ELSEVIER

Performance Evaluation 31 (1998) 163–182

PERFORMANCE EVALUATION
An International Journal

Waiting-time approximations for multiple-server polling systems ★

S.C. Borst [a,*], R.D. van der Mei [b]

[a] *Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, P.O. Box 636, Murray Hill, NJ 07974-0636, USA*
[b] *AT&T Labs, 101 Crawfords Corner Road, P.O. Box 3030, Holmdel, NJ 07733-3030, USA*

Received 3 October 1995; revised 15 July 1996

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 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. © 1998 Elsevier Science B.V.

Keywords: Cyclic service; Polling system; Multiple servers; Waiting-time approximations

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. 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.

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 load offered to the system, cf. [28]. 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 batch is served, the back-end system picks up the jobs from the next front-end system.

★ The work was done while the authors were with CWI – Center for Mathematics and Computer Science, Amsterdam, The Netherlands, and with Tilburg University, Center for Economic Research, The Netherlands, respectively.

* Corresponding author. E-mail: sem@research.bell-labs.com.

PII S0166-5316(96)00063-6

Examples also arise in communication networks, like the underlying communication medium in the above-mentioned distributed system (cf. also Takagi [32]). 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 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.

Multiple-server polling systems have received remarkably little attention in the vast literature on polling systems (cf. Takagi [31] for a comprehensive survey). One of the first studies is Morris and Wang [28] 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. A very interesting phenomenon observed by Morris and Wang is the tendency for the servers to cluster if they follow identical routes, especially in heavy traffic, cf. also [25]. 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 and Wang advocate the use of 'dispersive' schedules to improve the system performance. Levy et al. [22] propose bang–bang policies to avoid the bunching of servers.

Levy and Yechiali [23] and Kao and Narayanan [20] study a Markovian multiple-server queue, where the servers individually go on vacation when there are no waiting customers left. Mitrany and Avi-Itzhak [27] and Neuts and Lucantoni [29] analyze a Markovian multiple-server queue, where servers break down at exponential intervals and then get repaired.

In [6,19,21,30], mean response time approximations are developed to analyze the performance of LANs with multiple-token rings. Mean response time approximations oriented to LANs with a multiple-slotted ring are contained in [5,6,24,33]. Ajmone Marsan et al. [2–4] 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 and Weiss [13] 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 cycle length. Browne et al. [11] examine a completely symmetric two-queue system with an infinite number of coupled servers and deterministic service times. Browne and Kella [12] consider 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. Borst [7] explores the class of systems that allow an exact analysis in the case of coupled servers. Van der Mei and Borst [25] 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. Only the studies [20,23,27,29], considering single-queue systems, and [7,11–13], focusing on a limited class of models with *coupled* servers, present any exact results. To the best of the authors' knowledge, there are no exact results known for models with *independent* servers, 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 asymmetric systems with the exhaustive and gated service discipline, in which each of the servers visits the queues according to its own cyclic schedule.

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 be repeatedly used throughout the subsequent sections. In

Sections 4–6, 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 law-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. 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 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 occurred 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 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. 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 be repeatedly used throughout the subsequent sections. 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 carried and 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,

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

independent of i. Since $p_{ij} = \mathrm{E}\mathbf{A}_{ij}/\mathrm{E}\mathbf{A}_{ij}^*$,

$$\mathrm{E}\mathbf{A}_{ij}^* = \frac{s/p_{ij}}{1 - r_j}. \tag{2}$$

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 should carry the same load. 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 marginally, although in case of highly asymmetric system configurations, the differences may slightly increase. However, as observed in [25,28], 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 (1) 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 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 rule out 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

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

So, from (1),

$$\mathrm{E}\mathbf{A}_i = \frac{s}{m - \rho}, \tag{3}$$

which is again like (1) independent of i. Moreover, the mean server interarrival time is completely insensitive to how the total load is divided among the individual servers. Since $p_i = \mathrm{E}\mathbf{A}_i/\mathrm{E}\mathbf{A}_i^*$,

$$\mathrm{E}\mathbf{A}_i^* = \frac{s/p_i}{m-\rho}, \quad i = 1, \ldots, n. \tag{4}$$

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. 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.,

$$\Pr\{\mathbf{RX} < t\} = \frac{1}{\mathrm{E}\mathbf{X}} \int_{u=0}^{t} (1 - \Pr\{\mathbf{X} < u\})\,\mathrm{d}u, \quad t \geq 0.$$

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

$$\mathrm{E}\mathbf{W}_i \approx \gamma_i \mathrm{E}\mathbf{RC}_i, \quad i = 1, \ldots, n, \tag{5}$$

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,

$$\mathrm{E}\mathbf{W}_i = (1 - \rho_i)\mathrm{E}\mathbf{RD}_i, \quad i = 1, \ldots, n, \tag{6}$$

with $\mathbf{D}_i$ the server inter*departure* time at Q_i, cf. [16,18]. (An alternative relationship for the exhaustive service discipline is $\mathrm{E}\mathbf{W}_i = \lambda_i \beta_i^{(2)}/2(1-\rho_i) + \mathrm{E}\mathbf{RI}_i$, with $\mathbf{I}_i$ the intervisit time at Q_i, cf. [15].)

For the gated service discipline,

$$\mathrm{E}\mathbf{W}_i = (1 + \rho_i)\mathrm{E}\mathbf{RA}_i, \quad i = 1, \ldots, n, \tag{7}$$

with, as before, $\mathbf{A}_i$ the server inter*arrival* time at Q_i, cf. [16,18].

To proceed, one turns to approximating $\mathrm{E}\mathbf{RC}_i$. Since $\mathrm{E}\mathbf{RC}_i = \mathrm{E}(\mathbf{C}_i^2)/2\mathrm{E}\mathbf{C}_i$, where $\mathrm{E}\mathbf{C}_i = s/(1-\rho)$, it remains to approximate $\mathrm{E}(\mathbf{C}_i^2)$ by using some 'additional' information. (Similarly, $\mathrm{E}\mathbf{RI}_i = \mathrm{E}(\mathbf{I}_i^2)/2\mathrm{E}\mathbf{I}_i$, where $\mathrm{E}\mathbf{I}_i = (1-\rho_i)s/(1-\rho)$.) One approach, followed by Bux and Truong [15] in the case of exhaustive service and deterministic switch-over times, is to derive an exact formula for $\mathrm{E}(\mathbf{I}_i^2)$ 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 [16], and further elaborated on by Groenendijk [18], is to approximate $\mathrm{E}\mathbf{RC}_i$

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. [9,10]. Substituting (5) into a pseudo-conservation law, assuming $\mathbf{ERC}_i \approx \mathbf{ERC}$, yields an approximation for $\mathbf{ERC}_i$. Note that the the approximation is exact for completely symmetric systems. For asymmetric systems, the approximation is asymptotically correct in heavy traffic, cf. Van der Mei and Levy [26].

We now return to the multiple-server case. The usual approach to obtain waiting-time approximations may be sketched as follows, cf. [6,19,21,28,30]. Like in the single-server case, one starts by deriving an (approximative) relationship of the form

$$\mathbf{EW}_i \approx \gamma_i \mathbf{ERC}_i^*, \quad i = 1, \ldots, n, \tag{8}$$

with $\mathbf{C}_i^*$ either the *effective* server interarrival or interdeparture time at Q_i. At that stage, the complications start, since 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, since there is very little 'additional' information available that can be used, neither in the form of any exact results for special cases, nor in the global form of a pseudo-conservation law. Thus, 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 the subscript j referring to S_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. [6,19,21,28,30].

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 law-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 were 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 (6), replacing ρ_i by ρ_i/α_i,

$$\mathbf{EW}_i \approx (1 - q_i)\mathbf{ERD}_i, \quad i = 1, \ldots, n, \tag{9}$$

with $\mathbf{D}_i$ the server interdeparture time at Q_i.

Similarly, we obtain from (7) for the gated service discipline,

$$\mathbf{EW}_i \approx (1 + q_i)\mathbf{ERA}_i, \quad i = 1, \ldots, n, \tag{10}$$

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, since 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.,

$$\mathbf{ERD}_i \approx \mathbf{ERD}\rho_i/q_i, \quad i = 1, \dots, n, \tag{11}$$

and

$$\mathbf{ERA}_i \approx \mathbf{ERA}\rho_i/q_i, \quad i = 1, \dots, n, \tag{12}$$

with **ERD** and **ERA** unknown constants. Note that in case $m = 1$, $q_i = \rho_i$, so that (11) and (12) reduce to $\mathbf{ERD}_i \approx \mathbf{ERD}$ and $\mathbf{ERA}_i \approx \mathbf{ERA}$, respectively, the usual assumptions in the single-server case.

From (9)–(12) we obtain

$$\mathrm{E}\mathbf{W}_i \approx \frac{\rho_i(1-q_i)}{q_i} \frac{\sum_{h=1}^{n} \rho_h \mathrm{E}\mathbf{W}_h}{\sum_{h=1}^{n} \rho_h^2(1-q_h)/q_h}, \tag{13}$$

and

$$\mathrm{E}\mathbf{W}_i \approx \frac{\rho_i(1+q_i)}{q_i} \frac{\sum_{h=1}^{n} \rho_h \mathrm{E}\mathbf{W}_h}{\sum_{h=1}^{n} \rho_h^2(1+q_h)/q_h} \tag{14}$$

for the exhaustive and gated service discipline, respectively. Thus, to complete the derivation of the approximations, it suffices to (i) find an expression for the weighted sum $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$ and (ii) determine the probabilities q_i, which we will do in Sections 5 and 6, respectively.

5. Approximating the weighted sum $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$

In this section, we describe a method for approximating $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$. Denote by $\mathbf{V}$ the steady-state total amount of work in the system. Applying Brumelle's formula [14],

$$\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i = \mathrm{E}\mathbf{V} - \frac{1}{2}\sum_{i=1}^{n} \lambda_i \beta_i^{(2)}. \tag{15}$$

So to find an expression for $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$, it suffices to find an expression for the mean amount of work $\mathrm{E}\mathbf{V}$. For reference, we first briefly review the single-server case, where the crucial property that facilitates the determination of $\mathrm{E}\mathbf{V}$ 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. [9,10]:

$$\mathbf{V} \overset{\mathrm{d}}{=} \mathbf{V}^0 + \mathbf{Y}, \tag{16}$$

with $\stackrel{\mathrm{d}}{=}$ indicating equality in distribution. When the amount of work in a switching interval is always zero, we may recognize in (16) the underlying property of *work conservation*, which in fact holds even in a sample-path sense. Note that $\mathbf{E}\mathbf{V}^0$ is simply known from the Pollaczek–Khintchine formula. For a broad class of service disciplines, including gated and exhaustive, $\mathbf{E}\mathbf{Y}$ may be determined along the lines of [9,10]. Taking expectations in (16), substituting into (15), 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 a strict work conservation property in a 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 (16) holds in the multiple-server case. Even if it were, we would face the problem that $\mathbf{E}\mathbf{V}^0$ is generally not known, not to mention the problem of determining $\mathbf{E}\mathbf{Y}$, 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

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

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 $\mathbf{E}\mathbf{V}$, 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 (16) to the two auxiliary systems,

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

From the Pollaczek–Khintchine formula,

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

with $\hat{\rho} = \rho/m$. From [9,10],

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

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

$$\gamma(\rho) = \frac{\sum_{l=0}^{m-1} \frac{\rho^l}{l!} l + \frac{\rho^m}{m!} \sum_{l=m}^{\infty} \left(\frac{\rho}{m}\right)^{l-m} l}{\left(\sum_{l=0}^{m-1} \frac{\rho^l}{l!} + \frac{\rho^m}{m!} \sum_{l=m}^{\infty} \left(\frac{\rho}{m}\right)^{l-m}\right) \frac{\hat{\rho}}{1-\hat{\rho}}} = \frac{(m-\rho)\sum_{l=0}^{m-1} l \frac{\rho^l}{l!} + \frac{\rho^m}{m!} m^2 + \frac{\rho^m}{m!} \frac{m\rho}{m-\rho}}{\rho \sum_{l=0}^{m-1} \frac{\rho^l}{l!} + \frac{\rho^m}{m!} \frac{m\rho}{m-\rho}}. \tag{17}$$

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 more 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 (17) that $\gamma(\rho) \downarrow 1$ 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. If the total switch-over time incurred during a cycle is deterministic, then $\mathrm{E}\mathbf{W}_i \downarrow s/(m+1)$ for $\rho \downarrow 0$ (denoting that $\mathrm{E}\mathbf{W}_i$ decreases to $s/(m+1)$ as ρ decreases to 0). If the total switch-over time during a cycle is exponentially distributed, then $\mathrm{E}\mathbf{W}_i \downarrow s/m$ 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

$$\mathrm{E}\mathbf{Y}/\mathrm{E}\mathbf{Y}_{\lambda/m} \to \frac{(m + s^{(2)}/s^2 - 1)s/m(m+1)}{s^{(2)}/2ms} = 2(1 + (m-1)s^2/s^{(2)})/(m+1) \quad \text{for } \rho \downarrow 0.$$

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

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_i - \rho_i/(1-(1-\rho_i/m)^m)}{m - \rho_i/(1-(1-\rho_i/m)^m)}. \tag{18}$$

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

Concluding,

$$\begin{aligned}\mathbf{EV} \approx\ & \frac{\gamma(\rho)}{m}\frac{\sum_{i=1}^{n}\lambda_i\beta_i^{(2)}}{2(1-\hat{\rho})} + \left((1-\alpha)\left(2\frac{\rho}{m}\frac{1+(m-1)s^2/s^{(2)}}{m+1} + \gamma(\rho)\left(1-\frac{\rho}{m}\right)\right) + \alpha m\right) \\ & \times \left(\hat{\rho}\frac{s^{(2)}}{2s} + \frac{s}{2(1-\hat{\rho})}\left[\hat{\rho}^2 - \sum_{i\in E}\hat{\rho}_i^2 + \sum_{i\in G}\hat{\rho}_i^2\right]\right)\end{aligned} \tag{19}$$

with $\gamma(\rho)$ and α as in (17) and (18), respectively. Substituting (19) into (15) yields an approximative pseudo-conservation law. Subsequently substituting (15) into (13) and (14) yields waiting-time approximations, still containing the probabilities q_i, which we will determine in Section 6. Note that the approximation is exact for 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

$$q_i = 1 - \Pr\{(\pi_j(\mathbf{H}_j), \mathbf{Z}_j) \neq (i, 1),\ j = 1, \ldots, m\}. \tag{20}$$

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

$$\Pr\{(\mathbf{H}_j, \mathbf{Z}_j) = (h, 1)\} = r_{\pi_j(h)j} \approx \frac{\rho_{\pi_j(h)}}{m}. \tag{21}$$

Also, from (1),

$$\Pr\{(\mathbf{H}_j, \mathbf{Z}_j) = (h, 0)\} = \frac{s_{\pi_j(h)}}{\mathbf{EC}_j} \approx (1 - \frac{\rho}{m})\frac{s_{\pi_j(h)}}{s} \tag{22}$$

with $\mathbf{EC}_j$ denoting the mean cycle time of S_j.

It is considerably harder, however, to approximate the *simultaneous* distribution of $(\mathbf{H},\mathbf{Z}) = (\mathbf{H}_1,\dots,\mathbf{H}_m,\mathbf{Z}_1,\dots,\mathbf{Z}_m)$ which is actually needed in (20). There are three types of transitions in $(\mathbf{H},\mathbf{Z})$.

First,

$$(h,z) \to (h+e_j, z-e_j), \quad z_j = 1, \tag{23}$$

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 jth m-dimensional unit vector; $h_j + 1$ is to be understood as $(h_j \bmod n) + 1$.

Second,

$$(h,z) \to (h, z+e_j), \quad z_j = 0, \tag{24}$$

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

Third,

$$(h,z) \to (h+e_j, z), \quad z_j = 0, \tag{25}$$

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, since 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 *were* 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,

$$\mu_{(h,z)\to(h+e_j,z-e_j)} = (m-\rho)/(\rho_{\pi_j(h_j)}s), \quad z_j = 1, \tag{26}$$

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., $r_{\pi_j(h_j),j}\mathrm{E}\mathbf{A}_{\pi_j(h_j),j} \approx \rho_{\pi_j(h_j)}s/(m-\rho)$).

Second,

$$\mu_{(h,z)\to(h,z+e_j)} = 1/s_{\pi_j(h_j)}, \quad z_j = 0, \tag{27}$$

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,

$$\mu_{(h,z)\to(h+e_j,z)} = 0, \quad z_j = 0, \tag{28}$$

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, since a server arrival is likely to lead to a concurrent server departure, which might suggest replacing (28) by

$$\mu_{(h,z)\to(h+e_j,z)} = 1/s_{\pi_j(h_j)}, \quad z_j = 0, \tag{29}$$

when $\rho \downarrow 0$. However, when $\rho \downarrow 0$, combining (27) with (26) has a similar effect as using (29) would have.

Because of the homogeneity in the transition rates, we would obtain from (26)–(28)

$$\Pr\{(\mathbf{H},\mathbf{Z}) = (h,z)\} = \prod_{j=1}^{m} \Pr\{(\mathbf{H}_j,\mathbf{Z}_j) = (h_j,z_j)\}, \tag{30}$$

where $\Pr\{(\mathbf{H}_j, \mathbf{Z}_j) = (h_j, z_j)\}$ satisfies (21) and (22). In other words, we would obtain complete independence in the server position distribution, while 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 (30), 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 (21) and (22).

More detailed clustering measures

Remember that we approximated the distribution of $(\mathbf{H}, \mathbf{Z})$ to determine the probabilities q_i that at least one of the servers is busy at Q_i, $i = 1, \dots, n$. 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. In the remainder of the present section, we briefly discuss the definition of those alternatives. In Section 7, 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

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

$b = 0, 1$, with $k = \pi_j^{-1}(i)$ such that $\pi_j(k) = i$. For given $(h, z) = (h_1, \dots, h_m, z_1, \dots, z_m)$, let $\mathrm{E}\mathbf{T}_{ij_l}^{(b)}(h_{j_l}, z_{j_l})$, $l = 1, \dots, m$, be the mean entrance times $\mathrm{E}\mathbf{T}_{ij}^{(b)}(h_j, z_j)$, $j = 1, \dots, m$, ordered in decreasing magnitude. Let $\Delta_{il}^{(b)}(h, z) = (\mathrm{E}\mathbf{T}_{ij_{l-1}}^{(b)}(h_{j_{l-1}}, z_{j_{l-1}}) - \mathrm{E}\mathbf{T}_{ij_l}^{(b)}(h_{j_l}, z_{j_l}))$, $l = 1, \dots, m$, with $\mathrm{E}\mathbf{T}_{ij_0}^{(b)}(h_{j_0}, z_{j_0}) = \mathrm{EC}$, $\mathrm{EC} = s/(m-\rho)$. For given (h, z), $\Delta_{il}^{(b)}(h, z)$ represents the mean of the lth of the m most recent server interarrival ($b = 0$) or interdeparture ($b = 1$) times at $Q_i, l = 1, \ldots, m$. Denote $\Delta_i^{(b)}(h, z) = \sum_{l=1}^m (\Delta_{il}^{(b)}(h, z))^2$. The ordinary sum of $\Delta_{il}^{(b)}(h, z), l = 1, \ldots, m$, being always equal to 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

$$\delta_i^{(b)} = \sum_{j=1}^{m} \sum_{(h,z)} P_{ij}^{(b)}(h, z) \Delta_i^{(b)}(h, z)/(\mathrm{EC})^2 \tag{31}$$

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 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 $[0, \mathrm{EC}]$, then $\delta_i^{(b)} \approx m\kappa$, with $\kappa = \int_{1=x_0 \geq \cdots \geq x_m = 0} \sum_{l=1}^m (x_{l-1} - x_l)^2 \,\mathrm{d}x_1 \cdots \mathrm{d}x_{m-1}$. If the servers even tend to repel each other, i.e., all the $\Delta_{il}^{(b)}(h, z)$'s are approximately equal to EC/m, then $\delta_i^{(b)} \approx 1$.

We may also refine the measure (18) for the *global* degree of clustering. In the spirit of (31), we define

$$\delta = \frac{\sum_{(h,z)} \left(\Pr\{(\mathbf{H}, \mathbf{Z}) = (h, z)\} - \prod_{l=1}^m \Pr\{(\mathbf{H}_l, \mathbf{Z}_l) = (h_l, z_l)\}\right) \sum_{j=1}^m \Delta_{\pi_j(h_j)}^{(z_j)}(h, z)}{m(\mathrm{EC})^2 - \sum_{(h,z)} \left(\prod_{l=1}^m \Pr\{(\mathbf{H}_l, \mathbf{Z}_l) = (h_l, z_l)\}\right) \sum_{j=1}^m \Delta_{\pi_j(h_j)}^{(z_j)}(h, z)}. \tag{32}$$

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

7. Numerical results

We now present an overview of the numerical experiments that we have performed to test the accuracy of the waiting-time approximations. The reader is referred to [8] for a more extensive discussion of the numerical results. We reemphasize that multiple-server polling systems are very complex, containing single-server polling models and ordinary multiple-server models as special cases, with the visit order constituting an additional complicating factor. The accuracy of the approximations should be judged from this perspective.

We have focused on four-queue two-server models with exponentially distributed service and switch-over times. 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. In order to test the accuracy of the approximations for a wide range of parameters, we have considered several variants of a set of models in which the ratios between the arrival rates, $(\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4)$ and the mean service times $(\beta_1, \beta_2, \beta_3, \beta_4)$, 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. $(1:1:1:1)$; $(0.5, 0.5, 1.5, 1.5)$,

Table 1
The mean waiting times for Model I with $s = 0.0$; exhaustive service

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 2
The mean waiting times for Model I with $s = 1.0$; exhaustive service

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)

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 $\pi_2 = (1, 2, 3, 4)$, $\pi_2 = (1, 4, 3, 2)$, and $\pi_2 = (1, 2, 4, 3)$. For each of the models, all switch-over times are assumed to have mean $s/n = s/4$ with either $s = 0.0$ or $s = 1.0$ [1]. 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 1–4 show the results for exhaustive service at each of the queues. Tables 5 and 6 show the results for systems 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 (18) 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 (31) for exhaustive service or $\delta_i^{(0)}$ for gated service, and α replaced by δ as in (32). The rows indicated by 'exact' contain the 'exact' mean waiting times obtained from either the PSA (for exhaustive service) or simulation (for gated service) [2]. The truncation error in the PSA (the width of the confidence interval in the simulation) is typically less than 1%. For compactness of the presentation, the confidence regions have been

[1] In evaluating the approximations, we actually took $s = 10^{-6}$, since the formal definition of $(\mathbf{H}, \mathbf{Z})$ is restricted to the case of non-zero switch-over times.

[2] We implemented the PSA only for Bernoulli service, including exhaustive service as a special case, but in principle the method may also be used for gated service.

Table 3
The mean waiting times at Q_1 for Models II–VI with $s = 0.0$; exhaustive service

ρ	π_2	(1, 2, 3, 4)					(1, 4, 3, 2)				
		II	III	IV	V	VI	II	III	IV	V	VI
0.8	exact	0.21	0.13	0.25	0.27	0.21	0.21	0.13	0.26	0.27	0.23
	α	0.20	0.13	0.25	0.27	0.24	0.20	0.13	0.25	0.26	0.24
	δ	0.21	0.13	0.26	0.29	0.26	0.21	0.13	0.26	0.27	0.26
1.6	Exact	2.24	1.25	2.79	3.24	2.96	2.17	1.21	2.74	3.02	3.13
	α	2.03	1.24	2.53	2.75	2.45	2.00	1.24	2.50	2.59	2.48
	δ	2.18	1.24	2.72	3.20	2.94	2.08	1.24	2.60	2.65	2.76
1.8	Exact	5.51	3.01	7.00	8.26	7.94	5.11	2.94	6.38	7.12	8.05
	α	5.06	2.98	6.33	7.01	9.14	4.86	2.98	6.08	6.36	8.50
	δ	5.41	2.98	6.68	7.97	10.44	5.04	2.98	6.30	6.50	9.39

Table 4
The mean waiting times at Q_1 for Models II–VI with $s = 1.0$; exhaustive service

ρ	π_2	(1, 2, 3, 4)					(1, 4, 3, 2)				
		II	III	IV	V	VI	II	III	IV	V	VI
0.8	Exact	0.83	0.72	0.87	0.91	0.84	0.81	0.70	0.86	0.89	0.86
	α	0.82	0.72	0.87	0.89	0.86	0.81	0.71	0.86	0.88	0.86
	δ	0.86	0.72	0.91	0.98	0.95	0.83	0.71	0.89	0.91	0.93
1.6	Exact	3.98	2.80	4.56	5.12	4.69	3.69	2.59	4.28	4.58	4.93
	α	3.72	2.83	4.23	4.43	4.05	3.43	2.59	3.93	3.99	3.93
	δ	4.23	2.98	4.78	5.43	5.19	3.65	2.61	4.17	4.19	4.55
1.8	Exact	9.04	6.43	10.55	11.94	11.44	7.73	5.55	9.04	9.88	10.96
	α	8.69	6.30	9.95	10.59	13.91	7.54	5.47	8.75	8.95	11.99
	δ	9.60	6.59	10.93	12.56	16.66	8.06	5.55	9.32	9.43	13.94

omitted in the presentation of the numerical results. If we denote the 'exact' value of a performance measure by z_{exact} and the approximated value by z_{app}, then the relative error is defined by $(z_{\text{app}} - z_{\text{exact}})/z_{\text{exact}} \times 100\%$.

Table 1 shows the mean waiting times at each of the queues for Model I for the case $s = 0.0$, and Table 2 gives the results for $s = 1.0$.

Tables 1 and 2 show that the approximations for Model I are very accurate, with relative errors typically well below 5%. In particular, Table 1 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.

Tables 3 and 4 present the results for Models II to VI, for $s = 0.0$ and $s = 1.0$, respectively. For ease of the presentation, only the mean waiting times at Q_1 are presented here; the accuracy of the approximations at the other queues is similar.

The numerical results presented in Tables 3 and 4 lead to a number of conclusions. First, the results are still accurate when the arrival rates are fairly asymmetrical, even for heavily loaded systems (Model II), with relative errors typically less than 10%. For the cases considered for Model III, 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, with relative errors typically less than 10%. In Model IV, the service times are asymmetrical, whereas the arrival rates are the same. The presented results indicate that the accuracy of the approximations is still acceptable, even in heavily loaded systems. In Model V, the arrival rates as well as the service times are asymmetrical. 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. 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 3 and 4 show that the accuracy of the waiting-time approximation presented in this paper also degrades somewhat when the model is very asymmetrical, but remains acceptable as long as the load is not too high.

We have also checked the accuracy of the approximation for multiple-server systems with gated service at all queues. As a typical example, we present the results for Model II. Tables 5 and 6 present the results for $s = 0.0$ and $s = 1.0$, respectively.

Tables 5 and 6 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.

Table 5
The mean waiting times for Model II with $s = 0.0$; gated service

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 6
The mean waiting times for Model II with $s = 1.0$; gated service

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)

Discussion of the numerical results

As discussed extensively in Sections 4–6, the waiting-time approximations presented in this paper are based on a series of assumptions, each of which, by definition, forms a potential 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. (9) and (10)), (ii) the ratios between the mean residual cycle times (cf. (11) and (12)), and (iii) the value of q_i (cf. (20)). The second error source is the estimation of the mean amount of work in the system (cf. (19)). Extensive simulation experiments have been performed to check the impact of each of these error sources on the 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, $\mathbf{E}V$, according to (19), is rather accurate. The error in the estimation of $\mathbf{E}V$ is typically less than 5% in fairly symmetrical systems, even in 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. (9) and (10)). For systems with the exhaustive service discipline, the mean waiting times are usually underestimated according to (9) (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 (10). Apparently, in the case of exhaustive service the approximation (9) is too optimistic and, in the case of gated service, the approximation (10) 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 (9) and (10) only have a marginal impact on the waiting-time approximations. The ratios between the mean residual cycle times are estimated by the ratios between the estimated average processing rates according to (11) and (12). Numerical experimentation has indicated that the quality of these estimations is quite good, with errors typically below 10%, except for very asymmetrical systems in heavy traffic.

In the general approach developed in Sections 4–6, we used $\alpha_i = \rho_i/q_i$ as a measure for the local degree of clustering and α in (18) 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. (31) and (32)). 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 somewhat.

8. Concluding remarks and suggestions for further research

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. In these systems, the servers have the tendency to coalesce, especially in heavy traffic when 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 wide range of 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 (9) and (10) 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, however, the analysis will 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 be harder, however, to approximate the simultaneous distribution of $(\mathbf{H}, \mathbf{Z})$, needed to determine these measures, since for $m_i < m$ there are also instantaneous passages through states with more than m_i servers at Q_i. The derivation of an approximative pseudo-conservation law will also 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 evaluation is efficient operation and optimization. The development of accurate waiting-time approximations may be very useful in solving a variety of optimization problems for multiple-server polling systems, opening up a very interesting area for further research.

Acknowledgements

The authors would like to thank JPC Blanc and OJ Boxma for several valuable discussions and the anonymous referees for their useful suggestions.

References

[1] M. Ajmone Marsan, S. Donatelli and F. Neri, GSPN models of Markovian multiserver multiqueue systems, *Performance Evaluation* **11** (1990) 227–240.

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

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

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

[5] B. van Arem, Queueing models for slotted transmission systems, Ph.D. Thesis, Twente University, Enschede, 1990.

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

[7] S.C. Borst, Polling systems with multiple coupled servers, *Queueing Systems* **20** (1995) 369–393.

[8] S.C. Borst and R.D. van der Mei, Waiting-time approximations for multiple-server polling systems, CWI Report BS-R9428 (1994).

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

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

[11] S. Browne, E.G. Coffman, Jr., E.N. Gilbert and P.E.W. Wright, Gated, exhaustive, parallel service, *Prob. Eng. Inf. Sci.* **6** (1992) 217–239.

[12] S. Browne and O. Kella, Parallel service with vacations, *Oper. Res.* **43** (1995) 870–878.

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

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

[15] W. Bux and H.L. Truong, Mean-delay approximations for cyclic-service queueing systems, *Performance Evaluation* **3** (1983) 187–196.

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

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

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

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

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

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

[22] H. Levy, G. Mahalal and M. Sidi, Multi server polling systems: the bang bang policies, in: *Proc. Experts on Networks Workshop*, UCLA, June 1994, submitted for publication (1994).

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

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

[25] R.D. van der Mei and S.C. Borst, Analysis of multiple-server polling systems by means of the power-series algorithm, *Stochastic Models* **13** (1997).

[26] R.D. van der Mei and H. Levy, Expected delay analysis of polling systems in heavy traffic, submitted for publication (1996).

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

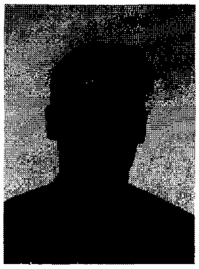
[28] R.J.T. Morris and Y.T. Wang, Some results for multi-queue systems with multiple cyclic servers, in: W. Bux and H. Rudin (Eds.), *Performance of Computer-Communication Systems* (North-Holland, Amsterdam, 1984) 245–258.

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

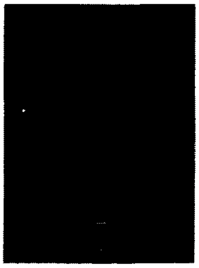
[30] T. Raith, Performance analysis of multi-bus interconnection networks in distributed systems, in: M. Akiyamma (Ed.), *Teletraffic Issues in an Advanced Information Society, Proc. ITC-11* (North-Holland, Amsterdam, 1985) 662–668.

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

[32] H. Takagi, Application of polling models to computer networks, *Comp. Netw. ISDN Syst.* **22** (1991) 193–211.
[33] M. Zafirovic-Vukotic, I.G. Niemegeers and D.S. Valk, Performance modelling of slotted ring protocols in HSLANs, *IEEE J. Sel. Areas Comm.* **6** (1988) 1001–1024.



Sem Borst received the M.Sc. degree in applied mathematics from the University of Twente, Netherlands, in 1990, and the Ph.D. degree from the University of Tilburg, Netherlands, in 1994. During the fall of 1994, he was a visiting scholar at the Statistical Laboratory of the University of Cambridge, England. Since 1995, he has been a member of technical staff in the Mathematics of Networks and Systems Department of Bell Laboratories, Lucent Technologies in Murray Hill, USA. His main research interests are in the performance evaluation of communication networks and computer systems.



Robert D. van der Mei (1966) received his M.Sc. degrees in mathematics and in decision sciences from the Free University of Amsterdam, Netherlands, in 1990, and his Ph.D. degree in operations research from Tilburg University, Netherlands, in 1995. Meanwhile, he worked for more than a year with the Center for Qantitative Methods, Philips, Eindhoven, Netherlands, and with the Department of Mathematics and Systems Engineering at Royal Shell Laboratories, Amsterdam, Netherlands. In 1995, he worked at the Research Division of the Royal PTT Netherlands. In 1996, he was a Visiting Scholar at Rutgers' Center for Operations Research at Rutgers University, New Brunswick, NJ, USA, and at the Department of Industrial Engineering and Operations Research at Columbia University, New York, USA. Since October 1996, Dr. van der Mei has been a Senior Technical Staff Member at the Teletraffic and Performance Analysis Department at AT&T Labs in Holmdel, NJ, USA. His research interests include performance analysis of computer and communication networks, queueing analysis, and stochastic systems.