Probability in the Engineering and Informational Sciences, **21**, 2007, 157–187. Printed in the U.S.A.
DOI: 10.1017/S0269964807070118

ON THE OPTIMAL OPEN-LOOP CONTROL POLICY FOR DETERMINISTIC AND EXPONENTIAL POLLING SYSTEMS

BRUNO GAUJAL
Lab. ID-IMAG (INRIA, CNRS, INPG, UJF)
38330 Montbonnot Saint Martin, France
E-mail: bruno.gaujal@imag.fr

ARIE HORDIJK
Department of Mathematics
Leiden University
2300 RA Leiden, The Netherlands
E-mail: hordijk@math.leidenuniv.nl

DINARD VAN DER LAAN
Faculty of Economics and Business Administration
Department of Econometrics
Vrije Universiteit, 1081 HV Amsterdam, The Netherlands
E-mail: dalaan@feweb.vu.nl

In this article, 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, we show that 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 noncontinuous in the parameters of the system (arrival rate and service rate) and shows a fractal behavior. Moreover, convexity properties are shown and are used in a generalization of the computation of the optimal control policy (in open loop) for the stochastic exponential case.

1. INTRODUCTION

A typical polling system consists of a number N of queues attended by a single server. There exists a huge body of literature on polling systems, starting in 1957

© 2007 Cambridge University Press 0269-9648/07 $25.00

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

[17,18]. Since then, polling systems have been used extensively in communication networks as well as manufacturing systems. In this article, we consider a polling system with N queues with infinite buffers. Time is slotted and the slot durations are either deterministic with unit size or exponentially 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 about 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-conserving. The control of the server is an open-loop control, which means that the decision about which queue is 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 nth element gives the queue that is served during the nth time slot. We derive an efficient algorithm for computing the optimal open-loop polling sequence with the objective being 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 work (see [1,2]). 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 regular sequence, say with density $\hat{d}_1$ (and 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 exponential models. Moreover, the optimal densities $(\hat{d}_1, \hat{d}_2)$ are obtained for both models.

In Sections 2–4 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 a regular service sequence as a 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 formulas for the average workload in the case that 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 optimal policy that is periodic.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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 in open loop is computed in an efficient way.

There is extensive literature on the many variants of controlled polling systems (see [6,22]). In several articles [4,5,15] 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 [2,14], and for the exponential model, see [10]). The results are similar; both cases exhibit a fractal behavior. However, there also exist subtle differences that are pointed out in this article, mainly due to the fact that a polling system has several arrival processes.

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. We assume that the input to the queues is deterministic, but we consider two slightly different models. In the first model the input to queue i, $i = 1, 2, \ldots, N$, is discrete and in fact we assume for $i = 1, 2, \ldots, N$ that at all of the integer times $T = 0, 1, 2, \ldots$, a job with constant workload λ_i arrives in queue i. In the second model we assume that the workload input to queue i, $i = 1, 2, \ldots, N$, is fluid and flows in with constant rate λ_i.

In both models we assume that at all of 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 the 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 $\bar{\lambda} := (\lambda_1, \lambda_2, \ldots, \lambda_N)$. An infinite sequence U corresponding to a polling policy for such a system is a so-called word on the alphabet $\{1, 2, \ldots, N\}$ (see [14,16]). The set of all words on the alphabet $\{1, 2, \ldots, N\}$ (corresponding to all of the possible polling policies for some $\bar{\lambda}$ 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

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

For $U \in \mathcal{A}(N)$ for $i = 1, 2, \ldots, N$, we define 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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

of zeros and ones, corresponding to the server assignment for one queue, a 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 λ_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 $\bar{\lambda}$ 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 $\bar{\lambda}$ system is given by

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

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 all of the subscripts 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 service 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*

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

and for the model with fluid input, we have

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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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

An infinite sequence of zeros and ones $u = (u_1, u_2, \ldots)$ is called regular with density $d \in [0,1]$ (see, e.g., [2] (where it is called balanced), and [3,23,24]) 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*

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

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 the 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, 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 λ and any service sequence u of upper density at most d, we have that $B(u) \ge B(d) = B(\pi(d))$.*

In the remainder 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 λ 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 $d = p/q$, with $p, q \in \mathbb{N}$, $\gcd(p,q) = 1$, then u is periodic with periodic q. For such rational density $d = p/q$, 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 consecutive time intervals of unit length corresponding to a finite sequence u of length k. Suppose that the first of these k inter-

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

vals 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 nonincreasing 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 nonincreasing if and only if $d \geq \lambda$.*

PROOF: We consider the model with discrete input and let $d = p/q$, 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

$$\begin{aligned} V(t_0 + q) - V(t_0) &= \lambda(t_0 + q) + m(t_0 + q) - (\lambda \cdot t_0 + m(t_0)) - \kappa_q(p(d)) \\ &= \lambda \cdot q - p + (m(t_0 + q) - m(t_0)). \end{aligned}$$

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 nonincreasing.

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 nonincreasing 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

$$\begin{aligned} V(t_0 + q) &= V(t_0 + q) - V(t') \\ &= \lambda \cdot (q - t^*) - (\kappa_q(p(d)) - \kappa_{t^*}(p(d))) \\ &= \lambda \cdot (q - t^*) - (p - \lfloor t^* \cdot d \rfloor) \leq d \cdot (q - t^*) - p + t^* \cdot d \\ &= 0. \end{aligned}$$

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

Suppose that $d \in \mathbb{Q}$ and $d \geq \lambda$. Then it follows from Lemma 3.4 that if the workload is zero at the beginning of an $p(d)$ factor, then it is also zero at the end of the factor and, thus, at the beginning of the next factor. So, 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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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 [14] (see also [13]).

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

In this subsection we use several notions defined in [14] and we summarize results that can be obtained analogously to results in [14]. We also recall that if d_1 and d_2 are rational numbers with $0 \leq d_1 \leq d_2 \leq 1$, $d_i = p_i/q_i$, and $\gcd(p_i, q_i) = 1$ for $i = 1, 2$, 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 $d_0 = (p_1 + p_2)/(q_1 + q_2)$. 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 denominators greater than or equal to $q_1 + q_2$.

The following result for the factorization of period words of lower-bracket sequences is proved in [14].

LEMMA 3.7 (Lemma 4.3 in [14]): *Let $I = [d_1, d_2]$ be a Farey interval and $d_0 = (p_1 + p_2)/(q_1 + q_2)$ as above. Then $p(d_0) = p(d_1)p(d_2)$.*

Using Lemma 3.7 we can show the following result using classical properties of lower-bracket sequences.

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 $w(d)$.*

Using this, one can further characterize the behavior of the function B.

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 λ is the input rate. Let $d \in I$, $d = \mu \cdot d_1 + (1 - \mu) \cdot d_2$, where $\mu \in [0, 1]$. Then $B(d) = \mu \cdot B(d_1) + (1 - \mu) \cdot B(d_2)$.*

PROOF (Sketch): Theorem 3.9 can be proved analogously to Theorem 5.9 in [14] 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 μ is the fraction (of time) taken by $p(d_1)$ factors and $1 - \mu$ is the fraction taken by $p(d_2)$ factors. Moreover, the workload is zero after every such factor. ■

According to Theorem 3.9, the function $B(d)$ is, in addition to being convex, also piecewise linear. More precisely, the theorem states 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 λ.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

3.2. Best Upper Approximations

DEFINITION 3.10: *Let $0 < x \leq 1$ be a given real number and let $s = p/q$, 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 $d = p/q$ be a best upper approximation of λ, 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 $|p(d)| = q$. Let $J(d) := \max\{t \in [0,q] : m(t) = 0\}$. For discrete input we have that*

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

For fluid input we have that

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

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 $\lambda \leq \lfloor dt \rfloor / t \leq d$. Since the rational number $\lfloor dt \rfloor / t$ has denominator $t < q$, this contradicts the fact that $d = p/q$ is a best upper approximation of λ. 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

$$\begin{aligned} V(q-1) &= \lambda(q-1) - \kappa_{p(d)}(q-1) \\ &= \lambda(q-1) - \left\lfloor \frac{p}{q}(q-1) \right\rfloor \\ &= \lambda(q-1) - (p-1). \end{aligned}$$

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 $J(d) - (q-1) = V(q-1)/(1-\lambda) = (\lambda(q-1) - (p-1))/(1-\lambda) \geq 0$. ■

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

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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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 $B_\lambda(d) = (1/q)\int_{t=0}^{q} V(t)\,dt$. 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 $A = \sum_{n=1}^{q-1}\int_{t=n-1}^{n}(\lambda\lceil t\rceil)\,dt = \lambda\cdot\sum_{n=1}^{q-1} n = (\lambda/2)(q-1)q$, and by Theorem 100 in [11],

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

$$= \sum_{n=0}^{q-1}\left\lfloor n\frac{p}{q}\right\rfloor - \left\lfloor (q-1)\frac{p}{q}\right\rfloor$$

$$= \frac{1}{2}(p-1)(q-1) - (p-1)$$

$$= \frac{1}{2}(p-1)(q-3).$$

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

$$C = \sum_{n=1}^{q-1}\int_{t=n-1}^{n} p(d)_{\lfloor t\rfloor+1}(t-\lfloor t\rfloor))\,dt$$

$$= \kappa_{p(d)}(q-1)\int_{t=0}^{1} t\,dt = \frac{1}{2}(p-1)$$

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

and

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

Hence,

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

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

$$\begin{aligned} D^* &= \int_{t=q-1}^{J(d)} (\lambda \cdot (q-1) + q - p - (1-\lambda)t)\, dt \\ &= \int_{t=q-1}^{q-1+\frac{\lambda(q-1)-(p-1)}{1-\lambda}} (\lambda \cdot (q-1) + q - p - (1-\lambda)t)\, dt \\ &= \int_{t=0}^{\frac{\lambda(q-1)-(p-1)}{1-\lambda}} (\lambda(q-1) - (p-1) - (1-\lambda)t)\, dt \\ &= \frac{(\lambda(q-1) - (p-1))^2}{2(1-\lambda)}. \end{aligned}$$

Hence,

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

■

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

In addition to this closed formula for the value of $B(d)$ for all of 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 $B_\lambda(\lambda) = (\lambda + 1)/2$ if λ is irrational and $B_\lambda(\lambda) = (p + q - 1)/2q$ if $\lambda = p/q$, with $p, q \in \mathbb{N}$ and $\gcd(p,q) = 1$. For the model with fluid input, we have that $B_\lambda(\lambda) = \frac{1}{2}$ in the case when λ is irrational and $B_\lambda(\lambda) = \frac{1}{2} - (1/2q)$ if $\lambda = p/q$, with $p, q \in \mathbb{N}$, $\gcd(p,q) = 1$.*

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

In the sequel of this proof we suppose that λ 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 λ is irrational and, thus, λt 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 $A1 = \limsup_{T\to\infty}(1/T)\int_{t=0}^{T}(\lambda\lceil t\rceil - \lfloor \lambda\lfloor t\rfloor\rfloor)\, dt$ and $A2 = \lim_{T\to\infty}(1/T)$ $\int_{t=0}^{T}(\pi(\lambda)_{\lfloor t\rfloor+1}(t - \lfloor t\rfloor))\, dt$. By the ergodic theorem of Weyl and Von Neumann (see, e.g., [21]) we have

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

Moreover,

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

Hence, $B_\lambda(\lambda) = \lambda + \frac{1}{2} - (\lambda/2) = (\lambda + 1)/2$ if λ is irrational. For the model with fluid input and irrational λ, we have analogously to the model with discrete input

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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, $B_\lambda(\lambda) = A3 - A2 = A3 - (\lambda/2)$, where

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

We have

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

from which $B_\lambda(\lambda) = \frac{1}{2}$. ■

For given input rate $0 < \lambda \leq 1$ we have that $s_1 = 1$ is the first best upper approximation of λ. Consider the following iterative construction of best upper approximations of λ. Put $i := 1$ and do the following iteratively. If $s_i = \lambda$, then stop; otherwise 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 [14] 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 λ is rational, then there exists some $k \in \mathbb{N}$ such that*

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

If λ is irrational, then

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

If s_i, s_{i+1} are consecutive best upper approximations of λ, 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 and s_{i+1} are consecutive best upper approximations of λ. In fact, it turns out that the slope of the function $B_\lambda(d)$ changes precisely at all of the best upper approximations of λ. This implies that the exact value of $B(d) = B_\lambda(d)$ is easily computed if we can find the consecutive best

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

upper approximations s_i, s_{i+1} of λ 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 $d = \sqrt{2}/2$. It is easily seen that the best upper approximations of $\lambda = \frac{12}{17}$ are consecutively $s_1 = \frac{1}{1}$, $s_2 = \frac{3}{4}$, $s_3 = \frac{5}{7}$, and $s_4 = \frac{12}{17}$ 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 $B(\frac{12}{17}) = (12 + 17 - 1)/2 \cdot 17 = \frac{14}{17}$, and by Theorem 3.12, we have that

$$B\left(\frac{5}{7}\right) = \frac{\frac{12}{17}7^2 + \frac{12}{17}7 - 5\cdot 7 + 7 - 1 + \left(\frac{12}{17}\right)^2 7^2 - 2\frac{12}{17}5\cdot 7 + 5^2}{2\cdot 7} = \frac{1522}{2023}.$$

Putting $\mu = (\frac{5}{7} - d)/(\frac{5}{7} - \frac{12}{17}) = 85 - \frac{119}{2}\sqrt{2}$, we have $\mu \in [0,1]$ and $d = \mu\frac{12}{17} + (1-\mu)\frac{5}{7}$. Hence, by Theorem 3.9 we have

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

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 λ such that $d \in [s_{i+1}, s_i]$ in an efficient way. Following the arguments in [14] (where the same problem appears with the only difference being that best lower approximations have to be found instead of best upper approximations), it can be shown that this problem can be efficiently solved by using the continued fraction expansion of λ.

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

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

$$a_0 = \lfloor \alpha \rfloor; \quad \alpha_1 = \frac{1}{\alpha - a_0}$$

$$a_n = \lfloor \alpha_n \rfloor; \quad \alpha_{n+1} = \frac{1}{\alpha_n - a_n} \quad \text{for } n = 1, 2, \ldots.$$

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

Then

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

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 α 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 α is irrational, then the continued fraction expansion of α is infinite.

We define p_n and q_n recursively by

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

Then $x_n := p_n/q_n = [a_0, a_1, \ldots, a_n]$ is called the nth 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 α, we also define the so-called intermediate convergents.

DEFINITION 3.15: *Let $\alpha = [a_0, a_1, \ldots]$ if α is irrational or $\alpha = [a_0, a_1, \ldots, a_m]$ for some $m \in \mathbb{N}$ if α is rational. Then a rational number p/q is an intermediate convergent of α 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 α is rational) and $c \in \{1, 2, \ldots, a_n - 1\}$. Moreover, p/q is called an odd (even) intermediate convergent if n is odd (even).*

In [14] it is stated that all the best lower approximations of some positive real number α are either even convergents of α, even intermediate convergents of α or α itself in case α is an odd convergent of itself. Analogously we have that all the best upper approximations of some positive real number α are either odd convergents of α, odd intermediate convergents of α or α itself in case α is an even convergent of itself. Thus to obtain best lower approximations, a method which is similar to the one used in [14] to obtain best upper approximations can be used, 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.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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 $p_1/q_1, p_2/q_2, \ldots$ of λ until we have found an odd convergent p_{2n+1}/q_{2n+1} $(n \geq 0)$ that is smaller or equal than d or we have that $\lambda = p_N/q_N \leq d < p_{N-1}/q_{N-1}$ for some even positive integer N. If we have the latter case, then we put $s_i = p_{N-1}/q_{N-1}$ and $s_{i+1} = p_N/q_N = \lambda$. 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 $d = s_1 = p_1/q_1 = 1$ and we go to Step 2. If $n > 0$, then we have that $p_{2n+1}/q_{2n+1} \leq d < p_{2n-1}/q_{2n-1}$ 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 $k = \lceil (p_{2n-1} - dq_{2n-1})/(dq_{2n} - p_{2n}) \rceil - 1$. By putting $s_i = (k\cdot p_{2n} + p_{2n-1})/(k\cdot q_{2n} + q_{2n-1})$ and $s_{i+1} = ((k+1)\cdot p_{2n} + p_{2n-1})/((k+1)\cdot q_{2n} + q_{2n-1})$, 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 $B_\lambda(d) = \lambda^2/2$ in the case of discrete input and $B_\lambda(d) = 0$ in the 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 and s_{i+1} of λ such that $d \in [s_{i+1}, s_i]$. Next compute by the formulas 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. Otherwise we go to Step 3.

Step 3. We have that $d \in (s_{i+1}, s_i)$. Put $\mu = (s_i - d)/(s_i - s_{i+1})$. 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 $\lambda = 1/\pi$. 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 λ and we consecutively find $a_0 = 0$, $p_0/q_0 = \frac{0}{1}$, $a_1 = 3$, $p_1/q_1 = \frac{1}{3}$, $a_2 = 7$, $p_2/q_2 = \frac{7}{22}$, $a_3 = 15$, $p_3/q_3 = \frac{106}{333}$, $a_4 = 1$, $p_4/q_4 = \frac{113}{355}$, $a_5 = 292$, and $p_5/q_5 = \frac{33{,}102}{103{,}993}$. We now have that $\lambda < p_5/q_5 \leq d$ and we stop applying the continued fraction algorithm. Next we compute that $k = \lceil (p_3 - d\cdot q_3)/(d\cdot q_4 - p_4) \rceil - 1 = 55$. So, it follows that $s_i := (k\cdot p_4 + p_3)/(k\cdot q_4 + q_3) = \frac{6321}{19{,}858}$ and $s_{i+1} := ((k+1)\cdot p_4 + p_3)/((k+1)\cdot q_4 + q_3) = \frac{6434}{20{,}213}$ are consecutive best upper approximations of λ 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 $B_\lambda(s_{i+1}) = B_{1/\pi}(\frac{6434}{20{,}213} = (278{,}494{,}715 - 88{,}633{,}874\pi)/40{,}426(\pi - 1)$ and $B_\lambda(s_i) = B_{1/\pi}(\frac{6321}{19{,}858}) = (268{,}797{,}889 - 85{,}547{,}520\pi)/39{,}716(\pi - 1)$. We have that $d \in (s_{i+1}, s_i)$ and go to Step 3 of the algorithm. Putting $\mu = (s_i - d)/(s_i - s_{i+1}) = \frac{20{,}213}{50{,}000}$, we obtain

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

$$\begin{aligned} B_\lambda(d) &= B_{1/\pi}(0.31831) \\ &= \mu B_\lambda(s_{i+1}) + (1-\mu)B_\lambda(s_i) \\ &= \frac{1,363,383,097 - 433,910,308\pi}{200,000(\pi - 1)} \\ &\sim 0.4988368585346. \end{aligned}$$

Remark: 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 λ, 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 λ.

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 well 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 λ_i is the (expected) workload arriving in queue i per time unit. We recall that for the deterministic polling systems, the parameter λ_i was defined 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 λ_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, whereas each job arriving in queue i brings a workload that is exponentially distributed with parameter μ_i. Hence, for the exponential model we have that $\lambda_i = 1/\mu_i$ for $i = 1, 2, \ldots, N$.

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

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

Thus, $B_{\bar{\lambda}}(\bar{d})$ is the average workload in a $\bar{\lambda}$ system (the total over all queues) if the service sequence for queue i is 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. However, as we show next, it provides a lower bound. Indeed, let U be a polling policy applied in a $\bar{\lambda}$ 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 $\bar{d}$ is restricted to the compact and convex set

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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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,

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

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

From Appendix A, we have that this lower bound $B_{\bar{\lambda}}(\bar{d})$ is convex in the vector of polling densities $\bar{d} = (d_1, d_2, \ldots, d_N)$. So, $B_{\bar{\lambda}}(\bar{d})$ 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 a $\bar{\lambda}$ system,

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

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

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

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

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

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

Suppose that we have a $\bar{\lambda}$ system for which $\sum_{i=1}^{N} \lambda_i \geq 1$ in the case of exponential queues or $\sum_{i=1}^{N} \lambda_i > 1$ in the 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 $B^*(\bar{\lambda}) = \infty$ and, thus, $B(U) = \infty$ for every polling policy U. Thus, such systems are unstable and optimal policies do not exist. We say that polling systems for which $\sum_{i=1}^{N} \lambda_i < 1$ in the case of exponential queues or $\sum_{i=1}^{N} \lambda_i \leq 1$ in the case of deterministic queues are stable. For stable systems, it follows directly from the results on polling to one queue that $B^*(\bar{\lambda}) < \infty$. 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 $B_{\bar{\lambda}}(\bar{d})$ over the set H^N and finding the minimum value $B^*(\bar{\lambda})$ 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 $B_{\bar{\lambda}}(\bar{d})$ for a particular $\bar{d}$. A dual problem is considered for a routing system with one arrival process and N parallel queues in [9,10,14].

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

4.1. Optimality Results for Two Queues

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

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

we have by Proposition 4.1 that $\text{opt}(\bar{\lambda})$ is a nonempty closed subinterval of $[0,1]$. Moreover, given $d \in \text{opt}(\bar{\lambda})$, 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,

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

Hence, U is optimal according to Proposition 4.1. Thus, in the 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 (λ_1, λ_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 $B(U) = B^*(\bar{\lambda})$.*

The structural result of Proposition 4.2 on optimal policies holds for arrival-driven polling systems with general independent and identically distributed (i.i.d.) input and independent i.i.d. service times. This follows from Appendix A analogously to the reasoning we used to obtain Proposition 4.2. Moreover, a result similar to 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 articles. In [10] this problem is considered for Poisson arrivals and both servers have exponential service times. In [9,14] the problem is considered for a deterministic system with constant interarrival times and constant service times for both servers. Methods similar to those considered in these articles can also be applied for corresponding polling systems 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 (λ_1, λ_2) polling systems where we fix the ratio $\lambda_1/(\lambda_1 + \lambda_2)$ to be equal to 0.37 by putting $\lambda_1 = 0.37\rho$ and $\lambda_2 = 0.73\rho$, where the load ρ is varied from zero to one. For this family of polling systems we computed the value of the optimal polling density α_{opt} for both the discrete and the fluid case. For $0 < \rho < 1$ we define α_{opt} to be the rational

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

number of lowest denominator that is contained in the nonempty closed subinterval $\mathrm{opt}(\lambda_1, \lambda_2)$ of $(0,1)$. We note that in almost all cases for varying (λ_1, λ_2), the set $\mathrm{opt}(\lambda_1, \lambda_2)$ consists of only one (rational) point, which by the above definition is the point α_{opt}. Moreover, in the few cases in which $\mathrm{opt}(\lambda_1, \lambda_2)$ consists of more than one point, it still holds by Lemma 4.6.4 in [24] that α_{opt} is unique and thus well defined. In the case $\rho = 1$ it is easily seen that the interval of stability consists of the single point $\lambda_1/(\lambda_1 + \lambda_2)$, 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 α_{opt} has to be 0.37 for both the discrete and the fluid cases. For both the discrete and the fluid cases we have computed the value of α_{opt} for varying ρ 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 α_{opt} is plotted for varying loads for the discrete case, whereas 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 α_{opt} is always equal to 0.5, which corresponds to the round-robin polling policy. So, in Figures 1 and 2 we have restricted to an interval with high loads ρ, since this is, by far, the most interesting part of the interval.

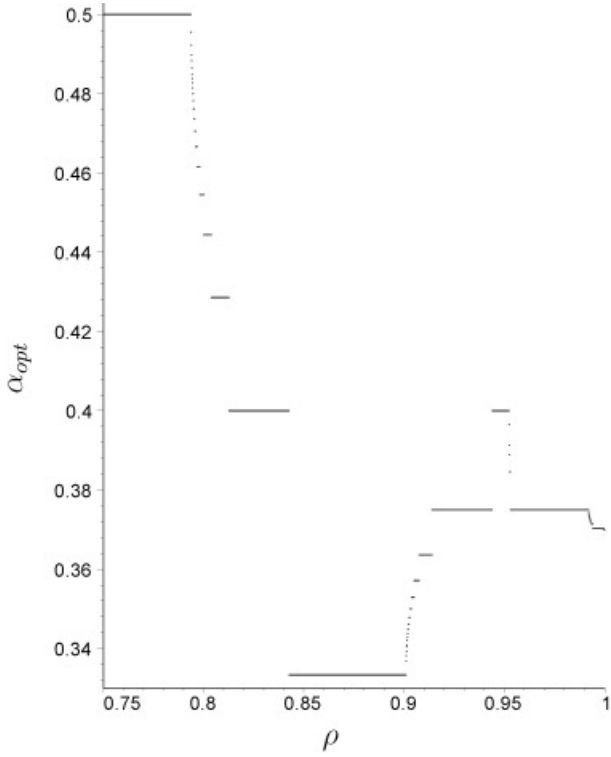


FIGURE 1. The optimal polling density for varying loads for discrete input.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

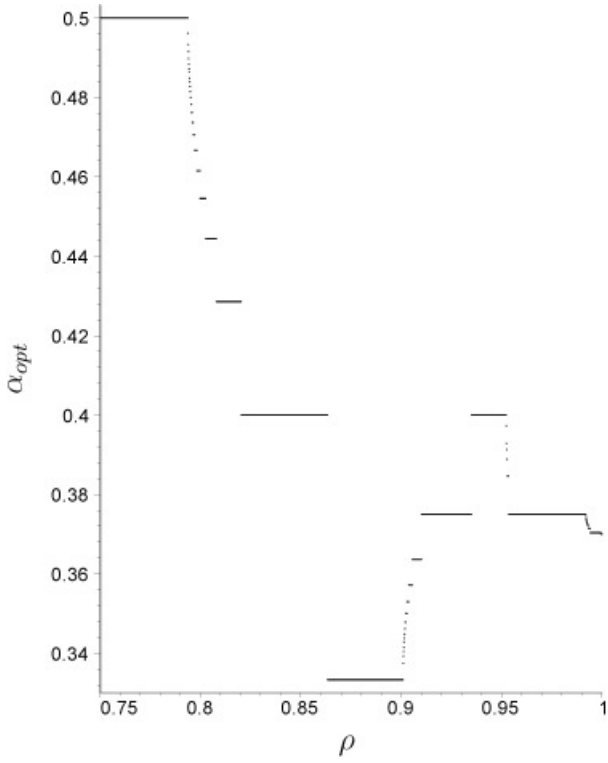


FIGURE 2. The optimal polling density for varying loads for fluid input.

Remark. We see in Figures 1 and 2 that for this example where $\lambda_1/(\lambda_1 + \lambda_2) = 0.37$ the optimal polling density α 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 ρ, the optimal ratio α is greater than 0.37, which means that the queue with the smaller arriving workload is served relatively often for small loads. In fact, $\alpha = \frac{1}{2}$, which corresponds to the round-robin policy, is optimal for quite a large part of the interval. We note also that α does not decrease monotonically from $\frac{1}{2}$ for a small load ρ 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 $\frac{1}{2}$ no longer gives a stable policy for such loads. If the load is increased even further, then also $\frac{1}{3}$ no longer gives a stable policy until, finally, for $\rho = 1$, we have that only $\alpha = 0.37$ gives a stable policy.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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

Another point of interest that we notice in Figures 1 and 2 is that for any given load, the α for the fluid model is equal or larger than the α for the discrete model. Intuitively this is explained from the fact that for the discrete model, the (whole) packets of workload enter the buffer at once, whereas 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 Figs. 3 and 4), in which the input rates for the two queues vary independently of each other. To plot Figures 3 and 4 we have calculated the optimal polling density α for various (λ_1, λ_2) systems in the triangular area $\{(x_1, x_2) : x_1, x_2 \geq 0, x_1 + x_2 \leq 1\}$. In Figures 3 and 4 there is a black dot at a point (λ_1, λ_2) within these triangular areas if one or more of the neighboring points has another value of α.

The resulting Figures 3 and 4 look quite similar to fractal type pictures. It is obvious that both pictures are symmetrical in the diagonal $\lambda_1 = \lambda_2$. In both figures

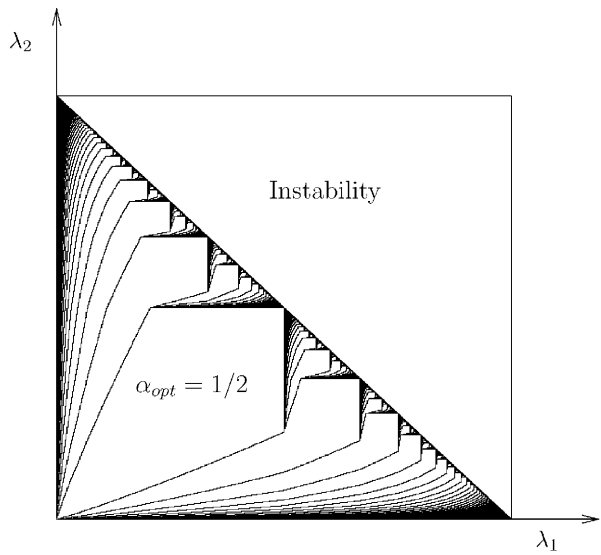


FIGURE 3. The regions of optimality for discrete input.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

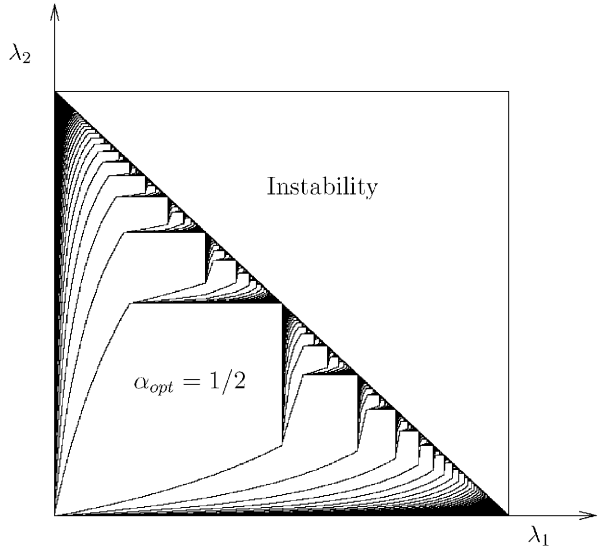


FIGURE 4. The regions of optimality for fluid input.

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 α with a low denominator are somewhat larger in Figure 4 (the fluid model) than in Figure 3 (the discrete model). This supports our conjecture that for the discrete model, the value of $|\alpha - \frac{1}{2}|$ is always greater than or equal to that of 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 (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 λ, which are rational numbers.
- $\lim_{\varepsilon \downarrow 0}[(B_\lambda(\lambda) - B_\lambda(\lambda + \varepsilon))/\varepsilon] = \infty$ if λ is irrational.

From these properties the following theorem follows.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

THEOREM 4.3: *Consider a deterministic* $\bar{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ *polling system with* $N \geq 2$ *queues such that* $\sum_{i=1}^{N} \lambda_i < 1$. *Then there exists some* $x = (x_1, x_2, \ldots, x_N) \in \text{opt}(\bar{\lambda})$ *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* λ_i.

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

Remark 4.5: In Section 7 of [14] 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 $\text{opt}(\bar{\lambda})$ for deterministic polling systems. Moreover, we note that in most cases the set $\text{opt}(\bar{\lambda})$ 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 two 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* (λ_1, λ_2) *polling system with* $\lambda_1 + \lambda_2 < 1$ *there exists an optimal regular and periodic polling policy* U *with* $B(U) = B^*(\bar{\lambda})$.

We note that a similar result on the optimality of periodic policies for deterministic parallel routing systems with only two queues was obtained in [9]. However, for deterministic $\bar{\lambda}$ polling systems with more than two queues we do not have such a result. Indeed, in the case for a rational point $(d_1, d_2, \ldots, d_N) \in \text{opt}(\bar{\lambda})$, there does not, in general, 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 existed, then U would be periodic and also optimal, since $B(U) = B^*(\bar{\lambda})$. However, if such a policy does not exist, it is possible that for an optimal policy V, it holds that $B(V) > B^*(\bar{\lambda})$. Moreover, such optimal policy V does not necessarily have rational densities $(d_1, d_2, \ldots, d_N) \in \text{opt}(\bar{\lambda})$ and, thus, it might not be periodic. Nevertheless, we think that in most cases there also exists an optimal periodic policy for deterministic polling systems with more than two queues. Additionally, the following can be proved analogously to the similar result for deterministic parallel routing systems considered in Chapter 5 of [24].

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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

Acknowledgment

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

References

1. Altman, E., Gaujal, B., & Hordijk, A. (2000). Optimal open-loop control of vacations, polling and service assignment. *Queueing Systems* 36: 303–325.
2. Altman, E., Gaujal, B., & Hordijk, A. (2003). *Discrete-event control of stochastic networks: Multimodularity and regularity.* New York: Springer-Verlag.
3. Arian, Y. & Levy, Y. (1992). Algorithms for generalized round robin routing. *Operations Research Letters* 12: 313–319.
4. Borst, S. (1994). Polling systems. PhD thesis, Katholieke Universiteit Brabant, Tilburg.
5. Boxma, O.J., Levy, H., & Weststrate, J.A. (1991). Efficient vist frequencies for polling tables: Minimization of waiting cost. *Queueing Systems* 9: 133–162.
6. Boxma, O.J. & Tagaki, H. (eds.) (1992). *Queueing Systems* [Special issue on polling systems].
7. Coffman, E.G., Puhalskii, A.A., & Reiman, M.I. (1995). Polling systems with zero switchover times: A heavy-traffic averaging principle. *Annals of Applied Probability* 5: 681–719.
8. Combé, M.B. & Boxma, O.J. (1994). Optimization of static traffic allocation policies. *Theoretical Computer Science* 125: 17–43.
9. Gaujal, B. & Hyon, E. (2000). Optimal routing policy in two deterministic queues. *Calculateurs Parallèles* 13: 601–634.
10. Gaujal, B., Hyon, E., & Jean-Marie, A. (2004). Optimal open-loop routing in two parallel queues with exponential service times. In IEEE, *Wodes*, Reims, 2004. Long version available from http://www.inria.fr/rrrt/rr-5109.html.
11. Hardy, G.H. & Wright, E.M. (1960). *An introduction to the theory of numbers*, 4th ed. Oxford: Oxford University Press.
12. Hordijk, A. & Loeve, J.A. (1997). Optimal noncyclic server allocation in a polling model. In IEEE, *Conference on Decision and Control*, vol. 36, pp. 2941–2945.
13. Hordijk, A. & van der Laan, D.A. (2003). Note on the convexity of the stationary waiting time as a function of the density. *Probability in the Engineering and Information Sciences* 17: 503–508.
14. Hordijk, A. & van der Laan, D.A. (2005). On the average waiting time for regular routing to deterministic queues. *Mathematics of Operations Research* 30: 521–544.
15. Kruskal, J.B. (1969). Work-scheduling algorithms: A non-probabilistic queueing study. *Bell System Technical Journal* 48: 2963–2974.
16. Lothaire, M. (2002). *Algebraic combinatorics on words*. Cambridge: Cambridge University Press.
17. Mack, C. (1957). The efficiency of *n* machines uni-directionally controlled by one operative when walking time and repair times are variable. *Journal of the Royal Statistical Society Press, Series B* 19(1): 173–178.
18. Mack, C., Murphy, T., & Webb, N.L. (1957). The efficiency of *n* machines uni-directionally controlled by one operative when walking time and repair times are constants. *Journal of the Royal Statistical Society Press, Series B* 19(1): 166–172.
19. Neuts, M.F. (1989). *Structured stochastic matrices of M/G/1 type and their applications*. New York: Marcel Dekker.
20. Perron, O. (1954). *Die Lehre von den Kettenbrüchen*. Stuttgart: Teubner.
21. Sinai, Y.G. (1976). *Introduction to ergodic theory*. Princeton, NJ: Princeton University Press.
22. Tagaki, H. (1986). *Analysis of polling systems*. Cambridge, MA: MIT Press.
23. Tijdeman, R. (2000). Fraenkel's conjecture for six sequences. *Discrete Mathematics* 222: 223–234.
24. van der Laan, D.A. (2003). The structure and performance of optimal routing sequences. PhD thesis, Leiden University.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

APPENDIX A
On the Optimality of the Bracket Sequence and Its Convexity for the Arrival-Driven Polling Model

Consider an arrival-driven polling model with i.i.d. service and independent i.i.d. interarrival times in every queue. Note that both the deterministic and exponential models that we analyze in this article can be described as such an arrival-driven model with i.i.d. service and independent i.i.d. interarrival times. Moreover, if for queue i in a polling system with these assumptions we have that τ is the mean interarrival time and $1/\mu_i$ is the mean service time of a job arriving in queue i, then λ_i, the (expected) workload arriving in queue i per time unit as described in the beginning of Section 4, is given by $\lambda_i = 1/\tau\mu_i$.

For one queue, Lemma 54 in [2] gives for service sequence a that $V_n(a)$, the expected workload in the queue for the nth 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 [13] can be applied. Hence, the average workload for the bracket sequence with initial phase θ is independent of θ and it is convex in the density p, since

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} V_n(a_1^p(\theta), \ldots, a_n^p(\theta)) = \lim_{n\to\infty} \int_0^1 V_n(a_1^p(\theta), \ldots, a_n^p(\theta))\, d\theta. \tag{A1}$$

For queue i, let τ_n^i denote the nth interarrival time and σ_n^i denote the nth service time. Let us further assume that the interarrival (resp. service) time is almost surely bounded from below (resp. above). Hence, 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 θ^i is optimal in the class of policies with upper density smaller than or equal to p^i. For service sequences $a = (a^1, a^2)$ with densities for queue 1 (resp. 2) equal to p^1 (resp. $p^2 = 1 - p^1$),

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

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 interarrival 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.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

THEOREM A1: *The optimal polling policy for $N = 2$ queues can be obtained by minimizing the convex function $V_\infty(p^1, p^2)$.*

APPENDIX B
Computation of the Optimal Policy in the Markovian Case

Here we show how the optimal polling sequence can be computed in a stochastic system consisting of two Poisson arrival queues and one exponential server. As in Appendix A, we consider an arrival-driven polling model with i.i.d. service and independent i.i.d. interarrival times in every queue. We show how the optimal polling between two queues can be computed when the service is exponential (with rate μ_i in queue i) and the interarrivals are synchronous and exponential (with rate λ in both queues). Other cases can be treated analogously.

We call $m(k)$ the kth 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 ℓ, 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 whose 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 and the second entry represents the current letter of the polling m (modulo ℓ).

The continuous-time Markov chain X_t is a quasi-birth-and-death process with generator Q 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 $A_0[i, (i+1) \bmod \ell] = \lambda$, 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 B1.

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

$$\pi Q = 0. \tag{B1}$$

We now refine the notation by introducing block vectors π_n of dimension ℓ whose ith 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 of having k customers in the system. We will not try to compute π directly, which can be quite hard; instead, we will determine its generating function. Let $\bar{D}(0,1)$ be the closed-unit disk. The generating function of π is the function $\Pi(z)$ from $\bar{D}(0,1)$ to $\mathbb{C}^\ell$ defined by

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

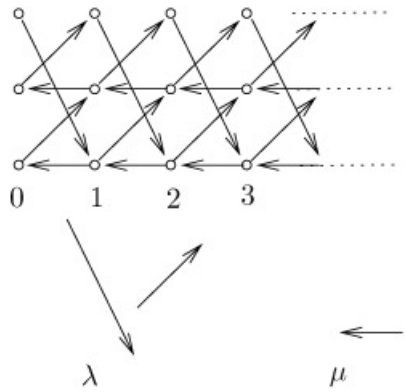


FIGURE B.1. Graph of the infinitesimal generator when the polling is $m = (110)^{\infty}$.

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

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

LEMMA B1: *The process X_t is positive recurrent if and only if $\ell\lambda/a\mu < 1$.*

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 $p_1 = p_2 = \cdots = p_\ell = 1/\ell$. Hence, the stability condition becomes $pA_2\mathbf{1} = a\mu/\ell > pA_0\mathbf{1} = \lambda$. ■

LEMMA B2: *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 $K(z)$:

$$\Pi(z)K(z) = \pi_0\mu(1-z)M. \tag{B2}$$

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

PROOF: Using the global balance equation (B1) we get the induction

$$\pi_0 C + \pi_1 A_2 = 0, \tag{B3}$$

$$\pi_{n-1} A_0 + \pi_n A_1 + \pi_{n+1} A_2 = 0, \quad \forall n \geq 1. \tag{B4}$$

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

$$\sum_{n=1}^{\infty} \pi_{n-1} z^{n-1} z^2 A_0 + \pi_n z^n z A_1 + z^{n+1} \pi_{n+1} A_2 = 0,$$

$$\Pi(z)(z^2 A_0 + z A_1 + A_2) - \pi_1 A_2 z - \pi_0 A_2 - \pi_0 A_1 z = 0,$$

which gives $\Pi(z)(z^2 A_0 + z A_1 + A_2) = \pi_0 \mu (1 - z) M$, 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 of 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, after direct computations one gets

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

LEMMA B3: *If $\ell\lambda/a\mu < 1$, then the number of nonnull 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 ℓ 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 ε is small enough. Then, the result follows from Rouché's theorem. ■

To illustrate the previous lemma, Figure B2 displays all of 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 five.

THEOREM B4: *If z_i is the ith nonnull root of $\Delta(z)$ in the unit disk and v_i is the right eigenvector of the eigenvalue 0 of $K(z_i)$, then π_0 is a solution of the system:*

$$\begin{aligned} \pi_0(j) &= 0 && \text{if } m(j) = 0 \\ \pi_0 v_i &= 0 && \forall i \in \{1, \ldots, a\} \text{ s.t. } z_i \neq 1 \\ \pi_0 \mathbf{1} &= a/\ell - \lambda/\mu && \text{when } z_i = 1, \end{aligned} \tag{B5}$$

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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

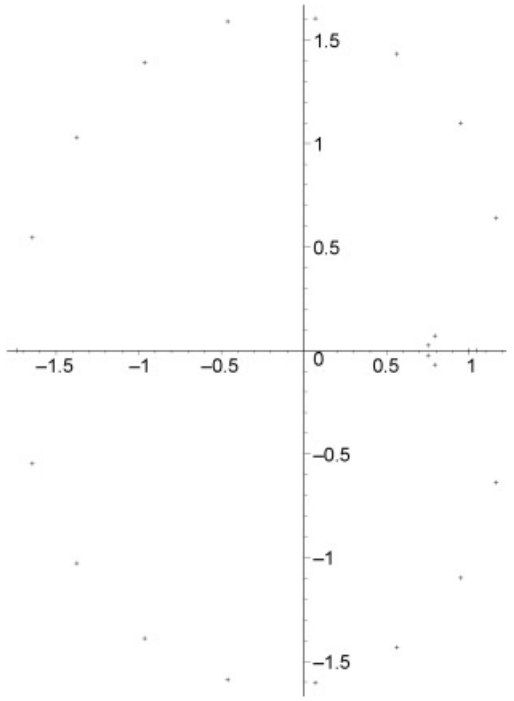


FIGURE B.2. Roots of $\Delta(z)$ when $a = 5, \ell = 18$ and $\lambda = 1, \mu = 4$.

PROOF: If $|z_i| < 1$, then it follows 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 $M_{jj} = 0$, it holds that $\pi_0 v_i = 0$. Note that the rank of the matrix $K(z_i)$ is $\ell - 1$; thus, 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 $K(1)\mathbf{1} = 0$.

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

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

Now we turn back to the two-queue system. Once π_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 $\mathbb{E}(N_j(m)) = d\Pi(z)\mathbf{1}/dz|_{z=1}$, introducing the vector $\hat{k}(z)$, which verifies $K(z)\hat{k}(z) = \mathbf{1}$, yields

$$\mathbb{E}(N_j(m)) = \frac{d}{dz}\left(\Pi(z)\mathbf{1}\right)|_{z=1} = \mu_j \pi_0 M_j \frac{d}{dz}\left((1-z)\hat{k}(z)\right)|_{z=1}. \tag{B6}$$

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

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

$$\mathbb{E}(V_j(m)) = \frac{1}{\mu_j}\mathbb{E}(N_j(m)). \tag{B7}$$

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

$$\alpha_{\text{opt}} = \text{argmin}_d(\mathbb{E}(V_1(w(d))) + \mathbb{E}(V_2(\overline{w(d)}))).$$

Using the fact that this function is convex in the density d (see Appendix A), the optimal density α_{opt} can be computed numerically by gradient descent in a fashion similar to that in [8,10]. We have run several computations using a Maple program implementing the algorithm above (For each triple (λ, μ_1, μ_2), compute the roots of $\Delta(z)$, then compute the corresponding eigenvectors, then compute π_0, and, finally, compute the corresponding workload.)

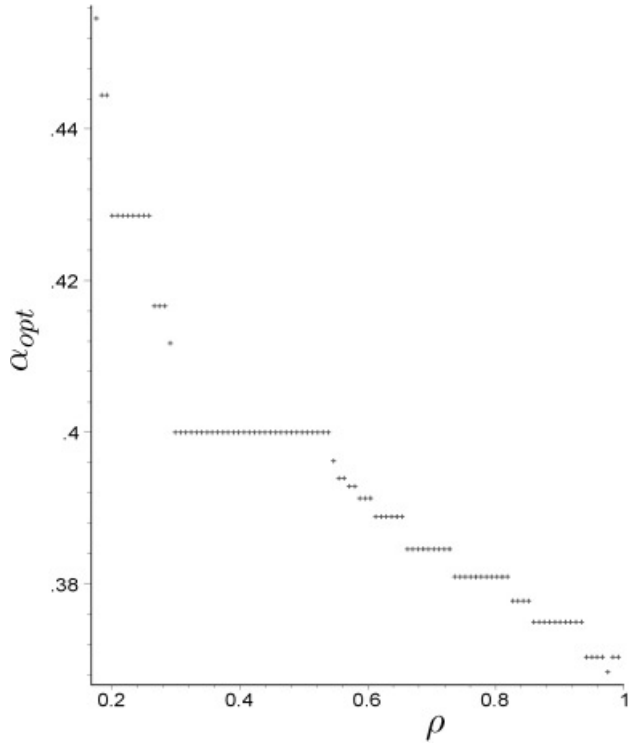


FIGURE B.3. Computation of the optimal polling ratio α_{opt} when the load ρ varies from 0.2 to 1.

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press

Figure B3 shows the values of α_{opt}, the optimal frequency with which the first queue is served (as in the deterministic case). We have chosen $\lambda = 1$ (in both queues) and $\mu_1/(\mu_2 + \mu_1) = \frac{37}{100}$ 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 α_{opt}, the numerical experiments in Figure B3 still suggest that α_{opt} is highly nondifferentiable with many flat zones and many cusps. Moreover, in this case it remains an open problem to (dis)-prove the continuity of α_{opt} with respect to varying load ρ.

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

https://doi.org/10.1017/S0269964807070118 Published online by Cambridge University Press