The Use of Service Limits for Efficient Operation of Multi-Station Single-Medium Communication Systems

S.C. Borst
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
O.J. Boxma*
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
H. Levy
Rutgers University
New Brunswick, NJ 08903, USA

March 17, 1995

Abstract

Time limits are the major mechanisms used for controlling a large variety of multi-station single-medium computer-communication systems like the FDDI network and the IEEE 802.4 Token Bus. The proper use of these mechanisms is still not understood and rules for efficient system operation are not available. Our objective is the derivation of such rules. We use a cyclic polling model with different service limits (k-limited service) at the different queues, thus emulating time limits. We are interested in determining these k-limit values so as to minimize the mean waiting cost of messages in the system. A simple approximative approach is proposed for two major problems: One in which a limit is set on the token rotation time and one in which no limits are imposed. The approach is tested for a variety of cases and is shown to be very effective.

1 Introduction

Timed-token passing protocols are the medium access control protocols used in many local area networks, such as the FDDI network [1] and the IEEE 802.4 Token Bus [2]. In timed-token protocols, the time during which a station can continue to transmit may depend on the congestion of the network as well as the priority of messages under transmission (cf. [29]). The option of setting *different* limits to different stations is the main mechanism available for *prioritizing* the stations and achieving good system performance. Timer-based service disciplines have also been commonly used in other systems with forms of resource sharing (e.g., several processors on the AT&T 5ESS Switch network and control point).

While the use of sophisticated time-limited service mechanisms in the control of multi-station systems is widespread, little is known about how to operate these mechanisms in order to achieve desired performance. Our objective is to address this problem by investigating which choices for the time limits lead to the optimal performance of the access protocol. The system performance

*Part of the second author's research was supported by the European Grant BRA-QMIPS of CEC DG XIII.

is expressed in the mean waiting cost, where the cost parameters of the different stations are set differently according to their relative importance.

The performance of token passing protocols has often been studied by analyzing a cyclic polling model, i.e., a queueing model in which a single server S visits a set of queues $Q_1, \ldots, Q_N$ in cyclic order. Restrictions on the token rotation time, which may be invoked in times of congestion, are hardly handled in the performance literature. Fixed time limits, too, usually represent unsurmountable mathematical difficulties (cf. De Souza e Silva et al. [17]). Therefore, one typically finds the following emulations of time-limited service: (i) exponential timers: Coffman et al. [15]; (ii) sum of exponential-phase timers: Leung & Lucantoni [26]; (iii) probabilistically-limited service: Leung [25]; (iv) Bernoulli service: Blanc & Van der Mei [5]; (v) k-limited service: Fuhrmann & Wang [23].

In the present study we also use k-limited as an emulation of time-limited service. Under k-limited service S serves, upon each visit to Q_i, at most k_i customers; $k_i \in \{1, 2, \ldots\}$, $i = 1, \ldots, N$. Leung [26] numerically analyzes a polling model with exponential timers, and compares the mean waiting times with those for k-limited service and those for fixed timers. Taking exponential service times, it turns out that both an exponential timer and k-limited give very good mean waiting time approximations for fixed time limits. k-Limited service is somewhat worse for relatively small timers, and (naturally) somewhat better for large timers; one may expect that k-limited has the edge when service times become less variable. Note that k-limited service coincides with time-limited in the practically relevant case of constant service times (fixed-length packets). Hence, the rules for setting the k_i-values optimally, as derived in the present study, will give very useful indications for setting efficient time limits.

k-Limited is also of interest in its own right; there is a large variety of non-generic systems in telecommunications which use polling strategies to provide service to several entities (e.g., collecting messages, which arrive on several incoming links and queue up in the incoming queues, in a telecommunications switch). Many of these systems use a limited-service mechanism to provide different service to the different queues in order to improve system performance, cf. also the recently introduced 'weighted fair queueing' (or 'weighted processor sharing' or 'weighted round robin') policy for ATM.

Our objective is to find the k_i-values that minimize the mean waiting cost (or the weighted mean waiting time):

$$\min_{k_1,\ldots,k_N} \sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i,$$

where $\mathrm{E}\mathbf{W}_i$ is the mean waiting time at Q_i, λ_i is the rate at which customers arrive at Q_i, and c_i is the waiting cost parameter of that queue (the cost imposed by having a customer wait one time unit). This problem may be denoted as the *unconstrained* optimization problem. To consider systems in which an additional control mechanism is to set up a limit on the server rotation time, namely the time it takes for the server to complete a cycle, we consider the same optimization problem but under the constraint

$$\sum_{i=1}^{N} \gamma_i k_i \leq K,$$

where γ_i are arbitrary parameters. We denote this problem as the *constrained* optimization problem. Setting $\gamma_i = \beta_i$, with β_i the mean service time at Q_i, translates the constraint to a limit on the expected value of the rotation time. If the service times are constant (transmission of fixed-length packets) as well as the switchover times, then the constrained case may reflect an upper bound on the *actual* cycle time.

Unfortunately, polling systems with k-limited service are very hard to analyze (let alone optimize); an exact analysis is only available for very few special cases. Eisenberg [18] and Cohen &

Boxma [16] study the 2-queue model with 1-limited service at both queues and zero switchover times; both papers use methods from complex function theory. Boxma & Groenendijk [10] analyze a similar model with non-zero switchover times by solving a Riemann boundary value problem, the mean waiting times being expressed as singular integrals. The 2-queue model with 1-limited service at Q_1 and *exhaustive* service at Q_2 ($k_2 = \infty$) has turned out to be a relatively simple model [24], Section 6.3.

The fact that even *mean* waiting times in polling models with limited-service policies are generally not known, adds to the importance of the so-called *pseudo-conservation law (pcl)*. The pcl provides an exact expression for a specific weighted sum of the mean waiting times. For the polling model with 1-limited service at all queues, such a pcl has first been derived by Watson [32]; a more general pcl is derived in [9], [7], using a simple probabilistic argument. For the case of k-limited service with $k_i > 1$ for some i-values, the pcl still contains some unknown quantities. Everitt [19], [20] approximates that term, while Fuhrmann & Wang [23] give bounds for it.

One of the advantages of the pcl is that it is useful in developing simple and reasonably accurate approximations for the individual mean waiting times; Fuhrmann & Wang [23] have provided such an approximation for k-limited service. Such simple approximations may subsequently be used for optimization purposes; that will be exploited in the present paper.

Motivated by the fact that relatively 'rough' approximations for the mean waiting times have led to quite good operational rules for polling systems, [11], [12], we use approximations for the mean waiting times which are relatively simple but which capture the major factors important for efficient operation. For the problem of finding optimal service limits under rotation time constraints, we develop and investigate several such approximations, leading to various operational rules. For the unconstrained problem we start our analysis by deriving some properties of polling systems with k-limited service. In particular we derive a $c\mu$-like rule for systems with switchover periods. The (partially conjectured) derived rule states that for optimal operation of these systems the queues with the highest value of the ratio c_i/β_i must have their k_i set at infinity, i.e., receive exhaustive service. We then study an approximation which possesses similar properties, to suggest operational rules for the system. The resulting operational rules (for both problems) are numerically tested for a wide range of cases and are shown to be very effective in optimizing the system performance.

The paper is organized as follows. Section 2 contains a detailed model description and some preliminary results on mean waiting times. In Section 3 we propose four approaches to the optimization problem under rotation time constraints. These approaches are numerically tested in Section 4. In Section 5 we derive properties of polling systems with k-limited service and a (partially conjectured) $c\mu$-like rule for the unconstrained optimization problem. We then also study an approximative approach to this problem. This approach is numerically examined in Section 6. Some conclusions are presented in Section 7.

Remark 1.1 The vast polling literature contains only a few optimization studies. At ITC-13, two surveys on optimal *server routing* were presented: [33] on semi-dynamic routing, and [8] on static routing. For *given* server routing, Levy et al. [28] prove that the service strategy that minimizes the amount of work in the system is to serve as many customers as possible at each queue visit.

Blanc & Van der Mei [5] study an optimization problem that is related to ours. They consider the Bernoulli service policy at each queue: when S visits a non-empty queue, it serves one customer; and at each service completion which does not leave the queue empty, S serves yet another customer with probability q_i and proceeds to the next queue with probability $1 - q_i$. Blanc & Van der Mei try to find those q_i, $i = 1, \ldots, N$, which minimize a weighted sum of the mean waiting times. Their main approach is a numerical one, based on the use of the power series algorithm (psa). The psa allows an accurate numerical determination of the mean

waiting times in polling models for which the joint queue-length process has the structure of a multi-dimensional quasi birth-death process [3], [4]. The psa has a quite wide applicability for multi-dimensional queueing problems, its main disadvantage being that time and memory requirements grow exponentially with the number of queues. In view of this drawback Blanc & Van der Mei [5] subsequently propose and investigate a simple approximation for the mean waiting times in polling models with Bernoulli service.
The Bernoulli service policy is the stochastic counterpart of the k-limited service policy, having mainly been devised to emulate the behavior of the k-limited discipline. In that sense, our paper can be viewed as a companion paper to [5].

□

2 Model description and preliminaries

A single server S serves N infinite-capacity queues (stations) $Q_1, \ldots, Q_N$ in cyclic order, switching from queue to queue. Customers arrive at all queues according to independent Poisson processes. The arrival intensity at Q_i is λ_i, $i = 1, \ldots, N$, and the total arrival rate is $\lambda := \sum_{i=1}^{N} \lambda_i$. Customers arriving at Q_i are called class-i customers; their service times are independent random variables with mean β_i and second moment $\beta_i^{(2)}$, $i = 1, \ldots, N$. The offered traffic load, ρ_i, at Q_i is defined as $\rho_i := \lambda_i \beta_i$, $i = 1, \ldots, N$, and the total offered load is $\rho := \sum_{i=1}^{N} \rho_i$.
When visiting Q_i, S works until either k_i customers have been served or the queue becomes empty, whichever comes first. Note that $k_i = \infty$ amounts to exhaustive service. Fuhrmann & Wang [23] call this policy E-limited service, as opposed to G-limited service in which S, when meeting m_i customers at Q_i upon arrival, only serves $\min(m_i, k_i)$ customers. In G-limited service, $k_i = \infty$ amounts to gated service. When swapping out of Q_i (moving towards $Q_{(i \bmod N)+1}$) the server incurs a switchover period of type i; the switchover durations are independent random variables with mean s_i and second moment $s_i^{(2)}$, $i = 1, \ldots, N$. The total switchover time during one cycle of the server has mean $s = \sum_{i=1}^{N} s_i$ and second moment $s^{(2)}$. The interarrival, service and switchover processes are independent stochastic processes.
Fricker & Jaïbi [21] have recently provided a mathematically rigorous presentation of necessary and sufficient stability conditions for a large class of cyclic polling systems, which includes the one of this paper. Their condition reads here:

$$\rho + \max_{i=1,\ldots,N} \{\lambda_i s / k_i\} < 1. \tag{1}$$

In the sequel this condition is assumed to be fulfilled.
Let $\mathbf{W}_i$ denote the steady-state waiting time at Q_i, and let c_i denote the cost imposed on the system of having a customer wait one unit of time at Q_i. The expected cost of operating the system per unit of time is thus $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$. The problem of interest in this paper is that of finding a vector $(k_1, \ldots, k_N)$ which minimizes the expected operating cost $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$. This cost is minimized both for the case of a linear constraint $\sum_{i=1}^{N} \gamma_i k_i \leq K$ and for the unconstrained case. The choice $\gamma_i \equiv 1$ puts a limit on the number of services in a cycle, and the choice $\gamma_i \equiv \beta_i$ yields a bound on the mean cycle time.

Everitt [19] has derived the following pseudo-conservation law for the E-limited service discipline:

$$\sum_{i=1}^{N} \rho_i \left(1 - \frac{\lambda_i s}{k_i(1-\rho)}\right) \mathrm{E}\mathbf{W}_i = D + \frac{s}{1-\rho} \sum_{i=1}^{N} \frac{\rho_i^2}{k_i} - \sum_{i=1}^{N} \frac{\rho_i(1-\rho_i) g_i^{(2)}}{2\lambda_i k_i}, \tag{2}$$

with

$$D = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)} \left[\rho^2 - \sum_{i=1}^{N} \rho_i^2\right], \tag{3}$$

and $g_i^{(2)}$ denoting the second factorial moment of the number of customers served during a visit to Q_i. For $k_i = 1$, $g_i^{(2)} = 0$, but for $k_i \neq 1$, $g_i^{(2)}$ is not known exactly.

3 Waiting cost minimization under a rotation time constraint

In the present section we study the problem of finding the service limits $k_1, \ldots, k_N$, constrained to $\sum_{i=1}^{N} \gamma_i k_i \leq K$, that minimize the waiting cost $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$. The constraint reflects some rotation time restriction. We successively consider the following four mean waiting time approximations:

I. an approximation based on a 1-limited polling table;
II. a simple k-limited approximation;
III. a Fuhrmann & Wang-like k-limited approximation;
IV. the original Fuhrmann & Wang k-limited approximation.

Each of the first three approximations for $\mathrm{E}\mathbf{W}_i$ is convex decreasing in k_i while being insensitive to k_j, $i \neq j$. Thus, ignoring integrality constraints, the optimal service limits $k_1^*, \ldots, k_N^*$ may be determined by putting (with some abuse of notation) $c_i \lambda_i \frac{\partial}{\partial k_i} \mathrm{E}\mathbf{W}_i = -\gamma_i \ell$, $i = 1, \ldots, N$, with ℓ denoting a Lagrangean multiplier. The optimal service limits based on the fourth approximation cannot be solved analytically but have to be determined numerically.

I. An approximation based on a 1-limited polling table.
A generalization of the cyclic visit order considered so far is a fixed, generally non-cyclic, visit order. Such a visit order may be described in a *polling table*, which may contain $m_i \geq 1$ visits to Q_i.
Our approximation idea is the following. There is some resemblance between adopting the k_i-limited service discipline at Q_i, visiting Q_i once, and adopting the 1-limited service discipline at Q_i, visiting Q_i k_i times; in either case the server is allowed to serve at most k_i customers in one 'cycle'. So the optimal visit numbers $m_1, \ldots, m_N$ for the 1-limited service discipline may provide an indication for the optimal $k_1, \ldots, k_N$.
Boxma, Levy, & Weststrate [11] study the problem of finding those polling table visit numbers $m_1, \ldots, m_n$, that minimize $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$. They propose the following mean waiting time approximation, under the assumption that the m_i visits to Q_i are spaced as evenly as possible:

$$\mathrm{E}\mathbf{W}_i \approx A \frac{1-\rho+\rho_i}{1-\rho-\lambda_i \frac{\sum m_j s_j}{m_i}} \frac{\mathrm{E}\mathbf{C}}{m_i}, \qquad i = 1, \ldots, N, \tag{4}$$

with $\mathrm{E}\mathbf{C} = \sum_{j=1}^{N} m_j s_j / (1-\rho)$ the mean cycle time, and B some unknown constant. B could be determined using the pseudo-conservation law for polling tables, but its value is not relevant for the determination of the optimal values of m_i (denoted by m_i^*), which follow easily from (4) by using the Lagrangean multiplier technique:

$$m_i^* = \lambda_i R + (1 - \rho - \sum_{j=1}^{N} \lambda_j s_j) R \frac{\sqrt{c_i \lambda_i (1 - \rho + \rho_i)/s_i}}{\sum_{j=1}^{N} s_j \sqrt{c_j \lambda_j (1 - \rho + \rho_j)/s_j}}, \qquad i = 1, \ldots, N. \tag{5}$$

Here R represents an arbitrary scaling factor, reflecting the homogeneity of the objective function in $m_1, \ldots, m_N$.

As remarked before, the optimal visit numbers $m_1, \ldots, m_N$ for the 1-limited service discipline may provide an indication for the optimal $k_1, \ldots, k_N$. However, visiting Q_i k_i times differs from visiting Q_i only once in the respect of the switchover time incurred. In the former (latter) case the switchover time corresponding to Q_i is incurred k_i times (once) per cycle. So the optimal visit numbers $m_1, \ldots, m_N$ may be better candidates to provide an indication for the optimal $k_1, \ldots, k_N$, when the mean switchover times in (4) are scaled by a factor $1/k_i$; this yields

$$\mathrm{E}\mathbf{W}_i \approx B \frac{1 - \rho + \rho_i}{1 - \rho - \lambda_i \frac{s}{k_i}} \frac{\mathrm{E}\mathbf{C}}{k_i}, \qquad i = 1, \ldots, N, \tag{6}$$

with $\mathrm{E}\mathbf{C} = s/(1-\rho)$. This leads to:

$$k_i^* = \frac{\lambda_i s}{1 - \rho} + (K - \sum_{j=1}^{N} \gamma_j \frac{\lambda_j s}{1 - \rho}) \frac{\sqrt{c_i \lambda_i (1 - \rho + \rho_i)/\gamma_i}}{\sum_{j=1}^{N} \gamma_j \sqrt{c_j \lambda_j (1 - \rho + \rho_j)/\gamma_j}}, \qquad i = 1, \ldots, N. \tag{7}$$

One may interpret (7) as follows. The server should be allowed to serve at least $\frac{\lambda_i s}{1-\rho}$ customers during a visit to Q_i, to satisfy the stability condition (1). The remaining service capacity, $K - \sum_{j=1}^{N} \gamma_j \frac{\lambda_j s}{1-\rho}$, should be assigned proportionally to $\sqrt{c_i \lambda_i (1 - \rho + \rho_i)/\gamma_i}$. Some reflection convinces one that indeed a station with relatively high c_i, λ_i, ρ_i, or $1/\gamma_i$ should be assigned a relatively high capacity.

(7) is just as simple as (5) and yields better results. Still, the numerical results in Section 4 reveal that it does not always perform well. Below we investigate a quite different idea.

II. A simple k-limited approximation.

We now imitate the derivation of the mean waiting time approximation for cyclic polling systems with 1-limited service [13] and for polling tables with 1-limited service ([11], leading to (4)). The waiting time of a (tagged) class-i customer is composed of:

i. the time from its arrival to the start of the subsequent visit of the server to Q_i, i.e., a residual cycle time $\mathbf{RC}_i$ with regard to Q_i;

ii. the time from the start of the latter visit to its service, i.e., approximately $\mathbf{X}_i/k_i$ cycle times $\mathbf{C}_i^+$ with regard to Q_i (atypical cycles, as each contains k_i services at Q_i), when the (tagged) customer finds $\mathbf{X}_i$ waiting class-i customers upon arrival.

Applying a traffic balance argument, $\mathrm{E}\mathbf{C}_i^+ \approx k_i \beta_i + s + (\rho - \rho_i)\mathrm{E}\mathbf{C}_i^+$. Noting that $\mathrm{E}\mathbf{X}_i = \lambda_i \mathrm{E}\mathbf{W}_i$, we thus obtain

$$\mathrm{E}\mathbf{W}_i \approx \frac{1 - \rho + \rho_i}{1 - \rho - \frac{\lambda_i}{k_i} s} \mathrm{E}\mathbf{RC}_i, \qquad i = 1, \ldots, N. \tag{8}$$

For $k_i = 1$ (8) reduces to $\mathrm{E}\mathbf{W}_i \approx \dfrac{1-\rho+\rho_i}{1-\rho-\lambda_i s}\mathrm{E}\mathbf{RC}_i$, the known approximation [13] for the 1-limited service discipline. However, for $k_i = \infty$ (8) reduces to $\mathrm{E}\mathbf{W}_i \approx \dfrac{1-\rho+\rho_i}{1-\rho}\mathrm{E}\mathbf{RC}_i$, rather than $\mathrm{E}\mathbf{W}_i = (1-\rho_i)\mathrm{E}\mathbf{RC}_i$, the known exact result for the exhaustive service discipline (defining a cycle with regard to Q_i as the interval between two successive departures of S from Q_i). The reason for this discrepancy is that the derivation of (8) ignores the possibility that a class-i customer upon arrival finds S visiting Q_i and can still be served during that visit.
Starting from (8), assuming $\mathrm{E}\mathbf{RC}_i \approx \mathrm{E}\mathbf{RC} = B\,\mathrm{E}\mathbf{C} = Bs/(1-\rho)$ with B some unknown constant,

$$k_i^* = \frac{\lambda_i s}{1-\rho} + (K - \sum_{j=1}^{N}\gamma_j\frac{\lambda_j s}{1-\rho})\frac{\lambda_i\sqrt{c_i(1-\rho+\rho_i)/\gamma_i}}{\sum_{j=1}^{N}\gamma_j\lambda_j\sqrt{c_j(1-\rho+\rho_j)/\gamma_j}}, \qquad i = 1,\ldots,N. \tag{9}$$

One may interpret (9) similarly to (7).
Note that (9) slightly differs from (7) in the proportional assignment of the remaining service capacity, $K - \sum_{j=1}^{N}\gamma_j\dfrac{\lambda_j s}{1-\rho}$, which may be explained as follows. Visiting Q_i k_i times differs from visiting Q_i only once not only in the respect of the switchover time incurred, as remarked before, but also differs in the respect of the residual time until visiting Q_i. In the former case the residual subcycle time is assumed to behave inversely proportional to k_i, whereas in the latter case the residual cycle time is assumed to behave constantly.

III. A Fuhrmann & Wang-like k-limited approximation.
To remedy the weakness of (8) indicated above, a natural heuristic approach is to take a weighted sum of the 1-limited mean waiting time approximation $\mathrm{E}\mathbf{W}_i \approx \dfrac{1-\rho+\rho_i}{1-\rho-\lambda_i s}\mathrm{E}\mathbf{RC}_i$ and the exhaustive mean waiting time result $\mathrm{E}\mathbf{W}_i = (1-\rho_i)\mathrm{E}\mathbf{RC}_i$, with weight factors $u_i(k_i)$ and $1-u_i(k_i)$. The choice $u_i(k_i) = \dfrac{1-\rho-\lambda_i s}{k_i(1-\rho)-\lambda_i s}$ has the desirable properties that $u_i(1) = 1$, $u_i(\infty) = 0$, and $\mathrm{E}\mathbf{W}_i \to \infty$ for $k_i \to \dfrac{\lambda_i s}{1-\rho}$.
This in fact yields the approximation (30) of Fuhrmann & Wang [23]:

$$\mathrm{E}\mathbf{W}_i \approx \frac{(1-\rho_i)(1-\rho) + \dfrac{\rho_i}{k_i}(2-\rho)}{1-\rho-\dfrac{\lambda_i s}{k_i}}\mathrm{E}\mathbf{RC}_i, \qquad i = 1,\ldots,N. \tag{10}$$

Starting from (10), assuming $\mathrm{E}\mathbf{RC}_i \approx \mathrm{E}\mathbf{RC} = B\,\mathrm{E}\mathbf{C}$, with B some unknown constant,

$$k_i^* = \frac{\lambda_i s}{1-\rho} + (K - \sum_{j=1}^{N}\gamma_j\frac{\lambda_j s}{1-\rho})\frac{\sqrt{c_i\lambda_i[\rho_i(2-\rho)+\lambda_i s(1-\rho_i)]/\gamma_i}}{\sum_{j=1}^{N}\gamma_j\sqrt{c_j\lambda_j[\rho_j(2-\rho)+\lambda_j s(1-\rho_j)]/\gamma_j}}, \tag{11}$$

One may interpret (11) similarly to (7).

IV. The original Fuhrmann & Wang k-limited approximation.
Fuhrmann & Wang [23] also assume $\mathrm{E}\mathbf{RC}_i \approx \mathrm{E}\mathbf{RC}$ but they do not assume $\mathrm{E}\mathbf{RC} = B\,\mathrm{E}\mathbf{C}$.

Instead they approximate **ERC** by substituting (10) into (2), taking $g_i^{(2)} = 0$,

$$\mathrm{E}\mathbf{RC} \approx \frac{D + \frac{s}{1-\rho}\sum_{j=1}^{N}\frac{\rho_j^2}{k_j}}{\sum_{j=1}^{N}[\rho_j(1-\rho_j) + \frac{\rho_j^2}{k_j}\frac{2-\rho}{1-\rho}]}. \tag{12}$$

Taking $g_i^{(2)} = \max\{0, \left(\frac{\lambda_i s}{1-\rho}\right)^2 - \frac{\lambda_i s}{1-\rho}\}$ in (2) would probably improve the accuracy of (12). We did however not consider this option, as the numerical results in Section 4 reveal that the rule based on (12) performs already very well.
Substituting (12) back into (10),

$$\mathrm{E}\mathbf{W}_i \approx \frac{(1-\rho_i)(1-\rho) + \frac{\rho_i}{k_i}(2-\rho)}{1-\rho-\frac{\lambda_i s}{k_i}} \frac{D + \frac{s}{1-\rho}\sum_{j=1}^{N}\frac{\rho_j^2}{k_j}}{\sum_{j=1}^{N}[\rho_j(1-\rho_j) + \frac{\rho_j^2}{k_j}\frac{2-\rho}{1-\rho}]}. \tag{13}$$

The optimal service limits based on (13) cannot be solved analytically but have to be determined numerically.

In the next section we test the simple rules (7), (9), (11), and the rule based on (13).

Remark 3.1 Fuhrmann & Wang [23] concentrate on k-limited service under a *gated* regime at all queues. They use the reasoning leading to our approximation II, observe the discrepancy for $k_i = \infty$ (the reason for which is explained above (9)) and then modify their approximation in a way that amounts to our taking a weighted sum. Tedijanto [31] considers cyclic polling systems with a Bernoulli service policy. He proposes a mean waiting time approximation which coincides with (10) when one replaces the Bernoulli parameters q_i by $1 - 1/k_i$. His approximation is used by Blanc & Van der Mei [5] to find those q_i that minimize a weighted sum of the mean waiting times, cf. Remark 1.1.

□

Remark 3.2 Setting $\gamma_i = \beta_i$, $K = L - s$, i.e., imposing a limit L on the mean cycle time at periods of overload (namely, when all queues are loaded), (11) reduces to

$$k_i^*\beta_i = \frac{\rho_i s}{1-\rho} + (L - \frac{s}{1-\rho})\frac{\sqrt{c_i\rho_i[\rho_i(2-\rho)+\lambda_i s(1-\rho_i)]}}{\sum_{j=1}^{N}\sqrt{c_j\rho_j[\rho_j(2-\rho)+\lambda_j s(1-\rho_j)]}}. \tag{14}$$

One may interpret (14) as follows. The server should be allowed to visit Q_i at least for a time $\frac{\rho_i s}{1-\rho}$, to satisfy the stability condition. The remaining non-switchover time, $L - \frac{s}{1-\rho}$, should be assigned proportionally to $\sqrt{c_i\rho_i[\rho_i(2-\rho)+\lambda_i s(1-\rho_i)]}$. This suggests a rule for the optimal setting of time-limits in polling models with a time-limited service discipline. Note that in the case of constant service times, the k-limited and time-limited service disciplines coincide.

□

4 Numerical results for the constrained waiting cost minimization

For a wide variety of cases we compared the optimal values of the service limits and the waiting cost with the values achieved by the rules (7), (9), (11), and the rule based on (13) proposed in the previous section. Due to space limitations we give here only a brief overview of the numerical results gathered; more extensive numerical results are presented in [6].
To evaluate the mean waiting times we used the power series algorithm (psa). The psa allows an accurate numerical determination of the mean waiting times in polling models for which the joint queue length process has the structure of a multi-dimensional quasi birth-death process, cf. [3], [4]. (Alternatively, the mean waiting times for k-limited service could also be evaluated by the numerical approach developed by Leung [25].) The main drawback from which the psa and the approach of Leung [25] suffer is that the time and memory requirements grow exponentially with the number of queues. We therefore confined ourselves to cases with only a few queues. We have confidence however that the various approaches will perform at least as good for a larger number of queues. In Remark 5.3 we shall discuss the case of a large number of queues in some detail.
A further drawback from which the psa suffers is that the time and memory requirements grow rapidly with the number of stages of the service and switchover time distributions. For this reason, most of the numerical tests are conducted for cases with exponential service and switchover times. The following arguments support our belief that the results for other service and switchover time distributions will be similar in general. It should be noted that the k_i-values prescribed by the rules (7), (9), and (11) are completely insensitive to the form of the service time and switchover time distributions. The Fuhrmann & Wang approximation (13) suggests that $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$ depends on the second moments of these distributions mainly through $\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}$ and $\rho s^{(2)}/2s$ (both hidden in the term D), and that the influence of the k_j's on $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$ only marginally depends on the second moments. Thus the optimal k_j's will be almost insensitive to those second moments. Limited numerical experience with Erlang and hyperexponential service time distributions (cf. Tables 5, 6 and 8 of Section 6) support this view completely. Numerical experiments of Blanc & Van der Mei [5] for cyclic polling with the closely related Bernoulli service policy (cf. our Remark 1.1) also point in the direction of a robustness of the optimal policy w.r.t. service time distributions; a robustness that was also observed in designing optimal polling tables, cf. pp. 152, 153, and 161 of [11]. Finally it should be observed that while Fuhrmann & Wang [23] p. 50 test their approximation only for exponentially distributed service times, they state that 'limited experience indicates that the accuracy for other service time distributions seems to be similar in general'.

The numerical results are presented in Tables 1-3. Table 1 contains 7 two-queue cases, Table 2 contains 3 three-queue cases, and Table 3 presents a five-queue case. Most of the parameter combinations are taken from [11]. In the two-queue cases we imposed the constraint $k_1+k_2 \leq 12$, in the three-queue cases $k_1 + k_2 + k_3 \leq 12$, and in the five-queue case $k_1 + k_2 + \ldots + k_5 \leq 20$. The constraint may be interpreted as a limit on the maximal number of services in a cycle. The k_i-values for the rules (7), (9), and (11) were computed by rounding the values obtained from the corresponding formula to the nearest integers. The k_i-values based on (13) were calculated by an exhaustive search over the feasible integer vectors. The displayed cost figures are the *'exact'* waiting cost figures obtained from the psa. We have only displayed the percentage errors for the rule (11) and the rule based on (13), as in most cases they outperform the rules (7) and (9).

Table 1 Two-queue cases.

a. $\lambda_1 = \lambda_2 = 0.75$; $\beta_1 = 0.9$; $\beta_2 = 0.1$; $s_1 = s_2 = 0.1$; $\rho = 0.75$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(0.9, 0.1)	(8, 4)	2.169	(10, 2)	2.307	(10, 2)	2.307	(9, 3)	2.209	1.8	(8, 4)	2.169	0.0
(1, 0.1)	(8, 4)	2.355	(10, 2)	2.476	(10, 2)	2.476	(9, 3)	2.386	1.3	(9, 3)	2.386	1.3
(1, 0.2)	(6, 6)	2.676	(9, 3)	3.006	(9, 3)	3.006	(8, 4)	2.842	6.2	(7, 5)	2.738	2.3
(1, 0.5)	(4, 8)	3.176	(8, 4)	4.301	(8, 4)	4.301	(7, 5)	3.819	20.2	(4, 8)	3.176	0.0
(1, 2)	(3, 9)	4.972	(6, 6)	6.996	(6, 6)	6.996	(6, 6)	6.996	40.7	(2, 10)	5.033	1.2
(1, 5)	(2, 10)	7.693	(5, 7)	10.99	(5, 7)	10.99	(4, 8)	9.391	22.1	(2, 10)	7.693	0.0
(1, 10)	(2, 10)	12.13	(4, 8)	16.30	(4, 8)	16.30	(4, 8)	16.30	34.4	(1, 11)	12.80	5.5

b. $\lambda_1 = \lambda_2 = 0.8$; $\beta_1 = 0.9$; $\beta_2 = 0.1$; $s_1 = s_2 = 0.4$; $\rho = 0.8$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(1, 10)	(5, 7)	34.19	(6, 6)	76.59	(6, 6)	76.59	(5, 7)	34.19	0.0	(4, 8)	38.96	13.9
(1, 1)	(5, 7)	12.06	(7, 5)	17.69	(7, 5)	17.69	(7, 5)	17.69	46.6	(6, 6)	13.99	15.9
(10, 1)	(7, 5)	68.78	(8, 4)	70.43	(8, 4)	70.43	(8, 4)	70.43	2.4	(7, 5)	68.78	0.0

c. $\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1$; $s_1 = s_2 = 0.1$; $\rho = 0.85$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(1, 0.1)	(11, 1)	4.333	(11, 1)	4.333	(11, 1)	4.333	(11, 1)	4.333	0.0	(11, 1)	4.333	0.0
(1, 1)	(11, 1)	5.529	(11, 1)	5.529	(10, 2)	5.565	(11, 1)	5.529	0.0	(10, 2)	5.565	6.8
(1, 10)	(6, 6)	8.428	(10, 2)	10.29	(8, 4)	8.692	(9, 3)	9.055	7.4	(4, 8)	8.630	2.4

d. $\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1$; $s_1 = s_2 = 0.4$; $\rho = 0.85$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(1, 10)	(10, 2)	12.65	(11, 1)	26.02	(9, 3)	12.81	(10, 2)	12.65	0.0	(9, 3)	12.81	1.2
(1, 1)	(11, 1)	4.92	(11, 1)	4.92	(11, 1)	4.92	(11, 1)	4.92	0.0	(11, 1)	4.92	0.0
(10, 1)	(11, 1)	28.11	(11, 1)	28.11	(11, 1)	28.11	(11, 1)	28.11	0.0	(11, 1)	28.11	0.0

e. $\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1$; $s_1 = 0.1$; $s_2 = 0.7$; $\rho = 0.85$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(1, 10)	(10, 2)	13.00	(11, 1)	26.65	(9, 3)	13.14	(10, 2)	13.00	0.0	(9, 3)	13.14	1.0
(1, 1)	(11, 1)	5.060	(11, 1)	5.060	(11, 1)	5.060	(11, 1)	5.060	0.0	(11, 1)	5.060	0.0
(10, 1)	(11, 1)	29.01	(11, 1)	29.01	(11, 1)	29.01	(11, 1)	29.01	0.0	(11, 1)	29.01	0.0

f. $\lambda_1 = 0.5$; $\lambda_2 = 0.25$; $\beta_1 = \beta_2 = 1$; $s_1 = 0.1$; $s_2 = 0.2$; $\rho = 0.75$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(1, 7.6)	(3, 9)	6.278	(6, 6)	7.434	(5, 7)	6.885	(6, 6)	7.434	18.4	(2, 10)	6.293	0.2
(1, 3.95)	(3, 9)	4.537	(7, 5)	5.243	(6, 6)	4.813	(6, 6)	4.813	6.1	(3, 9)	4.537	0.0
(1, 0.7)	(10, 2)	2.265	(9, 3)	2.294	(9, 3)	2.294	(9, 3)	2.294	1.3	(9, 3)	2.294	1.3

g. $\lambda_1 = 0.5$; $\lambda_2 = 1$; $\beta_1 = 1$; $\beta_2 = 0.3$; $s_1 = 0.2$; $s_2 = 0.6$; $\rho = 0.8$.

	optimum		(7)		(9)		(11)			(13)		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	(k_1, k_2)	cost	%	(k_1, k_2)	cost	%
(1, 10)	(3, 9)	32.09	(3, 9)	32.09	(4, 8)	69.06	(4, 8)	69.06	115.2	(3, 9)	32.09	0.0
(1, 1)	(3, 9)	10.13	(5, 7)	12.18	(5, 7)	12.18	(5, 7)	12.18	20.3	(4, 8)	10.83	6.9
(10, 1)	(6, 6)	32.79	(6, 6)	32.79	(7, 5)	41.42	(7, 5)	41.42	26.3	(6, 6)	32.79	0.0

Table 2 Three-queue cases.

a. $\lambda_1 = \lambda_2 = \lambda_3 = 0.7$; $\beta_1 = 0.8$; $\beta_2 = \beta_3 = 0.1$; $s_1 = s_2 = s_3 = 0.05$; $\rho = 0.7$.

	optimum		(7)		(9)		(11)		(13)		
$(c_1, c_{2,3})$	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(3.5, 1)	(3, 4, 5)	7.004	(7, 2, 3)	8.547	(7, 2, 3)	8.547	(6, 3, 3)	7.717	(4, 4, 4)	7.198	2.8
(14.4, 1)	(9, 2, 1)	18.09	(8, 2, 2)	18.20	(8, 2, 2)	18.20	(8, 2, 2)	18.20	(8, 2, 2)	18.20	0.6
(51.8, 1)	(9, 2, 1)	50.50	(10, 1, 1)	51.80	(10, 1, 1)	51.80	(9, 1, 2)	52.07	(10, 1, 1)	51.80	2.6

b. $\lambda_1 = 0.531$; $\lambda_2 = 0.212$; $\lambda_3 = 0.106$; $\beta_1 = \beta_2 = \beta_3 = 0.9$; $s_1 = s_2 = s_3 = 0.3$; $\rho = 0.7641$.												
	optimum		(7)		(9)		(11)		(13)			
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(1, 0.1, 1)	(9, 1, 2)	2.660	(9, 2, 1)	2.780	(8, 2, 2)	2.668	(8, 2, 2)	2.668	(8, 2, 2)	2.668	0.3
(1, 5, 1)	(5, 5, 2)	7.796	(7, 4, 1)	8.498	(6, 5, 1)	8.541	(6, 5, 1)	8.541	(5, 6, 1)	8.055	3.3
(1, 1, 5)	(6, 3, 3)	6.252	(7, 3, 2)	6.918	(6, 3, 3)	6.252	(6, 3, 3)	6.252	(6, 3, 3)	6.252	0.0

Table 3 A five-queue case.

$\lambda_1 = 0.35$; $\lambda_2 = \ldots = \lambda_5 = 0.1$; $\beta_1 = 1$; $\beta_2 = \ldots = \beta_5 = 1$; $s_1 = 0.1$; $s_2 = \ldots = s_5 = 0.05$; $\rho = 0.75$.													
	optimum		(7)		(9)		(11)			(13)		
(c_1, c_{2-5})	(k_1, k_{2-5})	cost	(k_1, k_{2-5})	cost	(k_1, k_{2-5})	cost	(k_1, k_{2-5})	cost	%	(k_1, k_{2-5})	cost	%
(1, 0.1)	(16, 1)	0.883	(12, 2)	1.015	(16, 1)	0.883	(16, 1)	0.883	0.0	(16, 1)	0.883	0.0
(1, 0.5)	(16, 1)	1.804	(12, 2)	1.807	(8, 3)	1.987	(12, 2)	1.807	0.2	(16, 1)	1.804	0.0
(1, 1)	(4, 4)	2.420	(12, 2)	2.797	(8, 3)	2.934	(8, 3)	2.934	21.6	(8, 3)	2.934	21.6
(1, 2)	(4, 4)	3.631	(8, 3)	4.827	(4, 4)	3.631	(8, 3)	4.827	32.9	(4, 4)	3.631	0.0

Discussion of the numerical results.
The various rules perform reasonably well; in the majority of the examples the waiting cost achieved is less than 10% larger than the minimal waiting cost. It is however interesting to compare how the various rules perform. On average the rules (7) and (9) perform similarly. Sometimes (7) performs better, sometimes (9).
The rule (11) performs slightly better than the rule (9). The underlying approximation of (11) is theoretically indeed better than the underlying approximation of (9). The former shows the correct exact behavior when $k_i \to \infty$, the latter does not. As the rotation time constraint prevents that $k_i \to \infty$, the difference in performance is however minor.
The rule based on the approximation (13) performs by far the best; only 6 out of the 35 times the relative error exceeds 5%. The approximation (13) is indeed theoretically better than the underlying approximation of (11). The former catches the influence of k_i on $\mathrm{E}\mathbf{W}_j$, the latter does not. Thus the difference in performance will especially be dramatic in cases where that influence plays a crucial role, as is illustrated by the numerical results. When we take a closer look at e.g. the two-queue cases where the rules (7), (9), and (11) perform poorly, we notice that β_1 is typically larger than β_2 and c_1 is usually smaller than c_2. These rules, which completely ignore the influence of k_1 on $\mathrm{E}\mathbf{W}_2$, then choose k_1 too large and k_2 too small. These two-queue cases are typically cases where the influence of k_1 on $\mathrm{E}\mathbf{W}_2$ plays a crucial role: for large β_1 the influence of k_1 on $\mathrm{E}\mathbf{W}_2$ is large and for large c_2 this influence is heavily weighted in the waiting cost. Concluding, when very high accuracy is not needed, we recommend to use the simple rule (11); otherwise the rule based on (13) should be used.

5 Monotonicity properties and waiting cost minimization under no constraint

In the present section we study the problem of finding the (unconstrained) service limits $k_1, \ldots, k_N$ that minimize the waiting cost $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$. We first derive several monotonicity properties of polling systems with k-limited service and switchover periods. The main result is a (partially conjectured) rule stating that for minimizing the waiting cost in such systems the queues with the highest value of c_i/β_i should be assigned $k_i = \infty$, i.e., receive exhaustive service. This property is very similar to the well-known $c\mu$-rule derived for systems with no switchover periods and in which the server is free to move from queue to queue dynamically. We subsequently propose to use the Fuhrmann & Wang approximation for the unconstrained waiting cost minimization. We specifically investigate to what extent the Fuhrmann & Wang approximation satisfies the above-mentioned properties.

Proposition 1 *In a stable polling system with cyclic visit order and k-limited service the sum* $\sum_{j=1}^{N} \rho_j \mathrm{E}\mathbf{W}_j$ *is non-increasing in each of the service limits* k_i, $i = 1, \ldots, N$.

Proof: Let $\mathbf{V}_t$ be the total amount of work in the system at time t. Let $\mathbf{V}$ be a random variable with distribution the steady-state distribution of the total amount of work in the system. As shown in [28] $\mathbf{V}_t$ is non-increasing in k_i (the proof in [28] is a path-wise proof). Hence $\mathrm{E}\mathbf{V}$ is non-increasing in k_i. Now, it is known (e.g. [7], [9]) that

$$\mathrm{E}\mathbf{V} = \sum_{j=1}^{N} \rho_j \mathrm{E}\mathbf{W}_j + \sum_{j=1}^{N} \rho_j \frac{\beta_j^{(2)}}{2\beta_j}$$

and thus $\sum_{j=1}^{N} \rho_j \mathrm{E}\mathbf{W}_j$ is non-increasing in k_i.

□

Conjecture 2 *In a stable polling system with cyclic visit order and k-limited service the mean waiting time at* Q_i, $\mathrm{E}\mathbf{W}_i$, *is decreasing in its service limit* k_i *and increasing in* k_j *for every* $j \neq i$.

While the claim made in Conjecture 2 is very appealing and intuitive, it seems difficult to prove it. A reasonable line of argument can nonetheless be provided as follows. To see the effect of k_j on $\mathrm{E}\mathbf{W}_i$ one can view the services given at Q_j as switchover periods (whose durations are distributed as the sum of several independent random variables, whose number is the number of customers being served at Q_j). It is easy to see that as long as the system is stable the mean number of services given at Q_j per visit is constant ($\frac{\lambda_j s}{1-\rho}$, which does not depend on k_j). On the other hand, the second moment of the number of services *does* depend on k_j. Increasing k_j will increase the number of services given at Q_j when that queue is loaded (has more than k_j customers), and thus is likely to decrease the number of services given at Q_j when it has less than k_j customers. This implies that the variance of the number of services given at Q_j increases with k_j.
Going back to the viewpoint of Q_i, we see that when k_j increases, the customers at Q_i observe switchover periods of the same mean but of higher variance. Viewing Q_i as an $M/G/1$ queue with vacations (where the services of the other queues and the switchover periods together constitute one large vacation), it is likely that increasing the vacation second moment while keeping its mean the same increases the mean waiting time at Q_i. We thus conclude that the mean waiting time at Q_i is increasing in k_j for any $j \neq i$. Combining this fact with Proposition 1 implies that the mean waiting time at Q_i must be decreasing in k_i.

Remark 5.1 The assumption that the system is stable plays an essential role in Conjecture 2. If Q_j is unstable, then increasing k_j will not only increase the variance of the intervisit time of Q_i, but also the mean. If the increment of the first moment is larger than the increment of the second moment, then the residual intervisit time of Q_i will *decrease*. The mean waiting time at Q_i will then also *decrease*.

□

Proposition 1 and Conjecture 2 lead to the main result of this section, namely that for optimality at least one of the queues, viz. the one whose c_i/β_i achieves the maximum value, must be served without limits.

Theorem 3 *If* $c_i/\beta_i = \max_{j=1,\ldots,N} c_j/\beta_j$, *then the optimal service limit* $k_i^* = \infty$, *provided Conjecture 2 holds true.*

Proof: We need to show that for every $k = (k_1, \ldots, k_N)$ the sum $\sum_{j=1}^{N} c_j \lambda_j \mathrm{E}\mathbf{W}_j$ is decreasing in k_i, namely that $\frac{\Delta}{\Delta k_i} \sum_{j=1}^{N} \lambda_j c_j \mathrm{E}\mathbf{W}_j \leq 0$ ($\frac{\Delta}{\Delta k_i}$ denoting partial difference):

$$\begin{aligned}
\frac{\Delta}{\Delta k_i} \sum_{j=1}^{N} \lambda_j c_j \mathrm{E}\mathbf{W}_j &= \frac{c_i}{\beta_i} \frac{\Delta}{\Delta k_i} \rho_i \mathrm{E}\mathbf{W}_i + \sum_{j \neq i} \frac{c_j}{\beta_j} \frac{\Delta}{\Delta k_i} \rho_j \mathrm{E}\mathbf{W}_j \\
&\leq \frac{c_i}{\beta_i} \frac{\Delta}{\Delta k_i} \rho_i \mathrm{E}\mathbf{W}_i + \frac{c_i}{\beta_i} \sum_{j \neq i} \frac{\Delta}{\Delta k_i} \rho_j \mathrm{E}\mathbf{W}_j \\
&= \frac{c_i}{\beta_i} \frac{\Delta}{\Delta k_i} \sum_{j=1}^{N} \rho_j \mathrm{E}\mathbf{W}_j \leq 0.
\end{aligned}$$

The inequality in the second line follows from the facts that $\frac{\Delta}{\Delta k_i} \mathrm{E}\mathbf{W}_j$ is non-negative (assuming Conjecture 2 holds) and that $c_i/\beta_i \geq c_j/\beta_j$ (condition of the theorem). The inequality in the third line follows from Proposition 1.

□

Remark 5.2 Theorem 3 implies that if the service limit policies are of the gated-limited type (namely, serve up to k_i customers but only of those present at the queue at the polling instant), then the queues with the maximal value of c_i/β_i should be served according to the gated policy with $k_i^* = \infty$. If the service limit policies are of the exhaustive-limited type (namely serve up to k_i customers but allow to include in these services customers that arrived during the service of the queue), then the queues with the maximal value of c_i/β_i should be served according to the exhaustive policy with $k_i^* = \infty$. This reminds of the $c\mu$-rule derived for systems with no switchover periods and in which the server is free to move from queue to queue. According to the $c\mu$-rule the queues with the highest value of c_i/β_i should receive the highest priority in the system, which implies, in particular, an exhaustive service at those queues.

□

Having in mind the properties discussed above we now study the problem of finding the service limits $k_1, \ldots, k_N$ that minimize the waiting cost $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$. As observed in the previous section, for the constrained waiting cost minimization the Fuhrmann & Wang approximation outperforms the simpler approximations. For the unconstrained waiting cost minimization the simpler approximations would be useless anyhow, as they completely ignore the influence of k_j on $\mathrm{E}\mathbf{W}_i$, which would always lead to $k_1^* = \infty, \ldots, k_N^* = \infty$. Therefore we restrict ourselves here to the Fuhrmann & Wang approximation:

$$\mathrm{E}\mathbf{W}_i \approx \frac{(1-\rho_i)(1-\rho) + \frac{\rho_i}{k_i}(2-\rho)}{1 - \rho - \frac{\lambda_i s}{k_i}} \frac{D + \frac{s}{1-\rho} \sum_{j=1}^{N} \frac{\rho_j^2}{k_j}}{\sum_{j=1}^{N} [\rho_j(1-\rho_j) + \frac{\rho_j^2}{k_j} \frac{2-\rho}{1-\rho}]}. \tag{15}$$

We now specifically investigate to what extent the Fuhrmann & Wang approximation (15) satisfies the properties discussed above.

Proposition 4 *The approximation (15) of* $\mathrm{E}\mathbf{W}_i$ *is: a) decreasing in* k_i*, and b) increasing in* k_j*,* $j \neq i$*.*

Proof: A straightforward computation shows that $\mathrm{E}\mathbf{W}_i \mid_{k_i=r} \geq \mathrm{E}\mathbf{W}_i \mid_{k_i=r+1}$ and that $\mathrm{E}\mathbf{W}_i \mid_{k_j=r} \leq \mathrm{E}\mathbf{W}_i \mid_{k_j=r+1}$, $j \neq i$. (The latter inequality holds provided $D \geq \frac{s}{2-\rho} \sum_{j=1}^{N} \rho_j(1-\rho_j)$, which may be shown to hold by substitution of the definition of D.)

□

Proposition 4 supports the use of (15) in trying to obtain the optimal service limit values for the actual polling system. Moreover, in the numerical experiments that will be presented in the next section we will find that the minimal value of $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$, where (15) is used to evaluate $\mathrm{E}\mathbf{W}_i$, is always achieved when the service limit of the queue with the maximal value of c_i/β_i is set to $k_i = \infty$ (exhaustive service). (Here the service limit is set to $k_i = \infty$ when this yields a smaller value for $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$ according to (15) than all values $k_i = 1, \ldots, 20$. Similarly, the true optimal service limit is supposed to be $k_i = \infty$ when the psa produces for exhaustive service at Q_i a smaller value for $\sum_{i=1}^{N} c_i \lambda_i \mathrm{E}\mathbf{W}_i$ than for all values $k_i = 1, \ldots, 20$.) These findings suggest that the approximation possesses the property derived in Theorem 3 for the real polling system. This can however not be proved along the same lines as Theorem 3, as the approximation does not always possess the property derived in Proposition 1.

Remark 5.3 In the numerical experiments for both the constrained case (Section 4) and the unconstrained case (Section 6), time and memory requirements of the psa have forced us to confine ourselves to models with only a few queues. Let us now discuss what happens when the number of queues, N, approaches infinity, distinguishing four cases: for all j:

I. s_j fixed, $\beta_j = O(1/N)$, λ_j fixed;
II. $s_j = O(1/N)$, β_j fixed, $\lambda_j = O(1/N)$;
III. $s_j = O(1/N)$, $\beta_j = O(1/N)$, λ_j fixed;
IV. s_j fixed, β_j fixed, $\lambda_j = O(1/N)$.

In case I, $\lambda_i s/(1-\rho) \to \infty$ and hence necessarily $k_i \to \infty$; this is not an interesting case. Case II reduces to continuous polling on a circle; cf. Fuhrmann & Cooper [22]. Each customer will be served in the cycle in which it arrives, even if the k_i values equal one; the actual choice of the k_i is irrelevant. Cases III and IV are equivalent up to a scaling of time by a factor N. Let us discuss case III in more detail. For the constrained situation, (8), (10) and (13) all reduce to

$$\mathrm{E}\mathbf{W}_i \approx B \frac{1-\rho}{1-\rho-\frac{\lambda_i s}{k_i}}, \tag{16}$$

with B some constant, leading to

$$k_i = \frac{\lambda_i s}{1-\rho} + (K - \sum_{j=1}^{N} \gamma_j \frac{\lambda_j s}{1-\rho}) \frac{\lambda_i \sqrt{c_i/\gamma_i}}{\sum_{j=1}^{N} \gamma_j \lambda_j \sqrt{c_j/\gamma_j}}. \tag{17}$$

Note that the weakness of approximation (8), indicated above (9), disappears when $N \to \infty$. Approximation (16) may be expected to perform very well. For the unconstrained situation, (15) also reduces to (16). Hence the waiting cost is minimized by taking $k_i = \infty$ for all i. Indeed, for large finite N an increment of k_i by one reduces $c_i \lambda_i \mathrm{E}\mathbf{W}_i$ much stronger than it increases $\sum_{j \neq i} c_j \lambda_j \mathrm{E}\mathbf{W}_j$, as is indicated by the following rough reasoning.

To make things simple, let us assume that $k_1 < \infty$; $k_2 = \ldots = k_N = \infty$; now increase k_1 by one. Customers in Q_1 only notice this increment at a server visit when at least $k_1 + 1$ customers are present. Suppose such an event occurs in the nth cycle. Now this saves one Q_1 customer one cycle time $\mathbf{C}_{1,n+1}$, which is $O(1)$. What is the effect on some other queue Q_j? First the bad effect. S reaches Q_j Δ_j later; this delay consists of a service time at Q_1 (of $O(1/N)$) and of extended visit times at $Q_2, \ldots, Q_{j-1}$; $\Delta_j = O(1/N)$. Each of the customers at Q_j experiences this additional delay as an addition to its waiting time. There are on the average $\lambda_j \mathrm{E}\mathbf{C}_{j,n}$ such customers. Here $\mathbf{C}_{j,n}$ denotes the nth cycle time for Q_j. The total mean 'loss' for Q_j is $\Delta_j \lambda_j \mathrm{E}\mathbf{C}_{j,n}$. Here we ignore an $O(N^{-2})$ contribution: compared to an ordinary cycle, this one lasted already Δ_j longer, during which additional period also on the average $\lambda_j \Delta_j$ customers have arrived who each experience an extra delay Δ_j. Now there is also a *benefit* for Q_j. During the extra delay also customers arrive at Q_j who are just in time to be served in *this* cycle; without the extra delay they would have arrived just after the departure of S from Q_j and would have had to wait a full cycle. The total mean 'gain' for Q_j is $\Delta_j \lambda_j \mathrm{E}\mathbf{C}_{j,n+1} + O(N^{-2})$. The result on Q_j of these two counteracting effects is $\lambda_j \Delta_j (\mathrm{E}\mathbf{C}_{j,n} - \mathrm{E}\mathbf{C}_{j,n+1}) + O(N^{-2})$ (the propagation of an extra service in Q_1 in later cycles should also have an $O(N^{-2})$ effect). It seems obvious that $\mathrm{E}\mathbf{C}_{j,n} - \mathrm{E}\mathbf{C}_{j,n+1} = O(1)$ and likely that $\mathrm{E}\mathbf{C}_{j,n} - \mathrm{E}\mathbf{C}_{j,n+1} = O(1/N)$. In the latter case increasing k_1 by one has an $O(N^{-2})$ effect on Q_j, which agrees with (16).

$\square$

6 Numerical results for the unconstrained waiting cost minimization

For a wide variety of cases we compared the optimal values of the service limits and the waiting cost with the values achieved by the approximative approach proposed in the previous section. Due to space limitations we present here only a small subset of the numerical results obtained; more extensive numerical results are reported in [6].
Just like in Section 4 we used the power series algorithm (psa) to evaluate the mean waiting times and we confined ourselves to cases with only a few queues. We further focused again on cases with an exponential service and switchover time distribution, although we did investigate some cases with Erlang and hyperexponential service time distributions as well. The results for Erlang and hyperexponential service time distributions appear to be similar to the results for an exponential service time distribution.

The numerical results are presented in Tables 4-9. Table 4 contains the same 7 two-queue cases as Table 1 of Section 4, Table 5 contains a two-queue case with exponential service times at Q_1 and hyperexponential service times at Q_2, Table 6 contains a two-queue case with hyperexponential service time distributions at both queues, Table 7 contains 4 three-queue cases with exponential service time distributions, Table 8 contains the same three-queue cases but with Erlang-2 (a, b, and c) and Erlang-3 (d) service time distributions, and Table 9 contains the same five-queue case as Table 3 of Section 4. In case of hyperexponential distributions we assumed that the service times are exponentially distributed with mean either 0.5 or 1.5, both with probability 0.5. The displayed cost figures are the *'exact'* waiting cost figures obtained from the psa.

Table 4 Two-queue cases; exponential service times.

a. $\lambda_1 = \lambda_2 = 0.75$; $\beta_1 = 0.9$; $\beta_2 = 0.1$; $s_1 = s_2 = 0.1$; $\rho = 0.75$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 17)$	2.158	(∞, ∞)	2.162	0.2
(1, 0.2)	$(12, \infty)$	2.535	$(20, \infty)$	2.558	0.9
(1, 0.5)	$(5, \infty)$	3.119	$(5, \infty)$	3.119	0.0
(1, 1)	$(4, \infty)$	3.817	$(3, \infty)$	3.836	0.5
(1, 2)	$(3, \infty)$	4.949	$(2, \infty)$	5.030	1.6
(1, 5)	$(2, \infty)$	7.685	$(2, \infty)$	7.685	0.0
(1, 10)	$(2, \infty)$	12.11	$(1, \infty)$	12.80	5.7

b. $\lambda_1 = \lambda_2 = 0.8$; $\beta_1 = 0.9$; $\beta_2 = 0.1$; $s_1 = s_2 = 0.4$; $\rho = 0.8$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	(∞, ∞)	3.600	(∞, ∞)	3.600	0.0
(1, 1)	$(11, \infty)$	8.174	$(13, \infty)$	8.207	0.4
(1, 10)	$(5, \infty)$	32.83	$(5, \infty)$	32.83	0.0

c. $\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1$; $s_1 = s_2 = 0.1$; $\rho = 0.85$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 1)$	3.139	$(\infty, 1)$	3.139	0.0
(1, 1)	(∞, ∞)	5.031	(∞, ∞)	5.031	0.0
(1, 10)	$(6, \infty)$	8.426	$(4, \infty)$	8.630	2.4

d. $\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1$; $s_1 = s_2 = 0.4$; $\rho = 0.85$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 2)$	3.685	$(\infty, 1)$	3.730	1.2
(1, 1)	(∞, ∞)	5.673	(∞, ∞)	5.673	0.0
(1, 10)	$(14, \infty)$	12.59	$(14, \infty)$	12.59	0.0

e. $\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1$; $s_1 = 0.1$; $s_2 = 0.7$; $\rho = 0.85$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 2)$	3.740	$(\infty, 1)$	3.789	1.3
(1, 1)	(∞, ∞)	5.769	(∞, ∞)	5.769	0.0
(1, 10)	$(14, \infty)$	12.86	$(14, \infty)$	12.86	0.0

f. $\lambda_1 = 0.5$; $\lambda_2 = 0.25$; $\beta_1 = \beta_2 = 1$; $s_1 = 0.1$; $s_2 = 0.2$; $\rho = 0.75$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 1)$	1.137	$(\infty, 1)$	1.137	0.0
(1, 1)	(∞, ∞)	2.575	(∞, ∞)	2.575	0.0
(1, 10)	$(2, \infty)$	7.262	$(2, \infty)$	7.262	0.0

g. $\lambda_1 = 0.5$; $\lambda_2 = 1$; $\beta_1 = 1$; $\beta_2 = 0.3$; $s_1 = 0.2$; $s_2 = 0.6$; $\rho = 0.8$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 20)$	2.188	(∞, ∞)	2.342	7.0
(1, 1)	$(8, \infty)$	6.611	$(15, \infty)$	6.928	4.8
(1, 10)	$(3, \infty)$	31.66	$(3, \infty)$	31.66	0.0

Table 5 A two-queue case; exponential service times at Q_1 and hyperexponential service times at Q_2.

$\lambda_1 = 0.675$; $\lambda_2 = 0.075$; $\beta_1 = \beta_2 = 1.0$; $s_1 = s_2 = 0.1$; $\rho = 0.75$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 1)$	2.026	$(\infty, 1)$	2.026	0.0
(1, 0.2)	$(\infty, 1)$	2.118	$(\infty, 1)$	2.118	0.0
(1, 0.5)	$(\infty, 1)$	2.391	$(\infty, 1)$	2.391	0.0
(1, 1)	(∞, ∞)	2.787	(∞, ∞)	2.787	0.0
(1, 2)	$(10, \infty)$	3.217	$(17, \infty)$	3.252	1.1
(1, 5)	$(5, \infty)$	3.920	$(7, \infty)$	3.977	1.5
(1, 10)	$(3, \infty)$	4.785	$(4, \infty)$	4.805	0.4

Table 6 A two-queue case; hyperexponential service times.

$\lambda_1 = 0.765$; $\lambda_2 = 0.085$; $\beta_1 = \beta_2 = 1.0$; $s_1 = s_2 = 0.1$; $\rho = 0.85$.

	optimum		approximation		
(c_1, c_2)	(k_1, k_2)	cost	(k_1, k_2)	cost	%
(1, 0.1)	$(\infty, 1)$	3.889	$(\infty, 1)$	3.889	0.0
(1, 0.2)	$(\infty, 1)$	4.188	$(\infty, 1)$	4.188	0.0
(1, 0.5)	$(\infty, 2)$	5.034	$(\infty, 1)$	5.087	1.1
(1, 1)	(∞, ∞)	6.235	(∞, ∞)	6.235	0.0
(1, 2)	$(17, \infty)$	7.157	$(17, \infty)$	7.157	0.0
(1, 5)	$(9, \infty)$	8.501	$(7, \infty)$	8.547	0.5
(1, 10)	$(6, \infty)$	10.085	$(4, \infty)$	10.432	3.4

Table 7 Three-queue cases; exponential service times.

a. $\lambda_1 = \lambda_2 = \lambda_3 = 0.25$; $\beta_1 = 0.2$; $\beta_2 = 0.6$; $\beta_3 = 2.2$; $s_1 = s_2 = s_3 = 0.1$; $\rho = 0.75$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	$(\infty, \infty, 2)$	34.26	$(\infty, \infty, 2)$	34.26	0.0
(10, 3, 3)	$(\infty, 8, 1)$	14.71	$(\infty, \infty, 1)$	14.72	0.0
(10, 1, 1)	$(\infty, 3, 1)$	8.26	$(\infty, \infty, 1)$	8.28	0.3

b. $\lambda_1 = 0.6$; $\lambda_2 = 0.2$; $\lambda_3 = 0.05$; $\beta_1 = 0.2$; $\beta_2 = 0.6$; $\beta_3 = 2.2$; $s_1 = s_2 = s_3 = 0.5$; $\rho = 0.35$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	$(\infty, \infty, 1)$	14.78	$(\infty, \infty, 1)$	14.78	0.0
(10, 3, 3)	$(\infty, 3, 1)$	11.47	$(\infty, \infty, 1)$	11.49	0.2
(10, 1, 1)	$(\infty, 2, 1)$	10.41	$(\infty, 1, 1)$	10.54	1.2

c. $\lambda_1 = 0.6$; $\lambda_2 = 0.2$; $\lambda_3 = 0.05$; $\beta_1 = \beta_2 = \beta_3 = 1$; $s_1 = s_2 = s_3 = 0.1$; $\rho = 0.85$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	(∞, ∞, ∞)	53.05	(∞, ∞, ∞)	53.05	0.0
(10, 3, 3)	$(\infty, 3, 1)$	30.04	$(\infty, 2, 1)$	30.13	0.3
(10, 1, 1)	$(\infty, 2, 1)$	21.26	$(\infty, 1, 1)$	21.35	0.4

d. $\lambda_1 = 0.3$; $\lambda_2 = 0.8$; $\lambda_3 = 0.1$; $\beta_1 = 0.2$; $\beta_2 = 0.5$; $\beta_3 = 2$; $s_1 = 2$; $s_2 = 0.1$; $s_3 = 0.5$; $\rho = 0.66$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	$(\infty, \infty, 2)$	63.47	$(\infty, \infty, 4)$	65.59	3.3
(10, 3, 3)	$(\infty, \infty, 2)$	31.60	$(\infty, \infty, 2)$	31.60	0.0
(10, 1, 1)	$(\infty, 16, 2)$	22.46	$(\infty, \infty, 2)$	22.50	0.2

Table 8 Three-queue cases; Erlang service times.

a. $\lambda_1 = \lambda_2 = \lambda_3 = 0.25$; $\beta_1 = 0.2$; $\beta_2 = 0.6$; $\beta_3 = 2.2$; $s_1 = s_2 = s_3 = 0.1$; $\rho = 0.75$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	$(\infty, \infty, 2)$	27.23	$(\infty, \infty, 2)$	27.23	0.0
(10, 3, 3)	$(\infty, 8, 1)$	11.70	$(\infty, \infty, 1)$	11.70	0.0
(10, 1, 1)	$(\infty, 3, 1)$	6.635	$(\infty, \infty, 1)$	6.652	0.3

b. $\lambda_1 = 0.6$; $\lambda_2 = 0.2$; $\lambda_3 = 0.05$; $\beta_1 = 0.2$; $\beta_2 = 0.6$; $\beta_3 = 2.2$; $s_1 = s_2 = s_3 = 0.5$; $\rho = 0.35$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	$(\infty, \infty, 1)$	13.79	$(\infty, \infty, 1)$	13.79	0.0
(10, 3, 3)	$(\infty, 3, 1)$	10.72	$(\infty, \infty, 1)$	10.74	0.1
(10, 1, 1)	$(\infty, 2, 1)$	9.751	$(\infty, 1, 1)$	9.875	1.3

c. $\lambda_1 = 0.6$; $\lambda_2 = 0.2$; $\lambda_3 = 0.05$; $\beta_1 = \beta_2 = \beta_3 = 1$; $s_1 = s_2 = s_3 = 0.1$; $\rho = 0.85$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	(∞, ∞, ∞)	41.03	(∞, ∞, ∞)	41.03	0.0
(10, 3, 3)	$(\infty, 2, 1)$	23.57	$(\infty, 2, 1)$	23.57	0.0
(10, 1, 1)	$(\infty, 1, 1)$	16.67	$(\infty, 1, 1)$	16.67	0.0

d. $\lambda_1 = 0.3$; $\lambda_2 = 0.8$; $\lambda_3 = 0.1$; $\beta_1 = 0.2$; $\beta_2 = 0.5$; $\beta_3 = 2$; $s_1 = 2$; $s_2 = 0.1$; $s_3 = 0.5$; $\rho = 0.66$.					
	optimum		approximation		
(c_1, c_2, c_3)	(k_1, k_2, k_3)	cost	(k_1, k_2, k_3)	cost	%
(10, 10, 10)	$(\infty, \infty, 3)$	55.40	$(\infty, \infty, 5)$	56.39	1.8
(10, 3, 3)	$(\infty, \infty, 2)$	28.69	$(\infty, \infty, 2)$	28.69	0.0
(10, 1, 1)	$(\infty, 16, 2)$	20.48	$(\infty, \infty, 2)$	20.98	2.4

Table 9 A five-queue case; exponential service times.

$\lambda_1 = 0.35$; $\lambda_2 = \ldots = \lambda_5 = 0.1$; $\beta_1 = 1$; $\beta_2 = \ldots = \beta_5 = 1$; $s_1 = 0.1$; $s_2 = \ldots = s_5 = 0.05$; $\rho = 0.75$.					
	optimum		approximation		
(c_1, c_{2-5})	(k_1, k_{2-5})	cost	(k_1, k_{2-5})	cost	%
(1, 0.1)	$(\infty, 1)$	0.882	$(\infty, 1)$	0.882	0.0
(1, 0.5)	$(\infty, 3)$	1.776	$(\infty, 4)$	1.776	0.2
(1, 1)	(∞, ∞)	2.263	(∞, ∞)	2.623	0.0
(1, 2)	$(3, \infty)$	3.898	$(4, \infty)$	3.898	0.0

Discussion of the numerical results.
The proposed approach performs extremely well; in the majority of the 67 examples the achieved waiting cost is less than 1% larger than the minimal waiting cost. Only twice the achieved waiting cost is more than 5% larger than the minimal waiting cost, not once more than 10% larger. The optimal service limits as well as the service limits obtained from the Fuhrmann & Wang approximation always satisfied the property stated in Theorem 3, i.e., if $c_i/\beta_i = \max_{j=1,\ldots,N} c_j/\beta_j$, then $k_i^* = \infty$. Recall that neither the optimal service limits nor the service limits obtained from the approximation were actually proved to satisfy the property stated in Theorem 3.
The results for Erlang and hyperexponential service time distributions are similar to the results for an exponential service time distribution. The waiting cost for an Erlang (hyperexponential) service time distribution is always smaller (larger) than the waiting cost for an exponential service time distribution with the same mean. Intuitively the waiting times are indeed likely to be smaller (larger) when the variance of the service time distribution is smaller (larger). The optimal service limits for Erlang and hyperexponential service time distributions however hardly differ from the optimal service limits for an exponential service time distribution with the same mean.

7 Conclusion

We have studied the problem of finding the optimal service limits in a cyclic polling system with the k-limited service discipline. The use of the Fuhrmann & Wang approximation is shown to be very effective in finding the optimal service limits. In the numerical experiments we have observed that the waiting cost according to the Fuhrmann & Wang approximation sometimes differs dramatically from the 'true' waiting cost obtained from the psa, but that still the optimal service limits according to the Fuhrmann & Wang approximation agree with the 'true' optimal service limits obtained from the psa. Even when completely misjudging the mean waiting time, the Fuhrmann & Wang approximation apparently *does* capture the major factors important for efficient operation of the system.
In this context it is worth noting that there are also some other approximation procedures for k-limited service available, like the one proposed by Chang & Sandhu [14], that are more accurate than the Fuhrmann & Wang approximation. In principle such more sophisticated approximation procedures may be used for optimization purposes as well. As they are more involved they will however also demand more complicated optimization techniques. Moreover, they will yield only marginally better results, as the results obtained from the Fuhrmann & Wang approximation already tend to be very close to the true optimal value.
The mean waiting time approximation for the Bernoulli service discipline that Blanc & Van der Mei [5] use to find the optimal Bernoulli parameters q_i, and the Fuhrmann & Wang approximation that we use, coincide when $q_i = 1 - 1/k_i$, cf. Remark 1.1 and Remark 3.1. The effectiveness of both approximations suggests that, as far as optimization is concerned, the Bernoulli service discipline is a very good emulation of the k-limited service discipline. Yet, as far as evaluation of the mean waiting time is concerned, the Bernoulli service discipline is often not a very good ap-

proximation of the k-limited service discipline. In general the stochastic nature of the Bernoulli service discipline tends to cause the mean waiting times to be larger than for the k-limited service discipline, cf. Tedijanto [30], [31] Chapter 5.
In the present study we have been concerned with optimization of the *service discipline*, ranging from 1-limited to exhaustive, for a given cyclic server routing. Earlier studies mostly were concerned with optimization of the *server routing* for a given service discipline, like 1-limited, gated, or exhaustive, cf. Remark 1.1. We feel that it would also be worthwhile to consider simultaneous optimization of the server routing and the service discipline. Simultaneous optimization of the number of visits and the amount of service per visit would enable more flexible priorization of the various stations.
At a few instances we faced difficult monotonicity questions: monotonicity of $\mathrm{E}\mathbf{W}_i$ in k_j, monotonicity of the mean waiting time for an $M/G/1$ queue with vacations in the vacation time variance. Relatively few monotonicity results for polling and vacation models have been obtained; this seems an interesting area for research.
In the present study we have focused on a static setting. An interesting topic for further research might be to use the insights obtained here to investigate the optimal control of a polling system with k-limited service in a dynamic context, e.g., fluctuating arrival rates, varying service limits. Evidently, in principle a dynamic control scheme may improve the performance of the system substantially, although probably little performance will be lost by simply setting the k_i-values according to a heavy-traffic scenario, as the gain from tightening the k_i-values in light-traffic conditions will be modest. Furthermore, keeping track of the arrival rates and queue lengths, and implementing a sophisticated dynamic control scheme may involve a considerable measurement and communication overhead, and may complicate the operation of the system significantly. Therefore, dynamic control is not necessarily preferable to static control. Moreover, the combination of dynamic control with k-limited service may not be so natural as the motives behind k-limited reflect rather a static point-of view: (i) to effectuate some kind of priorization and (ii) to accomplish an upper bound on the rotation time in temporary periods of overload.

Acknowledgment We are grateful to J.P.C. Blanc for providing the numerical procedures and programs of the power series algorithm, and to D. Aiger and A. Barel for their assistance in running the numerical tests. We also thank E. Altman and Y. Levy for helpful discussions.

References

[1] ANSI Standard, *FDDI Token Ring – Media Access Control* (ANSI X3.139-1987, 1986).

[2] ANSI/IEEE Standard 802.4, *Token-Passing Bus Access Method* (IEEE Press, New York, 1985).

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

[4] Blanc, J.P.C. [1991]. The power-series algorithm applied to cyclic polling systems. *Commun. Stat. - Stoch. Mod.* **7**, 527-545.

[5] Blanc, J.P.C., Van der Mei, R.D. [1994]. Optimization of polling systems with Bernoulli schedules. To appear in *Perf. Eval.*

[6] Borst, S.C., Boxma, O.J., Levy, H. [1993]. The use of service limits for efficient operation of multi-station single-medium communication systems. CWI Report BS-R9312.

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

[8] Boxma, O.J. [1991]. Analysis and optimization of polling systems. In: *Queueing, Performance and Control in ATM*, eds. J.W. Cohen and C.D. Pack, North-Holland Publ. Cy., Amsterdam, 173-183.

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

[10] Boxma, O.J., Groenendijk, W.P. [1988]. Two queues with alternating service and switching times. In: *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen*, eds. O.J. Boxma and R. Syski, North-Holland Publ. Cy., Amsterdam, 261-282.

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

[12] Boxma, O.J., Levy, H., Weststrate, J.A. [1993]. Efficient visit orders for polling systems. *Perf. Eval.* **18**, 103-123.

[13] Boxma, O.J., Meister, B.W. [1987]. Waiting-time approximations for cyclic-service systems with switchover times. *Perf. Eval.* **7**, 299-308.

[14] Chang, K.C., Sandhu, D. [1992]. Mean waiting time approximations in cyclic-service systems with exhaustive limited service policy. *Perf. Eval.* **15**, 21-40.

[15] Coffman, E.G., Jr., Mitrani, I., Fayolle, G. [1988]. Two queues with alternating service periods. In: *Performance '87*, eds. P.-J. Courtois and G. Latouche, North-Holland Publ. Cy., Amsterdam, 227-239.

[16] Cohen, J.W., Boxma, O.J. [1981]. The $M/G/1$ queue with alternating service formulated as a Riemann-Hilbert boundary value problem. In: *Performance '81*, ed. F.J. Kylstra (North-Holland, Amsterdam), 181-199.

[17] De Souza e Silva, E., Gail, H.R., Muntz, R.R. [1994]. Polling systems with server timeouts. Preprint.

[18] Eisenberg, M. [1979]. Two queues with alternating service. *SIAM J. Appl. Math.* **36**, 287-303.

[19] Everitt, D.E. [1986]. A conservation-type law for the token ring with limited service. *Br. Telecom Technol. J.* **4**, 51-61.

[20] Everitt, D.E. [1989]. An approximation procedure for cyclic service queues with limited service. In: *Performance of Distributed and Parallel Systems*, eds. T. Hasegawa, H. Takagi, and Y. Takahashi, North-Holland Publ. Cy., Amsterdam, 141-156.

[21] Fricker, C., Jaïbi, M.R. [1994]. Monotonicity and stability of periodic polling models. *Queueing Systems* **15**, 211-238.

[22] Fuhrmann, S.W., Cooper, R.B. [1985]. Application of decomposition principle in $M/G/1$ vacation model to two continuum cyclic queueing models - especially token-ring LANs. *AT & T Techn. J.* **64**, 1091-1099.

[23] Fuhrmann, S.W., Wang, Y.T. [1988]. Analysis of cyclic service systems with limited service: bounds and approximations. *Perf. Eval.* **9**, 35-54.

[24] Groenendijk, W.P. [1990]. *Conservation Laws in Polling Systems.* PhD Thesis, University of Utrecht, Utrecht.

[25] Leung, K.K. [1991]. Cyclic-service systems with probabilistically-limited service. *IEEE J. Sel. Areas Commun.* **9**, 185-193.

[26] Leung, K.K. [1994]. Cyclic-service systems with non-preemptive, time-limited service. *IEEE Trans. Commun.*, 2521-2524.

[27] Leung, K.K., Lucantoni, D.M. [1994]. Two vacation models for token-ring networks where service is controlled by timers. *Perf. Eval.* **20**, 165-184.

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

[29] Takagi, H., Leung, K.K. [1994]. Analysis of a discrete-time queueing system with time-limited service. *Queueing Systems* **18**, 183-197.

[30] Tedijanto, T.E. [1992]. A note on the comparison between Bernoulli and limited policies in vacation models. *Perf. Eval.* **15**, 89-97.

[31] Tedijanto, T.E. [1990]. *Nonexhaustive policies in polling systems and vacation models.* PhD Thesis, University of Maryland, College Park, MD.

[32] Watson, K. S. [1985]. Performance evaluation of cyclic service strategies - a survey. In: *Performance '84*, ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 521-533.

[33] Yechiali, U. [1991]. Optimal dynamic control of polling systems. In: *Queueing, Performance and Control in ATM*, eds. J.W. Cohen and C.D. Pack, North-Holland Publ. Cy., Amsterdam, 205-217.