EURANDOM PREPRINT SERIES
2010-050

Fairness and Efficiency for Polling Models with
the $\kappa$ 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 $\kappa$-Gated Service Discipline

A.C.C. van Wijk$^1$, I.J.B.F. Adan$^*$, 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 $\kappa$-gated service discipline. It is a hybrid of the classical gated and exhausted disciplines, and consists of using $\kappa_i$ 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 $\kappa_i$’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 $\kappa_i$’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
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. \cite{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 \cite{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 \( \kappa \)-gated service discipline.

The \( \kappa \)-gated discipline consists of using \( \kappa_i \) 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) \( \kappa_i \) 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 \( \kappa_i \) times, or until the queue becomes empty. The parameters \( \kappa_i \) 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 \( \kappa_i \)'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 $\kappa$-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, ... $N-1$, $N$, $N-1$, 2, 1, 2, ... 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: $E(W_i^{gat}) = (1 + \rho_i)E(R_C)$ respectively $E(W_i^{exh}) = (1 - \rho_i)E(R_C^*)$, where $\rho_i$ is the workload. Furthermore $E(R_C)$ and $E(R_C^*)$ denote the mean residual cycle duration from the visit completion respectively beginning of $Q_i$ on. These can be approximated by $E(R_C) \approx E(R_C) \approx E(R_C^*)$. From the resulting approximations for $E(W_i^{gat})$ and $E(W_i^{exh})$, one might expect the mean waiting time in the Gated/Exhaustive discipline to become $E(W_i^{g/e}) \approx E(R_C)$, 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 $\kappa$-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 $\kappa$-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 $\kappa_i$'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 $\kappa$. 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 $\kappa$ 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, $Q_1, \ldots, Q_N$, where each queue has infinite capacity. The queues are served by a single server, in fixed cyclic order $Q_1, Q_2, \ldots, Q_N, Q_1, Q_2, \ldots$. 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 $\lambda_i$ at $Q_i$, $i = 1, \ldots, N$. 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 $\beta_i(\cdot)$. 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 $\sigma_i(\cdot)$, 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 $E(S) > 0$. 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 $\rho_i$ and is given by $\rho_i = \lambda_i E(B_i)$. 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 $E(C) = E(S)/(1 - \rho)$. $E(C)$ does not depend on the service disciplines at the queues.

We now describe the $\kappa$-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) $\kappa_i$ 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 $\kappa$-gated service discipline.

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

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

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:

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

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:

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

for some $\alpha \in [0, \infty)$. Note that (1) depends on the service discipline at each of the queues. Under the $\kappa$-gated discipline, for a given $\alpha$, the $\kappa$ 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 $\kappa$-Gated Discipline

In this section we present the analysis of the $\kappa$-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 $\kappa$.

3.1 Mean Visit Times

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

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

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:

$$E(V_i^k) = \rho_i E(V_i^{k-1})$$
$$= \rho_i^{k-1} E(V_i^1), \quad \text{for } k = 2, \ldots, \kappa_i.$$

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

$$E(V_i^1) = \rho_i [\rho_i^{\kappa_i-1} E(V_i^1) + (1 - \rho_i) E(C)].$$
Solving this leads to $E(V^k_i) = \rho_i \frac{1 - \rho_i}{1-\rho_i} E(C)$, and hence
\[ E(V^k_i) = \rho_i^k \frac{1 - \rho_i}{1-\rho_i} E(C), \quad k = 1, \ldots, \kappa_i. \] (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:
\[ \sum_{k=1}^{\kappa_i} E(V^k_i) + E(I_i) = E(C), \quad i = 1, \ldots, N. \]

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: $\sum_{i=1}^{N} \rho_i E(W_i)$. 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)]:
\[ \sum_{i=1}^{N} \rho_i E(W_i) = \frac{\rho}{1 - \rho} \sum_{i=1}^{N} \rho_i E(R_{B_i}) + \rho E(R_S) + \frac{E(S)}{2(1-\rho)} \left( \rho^2 - \sum_{i=1}^{N} \rho_i^2 \right) + \sum_{i=1}^{N} E(M_i), \] (4)
where $E(M_i)$ 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 $E(M_i^{exh})$ trivially equals zero (cf. [7, (3.11)]), and for gated it holds that $E(M_i^{gat}) = \rho_i E(V_i) = \rho_i^2 E(S)/(1-\rho)$ (cf. [7, (3.12)]). The workload decomposition result in [7] is also valid for the $\kappa$-gated discipline, and we find, using (3):
\[ E(M_i^{\kappa-gat}) = \rho_i E(V^{\kappa_i}) = \rho_i^{\kappa_i+1} \frac{1 - \rho_i}{1-\rho_i} \frac{E(S)}{1-\rho}. \]
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 $E(M_i)$ terms for the different strategies, we find the following:
\[ 0 = E(M_i^{exh}) \leq E(M_i^{\kappa-gat}) \leq E(M_i^{gat}), \]
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 $(\kappa)$-gated. Gated (i.e. $\kappa_i = 1$) is less efficient than $\kappa$-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 $\kappa$-gated discipline is always between that of exhaustive and gated.
By substituting the expression for \( E(M_{\kappa-gat}^i) \) into (4) we find the pseudo conservation law for the \( \kappa \)-gated discipline:

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

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

\[
\gamma(\alpha) := \max_{i,j} (E(W_i) - E(W_j)) + \alpha \sum_{i=1}^{N} E(M_i),
\]

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 \( \kappa \)-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:

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

For \( \kappa \)-gated we can recursively express \( h_i \) as follows:

\[
h_i^{(1-gated)}(z) = h_i^{(gated)}(z),
\]

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

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

\[
h_i^{(\infty-gated)}(z) = h_i^{(exhaustive)}(z) = \theta_i \left( \sum_{j=1, j \neq i}^{N} \lambda_j (1 - z_j) \right),
\]
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 $V_{b_i}(z)$ in itself, by repeated application of the following relation, cf. [4, (2.2)]:

$$V_{b_{i+1}}(z) = V_{c_i}(1 - \sigma_i \left( \sum_{j=1}^{N} \lambda_j (1 - z_j) \right))$$

$$= V_{b_i}(z_1, \ldots, z_{i-1}, h_i^{(\kappa_i\text{-gated})}(z), z_{i+1}, \ldots, z_N) \sigma_i \left( \sum_{j=1}^{N} \lambda_j (1 - z_j) \right), \ i = 1, 2, \ldots, N,$$

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)]:

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

where $\bar{V}_{b_i}(\cdot)$ is the pgf of the steady-state marginal queue length distribution at a visit beginning of $Q_i$, given by $\bar{V}_{b_i}(z) = V_{b_i}(1, \ldots, 1, z, 1, \ldots, 1)$, with $z$ as the $i$th argument, and $\bar{V}_{c_i}(\cdot)$ 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, $E(W_i)$, can easily be obtained in a more efficient way. For this, we show that the $\kappa$-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 $\lambda_i$ 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 $\kappa$-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 \( \kappa \)-gated model, arriving customers queue behind a gate, which only opens when the server starts one of the \( \kappa_i \) serving phases. In the corresponding model, each of these phases now becomes a separate queue. Hence, we create a polling model with \( \kappa_i \) 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 \( Q_i^{(1)} \)), becomes:

\[
V_i^{(1)} - V_i^{(2)} - \ldots - V_i^{(\kappa_i)} - S_i - V_i^{(2)} - V_i^{(3)} - \ldots - V_i^{(\kappa_i)} - S_2 - \ldots - S_{N-1} - V_i^{(1)} - \ldots - V_i^{(\kappa_i)} - S_N.
\]

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 \( O\left(N^2\right) \) 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 \( O\left((\sum_i \kappa_i)^2\right) \) 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 \( \kappa \)-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 \( \kappa_i \)'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 \( \kappa_i \)'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 \( \rho_i \), and
during the visit time work is removed at rate 1. So, during the intervisit time of mean length $E(I_i) = (1 - \rho_i)E(C)$, the amount of work increases at rate $\rho_i$, and during the visit time, with mean length $E(V_i) = \rho_iE(C)$, 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 $\kappa$-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 $\rho_iE(V_i^{\kappa_i}) =: m$. 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 $m + \rho_iE(I_i) =: M$. Consequently, the average fluid level during a cycle, i.e. the mean workload $E(H_i)$, is given by:

$$E(H_i) = \frac{m + M}{2} = m + \frac{\rho_iE(I_i)}{2} = (1 + \rho_i^{\kappa_i})\frac{\rho_i(1 - \rho_i)}{2(1 - \rho_i^{\kappa_i})}E(C).$$

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

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

Figure 2 shows these fluid limits for different $\kappa_i$.

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

We now want to choose $\kappa$ 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 $\kappa$ 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 $\kappa_i$. However, we use the fluid limits (6) as approximation for the mean waiting times in the optimization:

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

where

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

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

$$\min_\kappa \sum_{i=1}^N E(M_i^{\kappa-gat}),$$

such that $E(W_i^{\text{fluid}}) = \ldots = E(W_N^{\text{fluid}})$. For the moment we allow the $\kappa_i$’s to be fractional, later we round them to integers. Note that the problem in (8) does not depend on $\alpha$. 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 $E(W_{1}^{\text{fluid}}) = E(W_{2}^{\text{fluid}})$ for $\kappa_2$ in terms of $\kappa_1, \rho_1$ and $\rho_2$:

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

where we have divided by $E(C) \neq 0$. Solving for $\kappa_2$, denoted by $\kappa_2^{\text{opt}}$, gives:

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

So, this $\kappa_2$ achieves optimal fairness (recall that we allowed $\kappa_2$ to be fractional). Using this $\kappa_2$, we now optimize the efficiency, i.e. we minimize:

$$\sum_{i=1}^{2} E(M_i) = \rho_1^{\kappa_1 + 1} \frac{1 - \rho_1}{1 - \rho_1^{\kappa_1}} + \rho_2^{\kappa_2^{\text{opt}} + 1} \frac{1 - \rho_2}{1 - \rho_2^{\kappa_2^{\text{opt}}}}$$

$$= \frac{(\rho_1 - \rho_2)\rho_2 + \rho_1^{\kappa_1}(2(1 - \rho_1)\rho_1 + (2 - \rho_1)\rho_2 - \rho_2^2)}{2(1 - \rho_1^{\kappa_1})}. \quad (10)$$

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), $\kappa_2^{\text{opt}}$ becomes:

$$\kappa_2^{\text{opt}} = \log_{\rho_2} \frac{\rho_2 - \rho_1}{2 - \rho_1 - \rho_2}. \quad (11)$$

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 $\kappa_1$ and $\kappa_2$:

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

In order to get integer $\kappa$’s, we have three possibilities: rounding to the nearest integer denoted by $\lceil x \rceil$; 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 $\kappa$ set according to the heuristic by $[\kappa]$, $[\kappa]$ 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 $\kappa_2$. From the figure it becomes clear that $\kappa_2 = 2$ almost always is a proper choice.
Figure 3: Optimal value of $\kappa_2$ (rounded to the nearest integer), given by $\kappa_2^{opt} = \left[ \log_{\rho_2} \frac{\rho_2 - \rho_1}{2 - \rho_1 - \rho_2} \right]$, for $\rho_1 < \rho_2$ and $\rho_1 + \rho_2 < 1$.

### 4.2 $N$ queues

For $N$ queues we first determine which $\kappa_i$’s give equal mean waiting times. We solve $\mathbb{E}(W_i^{\text{fluid}}) = \mathbb{E}(W_j^{\text{fluid}})$ 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 $\kappa_1$, and on all $\rho_i$’s. It only makes sense if $\rho_1$ is the smallest of all $\rho_i$, 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 $\kappa_i$’s, $i = 1, \ldots, N$:

\[
\begin{cases}
\text{For all } i \text{ such that } i = \arg \min \rho_i, \text{ let } \kappa_i = \infty; \\
\text{For all } j = 1, 2, \ldots, N \text{ where } j \neq i, \text{ let } \kappa_j = \log_{\rho_j} \frac{\rho_j - \rho_1}{2 - \rho_1 - \ldots - \rho_N}.
\end{cases}
\]

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

14
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 $\kappa$. For each instance we determine $\gamma(\alpha)$ as defined in (5).

For brevity of notation we define:

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

We compare the results of the $\kappa$-gated discipline with the elevator strategy in a globally gated regime, cf. [1, 15]. For this strategy all mean waiting times are equal:

$$
E(W_{elev.GG}^1) = E(W_{elev.GG}^2) = \ldots = E(W_{elev.GG}^N),
$$

and given by, cf. [1, (6), (10)]:

$$
E(W_{elev.GG}^1) = \frac{1}{\lambda_1} - \sum_{i=1}^{N} \rho_i E(R_{B_i}) + E(R_S) + \frac{1 + \rho}{2(1-\rho)} E(S).
$$

The PCL easily follows:

$$
\sum_{i=1}^{N} \rho_i E(W_i) = \rho E(W_1).
$$

Using (4) we then derive that:

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

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 $\log_{\rho_1} \frac{\rho_1 - \rho_2}{\rho_1 - \rho_2} \approx 2.15$. Hence the heuristic settings are $[\kappa] = [\kappa] = (2, \infty)$ and $[\kappa] = (3, \infty)$.

For the $\kappa$-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 $\kappa$ perform quite well. Although suboptimal for small $\alpha$, the performance seems to be rather robust with respect to $\alpha$. Despite $\kappa = (2, 2)$ performs better in this example for the four values of $\alpha$ chosen, it follows from the values of $\Delta$ and $\Sigma$ that for large $\alpha$ the heuristic setting will dominate in performance. In general, however, the heuristic settings outperform the $(2, 2)$ (unless $\alpha$ is small, i.e. less than 1), as the numerical study in Section 5.2 shows.

The difference $\Delta$ 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 $\kappa$-gated discipline closely resembles the elevator strategy in a globally gated regime (cf. [1, 15]). In this discipline the visit order is $1, 2, \ldots, N-1, N, N, N,$
leads to almost gated regime leads to $E$ where the gate is closed when each service phase starts. As the elevator strategy in a globally
the queues are served as:

Now consider the following setting. Again we have $N = 2$ queues, the queues are served as:

<table>
<thead>
<tr>
<th>$\kappa_1$</th>
<th>$\kappa_2$</th>
<th>$E(W_1)$</th>
<th>$E(W_2)$</th>
<th>$\Delta$</th>
<th>$\Sigma$</th>
<th>$\gamma(0)$</th>
<th>$\gamma(1)$</th>
<th>$\gamma(2)$</th>
<th>$\gamma(5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>12.77</td>
<td>9.69</td>
<td>3.08</td>
<td>4.00</td>
<td>3.1</td>
<td>7.1</td>
<td>11.1</td>
<td>23.1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8.42</td>
<td>11.49</td>
<td>3.07</td>
<td>1.75</td>
<td>3.1</td>
<td>4.8</td>
<td>6.6</td>
<td>11.8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>6.90</td>
<td>12.60</td>
<td>5.70</td>
<td>1.06</td>
<td>5.7</td>
<td>6.8</td>
<td>7.8</td>
<td>11.0</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1</td>
<td>5.04</td>
<td>14.88</td>
<td>9.84</td>
<td>0.40</td>
<td>9.8</td>
<td>10.2</td>
<td>10.6</td>
<td>11.8</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>13.00</td>
<td>7.21</td>
<td>5.83</td>
<td>3.67</td>
<td>5.8</td>
<td>9.5</td>
<td>13.2</td>
<td>24.2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8.77</td>
<td>8.77</td>
<td>0.0038</td>
<td>1.42</td>
<td>0.0</td>
<td>1.4</td>
<td>2.8</td>
<td>7.1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7.27</td>
<td>9.82</td>
<td>2.55</td>
<td>0.73</td>
<td>2.6</td>
<td>3.3</td>
<td>4.0</td>
<td>6.2</td>
</tr>
<tr>
<td>$\infty$</td>
<td>2</td>
<td>5.38</td>
<td>12.20</td>
<td>6.82</td>
<td>0.07</td>
<td>6.8</td>
<td>6.9</td>
<td>7.0</td>
<td>7.2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>13.10</td>
<td>6.76</td>
<td>6.34</td>
<td>3.61</td>
<td>6.3</td>
<td>10.0</td>
<td>13.6</td>
<td>24.4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>8.85</td>
<td>8.25</td>
<td>0.60</td>
<td>1.36</td>
<td>0.6</td>
<td>2.0</td>
<td>3.3</td>
<td>7.4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>7.36</td>
<td>9.29</td>
<td>1.92</td>
<td>0.67</td>
<td>1.9</td>
<td>2.6</td>
<td>3.3</td>
<td>5.3</td>
</tr>
<tr>
<td>$\infty$</td>
<td>3</td>
<td>5.47</td>
<td>11.66</td>
<td>6.19</td>
<td>0.01</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2</td>
<td>6.3</td>
</tr>
</tbody>
</table>

$\nu : [\kappa] : [\kappa] :$

$\kappa :$

<table>
<thead>
<tr>
<th>$[\kappa]$</th>
<th>$[\kappa]$ :</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$ :</td>
<td>$\infty$</td>
</tr>
<tr>
<td>1</td>
<td>13.12</td>
</tr>
<tr>
<td>2</td>
<td>8.88</td>
</tr>
<tr>
<td>3</td>
<td>7.39</td>
</tr>
<tr>
<td>$\infty$</td>
<td>5.50</td>
</tr>
</tbody>
</table>

Elev.GG 15.00 15.00 0.00 6.00 0.00 6.00 12.0 30.0

Table 1: Results for Example 1 for the $\kappa$-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, . . . , 2, 1, 1, 2, . . . , and all gates are closed when turning around at 1 and at $N$. Hence, for $N = 2$ the queues are served as:

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

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

... $- Q^{(1)}_1 - Q^{(2)}_1 - S_1 - Q^{(1)}_2 - Q^{(2)}_2 - S_2 - Q^{(1)}_1 - Q^{(2)}_1 - S_1 - \ldots - \ldots$

where the gate is closed when each service phase starts. As the elevator strategy in a globally gated regime leads to $E(W_1) = E(W_2)$, 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 $[\kappa] = [\kappa] = (3, \infty)$ and $[\kappa] = (2, \infty)$. The results are given in Table 2. The heuristic setting of $\kappa$ performs very well, and is even optimal for $\alpha = 1, 2$, and 5. Note that $\Delta$ is again small for $\kappa = (2, 2)$. 

16
In a numerical experiment we study the performance of the heuristic settings for the $\kappa$.

**Performance of fluid based heuristic**

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

<table>
<thead>
<tr>
<th>$\kappa_1$</th>
<th>$\kappa_2$</th>
<th>$E(W_1)$</th>
<th>$E(W_2)$</th>
<th>$\Delta$</th>
<th>$\Sigma$</th>
<th>$\gamma(0)$</th>
<th>$\gamma(1)$</th>
<th>$\gamma(2)$</th>
<th>$\gamma(5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>9.30</td>
<td>8.68</td>
<td>0.63</td>
<td>1.85</td>
<td>0.6</td>
<td>2.5</td>
<td>4.3</td>
<td>9.9</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6.36</td>
<td>9.16</td>
<td>2.80</td>
<td>0.94</td>
<td>2.8</td>
<td>3.7</td>
<td>4.7</td>
<td>7.5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5.60</td>
<td>9.37</td>
<td>3.77</td>
<td>0.73</td>
<td>3.8</td>
<td>4.5</td>
<td>5.2</td>
<td>7.4</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1</td>
<td><strong>5.19</strong></td>
<td>9.53</td>
<td>4.33</td>
<td>0.63</td>
<td>4.3</td>
<td>5.0</td>
<td>5.6</td>
<td>7.5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9.57</td>
<td>6.30</td>
<td>3.28</td>
<td>1.35</td>
<td>3.3</td>
<td>4.6</td>
<td>6.0</td>
<td>10.0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6.65</td>
<td>6.77</td>
<td>0.12</td>
<td>0.44</td>
<td>0.1</td>
<td>0.6</td>
<td>1.0</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5.87</td>
<td>6.98</td>
<td>1.11</td>
<td>0.23</td>
<td>1.1</td>
<td>1.3</td>
<td>1.6</td>
<td>2.2</td>
</tr>
<tr>
<td>$\infty$</td>
<td>2</td>
<td>5.46</td>
<td>7.16</td>
<td>1.71</td>
<td>0.13</td>
<td>1.7</td>
<td>1.8</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9.66</td>
<td>5.80</td>
<td>3.86</td>
<td>1.25</td>
<td>3.9</td>
<td>5.1</td>
<td>6.4</td>
<td>10.1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6.74</td>
<td>6.25</td>
<td>0.49</td>
<td>0.35</td>
<td>0.5</td>
<td>0.8</td>
<td>1.2</td>
<td>2.2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>5.97</td>
<td>6.47</td>
<td>0.50</td>
<td>0.13</td>
<td>0.5</td>
<td>0.6</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>$\infty$</td>
<td>3</td>
<td>5.55</td>
<td>6.65</td>
<td>1.10</td>
<td>0.03</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>1</td>
<td>$\infty$</td>
<td>9.70</td>
<td><strong>5.63</strong></td>
<td>4.07</td>
<td>1.23</td>
<td>4.1</td>
<td>5.3</td>
<td>6.5</td>
<td>10.2</td>
</tr>
<tr>
<td>2</td>
<td>$\infty$</td>
<td>6.79</td>
<td>6.07</td>
<td>0.72</td>
<td>0.32</td>
<td>0.7</td>
<td>1.0</td>
<td>1.4</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>$\infty$</td>
<td>6.02</td>
<td>6.28</td>
<td>0.26</td>
<td>0.10</td>
<td>0.3</td>
<td><strong>0.4</strong></td>
<td><strong>0.5</strong></td>
<td><strong>0.8</strong></td>
</tr>
<tr>
<td>$\infty$</td>
<td>$\infty$</td>
<td>5.60</td>
<td>6.46</td>
<td>0.87</td>
<td><strong>0.00</strong></td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>

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 $\kappa$-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 $\kappa_i$'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;
- $\kappa$-gated with $\kappa$ as in the heuristic (cf. (12));
- $\kappa$-gated with $\kappa$ 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 $\kappa$-gated discipline, we optimize the $\kappa$. 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
<table>
<thead>
<tr>
<th>$N$</th>
<th>settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2,925</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
</tr>
<tr>
<td>4</td>
<td>552</td>
</tr>
<tr>
<td>5</td>
<td>894</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2,925</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
</tr>
<tr>
<td>4</td>
<td>552</td>
</tr>
<tr>
<td>5</td>
<td>894</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>discipline \ averages</th>
<th>$E(W_1)$</th>
<th>$E(W_2)$</th>
<th>$\Delta$</th>
<th>$\Sigma$</th>
<th>$\gamma(0)$</th>
<th>$\gamma(1)$</th>
<th>$\gamma(2)$</th>
<th>$\gamma(5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁ exh - Q₂ exh</td>
<td>9.8</td>
<td>30.7</td>
<td>25.6</td>
<td>0.0</td>
<td>25.6</td>
<td>25.6</td>
<td>25.6</td>
<td>25.6</td>
</tr>
<tr>
<td>Q₁ exh - Q₂ gat</td>
<td>7.5</td>
<td>35.8</td>
<td>28.2</td>
<td>1.7</td>
<td>28.2</td>
<td>29.9</td>
<td>31.6</td>
<td>36.7</td>
</tr>
<tr>
<td>Q₁ gat - Q₂ exh</td>
<td>21.7</td>
<td>10.5</td>
<td>11.3</td>
<td>7.9</td>
<td>11.3</td>
<td>19.1</td>
<td>27.0</td>
<td>50.6</td>
</tr>
<tr>
<td>Q₁ gat - Q₂ gat</td>
<td>19.5</td>
<td>15.2</td>
<td>6.2</td>
<td>9.6</td>
<td>6.2</td>
<td>15.7</td>
<td>25.3</td>
<td>54.0</td>
</tr>
<tr>
<td>elevator g₂</td>
<td>22.7</td>
<td>22.7</td>
<td>0.0</td>
<td>11.9</td>
<td>11.9</td>
<td>23.8</td>
<td>59.5</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (round)</td>
<td>13.2</td>
<td>12.9</td>
<td>0.7</td>
<td>4.4</td>
<td>0.7</td>
<td>4.4</td>
<td>8.1</td>
<td>19.3</td>
</tr>
<tr>
<td>$\kappa$-gat heur (floor)</td>
<td>15.3</td>
<td>12.9</td>
<td>3.9</td>
<td>5.8</td>
<td>3.9</td>
<td>9.7</td>
<td>15.4</td>
<td>32.8</td>
</tr>
<tr>
<td>$\kappa$-gat heur (ceiling)</td>
<td>12.3</td>
<td>13.5</td>
<td>1.9</td>
<td>2.9</td>
<td>1.9</td>
<td>4.8</td>
<td>7.7</td>
<td>16.4</td>
</tr>
<tr>
<td>$\kappa = (2, 2)$</td>
<td>13.5</td>
<td>13.6</td>
<td>0.4</td>
<td>4.1</td>
<td>0.4</td>
<td>4.5</td>
<td>8.6</td>
<td>20.9</td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 0$</td>
<td>13.5</td>
<td>13.5</td>
<td>0.3</td>
<td>4.1</td>
<td>0.3</td>
<td>4.4</td>
<td>8.5</td>
<td>20.6</td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 1$</td>
<td>13.1</td>
<td>13.2</td>
<td>0.5</td>
<td>3.7</td>
<td>0.5</td>
<td>4.1</td>
<td>7.8</td>
<td>18.7</td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 2$</td>
<td>12.0</td>
<td>13.5</td>
<td>2.0</td>
<td>2.7</td>
<td>2.0</td>
<td>4.7</td>
<td>7.4</td>
<td>15.4</td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 5$</td>
<td>10.4</td>
<td>15.3</td>
<td>6.7</td>
<td>1.1</td>
<td>6.7</td>
<td>7.8</td>
<td>8.9</td>
<td>12.2</td>
</tr>
</tbody>
</table>

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 $\kappa$ by exhaustive search over $\kappa_i \in \{1, 2, 3, 4, 5, 6, \infty\}$, $i = 1, 2$. 

18
Table 6: Results of numerical study for $N = 3$: average values over 243 cases (as described in Table 3). Optimization of $\kappa$ by exhaustive search over $\kappa_i \in \{1, 2, 3, \infty\}$, $i = 1, 2, 3$.

<table>
<thead>
<tr>
<th>discipline \ averages</th>
<th>$E(W_1)$</th>
<th>$E(W_2)$</th>
<th>$E(W_3)$</th>
<th>$E(W_4)$</th>
<th>$\Delta$</th>
<th>$\Sigma$</th>
<th>$\gamma(0)$</th>
<th>$\gamma(1)$</th>
<th>$\gamma(2)$</th>
<th>$\gamma(5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>exhaustive</td>
<td>11.2</td>
<td>12.2</td>
<td>12.3</td>
<td>6.2</td>
<td>0.0</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2</td>
<td>6.2</td>
</tr>
<tr>
<td>gated</td>
<td>16.0</td>
<td>15.3</td>
<td>15.4</td>
<td>4.2</td>
<td>5.6</td>
<td>4.2</td>
<td>9.9</td>
<td>15.5</td>
<td>32.4</td>
<td></td>
</tr>
<tr>
<td>elevator gg</td>
<td>21.4</td>
<td>21.4</td>
<td>21.4</td>
<td>0.0</td>
<td>10.2</td>
<td>0.0</td>
<td>10.2</td>
<td>20.3</td>
<td>59.8</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (round)</td>
<td>12.2</td>
<td>12.1</td>
<td>12.2</td>
<td>0.7</td>
<td>1.6</td>
<td>0.7</td>
<td>2.4</td>
<td>4.0</td>
<td>8.9</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (floor)</td>
<td>14.2</td>
<td>13.8</td>
<td>13.8</td>
<td>6.0</td>
<td>4.9</td>
<td>6.0</td>
<td>10.9</td>
<td>15.8</td>
<td>30.5</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (ceiling)</td>
<td>12.4</td>
<td>12.1</td>
<td>12.2</td>
<td>1.2</td>
<td>1.6</td>
<td>1.2</td>
<td>2.8</td>
<td>4.4</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 0$</td>
<td>12.5</td>
<td>12.7</td>
<td>10.5</td>
<td>2.4</td>
<td>1.2</td>
<td>2.4</td>
<td>3.6</td>
<td>4.9</td>
<td>8.6</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 1$</td>
<td>12.2</td>
<td>12.6</td>
<td>11.1</td>
<td>2.3</td>
<td>1.2</td>
<td>2.3</td>
<td>3.5</td>
<td>4.7</td>
<td>8.3</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 2$</td>
<td>11.6</td>
<td>12.9</td>
<td>11.6</td>
<td>3.3</td>
<td>0.9</td>
<td>3.2</td>
<td>4.2</td>
<td>5.0</td>
<td>7.7</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat opt $\alpha = 5$</td>
<td>10.8</td>
<td>13.2</td>
<td>12.0</td>
<td>4.5</td>
<td>0.6</td>
<td>4.5</td>
<td>5.1</td>
<td>6.7</td>
<td>7.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: 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.

<table>
<thead>
<tr>
<th>discipline \ averages</th>
<th>$E(W_1)$</th>
<th>$E(W_2)$</th>
<th>$E(W_3)$</th>
<th>$E(W_4)$</th>
<th>$E(W_5)$</th>
<th>$\Delta$</th>
<th>$\Sigma$</th>
<th>$\gamma(0)$</th>
<th>$\gamma(1)$</th>
<th>$\gamma(2)$</th>
<th>$\gamma(5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>exhaustive</td>
<td>20.3</td>
<td>22.4</td>
<td>22.4</td>
<td>22.6</td>
<td>9.7</td>
<td>0.0</td>
<td>9.7</td>
<td>9.7</td>
<td>9.7</td>
<td>9.7</td>
<td>9.7</td>
</tr>
<tr>
<td>gated</td>
<td>30.6</td>
<td>29.0</td>
<td>29.0</td>
<td>29.1</td>
<td>8.0</td>
<td>11.1</td>
<td>8.0</td>
<td>19.0</td>
<td>30.1</td>
<td>63.2</td>
<td></td>
</tr>
<tr>
<td>elevator gg</td>
<td>43.5</td>
<td>43.5</td>
<td>43.5</td>
<td>43.5</td>
<td>0.0</td>
<td>23.0</td>
<td>0.0</td>
<td>23.0</td>
<td>46.0</td>
<td>115.0</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (round)</td>
<td>22.9</td>
<td>23.1</td>
<td>23.2</td>
<td>23.2</td>
<td>2.3</td>
<td>2.9</td>
<td>2.3</td>
<td>5.2</td>
<td>8.2</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (floor)</td>
<td>29.5</td>
<td>27.6</td>
<td>27.6</td>
<td>27.6</td>
<td>11.6</td>
<td>10.5</td>
<td>11.6</td>
<td>22.0</td>
<td>32.5</td>
<td>63.8</td>
<td></td>
</tr>
<tr>
<td>$\kappa$-gat heur (ceiling)</td>
<td>22.9</td>
<td>23.1</td>
<td>23.2</td>
<td>23.2</td>
<td>2.3</td>
<td>2.9</td>
<td>2.3</td>
<td>5.2</td>
<td>8.2</td>
<td>17.0</td>
<td></td>
</tr>
</tbody>
</table>

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 $\alpha$ near zero as well. The exhaustive strategy, leading to $\Sigma = 0$ (maximal efficiency), hence would be optimal for large values of $\alpha$. The $\kappa$-gated discipline, using the heuristic settings for $\kappa$, seems to perform very well in the
range of $\alpha$’s in between. For a specific $\alpha$ (and specific setting of the parameters), one can typically find a better performing $\kappa$, 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 $\alpha$’s. For $N = 2$ it outperforms $(2, 2)$ for all $\alpha$ except for $\alpha$ 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 $\kappa$-gated discipline already for small $\alpha$, by all indicated choices for the setting of the $\kappa$ (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 $\kappa$-gated discipline with $\kappa$ set according to the heuristic, either rounding or ceiling, performs very well. It is robust against the setting of $\alpha$ and it performs well over a wide range of values for $\mathbb{E}(S)$.

### 6 Conclusion

We introduced the $\kappa$-gated service discipline for a polling model. It is a hybrid of the classical gated and exhausted disciplines, and consists of using $\kappa_i$ 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 $\alpha$. The $\kappa_i$’s can then be

<table>
<thead>
<tr>
<th>discipline \ averages</th>
<th>$\Delta$</th>
<th>$\Sigma$</th>
<th>$\gamma(0)$</th>
<th>$\gamma(1)$</th>
<th>$\gamma(2)$</th>
<th>$\gamma(5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>exhaustive</td>
<td>19.1</td>
<td><strong>0.0</strong></td>
<td>19.1</td>
<td>19.1</td>
<td>19.1</td>
<td>19.1</td>
</tr>
<tr>
<td>gated</td>
<td>6.4</td>
<td>9.3</td>
<td>6.4</td>
<td>15.6</td>
<td>24.9</td>
<td>52.7</td>
</tr>
<tr>
<td>elevator gg</td>
<td><strong>0.0</strong></td>
<td>14.7</td>
<td><strong>0.0</strong></td>
<td>14.7</td>
<td>29.5</td>
<td>73.7</td>
</tr>
<tr>
<td>$\kappa$-gat heur (round)</td>
<td>1.2</td>
<td>3.2</td>
<td>1.2</td>
<td><strong>4.4</strong></td>
<td>7.6</td>
<td>17.1</td>
</tr>
<tr>
<td>$\kappa$-gat heur (floor)</td>
<td>5.9</td>
<td>6.7</td>
<td>5.9</td>
<td>12.6</td>
<td>19.3</td>
<td>39.5</td>
</tr>
<tr>
<td>$\kappa$-gat heur (ceiling)</td>
<td>1.9</td>
<td>2.6</td>
<td>1.9</td>
<td>4.5</td>
<td><strong>7.2</strong></td>
<td><strong>15.0</strong></td>
</tr>
</tbody>
</table>

*Table 8: Results over all 4,614 instances.*
Table 9: Results for $N = 2$ split out according to $E(S)$ (the mean total switchover time during a cycle), where $E(S) \in [0.4,10]$ for the testbed of Table 3.

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 $\kappa$ (not depending on $\alpha$). In an extensive numerical study we showed that the heuristics perform very well. Typically when $\alpha$ is given, one can find (e.g. by an exhaustive search) a better setting, but the heuristic setting is robust against the value of $\alpha$, that is, for all $\alpha$ it performs close to optimal. So, the factor $\alpha$ typically does not play a significant role in the choice of $\kappa$.

We have chosen here to set $\kappa$ so as to optimize the fairness and efficiency. However, the $\kappa$-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 $\sum_{i=1}^{N} c_i E(W_i)$, 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 $\kappa_i$’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