TI 2005-066/4
Tinbergen Institute Discussion Paper

# On the Optimal Policy for Deterministic and Exponential Polling Systems

*Bruno Gaujal*[1]
*Arie Hordijk*[2]
*Dinard van der Laan*[3]

[1] *INRIA Rhône-Alpes, Montbonnot Saint Martin, France,*
[2] *Dept of Mathematics, Leiden University, Leiden,*
[3] *Dept of Econometrics & Operations Research, Vrije Universiteit Amsterdam, and Tinbergen Institute.*

**Tinbergen Institute**
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

**Tinbergen Institute Amsterdam**
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

**Tinbergen Institute Rotterdam**
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

# On the optimal policy for deterministic and exponential polling systems*

Bruno Gaujal † Arie Hordijk ‡ Dinard van der Laan §

April 26, 2005

**Abstract**

In this paper, we consider deterministic (both fluid and discrete) polling systems with $N$ queues with infinite buffers and we show how to compute the best polling sequence (minimizing the average total workload). With two queues, the best polling sequence is always periodic when the system is stable and forms a regular sequence. The fraction of time spent by the server in the first queue is highly non continuous in the parameters of the system (arrival rate and service rate) and shows a fractal behavior. Convexity properties are shown in Appendix as well as a generalization of the computations to the stochastic exponential case.
**Keywords** Polling systems, regular sequences, multimodularity, optimal control.

## 1 Introduction

In this paper we consider a polling system with $N$ queues with infinite buffers. Time is slotted and the time-slot sizes are either deterministic with unit size or exponential distributed with mean one. In each queue one customer arrives in each time-slot, we assume that its required service time is queue dependent, it is a fixed amount in the deterministic polling model and it is exponentially distributed in the exponential model. For the deterministic model we also consider the model with fluid input. There is one server, and at the beginning of each time-slot the decision has to be made which queue is being served during the next time-slot, we assume zero switching times as is usual in the performance analysis of communication networks. The service-discipline is first-in-first-out for each queue and it is work-conservative. The control of the server is an open-loop control, which means that the decision, which queue to be served in the next time-slot is independent of the actual workloads in the nodes. The open-loop polling sequence is an infinite sequence where its $n-th$ element gives the queue which is served during the $n-th$ time-slot. We derive an efficient algorithm for computing the optimal open-loop polling sequence with objective the sum of the average workloads in the queues, for the deterministic and for the exponential polling model with $N = 2$ queues. We exploit the theory on multimodularity of the average workload and the optimality of regular sequences as it has been derived in recent papers (see [1] and [2]). It follows from this theory that for $N = 2$ queues the optimal polling policy assigns time-slots to the first queue (service sequence of the first queue as we will call it) as a

*This work was partially supported by the Van Gogh grant on discrete event systems

†Lab. ID-IMAG, (INRIA, CNRS, INPG, UJF) 51 avenue Jean Kuntzmann, 38330 Montbonnot Saint Martin, France. E-mail: bruno.gaujal@imag.fr Tel: (+33)476612058

‡Department of Mathematics, Leiden University, Niels Bohrweg 1, P.O. Box 9512, 2300 RA Leiden, The Netherlands. E-mail: hordijk@math.leidenuniv.nl Tel: (+31)715277146

§Faculty of Economics and Business Administration, Department of Econometrics, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: dalaan@feweb.vu.nl Tel: (+31)204446022

regular sequence, say with density $\hat{d}_1$ ( and also the service sequence for the second queue is also regular with density $\hat{d}_2 = 1 - \hat{d}_1$). As we show in Appendix A, the sum (in general of any linear combination) of the average workloads in both queues say $B(d_1, d_2)$ is a convex function of the densities $(d_1, d_2)$ if the service sequences are regular for both queues. Our algorithms compute this convex function for the deterministic and for the exponential model, and it finds the optimal densities $(\hat{d}_1, \hat{d}_2)$ for both models.

In the sections 2 to 4 we analyze the two deterministic models with discrete and fluid input. For fixed densities $(d_1, d_2)$ the average workloads in both queues become independent of each other and the average workloads can be studied separately. In Section 3 the average workload for one queue with regular service sequence as function of its density is analyzed, it is shown that it is a convex piecewise-linear function. Using results from number theory we derive explicit formulae for the average workload in case the service density is a best approximation point of the input rate. With the use of the continued fraction expansion of the input rate we then derive an efficient algorithm for calculating the average workload for any density. Doing these calculations for both queues gives an efficient algorithm for calculating $\min_{d_1+d_2=1} B(d_1, d_2)$, which provides the exact optimal polling policy. In section 4 we illustrate the algorithm with numerical experiments. Also in section 4 we derive results on the structure of the optimal policy. We prove that for deterministic polling systems with $N \geq 2$ queues and rational input rates there is always an periodic optimal policy.

In Appendix B, the algorithm for calculating the optimal polling sequence for the exponential system is derived. The sum of the workloads in both queues is again a convex function in the densities $(d_1, d_2)$. Using the Kernel method the exact optimal policy is computed in an efficient way.

There is an extensive literature on the many variants of polling systems (see [6, 21]). In several papers [4, 17, 5] an algorithm is derived for calculating efficient visit orders to the queues, also called polling tables.. In [12] the exponential polling model is converted to a Markov decision chain with no state information and an algorithm to compute a nearly optimal polling policy is given. In [7] a heavy-traffic averaging principle is derived. To the knowledge of the authors no algorithm for computing the exact optimal polling sequence is available in the literature. Also it seems that our structural results are new. Since a polling can be seen as a server routing model there is a duality with the customer routing model. The authors have studied the latter model recently (for the deterministic model see [2] and [13], and for the exponential model see [10]).

## 2 Description of deterministic polling systems

We consider a polling model in which queues are served by one server, which serves at a constant rate of 1. So, if a queue is served by this server (and not considering any new input of workload in the queue) then the workload in the queue decreases by 1 per time-unit until the queue is empty or the server stops serving the queue. We assume that the input in the queues is deterministic, but we consider two slightly different models. In the first model the input in queue $i$, $i = 1, 2, \ldots, N$, is discrete and in fact we assume for $i = 1, 2, \ldots, N$ that at all the integer times $T = 0, 1, 2, \ldots$ a job with constant workload $\lambda_i$ arrives in queue $i$. In the second model we assume that the workload input in queue $i$, $i = 1, 2, \ldots, N$ is fluid which flows in with constant rate $\lambda_i$.

In both models we assume that at all the integer times $T = 0, 1, 2, \ldots$ a queue is chosen to be served by the server for the next time-unit. So, in case of $N$ queues numbered $1, 2, \ldots, N$ the polling policy can be described by the infinite sequence $U = (U_1, U_2, \ldots)$ where $U_n$ is the queue to be served by the server during the time interval $[n-1, n]$. For both models we describe the system with $N$ queues by the $N$ -dimensional vector $\overline{\lambda} := (\lambda_1, \lambda_2, \ldots, \lambda_N)$. An infinite sequence $U$ corresponding to an polling policy for such a system is a so-called word on the alphabet $\{1, 2, \ldots, N\}$ (see [13] and [18]). The set

of all words on the alphabet $\{1, 2, \ldots, N\}$ (corresponding to all the possible polling policies for some $\overline{\lambda}$ system) is denoted by $\mathcal{A}(N)$.

For $i = 1, 2, \ldots, N$ and $t \in \mathbb{R}_{\geq 0}$ let $V_i(t)$ be the (remaining) workload in queue $i$ at time $t$. Then the long-run average workload in queue $i$ is given by the Cesaro integral

$$B_i := \limsup_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} V_i(t) dt.$$

We define for $U \in \mathcal{A}(N)$ for $i = 1, 2, \ldots, N$ an infinite sequence $u^i = (u_1^i, u_2^i, \ldots)$ of zeros and ones by $u_n^i = 1$ if $U_n = i$ and $u_i^n = 0$ if $U_n \neq i$. We call such a sequence of zeros and ones corresponding to the server-assignment for one queue for short service sequence. This in contrast to Chapter 9 of [2] and [1], where such a sequence is called a vacation sequence. Note that given the fact that we consider a model with fluid or discrete input the value of $B_i$ for a given queue $i$ depends only on the arrival rate $\lambda_i$ in the queue and the service sequence $u^i$ for this queue.

For both the discrete model and the fluid model the objective of the polling is to minimize the total long-run average workload which is given by

$$B = B(U) := \sum_{i=1}^{N} B_i,$$

where $U$ is the polling policy. So, a polling policy $U'$ is called optimal for a given $\overline{\lambda}$ system (either for the discrete or the fluid model) if $B(U') = \min_{U \in \mathcal{A}(N)} B(U)$. The minimal total long-run average workload for some $\overline{\lambda}$ system is given by

$$\widetilde{B} = \widetilde{B}(\overline{\lambda}) := \inf_{U \in \mathcal{A}(N)} B(U).$$

## 3 On the average workload in a single queue

In this section we consider a single queue of the polling system with given input rate $\lambda > 0$. In the sequel the input can be both discrete and fluid unless it is specified which model we consider. Since we consider in this section a single queue of the system we omit in notation all the sub-indices $i$ referring to the queue. Let $u = (u_1, u_2, \ldots)$ be the infinite service sequence of zeros and ones for this queue. Then for $n = 0, 1, 2, \ldots$ we denote by $\kappa_u(n) := \sum_{i=1}^{n} u_i$ the partial sum of the first $n$ terms of $u$. So, $\kappa_u(n)$ is the number from the first $n$ time intervals of unit length that the server is serving this particular queue.

We always assume that the queue is empty at time $t = 0$ which means that $V(0) = 0$. Moreover, for the model with discrete arrivals of workload $\lambda > 0$ at $t = 0, 1, 2, \ldots$ we make the convention that $V(t) = \lim_{t' \uparrow t} V(t')$ for $t > 0$. Hence for $t = 0, 1, 2, \ldots$ we have $\lim_{t' \downarrow t} V(t') = V(t) + \lambda$ in this model. For the model with fluid input it is easily seen that the function $V(t)$ is continuous for $t \geq 0$. For both models let $G(t) = \{t' \in [0, t] : V(t) = 0\}$ and let $m(t)$ be the Lebesgue measure of $G(t)$. Then for a given vacancy sequence $u = (u_1, u_2, \ldots)$ we have the following formulas for $V(t)$ for every $t \geq 0$.

**Lemma 3.1** *For the model with discrete input we have*

$$V(t) = \lambda \lceil t \rceil - \kappa_u(\lfloor t \rfloor) - u_{\lfloor t \rfloor + 1} \cdot (t - \lfloor t \rfloor) + m(t) \text{ for every } t \geq 0 \tag{1}$$

*and for the model with fluid input we have*

$$V(t) = \lambda t - \kappa_u(\lfloor t \rfloor) - u_{\lfloor t \rfloor + 1} \cdot (t - \lfloor t \rfloor) + m(t) \cdot (1 - \lambda) \text{ for every } t \geq 0. \tag{2}$$

We denote by $\overline{d} := \limsup_{t\to\infty} \frac{\kappa_u(t)}{t}$ the upper density of $u$ and by $\underline{d} := \liminf_{t\to\infty} \frac{\kappa_u(t)}{t}$ the lower density of $u$. If $\overline{d} = \underline{d}$ then we say that $u$ has a density $d = \lim_{t\to\infty} \frac{\kappa_u(t)}{t}$.

An infinite sequence of zeros and ones $u = (u_1, u_2, \ldots)$ is called regular with density $d \in [0,1]$ (see for example [2](where it is called balanced), [3], [23] and [22]) if for every $n \in \mathbb{N}$ we have for every subsequence $v = (u_k, u_{k+1}, \ldots, u_{k+n-1})$ of $u$ of length $n$ that the number of ones in $v$ is equal to $\lfloor nd \rfloor$ or $\lceil nd \rceil$. For $d \in [0,1]$ let $\mathcal{S}(d)$ be the set of all infinite regular sequences of zeros and ones with density $d$. Moreover, let $\omega(d) = (\omega_1, \omega_2, \ldots)$ be the sequence defined by $\kappa_\omega(n) = \lceil nd \rceil$ for all $n \in \mathbb{N}$ and let $\pi(d) = (\pi_1, \pi_2, \ldots)$ be the sequence defined by $\kappa_{\pi(d)}(n) = \lfloor nd \rfloor$ for all $n \in \mathbb{N}$. It is easily seen that the following lemma holds which gives a characterization of $\omega(d)$ and $\pi(d)$ in the set of regular sequences of density $d$.

**Lemma 3.2** *For every $d \in [0,1]$ we have that $\omega(d) \in \mathcal{S}(d)$ and $\pi(d) \in \mathcal{S}(d)$. Moreover, for every $u \in \mathcal{S}(d)$ we have that*

$$\kappa_{\pi(d)}(n) \leq \kappa_u(n) \leq \kappa_\omega(n) \text{ for } n = 0, 1, 2, \ldots.$$

Therefore $\omega(d)$ is called the upper bracket sequence sequence of density $d$ and $\pi(d)$ is called the lower bracket sequence of density $d$.

By results in [2] (see Appendix A) it follows that assigning the server to the queue according to a regular service sequence of density $d$ is optimal (thus minimizes the long-run average workload in the queue) among all polling sequences of upper density at most $d$. Note that if $u, v \in \mathcal{S}(d)$ are two regular sequences of the same density $d$ that then $u$ and $v$ have the same performance. So, if we denote with slight abuse of notation the long-run average workload in the queue for any service sequence $u$ with $B(u) = B_\lambda(u)$ and we define for $d \in [0,1]$ the long-run average workload in the queue for regular service sequences with density $d$ as $B(d) := B(\pi(d))$ then we can summarize this with the following lemma.

**Lemma 3.3** *For every input rate $\lambda$ and any service sequence $u$ of upper density at most $d$ we have that $B(u) \geq B(d) = B(\pi(d))$.*

In the remaining of this section we obtain properties of $B(d)$ as function of $d$ and we give an algorithm for calculating the value of $B(d)$ for any given input rate $\lambda$ and density $d$.

An infinite sequence $u = (u_1, u_2, \ldots)$ is periodic with period $T$ if $u_n = u_{n+T}$ for $n = 1, 2, \ldots$ and $T$ is the minimal positive integer with this property. If $u$ is periodic with period $T$ then the finite sequence $(u_1, u_2, \ldots, u_T)$ is called the period word of $u$. It is easily seen that if $u = (u_1, u_2, \ldots)$ is a regular sequence of zeros and ones with a rational density $d = \frac{p}{q}$ with $p, q \in \mathbb{N}$, $\gcd(p,q) = 1$ that then $u$ is periodic with periodic $q$. For such rational density $d = \frac{p}{q}$ the period word $(\omega_1, \omega_2, \ldots, \omega_q)$ of the upper bracket sequence is denoted by $w(d)$ and the period word $(\pi_1, \pi_2, \ldots, \pi_q)$ of the lower bracket sequence is denoted by $p(d)$.

Consider a service sequence for $k$ consecutively time-intervals of unit length corresponding to a finite sequence $u$ of length $k$. Suppose that the first of these $k$ intervals starts at time $t_0 \in \mathbb{Z}_{\geq 0}$ and thus the last at time $t_0 + k - 1$. Then we say that $u$ lasts from $t_0$ to $t_0 + k$ and we say that $u$ is workload non-increasing if for every initial workload $V(t_0)$ it holds that $V(t_0 + k) \leq V(t_0)$. The following lemma will be useful for proving properties of $B(d)$.

**Lemma 3.4** *Let $d \in \mathbb{Q}$, $0 \leq d \leq 1$. Then $p(d)$, the period word of the lower bracket sequence of density $d$, is workload non-increasing if and only if $d \geq \lambda$.*

**Proof**. We consider the model with discrete input and let $d = \frac{p}{q}$ with $\gcd(p, q) = 1$. Suppose that $p(d)$ lasts from $t_0 \in \mathbb{Z}$ to $t_0 + q \in \mathbb{Z}$. By (1) we have that

$$V(t_0+q)-V(t_0) = \lambda(t_0+q)+m(t_0+q)-(\lambda\cdot t_0+m(t_0))-\kappa_q(p(d)) = \lambda\cdot q-p+(m(t_0+q)-m(t_0)).$$

It is obvious that $m(t_0 + q) - m(t_0) \geq 0$. So, if $d < \lambda$ then $V(t_0 + q) - V(t_0) \geq \lambda \cdot q - p > d \cdot q - p = 0$ and thus $p(d)$ is not workload non-increasing.

It is obvious that the value of $m(t_0 + k) - m(t_0)$ is monotonically decreasing with the value of $V(t_0)$. Hence, by Lemma 3.1 it follows that some finite factor $u = (u_1, u_2, \ldots, u_k)$ is workload non-increasing if for this factor $V(t_0) = 0$ implies that $V(t_0 + k) = 0$. So, suppose that $d \geq \lambda$ and $V(t_0) = 0$. Put $t' = \max_{t\in[t_0,t_0+q]:V(t)=0}$ and $t^* = t' - t_0$. Since the workload only increases at integer times it follows that $t'$ is an integer number and by definition we have that $m(t') = m(t_0 + q)$. Moreover, $t^* := t' - t_0 \in [0, 1, \ldots, q]$. Hence by (1) we have that ,

$$V(t_0 + q) = V(t_0 + q) - V(t') = \lambda \cdot (q - t^*) - (\kappa_q(p(d)) - \kappa_{t^*}(p(d)) =$$

$$\lambda \cdot (q - t^*) - (p - \lfloor t^* \cdot d \rfloor) \leq d \cdot (q - t^*) - p + t^* \cdot d = 0.$$

So, $V(t_0 + q) = 0$ and thus $p(d)$ is workload non-increasing if $d \geq \lambda$. For the model with fluid this can be proved analogously. □

Suppose that $d \in \mathbb{Q}$, $d \geq \lambda$ and that the server is serving the queue according to the periodic lower bracket sequence $\pi(d) = p(d)^\infty$. Then it follows from Lemma 3.4 that if the workload is 0 at the beginning of an $p(d)$ factor then it is also 0 at the end of the factor and thus at the beginning of the next factor. So, since the workload is 0 at $t = 0$, the beginning of the first $p(d)$ factor, it follows that the workload is 0 at the end of every $p(d)$ factor. Thus the workload process is renewed after every $p(d)$ factor. Hence we have the following corollary of Lemma 3.4.

**Corollary 3.5** *If $d \in \mathbb{Q}$, $d \geq \lambda$ then $B(d)$ is equal to the average workload in the queue during any period $p(d)$.*

On the other hand if $d < \lambda$ then it is easily seen that the workload goes to infinity if the queue is served according to the lower bracket sequence $\pi(d)$ and thus $B(d) = \infty$ in that case. We will call $[\lambda, 1]$ the interval of stability since the workload remains bounded and thus $B(d)$ is finite if $d \in [\lambda, 1]$. We will examine some properties of the function $B(d)$ on the interval of stability. By Lemma 3.4 and Corollary 3.5 the following property follows analogously to Theorem 5.8 in [13] (see also [14]).

**Theorem 3.6** *For given input rate $0 \leq \lambda \leq 1$ we have that the function $B(d)$ is convex on the interval of stability $[\lambda, 1]$.*

### 3.1 Farey intervals

We use in this subsection several notions defined in [13] and we summarize results which can be obtained analogously to results in [13]. We also recall that if $d_1, d_2$ are rational numbers with $0 \leq d_1 \leq d_2 \leq 1$, $d_i = \frac{p_i}{q_i}$ and $\gcd(p_i, q_i) = 1$ for $i = 1, 2$, that then $I = [d_1, d_2]$ is called a Farey interval if and only if $q_1 \cdot p_2 - p_1 \cdot q_2 = 1$. Put $d_0 = \frac{p_1+p_2}{q_1+q_2}$. If $I = [d_1, d_2]$ is a Farey interval then $I' = [d_1, d_0]$ and $I'' = [d_0, d_2]$ are also Farey intervals and all rational numbers in $(d_1, d_2)$ have denominator greater than or equal to $q_1 + q_2$.

The following result for the factorization of period words of lower bracket sequences follows analogously to lemma 4.3 in [13].

**Lemma 3.7** *Let $I = [d_1, d_2]$ be a Farey interval and $d_0 = \frac{p_1+p_2}{q_1+q_2}$ as above. Then $p(d_0) = p(d_1)p(d_2)$.*

By Lemma 3.7 we have analogously to theorem 4.4 in [13] the following theorem.

**Theorem 3.8** *Let $I = [d_1, d_2]$ be a Farey interval and put $X := \{p(d_1), p(d_2)\}$. Then for every $d \in (d_1, d_2)$ there exists a unique $X$-factorization of the lower bracket sequence $\pi(d)$ of density $d$. Moreover, if $d$ is rational then there exists a unique finite $X$-factorization of the period word $p(d)$ of $w(d)$.*

According to the following theorem the function $B(d)$ is besides being convex also piecewise linear. More precisely, the theorem says that the function $B(d)$ is linear on Farey intervals contained in the stability interval $[\lambda, 1]$. This property will be very useful for computing the value of $B(d)$ for any $d$ and input rate $\lambda$.

**Theorem 3.9** *Let $d_1, d_2 \in [0, 1]$ be rational numbers such that $I = [d_1, d_2]$ is a Farey interval and $\lambda \leq d_1 < d_2$, where $\lambda$ is the input rate. Let $d \in I$, $d = \mu \cdot d_1 + (1 - \mu) \cdot d_2$, where $\mu \in [0, 1]$. Then $B(d) = \mu \cdot B(d_1) + (1 - \mu) \cdot B(d_2)$.*

Theorem 3.9 can be proved analogously to theorem 5.9 in [13] by combining Lemma 3.4, Corollary 3.5 and Theorem 3.8. Essential in the proof is that if the (unique) factorization of $\pi(d)$ in $p(d_1)$ and $p(d_2)$ factors is considered, then $\mu$ is the fraction (of time) taken by $p(d_1)$ factors and $1-\mu$ the fraction taken by $p(d_2)$ factors. Moreover, the workload is zero after every such a factor.

## 3.2 Best upper approximations

**Definition 3.10** *Let $0 < x \leq 1$ be a given real number and let $s = \frac{p}{q}$ with $p \in \mathbb{N}$, $q \in \mathbb{N}$ with $\gcd(p, q) = 1$ be a rational number such that $s \geq x$. Then $s$ is called a best upper approximation of $x$ if there does not exist a rational number in the interval $[x, s)$ with denominator smaller than or equal to $q$.*

Note that $x$ is a best lower approximation of $x$ itself if and only if $x$ is a rational number.

**Lemma 3.11** *Let $0 < \lambda < 1$ be the input rate and $d = \frac{p}{q}$ be a best upper approximation of $\lambda$, where $p, q \in \mathbb{N}$ with $\gcd(p, q) = 1$ . Suppose that the server is serving according to the lower bracket sequence $\pi(d)$ with period $\mid p(d) \mid = q$. Let $J(d) := \max\{t \in [0, q] : m(t) = 0\}$. For discrete input we have that*

$$V(q-1) = \lambda(q-1) - (p-1) \text{ and } J(d) - (q-1) = \lambda q - (p-1) > 0.$$

*For fluid input we have that*

$$V(q-1) = \lambda(q-1) - (p-1) \text{ and } J(d) - (q-1) = \frac{\lambda(q-1) - (p-1)}{1-\lambda} \geq 0.$$

**Proof.** By Lemma 3.1 we have for $t = 0, 1, 2, \ldots$ that $V(t) \geq \lambda t - \kappa_{p(d)}(t) = \lambda t - \lfloor dt \rfloor$. Suppose $V(t) = 0$ for some $t \in [1, 2, \ldots, q-1]$. Then $\lambda t - \lfloor dt \rfloor \geq 0$ and thus $\lambda \leq \frac{\lfloor dt \rfloor}{t} \leq d$. Since the rational number $\frac{\lfloor dt \rfloor}{t}$ has denominator $t < q$, this contradicts the fact that $d = \frac{p}{q}$ is a best upper approximation of $\lambda$. Hence $V(t) > 0$ for $t = 1, 2, \ldots, q-1$ and from this it is easily seen that $V(t) > 0$ for every $t \in (0, q-1]$. Hence $m(q-1) = 0$ and thus we have by Lemma 3.1 that

$$V(q-1) = \lambda(q-1) - \kappa_{p(d)}(q-1) = \lambda(q-1) - \lfloor \frac{p}{q}(q-1) \rfloor = \lambda(q-1) - (p-1).$$

So, for the model with discrete input we have that $J(d) - (q-1) = V(q-1) + \lambda = \lambda q - (p-1) > 0$ and for the model with fluid input we have that $J(d) - (q-1) = \frac{V(q-1)}{1-\lambda} = \frac{\lambda(q-1)-(p-1)}{1-\lambda} \geq 0$. □

**Theorem 3.12** *Let $0 < \lambda < 1$ be the input rate and $d = \frac{p}{q}$ be a best upper approximation of $\lambda$, where $p, q \in \mathbb{N}$ with $\gcd(p, q) = 1$ Then for discrete input we have*

$$B_\lambda(d) = \frac{\lambda q^2 + \lambda q - pq + q - 1 + \lambda^2 q^2 - 2\lambda pq + p^2}{2q}$$

*and for fluid input we have*

$$B_\lambda(d) = \frac{\lambda q^2 - \lambda q + \lambda - pq - \lambda pq + q + p^2 - 1}{2q(1-\lambda)}.$$

**Proof.** According to Corollary 3.5 we have that $B_\lambda(d) = \frac{1}{q}\int_{t=0}^{q} V(t)dt$. For the model with discrete input we have by (1) and Lemma 3.11 that

$$V(t) = \lambda\lceil t\rceil - \lfloor \lfloor t\rfloor d\rfloor - p(d)_{\lfloor t\rfloor+1}\cdot(t - \lfloor t\rfloor)$$

for every $t \in [0, J(d)]$ and in particular for every $t \in [0, q-1]$. Moreover, it is easily seen that $V(t) = V(q-1) + \lambda - (t-(q-1)) = \lambda\cdot q + q - p - t$ for $t \in (q-1, J(d)]$ and $V(t) = 0$ for every $t \in [J(d), q]$. Hence, by putting $A := \int_{t=0}^{q-1}(\lambda\lceil t\rceil)dt$, $B := \int_{t=0}^{q-1}(\lfloor\lfloor t\rfloor d\rfloor)dt$, $C := \int_{t=0}^{q-1}(p(d)_{\lfloor t\rfloor+1}(t-\lfloor t\rfloor))dt$ and $D := \int_{t=q-1}^{J(d)}(\lambda\cdot q + q - p - t)dt$ we have

$$B_\lambda(d) = \frac{1}{q}(A - B - C + D).$$

We have $A = \sum_{n=1}^{q-1}\int_{t=n-1}^{n}(\lambda\lceil t\rceil)dt = \lambda\cdot\sum_{n=1}^{q-1} n = \frac{\lambda}{2}(q-1)q$ and by theorem 100 in [11]

$$B = \sum_{n=0}^{q-2}\int_{t=n}^{n+1}(\lfloor\lfloor t\rfloor d\rfloor)dt = \sum_{n=0}^{q-2}\lfloor nd\rfloor = \sum_{n=0}^{q-1}\lfloor n\frac{p}{q}\rfloor - \lfloor(q-1)\frac{p}{q}\rfloor = \frac{1}{2}(p-1)(q-1) - (p-1) = \frac{1}{2}(p-1)(q-3).$$

Moreover, since $p(d)$ has an 1 as last component,

$$C = \sum_{n=1}^{q-1}\int_{t=n-1}^{n} p(d)_{\lfloor t\rfloor+1}(t-\lfloor t\rfloor))dt = \kappa_{p(d)}(q-1)\int_{t=0}^{1} tdt = \frac{1}{2}(p-1)$$

and

$$D = \int_{t=q-1}^{q-1+\lambda q-(p-1)}(\lambda\cdot q + q - p - t)dt = \int_{t=0}^{\lambda q-(p-1)}(\lambda\cdot q - (p-1) - t)dt = \frac{1}{2}(\lambda q - (p-1))^2.$$

Hence,

$$B_\lambda(d) = \frac{1}{q}(A - B - C + D) = \frac{\lambda q^2 + \lambda q - pq + q - 1 + \lambda^2 q^2 - 2\lambda pq + p^2}{2q}.$$

Analogously we have for the model with fluid input that

$$B_\lambda(d) = \frac{1}{q}(A^* - B - C + D^*),$$

where $A^* = \int_{t=0}^{q-1}\lambda t dt = \frac{1}{2}\lambda(q-1)^2$ and

$$D^* = \int_{t=q-1}^{J(d)}(\lambda\cdot(q-1) + q - p - (1-\lambda)t)dt = \int_{t=q-1}^{q-1+\frac{\lambda(q-1)-(p-1)}{1-\lambda}}(\lambda\cdot(q-1) + q - p - (1-\lambda)t)dt =$$

$$\int_{t=0}^{\frac{\lambda(q-1)-(p-1)}{1-\lambda}} (\lambda(q-1)-(p-1)-(1-\lambda)t)dt = \frac{(\lambda(q-1)-(p-1))^2}{2(1-\lambda)}.$$

Hence,

$$B_\lambda(d) = \frac{1}{q}(A^* - B - C + D^*) = \frac{\lambda q^2 - \lambda q + \lambda - pq - \lambda pq + q + p^2 - 1}{2(1-\lambda)}.$$

□

Besides this closed formula of the value of $B(d)$ for all the best upper approximations we give a separate formula for $d = \lambda$, the initial point of the interval of stability

**Lemma 3.13** *Let $0 < \lambda \leq 1$ be the input rate. For the model with discrete input we have that $B_\lambda(\lambda) = \frac{\lambda+1}{2}$ if $\lambda$ is irrational, and $B_\lambda(\lambda) = \frac{p+q-1}{2q}$ if $\lambda = \frac{p}{q}$ with $p, q \in \mathbb{N}$, $\gcd(p,q) = 1$. For the model with fluid input we have that $B_\lambda(\lambda) = \frac{1}{2}$ in case $\lambda$ is irrational, and $B_\lambda(\lambda) = \frac{1}{2} - \frac{1}{2q}$ if $\lambda = \frac{p}{q}$ with $p, q \in \mathbb{N}$, $\gcd(p,q) = 1$.*

**Proof.** If $\lambda$ is rational then $\lambda = \frac{p}{q}$ with $p, q \in \mathbb{N}$, $\gcd(p,q) = 1$ is a best upper approximation of $\lambda$. Therefore the formulae for $B_\lambda(\lambda)$ follow directly from Theorem 3.12 by substituting $\frac{p}{q}$ for $\lambda$.

In the sequel of this proof we suppose that $\lambda$ is irrational and we first consider the model with discrete input. For $t = 1, 2, \ldots$ we have that

$$\lambda\lceil t\rceil - \kappa_{\pi(\lambda)}(\lfloor t\rfloor) - \pi(\lambda)_{\lfloor t\rfloor+1}(t - \lfloor t\rfloor) = \lambda t - \lfloor \lambda t\rfloor > 0,$$

since $\lambda$ is irrational and thus $\lambda t$ is not an integer for $t = 1, 2, \ldots$. Hence for every $t > 0$ we have that $m(t) = 0$ and $V(t) = \lambda\lceil t\rceil - \lfloor \lambda\lfloor t\rfloor\rfloor - \pi(\lambda)_{\lfloor t\rfloor+1}(t - \lfloor t\rfloor) > 0$. So,

$$B_\lambda(\lambda) = \limsup_{T\to\infty} \frac{1}{T}\int_{t=0}^{T} V(t)dt = A1 - A2,$$

where $A1 = \limsup_{T\to\infty} \frac{1}{T}\int_{t=0}^{T}(\lambda\lceil t\rceil - \lfloor\lambda\lfloor t\rfloor\rfloor)dt$ and $A2 = \lim_{T\to\infty} \frac{1}{T}\int_{t=0}^{T}(\pi(\lambda)_{\lfloor t\rfloor+1}(t - \lfloor t\rfloor))dt$. By the ergodic theorem of Weyl and von Neumann we have

$$A1 = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}(\lambda(t+1) - \lfloor\lambda t\rfloor) = \lambda + \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}(\lambda t - \lfloor\lambda t\rfloor) = \lambda + \int_{x=0}^{1} 1dx = \lambda + \frac{1}{2}.$$

Moreover,

$$A2 = \lim_{T\to\infty} \frac{\kappa_{\pi(\lambda)}(T)}{T} \cdot \int_{t=0}^{1}(t - \lfloor t\rfloor))dt = \lambda\int_{t=0}^{1} tdt = \frac{\lambda}{2}.$$

Hence $B_\lambda(\lambda) = \lambda + \frac{1}{2} - \frac{\lambda}{2} = \frac{\lambda+1}{2}$ if $\lambda$ is irrational. For the model with fluid input and irrational $\lambda$ we have analogously to the model with discrete input that $V(t) = \lambda t - \lfloor\lambda\lfloor t\rfloor\rfloor - \pi(\lambda)_{\lfloor t\rfloor+1}(t - \lfloor t\rfloor) > 0$ for every $t > 0$ and thus $B_\lambda(\lambda) = A3 - A2 = A3 - \frac{\lambda}{2}$ where

$$A3 = \limsup_{T\to\infty} \frac{1}{T}\int_{t=0}^{T}(\lambda t - \lfloor\lambda\lfloor t\rfloor\rfloor)dt.$$

We have

$$A3 = A1 - \lim_{T\to\infty} \frac{1}{T}\int_{t=0}^{T}(\lambda\lceil t\rceil - \lambda t)dt = (\lambda + \frac{1}{2}) - \lambda\int_{t=0}^{1}(1-t)dt = \frac{\lambda+1}{2},$$

whence $B_\lambda(\lambda) = \frac{1}{2}$. □

For given input rate $0 < \lambda \leq 1$ we have that $s_1 = 1$ is the first best upper approximation of $\lambda$. Consider the following iterative construction of best upper approximations of $\lambda$. Put $i := 1$ and do the following iteratively. If $s_i = \lambda$ then stop. Else let $s_{i+1}$ be the rational number of lowest denominator in the interval $[\lambda, s_i)$ and let $i := i + 1$. Since every subinterval of positive Lebesgue measure of the open interval $(0, 1)$ contains a unique rational number of lowest denominator the $s_i$ are well defined. Moreover, analogously to the properties for best lower approximations in [13] we have the following properties for the best upper approximations.

**Lemma 3.14** *Some number $d$ is a best upper approximation of $0 < \lambda \leq 1$ if and only if $d = s_i$ for some $i \in \mathbb{N}$. Moreover, if $\lambda$ is rational then there exists some $k \in \mathbb{N}$ such that*

$$1 = s_1 > s_2 > \ldots > s_k = \lambda.$$

*If $\lambda$ is irrational then*

$$1 = s_1 > s_2 > \ldots \text{ and } \lim_{i\to\infty} s_i = \lambda.$$

*If $s_i, s_{i+1}$ are consecutive best upper approximations of $\lambda$ then $[s_{i+1}, s_i]$ is a Farey interval.*

Note that by Theorem 3.9 and Lemma 3.14 it follows that the function $B(d) = B_\lambda(d)$ is linear on any interval $[s_{i+1}, s_i]$ where $s_i, s_{i+1}$ are consecutive best upper approximations of $\lambda$. In fact it turns out that the slope of the function $B_\lambda(d)$ changes precisely at all the best upper approximations of $\lambda$. This implies that the exact value of $B(d) = B_\lambda(d)$ is easily computed if we can find the consecutive best upper approximations $s_i, s_{i+1}$ of $\lambda$ such that $d \in [s_{i+1}, s_i]$. We give an example to illustrate this.

**Example.** We calculate $B_\lambda(d)$ for $\lambda = \frac{12}{17}$ and $d = \frac{\sqrt{2}}{2}$. It is easily seen that the best upper approximations of $\lambda = \frac{12}{17}$ are consecutively $s_1 = \frac{1}{1}$, $s_2 = \frac{3}{4}$ , $s_3 = \frac{5}{7}$ , $s_4 = \frac{12}{17}$ and we have that $d \in [s_4, s_3] = [\frac{12}{17}, \frac{5}{7}]$, which is a Farey interval. We consider the model with discrete input (the computation for the model with fluid input is similar). Then Lemma 3.13 gives $B(\frac{12}{17}) = \frac{12+17-1}{2\cdot 17} = \frac{14}{17}$ and by Theorem 3.12 we have that $B(\frac{5}{7}) = \frac{\frac{12}{17}7^2+\frac{12}{17}7-5\cdot 7+7-1+\frac{12}{17}^2 7^2-2\frac{12}{17}5\cdot 7+5^2}{2\cdot 7} = \frac{1522}{2023}$. Putting $\mu = \frac{\frac{5}{7}-d}{\frac{5}{7}-\frac{12}{17}} = 85 - \frac{119}{2}\sqrt{2}$ we have $\mu \in [0, 1]$ and $d = \mu\frac{12}{17} + (1 - \mu)\frac{5}{7}$. Hence by Theorem 3.9 we have

$$B_{\frac{12}{17}}(\frac{\sqrt{2}}{2}) = \mu B(\frac{12}{17}) + (1-\mu)B(\frac{5}{7}) = (85 - \frac{119}{2}\sqrt{2})\frac{14}{17} + (\frac{119}{2}\sqrt{2} - 84)\frac{1522}{2023} = \frac{1966}{289} - \frac{72}{17}\sqrt{2}.$$

### 3.3 An efficient algorithm for calculating $B_\lambda(d)$

The only remaining problem for the efficiency of the algorithm to calculate $B_\lambda(d)$ for any given $0 < \lambda < 1$ and $d \in [\lambda, 1]$ is to find in general (the) two consecutive best upper approximations $s_i$ and $s_{i+1}$ of $\lambda$ such that $d \in [s_{i+1}, s_i]$ in an efficient way. Following the arguments in [13] (where the same problem appears with the only difference that best lower approximations have to found in stead of best upper approximations), it can be shown that this problem can be efficiently solved by using the continued fraction expansion of $\lambda$.

The following facts about the continued fraction expansion and the convergents of some real number $\alpha > 0$ are well known (see for example [11] and [20]).

The partial quotients $a_0, a_1, \ldots$ of the (simple) continued fraction expansion of $\alpha > 0$ are recursively defined by:

$$\left\{ \begin{array}{ll} a_0 = \lfloor \alpha \rfloor; & \alpha_1 = \frac{1}{\alpha - a_0} \\ a_n = \lfloor \alpha_n \rfloor; & \alpha_{n+1} = \frac{1}{\alpha_n - a_n} \text{ for } n = 1, 2, \ldots \end{array} \right\}.$$

Then

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ldots \cfrac{1}{a_n + \ldots}}} := [a_0, a_1, \ldots, a_n, \ldots].$$

Note that applying the continued fraction algorithm for some input rate $0 < \lambda < 1$ we have that $a_0 = 0$. Moreover, note that $a_1, a_2, \ldots$ are positive integers. If $\alpha$ is rational then $\alpha_m - a_m = 0$ for some $m \in \mathbb{N}$ and the process of computing the partial quotients stops for $n = m$, $\alpha = [0, a_1, a_2, \ldots, a_m]$. If $\alpha$ is irrational then the continued fraction expansion of $\alpha$ is infinite.
We define $p_n, q_n$ recursively by

$$\begin{array}{lll} p_0 = a_0, & p_1 = a_0 a_0 + 1, & p_n = a_n p_{n-1} + p_{n-2} (n \geq 2), \\ q_0 = 1, & q_1 = a_1, & q_n = a_n q_{n-1} + q_{n-2} (n \geq 2) \end{array} .$$

Then $x_n := \frac{p_n}{q_n} = [a_0, a_1, \ldots, a_n]$ is called the $n$th convergent of $\alpha = [a_0, a_1, \ldots]$. If $n$ is odd then $x_n \geq \alpha$ is called an odd convergent and if $n$ is even then $x_n \leq \alpha$ is called an even convergent.

Now that we have defined the convergents of $\alpha$ we also define the so-called intermediate convergents.

**Definition 3.15** *Let $\alpha = [a_0, a_1, \ldots]$ if $\alpha$ is irrational or $\alpha = [a_0, a_1, \ldots, a_m]$ for some $m \in \mathbb{N}$ if $\alpha$ is rational. Then a rational number $\frac{p}{q}$ is an intermediate convergent of $\alpha$ if and only if $p = p_{n-2} + c \cdot p_{n-1}$ and $q = q_{n-2} + c \cdot q_{n-1}$ for some positive integer $n$ (with $n$ less than or equal to $m$ if $\alpha$ is rational) and $c \in \{1, 2, \ldots, a_n - 1\}$. Moreover, $\frac{p}{q}$ is called an odd (even) intermediate convergent if $n$ is odd (even).*

In [13] it is stated that all the best lower approximations of some positive real number $\alpha$ are either even convergents of $\alpha$, even intermediate convergents of $\alpha$ or $\alpha$ itself in case $\alpha$ is an odd convergent of itself. Analogously we have that all the best upper approximations of some positive real number $\alpha$ are either odd convergents of $\alpha$, odd intermediate convergents of $\alpha$ or $\alpha$ itself in case $\alpha$ is an even convergent of itself. So, we can use exactly the same algorithm as is used in [13] except that everything which was 'even' becomes 'odd' and vice versa.

We summarize below all the steps of the algorithm to calculate $B_\lambda(d)$ for any given $0 < \lambda < 1$ and $d \in [\lambda, 1]$, which is obtained by combining the foregoing results of this section.

**Algorithm 3.16** *Let $0 < \lambda < 1$ and $d \in [\lambda, 1]$ be given.*

**step 1.** *Apply the continued fraction algorithm to find consecutively the partial quotients $a_1, a_2, \ldots$ and the corresponding convergents $\frac{p_1}{q_1}, \frac{p_2}{q_2}, \ldots$ of $\lambda$ until we have found an odd convergent $\frac{p_{2n+1}}{q_{2n+1}}$ $(n \geq 0)$, that is smaller or equal than $d$ or we have that $\lambda = \frac{p_N}{q_N} \leq d < \frac{p_{N-1}}{q_{N-1}}$ for some even positive integer $N$. If we have the latter case then we put $s_i = \frac{p_{N-1}}{q_{N-1}}$ and $s_{i+1} = \frac{p_N}{q_N} = \lambda$. Then $d \in [s_{i+1}, s_i)$ and we go to step 2 of the algorithm. So, suppose the former case. If $n = 0$ it follows that $d = s_1 = \frac{p_1}{q_1} = 1$ and we go to step 2. If $n > 0$ then we have that $\frac{p_{2n+1}}{q_{2n+1}} \leq d < \frac{p_{2n-1}}{q_{2n-1}}$ and there exists some unique integer $k$, $0 \leq k < a_{2n+1}$ such that $\frac{(k+1) \cdot p_{2n} + p_{2n-1}}{(k+1) \cdot q_{2n} + q_{2n-1}} \leq d < \frac{k \cdot p_{2n} + p_{2n-1}}{k \cdot q_{2n} + q_{2n-1}}$. It is easily seen that this holds for $k = \lceil \frac{p_{2n-1} - d q_{2n-1}}{d q_{2n} - p_{2n}} \rceil - 1$. By putting $s_i = \frac{k \cdot p_{2n} + p_{2n-1}}{k \cdot q_{2n} + q_{2n-1}}$ and $s_{i+1} = \frac{(k+1) \cdot p_{2n} + p_{2n-1}}{(k+1) \cdot q_{2n} + q_{2n-1}}$ we have that $d \in [s_{i+1}, s_i)$ and we go to step 2.*

**step 2.** *If $d = s_1 = 1$ then we have by Theorem 3.12 that $B_\lambda(d) = \frac{\lambda^2}{2}$ in case of discrete input and $B_\lambda(d) = 0$ in case of fluid input. Thus we are finished in that case. If $d < 1$ then we have found in step 1 consecutive best upper approximations $s_i, s_{i+1}$ of $\lambda$ such that $d \in [s_{i+1}, s_i]$. Next compute by the formulae given in Theorem 3.12 (or eventually Lemma 3.13 if applicable) the values of $B_\lambda(s_i)$ and $B_\lambda(s_{i+1})$. If $d = s_i$ or $d = s_{i+1}$ then we have calculated the value of $B_\lambda(d)$ and we are finished. Else we go to step 3.*

**step 3.** *We have that* $d \in (s_{i+1}, s_i)$. *Put* $\mu = \frac{s_i - d}{s_i - s_{i+1}}$. *Then we have* $d = \mu s_{i+1} + (1-\mu)s_i$ *with* $\mu \in (0,1)$. *Thus by Theorem 3.9 we compute* $B_\lambda(d) = \mu B(s_{i+1}) + (1-\mu)B(s_i)$ *and we are finished.*

**Example.** We consider the model with fluid input and suppose that the input rate $\lambda = \frac{1}{\pi}$. We apply Algorithm 3.16 to compute $B_\lambda(d)$ for $d = 0.31831$. In step 1 of the algorithm we start applying the continued fraction algorithm to $\lambda$ and we consecutively find $a_0 = 0$ , $\frac{p_0}{q_0} = \frac{0}{1}$, $a_1 = 3$, $\frac{p_1}{q_1} = \frac{1}{3}$, $a_2 = 7$, $\frac{p_2}{q_2} = \frac{7}{22}$, $a_3 = 15$, $\frac{p_3}{q_3} = \frac{106}{333}$, $a_4 = 1$, $\frac{p_4}{q_4} = \frac{113}{355}$, $a_5 = 292$ and $\frac{p_5}{q_5} = \frac{33102}{103993}$. We now have that $\lambda < \frac{p_5}{q_5} \leq d$ and we stop applying the continued fraction algorithm. Next we compute that $k = \lceil \frac{p_3 - d \cdot q_3}{d \cdot q_4 - p_4} \rceil - 1 = 55$. So, it follows that $s_i := \frac{k \cdot p_4 + p_3}{k \cdot q_4 + q_3} = \frac{6321}{19858}$ and $s_{i+1} := \frac{(k+1) \cdot p_4 + p_3}{(k+1) \cdot q_4 + q_3} = \frac{6434}{20213}$ are consecutive best upper approximations of $\lambda$ such that $d \in [s_{i+1}, s_i]$. We go to step 2 of the algorithm. By the formula for fluid input in Theorem 3.12 we find that $B_\lambda(s_{i+1}) = B_{\frac{1}{\pi}}(\frac{6434}{20213}) = \frac{278494715 - 88633874\pi}{40426(\pi - 1)}$ and $B_\lambda(s_i) = B_{\frac{1}{\pi}}(\frac{6321}{19858}) = \frac{268797889 - 85547520\pi}{39716(\pi - 1)}$. We have that $d \in (s_{i+1}, s_i)$ and go to step 3 of the algorithm. Putting $\mu = \frac{s_i - d}{s_i - s_{i+1}} = \frac{20213}{50000}$ we obtain

$$B_\lambda(d) = B_{\frac{1}{\pi}}(0.31831) = \mu B_\lambda(s_{i+1}) + (1-\mu)B_\lambda(s_i) = \frac{1363383097 - 433910308\pi}{200000(\pi - 1)} \sim 0.4988368585346.$$

**Remark.** Analogously to the algorithm used in [13] it follows that the number of operations needed for applying Algorithm 3.16 is of order $\log(q)$, where $q$ is the denominator of the best upper approximation $s_{i+1}$ of $\lambda$, which is obtained in step 1 of the algorithm. This follows from the fact that the algorithm is based on the continued fraction expansion of $\lambda$.

# 4 Optimal polling with multiple queues

In this section we obtain results on the minimal long-run average workload for polling systems with $N$ parallel queues. For the moment it is not specified whether the queues are deterministic or exponential. We denote for both the deterministic model as the exponential model with $B_{\lambda_i}(d_i)$ the (long-run) average workload for regular polling with density $d_i$ for queue $i$, where $\lambda_i$ is the (expected) workload arriving in queue $i$ per time-unit. We recall that for the deterministic polling systems the parameter $\lambda_i$ was both for the fluid model and the discrete model defined in the beginning of Section 2. For the exponential model (see Appendix B) the parameter $\lambda_i$ is not defined explicitly. However, we recall that the expected number of jobs arriving per time-unit according to a Poisson process was scaled to be 1, while each job arriving in queue $i$ brings a workload which is exponentially distributed with parameter $\mu_i$. Hence, for the exponential model we have that $\lambda_i = \frac{1}{\mu_i}$ for $i = 1, 2, \ldots, N$.

For a given $\overline{\lambda} := (\lambda_1, \lambda_2, \ldots, \lambda_N)$ system and vector of polling densities $\overline{d} = (d_1, d_2, \ldots, d_N) \in [0,1]^N$ we put

$$B_{\overline{\lambda}}(\overline{d}) := \sum_{i=1}^{N} B_{\lambda_i}(d_i).$$

Thus $B_{\overline{\lambda}}(\overline{d})$ is the average workload in an $\overline{\lambda}$ system (the total over all queues) if the service sequence for queue $i$ would be regular with density $d_i$ for $i = 1, 2, \ldots, N$. However, note that in general the composition of regular sequences is not a feasible polling sequence. But as we show next it provides a lower bound. Indeed, let $U$ be a polling policy applied in an $\overline{\lambda}$ polling system such that service sequence

$u^i$ has density $d_i$ for $i = 1, 2, \ldots, N$. Then it is easily seen that $\sum_{i=1}^{N} d_i = 1$. So, the possible vector of polling densities $\overline{d}$ is restricted to the compact and convex set

$$H^N := \{(x_1, x_2, \ldots, x_N) \in \mathbb{R}^N : x_i \geq 0, i = 1, 2, \ldots, N \text{ and } \sum_{i=1}^{N} x_i = 1\}.$$

Moreover, by Lemma 3.3 we have for $i = 1, 2, \ldots, N$ that $B_i(u^i)$, the average workload in queue $i$, is greater or equal than $B_{\lambda_i}(d_i)$. Thus

$$B(U) = \sum_{i=1}^{N} B_i(u^i) \geq \sum_{i=1}^{N} B_{\lambda_i}(d_i) = B_{\overline{\lambda}}(\overline{d}), \tag{3}$$

which implies that $B_{\overline{\lambda}}(\overline{d})$ is a lower bound on the average workload for any policy $U$ with polling densities $\overline{d}$.

By Appendix A, we have that this lower bound $B_{\overline{\lambda}}(\overline{d})$ is convex in the vector of polling densities $\overline{d} = (d_1, d_2, \ldots, d_N)$. So, $B_{\overline{\lambda}}(\overline{d})$ has a minimum over the convex and compact set $H^N$ and it follows that this minimum is a lower bound on the average workload for any polling policy $U$ for which each corresponding service sequence $u^i$ has some density. Moreover, analogously to Theorem 25 of [2] it follows that this minimum is a lower bound on the average workload for any polling policy $U$. By putting for an $\overline{\lambda}$ system

$$D^*(\overline{\lambda}) := \{\overline{d} \in H^N : B_{\overline{\lambda}}(\overline{d}) = \min_{\overline{x} \in H^N} B_{\overline{\lambda}}(\overline{x})\} \tag{4}$$

as the set of possible densities for which the lower bound $B_{\overline{\lambda}}(\overline{d})$ is minimal and

$$B^*(\overline{\lambda}) := B_{\overline{\lambda}}(\overline{d}), \text{ where } \overline{d} \in D^*(\overline{\lambda})$$

as the minimal lower bound, we can summarize with the following proposition.

**Proposition 4.1** *For any $\overline{\lambda}$ polling system we have that $D^*(\overline{\lambda})$ is a nonempty compact and convex subset of $H^N$. Moreover, for any polling policy $U$ we have that*

$$B(U) \geq B^*(\overline{\lambda}).$$

Suppose that we have an $\overline{\lambda}$ system for which $\sum_{i=1}^{N} \lambda_i \geq 1$ in case of exponential queues or $\sum_{i=1}^{N} \lambda_i > 1$ in case of deterministic queues. Then for every $(x_1, x_2, \ldots, x_N) \in H^N$ there exists some $i$ for which $x_i \leq \lambda_i$ (respectively $x_i < \lambda_i$) which implies that $B^*(\overline{\lambda}) = \infty$ and thus $B(U) = \infty$ for every polling policy $U$. Thus such system are unstable and optimal policies do not exist. We say that polling systems for which $\sum_{i=1}^{N} \lambda_i < 1$ in case of exponential queues or $\sum_{i=1}^{N} \lambda_i \leq 1$ in case of deterministic queues are stable. For stable systems it follows directly from the results on polling to one queue that $B^*(\overline{\lambda}) < \infty$. Moreover, there exist policies $U$ for which $B(U) < \infty$. In the sequel we consider only stable polling systems.

Note that the problem of minimizing the lower bound $B_{\overline{\lambda}}(\overline{d})$ over the set $H^N$ and finding the minimum value $B^*(\overline{\lambda})$ is a problem of minimizing a convex functions in multiple variables over a convex and compact set. There are standard techniques for this, but the best way depends on the time it takes to compute the function value $B_{\overline{\lambda}}(\overline{d})$ for particular $\overline{d}$. The similar problem is considered for a routing system with parallel queues in the papers [9], [10] and [13]. For the polling model most results follow in an analogously way.

### 4.1 Optimality results for two queues

We first consider the case $N = 2$ in particular. Then we have for any $\overline{d} \in H^N$ that $\overline{d} = (d, 1-d)$ for some $d \in [0,1]$. Thus minimizing the function $B_{\overline{\lambda}}(\overline{d})$ over the set $H^N$ comes down to minimizing the function $B_{\overline{\lambda}}(d, 1-d)$, which is convex in the single variable $d$, over the interval $[0,1]$. Putting

$$opt(\overline{\lambda}) := \{d \in [0,1] : B_{\overline{\lambda}}(d, 1-d) \text{ is minimal } \}$$

we have by Proposition 4.1 that $opt(\overline{\lambda})$ is a nonempty closed subinterval of $[0,1]$. Moreover, given $d \in opt(\overline{\lambda})$, an optimal polling policy $U$ for the system is obtained in the following way. Let $U$ be such that the serving of queue 1 is according to a regular service sequence $u^1$ of density $d$. Then the service sequence $u^2$ for queue 2 is regular with density $1-d$ and thus

$$B(U) = B_{\lambda_1}(d) + B_{\lambda_2}(1-d) = B_{\overline{\lambda}}(\overline{d}) = B^*(\overline{\lambda}).$$

Hence $U$ is optimal according to Proposition 4.1. Thus in case of only two queues the lower bound of Proposition 4.1 is attained and we have the following proposition.

**Proposition 4.2** *For every stable $(\lambda_1, \lambda_2)$ polling system there exists a nonempty closed interval $I \subseteq [0,1]$ such that for every $d \in I$ any regular polling policy $U$, for which the corresponding service sequences $u^1$ and $u^2$ are regular with densities $d$ and $1-d$ respectively, is optimal and $B(U) = B^*(\overline{\lambda})$.*

Th structural result of Proposition 4.2 on optimal policies holds for polling systems with general input and service times. A similar result as Proposition 4.2 holds for parallel routing systems (see [2]). For these parallel routing systems the problem of computing an optimal (routing) density $d$ (and thus a corresponding optimal routing policy $U$) is dealt with in several papers. In [10] this problem is considered for Poisson arrivals and both servers have exponential service times. In [9] and [13] the problem is considered for a deterministic system with constant inter-arrival times and constant service times for both servers. Similar methods as considered in these papers can also be applied for the polling system with two queues to determine an optimal density and thus an optimal policy.

### 4.2 Numerical experiments

We have used such methods to calculate the optimal polling density $d$ (and thus also the corresponding optimal regular polling policy) for deterministic $(\lambda_1, \lambda_2)$ polling systems where we fix the ratio $\frac{\lambda_1}{\lambda_1+\lambda_2}$ to be equal to 0.37 by putting $\lambda_1 = 0.37\rho$ and $\lambda_2 = 0.73\rho$, where the load $\rho$ is varied from 0 to 1. For this family of polling systems we computed for both the discrete and the fluid case the value of the optimal polling density $\alpha_{opt}$. For $0 < \rho < 1$ we define $\alpha_{opt}$ to be the rational number of lowest denominator which is contained in the nonempty closed subinterval $opt(\lambda_1, \lambda_2)$ of $(0,1)$. We note that in almost all cases for varying $(\lambda_1, \lambda_2)$ the set $opt(\lambda_1, \lambda_2)$ consists of only one (rational)point, which by the above definition is the point $\alpha_{opt}$. Moreover, in the few cases that $opt(\lambda_1, \lambda_2)$ consists of more than one point, it still holds by lemma 4.6.4 in [23] that $\alpha_{opt}$ is unique and thus well defined. In case $\rho = 1$ it is easily seen that the interval of stability consists of the single point $\frac{\lambda_1}{\lambda_1+\lambda_2}$ which is fixed to be 0.37 for this family of polling systems. Thus it is clear that for $\rho = 1$ the value of the optimal polling density $\alpha_{opt}$ has to be 0.37 for both the discrete and the fluid case. For both the discrete and the fluid case we have computed the value of $\alpha_{opt}$ for varying $\rho$ by implementing Algorithm 3.16 and the appropriate standard techniques for minimizing a convex function in a Maple program using exact computations. In Figure 1 the value of $\alpha_{opt}$ is plotted for varying load for the discrete case, while in Figure 2 this is plotted for the fluid case. In both figures the load varies from 0.75 to 1, since it turned out that for smaller loads the value of $\alpha_{opt}$ is always equal to 0.5, which corresponds to the round robin

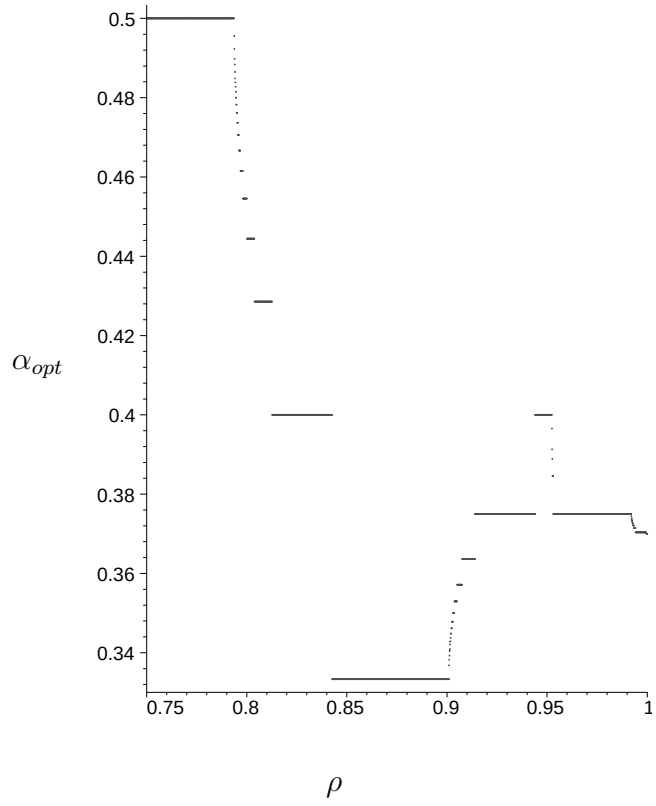


Figure 1: The optimal polling density for varying load for discrete input

polling policy. So, in the figures we have restricted to an interval with high loads $\rho$, since this is by far the most interesting part of the interval.

**Remark.** We see in Figure 1 and Figure 2 that for this example where $\frac{\lambda_1}{\lambda_1+\lambda_2} = 0.37$ the optimal polling density $\alpha$ takes several rational values varying between $\frac{1}{3}$ and $\frac{1}{2}$. We notice that only at the very end of the interval, where $\rho = 1$ we have that $\alpha = 0.37$, which is the fraction from the total arriving workload that arrives at queue 1. It is also clearly seen that for low traffic intensity $\rho$ the optimal ratio $\alpha$ is greater than 0.37, which means that the queue with smaller arriving workload is served relatively often for small load. In fact $\alpha = \frac{1}{2}$, which corresponds to the round robin policy, is optimal for quite a large part of the interval. Another thing we note is that $\alpha$ does not decrease monotonically from $\frac{1}{2}$ for small load $\rho$ to 0.37 for $\rho = 1$. For example for both the discrete model and the fluid model there is some part of the interval where $\alpha = \frac{1}{3}$, which is smaller than 0.37. In that case the queue with the higher workload input is served relatively often. The fact that the optimal polling density $\alpha = \frac{1}{3}$ on some part of the interval can be explained from the fact that optimal densities tend to have a low denominator in general, but that for example density $\frac{1}{2}$ gives no longer a stable policy for such loads. If the load is increased even further than also $\frac{1}{3}$ gives no longer a stable policy until finally for $\rho = 1$ we have that only $\alpha = 0.37$ gives a stable policy.

However, not every change of the optimal $\alpha$ for increasing load is explained by such stability considerations. For example it is clearly seen in both Figure 1 and Figure 2 that $\frac{3}{8}$ is the optimal polling density on two disjunct intervals of varying load, while for loads between these intervals we have that $\alpha$ has (among other values) a maximum of $\frac{2}{5}$.

Another point of interest we notice in the figures is that for any given load the $\alpha$ for the fluid model is equal or larger than the $\alpha$ for the discrete model. Intuitively this is explained from the fact that for

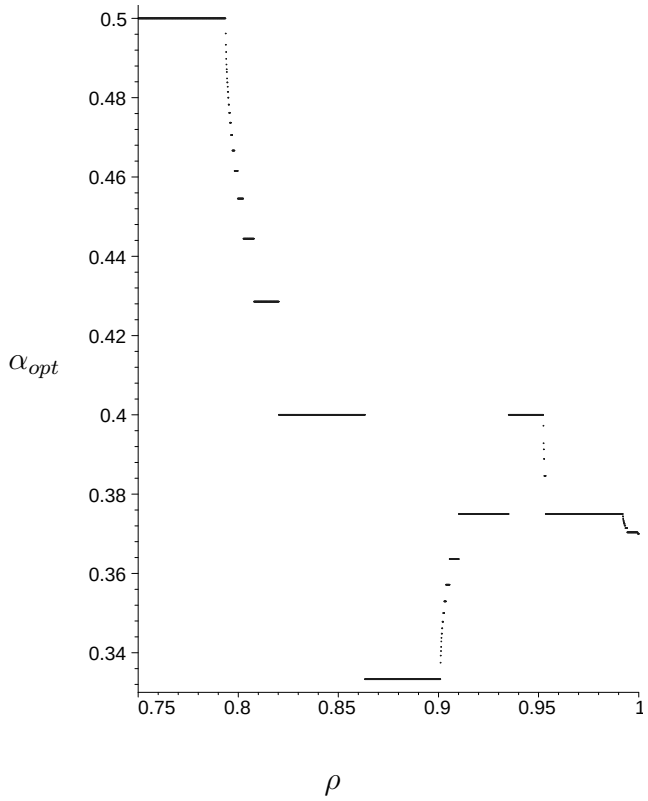


Figure 2: The optimal polling density for varying load for fluid input.

the discrete model the (whole) packets of workload enter the buffer at once, while in the fluid model it takes more time before the same amount of workload has entered the buffer. So, for the discrete case there should be more urgency to serve the queue where the packets of larger workload arrive. This explains that for the discrete model the optimal service rate for the queue with larger input is never lower than for the fluid model and in some cases higher.

We have also some two-dimensional plots (see Figure 3 and Figure 4), in which the input rates for the two queues vary independently of each other. To plot these figures we have calculated the optimal polling density $\alpha$ for various $(\lambda_1, \lambda_2)$ systems in the triangle area $\{(x_1, x_2) : x_1, x_2 \geq 0, x_1 + x_2 \leq 1)\}$. In Figure 3 and Figure 4 there is a black dot at a point $(\lambda_1, \lambda_2)$ within these triangle area if one or more of the neighboring points has another value of $\alpha$.
The resulting Figure 3 and Figure 4 are quite similar looking fractal type pictures. It is obvious that both pictures are symmetrical in the diagonal $\lambda_1 = \lambda_2$. In both figures it is easy to identify the largest (and also most central containing the diagonal $\lambda_1 = \lambda_2$) white area within the triangle which corresponds to $\alpha = \frac{1}{2}$. The second largest white areas which are symmetrically situated with respect to the diagonal $\lambda_1 = \lambda_2$ correspond to $\alpha = \frac{1}{3}$ and $\alpha = \frac{2}{3}$ respectively. We note that these large white areas corresponding to $\alpha$ with low denominator are somewhat bigger in Figure 4 (the fluid model) than in Figure 3 (the discrete model). This confirms our conjecture that for the discrete model the value of $|\alpha - \frac{1}{2}|$ is always greater or equal than for the fluid model.

### 4.3 Structural results for deterministic polling systems

To obtain more structural results like Proposition 4.2 we restrict ourselves to deterministic polling systems in the sequel of this section. The first result in that case follows from the following properties

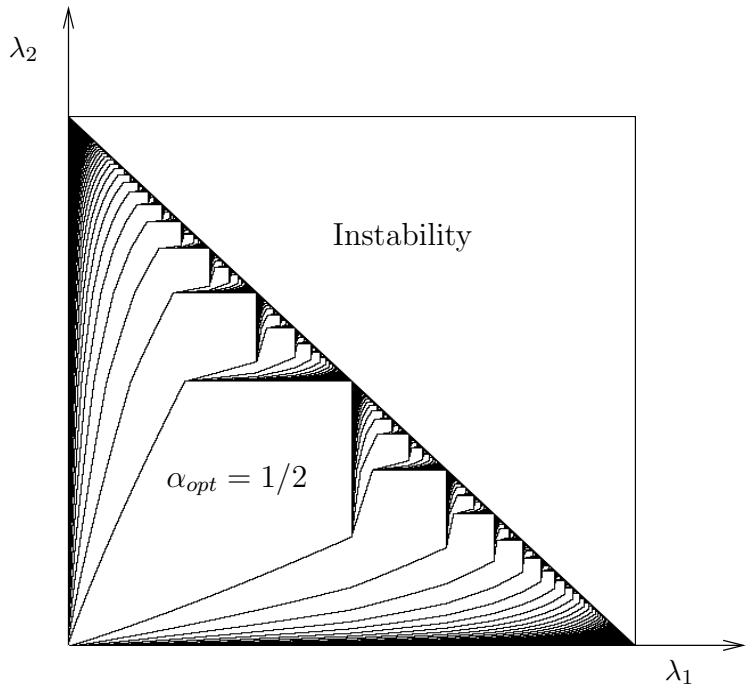


Figure 3: The regions of optimality for discrete input

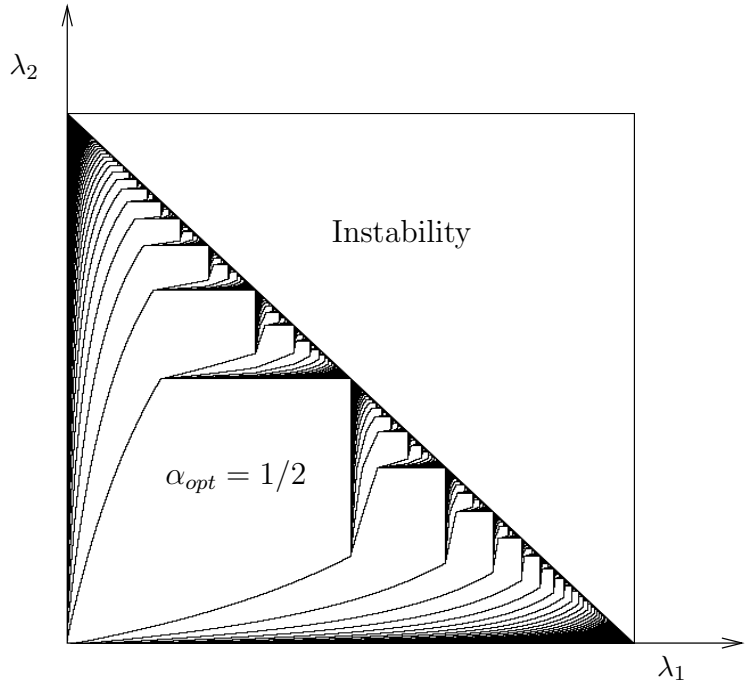


Figure 4: The regions of optimality for fluid input

(which all follow from Section 3) of the function $B_\lambda(d)$, where $\lambda \in [0,1]$ is the input rate in some queue.

- $B_\lambda(d) =$ is convex for $d \in [\lambda, 1]$, moreover $B_\lambda(d) = \infty$ for $d < \lambda$.
- $B_\lambda(d)$ is piecewise linear in $d$ and the slope changes only in the best upper approximations of $\lambda$ which are rational numbers.
- $\lim_{\varepsilon \downarrow 0} \frac{B_\lambda(\lambda) - B_\lambda(\lambda+\varepsilon)}{\varepsilon} = \infty$ if $\lambda$ is irrational.

From these properties the following theorem follows analogously as the similar result (Theorem 7.9) for deterministic routing systems in [13].

**Theorem 4.3** *Consider a deterministic $\overline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ polling system with $N \geq 2$ queues such that $\sum_{i=1}^N \lambda_i < 1$. Then there exists some $x = (x_1, x_2, \ldots, x_N) \in opt(\overline{\lambda})$ and some $j \in \{1, 2, \ldots, N\}$ such that for every $i \neq j$ it holds that $x_i$ is a best upper approximation of $\lambda_i$.*

**Corollary 4.4** *Let $\overline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ be a deterministic polling system with $\sum_{i=1}^N \lambda_i < 1$. Then the set $opt(\overline{\lambda})$ contains a point with rational coordinates.*

**Remark 4.5** In section 7 of [13] an algorithm is given to obtain such a rational point for deterministic parallel routing systems. With a slight modification the same algorithm can be used to find a rational point in the set $opt(\overline{\lambda})$ for deterministic polling systems. Moreover, we note that in most cases the set $opt(\overline{\lambda})$ consists of only one point, which according to Corollary 4.4 and Theorem 4.3 has rational coordinates and all coordinates except at most one are best upper approximations of the input rate in the corresponding queue.

Combining Theorem 4.2 and Corollary 4.4 it follows that for a deterministic polling system with only 2 queues and $\lambda_1 + \lambda_2 < 1$ there exists an optimal regular polling policy with rational densities. It is easily seen that such a policy is periodic. So, we have obtained the following structural result on optimal policies.

**Theorem 4.6** *For every deterministic $(\lambda_1, \lambda_2)$ polling system with $\lambda_1 + \lambda_2 < 1$ there exists an optimal regular and periodic polling policy $U$ with $B(U) = B^*(\overline{\lambda})$.*

We note that a similar result on the optimality of periodic policies for deterministic parallel routing systems with only 2 queues was obtained in [9]. However, for a deterministic $\overline{\lambda}$ polling systems with more than 2 queues we do not have such a result. Indeed, in that case for a rational point $(d_1, d_2, \ldots, d_N) \in opt(\overline{\lambda})$ there does in general not exist some policy $U$ such that service sequence $u^i$ is regular with density $d_i$ for $i = 1, 2, \ldots, N$. Note that if such a policy $U$ would exist than $U$ would be periodic and also optimal, since $B(U) = B^*(\overline{\lambda})$. However, if such a policy does not exist it is possible that for an optimal policy $V$ it holds that $B(V) > B^*(\overline{\lambda})$. Moreover, such optimal policy $V$ does not necessarily have rational densities $(d_1, d_2, \ldots, d_N) \in opt(\overline{\lambda})$ and thus it may not be periodic. Nevertheless we think that in most cases there also exists an optimal periodic policy for deterministic polling systems with more than 2 queues. Besides, the following can be proved analogously to the similar result for deterministic parallel routing systems considered in chapter 5 of [23].

**Theorem 4.7** *For a deterministic and stable $\overline{\lambda}$ polling system with rational input rates $\lambda_i$ for $i = 1, 2, \ldots, N$ there exists an optimal routing policy $U$ which is periodic.*

# A Appendix 1: On the optimality of the bracket sequence and its convexity for the arrival-driven polling model

Consider an arrival driven polling model (as in section 9.5 of [2]) with i.i.d. service and independent i.i.d. inter-arrival times in every queue. Note that both the deterministic and exponential model which we analyze in this paper can be described as such arrival driven model with i.i.d. service and independent i.i.d. inter-arrival times. Moreover, if for queue $i$ in a polling system with these assumptions we have that $\tau$ is the mean inter-arrival time and $\frac{1}{\mu_i}$ is the mean service time of a job arriving in queue $i$, then $\lambda_i$, the (expected) workload arriving in queue $i$ per time-unit as described in the beginning of Section 4 is given by $\lambda_i = \frac{1}{\tau\mu_i}$.

For one queue Lemma 54 in [2] gives for service sequence $a$ that $V_n(a)$, the expected workload in the queue for the $n$-th arrival starting with an empty queue, is multimodular in $a$. Since for any service sequence hence also for the bracket sequence $a^p(\theta)$ it holds that

$$V_n(a_n^p(\theta), \ldots, a_1^p(\theta)) \leq V_{n+1}(a_{n+1}^p(\theta), \ldots, a_1^p(\theta)),$$

we have that Lemma 1 of [14] can be applied. Hence the average workload for the bracket sequence with initial phase $\theta$ is independent of $\theta$ and it is convex in the density $p$, since

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} V_n(a_1^p(\theta), \ldots, a_n^p(\theta)) = \lim_{n\to\infty} \int_0^1 V_n(a_1^p(\theta), \ldots, a_n^p(\theta)) d\theta. \tag{5}$$

For the optimality of the bracket sequence we can apply Theorem 50 of [2], however it has the nonstandard assumption that the initial distribution is the stationary regime that corresponds to the policy that never takes vacation, and we would prefer to start with an initially empty queue. The approach to prove this is similarly to the way it is done in the sections 4.6.3 and 4.6.4. of [2]. Let us first assume that the inter-arrival(service) time is almost surely bounded from below(above), hence for some $k \in \mathbb{N}$ we have (we use the notation of [2] where for queue $i$, $\tau_n^i$ denotes the $n$ -th inter-arrival time and $\sigma_n^i$ denotes the $n$-th service time) for queue $i = 1, 2$ and all $n$, almost surely

$$\sum_{l=1}^{k} \tau_{n+l}^i \geq \sigma_n^i.$$

Then for queues $i = 1, 2$

$$V_{n+1}^i(k, a_1^i, \ldots, a_n^i) = V_n^i(a_1^i, \ldots, a_n^i)$$

and hence both satisfy the conditions of Theorem 7 in [2], from which it follows that for queue $i = 1, 2$ the bracket sequence with density $p^i$ and any initial phase $\theta^i$ is optimal in the class of policies with upper density smaller than or equal to $p^i$. This gives together with (5) for service sequence $a = (a^1, a^2)$ with densities for queue 1 (2) equal to $p^1 (p^2 = (1- p^1))$ that

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{2} V_n^i(a_1^i, \ldots, a_n^i) \geq \lim_{n\to\infty} \sum_{i=1}^{2} \int_0^1 V_n^i(a_1^{p^i}(\theta), \ldots, a_n^{p^i}(\theta)) d\theta.$$

The right-hand side, let us denote it with $V_\infty(p^1, p^2)$, is a limit of convex functions and hence $V_\infty(p^1, p^2)$ is convex in $(p^1, p^2)$. The proof for unbounded inter-arrival and service times is similar to the proof of Theorem 22 in [2].

Combining the above results gives the following theorem for an arrival-driven polling model with the i.i.d. and independence assumptions.

**Theorem A.1** *The optimal polling policy for $N = 2$ queues can be obtained by minimizing the convex function $V_\infty(p^1, p^2)$.*

# B Appendix 2: Computation of the optimal policy in the Markovian case

In this section, we show how the optimal polling sequence can be computed in a stochastic system made of two Poisson arrival queues and one exponential server.

As in Appendix A, we consider an arrival driven polling model (as in section 9.5 of [2]) with i.i.d. service and independent i.i.d. inter-arrival times in every queue. In this section, we show how the optimal polling between two queues can be computed when the service is exponential (with rate $\mu_i$ in queue $i$) and the inter-arrivals are synchronous and exponential (with rate $\lambda$ in both queues).

We call $m(k)$ the $k-th$ polling decision (at arrival $k$). We set $m(k)=1$ if the server is allocated to queue 1 and $m(k)=0$ if the server is allocated to queue 2.

We now show that under the polling by a periodic decision word $m$ of period $\ell$, the number of customers in a queue can be modeled by a Markov Process. The behavior of the number of customers in one queue (say queue 1, and index 1 will be omitted in the following) of the system is given by a continuous time Markov chain $X_t$ which state space is equal to $\mathbb{N}\times\{1,\ldots,\ell\}$. The first entry represents the number of customers in the queue at time $t$ while the second entry represents the current letter of the polling $m$ (modulo $\ell$).

The continuous time Markov chain $X_t$ is a quasi birth and death process whose generator $Q$ is given by

$$Q=\begin{bmatrix} C & A_0 & 0 & 0 & \ldots \\ A_2 & A_1 & A_0 & 0 & \ldots \\ 0 & A_2 & A_1 & A_0 & \ldots \\ 0 & 0 & A_2 & A_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where matrices $A_0, A_1, A_2$ and $C$ are of size $\ell\times\ell$, with $A_0[i,(i+1) \mod \ell]=\lambda$, and is null everywhere else, $A_2[i,i]=m(i)\mu$, and is null everywhere else, $C[i,i]=-\lambda$, and is null everywhere else, and $A_1[i,i]=-\lambda-m(i)\mu$ and is null everywhere else.

An example of the infinitesimal generator $Q$ of $X_t$ when the polling is $m=(110)^\infty$ is displayed in Figure 5

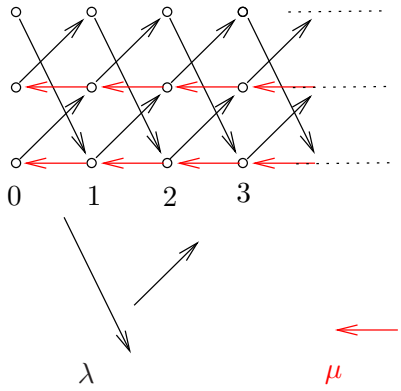


Figure 5: graph of the infinitesimal generator when the polling is $m=(110)^\infty$

Let $\pi$ be the invariant measure of the process $X_t$ (when it exists). This probability satisfies

$$\pi Q=0. \tag{6}$$

We now refine the notation by introducing block vectors $\pi_n$ of dimension $\ell$ whose $i$th entry $(\pi_n(i))$ represents the stationary probability to have $n$ customers in the system when the current polling decision is $m(i)$. Hence $\pi_k(1) + \cdots + \pi_k(\ell)$ is the stationary probability to have $k$ customer in the system. We will not try to compute $\pi$ directly which can be quite hard, but we will rather determine its generating function. Let $\overline{D}(0,1)$ be the closed unit disk. The generating function of $\pi$ is the function $\Pi(z)$ from $\overline{D}(0,1)$ to $\mathbb{C}^\ell$ defined by

$$\Pi(z) = \sum_{n=0}^{\infty} z^n \pi_n.$$

The following theorem will be used to make sure that the stationary distribution (as well as the function $\Pi(z)$) exists.

**Lemma B.1** *The process $X_t$ is positive recurrent if and only if $\frac{\ell\lambda}{a\mu} < 1$.*

**Proof.** This proof is based on Theorem 1.3.2 of [19] which states that $X_t$ is positive recurrent if and only if $pA_2\mathbf{1} > pA_0\mathbf{1}$, where $\mathbf{1}$ is the column vector with all entries equal to one, and $p$ is the stationary distribution vector of the finite generator $A = A_0 + A_1 + A_2$, (*i.e* $pA = 0$ and $p\mathbf{1} = 1$).

Let us compute $p$. Using formulas of $A_0$, $A_1$ and $A_2$, we get $A = -\lambda I + A_0$. It should be clear that $p_1 = p_2 = \ldots = p_\ell = 1/\ell$. Hence, the stability condition becomes $pA_2\mathbf{1} = a\mu/\ell > pA_0\mathbf{1} = \lambda$. ■

**Lemma B.2** *Let $K(z)$ be the $\ell \times \ell$ matrix defined by*

$$K(z) = \begin{bmatrix} \mu(1-z)m(1) - \lambda z & \lambda z^2 & & \\ & \ddots & & \\ & \ddots & & \\ & & & \lambda z^2 \\ \lambda z^2 & & & \mu(1-z)m(\ell) - \lambda z \end{bmatrix}.$$

*Then the generating function satisfies the functional equation with kernel $K(z)$,*

$$\Pi(z)K(z) = \pi_0\mu(1-z)M. \tag{7}$$

**Proof.** Using the global balance equation (6) we get the induction

$$\pi_0 C + \pi_1 A_2 = 0, \tag{8}$$
$$\pi_{n-1}A_0 + \pi_n A_1 + \pi_{n+1}A_2 = 0, \quad \forall n \geq 1. \tag{9}$$

By multiplying the second equation by $z^{n+1}$ and by summing it follows

$$\begin{aligned} \Sigma_{n=1}^{\infty}\pi_{n-1}z^{n-1}z^2A_0 + \pi_n z^n z A_1 + z^{n+1}\pi_{n+1}A_2 &= 0, \\ \Pi(z)(z^2A_0 + zA_1 + A_2) - \pi_1 A_2 z - \pi_0 A_2 - \pi_0 A_1 z &= 0, \end{aligned}$$

which gives $\Pi(z)(z^2A_0 + zA_1 + A_2) = \pi_0\mu(1-z)M$, where

$$M = \begin{bmatrix} m(1) & & 0 \\ & \ddots & \\ 0 & & m(\ell) \end{bmatrix}.$$

■

Let us now study the zeros of $K(z)$. More precisely we will focus on the zeros inside the unit disk since $\Pi(z)$ is a power series with radius of convergence one. Let us call $\Delta(z)$ the determinant of the matrix $K(z)$. Using the definition of the matrices $A_0, A_1$ and $A_2$, one gets after direct computations

$$\Delta(z) = (-1)^{\ell+1}\lambda^\ell z^{2\ell} + (-\lambda z)^{\ell-a}(\mu - (\mu+\lambda)z)^a.$$

**Lemma B.3** *If $\frac{\ell\lambda}{a\mu} < 1$ then the number of non-null roots of $\Delta(z)$ inside the unit disk is $a$. Moreover, 0 is a root with multiplicity $\ell - a$.*

**Proof.** It is obvious that 0 is a root with multiplicity $\ell - a$. Let $f(z) = (-\lambda z)^{\ell-a}(\mu - (\mu+\lambda)z)^a$. Obviously, $f$ has exactly $\ell$ roots inside the unit disk.

Now, let $|z| = 1+\varepsilon$. $|\Delta(z) - f(z)| = \lambda^\ell(1+2\ell\varepsilon) + o(\varepsilon)$ and $|f(z)| \geq \lambda^\ell + (\lambda^\ell(\ell-a) + \mu^{a-1}(\lambda+\mu)\lambda^{l-a}a)\varepsilon + o(\varepsilon)$. A direct computation using the stability condition $\mu > \ell\lambda/a$, shows that $|\Delta(z) - f(z)| < |f(z)|$ if $\varepsilon$ is small enough. Then, the result follows from Rouché's theorem. ■

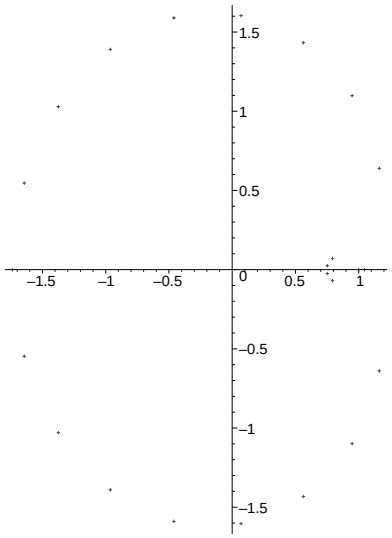


Figure 6: Roots of $\Delta(z)$ when $a = 5, \ell = 18$ and $\lambda = 1, \mu = 4$.

To illustrate the previous lemma, Figure 6 displays all the roots of $\Delta(z)$ when $a = 5, \ell = 18$ and $\lambda = 1, \mu = 4$. The number of roots inside the unit disk (including 1, which is always a root of $\Delta(z)$) is exactly 5.

**Theorem B.4** *If $z_i$ is the $i$th non-null root of $\Delta(z)$ in the unit disk and $v_i$ is the right eigenvector of the eigenvalue 0 of $K(z_i)$, then $\pi_0$ is a solution of the system :*

$$\left\{ \begin{array}{cc} \pi_0(j) = 0, & \text{if } m(j) = 0. \\ \pi_0 v_i = 0, & \forall\, i \in \{1,\ldots,a\} \quad s.t. \quad z_i \neq 1 \\ & \\ \pi_0 \mathbf{1} = a/\ell - \lambda/\mu, & \text{when } z_i = 1, \end{array} \right\}. \tag{10}$$

*where $\mathbf{1}$ is the column vector with all its components equal to 1.*

**Proof.** If $|z_i| < 1$, then it comes by definition of $v_i$ that $(1-z_i)\mu\pi_0 M v_i = 0$. Since $z_i \neq 1$ and $\pi_0(j) = 0$ if $M_{jj} = 0$, yields $\pi_0 v_i = 0$. Note that the rank of the matrix $K(z_i)$ is $\ell - 1$ so that the vector $v_i$ is unique up to a multiplicative constant.

The case $z_i = 1$ has to be handled differently since $(1 - z_i) = 0$ and $K(1)\mathbf{1} = 0$.

$$\mu\pi_0 M\mathbf{1} = \lim_{z\to 1} \frac{\Pi(z)K(z)\mathbf{1}}{1-z} = \lim_{z\to 1} \Pi(z)\frac{K(z)\mathbf{1} - K(1)\mathbf{1}}{1-z} = -\Pi(1)K'(1) = \mu a/\ell - \lambda,$$

since $\Pi(1) = (1/\ell \cdots 1/\ell)$ (easy computation). ■

Now. we turn back to the two queue system. Once $\pi_0$ has been computed, it is possible to compute $\mathbb{E}(N_j(m))$ and $\mathbb{E}(V_j(m))$ $(j = 1, 2)$ being respectively the expected number of customers and the expected workload in the queue $Q_j$ when the polling is made according to $m$.

Since $\mathbb{E}(N_j(m)) = \frac{d\Pi(z)\mathbf{1}}{dz}|_{z=1}$, introducing the vector $\hat{k}(z)$ which verifies $K(z)\hat{k}(z) = \mathbf{1}$ yields

$$\mathbb{E}(N_j(m)) = \frac{d}{dz}\big(\Pi(z)\mathbf{1}\big)|_{z=1} = \mu_j\pi_0 M_j \frac{d}{dz}\big((1-z)\hat{k}(z)\big)|_{z=1}. \tag{11}$$

In turn, this gives a way to compute the expected workload using the fact that

$$\mathbb{E}(V_j(m)) = \frac{1}{\mu_j}\mathbb{E}(N_j(m)) \tag{12}$$

Now, if we consider the system made of two synchronous queues, the polling sequence in the second queue is the complementary sequence $(\overline{m})$ of the polling in the first one $(m)$. The total workload is $\mathbb{E}(V_1(m)) + (\mathbb{E}(V_2(\overline{m}))$. The optimal polling is a bracket sequence $\omega(\alpha_{opt})$, which density $\alpha_{opt}$ has to be computed by:

$$\alpha_{opt} = argmin_d \left(\mathbb{E}(V_1(w(d))) + \mathbb{E}\left(V_2(\overline{w(d)})\right)\right).$$

Using the fact that this function is convex in the density $d$ (see Appendix A), the optimal density $\alpha_{opt}$ can be computed numerically by gradient descent in a similar fashion as in [10, 8]. We have run several computations using a Maple program implementing the algorithm above (For each triple $(\lambda, mu_1, \mu_2)$, compute the roots of $\Delta(z)$, then compute the corresponding eigenvectors, then compute $\pi_0$ and finally the corresponding workload).

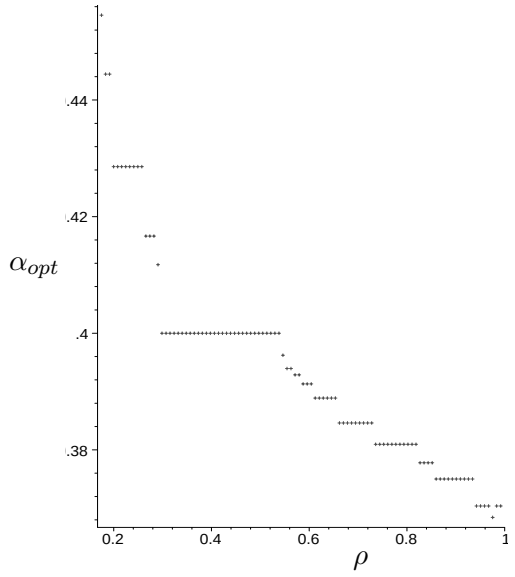


Figure 7: Computation of the optimal polling ratio $\alpha_{opt}$ when the load $\rho$ varies from .2 to 1.

Figure 7 shows the values of $\alpha_{opt}$, the optimal frequency with which the first queue is served (as in the deterministic case). We have chosen $\lambda = 1$ (in both queues) while $\mu_1/(\mu_2 + \mu_1) = 37/100$ to match the intensities of the deterministic cases. Although one could expect that the stochastic assumption should have a smoothing effect on the value of $\alpha_{opt}$, the numerical experiments in Figure 7 still suggest that $\alpha_{opt}$ is highly non-differentiable with many flat zones and many cusps. It is an open problem to state about the continuity of $\alpha_{opt}$.

One can also observe that even when the two queues are not identical, $\alpha_{opt} = 1/2$ when the system is lightly loaded, as in the deterministic case, although this is only true for very small loads here ($\rho < 0.18$, to be compared with $\rho < 0.78$ in the deterministic case). Finally, when the system is heavily loaded, $\alpha_{opt}$ converges to the ratio of the service intensities, namely 0.37, as expected.

## References


[1] E. Altman, B. Gaujal, and A. Hordijk. Optimal open-loop control of vacations, polling and service assignment. *Queueing systems*, 36:303–325, 2000.

[2] E. Altman, B. Gaujal, and A. Hordijk. *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity.* LNM. Springer Verlag, 2003.

[3] Y. Arian and Y. Levy. Algorithms for generalized round robin routing. *Oper. Res. Lett.*, 12:313–319, 1992.

[4] S. Borst. *Polling systems*. PhD thesis, Tilburg : Katholieke Universiteit Brabant, 1994.

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

[6] O.J. Boxma and H. Tagaki, editors. *Special Issue on Polling systems*. Queueing Systems, 1992.

[7] E.G. Coffman, A.A. Puhalskii, and M.I. Reiman. Polling systems with zero switchover times: a heavy-traffic averaging principle. *Annals of Applied Probability*, 5:681–719, 1995.

[8] M.B. Combé and O.J. Boxma. Optimization of static traffic allocation policies. *Theoretical Computer Science*, 125:17–43, 1994.

[9] B. Gaujal and E. Hyon. Optimal routing policy in two deterministic queues. *Calculateurs Parallèles*, 13:601–634, 2000.

[10] B. Gaujal, E. Hyon, and A. Jean-Marie. Optimal open-loop routing in two parallel queues with exponential service times. In IEEE, editor, *Wodes*, Reims, 2004. Long version available in http://www.inria.fr/rrrt/rr-5109.html.

[11] G.H. Hardy and E.M. Wright. *An introduction to the theory of numbers*. Oxford University Press, 1960. 4th edition.

[12] A. Hordijk and J.A. Loeve. Optimal noncyclic server allocation in a polling model. In IEEE, editor, *Conference on Decision and Control*, volume 36, pages 2941–2945, 1997.

[13] A. Hordijk and D.A. van der Laan. On the average waiting time for regular routing to deterministic queues. Technical Report 2002-24, Leiden University, 2002. Available on www.math.leidenuniv.nl/reports/2002-24.shtml.

[14] A. Hordijk and D.A. van der Laan. Note on the convexity of the stationary waiting time as a function of the density. *Prob. Engin. Inform. Sci.*, 17:503–508, 2003.

[15] A. Hordijk and D.A. van der Laan. Periodic routing to parallel queues and billiard sequences. *Math. Methods Oper. Res.*, 2004. To appear.

[16] A. Hordijk and D.A. van der Laan. The unbalance and bounds on the average waiting time for periodic routing to one queue. *Math. Methods Oper. Res.*, 59:1–23, 2004.

[17] J.B. Kruskal. Work-scheduling algorithms: a non-probabilistic queueing study. *Bell Syst. Techn. Journal*, 48:2963–2974, 1969.

[18] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002.

[19] M.F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications*. Marcel Dekker, 1989.

[20] O. Perron. *Die Lehre von den Kettenbrüchen*. Stuttgart: B.G. Teubner Verlagsgesellschaft, 1954.

[21] H. Tagaki. *Analysis of Polling Systems*. MIT Press, Cambridge, 1986.

[22] R. Tijdeman. Fraenkel's conjecture for six sequences. *Discrete Math.*, 222:223–234, 2000.

[23] D.A. van der Laan. *The structure and perfomance of optimal routing sequences*. PhD thesis, Leiden University, 2003.