EURANDOM PREPRINT SERIES
2010-050

Fairness and Efficiency for Polling Models with the κ Gated Service Discipline.

A.C.C. van Wijk, I.J.B.F. Adan, O.J. Boxma and A. Wierman
ISSN 1389-2355

Fairness and Efficiency for Polling Models with the κ-Gated Service Discipline

A.C.C. van Wijk[1], I.J.B.F. Adan*[1], O.J. Boxma[1], and A. Wierman[2]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands

[2]California Institute of Technology, Pasadena, CA, USA

December 17, 2010

Abstract

We study a polling model where we want to achieve a balance between the fairness of the waiting times and the efficiency of the system. For this purpose, we introduce the κ-gated service discipline. It is a hybrid of the classical gated and exhausted disciplines, and consists of using κ_i gated service phases at queue i before the server switches to the next queue. We derive the distributions and means of the waiting times, a pseudo conservation law for the weighted sum of the mean waiting times, and the fluid limits of the waiting times. Our goal is to optimize the κ_i's so as to minimize the differences in the mean waiting times, i.e. to achieve maximal fairness, without giving up too much on the efficiency of the system. From the fluid limits we derive a heuristic rule for setting the κ_i's. In a numerical study the heuristic is shown to perform well.

Keywords: polling model, waiting times, fairness, efficiency, gated service discipline, exhaustive service discipline, optimization.

1 Introduction

Polling models are used in the modeling of many problems, for example computer systems, maintenance systems and telecommunication. In these models, multiple queues are served by a single

* *Corresponding author:* P.O. Box 513, 5600MB Eindhoven, The Netherlands, iadan@win.tue.nl

server, which cyclically visits the queues. A typical performance measure in such systems is the mean waiting time at each of the queues. In certain applications (see e.g. [13, 17]) it is important to maintain *fairness*, in the sense of the queues having (almost) equal mean waiting times. In achieving this, one usually has to sacrifice the efficiency of the system. In this paper, however, we introduce a strategy which on the one hand achieves fairness, while on the other hand still is efficient. Here the efficiency is given by the sum of the mean waiting times, weighted by the utilization rates, and fairness is understood as the maximal difference in the mean waiting times at each of the queues. In the literature multiple meanings have been associated to fairness, e.g. serving customers in order of arrival (see [2, 10]). These interpretations, however, are different from the fairness considered here.

In a polling model when the server switches to the next queue, a switchover time is incurred. There are many possible choices for deciding when the server should switch to the next queue. The rules studied most often are the exhaustive service discipline (when the server arrives at a queue, it serves its customers until the queue has become empty) and the gated service discipline (when the server arrives at a queue, a gate closes and only the customers who are before the gate, i.e., who are already present, will be served in this server visit).

The main advantage of the exhaustive strategy, is that it is optimally efficient. That is, it minimizes the sum of the mean waiting times at the queues weighted by their utilization rates. However, the differences between mean waiting times at the queues might be large. Typically, the heaviest loaded one has the smallest mean waiting time in this discipline. Conversely, the gated discipline leads in general to much smaller differences. But this is at the expense of the efficiency, which is much lower for this discipline. We aim to combine the best of both worlds into a new service discipline, by introducing a hybrid version of exhaustive and gated: the *κ-gated service discipline.*

The κ-gated discipline consists of using κ_i gated service phases at queue i before the server switches to the next queue. That is, upon arrival of the server, it servers the queue consecutively (at most) κ_i times, according to the gated discipline. So upon arrival of the server, a first gate closes and only the customers before this gate are served. After this, a second gate closes, and again only the customers before this gate are served, etcetera. This is done κ_i times, or until the queue becomes empty. The parameters κ_i are specified in the vector $\kappa = (\kappa_1, \ldots, \kappa_N)$, where N is the number of queues. Note that when $\kappa_i = 1$, queue i is served according to the gated discipline; when $\kappa_i \to \infty$, queue i is served according to the exhaustive discipline (as it is served until it becomes empty). One of the main questions studied in the current paper is whether the κ_i's can be optimized as to achieve both fairness and efficiency.

Fairness has frequently played a role in the choice of a service discipline in polling systems. For example, motivated by a dynamic bandwidth allocation problem of Ethernet Passive Optical Networks (EPON), in [13, 17] a *two-stage gated* service discipline is studied. In that case, a gate closes behind the customers in a stage-1 buffer at the moment the server arrives, the customers in the stage-2 buffer are being served, and then those present in stage-1 move to the stage-2 buffer. This was seen to give rise to relatively small differences between mean waiting times at the various queues, but at the expense of longer delays, i.e., at the expense of the efficiency of the system. The strategy was later generalized to multi-phase gated (see [18]). The κ-gated discipline can be seen as a variant of this discipline, where we have removed the extra cycles all customers have to wait for, in between moving to the next stage buffer. Hence, we expect it to lead to small differences between mean waiting times as well, but with significantly smaller total mean delays than for two- or multi-stage gated.

Besides the two- and multi-stage gated disciplines, a number of other disciplines have been proposed in the literature in order to achieve fairness (in the sense considered here). We mention a few in the following. Altman, Khamisy and Yechiali [1] (see also Shoham and Yechiali [15]) consider a so-called elevator strategy in a globally gated regime. In this setting the queues are visited in the order: $1, 2, \dots N-1, N, N, N-1, \dots 2, 1, 1, 2, \dots$ etc. When the server turns around at queue 1 or queue N, a gate closes at all queues: only those before the gate are served. This strategy turns out to be perfectly fair. However, it is far less efficient because of the globally gated regime. Our focus here is on cyclic models. Boxma, Van Wijk and Adan [11] introduce the Gated/Exhaustive discipline: the queues are visited cyclically, where in one cycle alternately all queues are served according to the gated discipline or all queues are served exhaustively. The incentive for this mixed strategy arose from the well-known expressions for the mean waiting time of queue i for gated respectively exhaustive systems: $\mathbb{E}(W_i^{gat}) = (1+\rho_i)\mathbb{E}(R_{C_i})$ respectively $\mathbb{E}(W_i^{exh}) = (1-\rho_i)\mathbb{E}(R^*_{C_i})$, where ρ_i is the workload. Furthermore $\mathbb{E}(R_{C_i})$ and $\mathbb{E}(R^*_{C_i})$ denote the mean residual cycle duration from the visit completion respectively beginning of Q_i on. These can be approximated by $\mathbb{E}(R_C) \approx \mathbb{E}(R_{C_i}) \approx \mathbb{E}(R^*_{C_i})$. From the resulting approximations for $\mathbb{E}(W_i^{gat})$ and $\mathbb{E}(W_i^{exh})$, one might expect the mean waiting time in the Gated/Exhaustive discipline to become $\mathbb{E}(W_i^{g/e}) \approx \mathbb{E}(R_C)$, which does not depend on i. However, it turns out that this guess was incorrect, as the exhaustive cycle dominates in the waiting times. The difference in waiting times only marginally decreases compared to exhaustive. To overcome this, [11] proposes the use of a polling table (see also [3, 19]), which prescribes the order in which queues are visited. This is related to [9], in which efficient visit orders are studied. Another option are efficient visit frequencies, see [8]. These options, however, do not focus on fairness.

Our contribution in this paper is as follows. We introduce the κ-gated discipline. Our motivation for this discipline is the search for a policy that achieves almost equal mean waiting times at the queues (fairness), without giving up too much of the efficiency. In earlier work in the literature, the focus has been solely on fairness, leading to inefficient policies [1, 17]. For the κ-gated discipline we derive the distributions and means of the waiting times, a pseudo conservation law for the weighted sum of the mean waiting times, and the fluid limits of the waiting times. We want to set the κ_i's so as to achieve maximal fairness without giving up too much on the efficiency of the system. To accomplish this, we use the fluid limits to derive a heuristic for setting κ. Finally, in a numerical study we extensively test the performance of the heuristic. It turns out to perform very well.

The structure of this paper is as follows. In Section 2 we introduce the model in more detail and give the notation that is being used. In Section 3 we derive the mean visit times of the queues, a *Pseudo Conservation Law* for the weighted sum of the mean waiting times, the waiting time distributions at all queues using *Multitype Branching Processes*, the mean waiting times using the *Mean Value Analysis* technique exploiting the concept of *Smart Customers*, and the *Fluid Limits* of the waiting times. In Section 4 we derive a heuristic rule for the setting of κ based on the fluid limits. Section 5 contains examples and a numerical study into the performance of the heuristic. We end with a conclusion and discussion of possible further work in Section 6.

2 Model and notation

We consider a polling system [16], with N queues, $Q_1, \dots, Q_N$, where each queue has infinite capacity. The queues are served by a single server, in fixed cyclic order $Q_1, Q_2, \dots, Q_N, Q_1, Q_2, \dots$. Customers in each queue are served in order of arrival (first come, first served). The arrival processes at the queues are independent Poisson processes with arrival rate λ_i at Q_i, $i = 1, \dots, N$. The service times at Q_i are independent and identically distributed (i.i.d.) random variables, denoted by B_i, having finite first and second moment, and Laplace–Stieltjes transform $\beta_i(.)$. By R_{B_i} we denote a residual service time at Q_i. The switch of the server from Q_i to Q_{i+1} lasts for a switchover time S_i, these being i.i.d. random variables, with finite first two moments, Laplace–Stieltjes transform $\sigma_i(.)$, and residual duration R_{S_i}. The sum of the switchover times is denoted by $S = \sum_{i=1}^{N} S_i$, where we assume $\mathbb{E}(S) > 0$. Its residual duration is denoted by R_S. Furthermore, we assume that the arrival processes, the service times and the switchover times are all mutually independent. Customers at Q_i are referred to as type i customers. Indices are understood to be modulo N: Q_{N+1} actually refers to Q_1.

The traffic offered per time unit at Q_i is denoted by ρ_i and is given by $\rho_i = \lambda_i \mathbb{E}(B_i)$. The total traffic offered to the system per time unit is $\rho = \sum_{i=1}^{N} \rho_i$. A necessary and sufficient condition for stability in case of gated and exhaustive services, is $\rho < 1$, see [12]. In the sequel we assume $\rho < 1$ and we concentrate on the steady-state behavior of the system. We are mainly interested in the waiting times of customers. By W_i we denote the steady-state waiting time of a customer at Q_i, excluding its own service time.

The cycle time starting from Q_i, denoted by C_i, consists of the visit times to each of the queues and all switchover times incurred. A well-known result [16] is that its first moment does not depend on i, and is given (for a stable system) by $\mathbb{E}(C) = \mathbb{E}(S)/(1-\rho)$. $\mathbb{E}(C)$ does not depend on the service disciplines at the queues.

We now describe the *κ-gated service discipline.* Upon arrival at Q_i, the server serves exactly those customers present on arrival (phase 1); when this is done, it serves exactly those customers present in Q_i at that moment (phase 2); and so on, until (at most) κ_i phases are completed, and then the server switches to the next queue. If the queue is empty at the start of a phase, the server also switches. This discipline consists of the prescription of $\kappa = (\kappa_1, \ldots, \kappa_N)$, with $\kappa_i \in \{1, 2, \ldots\} \cup \{\infty\}$ for all $i = 1, \ldots, N$. For $\kappa_i = 1$ the discipline at Q_i is equivalent to the well-known gated service discipline, and for $\kappa_i = \infty$ it is equivalent to the exhaustive service discipline. It is readily verified that the condition $\rho < 1$ is also necessary and sufficient for the stability in case of the κ-gated service discipline.

We want to achieve fairness in the waiting times, that is, we want the $\mathbb{E}(W_i)$ for $i = 1, \ldots, N$ to be (almost) equal. Hence, we want to minimize

$$\max_{i,j}\big(\mathbb{E}(W_i) - \mathbb{E}(W_j)\big).$$

On the other hand, we do not want to give up too much of the efficiency of the system. For the efficiency, we use the weighted sum over all mean waiting times:

$$\sum_{i=1}^{N} \rho_i \mathbb{E}(W_i).$$

This is a measure for the total workload in the system: it is the expected value of the waiting work in the system at an arbitrary moment. Hence, we focus on the following performance characteristic of the system:

$$\tilde{\gamma}(\alpha) := \max_{i,j}\big(\mathbb{E}(W_i) - \mathbb{E}(W_j)\big) + \alpha \sum_{i=1}^{N} \rho_i \mathbb{E}(W_i), \tag{1}$$

for some $\alpha \in [0, \infty)$. Note that (1) depends on the service discipline at each of the queues. Under the κ-gated discipline, for a given α, the κ can be optimized to minimize $\tilde{\gamma}$. This optimization is a trade-off between the fairness (maximal difference in mean waiting times) and the efficiency (weighted sum of mean waiting times). One can distinguish two extreme cases. For $\alpha = 0$ only the fairness of the discipline counts. In that case, the elevator strategy in a globally gated regime is the best choice, as it leads to equal mean waiting times. For $\alpha \to \infty$ only the efficiency of the system is important. The exhaustive discipline is optimal in that case. We remark that for the term measuring the efficiency, a so-called pseudo conservation law holds, and it is easily determined without having to calculate all individual mean waiting times (see Section 3.2).

3 Analysis of the κ-Gated Discipline

In this section we present the analysis of the κ-gated discipline. First we derive the mean visit times at each of the queues. Then we give a pseudo conservation law for the weighted sum of the mean waiting times. Next we present the derivation of the waiting time distributions, using multi–type branching processes. Following that, we briefly indicate a simpler way to compute the mean waiting times. For this we show that the discipline fits into the framework of smart customers, and then we apply mean value analysis for polling models. We end this section by presenting the fluid limits of the waiting times. These fluid limits are used in the next section to derive a heuristic for the optimal setting of κ.

3.1 Mean Visit Times

For the κ-gated discipline, we derive the expected duration of each of the visits and visit phases to a queue. The expected cycle duration is $\mathbb{E}(C) = \mathbb{E}(S)/(1-\rho)$. A fraction ρ_i of the cycle the server is working on Q_i, hence the expected duration of a visit to Q_i, denoted by $\mathbb{E}(V_i)$, is given by $\mathbb{E}(V_i) = \rho_i \mathbb{E}(C)$. This gives that the mean intervisit time, denoted by $\mathbb{E}(I_i)$, is given by $\mathbb{E}(I_i) = (1-\rho_i)\mathbb{E}(C)$. To further specify the visit times, let $\mathbb{E}(V_i^k)$ be the mean visit time of phase k at Q_i, for $k = 1, \dots, \kappa_i$. Then $\mathbb{E}(V_i) = \sum_{k=1}^{\kappa_i} \mathbb{E}(V_i^k)$. In the first phase, all work that arrived during the last phase of the previous cycle and the intervisit time has to be served. This gives for the mean durations:

$$\mathbb{E}(V_i^1) = \rho_i(\mathbb{E}(V_i^{\kappa_i}) + \mathbb{E}(I_i)). \tag{2}$$

In the second phase, the work that arrived during the first phase is served; in the third phase that of the second, and so on. This leads to:

$$\begin{aligned} \mathbb{E}(V_i^k) &= \rho_i \, \mathbb{E}(V_i^{k-1}) \\ &= \rho_i^{k-1} \mathbb{E}(V_i^1), \quad \text{for } k = 2, \dots, \kappa_i. \end{aligned}$$

Substituting this expression for $k = \kappa_i$ into (2) gives

$$\mathbb{E}(V_i^1) = \rho_i[\rho_i^{\kappa_i - 1}\mathbb{E}(V_i^1) + (1-\rho_i)\mathbb{E}(C)].$$

Solving this leads to $\mathbb{E}(V_i^1) = \rho_i \frac{1-\rho_i}{1-\rho_i^{\kappa_i}} \mathbb{E}(C)$, and hence

$$\mathbb{E}(V_i^k) = \rho_i^k \frac{1-\rho_i}{1-\rho_i^{\kappa_i}} \mathbb{E}(C), \quad k = 1, \ldots, \kappa_i. \tag{3}$$

Note that the mean duration of subsequent phases decreases, as is to be expected. It is readily verified that with (3), it indeed holds that: $\sum_{k=1}^{\kappa_i} \mathbb{E}(V_i^k) + \mathbb{E}(I_i) = \mathbb{E}(C)$, for $i = 1, \ldots, N$.

3.2 Pseudo Conservation Law

Boxma and Groenendijk [7] derive a so-called Pseudo Conservation Law (PCL) for the case of cyclic order polling systems. These pseudo conservation laws give an expression for the weighted sum of the mean waiting times at each of the queues: $\sum_{i=1}^N \rho_i \mathbb{E}(W_i)$. It is in that way a measure for the efficiency of the discipline. Based on a workload decomposition result, the following expression is derived in [7, (3.10)]:

$$\sum_{i=1}^N \rho_i \mathbb{E}(W_i) = \frac{\rho}{1-\rho} \sum_{i=1}^N \rho_i \mathbb{E}(R_{B_i}) + \rho \mathbb{E}(R_S) + \frac{\mathbb{E}(S)}{2(1-\rho)} \left(\rho^2 - \sum_{i=1}^N \rho_i^2 \right) + \sum_{i=1}^N \mathbb{E}(M_i), \tag{4}$$

where $\mathbb{E}(M_i)$ is the mean amount of work in Q_i at a departure epoch of the server from Q_i. This is the only term that depends on the service discipline at the queues. For the exhaustive discipline $\mathbb{E}(M_i^{exh})$ trivially equals zero (cf. [7, (3.11)]), and for gated it holds that $\mathbb{E}(M_i^{gat}) = \rho_i \mathbb{E}(V_i) = \rho_i^2 \mathbb{E}(S)/(1-\rho)$ (cf. [7, (3.12)]). The workload decomposition result in [7] is also valid for the κ-gated discipline, and we find, using (3):

$$\mathbb{E}(M_i^{\kappa-gat}) = \rho_i \mathbb{E}(V_i^{\kappa_i}) = \rho_i^{\kappa_i+1} \frac{1-\rho_i}{1-\rho_i^{\kappa_i}} \frac{\mathbb{E}(S)}{1-\rho}.$$

Remark that for the two extreme cases $\kappa_i = 1$ and $\kappa_i = \infty$ this expression simplifies to that of the gated respectively exhaustive discipline.

Comparing the $\mathbb{E}(M_i)$ terms for the different strategies, we find the following:

$$0 = \mathbb{E}(M_i^{exh}) \leq \mathbb{E}(M_i^{\kappa-gat}) \leq \mathbb{E}(M_i^{gat}),$$

with equality for the first '$\leq$' if and only if $\kappa_i = \infty$, and equality for the second '$\leq$' if and only if $\kappa_i = 1$. Exhaustive is the most efficient service discipline, as the server never switches when there are still customers in the queue it is serving. So, it leaves no customers behind that have to wait for an entire cycle. The latter is the case for (κ-) gated. Gated (i.e. $\kappa_i = 1$) is less efficient than κ-gated for $\kappa_i \geq 2$, since more customers will be left behind when the server switches to the next queue. It follows that the efficiency of the κ-gated discipline is always between that of exhaustive and gated.

By substituting the expression for $\mathbb{E}(M_i^{\kappa-gat})$ into (4) we find the pseudo conservation law for the κ-gated discipline:

$$\sum_{i=1}^{N} \rho_i \mathbb{E}(W_i) = \rho \frac{\sum_{i=1}^{N} \rho_i \mathbb{E}(R_{B_i})}{1-\rho} + \rho \mathbb{E}(R_S) + \frac{\mathbb{E}(S)}{2(1-\rho)} \left(\rho^2 - \sum_{i=1}^{N} \rho_i^2 + 2 \sum_{i=1}^{N} \rho_i^{\kappa_i+1} \frac{1-\rho_i}{1-\rho_i^{\kappa_i}} \right).$$

As only the terms $\mathbb{E}(M_i)$ depend on the service discipline (and hence on κ for the κ-gated discipline), we can restrict our attention to $\sum_{i=1}^{N} \mathbb{E}(M_i)$ instead of $\sum_{i=1}^{N} \rho_i \mathbb{E}(W_i)$. So in the sequel, instead of (1), we concentrate on optimizing:

$$\gamma(\alpha) := \max_{i,j} \big(\mathbb{E}(W_i) - \mathbb{E}(W_j) \big) + \alpha \sum_{i=1}^{N} \mathbb{E}(M_i), \tag{5}$$

for some $\alpha \in [0, \infty)$.

3.3 Waiting time distributions

We determine the Laplace–Stieltjes transform (LST) of the waiting times W_i analogously to Resing [14]. In [14] it is shown, that if the service discipline in each queue satisfies the so-called *branching property*, then the queue length process at polling instants of a fixed queue is a multitype branching process (MTBP) with immigration in each state. This leads to expressions for the generating function of the joint queue length process at polling instants. Conform e.g. [4] the LST of the waiting time then follows.

The κ-gated service discipline does satisfy the branching property [14, Property 1]. Let the start of the visit to Q_1 be the start of the cycle, then by the branching property, each customer present will during the cycle be replaced in an i.i.d. manner by customers of type $1, \ldots, N$, according to the probability generating function (pgf) $h_i(z)$, where $z = (z_1, \ldots, z_N)$. For the gated service discipline, this h_i is given by:

$$h_i^{(\text{gated})}(z) = \beta_i \left(\sum_{j=1}^{N} \lambda_j (1 - z_j) \right).$$

For κ-gated we can recursively express h_i as follows:

$$h_i^{(\text{1-gated})}(z) = h_i^{(\text{gated})}(z),$$

$$h_i^{(m\text{-gated})}(z) = \beta_i \left(\sum_{j=1, j\neq i}^{N} \lambda_j (1 - z_j) + \lambda_i \left(1 - h_i^{((m-1)\text{-gated})}(z) \right) \right), \text{ for } m = 2, 3, \ldots.$$

For $\kappa_i = \infty$, the pgf h_i coincides with that of the exhaustive service discipline, which is given by:

$$h_i^{(\infty\text{-gated})}(z) = h_i^{(\text{exhaustive})}(z) = \theta_i \left(\sum_{j=1, j\neq i}^{N} \lambda_j (1 - z_j) \right),$$

where $\theta_i(\cdot)$ is the LST of a busy period triggered by one type i customer in Q_i in isolation.

Analogously to [4], let $V_{b_i}(z)$ and $V_{c_i}(z)$ be the pgfs of the steady-state joint queue length distributions at the beginning, respectively completion of a visit to Q_i. We can express $V_{b_1}(z)$ in itself, by repeated application of the following relation, cf. [4, (2.2)]:

$$\begin{aligned} V_{b_{i+1}}(z) &= V_{c_i}(z)\sigma_i\Big(\sum_{j=1}^{N}\lambda_j(1-z_j)\Big) \\ &= V_{b_i}(z_1,\dots,z_{i-1},h_i^{(\kappa_i\text{-gated})}(z),z_{i+1},\dots,z_N)\sigma_i\Big(\sum_{j=1}^{N}\lambda_j(1-z_j)\Big),\ i=1,2,\dots,N, \end{aligned}$$

where $N+1$ is understood to be 1.

The LST of the steady-state waiting time distribution of a type i customer is given by, cf. [4, (2.8)]:

$$\mathbb{E}(e^{-\omega W_i}) = \frac{\tilde{V}_{c_i}(1-\omega/\lambda_i) - \tilde{V}_{b_i}(1-\omega/\lambda_i)}{(\omega-\lambda_i(1-\beta_i(\omega)))\mathbb{E}(C)},$$

where $\tilde{V}_{b_i}(\cdot)$ is the pgf of the steady-state marginal queue length distribution at a visit beginning of Q_i, given by $\tilde{V}_{b_i}(z) = V_{b_i}(1,\dots,1,z,1,\dots,1)$, with z as the ith argument, and $\tilde{V}_{c_i}(\cdot)$ is defined analogously. By differentiation, moments of the steady-state waiting time for an arbitrary type i customer can be derived.

3.4 Mean waiting times

We briefly discuss how the first moments of the waiting times, $\mathbb{E}(W_i)$, can easily be obtained in a more efficient way. For this, we show that the κ-gated discipline fits into the framework of a polling model with smart customers (introduced in [6]). We can then use mean value analysis (MVA) for polling systems (introduced by Winands, Adan and Van Houtum [20]), adapted for smart customers (cf. Boon et al. [5]).

In an ordinary polling model, customers arrive according to a Poisson process at a constant rate λ_i at Q_i. However, in the case of a model with smart customers, the arrival rate depends on the position of the server. This rate is $\lambda_{i,j}$ at Q_i when the server is serving (or switching to) Q_j. This concept can be used to route arriving customers to a specific queue, depending on the position of the server.

We use this routing in the following way. We introduce a polling model with the gated discipline that is related to one served according to the κ-gated discipline. In that model we create multiple copies of the same queue. We refer to this as the corresponding model. By routing the customers properly, we send them to the correct queue. That is, a customer arriving at Q_i in the original model is routed in the corresponding model to the copy of Q_i that will be served first. The underlying idea of this is the following. In the κ-gated model, arriving customers queue behind a gate, which only opens when the server starts one of the κ_i serving phases. In the corresponding model, each of these phases now becomes a separate queue. Hence, we create a polling model with κ_i copies of queue Q_i, denoted by $Q_i^{(1)}, \ldots, Q_i^{(\kappa_i)}$. No switchover times are incurred between these copies. Denoting phase k of a visit to Q_i by $V_i^{(k)}$, then the cycle, including the switchover times S_i (between $Q_i^{(\kappa_i)}$ and $Q_{i+1}^{(1)}$) becomes:

$$V_1^{(1)} - V_1^{(2)} - \ldots - V_1^{(\kappa_1)} - S_1 - V_2^{(1)} - V_2^{(2)} - \ldots - V_2^{(\kappa_2)} - S_2 - \ldots - S_{N-1} - V_N^{(1)} - \ldots - V_N^{(\kappa_N)} - S_N.$$

We now have an 'ordinary' cyclic polling model with $\sum_{i=1}^N \kappa_i$ queues, each of which is served according to the gated discipline. We want this system to have the same arrival process as the original one. For that, we have to route the arriving customers, depending on the position of the server. A customer arriving at Q_i in the original model is now routed to $Q_i^{(j)}$ during $V_i^{(j-1)}$, for $j = 2, \ldots, \kappa_i$; and to $Q_i^{(1)}$ otherwise.

The corresponding model is a polling model with smart customers. These are studied by Boon et al. [5]. In [5, Section 6] a system of $\mathcal{O}\left(N^2\right)$ linear equations is derived for an N queue polling model with the exhaustive service discipline, from which the $\mathbb{E}(W_i)$ can immediately be solved. Analogously, we can write down a system of $\mathcal{O}\left((\sum_i \kappa_i)^2\right)$ linear equations, from which the $\mathbb{E}(W_i)$ directly follow.

Remark: Boon et al. [5, Section 8.2] also give the MTBP approach for polling models with smart customers. In case that some of the arrival rates are equal to zero, they have to introduce extra queues requiring zero service times. However, by the structure of the κ-gated discipline, the MTBP analysis can be reduced to that presented in Section 3.3.

3.5 Fluid limits

The exact expressions for the mean waiting times, following from Sections 3.3 and 3.4, do not provide an easy way to determine the κ_i's minimizing $\gamma(\alpha)$. Therefore, we derive the fluid limit approximations of the mean waiting times. These approximations yield closed form expressions, and can hence easily be used to (approximately) optimize the κ_i's.

By taking the fluid limits, we scale the interarrival and service times. For this, we let $\lambda_i \to \infty$ and $\mathbb{E}(B_i) \to 0$ while keeping the workload $\lambda_i \mathbb{E}(B_i) = \rho_i$ fixed. We concentrate on the amount of work present at a queue, denoted by H_i at queue Q_i. By the use of this scaling, we smoothen the discrete process H_i into a continuous one. In this way, work arrives at a constant rate ρ_i, and

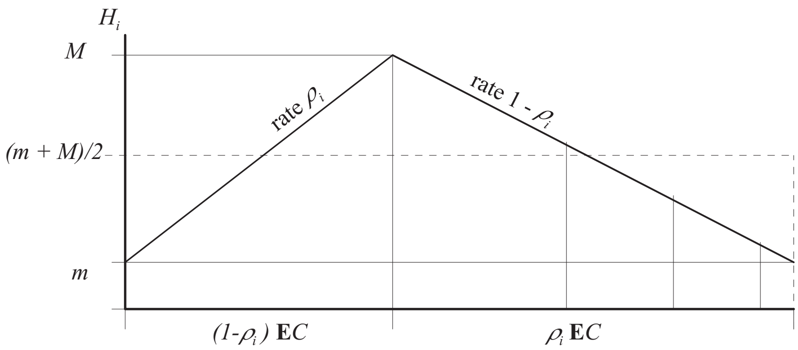

Figure 1: The fluid limit of the workload H_i at Q_i during one cycle.

during the visit time work is removed at rate 1. So, during the intervisit time of mean length $\mathbb{E}(I_i) = (1 - \rho_i)\,\mathbb{E}(C)$, the amount of work increases at rate ρ_i, and during the visit time, with mean length $\mathbb{E}(V_i) = \rho_i\,\mathbb{E}(C)$, the amount of work decreases at rate $1 - \rho_i$. This cyclic pattern repeats itself in every cycle. Hence, the workload H_i during a cycle in the κ-gated discipline becomes as depicted in Figure 1.

At the *end* of the visit to Q_i, the amount of work present is equal to that built up during the last visit phase $V_i^{\kappa_i}$. So, it is $\rho_i\,\mathbb{E}(V_i^{\kappa_i}) =: m$. At the *start* of the visit time it is equal to the work already present at the beginning of Q_i, which is m, plus the work built up during the intervisit time. Hence, it is $m + \rho_i\,\mathbb{E}(I_i) =: M$. Consequently, the average fluid level during a cycle, i.e. the mean workload $\mathbb{E}(H_i)$, is given by:

$$\begin{aligned}\mathbb{E}(H_i) = \frac{m+M}{2} &= m + \frac{\rho_i\,\mathbb{E}(I_i)}{2} \\ &= (1+\rho_i^{\kappa_i})\frac{\rho_i(1-\rho_i)}{2\,(1-\rho_i^{\kappa_i})}\mathbb{E}(C).\end{aligned}$$

Using Little's law formulated for the workload, $\mathbb{E}(H_i) = \rho_i\,\mathbb{E}(W_i)$, the fluid limit of the mean waiting time of a type i customer directly follows:

$$\mathbb{E}(W_i^{fluid}) = \frac{m+M}{2\,\rho_i} = (1+\rho_i^{\kappa_i})\frac{1-\rho_i}{2\,(1-\rho_i^{\kappa_i})}\mathbb{E}(C). \tag{6}$$

Figure 2 shows these fluid limits for different κ_i.

It is easily checked that for $\kappa_i = 1$, (6) reduces to $\mathbb{E}(W_i^{fluid}) = \frac{1+\rho_i}{2}\mathbb{E}(C)$, which is indeed the fluid limit for the gated discipline. For $\kappa_i = \infty$, (6) reduces to $\mathbb{E}(W_i^{fluid}) = \frac{1-\rho_i}{2}\mathbb{E}(C)$, which is indeed the fluid limit for the exhaustive discipline.

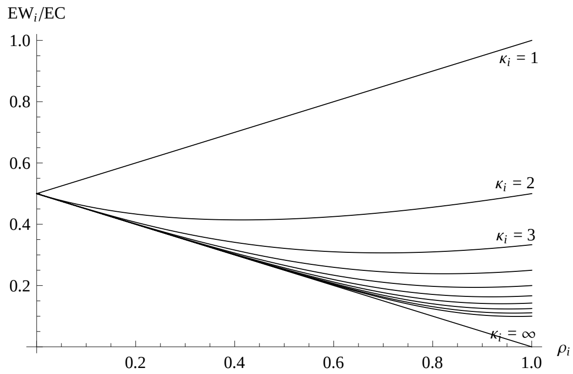

Figure 2: Fluid limits of the waiting times: $\mathbb{E}(W_i^{fluid})/\mathbb{E}(C)$ *plotted versus* ρ_i *for* $\kappa_i = 1, 2, 3, \ldots, 10$ *and for* $\kappa_i = \infty$.

4 Balancing fairness and efficiency

We now want to choose κ such that on one hand we achieve fairness, while on the other hand the system is still efficient. For that, we want to determine the κ that minimizes $\gamma(\alpha)$ as given in (5). As we do not have closed form expressions for the mean waiting times, optimization could be done by an exhaustive search over all κ_i. However, we use the fluid limits (6) as approximation for the mean waiting times in the optimization:

$$\min_{\kappa} \gamma^{fluid}(\kappa, \alpha) \tag{7}$$

where

$$\gamma^{fluid}(\kappa, \alpha) = \max_{i,j} \big(\mathbb{E}(W_i^{fluid}) - \mathbb{E}(W_j^{fluid})\big) + \alpha \sum_{i=1}^{N} \mathbb{E}(M_i^{\kappa-gat}).$$

For deriving a heuristic rule for the optimal setting of κ, we take the following approach. First we determine the κ_i's such that all mean waiting times are equal (optimal fairness), then, using these κ_i's, we minimize the term $\sum_i \mathbb{E}(M_i)$ (maximal efficiency given optimal fairness). That is, we consider the following optimization problem:

$$\begin{aligned} \min_{\kappa} \ & \sum_{i=1}^{N} \mathbb{E}(M_i^{\kappa-gat}), \\ \text{such that} \ & \mathbb{E}(W_1^{fluid}) = \ldots = \mathbb{E}(W_N^{fluid}). \end{aligned} \tag{8}$$

For the moment we allow the κ_i's to be fractional, later we round them to integers. Note that the problem in (8) does not depend on α. In an extensive numerical study in the next section we compare the performance of this heuristic setting to that of the optimal setting solving (7). We now solve (8), first for 2 queues, and then for N queues.

4.1 2 queues

For simplicity we start with the case of 2 queues. In this case we can explicitly solve $\mathbb{E}(W_1^{fluid}) = \mathbb{E}(W_2^{fluid})$ for κ_2 in terms of κ_1, ρ_1 and ρ_2:

$$(1+\rho_1^{\kappa_1})\frac{1-\rho_1}{2(1-\rho_1^{\kappa_1})} = (1+\rho_2^{\kappa_2})\frac{1-\rho_2}{2(1-\rho_2^{\kappa_2})},$$

where we have divided by $\mathbb{E}(C) \neq 0$. Solving for κ_2, denoted by κ_2^{opt}, gives:

$$\rho_2^{\kappa_2^{opt}} = \frac{(1-\rho_1)\,(1+\rho_1^{\kappa_1}) - (1-\rho_2)\,(1-\rho_1^{\kappa_1})}{(1-\rho_1)\,(1+\rho_1^{\kappa_1}) + (1-\rho_2)\,(1-\rho_1^{\kappa_1})}. \tag{9}$$

So, this κ_2 achieves optimal fairness (recall that we allowed κ_2 to be fractional). Using this κ_2, we now optimize the efficiency, i.e. we minimize:

$$\begin{aligned}\sum_{i=1}^{2}\mathbb{E}(M_i) &= \rho_1^{\kappa_1+1}\frac{1-\rho_1}{1-\rho_1^{\kappa_1}} + \rho_2^{\kappa_2^{opt}+1}\frac{1-\rho_2}{1-\rho_2^{\kappa_2^{opt}}} \\ &= \frac{(\rho_1-\rho_2)\rho_2 + \rho_1^{\kappa_1}\left(2(1-\rho_1)\rho_1 + (2-\rho_1)\rho_2 - \rho_2^2\right)}{2(1-\rho_1^{\kappa_1})}. \end{aligned} \tag{10}$$

In (10) we have substituted (9) and simplified the expression.

The minimum of (10) (where $\kappa_1 > 0$, for $\rho_1 \neq \rho_2$) is found for $\kappa_1 \to \infty$. In this way, from (9), κ_2^{opt} becomes:

$$\kappa_2^{opt} = \log_{\rho_2}\frac{\rho_2-\rho_1}{2-\rho_1-\rho_2}. \tag{11}$$

This only makes sense for $\rho_1 < \rho_2$; if $\rho_1 > \rho_2$, we interchange the indices. In case $\rho_1 = \rho_2$ all $\kappa_1 = \kappa_2$ give equal mean waiting times. However, $\kappa_1 = \kappa_2 = \infty$ optimizes the efficiency. So, we come up with the following heuristic for the choice of κ_1 and κ_2:

$$\begin{cases} \text{if } \rho_1 < \rho_2\text{:} & \kappa_1 = \infty,\ \kappa_2 = \log_{\rho_2}\frac{\rho_2-\rho_1}{2-\rho_1-\rho_2}, \\ \text{if } \rho_1 = \rho_2\text{:} & \kappa_1 = \kappa_2 = \infty, \\ \text{if } \rho_1 > \rho_2\text{:} & \kappa_1 = \log_{\rho_1}\frac{\rho_1-\rho_2}{2-\rho_1-\rho_2},\ \kappa_2 = \infty. \end{cases}$$

In order to get integer κ_i's, we have three possibilities: rounding to the nearest integer denoted by $[x]$; using the integer floor function, $\lfloor x \rfloor$; and using the integer ceiling function, $\lceil x \rceil$. We study all three options in the numerical study in Section 5. We denote a κ set according to the heuristic by $[\kappa]$, $\lfloor \kappa \rfloor$ respectively $\lceil \kappa \rceil$.

We plot (11) in Figure 3, for $\rho_1 < \rho_2$ (and $\rho_1 + \rho_2 < 1$ for stability) where we round κ_2. From the figure it becomes clear that $\kappa_2 = 2$ almost always is a proper choice.

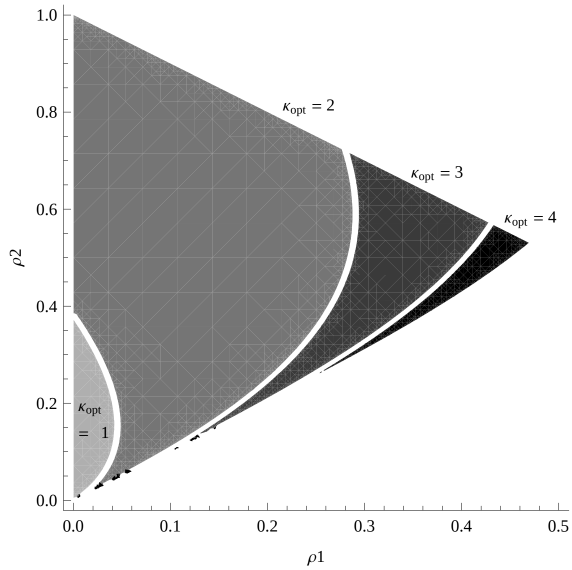

Figure 3: Optimal value of κ_2 (rounded to the nearest integer), given by $\kappa_2^{opt} = \left[\log_{\rho_2} \dfrac{\rho_2 - \rho_1}{2 - \rho_1 - \rho_2}\right]$, *for $\rho_1 < \rho_2$ and $\rho_1 + \rho_2 < 1$.*

4.2 N queues

For N queues we first determine which κ_i's give equal mean waiting times. We solve $\mathbb{E}(W_1^{fluid}) = \mathbb{E}(W_j^{fluid})$ for $j = 2, \ldots, N$, which leads to an expression analogous to (9), with 2 replaced by j everywhere. We plug these into $\sum_i \mathbb{E}(M_i)$. The resulting expression depends only on κ_1, and on all ρ_i's. It only makes sense if ρ_1 is the smallest of all ρ_i, and it is minimized for $\kappa_1 \to \infty$. From this, the expressions for the optimal $\kappa_2, \ldots, \kappa_N$ directly follow: $\kappa_j^{opt} = \log_{\rho_j} \frac{\rho_j - \rho_1}{2 - \rho_1 - \ldots - \rho_N}$, for $j = 2, \ldots, N$.

Hence, we come up with the following heuristic for the choice of the κ_i's, $i = 1, \ldots, N$:

$$\begin{cases} \text{For all } i \text{ such that } i = \arg\min \rho_i, \text{ let } \kappa_i = \infty; \\ \text{For all } j = 1, 2, \ldots, N \text{ where } j \neq i, \text{ let } \kappa_j = \log_{\rho_j} \frac{\rho_j - \rho_i}{2 - \rho}. \end{cases} \tag{12}$$

Recall that we have three options to get integer κ_i's (round, floor, ceiling). An important notion here is that, by construction, this heuristic does *not* depend on α. The numerical results in the next section, however, show that it performs well for a wide range of α's. So, this heuristic is robust against the value of α.

5 Numerical analysis

In this section we first consider two examples, followed by an extensive numerical study into the performance of the heuristic setting of κ. For each instance we determine $\gamma(\alpha)$ as defined in (5). For brevity of notation we define:

$$\Delta = \max_{i,j}\big(\mathbb{E}(W_i) - \mathbb{E}(W_j)\big),$$
$$\Sigma = \sum_{i=1}^{N} \mathbb{E}(M_i).$$

We compare the results of the κ-gated discipline with the elevator strategy in a globally gated regime, cf. [1, 15]. For this strategy all mean waiting times are equal: $\mathbb{E}(W_1^{elev.GG}) = \mathbb{E}(W_2^{elev.GG}) = \ldots = \mathbb{E}(W_N^{elev.GG})$, and given by, cf. [1, (6), (10)]:

$$\mathbb{E}(W_1^{elev.GG}) = \frac{1}{1-\rho}\sum_{i=1}^{N} \rho_i \mathbb{E}(R_{B_i}) + \mathbb{E}(R_S) + \frac{1+\rho}{2(1-\rho)}\mathbb{E}(S).$$

The PCL easily follows: $\sum_{i=1}^{N} \rho_i \mathbb{E}(W_i) = \rho\, \mathbb{E}(W_1)$. Using (4) we then derive that:

$$\sum_{i=1}^{N} \mathbb{E}(M_i^{elev.GG}) = \frac{\rho + \sum_{i=1}^{N} \rho_i^2}{2(1-\rho)}\, \mathbb{E}(S).$$

5.1 Examples

Example 1. Consider a polling model with $N = 2$ queues, $S_i, B_i \sim \exp(1)$, $i = 1, 2$, and $\lambda_1 = 0.6$, $\lambda_2 = 0.2$. Hence $\rho_1 = 0.6$ and $\rho_2 = 0.6$. We have $\rho_1 > \rho_2$ and $\log_{\rho_1} \frac{\rho_1 - \rho_2}{2-\rho_1-\rho_2} \approx 2.15$. Hence the heuristic settings are $[\kappa] = \lfloor\kappa\rfloor = (2, \infty)$ and $\lceil\kappa\rceil = (3, \infty)$.

For the κ-gated discipline, taking $\kappa_1, \kappa_2 \in \{1, 2, 3, \infty\}$, the results are given in Table 1. It turns out that the heuristic settings for κ perform quite well. Although suboptimal for small α, the performance seems to be rather robust with respect to α. Despite $\kappa = (2, 2)$ performs better in this example for the four values of α chosen, it follows from the values of Δ and Σ that for large α the heuristic setting will dominate in performance. In general, however, the heuristic settings outperform the $(2, 2)$ (unless α is small, i.e. less than 1), as the numerical study in Section 5.2 shows.

The difference Δ in the example turns out to be minimal for $\kappa = (2, 2)$. This is not surprising as for $N = 2$ and $\kappa = (2, 2)$ the κ-gated discipline closely resembles the elevator strategy in a globally gated regime (cf. [1, 15]). In this discipline the visit order is $1, 2, \ldots, N-1, N, N, N-$

	κ_1	κ_2	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
	1	1	12.77	9.69	3.08	4.00	3.1	7.1	11.1	23.1
	2	1	8.42	11.49	3.07	1.75	3.1	4.8	6.6	11.8
	3	1	6.90	12.60	5.70	1.06	5.7	6.8	7.8	11.0
	∞	1	**5.04**	14.88	9.84	0.40	9.8	10.2	10.6	11.8
	1	2	13.00	7.21	5.83	3.67	5.8	9.5	13.2	24.2
	2	2	8.77	8.77	0.0038	1.42	0.0	**1.4**	**2.8**	7.1
	3	2	7.27	9.82	2.55	0.73	2.6	3.3	4.0	6.2
	∞	2	5.38	12.20	6.82	0.07	6.8	6.9	7.0	7.2
	1	3	13.10	6.76	6.34	3.61	6.3	10.0	13.6	24.4
	2	3	8.85	8.25	0.60	1.36	0.6	2.0	3.3	7.4
	3	3	7.36	9.29	1.92	0.67	1.9	2.6	3.3	5.3
	∞	3	5.47	11.66	6.19	0.01	6.2	6.2	6.2	6.3
	1	∞	13.12	**6.64**	6.48	3.60	6.5	10.1	13.7	24.5
$\lfloor\kappa\rceil, \lfloor\kappa\rfloor$:	$\underline{2}$	$\underline{\infty}$	$\underline{8.88}$	$\underline{8.11}$	$\underline{0.77}$	$\underline{1.35}$	$\underline{0.8}$	$\underline{2.1}$	$\underline{3.5}$	$\underline{7.5}$
$\lceil\kappa\rceil$:	$\underline{3}$	$\underline{\infty}$	$\underline{7.39}$	$\underline{9.13}$	$\underline{1.74}$	$\underline{0.66}$	$\underline{1.7}$	$\underline{2.4}$	$\underline{3.1}$	$\underline{\mathbf{5.0}}$
	∞	∞	5.50	11.50	6.00	**0.00**	6.0	6.0	6.0	6.0
	Elev.GG		15.00	15.00	**0.00**	6.00	**0.0**	6.0	12.0	30.0

Table 1: Results for Example 1 for the κ-gated strategy, where $\kappa_1, \kappa_2 \in \{1, 2, 3, \infty\}$. Smallest values per column are given in bold; the optimal settings from the heuristic are underlined. Recall that $\kappa_i = \infty$ is equivalent to the exhaustive service discipline; $\kappa_i = 1$ to the gated service discipline. Elevator strategy in a globally gated regime is added for comparison.

$1, \ldots, 2, 1, 1, 2, \ldots$, and *all* gates are closed when turning around at 1 and at N. Hence, for $N = 2$ the queues are served as:

$$\ldots - Q_1 \overset{(*)}{-} Q_1 - S_1 - Q_2 \overset{(*)}{-} Q_2 - S_2 - Q_1 \overset{(*)}{-} Q_1 - S_1 - \ldots - \ldots$$

where $(*)$ denotes that the gate is closed at *both* queues. In the κ-gated strategy where $\kappa = (2, 2)$ the queues are served as:

$$\ldots - Q_1^{(1)} - Q_1^{(2)} - S_1 - Q_2^{(1)} - Q_2^{(2)} - S_2 - Q_1^{(1)} - Q_1^{(2)} - S_1 - \ldots - \ldots$$

where the gate is closed when each service phase starts. As the elevator strategy in a globally gated regime leads to $\mathbb{E}(W_1) = \mathbb{E}(W_2)$, it should not be surprising that the $(2, 2)$-gated strategy leads to *almost* equal mean waiting times.

Example 2. Now consider the following setting. Again we have $N = 2$ queues, $S_i \sim \exp(2)$, $B_i \sim \exp(1)$, $i = 1, 2$ and $\lambda_1 = 0.35$, $\lambda_2 = 0.25$. The heuristic settings are $\lfloor\kappa\rceil = \lceil\kappa\rceil = (3, \infty)$ and $\lfloor\kappa\rfloor = (2, \infty)$. The results are given in Table 2. The heuristic setting of κ performs very well, and is even optimal for $\alpha = 1, 2$, and 5. Note that Δ is again small for $\kappa = (2, 2)$.

	κ_1	κ_2	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
	1	1	9.30	8.68	0.63	1.85	0.6	2.5	4.3	9.9
	2	1	6.36	9.16	2.80	0.94	2.8	3.7	4.7	7.5
	3	1	5.60	9.37	3.77	0.73	3.8	4.5	5.2	7.4
	∞	1	**5.19**	9.53	4.33	0.63	4.3	5.0	5.6	7.5
	1	2	9.57	6.30	3.28	1.35	3.3	4.6	6.0	10.0
	2	2	6.65	6.77	0.12	0.44	0.1	0.6	1.0	2.3
	3	2	5.87	6.98	1.11	0.23	1.1	1.3	1.6	2.2
	∞	2	5.46	7.16	1.71	0.13	1.7	1.8	2.0	2.3
	1	3	9.66	5.80	3.86	1.25	3.9	5.1	6.4	10.1
	2	3	6.74	6.25	0.49	0.35	0.5	0.8	1.2	2.2
	3	3	5.97	6.47	0.50	0.13	0.5	0.6	0.8	1.2
	∞	3	5.55	6.65	1.10	0.03	1.1	1.1	1.2	1.2
	1	∞	9.70	**5.63**	4.07	1.23	4.1	5.3	6.5	10.2
$\lfloor\kappa\rfloor$:	$\underline{2}$	$\underline{\infty}$	$\underline{6.79}$	$\underline{6.07}$	$\underline{0.72}$	$\underline{0.32}$	$\underline{0.7}$	$\underline{1.0}$	$\underline{1.4}$	$\underline{2.3}$
$[\kappa]$, $\lceil\kappa\rceil$:	$\underline{3}$	$\underline{\infty}$	$\underline{6.02}$	$\underline{6.28}$	$\underline{0.26}$	$\underline{0.10}$	$\underline{0.3}$	$\underline{\mathbf{0.4}}$	$\underline{\mathbf{0.5}}$	$\underline{\mathbf{0.8}}$
	∞	∞	5.60	6.46	0.87	**0.00**	0.9	0.9	0.9	0.9
	Elev.GG		11.50	11.50	**0.00**	3.93	**0.0**	3.9	7.9	39.3

Table 2: Results for Example 2 for the κ-gated strategy. Smallest values per column are given in bold; the optimal settings from the heuristic are underlined.

5.2 Performance of fluid based heuristic

In a numerical experiment we study the performance of the heuristic settings for the κ_i's. We use a testbed with 4,614 instances (see Table 3) with $N = 2, 3, 4$, and 5 queues. For $\alpha = 0, 1, 2$, and 5 we calculate the mean waiting times in case of:

- exhaustive;
- gated;
- κ-gated with κ as in the heuristic (cf. (12));
- κ-gated with κ optimal, found by enumeration of all possibilities (for $N = 2, 3$);
- elevator strategy in a globally gated regime (cf [1, 15]).

The elevator strategy in a globally gated regime is added for comparison as it is known to give identical mean waiting times. However, it is in general far less efficient. On the contrary, the exhaustive discipline is optimally efficient, however, it might be less fair. For the κ-gated discipline, we optimize the κ. This is done by enumerating over all combinations of $\kappa_i \in \{1, 2, 3, 4, 5, 6, \infty\}$ (for $N = 2$) or $\kappa_i \in \{1, 2, 3, \infty\}$ (for $N = 3$), for $i = 1, \ldots, N$. For $N = 4$ and 5 we leave this out.

The results for $N = 2, 3, 4$, and 5 are respectively given in Tables 4, 5, 6, and 7. The results over all these N are in Table 8. From the tables we can make the following observations. The elevator

TESTBED	
$N = 2$ (2,925 settings)	
λ_1, λ_2	$\in \{0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95\}$,
B_1	$\sim \exp(1)$,
B_2, S_1, S_2	$\sim \exp(.)$ with mean $\in \{0.2, 0.5, 1., 2., 5.\}$,
$N = 3$ (243 settings)	
$\lambda_1, \lambda_2, \lambda_3$	$\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$,
B_1	$\sim \exp(1)$,
B_2, B_3	$\sim \exp(.)$ with same mean $\in \{0.5, 1., 2.\}$,
S_1, S_2, S_3	$\sim \exp(.)$ with same mean $\in \{0.5, 1., 2.\}$,
$N = 4$ (552 settings)	
$\lambda_1, \ldots, \lambda_4$	$\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$,
B_1	$\sim \exp(1)$,
$B_2, \ldots, B_4$	$\sim \exp(.)$ with same mean $\in \{0.5, 1., 2.\}$,
$S_1, \ldots, S_4$	$\sim \exp(.)$ with same mean $\in \{0.5, 1., 2.\}$,
$N = 5$ (894 settings)	
$\lambda_1, \ldots, \lambda_5$	$\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$,
B_1	$\sim \exp(1)$,
$B_2, \ldots, B_5$	$\sim \exp(.)$ with same mean $\in \{0.5, 1., 2.\}$,
$S_1, \ldots, S_5$	$\sim \exp(.)$ with same mean $\in \{0.5, 1., 2.\}$,

Table 3: Test bed for numerical study: full factorial design of the given possibilities which are stable ($\sum_{i=1}^{N} \rho_i < 1$). In total 4,614 settings.

discipline \ averages	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
Q_1 exh - Q_2 exh	9.8	30.7	25.6	**0.0**	25.6	25.6	25.6	25.6
Q_1 exh - Q_2 gat	**7.5**	35.8	28.2	1.7	28.2	29.9	31.6	36.7
Q_1 gat - Q_2 exh	21.7	**10.5**	11.3	7.9	11.3	19.1	27.0	50.6
Q_1 gat - Q_2 gat	19.5	15.2	6.2	9.6	6.2	15.7	25.3	54.0
elevator gg	22.7	22.7	**0.0**	11.9	**0.0**	11.9	23.8	59.5
κ-gat heur (round)	13.2	12.9	0.7	4.4	0.7	4.4	8.1	19.3
κ-gat heur (floor)	15.3	12.9	3.9	5.8	3.9	9.7	15.4	32.8
κ-gat heur (ceiling)	12.3	13.5	1.9	2.9	1.9	4.8	7.7	16.4
$\kappa = (2, 2)$	13.5	13.6	0.4	4.1	0.4	4.5	8.6	20.9
κ-gat opt $\alpha = 0$	13.5	13.5	0.3	4.1	0.3	4.4	8.5	20.6
κ-gat opt $\alpha = 1$	13.1	13.2	0.5	3.7	0.5	**4.1**	7.8	18.7
κ-gat opt $\alpha = 2$	12.0	13.5	2.0	2.7	2.0	4.7	**7.4**	15.4
κ-gat opt $\alpha = 5$	10.4	15.3	6.7	1.1	6.7	7.8	8.9	**12.2**

Table 4: Results of numerical study for $N = 2$: average values over 2,925 cases (as described in Table 3). Minimum value per column in bold. Optimization of κ by exhaustive search over $\kappa_i \in \{1, 2, 3, 4, 5, 6, \infty\}$, $i = 1, 2$.

discipline \ averages	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	$\mathbb{E}(W_3)$	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
exhaustive	11.2	12.2	12.3	6.2	**0.0**	6.2	6.2	6.2	**6.2**
gated	16.0	15.3	15.4	4.2	5.6	4.2	9.9	15.5	32.4
elevator gg	21.4	21.4	21.4	**0.0**	10.2	**0.0**	10.2	20.3	50.8
κ-gat heur (round)	12.2	12.1	12.2	0.7	1.6	0.7	**2.4**	**4.0**	8.9
κ-gat heur (floor)	14.2	13.8	13.8	6.0	4.9	6.0	10.9	15.8	30.5
κ-gat heur (ceiling)	12.4	**12.1**	12.2	1.2	1.6	1.2	2.8	4.4	9.2
κ-gat opt $\alpha = 0$	12.5	12.7	**10.5**	2.4	1.2	2.4	3.6	4.9	8.6
κ-gat opt $\alpha = 1$	12.2	12.6	11.1	2.3	1.2	2.3	3.5	4.7	8.3
κ-gat opt $\alpha = 2$	11.6	12.9	11.6	3.3	0.9	3.2	4.2	5.0	7.7
κ-gat opt $\alpha = 5$	**10.8**	13.2	12.0	4.5	0.6	4.5	5.1	6.7	7.4

Table 5: Results of numerical study for $N = 3$: average values over 243 cases (as described in Table 3). Optimization of κ by exhaustive search over $\kappa_i \in \{1, 2, 3, \infty\}$, $i = 1, 2, 3$.

discipline \ averages	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	$\mathbb{E}(W_3)$	$\mathbb{E}(W_4)$	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
exhaustive	**20.3**	**22.4**	**22.4**	**22.6**	9.7	**0.0**	9.7	9.7	9.7	**9.7**
gated	30.6	29.0	29.0	29.1	8.0	11.1	8.0	19.0	30.1	63.2
elevator gg	43.5	43.5	43.5	43.5	**0.0**	23.0	**0.0**	23.0	46.0	115.0
κ-gat heur (round)	22.9	23.1	23.2	23.2	2.3	2.9	2.3	5.2	8.2	17.0
κ-gat heur (floor)	29.5	27.6	27.6	27.6	11.6	10.5	11.6	22.0	32.5	63.8
κ-gat heur (ceiling)	22.9	23.1	23.2	23.2	2.3	2.9	2.3	**5.2**	**8.2**	17.0

Table 6: Results of numerical study for $N = 4$: average values over 552 cases (as described in Table 3). Note: Ceiling differs in only 3 instances from Round.

discipline \ averages	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	$\mathbb{E}(W_3)$	$\mathbb{E}(W_4)$	$\mathbb{E}(W_5)$	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
exhaustive	**19.9**	**20.6**	**20.6**	**20.6**	**20.7**	7.4	**0.0**	7.4	7.4	7.4	**7.4**
gated	26.6	26.0	26.0	26.1	26.1	6.5	8.2	6.5	14.8	23.0	47.6
elevator gg	39.9	39.9	39.9	39.9	39.9	**0.0**	20.2	**0.0**	20.2	40.4	101.0
κ-gat heur (round)	21.1	21.3	21.3	21.3	21.4	2.4	1.9	2.4	**4.4**	**6.3**	12.1
κ-gat heur (floor)	26.1	25.0	25.0	25.0	25.0	9.2	7.9	9.2	17.1	25.0	48.8
κ-gat heur (ceiling)	21.1	21.3	21.3	21.3	21.4	2.4	1.9	2.4	**4.4**	**6.3**	12.1

Table 7: Results of numerical study for $N = 5$: average values over 894 cases (as described in Table 3). Note: Ceiling identical to Round in all tested instances.

strategy in a globally gated regime, having equal mean waiting times (maximal fairness), is hence always optimal for $\alpha = 0$. This would be the case for small values of α near zero as well. The exhaustive strategy, leading to $\Sigma = 0$ (maximal efficiency), hence would be optimal for large values of α. The κ-gated discipline, using the heuristic settings for κ, seems to perform very well in the

discipline \ averages	Δ	Σ	$\gamma(0)$	$\gamma(1)$	$\gamma(2)$	$\gamma(5)$
exhaustive	19.1	**0.0**	19.1	19.1	19.1	19.1
gated	6.4	9.3	6.4	15.6	24.9	52.7
elevator gg	**0.0**	14.7	**0.0**	14.7	29.5	73.7
κ-gat heur (round)	1.2	3.2	1.2	**4.4**	7.6	17.1
κ-gat heur (floor)	5.9	6.7	5.9	12.6	19.3	39.5
κ-gat heur (ceiling)	1.9	2.6	1.9	4.5	**7.2**	**15.0**

Table 8: Results over all 4,614 instances.

range of α's in between. For a specific α (and specific setting of the parameters), one can typically find a better performing κ, but this optimization by exhaustive search is very time-consuming. The performance of the heuristics turns out to perform close to optimal for all given α's. For $N = 2$ it outperforms $(2, 2)$ for all α except for α close to zero. When using the floor function in the heuristic, the results seem to be not that good. It is both less fair and less efficient on average than the rounding and ceiling. The performance of those two does not differ that much.

One might expect the performance of the different settings to depend heavily on the switchover times incurred during a cycle, as during those intervals all work in the system is waiting. For that reason, we separate the results according to the value of $\mathbb{E}(S)$ (the mean total switchover time during a cycle), see Table 9. We focused on $N = 2$, as this most clearly illustrates the results. From the table, we see that the performance of e.g. the elevator strategy in a globally gated regime is best for small values of $\mathbb{E}(S)$, as is to be expected, but it is outperformed by the κ-gated discipline already for small α, by all indicated choices for the setting of the κ (except the heuristic setting using the floor function). Note that it is also outperformed by $\kappa = (2, 2)$, although these settings closely resemble each other.

Summarizing, the κ-gated discipline with κ set according to the heuristic, either rounding or ceiling, performs very well. It is robust against the setting of α and it performs well over a wide range of values for $\mathbb{E}(S)$.

6 Conclusion

We introduced the κ-gated service discipline for a polling model. It is a hybrid of the classical gated and exhausted disciplines, and consists of using κ_i gated service phases at Q_i before the server switches to the next queue. The aim of this discipline is to provide fairness (almost equal mean waiting times at the queues), while not giving up efficiency (weighted sum of mean waiting times). For the trade-off between these two we introduced the factor α. The κ_i's can then be

value of $\mathbb{E}(S)$ (inst.):	all (2925)		$[0.4, 1]$ (468)		$(1, 2]$ (585)		$(2, 3]$ (702)		$(3, 5.5]$ (585)		$(5.5, 10]$ (585)	
discipline \ averages	Δ	Σ	Δ	Σ	Δ	Σ	Δ	Σ	Δ	Σ	Δ	Σ
exhaustive	25.6	0.0	20.8	0.0	22.1	0.0	24.0	0.0	28.7	0.0	31.6	0.0
gated	6.2	9.6	2.4	1.9	3.4	4.1	4.9	7.1	8.4	14.0	11.2	19.8
elevator gg	0.0	11.9	0.0	2.4	0.0	5.1	0.0	8.8	0.0	17.4	0.0	24.6
κ-gat heur (round)	0.7	3.7	0.3	0.8	0.4	1.6	0.6	2.7	1.0	5.4	1.2	7.7
κ-gat heur (floor)	3.9	5.8	1.6	1.2	2.2	2.5	3.1	4.3	5.3	8.4	7.0	12.0
κ-gat heur (ceiling)	1.9	2.9	0.9	0.6	1.2	1.2	1.6	2.2	2.5	4.3	3.1	6.0
κ-gat opt $\alpha = 0$	0.3	4.1	0.1	0.8	0.2	1.7	0.3	3.0	0.5	5.9	0.6	8.4
κ-gat opt $\alpha = 1$	0.5	3.7	0.1	0.8	0.2	1.6	0.3	2.7	0.7	5.3	0.9	7.5
κ-gat opt $\alpha = 2$	2.0	2.7	0.1	0.8	0.3	1.5	0.6	2.6	3.4	3.7	5.6	4.5
κ-gat opt $\alpha = 5$	6.7	1.1	0.8	0.6	2.7	0.8	5.3	1.0	10.4	1.3	13.3	1.7
$\kappa = (2, 2)$	0.4	4.1	0.1	0.8	0.2	1.7	0.3	3.0	0.7	6.0	0.8	8.5

Table 9: Results for $N = 2$ split out according to $\mathbb{E}(S)$ (the mean total switchover time during a cycle), where $\mathbb{E}(S) \in [0.4, 10]$ for the testbed of Table 3.

optimized.

We showed how the mean visit times, the pseudo conservation law, the distribution of waiting times and the mean waiting times can be derived. We also derived the fluid limits. Further, using the fluid limits, we provided a heuristic to set the κ (not depending on α). In an extensive numerical study we showed that the heuristics perform very well. Typically when α is given, one can find (e.g. by an exhaustive search) a better setting, but the heuristic setting is robust against the value of α, that is, for all α it performs close to optimal. So, the factor α typically does not play a significant role in the choice of κ.

We have chosen here to set κ so as to optimize the fairness and efficiency. However, the κ-gated discipline can be used for other performance characteristics on the mean waiting times as well. Instead of the efficiency, one could for example consider the sum $\sum_{i=1}^{N} c_i \mathbb{E}(W_i)$, where each queue $i = 1, \ldots, N$ is assigned a cost factor c_i. This could e.g. reflect a difference in the importance of the customers in each queue.

An interesting option for further research is the handling of the fractional κ_i's. Instead of rounding, one might assign a probability, say p_i with which $\lfloor \kappa_i \rfloor$ phases are used, and otherwise $\lceil \kappa_i \rceil$. This, however, might lead to a more complicated exact analysis. Another question is in which order the queues should be placed, as to minimize the variance in waiting times or in the $\gamma(\alpha)$.

7 Acknowledgments

The authors would like to thank Marko Boon for assistance with the Mathematica implementation used in the numerical analysis, and for comments on an earlier version of this manuscript. The authors would also like to thank Erik Winands for his fruitful suggestion to use fluid heuristics.

References

[1] E. Altman, A. Khamisy, and U. Yechiali. On elevator polling with globally gated regime. *Queueing Systems*, 11(1):85–90, 1992.

[2] B. Avi-Itzhak, H. Levy, and D. Raz. Quantifying fairness in queuing systems. *Probability in the Engineering and Informational Sciences*, 22(04):495–517, 2008.

[3] J. Baker and I. Rubin. Polling with a General-Service Order Table. *IEEE Transactions on Communications*, 35(3):283–288, 1987.

[4] M.A.A. Boon, I.J.B.F. Adan, and O.J. Boxma. A Polling Model with Multiple Priority Levels. *Performance Evaluation*, 67(6):468–484, 2010.

[5] M.A.A. Boon, A.C.C. van Wijk, I.J.B.F. Adan, and O.J. Boxma. A Polling Model with Smart Customers. *Queueing Systems*, 66(3):1–36, 2010.

[6] O.J. Boxma. Polling systems. In *From universal morphisms to megabytes: A Baayen space odyssey. Liber amicorum for P.C. Baayen*, pages 215–230. CWI, Amsterdam, 1994.

[7] O.J. Boxma and W.P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability*, 24(4):949–964, 1987.

[8] O.J. Boxma, H. Levy, and J.A. Weststrate. Efficient visit frequencies for polling tables: minimization of waiting cost. *Queueing Systems*, 9(1):133–162, 1991.

[9] O.J. Boxma, H. Levy, and J.A. Weststrate. Efficient visit orders for polling systems. *Performance Evaluation*, 18(2):103–123, 1993.

[10] O.J. Boxma, H. Levy, and U. Yechiali. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research*, 35(3):187–208, 1992.

[11] O.J. Boxma, A.C.C. van Wijk, and I.J.B.F. Adan. Polling systems with a gated/exhaustive discipline. In *Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools*. ICST, 2008.

[12] C. Fricker and M.R. Jaibi. Monotonicity and stability of periodic polling models. *Queueing Systems*, 15(1):211–238, 1994.

[13] C.G. Park, D.H. Han, B. Kim, and H.S. Jun. Queueing analysis of symmetric polling algorithm for DBA scheme in an EPON. In B.D. Choi, editor, *Proceedings 1st Korea-Netherlands Joint Conference on Queueing Theory and its Applications to Telecommunication Systems*, pages 147–154, Korea University, Seoul, 2005.

[14] J.A.C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(4):409–426, 1993.

[15] R. Shoham and U. Yechiali. Elevator-type polling systems. *ACM Sigmetrics Performance Evaluation Review*, 20(1):255–257, 1992.

[16] H. Takagi. *Analysis of Polling Systems*. MIT Press Cambridge, MA, USA, 1986.

[17] R.D. van der Mei and J.A.C. Resing. Polling models with two-stage gated service: Fairness versus efficiency. *Lecture Notes in Computer Science*, 4516:544–555, 2007.

[18] R.D. van der Mei and A. Roubos. Polling systems with multi-phase gated service. *In Revision*, 2010.

[19] J.A. Weststrate. *Analysis and optimization of polling models*. PhD thesis, Catholic University of Brabant, Tilburg, the Netherlands, 1992.

[20] E.M.M. Winands, I.J.B.F. Adan, and G.-J. van Houtum. Mean value analysis for polling systems. *Queueing Systems*, 54(1):35–44, 2006.