 8 we conclude with some remarks and suggestions for further research.

2 MODEL DESCRIPTION

The model under consideration consists of n queues $Q_1, \ldots, Q_n$, each of infinite capacity, attended by m identical servers $S_1, \ldots, S_m$. Customers arrive at the queues according to independent Poisson processes. Customers arriving at Q_i will also be referred to as type-i customers, $i = 1, \ldots, n$. Denote by λ_i the arrival rate at Q_i, $i = 1, \ldots, n$. The total arrival rate is $\lambda := \sum_{i=1}^{n} \lambda_i$. Type-$i$ customers require service times with first moment β_i and second moment $\beta_i^{(2)}$, $i = 1, \ldots, n$. All service times are assumed to be independent. Define the traffic intensity at Q_i as $\rho_i := \lambda_i \beta_i$, $i = 1, \ldots, n$. The total traffic intensity is $\rho := \sum_{i=1}^{n} \rho_i$.

The servers move from queue to queue in a cyclic manner. Server j visits the queues in the order $Q_{\pi_j(1)}, \ldots, Q_{\pi_j(n)}$, with $(\pi_j(1), \ldots, \pi_j(n))$ a permutation of $(1, \ldots, n)$, $j = 1, \ldots, m$. Moving into Q_i a server incurs a switch-over time with first moment s_i and second moment $s_i^{(2)}$, $i = 1, \ldots, n$. All switch-over times are assumed to be independent. Note that the total switch-over time incurred during a cycle has the same distribution for each server, with first moment $s := \sum_{i=1}^{n} s_i$ and second moment $s^{(2)} := \sum_{i=1}^{n} s_i^{(2)} + \sum_{i \neq k} s_i s_k$. The arrival, service, and switch-over processes are assumed to be mutually independent.

The servers visit the queues independently of each other, under the restriction that at most m_i servers may visit Q_i simultaneously. In view of the latter restriction a server arrival will be called effective if there are less than m_i other servers already busy at Q_i. If an arrival at Q_i is not effective, then the server starts switching to the next queue immediately. If an arrival at Q_i *is* effective, then the server starts serving type-i customers (possibly none), as prescribed by the service discipline at Q_i. At each queue the service discipline may either be exhaustive or gated. Under the exhaustive service discipline a server leaves the queue when there are no waiting customers left. Under the gated service discipline a server leaves the queue when there are no waiting customers left whose arrival fell before the last server arrival. In other words, at each server arrival an imaginary gate opens to let waiting customers pass through. At each queue customers are taken into service in order of arrival. As soon as the server finishes serving type-i customers, as prescribed by the service discipline at Q_i, it starts switching to the next queue as specified in its schedule.

Finally some words on the stability conditions. Necessary conditions are of course that $\rho < m$, $\rho_i < m_i$, $i = 1, \ldots, n$. We strongly conjecture that these conditions are in fact also sufficient for service disciplines, like exhaustive and gated, that do not impose any (probabilistic) parametric restriction on the number of customers served during a server visit. Throughout the paper the stability conditions are assumed to hold.

3 THE SERVER INTERARRIVAL TIME

In this section we derive some preliminary results for the mean interarrival time of the various servers at the various queues, which will repeatedly be used throughout the next sections in obtaining waiting-time approximations. We first introduce some notation. Define r_{ij} as the load carried by S_j at Q_i, i.e., the fraction of time that S_j is busy at Q_i, $i = 1, \ldots, n$, $j = 1, \ldots, m$. The total load carried by S_j is $r_j = \sum_{i=1}^{n} r_{ij}$. In general the fractions r_{ij}

are unknown. However, the balance between the carried and the offered load at Q_i implies $\sum_{j=1}^{m} r_{ij} = \rho_i$, $i = 1, \ldots, n$.

Denote by $\mathbf{A}_{ij}$ ($\mathbf{A}^*_{ij}$) the interarrival (effective interarrival) time of S_j at Q_i, i.e., the time between two consecutive arrivals (effective arrivals) of S_j at Q_i, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Denote by p_{ij} the probability that an arbitrary arrival of S_j at Q_i is effective, i.e., the probability that at an arbitrary arrival of S_j there are less than m_i other servers already busy at Q_i, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Obviously, for $m_i = m$, $p_{ij} = 1$; for $m_i < m$ the probabilities p_{ij} are however not known.

Applying a traffic balance argument,

$$\mathrm{E}\mathbf{A}_{ij} = \frac{s}{1 - r_j}, \tag{3.1}$$

independent of i.

As $p_{ij} = \mathrm{E}\mathbf{A}_{ij}/\mathrm{E}\mathbf{A}^*_{ij}$,

$$\mathrm{E}\mathbf{A}^*_{ij} = \frac{s/p_{ij}}{1 - r_j}. \tag{3.2}$$

A question that arises here quite naturally is whether or not all the servers will carry the same load. If two servers follow the same visit order, then by symmetry considerations both will also carry the same load of course. In particular, if all the servers follow the same visit order, then $r_{ij} = \rho_i/m$, $j = 1, \ldots, m$. Numerical experiments indicate that, even when the servers follow different visit orders, at each individual queue the load carried by each of the servers tends to differ only slightly, although in case of highly asymmetric system configurations the differences may increase somewhat. However, as observed in [24], [26], even in case of highly asymmetric system configurations the *total* load carried by each of the servers does not appear to differ significantly.

The above observation may be explained as follows. Suppose that the total load r_1 carried by S_1 is larger than the total load r_2 carried by S_2. So by (3.1) also the mean interarrival time $\mathrm{E}\mathbf{A}_{i1}$ of S_1 is larger than the mean interarrival time $\mathrm{E}\mathbf{A}_{i2}$ of S_2. In other words, S_2 visits the queues more frequently than S_1, so that S_2 is also likely (but not absolutely sure) to meet more work at the queues than S_1. So the total load r_2 carried by S_2 is likely to be larger than the total load r_1 carried by S_1, in contradiction with the initial supposition. The above explanation does not exclude that some minor differences may occur in the total load carried by each of the servers. However, the reasoning supports the observation that such differences cannot grow dramatically.

Denote by $\mathbf{A}_i$ ($\mathbf{A}^*_i$) the server interarrival (effective server interarrival) time at Q_i, i.e., the time between two consecutive server arrivals (effective server arrivals) at Q_i, $i = 1, \ldots, n$. Denote by p_i the probability that an arbitrary server arrival at Q_i is effective, i.e., the probability that at an arbitrary server arrival there are less than m_i other servers busy at Q_i, $i = 1, \ldots, n$. Obviously, for $m_i = m$, $p_i = 1$; for $m_i < m$ the probabilities p_i are however not known.

The $\mathbf{A}_i$ ($\mathbf{A}^*_i$) process is the superposition of the $\mathbf{A}_{ij}$ ($\mathbf{A}^*_{ij}$) processes. So,

$$\frac{1}{\mathrm{E}\mathbf{A}_i} = \sum_{j=1}^{m} \frac{1}{\mathrm{E}\mathbf{A}_{ij}} \qquad \left(\frac{1}{\mathrm{E}\mathbf{A}^*_i} = \sum_{j=1}^{m} \frac{1}{\mathrm{E}\mathbf{A}^*_{ij}} \right), \qquad i = 1, \ldots, n.$$

So, from (3.1),

$$\mathrm{E}\mathbf{A}_i = \frac{s}{m-\rho}, \tag{3.3}$$

which is again like (3.1) independent of i. Moreover, the mean server interarrival time is completely insensitive to how the total load is divided among the individual servers.
As $p_i = \mathrm{E}\mathbf{A}_i/\mathrm{E}\mathbf{A}_i^*$,

$$\mathrm{E}\mathbf{A}_i^* = \frac{s/p_i}{m-\rho}, \qquad i = 1, \ldots, n. \tag{3.4}$$

Note that the mean-value results obtained here for the (effective) server interarrival time also hold for the (effective) server interdeparture time. (A departure is called effective if it corresponds to an effective arrival.)

4 The waiting time

In this section we derive waiting-time approximations for systems with the exhaustive and gated service discipline. We first introduce some notation. Denote by $\mathbf{W}_i$ the waiting time of an arbitrary type-i customer, $i = 1, \ldots, n$. For any non-negative continuous stochastic variable $\mathbf{X}$, denote by $\mathbf{RX}$ a stochastic variable with as distribution the residual-lifetime distribution of $\mathbf{X}$, i.e., $\Pr\{\mathbf{RX} < t\} = \dfrac{1}{\mathrm{E}\mathbf{X}} \displaystyle\int_{u=0}^{t} (1 - \Pr\{\mathbf{X} < u\})du$, $t \geq 0$.

For reference, we first briefly review the single-server case. The usual approach to obtain waiting-time approximations may be outlined as follows. To start with, one derives an (approximative) relationship of the form

$$\mathrm{E}\mathbf{W}_i \approx \gamma_i \mathrm{E}\mathbf{RC}_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

with $\mathbf{C}_i$ either the interarrival or the interdeparture time at Q_i, depending on the service discipline at Q_i. The symbol γ_i represents some coefficient in terms of the system parameters, which reflects the influence of the service discipline at Q_i.
For the exhaustive service discipline,

$$\mathrm{E}\mathbf{W}_i = (1-\rho_i)\mathrm{E}\mathbf{RD}_i, \qquad i = 1, \ldots, n, \tag{4.2}$$

with $\mathbf{D}_i$ the server inter*departure* time at Q_i, cf. [15], [18]. (An alternative relationship for the exhaustive service discipline is $\mathrm{E}\mathbf{W}_i = \dfrac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \mathrm{E}\mathbf{RI}_i$, with $\mathbf{I}_i$ the intervisit time at Q_i, cf. [14].)
For the gated service discipline,

$$\mathrm{E}\mathbf{W}_i = (1+\rho_i)\mathrm{E}\mathbf{RA}_i, \tag{4.3}$$

with, as before, $\mathbf{A}_i$ the server inter*arrival* time at Q_i, cf. [15], [18].
To proceed, one turns to approximating $\mathrm{E}\mathbf{RC}_i$ (or $\mathrm{E}\mathbf{RI}_i$). Since $\mathrm{E}\mathbf{RC}_i = \mathrm{E}(\mathbf{C}_i)^2/2\mathrm{E}\mathbf{C}_i$ (similarly for $\mathrm{E}\mathbf{RI}_i$), cf. [16], where $\mathrm{E}\mathbf{C}_i = s/(1-\rho)$ (respectively $\mathrm{E}\mathbf{I}_i = (1-\rho_i)s/(1-\rho)$),

it remains to approximate $E(\mathbf{C}_i)^2$ (or $E(\mathbf{I}_i)^2$) by using some additional information. One approach, followed by Bux & Truong [14] in the case of exhaustive service and deterministic switch-over times, is to derive an exact formula for $E(\mathbf{I}_i^2)$ in the case of two queues, subsequently applying a 'heuristic extrapolation' to the case of an arbitrary number of queues. Another approach, proposed by Everitt [15] and further elaborated on by Groenendijk [18] is to approximate $\mathbf{ERC}_i$ in a direct manner, by invoking a so-called pseudo-conservation law, which provides an exact explicit expression for a weighted sum of the mean waiting times, typically $\sum_{i=1}^{n} \rho_i \mathbf{EW}_i$, cf. [8], [9]. Substituting (4.1) into a pseudo-conservation law, assuming $\mathbf{ERC}_i \approx \overline{\mathbf{ERC}}$, yields an approximation for $\mathbf{ERC}_i$. Note that the latter method in particular yields an exact expression for the mean waiting time in completely symmetric systems.

We now return to the multiple-server case. The usual approach to obtain waiting-time approximations in the multiple-server case may be sketched as follows, cf. [6], [19], [21], [26], [28], [29]. Like in the single-server case, one starts by deriving an (approximative) relationship of the form

$$\mathbf{EW}_i \approx \gamma_i \mathbf{ERC}_i^*, \tag{4.4}$$

with $\mathbf{C}_i^*$ either the *effective* server interarrival or interdeparture time at Q_i. At that stage the complications already start, as for most service disciplines at best a very rough approximation for γ_i can be found. Next, like in the single-server case, one proceeds by approximating $\mathbf{ERC}_i^*$. The complications then grow even worse, as there is very little additional information available that can be used, neither in the form of exact results for special cases, nor in the global form of a pseudo-conservation law. Having little choice left, one typically considers the $\mathbf{C}_i^*$-process as resulting from the $\mathbf{C}_i$-process after a 'filtering' with probability p_i (the probability of an arrival at Q_i being effective), and then the $\mathbf{C}_i$-process in its turn as the superposition of the $\mathbf{C}_{ij}$-processes, with j indicating server j. Subsequently one approximates p_i and fits some distribution to the $\mathbf{C}_{ij}$-processes, assuming that the $\mathbf{C}_{ij}$-processes are independent and identically distributed. The motivation for fitting some particular distribution to the $\mathbf{C}_{ij}$-processes is at best questionable, but is usually even completely lacking. What is worse however, is that the assumption that the $\mathbf{C}_{ij}$-processes are independent and identically distributed completely ignores the tendency for the servers to cluster, which immediately explains why the resulting approximations only appear to be reasonably accurate for dispersive schedules or under conditions (like $m_i = 1$, or 1-limited service) with dispersive effects, cf. [6], [19], [21], [26], [28], [29].
We now describe an alternative approach to derive waiting-time approximations. From now on we focus on the case $m_i = m$, which we consider to be the most interesting case; in the last section of the paper we briefly discuss the case $m_i = 1$. Considering the above-mentioned objections, we intend

(i). to take into account the visit orders of the servers which in the multiple-server case, through the clustering effects, appear to have a major impact on the waiting times;
(ii). to avoid considering cycle-time processes, instead using pseudo-conservation-like concepts which have proven to be a very useful instrument in the single-server case.

Denote by q_i the steady-state probability that at least one of the servers is busy at Q_i. In

general the probabilities q_i are unknown. However, $\rho_i/m \leq q_i \leq \min\{\rho_i, 1\}$. To derive an approximative relationship of the form $\mathbf{EW}_i \approx \gamma_i \mathbf{ERC}_i$, we assume that the customers experience the presence of multiple servers as if there is a single server processing at speed $\alpha_i = \rho_i/q_i$, the exact average processing speed at Q_i.
For the exhaustive service discipline we then obtain from (4.2), replacing ρ_i by ρ_i/α_i,

$$\mathbf{EW}_i \approx (1 - q_i)\mathbf{ERD}_i, \tag{4.5}$$

with $\mathbf{D}_i$ the server interdeparture time at Q_i.
Similarly, we obtain from (4.3) for the gated service discipline,

$$\mathbf{EW}_i \approx (1 + q_i)\mathbf{ERA}_i, \tag{4.6}$$

with $\mathbf{A}_i$, as before, the server interarrival time at Q_i.
In the multiple-server case it is no longer reasonable to assume that the residual server interdeparture (interarrival) times are approximately equal, as the degree of clustering may differ significantly from queue to queue. Instead we assume that the residual server interdeparture (interarrival) times are proportional to the average processing speed $\alpha_i = \rho_i/q_i$, which may be seen as a measure for the degree of clustering at Q_i, i.e.,

$$\mathbf{ERD}_i \approx \mathbf{ERD}\rho_i/q_i, \tag{4.7}$$

and

$$\mathbf{ERA}_i \approx \mathbf{ERA}\rho_i/q_i, \tag{4.8}$$

with $\mathbf{ERD}$ and $\mathbf{ERA}$ some unknown constants. Note that in case $m = 1$, $q_i = \rho_i$, so that (4.7) and (4.8) reduce to $\mathbf{ERD}_i \approx \mathbf{ERD}$ and $\mathbf{ERA}_i \approx \mathbf{ERA}$, respectively, the usual assumptions in the single-server case.
To complete the derivation of the approximations, it suffices to (i) find an expression for the weighted sum $\sum_{i=1}^{n} \rho_i \mathbf{EW}_i$ and (ii) determine the probabilities q_i, which we will do in sections 5 and 6, respectively.

5 Approximating the weighted sum $\sum_{i=1}^{n} \rho_i \mathbf{EW}_i$

In this section we describe a method for approximating $\sum_{i=1}^{n} \rho_i \mathbf{EW}_i$. Denote by $\mathbf{V}$ the steady-state total amount of work in the system. Applying Brumelle's formula [13],

$$\sum_{i=1}^{n} \rho_i \mathbf{EW}_i = \mathbf{EV} - \frac{1}{2}\sum_{i=1}^{n} \lambda_i \beta_i^{(2)}. \tag{5.1}$$

So to find an expression for $\sum_{i=1}^{n} \rho_i \mathbf{EW}_i$ it suffices to find an expression for the mean amount of work $\mathbf{EV}$. For reference, we first briefly review the single-server case, where the crucial property that facilitates the determination of $\mathbf{EV}$ is work decomposition, which in its turn builds on the fundamental property of work conservation. To illuminate these concepts, denote by $\mathbf{V}^0$ the steady-state total amount of work in the 'corresponding $M/G/1$ system'. The

'corresponding $M/G/1$ system' is a single-server system with similar traffic characteristics but with zero switch-over times, i.e., without any interruptions by the switch-over process. Denote by $\mathbf{Y}$ the steady-state amount of work in the original system in a switching interval. Then the following *work decomposition* property holds, cf. [8], [9]:

$$\mathbf{V} \stackrel{d}{=} \mathbf{V}^0 + \mathbf{Y}, \tag{5.2}$$

with $\stackrel{d}{=}$ indicating equality in distribution. When the amount of work in a switching interval is always zero, we may recognize in (5.2) the underlying property of *work conservation*, which in fact holds even in sample-path sense. Note that $\mathrm{E}\mathbf{V}^0$ is simply known from the Pollaczek-Khintchine formula. For a broad class of service disciplines, including gated and exhaustive, $\mathrm{E}\mathbf{Y}$ may be determined along the lines of [8], [9]. Taking expectations in (5.2), substituting into (5.1), then yields a so-called *pseudo-conservation law* for the mean waiting times.
We now return to the multiple-server case, where deriving a pseudo-conservation law in an exact way involves serious complications. A simple interchange argument shows that in the multiple-server case a strict work conservation property in sample-path sense only holds if all the customers have the same (deterministic) service time. A weaker work conservation property in stochastic sense only holds if all the customers have the same service time distribution and the service discipline is regardless of the actual service times. Hence, since work conservation may be seen as the basis for work decomposition, it is not very likely that a property like (5.2) holds in the multiple-server case. Even if it were, we would face the problem that $\mathrm{E}\mathbf{V}^0$ is not known in the multiple-server case, not to mention the problem of determining $\mathrm{E}\mathbf{Y}$, so that the chances of deriving a pseudo-conservation law in an exact way appear to be negligible. Instead we therefore derive an approximative pseudo-conservation law. Although a work decomposition property probably does not hold, we can always write

$$\mathrm{E}\mathbf{V} = \mathrm{E}\mathbf{V}^0 + \mathrm{E}\mathbf{Y},$$

with $\mathbf{V}^0$ denoting the steady-state total amount of work in the 'corresponding $M/G/m$ system' and $\mathbf{Y}$ representing a stochastic variable whose mean satisfies the above equality but which further remains unspecified. (The 'corresponding $M/G/m$ system' is defined analogously as in the single-server case.) To approximate $\mathrm{E}\mathbf{V}$, we consider two auxiliary single-server systems with similar characteristics, for which the work decomposition property *does* hold, viz:
(i). the 'λ/m system', i.e., a single-server system with identical characteristics, but with the arrival rate decreased by a factor m;
(ii). the 'β/m system', i.e., a single-server system with identical characteristics, but with the service rate increased by a factor m.
For these auxiliary systems we adopt the notational convention introduced for the original system.
Applying (5.2) to the two auxiliary systems,

$$\mathbf{V}_{\lambda/m} \stackrel{d}{=} \mathbf{V}^0_{\lambda/m} + \mathbf{Y}_{\lambda/m}, \quad \mathbf{V}_{\beta/m} \stackrel{d}{=} \mathbf{V}^0_{\beta/m} + \mathbf{Y}_{\beta/m}.$$

From the Pollaczek-Khintchine formula,

$$\mathrm{E}\mathbf{V}^0_{\lambda/m} = \mathrm{E}\mathbf{V}^0_{\beta/m} = \frac{1}{m} \frac{\sum_{i=1}^{n} \lambda_i \beta_i^{(2)}}{2(1-\hat{\rho})},$$

with $\hat{\rho} = \rho/m$.
From [8], [9],

$$\mathrm{E}\mathbf{Y}_{\lambda/m} = \frac{1}{m}\mathrm{E}\mathbf{Y}_{\beta/m} = \hat{\rho}\frac{s^{(2)}}{2s} + \frac{s}{2(1-\hat{\rho})}\left[\hat{\rho}^2 - \sum_{i\in E}\hat{\rho}_i^2 + \sum_{i\in G}\hat{\rho}_i^2\right],$$

with $\hat{\rho}_i = \rho_i/m$. The symbols E and G indicate the index sets of the queues with the exhaustive and gated service discipline, respectively.
To approximate $\mathrm{E}\mathbf{V}^0$, we assume that the ratio of the mean amount of work in a multiple-server system and a single-server system with similar characteristics and proportional load, is rather insensitive to the service time distribution, i.e.,

$$\frac{\mathrm{E}\mathbf{V}^0}{\mathrm{E}\mathbf{V}^0_{\lambda/m}} = \frac{\mathrm{E}\mathbf{V}^0}{\mathrm{E}\mathbf{V}^0_{\beta/m}} \approx \gamma(\rho),$$

with $\gamma(\rho)$ denoting the known value of the ratio in question in case of identically exponentially distributed service times. In other words,

$$\mathrm{E}\mathbf{V}_0 \approx \gamma(\rho)\mathrm{E}\mathbf{V}^0_{\lambda/m} = \gamma(\rho)\mathrm{E}\mathbf{V}^0_{\beta/m},$$

with

$$\gamma(\rho) = \frac{(m-\rho)\sum_{l=0}^{m-1} l\frac{\rho^l}{l!} + \frac{\rho^m}{m!}m^2 + \frac{\rho^m}{m!}\frac{m\rho}{m-\rho}}{\rho\sum_{l=0}^{m-1}\frac{\rho^l}{l!} + \frac{\rho^m}{m!}\frac{m\rho}{m-\rho}}. \tag{5.3}$$

To approximate $\mathrm{E}\mathbf{Y}$, we assume

$$\mathrm{E}\mathbf{Y} \approx (1-\alpha)\zeta_{\lambda/m}(\rho)\mathrm{E}\mathbf{Y}_{\lambda/m} + \alpha\zeta_{\beta/m}(\rho)\mathrm{E}\mathbf{Y}_{\beta/m},$$

with α indicating whether the comparison with the 'λ/m system' or with the 'β/m system' is most appropriate. The interpretation of the coefficients $\zeta_{\lambda/m}(\rho)$ and $\zeta_{\beta/m}(\rho)$ is similar to that of the factor $\gamma(\rho)$ introduced above.
If the server clustering is strong, which will occur especially in heavy traffic if the servers follow identical routes, then the system will tend to behave as the 'β/m system', i.e., $\alpha \uparrow 1$ for a high degree of clustering. It also suggests choosing $\zeta_{\beta/m}(\rho) = 1$ (note from (5.3) that also $\gamma(\rho) \downarrow 1$ when $\rho \uparrow m$). On the other hand, if the server clustering is weak, which will occur in light traffic, or if a dispersive schedule is used, then the comparison with the 'λ/m system' is probably more appropriate, i.e., $\alpha \downarrow 0$ for a low degree of clustering. Choosing $\zeta_{\lambda/m}(\rho)$ in this case is however not so easy. In light traffic the switch-over times will tend to dominate the behavior of the system. In fact, if the total switch-over time incurred during a cycle is deterministic, $\mathrm{E}\mathbf{W}_i \downarrow s/(m+1)$ for $\rho \downarrow 0$. If the total switch-over time during a cycle is exponentially distributed, $\mathrm{E}\mathbf{W}_i \downarrow s/m$ for $\rho \downarrow 0$. Interpolating we obtain $\mathrm{E}\mathbf{W}_i \downarrow (m + s^{(2)}/s^2 - 1)s/m(m+1)$ for $\rho \downarrow 0$, implying that $\mathrm{E}\mathbf{Y} = \rho(m+s^{(2)}/s^2-1)s/m(m+1)+O(\rho^2)$ for $\rho \downarrow 0$. Note that $\mathrm{E}\mathbf{Y}_{\lambda/m} = \rho s^{(2)}/2ms+O(\rho^2)$ for $\rho \downarrow 0$. So $\mathrm{E}\mathbf{Y}/\mathrm{E}\mathbf{Y}_{\lambda/m} \to \dfrac{(m+s^{(2)}/s^2-1)s/m(m+1)}{s^{(2)}/2ms} = 2(1+(m-1)s^2/s^{(2)})/(m+1)$

for $\rho \downarrow 0$. In other words, $\zeta_{\lambda/m}(\rho) \to 2(1+(m-1)s^2/s^{(2)})/(m+1)$ for $\rho \downarrow 0$. On the other hand, in heavy traffic the switch-over times occupy only a negligible fraction of time, implying that $\zeta_{\lambda/m}(\rho) \to \gamma(\rho)$ for $\rho \uparrow m$. Interpolating we obtain $\zeta_{\lambda/m}(\rho) \approx 2(\rho/m)(1+(m-1)s^2/s^{(2)})/(m+1)+\gamma(\rho)(1-\rho/m)$.
To choose α, we consider again $\alpha_i = \rho_i/q_i$, the average processing speed at Q_i, as a measure for the degree of clustering at Q_i. We define

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_i - \rho_i/\left(1-(1-\rho_i/m)^m\right)}{m-\rho_i/\left(1-(1-\rho_i/m)^m\right)}. \tag{5.4}$$

Note that for a high degree of clustering, i.e., $q_i \downarrow \rho_i/m$, $\alpha \uparrow 1$. On the other hand, for a low degree of clustering, i.e., $q_i \uparrow 1-(1-\rho_i/m)^m$, $\alpha \downarrow 0$.
Concluding,

$$\begin{aligned} \mathrm{EV} \approx \frac{\gamma(\rho)}{m}\frac{\sum_{i=1}^{n}\lambda_i\beta_i^{(2)}}{2(1-\hat{\rho})} &+ \left((1-\alpha)\left(2\frac{\rho}{m}\frac{1+(m-1)s^2/s^{(2)}}{m+1}+\gamma(\rho)\left(1-\frac{\rho}{m}\right)\right)+\alpha m\right) \\ &\times \left(\hat{\rho}\frac{s^{(2)}}{2s}+\frac{s}{2(1-\hat{\rho})}\left[\hat{\rho}^2-\sum_{i\in E}\hat{\rho}_i^2+\sum_{i\in G}\hat{\rho}_i^2\right]\right), \end{aligned} \tag{5.5}$$

with $\gamma(\rho)$ and α as in (5.3) and (5.4), respectively. Substituting (5.5) into (5.1) yields an approximative pseudo-conservation law. Subsequently substituting (4.5), (4.6), (4.7), (4.8) into (5.1) yields waiting-time approximations, still containing the probabilities q_i, which we will determine in the next section. Note that the method in particular yields an exact expression for the mean waiting time in completely symmetric systems with exponential service times and zero switch-over times.

6 Approximating the probabilities q_i

In this section we describe a method for approximating the probabilities q_i that at least one of the servers is busy at Q_i, $i=1,\ldots,n$. We first introduce some notation. Denote by $\mathbf{H}_j(t)$ the entry in the polling table of S_j at time t. Indicate by $\mathbf{Z}_j(t)$ whether S_j is switching ($\mathbf{Z}_j(t)=0$) or serving ($\mathbf{Z}_j(t)=1$) at time t. So if $(\mathbf{H}_j(t),\mathbf{Z}_j(t))=(h,0)$, then S_j is switching to $Q_{\pi_j(h)}$ at time t; if $(\mathbf{H}_j(t),\mathbf{Z}_j(t))=(h,1)$, then S_j is serving at $Q_{\pi_j(h)}$ at time t. Denote by $(\mathbf{H},\mathbf{Z})$ a pair of stochastic variables with as joint distribution the joint stationary distribution of $(\mathbf{H}(t),\mathbf{Z}(t))$ with $(\mathbf{H}(t),\mathbf{Z}(t))=(\mathbf{H}_1(t),\ldots,\mathbf{H}_m(t),\mathbf{Z}_1(t),\ldots,\mathbf{Z}_m(t))$.
We now describe a method for approximating the distribution of $(\mathbf{H},\mathbf{Z})$. Note that the probabilities q_i follow immediately from the distribution of $(\mathbf{H},\mathbf{Z})$ as

$$q_i = 1-\Pr\{(\pi_j(\mathbf{H}_j),\mathbf{Z}_j)\neq(i,1), j=1,\ldots,m\}. \tag{6.1}$$

In fact it is not difficult to approximate each of the *marginal* distributions of $(\mathbf{H}_j,\mathbf{Z}_j)$, $j=1,\ldots,m$. As observed in section 3, at each individual queue the load carried by each of the servers tends to differ only rather slightly, i.e., $r_{ij}\approx\rho_i/m$, $i=1,\ldots,n$, $j=1,\ldots,m$. So,

$$\Pr\{(\mathbf{H}_j,\mathbf{Z}_j)=(h,1)\}=r_{\pi_j(h)j}\approx\frac{\rho_{\pi_j(h)}}{m}. \tag{6.2}$$

Also, from (3.1),

$$\Pr\{(\mathbf{H}_j, \mathbf{Z}_j) = (h, 0)\} = \frac{s_{\pi_j(h)}}{\mathbf{EC}_j} \approx (1 - \frac{\rho}{m})\frac{s_{\pi_j(h)}}{s}, \tag{6.3}$$

with $\mathbf{EC}_j$ denoting the mean cycle time of S_j.
However, it is considerably harder to approximate the *simultaneous* distribution of $(\mathbf{H}, \mathbf{Z}) = (\mathbf{H}_1, \ldots, \mathbf{H}_m, \mathbf{Z}_1, \ldots, \mathbf{Z}_m)$, which is actually needed in (6.1). There are three types of transitions in $(\mathbf{H}, \mathbf{Z})$.
First,

$$(h, z) \to (h + e_j, z - e_j), \qquad z_j = 1, \tag{6.4}$$

representing a departure of S_j from $Q_{\pi_j(h_j)}$ which does not result from an instantaneous passage; here e_j represents the j-th m-dimensional unit vector; $h_j + 1$ is in fact to be understood as $(h_j \bmod n) + 1$.
Second,

$$(h, z) \to (h, z + e_j), \qquad z_j = 0, \tag{6.5}$$

representing an arrival of S_j at $Q_{\pi_j(h_j)}$ which does not lead to an instantaneous passage.
Third,

$$(h, z) \to (h + e_j, z), \qquad z_j = 0, \tag{6.6}$$

representing an instantaneous passage of S_j at $Q_{\pi_j(h_j)}$.
Note that $\{(\mathbf{H}(t), \mathbf{Z}(t)), t \geq 0\}$ is *not* a Markov process, as the transitions are not independent of the past. To approximate the simultaneous distribution of $(\mathbf{H}, \mathbf{Z})$, we will however deal with the process as if it *is* Markov, i.e., as if the transitions in $(\mathbf{H}, \mathbf{Z})$ occur at a constant rate, independent of the past. The distribution of $(\mathbf{H}, \mathbf{Z})$ may then be determined, as soon as the transition rates $\mu_{(h,z)\to(h',z')}$ are specified, which we might do as follows.
First,

$$\mu_{(h,z)\to(h+e_j,z-e_j)} = (m - \rho)/(\rho_{\pi_j(h_j)} s), \qquad z_j = 1, \tag{6.7}$$

i.e., a departure of S_j from $Q_{\pi_j(h_j)}$ (which does not result from an instantaneous passage) occurs at a rate reciprocal to the approximate mean visit time of S_j at $Q_{\pi_j(h_j)}$ (i.e. $r_{\pi_j(h_j),j}\mathbf{EA}_{\pi_j(h_j),j} \approx \rho_{\pi_j(h_j)} s/(m - \rho)$).
Second,

$$\mu_{(h,z)\to(h,z+e_j)} = 1/s_{\pi_j(h_j)}, \qquad z_j = 0, \tag{6.8}$$

i.e., an arrival of S_j at $Q_{\pi_j(h_j)}$ (which does not lead to an instantaneous passage) occurs at a rate reciprocal to the mean switch-over time into $Q_{\pi_j(h_j)}$.
Third,

$$\mu_{(h,z)\to(h+e_j,z)} = 0, \qquad z_j = 0, \tag{6.9}$$

i.e., an instantaneous passage of S_j at $Q_{\pi_j(h_j)}$ (only occurring when there are no waiting customers at $Q_{\pi_j(h_j)}$, which cannot be deduced from $(\mathbf{H}, \mathbf{Z})$) does not occur. Note that in light traffic instantaneous passages in fact *do* frequently occur, as a server arrival is likely be attended by a concurrent server departure, which might suggest to replace (6.9) by

$$\mu_{(h,z)\to(h+e_j,z)} = 1/s_{\pi_j(h_j)}, \qquad z_j = 0, \tag{6.10}$$

when $\rho \downarrow 0$. However, when $\rho \downarrow 0$ in light traffic, combining (6.8) with (6.7) has a similar effect as using (6.10) would have.
Because of the homogeneity in the transition rates we would obtain from (6.7), (6.8), (6.9)

$$\Pr\{(\mathbf{H},\mathbf{Z}) = (h,z)\} = \prod_{j=1}^{m} \Pr\{(\mathbf{H}_j,\mathbf{Z}_j) = (h_j,z_j)\}, \tag{6.11}$$

where $\Pr\{(\mathbf{H}_j,\mathbf{Z}_j) = (h_j,z_j)\}$ satisfies (6.2), (6.3). In other words, we would obtain complete independence in the server position distribution, while in fact we attempted to capture the tendency for the servers to cluster.
The driving force behind the tendency for the servers to cluster is that the time that a server visits a queue depends on the time that the queue has not been visited by one of the other servers, so that the servers, somewhat depending on the visit orders, tend to be driven together. Once driven together, the servers do not disperse, as long as the visit orders do not direct them to different queues. To capture these phenomena, we slightly modify the transitions for the states in which more than one server is busy at the same queue simultaneously. For these states we replace the transitions where *one* server leaves the queue by a single transition of the same rate where *all* the visiting servers leave the queue simultaneously, reflecting that actually all the servers will tend to leave relatively shortly after one another.
The transition rates being specified, the distribution of $(\mathbf{H},\mathbf{Z})$ may then be determined by solving the balance equations, supplemented with the normalization condition. Because of the inhomogeneity introduced in the transition rates it is no longer possible to give the simultaneous distribution as explicitly as in (6.11), but it is easily verified from the balance equations that the marginal distribution $\Pr\{(\mathbf{H}_j,\mathbf{Z}_j) = (h_j,z_j)\}$ still satisfies (6.2), (6.3).

More detailed clustering measures
Remember that we approximated the distribution of $(\mathbf{H},\mathbf{Z})$ in the first place to determine the probabilities q_i that at least one of the servers is busy at Q_i, $i = 1,\ldots,n$. In their turn we used the probabilities q_i to determine $\alpha_i = \rho_i/q_i$, the average processing speed at Q_i, as a measure for the degree of clustering at Q_i. Having approximated the simultaneous distribution of $(\mathbf{H},\mathbf{Z})$, we may however refine the latter estimate for the degree of clustering, to deal with situations in which the average processing speed does not provide a good indication for the degree of clustering (e.g. in case of a lightly-loaded queue preceded by a heavily-loaded queue). In the remainder of the present section we briefly discuss the definition of those alternatives. In the next section, when testing the resulting waiting-time approximations, we will examine the impact of implementing these alternatives.
Denote by $P_{ij}^{(0)}(h,z)$ and $P_{ij}^{(1)}(h,z)$ the conditional probability that $(\mathbf{H},\mathbf{Z}) = (h,z)$ just after an arrival of S_j at Q_i and after a departure of S_j from Q_i, respectively. These conditional probabilities follow immediately from the distribution of $(\mathbf{H},\mathbf{Z})$. Denote by $\mathbf{T}_{ij}^{(0)}(h_j,z_j)$ and $\mathbf{T}_{ij}^{(1)}(h_j,z_j)$ the entrance time into $(\mathbf{H}_j,\mathbf{Z}_j) = (h_j,z_j)$ just after an arrival of S_j at Q_i and after a departure of S_j from Q_i, respectively. The mean values of these entrance times are

given by

$$\begin{aligned}\mathbf{ET}_{ij}^{(b)}(h_j,z_j) &= \frac{r_{ij}s}{1-r_j}(1-b)+\sum_{k=\pi_j^{-1}(i)+1}^{h_j-1}\left(s_{\pi_j(k)}+\frac{r_{\pi_j(k)j}s}{1-r_j}\right)+s_{\pi_j(h_j)}z_j \\ &\approx \frac{\rho_i s}{m-\rho}(1-b)+\sum_{k=\pi_j^{-1}(i)+1}^{h_j-1}\left(s_{\pi_j(k)}+\frac{\rho_{\pi_j(k)}s}{m-\rho}\right)+s_{\pi_j(h_j)}z_j,\end{aligned}$$

$b=0,1$, with $k=\pi_j^{-1}(i)$ such that $\pi_j(k)=i$. For given $(h,z)=(h_1,\ldots,h_m,z_1,\ldots,z_m)$, let $\mathbf{ET}_{ij_l}^{(b)}(h_{j_l},z_{j_l})$, $l=1,\ldots,m$, be the mean entrance times $\mathbf{ET}_{ij}^{(b)}(h_j,z_j)$, $j=1,\ldots,m$, ordered in decreasing magnitude. Let $\Delta_{il}^{(b)}(h,z)=\left(\mathbf{ET}_{ij_{l-1}}^{(b)}(h_{j_{l-1}},z_{j_{l-1}})-\mathbf{ET}_{ij_l}^{(b)}(h_{j_l},z_{j_l})\right)$, $l=1,\ldots,m$, with $\mathbf{ET}_{ij_0}^{(b)}(h_{j_0},z_{j_0})=\mathbf{EC}$, $\mathbf{EC}=s/(m-\rho)$. For given (h,z), $\Delta_{il}^{(b)}(h,z)$ represents the mean of the l-th of the m most recent server interarrival ($b=0$) or interdeparture ($b=1$) times at Q_i, $l=1,\ldots,m$. Denote $\Delta_i^{(b)}(h,z)=\sum_{l=1}^{m}\left(\Delta_{il}^{(b)}(h,z)\right)^2$. The ordinary sum of $\Delta_{il}^{(b)}(h,z)$, $l=1,\ldots,m$, being always equal to $\mathbf{EC}$, the sum of the *squares* provides a good indication for the *spacing* of the server arrivals at or departures from Q_i. Having this in mind, we define

$$\delta_i^{(b)}=\sum_{j=1}^{m}\sum_{(h,z)}P_{ij}^{(b)}(h,z)\Delta_i^{(b)}(h,z)/(\mathbf{EC})^2 \tag{6.12}$$

for $b=1$ and $b=0$ as a measure for the *local* degree of clustering at Q_i under the exhaustive and gated service discipline, respectively. If the degree of clustering at Q_i is high, then for the states (h,z) with large $P_{ij}^{(b)}(h,z)$, one of the $\Delta_{il}^{(b)}(h,z)$'s is approximately equal to $\mathbf{EC}$, while all the other $\Delta_{il}^{(b)}(h,z)$'s are approximately equal to 0, so that $\delta_i^{(b)}\approx m$. On the other hand, if the degree of clustering at Q_i is low, i.e., the $\Delta_{il}^{(b)}(h,z)$'s are the distances between approximately homogeneously distributed points on $[0,\mathbf{EC}]$, then $\delta_i^{(b)}\approx m\kappa$, with $\kappa=\int_{1=x_0\geq\ldots\geq x_m=0}\sum_{l=1}^{m}(x_{l-1}-x_l)^2dx_1\ldots dx_{m-1}$. If the servers even tend to repel each other, i.e, all the $\Delta_{il}^{(b)}(h,z)$'s are approximately equal to $\mathbf{EC}/m$, then $\delta_i^{(b)}\approx 1$.

We may also refine the measure (5.4) for the *global* degree of clustering. In the spirit of (6.12), we define

$$\delta=\frac{\sum_{(h,z)}\left(\Pr\{(\mathbf{H},\mathbf{Z})=(h,z)\}-\prod_{l=1}^{m}\Pr\{(\mathbf{H}_l,\mathbf{Z}_l)=(h_l,z_l)\}\right)\sum_{j=1}^{m}\Delta_{\pi_j(h_j)}^{(z_j)}(h,z)}{m(\mathbf{EC})^2-\sum_{(h,z)}\left(\prod_{l=1}^{m}\Pr\{(\mathbf{H}_l,\mathbf{Z}_l)=(h_l,z_l)\}\right)\sum_{j=1}^{m}\Delta_{\pi_j(h_j)}^{(z_j)}(h,z)}. \tag{6.13}$$

Note that for a high degree of clustering, i.e., for the states (h,z) with large $\Pr\{(\mathbf{H},\mathbf{Z})=(h,z)\}$, $\Delta_{\pi_j(h_j)}^{(z_j)}(h,z)\approx 1$, $\delta\uparrow 1$. On the other hand, for a low degree of clustering, i.e., $\Pr\{(\mathbf{H},\mathbf{Z})=(h,z)\}\approx\prod_{l=1}^{m}\Pr\{(\mathbf{H}_l,\mathbf{Z}_l)=(h_l,z_l)\}$, $\delta\downarrow 0$.

7 NUMERICAL RESULTS

We have tested the waiting-time approximations for a wide range of parameter combinations by comparison with either simulation results or exact numerical results obtained from the power-series algorithm (PSA), as elaborated upon in [24]. The large class of models considered in this paper has forced us to limit the examples that are used to demonstrate the accuracy of the approximations. Thus we have restricted ourselves to four-queue models with two servers and with exponentially distributed service and switch-over times. Numerous numerical experiments have indicated that for these models the accuracy of the approximations is acceptable, even for rather asymmetrical and heavily-loaded systems. In particular, the approximations rightly capture the clustering effects of the visit order, whereas most of the existing approximations completely ignore the considerable influence of the visit order on the waiting times.
The results of the numerical experiments are summarized below. Limited numerical experience (which is however not reported in any further detail below) suggests that the approximations perform similarly for non-exponentially distributed service and switch-over times). We reemphasize that the models considered here are very complex, containing single-server polling models and ordinary multiple-server models as special cases, while the visit order constitutes an additional complicating factor. The accuracy of the approximations should be judged from this perspective.

In order to test the accuracy of the approximations for a wide variety of models, we have considered various variants of a set of models in which the ratios between the arrival rates, (λ_1: λ_2: λ_3: λ_4) and the mean service times (β_1, β_2, β_3, β_4), respectively, are given as follows:

I. (1: 1: 1: 1); (1.0, 1.0, 1.0, 1.0);
II. (1: 1: 3: 3); (1.0, 1.0, 1.0, 1.0);
III. (2: 2: 5: 5); (1.0, 1.0, 0.4, 0.4);
IV. (1: 1: 1: 1); (0.5, 0.5, 1.5, 1.5);
V. (1: 1: 3: 3); (0.5, 1.5, 0.5, 1.5);
VI. (1: 1: 9: 1); (0.5, 0.5, 0.5, 2.5).

By convention, the queues are numbered such that π_1, the visit order of S_1, is always $(1,2,3,4)$. The visit order of S_2 is considered for the cases $\pi_2 = (1,2,3,4)$, $\pi_2 = (1,2,4,3)$, and $\pi_2 = (1,4,3,2)$. For each of the models, all switch-over times are assumed to have mean $s/n = s/4$ with either $s = 0$ or $s = 1$. (In evaluating the approximations we actually took $s = 10^{-6}$, as the formal definition of $(\mathbf{H}, \mathbf{Z})$ is restricted to the case of non-zero switch-over times.) The value of the total load is either $\rho = 0.8$, $\rho = 1.6$, or $\rho = 1.8$. In all considered cases we assume $m_i = m = 2$.

For the models listed above, Tables 7.1.A to 7.6.B show the results for exhaustive service at each of the queues. Tables 7.7.A to 7.8.B show the results for the models II and V with gated service. The rows indicated by 'α' contain the approximations obtained with $\alpha_i = \rho_i/q_i$ as a measure for the local degree of clustering at Q_i, and α as in (5.4) as a measure for the global degree of clustering. The rows marked with 'δ' give the approximations with α_i replaced by $\delta_i^{(1)}$ as in (6.12) for exhaustive service or $\delta_i^{(0)}$ for gated service, and α replaced by δ as in (6.13). The rows indicated by 'exact' contain the 'exact' mean waiting times obtained from

either the PSA (for exhaustive service) or simulation (for gated service). (We implemented the PSA only for Bernoulli service, including exhaustive service as special case, but in principle the method may also be developed to compute the mean waiting times for gated service.)

$(\lambda_1,\lambda_2,\lambda_3,\lambda_4)=\rho(0.25,0.25,0.25,0.25)$; $(\beta_1,\beta_2,\beta_3,\beta_4)=(1.0,1.0,1.0,1.0)$; $s=0.0$.				
		$(\mathbf{EW}_1,\mathbf{EW}_2,\mathbf{EW}_3,\mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.19, 0.19, 0.19, 0.19)	(0.19, 0.19, 0.19, 0.19)	(0.19, 0.19, 0.19, 0.19)
	α	(0.19, 0.19, 0.19, 0.19)	(0.19, 0.19, 0.19, 0.19)	(0.19, 0.19, 0.19, 0.19)
	δ	(0.19, 0.19, 0.19, 0.19)	(0.19, 0.19, 0.19, 0.19)	(0.19, 0.19, 0.19, 0.19)
1.6	exact	(1.78, 1.78, 1.78, 1.78)	(1.78, 1.85, 1.74, 1.74)	(1.78, 1.78, 1.78, 1.78)
	α	(1.78, 1.78, 1.78, 1.78)	(1.76, 1.93, 1.71, 1.71)	(1.78, 1.78, 1.78, 1.78)
	δ	(1.78, 1.78, 1.78, 1.78)	(1.78, 1.92, 1.71, 1.71)	(1.78, 1.78, 1.78, 1.78)
1.8	exact	(4.26, 4.26, 4.26, 4.26)	(4.41, 4.66, 3.98, 3.98)	(4.26, 4.26, 4.26, 4.26)
	α	(4.26, 4.26, 4.26, 4.26)	(4.19, 4.77, 4.04, 4.04)	(4.26, 4.26, 4.26, 4.26)
	δ	(4.26, 4.26, 4.26, 4.26)	(4.25, 4.71, 4.04, 4.04)	(4.26, 4.26, 4.26, 4.26)

Table 7.1.A The mean waiting times for Model I with $s=0.0$; exhaustive service.

$(\lambda_1,\lambda_2,\lambda_3,\lambda_4)=\rho(0.25,0.25,0.25,0.25)$; $(\beta_1,\beta_2,\beta_3,\beta_4)=(1.0,1.0,1.0,1.0)$; $s=1.0$.				
		$(\mathbf{EW}_1,\mathbf{EW}_2,\mathbf{EW}_3,\mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.77, 0.77, 0.77, 0.77)	(0.77, 0.77, 0.77, 0.77)	(0.76, 0.76, 0.76, 0.76)
	α	(0.78, 0.78, 0.78, 0.78)	(0.78, 0.78, 0.77, 0.77)	(0.78, 0.78, 0.78, 0.78)
	δ	(0.78, 0.78, 0.78, 0.78)	(0.77, 0.78, 0.77, 0.77)	(0.77, 0.77, 0.77, 0.77)
1.6	exact	(3.29, 3.29, 3.29, 3.29)	(3.24, 3.39, 3.09, 3.09)	(3.16, 3.16, 3.16, 3.16)
	α	(3.36, 3.36, 3.36, 3.36)	(3.14, 3.43, 3.05, 3.05)	(3.12, 3.12, 3.12, 3.12)
	δ	(3.52, 3.52, 3.52, 3.52)	(3.24, 3.49, 3.11, 3.11)	(3.14, 3.14, 3.14, 3.14)
1.8	exact	(7.55, 7.55, 7.55, 7.55)	(7.52, 7.86, 6.38, 6.38)	(6.87, 6.87, 6.87, 6.87)
	α	(7.58, 7.58, 7.58, 7.58)	(6.79, 7.72, 6.54, 6.54)	(6.75, 6.75, 6.75, 6.75)
	δ	(7.87, 7.87, 7.87, 7.87)	(7.09, 7.85, 6.74, 6.74)	(6.83, 6.83, 6.83, 6.83)

Table 7.1.B The mean waiting times for Model I with $s=1.0$; exhaustive service.

Tables 7.1.A and 7.1.B show that the approximations for Model I are very accurate. In particular Table 7.1.A confirms that for completely symmetric systems (including 'symmetric' visit order combinations, i.e., $\pi_2=(1,2,3,4)$ or $\pi_2=(1,4,3,2)$), the approximations are exact for exponentially distributed service times and zero switch-over times.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.0, 1.0, 1.0, 1.0)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.21, 0.21, 0.18, 0.18)	(0.21, 0.21, 0.18, 0.18)	(0.21, 0.21, 0.18, 0.18)
	α	(0.20, 0.20, 0.19, 0.19)	(0.20, 0.20, 0.19, 0.19)	(0.20, 0.20, 0.19, 0.19)
	δ	(0.21, 0.21, 0.18, 0.19)	(0.21, 0.21, 0.18, 0.18)	(0.21, 0.21, 0.18, 0.18)
1.6	exact	(2.24, 2.30, 1.63, 1.60)	(2.09, 2.26, 1.65, 1.65)	(2.17, 2.17, 1.65, 1.65)
	α	(2.03, 1.92, 1.67, 1.76)	(1.99, 2.06, 1.70, 1.70)	(2.00, 2.00, 1.70, 1.70)
	δ	(2.18, 2.13, 1.63, 1.68)	(2.05, 2.19, 1.66, 1.66)	(2.08, 2.08, 1.68, 1.68)
1.8	exact	(5.51, 5.92, 3.90, 3.75)	(4.82, 5.36, 4.00, 4.00)	(5.11, 5.11, 4.00, 4.00)
	α	(5.06, 4.83, 3.96, 4.12)	(4.83, 5.14, 4.02, 4.02)	(4.86, 4.86, 4.06, 4.06)
	δ	(5.41, 5.37, 3.86, 3.92)	(5.00, 5.43, 3.95, 3.95)	(5.04, 5.04, 4.00, 4.00)

Table 7.2.A The mean waiting times for Model II with $s = 0.0$; exhaustive service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.0, 1.0, 1.0, 1.0)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.83, 0.82, 0.74, 0.74)	(0.81, 0.81, 0.73, 0.73)	(0.81, 0.81, 0.74, 0.74)
	α	(0.82, 0.81, 0.75, 0.76)	(0.81, 0.81, 0.74, 0.74)	(0.81, 0.81, 0.75, 0.75)
	δ	(0.86, 0.85, 0.74, 0.75)	(0.83, 0.84, 0.73, 0.73)	(0.83, 0.83, 0.73, 0.73)
1.6	exact	(3.98, 4.04, 2.92, 2.88)	(3.58, 3.86, 2.88, 2.88)	(3.69, 3.69, 2.89, 2.89)
	α	(3.72, 3.53, 3.56, 3.23)	(3.42, 3.55, 2.92, 2.92)	(3.43, 3.43, 2.93, 2.93)
	δ	(4.23, 4.14, 3.16, 3.26)	(3.63, 3.87, 2.94, 2.94)	(3.65, 3.65, 2.95, 2.95)
1.8	exact	(9.04, 9.69, 6.53, 6.32)	(7.27, 8.04, 6.44, 6.44)	(7.73, 7.73, 6.45, 6.45)
	α	(8.69, 8.29, 6.79, 7.07)	(7.53, 8.01, 6.27, 6.27)	(7.54, 7.54, 6.30, 6.30)
	δ	(9.60, 9.43, 6.96, 7.12)	(8.05, 8.74, 6.36, 6.36)	(8.06, 8.06, 6.40, 6.40)

Table 7.2.B The mean waiting times for Model II with $s = 1.0$; exhaustive service.

Tables 7.2.A and 7.2.B show that the results are still accurate when the arrival rates are fairly asymmetrical, even for heavily-loaded systems.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.25, 0.25, 0.625, 0.625)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.0, 1.0, 0.4, 0.4)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.13, 0.12, 0.13, 0.13)	(0.13, 0.12, 0.13, 0.13)	(0.13, 0.13, 0.13, 0.13)
	α	(0.13, 0.13, 0.13, 0.13)	(0.13, 0.14, 0.13, 0.13)	(0.13, 0.13, 0.13, 0.13)
	δ	(0.13, 0.13, 0.13, 0.13)	(0.13, 0.14, 0.13, 0.13)	(0.13, 0.13, 0.13, 0.13)
1.6	exact	(1.25, 1.14, 1.22, 1.34)	(1.27, 1.20, 1.24, 1.24)	(1.21, 1.21, 1.26, 1.26)
	α	(1.24, 1.24, 1.24, 1.24)	(1.23, 1.35, 1.20, 1.20)	(1.24, 1.24, 1.24, 1.24)
	δ	(1.24, 1.24, 1.24, 1.24)	(1.25, 1.34, 1.20, 1.20)	(1.24, 1.24, 1.24, 1.24)
1.8	exact	(3.01, 2.75, 2.99, 3.22)	(3.15, 3.14, 2.83, 2.83)	(2.94, 2.94, 3.01, 3.01)
	α	(2.98, 2.98, 2.98, 2.98)	(2.94, 3.34, 2.83, 2.83)	(2.98, 2.98, 2.98, 2.98)
	δ	(2.98, 2.98, 2.98, 2.98)	(2.98, 3.30, 2.83, 2.83)	(2.98, 2.98, 2.98, 2.98)

Table 7.3.A The mean waiting times for Model III with $s = 0.0$; exhaustive service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.25, 0.25, 0.625, 0.625)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.0, 1.0, 0.4, 0.4)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.72, 0.70, 0.70, 0.72)	(0.71, 0.70, 0.70, 0.70)	(0.70, 0.70, 0.70, 0.70)
	α	(0.72, 0.72, 0.72, 0.72)	(0.72, 0.73, 0.72, 0.72)	(0.71, 0.71, 0.71, 0.71)
	δ	(0.72, 0.72, 0.72, 0.72)	(0.71, 0.72, 0.71, 0.71)	(0.71, 0.71, 0.71, 0.71)
1.6	exact	(2.80, 2.60, 2.79, 2.99)	(2.76, 2.72, 2.60, 2.60)	(2.59, 2.59, 2.65, 2.65)
	α	(2.83, 2.83, 2.83, 2.83)	(2.61, 2.85, 2.54, 2.54)	(2.59, 2.59, 2.59, 2.59)
	δ	(2.98, 2.98, 2.98, 2.98)	(2.71, 2.92, 2.60, 2.60)	(2.61, 2.61, 2.61, 2.61)
1.8	exact	(6.43, 5.97, 6.38, 6.70)	(6.42, 6.60, 5.24, 5.24)	(5.55, 5.55, 5.63, 5.63)
	α	(6.30, 6.30, 6.30, 6.30)	(5.53, 6.29, 5.33, 5.33)	(5.47, 5.47, 5.47, 5.47)
	δ	(6.59, 6.59, 6.59, 6.59)	(5.81, 6.43, 5.52, 5.52)	(5.55, 5.55, 5.55, 5.55)

Table 7.3.B The mean waiting times for Model III with $s = 1.0$; exhaustive service.

In the cases considered in Tables 7.3.A and 7.3.B the arrival rates and the service rates are rather asymmetrical, but the load offered to each of the queues is the same. By construction, the approximated ratios of the mean waiting times only depend on the λ_i's and β_i's through the ρ_i's. As the ρ_i's are all equal here, the approximated mean waiting times are also all equal for 'symmetric' visit order combinations, i.e., $\pi_2 = (1, 2, 3, 4)$ or $\pi_2 = (1, 4, 3, 2)$. The numerical results show that the *true* ratios of the mean waiting times *do* depend on the individual λ_i's and β_i's, but that the accuracy of the approximated mean waiting times is still acceptable.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.25, 0.25, 0.25, 0.25)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 0.5, 1.5, 1.5)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.25, 0.27, 0.23, 0.21)	(0.25, 0.27, 0.22, 0.22)	(0.26, 0.26, 0.22, 0.22)
	α	(0.25, 0.25, 0.23, 0.24)	(0.25, 0.26, 0.23, 0.23)	(0.25, 0.25, 0.23, 0.23)
	δ	(0.26, 0.26, 0.23, 0.23)	(0.26, 0.26, 0.23, 0.23)	(0.26, 0.26, 0.23, 0.23)
1.6	exact	(2.79, 3.07, 2.05, 1.88)	(2.55, 2.92, 2.03, 2.03)	(2.74, 2.74, 2.03, 2.03)
	α	(2.53, 2.40, 2.08, 2.20)	(2.48, 2.57, 2.12, 2.12)	(2.50, 2.50, 2.13, 2.13)
	δ	(2.72, 2.66, 2.03, 2.10)	(2.57, 2.74, 2.08, 2.08)	(2.60, 2.60, 2.10, 2.10)
1.8	exact	(7.00, 7.58, 4.84, 4.49)	(5.90, 6.77, 4.99, 4.99)	(6.38, 6.38, 4.96, 4.96)
	α	(6.33, 6.03, 4.95, 5.15)	(6.04, 6.43, 5.03, 5.03)	(6.08, 6.08, 5.08, 5.08)
	δ	(6.68, 6.56, 4.84, 4.96)	(6.25, 6.78, 4.93, 4.93)	(6.30, 6.30, 5.00, 5.00)

Table 7.4.A The mean waiting times for Model IV with $s = 0.0$; exhaustive service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.25, 0.25, 0.25, 0.25)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 0.5, 1.5, 1.5)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.87, 0.88, 0.79, 0.78)	(0.85, 0.88, 0.78, 0.78)	(0.86, 0.86, 0.78, 0.78)
	α	(0.87, 0.86, 0.79, 0.81)	(0.86, 0.87, 0.79, 0.79)	(0.86, 0.86, 0.79, 0.79)
	δ	(0.91, 0.90, 0.78, 0.80)	(0.88, 0.89, 0.78, 0.78)	(0.89, 0.89, 0.78, 0.78)
1.6	exact	(4.56, 4.91, 3.35, 3.13)	(4.05, 4.53, 3.28, 3.28)	(4.28, 4.28, 3.28, 3.28)
	α	(4.23, 4.01, 3.47, 3.67)	(3.92, 4.06, 3.35, 3.35)	(3.93, 3.93, 3.36, 3.36)
	δ	(4.78, 4.67, 3.56, 3.68)	(4.14, 4.42, 3.35, 3.35)	(4.17, 4.17, 3.37, 3.37)
1.8	exact	(10.55, 11.20, 7.23, 7.00)	(8.34, 9.51, 7.41, 7.41)	(9.04, 9.04, 7.39, 7.39)
	α	(9.95, 9.49, 7.78, 8.10)	(8.74, 9.30, 7.27, 7.27)	(8.75, 8.75, 7.32, 7.32)
	δ	(10.93, 10.75, 7.92, 8.11)	(9.30, 10.10, 7.35, 7.35)	(9.32, 9.32, 7.40, 7.40)

Table 7.4.B The mean waiting times for Model IV with $s = 1.0$; exhaustive service.

In Model IV the service times are asymmetrical, whereas the arrival rates are the same. Tables 7.4.A and 7.4.B show that the accuracy of the results is acceptable, even in heavily-loaded systems.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 1.5, 0.5, 1.5)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.27, 0.25, 0.26, 0.21)	(0.28, 0.26, 0.26, 0.21)	(0.27, 0.25, 0.26, 0.21)
	α	(0.27, 0.25, 0.25, 0.23)	(0.27, 0.25, 0.25, 0.23)	(0.26, 0.26, 0.25, 0.23)
	δ	(0.29, 0.26, 0.26, 0.22)	(0.28, 0.26, 0.26, 0.22)	(0.27, 0.27, 0.26, 0.22)
1.6	exact	(3.24, 2.65, 2.99, 1.67)	(3.13, 2.84, 2.76, 1.71)	(3.02, 2.81, 2.78, 1.72)
	α	(2.75, 2.63, 2.61, 1.90)	(2.72, 2.65, 2.48, 1.94)	(2.59, 2.72, 2.42, 1.95)
	δ	(3.20, 2.75, 2.76, 1.76)	(3.08, 2.82, 2.54, 1.82)	(2.65, 2.91, 2.50, 1.85)
1.8	exact	(8.26, 6.68, 7.53, 3.73)	(7.51, 7.17, 6.59, 4.03)	(7.12, 7.14, 6.61, 4.06)
	α	(7.01, 6.64, 6.62, 4.28)	(6.77, 6.74, 6.18, 4.42)	(6.36, 6.87, 6.04, 4.46)
	δ	(7.97, 6.84, 6.86, 4.02)	(7.74, 7.08, 6.34, 4.14)	(6.50, 7.29, 6.26, 4.24)

Table 7.5.A The mean waiting times for Model V with $s = 0.0$; exhaustive service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 1.5, 0.5, 1.5)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.91, 0.85, 0.86, 0.74)	(0.90, 0.86, 0.85, 0.74)	(0.89, 0.86, 0.85, 0.74)
	α	(0.89, 0.85, 0.85, 0.75)	(0.88, 0.85, 0.84, 0.75)	(0.88, 0.85, 0.84, 0.75)
	δ	(0.98, 0.89, 0.90, 0.74)	(0.94, 0.88, 0.86, 0.73)	(0.91, 0.89, 0.85, 0.73)
1.6	exact	(5.12, 4.28, 4.68, 2.69)	(4.82, 4.42, 4.19, 2.74)	(4.58, 4.41, 4.22, 2.76)
	α	(4.43, 4.23, 4.20, 3.06)	(4.19, 4.08, 3.81, 2.98)	(3.99, 4.19, 3.73, 3.00)
	δ	(5.43, 4.67, 4.69, 2.98)	(4.88, 4.47, 4.03, 2.89)	(4.19, 4.60, 3.95, 2.93)
1.8	exact	(11.94, 10.95, 11.09, 5.44)	(10.56, 10.29, 9.15, 5.89)	(9.88, 10.49, 9.27, 5.89)
	α	(10.59, 10.03, 10.01, 6.47)	(9.53, 9.48, 8.69, 6.22)	(8.95, 9.67, 8.50, 6.28)
	δ	(12.56, 10.79, 10.81, 6.34)	(11.28, 10.32, 9.24, 6.03)	(9.43, 10.57, 9.08, 6.15)

Table 7.5.B The mean waiting times for Model V with $s = 1.0$; exhaustive service.

In the models considered in Tables 7.5.A and 7.5.B the arrival rates as well as the service times are asymmetrical. It is shown that in these cases the approximations are less accurate than in the cases considered above, but still acceptable. In Models I-IV both approximations yielded similar results, but here the δ-approximation tends to outperform the α-approximation. Apparently the latter fails to detect the clustering at the lightly-loaded queues that are visited after the heavily-loaded Q_4.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 1.125, 0.125)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 0.5, 0.5, 2.5)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.21, 0.22, 0.18, 0.19)	(0.23, 0.23, 0.18, 0.19)	(0.23, 0.23, 0.18, 0.19)
	α	(0.24, 0.24, 0.20, 0.23)	(0.24, 0.24, 0.21, 0.22)	(0.24, 0.24, 0.21, 0.22)
	δ	(0.26, 0.26, 0.20, 0.23)	(0.26, 0.26, 0.20, 0.22)	(0.26, 0.25, 0.20, 0.22)
1.6	exact	(2.96, 3.17, 1.65, 2.04)	(2.99, 3.25, 1.62, 2.08)	(3.13, 3.12, 1.62, 2.08)
	α	(2.45, 2.31, 1.70, 2.39)	(2.43, 2.47, 1.77, 2.23)	(2.48, 2.41, 1.77, 2.23)
	δ	(2.94, 2.88, 1.61, 2.34)	(2.60, 2.75, 1.70, 2.28)	(2.76, 2.54, 1.70, 2.28)
1.8	exact	(7.94, 8.50, 3.78, 4.91)	(7.54, 8.18, 3.83, 5.05)	(8.05, 7.75, 3.84, 5.07)
	α	(9.14, 8.57, 5.21, 8.22)	(8.29, 8.55, 5.46, 7.92)	(8.50, 8.18, 5.47, 7.93)
	δ	(10.44, 10.30, 5.00, 8.01)	(8.88, 9.43, 5.21, 8.09)	(9.39, 8.66, 5.23, 8.11)

Table 7.6.A The mean waiting times for Model VI with $s = 0.0$; exhaustive service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 1.125, 0.125)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 0.5, 0.5, 2.5)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.84, 0.84, 0.69, 0.79)	(0.86, 0.86, 0.69, 0.77)	(0.86, 0.85, 0.69, 0.77)
	α	(0.86, 0.85, 0.72, 0.81)	(0.86, 0.86, 0.73, 0.79)	(0.86, 0.86, 0.73, 0.79)
	δ	(0.95, 0.94, 0.72, 0.84)	(0.92, 0.92, 0.72, 0.80)	(0.93, 0.90, 0.72, 0.80)
1.6	exact	(4.69, 5.05, 2.67, 3.35)	(4.63, 4.95, 2.62, 3.36)	(4.93, 4.86, 2.62, 3.34)
	α	(4.05, 3.81, 2.80, 3.95)	(3.85, 3.90, 2.80, 3.52)	(3.93, 3.81, 2.80, 3.52)
	δ	(5.19, 5.09, 2.84, 4.13)	(4.29, 4.52, 2.80, 3.75)	(4.55, 4.18, 2.80, 3.75)
1.8	exact	(11.44, 12.45, 5.63, 5.70)	(9.99, 11.10, 5.70, 7.36)	(10.96, 9.97, 5.74, 7.36)
	α	(13.91, 13.04, 7.93, 12.50)	(11.71, 12.08, 7.71, 11.18)	(11.99, 11.53, 7.72, 11.18)
	δ	(16.66, 16.43, 7.98, 12.78)	(13.19, 14.02, 7.75, 12.03)	(13.94, 12.85, 7.76, 12.04)

Table 7.6.B The mean waiting times for Model VI with $s = 1.0$; exhaustive service.

Model VI is a typical example of a very asymmetrical system. In such cases, the accuracy of all waiting-time approximations in the literature degrades significantly, even in single-server systems. Tables 7.6.A and 7.6.B show that the accuracy of the waiting-time approximation presented in this paper also degrades when the model is very asymmetrical, but remains acceptable as long as the load is not too high.

We now check the accuracy of the approximation for multiple-server systems with gated service at all queues.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.0, 1.0, 1.0, 1.0)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.18, 0.17, 0.19, 0.20)	(0.18, 0.17, 0.19, 0.20)	(0.18, 0.18, 0.19, 0.19)
	α	(0.16, 0.16, 0.20, 0.20)	(0.16, 0.16, 0.20, 0.20)	(0.16, 0.16, 0.20, 0.20)
	δ	(0.17, 0.17, 0.20, 0.20)	(0.17, 0.17, 0.20, 0.20)	(0.17, 0.17, 0.20, 0.20)
1.6	exact	(1.54, 1.52, 1.82, 1.85)	(1.46, 1.51, 1.85, 1.86)	(1.46, 1.46, 1.86, 1.86)
	α	(1.38, 1.33, 1.90, 1.94)	(1.29, 1.32, 1.94, 1.94)	(1.29, 1.29, 1.94, 1.94)
	δ	(1.56, 1.54, 1.84, 1.87)	(1.44, 1.53, 1.88, 1.88)	(1.44, 1.44, 1.89, 1.89)
1.8	exact	(3.60, 3.59, 4.34, 4.40)	(3.22, 3.42, 4.46, 4.46)	(3.29, 3.29, 4.47, 4.47)
	α	(3.40, 3.28, 4.53, 4.61)	(3.03, 3.16, 4.65, 4.65)	(3.03, 3.03, 4.67, 4.67)
	δ	(3.72, 3.67, 4.43, 4.48)	(3.36, 3.63, 4.52, 4.52)	(3.36, 3.36, 4.56, 4.56)

Table 7.7.A The mean waiting times for Model II with $s = 0.0$; gated service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.0, 1.0, 1.0, 1.0)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.84, 0.83, 0.91, 0.92)	(0.83, 0.83, 0.91, 0.91)	(0.83, 0.83, 0.91, 0.91)
	α	(0.75, 0.75, 0.95, 0.96)	(0.75, 0.75, 0.95, 0.95)	(0.75, 0.75, 0.95, 0.95)
	δ	(0.83, 0.82, 0.94, 0.95)	(0.80, 0.81, 0.93, 0.93)	(0.80, 0.80, 0.93, 0.93)
1.6	exact	(3.85, 3.79, 4.56, 4.61)	(3.27, 3.44, 4.28, 4.28)	(3.35, 3.35, 4.35, 4.35)
	α	(3.25, 3.12, 4.47, 4.58)	(2.80, 2.88, 4.21, 4.21)	(2.80, 2.80, 4.21, 4.21)
	δ	(3.95, 3.88, 4.66, 4.73)	(3.22, 3.43, 4.22, 4.22)	(3.22, 3.22, 4.21, 4.21)
1.8	exact	(8.38, 8.31, 10.15, 10.22)	(6.47, 6.88, 9.33, 9.37)	(6.94, 6.94, 9.84, 9.84)
	α	(7.69, 7.42, 10.23, 10.42)	(6.00, 6.26, 9.22, 9.22)	(5.96, 5.96, 9.19, 9.19)
	δ	(8.91, 8.81, 10.62, 10.74)	(6.96, 7.53, 9.37, 9.37)	(6.89, 6.89, 9.37, 9.37)

Table 7.7.B The mean waiting times for Model II with $s = 1.0$; gated service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 1.5, 0.5, 1.5)$; $s = 0.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.20, 0.21, 0.21, 0.25)	(0.20, 0.22, 0.21, 0.24)	(0.20, 0.21, 0.21, 0.24)
	α	(0.17, 0.20, 0.20, 0.27)	(0.17, 0.20, 0.19, 0.27)	(0.17, 0.21, 0.19, 0.27)
	δ	(0.20, 0.21, 0.22, 0.26)	(0.20, 0.22, 0.21, 0.26)	(0.19, 0.22, 0.21, 0.26)
1.6	exact	(1.71, 1.91, 1.86, 2.46)	(1.68, 1.94, 1.81, 2.52)	(1.54, 1.91, 1.80, 2.49)
	α	(1.39, 1.74, 1.73, 2.64)	(1.28, 1.68, 1.61, 2.71)	(1.23, 1.72, 1.60, 2.71)
	δ	(1.77, 1.93, 1.93, 2.47)	(1.64, 1.91, 1.74, 2.55)	(1.40, 1.95, 1.72, 2.57)
1.8	exact	(3.91, 4.45, 4.36, 5.87)	(3.76, 4.54, 4.04, 6.18)	(3.35, 4.62, 4.11, 6.37)
	α	(3.46, 4.28, 4.28, 6.24)	(2.96, 4.04, 3.84, 6.52)	(2.83, 4.10, 3.80, 6.52)
	δ	(4.19, 4.60, 4.59, 5.95)	(3.85, 4.55, 4.12, 6.16)	(3.21, 4.62, 4.08, 6.22)

Table 7.8.A The mean waiting times for Model V with $s = 0.0$; gated service.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \rho(0.125, 0.125, 0.375, 0.375)$; $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 1.5, 0.5, 1.5)$; $s = 1.0$.				
		$(\mathbf{EW}_1, \mathbf{EW}_2, \mathbf{EW}_3, \mathbf{EW}_4)$		
ρ	π_2	(1, 2, 3, 4)	(1, 2, 4, 3)	(1, 4, 3, 2)
0.8	exact	(0.85, 0.87, 0.89, 0.99)	(0.85, 0.87, 0.88, 0.98)	(0.84, 0.88, 0.90, 1.00)
	α	(0.70, 0.80, 0.80, 1.12)	(0.69, 0.80, 0.79, 1.11)	(0.69, 0.80, 0.79, 1.11)
	δ	(0.84, 0.89, 0.90, 1.07)	(0.81, 0.88, 0.85, 1.06)	(0.78, 0.88, 0.85, 1.06)
1.6	exact	(3.75, 4.18, 4.14, 5.42)	(3.51, 4.11, 3.73, 5.39)	(3.16, 4.09, 3.76, 5.40)
	α	(2.95, 3.69, 3.68, 5.61)	(2.54, 3.36, 3.22, 5.40)	(2.46, 3.43, 3.19, 5.41)
	δ	(4.06, 4.42, 4.42, 5.65)	(3.41, 3.98, 3.62, 5.32)	(2.91, 4.05, 3.56, 5.33)
1.8	exact	(8.56, 9.66, 9.56, 12.74)	(7.15, 8.84, 7.54, 12.18)	(6.26, 9.10, 7.79, 12.71)
	α	(7.07, 8.76, 8.75, 12.76)	(5.42, 7.41, 7.05, 11.95)	(5.18, 7.52, 6.98, 11.96)
	δ	(9.12, 10.02, 10.00, 12.95)	(7.45, 8.80, 7.97, 11.92)	(6.17, 8.88, 7.84, 11.95)

Table 7.8.B The mean waiting times for Model V with $s = 1.0$; gated service.

Tables 7.7.A to 7.8.B show similar results as for the corresponding models with exhaustive service: the accuracy is acceptable for systems which are not too asymmetrical, even for heavily-loaded systems in which the switch-over times are significant.

Discussion of the numerical results

We have tested the accuracy of the waiting-time approximations extensively for a broad set of parameter combinations, viz., for lightly-, medium- and heavily-loaded systems, with symmetrical and asymmetrical arrival and service rates, with negligible and non-negligible switch-over times, and with varying visit order combinations. In general, the results are fairly accurate. Because the clustering effects of the visit order are explicitly taken into account, the results are still accurate in cases where the server bunching is significant, whereas most of the existing approximations completely ignore the influence of the visit order on the waiting times.

As discussed extensively in sections 4, 5, and 6, the waiting-time approximations presented in this paper are based on a series of assumptions, each of which, by definition, forms a source of inaccuracy. The first source of inaccuracy stems from the estimation of the ratios between the mean waiting times which, in turn, is composed of a number of approximations for (i) the mean waiting times in terms of the mean residual cycle times (cf. (4.5), (4.6)), (ii) the ratios between the mean residual cycle times (cf. (4.7), (4.8)), and (iii) the value of q_i (cf. (6.1)). The second error source is the estimation of the mean amount of work in the system (cf. (5.5)). Extensive simulation experiments have been performed to check the impact of each of these error sources on the finally observed error in the approximated mean waiting times.

Inspection of the numerical results has revealed that the estimation of the mean amount of work in the system, $\mathbf{EV}$, according to (5.5), is rather accurate. The error in the estimation of $\mathbf{EV}$ is typically less than 5% in fairly symmetrical systems, even under heavy traffic, and remains well below 10% for rather asymmetrical systems, even when the offered load is high. The main source of inaccuracy stems from the estimation of the ratios between the mean waiting times. The approximation of q_i, i.e., the probability that at least one of the servers is busy at Q_i, is quite accurate in many cases, also when the clustering effect is significant, with

errors typically below 10%. We found that q_i is underestimated in most of the cases. This is probably due to the fact that the clustering effect in the approximative approach is somewhat exaggerated because of the assumption that all the visiting servers depart from a queue simultaneously. The approximation of q_i may become inaccurate for very asymmetrical systems under a heavy-traffic scenario. Other inaccuracies stem from the approximation of the mean waiting times in terms of the mean residual cycle times (cf. (4.5), (4.6)). For systems with the exhaustive service discipline, the mean waiting times are usually underestimated according to (4.5) (where q_i and $\mathbf{ERD}_i$ are taken to be their respective true (simulated) values), whereas in case of the gated service discipline, the mean waiting times are somewhat overestimated according to (4.6). Apparently, in the case of exhaustive service the approximation (4.5) is too optimistic and, in the case of gated service, the approximation (4.6) is rather pessimistic. However, in both cases the *ratios* between the overestimated mean waiting times appear to be rather robust with respect to these errors, so that the errors resulting from (4.5) and (4.6) only have a marginal impact on the finally obtained waiting-time approximations. The ratios between the mean residual cycle times are estimated by the ratios between the estimated average processing rates according to (4.7) and (4.8). Numerical experience has taught us that the quality of these estimations is quite good, with errors typically up to 10%, except for very asymmetrical systems under heavy traffic, in which cases the accuracy of the approximations (4.7) and (4.8) may degrade significantly.

In the general approach developed in sections 4, 5, and 6 we used $\alpha_i = \rho_i/q_i$ as a measure for the local degree of clustering and α in (5.4) as a measure for the global degree of clustering, where both degrees of clustering are based on estimation of the average processing speed. However, for situations in which the average processing speed does not provide a good indication for the degree of clustering, we defined in the second part of section 6 alternative clustering measures, viz., sum-of-square-like spacing measures for the positions of the servers in the system (cf. (6.12), (6.13)). Comparing the accuracy of the waiting-time approximations based on both clustering measures (in the tables indicated by α and δ) has not indicated a clear superiority of one of the two measures; the α-approximation is 122 out of the 468 times more than 10% off; the δ-approximation 96 times.

Summarizing, in general the approximations presented in this paper lead to fairly accurate results when the system load is not too high and the system parameters are not too asymmetrical. Apparently, the approximations cover the main characteristics of the extremely complicated behavior of multiple-server polling systems. When the system is very asymmetrical and the offered load is very high, the accuracy of the approximations may however degrade significantly.

8 Concluding remarks and suggestions for further research

In this paper we have presented waiting-time approximations for asymmetric multiple-server polling systems with the exhaustive and gated service discipline, in which each of the servers visits the queues according to its own cyclic schedule. It is known that the servers have the tendency to coalesce, especially in heavily-loaded systems in which the servers follow the same route. While most of the existing waiting-time approximations completely ignore the clustering effects, we have explicitly taken the server bunching into account by approximating

the evolution of the joint server-position $(\mathbf{H}, \mathbf{Z})$ as a continuous-time Markov process. The approximations have been tested for a broad set of system parameters and have been found to be fairly accurate in the vast majority of the cases.
In the present paper we have focused on the case $m_i = m$, i.e., all the m servers may visit Q_i simultaneously. It would be interesting to derive waiting-time approximations for the case $m_i < m$, in particular for $m_i = 1$. In some respects the analysis will be somewhat facilitated then. Formulae (4.5), (4.6) e.g. are exact for $m_i = 1$. Also the probabilities q_i that show up in these formulae are simply known to be ρ_i for $m_i = 1$. In certain other respects the analysis will however be more complicated. The average processing speed $\alpha_i = \rho_i/q_i$ will always be equal to 1 for $m_i = 1$, so that it can no longer be used as a measure for the degree of clustering at Q_i. The more detailed measures $\delta_i^{(b)}$ can still be used. It will however be harder to approximate the simultaneous distribution of $(\mathbf{H}, \mathbf{Z})$ needed to determine these measures, as for $m_i < m$ there are also instantaneous passages through states with more than m_i servers at Q_i. Also the derivation of an approximative pseudo-conservation law will be considerably harder.
In the present paper we have considered systems in which each of the servers visits the queues cyclically, in a fixed order, and where the switch-over times only depend on the next queue to be visited. It would be interesting to explore the same ideas to derive approximations for the mean waiting times in multiple-server polling systems with a non-cyclic polling table, with probabilistic server routing, or where the switch-over times also depend on the previous queue visited.
The ultimate goal of performance modeling is in fact efficient operation and optimization. The development of accurate waiting-time approximations may be very useful for (approximately) solving a variety of optimization problems for multiple-server polling systems, opening up a very interesting area for further research.

Acknowledgement The authors are grateful to J.P.C. Blanc and O.J. Boxma for several valuable discussions and useful suggestions.

REFERENCES

[1] Ajmone Marsan, M., Donatelli, S., Neri, F. (1990). GSPN models of Markovian multiserver multiqueue systems. *Perf. Eval.* **11**, 227-240.

[2] Ajmone Marsan, M., Donatelli, S., Neri, F. (1991). Multiserver multiqueue systems with limited service and zero walk time. In: *Proc. INFOCOM '91*, 1178-1188.

[3] Ajmone Marsan, M., De Moraes, L.F., Donatelli, S., Neri, F. (1990). Analysis of symmetric nonexhaustive polling with multiple servers. In: *Proc. INFOCOM '90*, 284-295.

[4] Ajmone Marsan, M., De Moraes, L.F., Donatelli, S., Neri, F. (1992). Cycles and waiting times in symmetric exhaustive and gated multiserver multiqueue systems. In: *Proc. INFOCOM '92*, 2315-2324.

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

[6] Bhuyan, L.N., Ghosal, D., Yang, Q. (1989). Approximate analysis of single and multiple ring networks. *IEEE Trans. Comp.* **38**, 1027-1040.

[7] Borst, S.C. (1994). Polling systems with multiple coupled servers. CWI Report BS-R9408.

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

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

[10] Browne, S., Coffman, E.G. jr., Gilbert, E.N., Wright, P.E.W. (1992). Gated, exhaustive, parallel service. *Prob. Eng. Inf. Sc.* **6**, 217-239.

[11] Browne, S., Kella, O. (1992). Parallel service with vacations. To appear in *Oper. Res.*

[12] Browne, S., Weiss, G. (1992). Dynamic priority rules when polling with multiple parallel servers. *Oper. Res. Lett.* **12**, 129-137.

[13] Brumelle, S.L. (1971). On the relation between customer and time averages in queues. *J. Appl. Prob.* **8**, 508-520.

[14] Bux, W., Truong, H.L. (1983). Mean-delay approximations for cyclic-service queueing systems. *Perf. Eval.* **3**, 187-196.

[15] Everitt, D.E. (1986). Simple approximations for token rings. *IEEE Trans. Comm.* **34**, 719-721.

[16] Franken, P., König, D., Arndt, U., Schmidt, V. (1982). *Queues and Point Processes* (Wiley, New-York).

[17] Gamse, B., Newell, G.F. (1982). An analysis of elevator operation in moderate height buildings - II. Multiple elevators. *Transp. Res., B* **16**, 321-335.

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

[19] Kamal, A.E., Hamacher, V.C. (1989). Approximate analysis of non-exhaustive multi-server polling systems with applications to local area networks. In: *Comp. Netw. ISDN Syst.* **17**, 15-27.

[20] Kao, E.P.C., Narayanan, K.S. (1991). Analyses of an $M/M/N$ queue with servers' vacations. *EJOR* **54**, 256-266.

[21] Karmarkar, V.V., Kuhl, J.G. (1989). An integrated approach to distributed demand assignment in multiple-bus local networks. *IEEE Trans. Comp.* **38**, 679-695.

[22] Levy, Y., Yechiali, U. (1976). An $M/M/s$ queue with servers' vacations. *INFOR* **14**, 153-163.

[23] Loucks, W.M., Hamacher, V.C., Preiss, B.R., Wong, L. (1985). Short-packet transfer performance in local area ring networks. *IEEE Trans. Comp.* **34**, 1006-1014.

[24] Mei, R.D. van der, Borst, S.C. (1994). Analysis of multiple-server polling systems by means of the power-series algorithm. CWI Report BS-R9410.

[25] Mitrany, I.L., Avi-Itzhak, B. (1968). A many-server queue with server interruptions. *Oper. Res.* **16**, 628-638.

[26] Morris, R.J.T., Wang, Y.T. (1984). Some results for multi-queue systems with multiple cyclic servers. In: *Performance of Computer-Communication Systems*, eds. W. Bux and H. Rudin (North-Holland, Amsterdam), 245-258.

[27] Neuts, M.F., Lucantoni, D.M. (1979). A Markovian queue with N servers subject to breakdowns and repairs. *Mgmt. Sc.* **25**, 849-861.

[28] Raith, T. (1985). Performance analysis of multibus interconnection networks in distributed systems. In: *Teletraffic Issues in an Advanced Information Society, Proc. ITC-11*, ed. M. Akiyama (North-Holland, Amsterdam), 662-668.

[29] Yang, Q., Ghosal, D., Bhuyan, L.N. (1986). Performance analysis of multiple token ring and multiple slotted ring networks. In: *Proc. 1986 Comp. Netw. Symp.*, 79-86.

[30] Zafirovic-Vukotic, M., Niemegeers, I.G., Valk, D.S. (1988). Performance modelling of slotted ring protocols in HSLAN's. *IEEE J. Sel. Areas Comm.* **6**, 1001-1024.