Polling systems with a Gated/Exhaustive discipline*

O.J. Boxma
EURANDOM and Dept. of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
boxma@win.tue.nl

A.C.C. van Wijk
EURANDOM and Dept. of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
a.c.c.v.wijk@tue.nl

I.J.B.F. Adan
EURANDOM and Dept. of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
iadan@win.tue.nl

ABSTRACT

We consider a polling system where the server cyclically serves the queues according to the following discipline: the server does one round of visits to the queues applying the gated service discipline at each of the queues, followed by one round of visits applying the exhaustive service discipline at each of the queues, and this alternating pattern repeats itself. We call this the Gated/Exhaustive service discipline. For this we derive (i) a Pseudo Conservation Law for the weighted sum of the mean waiting times, (ii) the mean steady state waiting times using Mean Value Analysis, (iii) queue length distributions making use of results for Multitype Branching Processes and the concept of so-called Smart Customers, and (iv) the sojourn time distributions.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Queueing theory

General Terms

Performance, Theory

Keywords

polling systems, gated service discipline, exhaustive service discipline, mean value analysis, multitype branching processes

1. INTRODUCTION

The classical polling system is a queueing system with multiple queues and one single server. The server cyclically visits all queues, where it serves the customers. Typically a so-called switchover time is incurred when the server switches from one queue to another. There are many possible choices for deciding when the server should switch to the next queue. Those most often studied 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 present paper considers the following variant of this classical model. First the server cyclically serves all queues according to the gated service discipline, and then the server cyclically serves all queues according to the exhaustive service discipline, and this alternating pattern repeats itself. We call this the *Gated/Exhaustive discipline*.

We present a detailed analysis of this model. We first aim for mean values, deriving both a *Pseudo Conservation Law* for the mean waiting times and exploiting the *Mean Value Analysis* (MVA) technique, that was recently [16] developed for polling systems, to obtain all mean waiting times. Subsequently we relate the joint queue length process to *Multitype Branching Processes*, by using the concept of *'smart customers'* (and basically doubling the number of queues in the system). The latter concept may be of independent interest, as it allows us to model and analyze a rather large class of polling systems. After having obtained the joint queue length distribution at server polling epochs, we use that result to derive the sojourn time distributions at all queues.

Our work was partly motivated by the question whether an alternation of gated and exhaustive cycles might lead to *fairness* to the queues, in the sense of having almost identical mean waiting times. Fairness is a topic that has frequently played a role in the choice of service discipline in polling systems. In some recent studies [12, 15] of dynamic bandwidth allocation of Ethernet Passive Optical Networks (EPON), polling models have been considered with two-stage gated service: 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. We conjectured that an alternation of gated and exhaustive cycles might also lead to small differences between mean waiting times (but with smaller delays than for two-stage

*The research was done in the framework of the BSIK/BRICKS project, and of the European Network of Excellence Euro-FGI.

ValueTools 2008, October 21 – 23, 2008, Athens, GREECE.

gated), as various mean waiting time approximations (see, e.g., [7]) have been based on the following facts: (i) for gated service, the mean waiting time at the ith queue is $(1+\rho_i)$ times the mean residual length of one server cycle for the ith queue, with ρ_i the traffic load at the ith queue; (ii) for exhaustive service, the mean waiting time is roughly $(1-\rho_i)$ times the mean residual length of one server cycle. Averaging with equal weights would yield roughly equal mean waiting times at all queues; our numerical results, however, will show that there can still be substantial differences between mean waiting times.

The structure of this paper is as follows. In Section 2 we introduce the model in more detail and give the notation that is used. In Section 3 we derive the mean visit times of the server at each of the queues. The Pseudo Conservation Law for the mean waiting times is derived in Section 4. After that, in Section 5, we use Mean Value Analysis to determine the individual mean waiting times. In Section 6 we use Multitype Branching Processes and the concept of smart customers to obtain exact expressions for the queue length distributions at server polling epochs, as well as the sojourn time distribution at each queue. We end with a discussion of possible further work in Section 7.

2. MODEL AND NOTATION

We consider a polling system [14], 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 λ_i at Q_i, $i=1,\ldots,N$. The service times at Q_i are i.i.d. random variables, denoted by B_i, having finite first and second moment, and Laplace-Stieltjes transform $\beta_i(.)$. The switch of the server from Q_{i-1} to Q_i lasts for a (non-zero) switchover time S_i, these being i.i.d. random variables, with finite first two moments, and Laplace-Stieltjes transform $\sigma_i(.)$. The sum of the switchover times is denoted by $S=\sum_{i=1}^{N} S_i$. The residual service times and switchover times are denoted by R_{B_i}, and R_{S_i} and R_S, respectively, with expected values

$$\mathrm{E}[R_{B_i}]=\frac{\mathrm{E}[B_i^2]}{2\mathrm{E}[B_i]},\quad \mathrm{E}[R_{S_i}]=\frac{\mathrm{E}[S_i^2]}{2\mathrm{E}[S_i]},\quad \mathrm{E}[R_S]=\frac{\mathrm{E}[S^2]}{2\mathrm{E}[S]}.$$

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.

The service disciplines applied at a queue are gated service (G) and exhaustive service (E). In case of gated service, the server serves exactly the customers present upon its arrival at the queue. In case of exhaustive service, the server serves customers until the queue where it is working on, is empty.

The traffic offered per time unit at Q_i is denoted by ρ_i and is given by $\rho_i=\lambda_i\mathrm{E}[B_i]$. 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 [8]. 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.

2.1 Gated/Exhaustive discipline

We now describe the Gated/Exhaustive service discipline (G/E). The server visits the queues in fixed cyclic order. A cycle consists of the visit of the server to each of the queues twice: once serving them according to the gated service discipline, and once to the exhaustive one. The first visit to Q_i is gated, denoted by Q_{i_G}, the second exhaustive, denoted by Q_{i_E}. Starting with the switchover to Q_1, a cycle is typically given by:

$$S_1-Q_{1_G}-S_2-Q_{2_G}-\ldots-S_N-Q_{N_G}-\\ S_1-Q_{1_E}-S_2-Q_{2_E}-\ldots-S_N-Q_{N_E}.$$

The cycle time, denoted by C, consists of the visit times to each of the queues twice, and all switchover times occurred. A well known result [14] for the mean cycle time in a system where the queues are visited once in a cycle is $\mathrm{E}[C_{1visit}]=\mathrm{E}[S]/(1-\rho)$. As a cycle now contains two visits to each of the queues, we have

$$\mathrm{E}[C]=\frac{2\mathrm{E}[S]}{1-\rho}. \tag{1}$$

3. MEAN VISIT TIMES

For the G/E discipline, we aim to find the expected duration of a visit to a certain queue. Denote by $\mathrm{E}[V_{i_G}]$ the expected duration of a visit period to Q_i when it is served gated, and by $\mathrm{E}[V_{i_E}]$ when it is served exhaustively, $i=1,\ldots,N$. Denote by $\mathrm{E}[V_i]$ the expected duration of the visit periods to Q_i per cycle, so

$$\mathrm{E}[V_i]=\mathrm{E}[V_{i_G}]+\mathrm{E}[V_{i_E}].$$

As the server is working a fraction ρ_i of the time on Q_i, it follows from (1) that, for $i=1,\ldots,N$:

$$\mathrm{E}[V_i]=\frac{2\,\mathrm{E}[S]\,\rho_i}{1-\rho}. \tag{2}$$

In order to determine the individual mean visit times $\mathrm{E}[V_{i_G}]$ and $\mathrm{E}[V_{i_E}]$ we set up a system of linear equations. For each of the $2N$ visits to a queue during one cycle, we have a single linear equation. This equation expresses the expected duration of that visit in terms of the other mean visit times.

For $\mathrm{E}[V_{i_G}]$ we make use of the fact that at the moment an exhaustive service to Q_i ends, there are no type i customers present in the system any more. After this, type i customers arrive at rate λ_i during the switchover and visit times at other queues, until the start of the next gated service at Q_i. At that moment the type i customers present in the system are placed before a gate and these are the only ones to be served in this visit period to the queue. Now the mean duration of this visit time is the mean number of customers present at the start of the service, times the mean service time per customer. The mean number of customers present at the start of the service is equal to the arrival rate λ_i times the expected amount of time that has passed since the previous exhaustive visit to the queue. This gives:

$$\begin{aligned}\mathrm{E}[V_{i_G}]&=\lambda_i\mathrm{E}[B_i]\Big(\mathrm{E}[S_{i+1}]+\mathrm{E}[V_{i+1_E}]+\mathrm{E}[S_{i+2}]+\ldots\\ &\qquad+\mathrm{E}[V_{N_E}]+\mathrm{E}[S_1]+\mathrm{E}[V_{1_G}]+\ldots+\mathrm{E}[S_i]\Big)\\ &=\rho_i\left(\mathrm{E}[S]+\sum_{k=i+1}^{N}\mathrm{E}[V_{k_E}]+\sum_{k=1}^{i-1}\mathrm{E}[V_{k_G}]\right),\end{aligned} \tag{3}$$

for $i = 1, \ldots, N$, where an empty sum equals zero.

A similar expression can be found for $\mathrm{E}[V_{i_E}]$. Note that at the beginning of a gated service to Q_i there are no type i customers behind the gate, as all are placed before. Newly arriving type i customers are not served during this visit period any more, but have to wait until the server returns to the queue. So the mean number of customers present at the beginning of the exhaustive service to Q_i is equal to the arrival rate λ_i times the expected amount of time that has passed since the *beginning* of the previous gated service. But as the service to the queue is now exhaustive, the newly arriving customers during this visit time are still to be served during this visit. This can be interpreted as that every customer present at the start of the visit time induces a busy period. The expected duration of a busy period of one type i customer is $\mathrm{E}[B_i]/(1-\rho_i)$. This gives

$$\begin{aligned}\mathrm{E}[V_{i_E}] &= \lambda_i \frac{\mathrm{E}[B_i]}{1-\rho_i}\Big(\mathrm{E}[V_{i_G}] + \mathrm{E}[S_{i+1}] + \mathrm{E}[V_{i+1_G}] + \ldots \\ &\qquad + \mathrm{E}[V_{N_G}] + \mathrm{E}[S_1] + \mathrm{E}[V_{1_E}] + \ldots + \mathrm{E}[S_i]\Big) \\ &= \frac{\rho_i}{1-\rho_i}\left(\mathrm{E}[S] + \sum_{k=i}^{N}\mathrm{E}[V_{k_G}] + \sum_{k=1}^{i-1}\mathrm{E}[V_{k_E}]\right), \quad (4)\end{aligned}$$

for $i = 1, \ldots, N$.

Now (3) and (4) give a system of $2N$ linear equations in the $2N$ unknowns $\mathrm{E}[V_{i_G}]$ and $\mathrm{E}[V_{i_E}]$, $i = 1, \ldots, N$. Solving this gives explicit expressions for $\mathrm{E}[V_{i_G}]$ and $\mathrm{E}[V_{i_E}]$, $i = 1, \ldots, N$, in terms of $\mathrm{E}[S]$ and ρ_i.

Notice that in equilibrium the rate at which type i customers enter the system is equal to the rate at which they leave the system. So for $i = 1, \ldots, N$:

$$\lambda_i \mathrm{E}[C] = \frac{\mathrm{E}[V_{i_E}] + \mathrm{E}[V_{i_G}]}{\mathrm{E}[B_i]}.$$

The left-hand side gives the mean number of type i customers that enters the system during a cycle; the right-hand side gives the mean number of type i customers that are served during a cycle. This observation is another way to derive (2).

4. PSEUDO CONSERVATION LAW

4.1 PCL for polling systems

Boxma and Groenendijk [4] 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 sums of the mean waiting times at each of the queues. In case of exhaustive service at each of the queues, this expression is

$$\begin{aligned}\sum_{i=1}^{N}\rho_i\mathrm{E}[W_i] &= \rho\frac{\sum_{i=1}^{N}\rho_i\mathrm{E}[R_{B_i}]}{1-\rho} + \rho\mathrm{E}[R_S] \\ &\quad + \frac{\mathrm{E}[S]}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{N}\rho_i^2\right), \quad (5)\end{aligned}$$

and in case of gated service at each of the queues, it is

$$\begin{aligned}\sum_{i=1}^{N}\rho_i\mathrm{E}[W_i] &= \rho\frac{\sum_{i=1}^{N}\rho_i\mathrm{E}[R_{B_i}]}{1-\rho} + \rho\mathrm{E}[R_S] \\ &\quad + \frac{\mathrm{E}[S]}{2(1-\rho)}\left(\rho^2 + \sum_{i=1}^{N}\rho_i^2\right). \quad (6)\end{aligned}$$

Next to that, an expression is given in [4] for the case where some queues are served exhaustively and some are served gated.

These expressions are derived by considering a workload decomposition. Denote by V_{with} the amount of work in the cyclic service system at an arbitrary epoch in time, by $V_{without}$ the amount of work in the same system but without switchover times at an arbitrary epoch in time, and by Y the amount of work in the system at an arbitrary epoch in a switchover interval. It is proven in [4] that $V_{without}$ and Y are independent and that the following relation holds:

$$V_{with} \stackrel{d}{=} V_{without} + Y,$$

where $\stackrel{d}{=}$ denotes equality in distribution. This gives that

$$\mathrm{E}[V_{with}] = \mathrm{E}[V_{without}] + \mathrm{E}[Y].$$

The mean amount of work in the system without switchover times is given by

$$\mathrm{E}[V_{without}] = \frac{\sum_{i=1}^{N}\rho_i\mathrm{E}[R_{B_i}]}{1-\rho}, \quad (7)$$

independent of the service strategies. Denoting by L_i the number of customers at Q_i, then we have, next to this,

$$\begin{aligned}\mathrm{E}[V_{with}] &= \sum_{i=1}^{N}\mathrm{E}[B_i]\mathrm{E}[L_i] + \sum_{i=1}^{N}\rho_i\mathrm{E}[R_{B_i}] \\ &= \sum_{i=1}^{N}\rho_i\mathrm{E}[W_i] + \sum_{i=1}^{N}\rho_i\mathrm{E}[R_{B_i}]. \quad (8)\end{aligned}$$

Combining (7) and (8) now gives

$$\sum_{i=1}^{N}\rho_i\mathrm{E}[W_i] = \rho\frac{\sum_{i=1}^{N}\rho_i\mathrm{E}[R_{B_i}]}{1-\rho} + \mathrm{E}[Y]. \quad (9)$$

By Y_i we denote the amount of work at an arbitrary epoch in a switchover interval when switching to Q_i. Now $\mathrm{E}[Y]$ is the weighted sum of $\mathrm{E}[Y_i]$:

$$\mathrm{E}[Y] = \sum_{i=1}^{N}\frac{\mathrm{E}[S_i]}{\mathrm{E}[S]}\mathrm{E}[Y_i]. \quad (10)$$

In order to find the PCL, it remains to determine $\mathrm{E}[Y_i]$ for $i = 1, \ldots, N$. Note that these depend on the service disciplines at the queues. In [4] this is done for the cases of purely exhaustive and purely gated services, resulting in (5) respectively (6), and also for mixtures of these. In the next section we will derive expressions for the $\mathrm{E}[Y_i]$ in case of the G/E discipline.

4.2 PCL for G/E

We derive the PCL for the G/E policy in the general case with N queues. Using (9) this reduces to determining $\mathrm{E}[Y]$. Recall that this is the expected amount of work in the system at an arbitrary epoch in a switchover interval.

By $\mathrm{E}[Y_{i_G}]$ we denote the mean amount of work at an arbitrary epoch in a switchover interval to Q_i served gated, and by $\mathrm{E}[Y_{i_E}]$ to Q_i served exhaustively. As both the switchover intervals have the same distribution, when looking at the system at an arbitrary epoch in a switchover interval to Q_i, it is with equal probability a switchover interval to Q_i served gated or to Q_i served exhaustively. Therefore, for $i = 1, \ldots, N$:

$$\mathrm{E}[Y_i] = \frac{1}{2}\mathrm{E}[Y_{i_G}] + \frac{1}{2}\mathrm{E}[Y_{i_E}],$$

and so, using (10)

$$\mathrm{E}[Y] = \frac{1}{2}\sum_{i=1}^{N} \frac{\mathrm{E}[S_i]}{\mathrm{E}[S]}\mathrm{E}[Y_{i_G}] + \frac{1}{2}\sum_{i=1}^{N} \frac{\mathrm{E}[S_i]}{\mathrm{E}[S]}\mathrm{E}[Y_{i_E}]. \qquad (11)$$

We look at the system at an arbitrary epoch in the switchover interval S_i. First we consider the amount of work that arrived to the system during the time already passed in this switchover time. As at Q_i work is arriving at rate $\rho_i = \lambda_i \mathrm{E}[B_i]$, the rate at which work is arriving to the system is $\rho = \sum_{i=1}^{N} \rho_i$. Now use that the expected amount of time already passed during the switchover interval is equal to the expected residual switchover time $\mathrm{E}[R_{S_i}]$. This gives that the amount of work arrived to the system during the time already passed in this switchover interval is equal to $\rho \mathrm{E}[R_{S_i}]$.

Next to this, at each of the Q_i work arrived at rate ρ_i during the period until the start of the switchover time which is considered. If the last visit to Q_i was exhaustive, then this mean amount of work is equal to ρ_i times the mean duration of all intervals after the end of the visit to Q_i until the start of the switchover interval considered. If the last visit to the queue was gated, then we also have to include the mean duration of this gated visiting time, as the type i customers arriving in this interval are still in the system. Deriving the expressions for $\mathrm{E}[Y_{i_G}]$ and $\mathrm{E}[Y_{i_E}]$ is now straightforward, but some tedious bookkeeping is needed to consider which switchover intervals and visit times this concerns. We find, for $i = 1, \ldots, N$:

$$\begin{aligned}
\mathrm{E}[Y_{i_G}] &= \rho\, \mathrm{E}[R_{S_i}] \\
&\quad + \sum_{k=1}^{i-1} \rho_k \Bigg(\mathrm{E}[V_{k_G}] + \sum_{j=k+1}^{i-1} (\mathrm{E}[S_j] + \mathrm{E}[V_{j_G}]) \Bigg) \\
&\quad + \sum_{k=i}^{N} \rho_k \Bigg(\sum_{j=1}^{i-1} (\mathrm{E}[S_j] + \mathrm{E}[V_{j_G}]) \\
&\qquad\qquad + \sum_{j=k+1}^{N} (\mathrm{E}[S_j] + \mathrm{E}[V_{j_E}]) \Bigg), \\
\mathrm{E}[Y_{i_E}] &= \rho\, \mathrm{E}[R_{S_i}] + \sum_{k=1}^{i-2} \rho_k \Bigg(\sum_{j=k+1}^{i-1} (\mathrm{E}[S_j] + \mathrm{E}[V_{j_E}]) \Bigg) \\
&\quad + \sum_{k=i}^{N} \rho_k \Bigg(\mathrm{E}[V_{k_G}] + \sum_{j=1}^{i-1} (\mathrm{E}[S_j] + \mathrm{E}[V_{j_E}]) \\
&\qquad\qquad + \sum_{j=k+1}^{N} (\mathrm{E}[S_j] + \mathrm{E}[V_{j_G}]) \Bigg),
\end{aligned}$$

where an empty sum equals zero. Substituting these expressions into (11) gives the expression for $\mathrm{E}[Y]$ in terms of ρ_i and $\mathrm{E}[S_i]$. Using (9) we then find the PCL for a polling system with N queues in the G/E policy.

For the case of $N = 1$ queue, served according to the G/E discipline, this gives

$$\mathrm{E}[W_1] = \frac{\rho^2 \mathrm{E}[R_{B_1}]}{1-\rho} + \rho \mathrm{E}[R_{S_1}] + \frac{\mathrm{E}[S_1]}{2(1-\rho)}\left(\rho - \rho^2\right),$$

where for gated services the last term is (cf. (6)) $\frac{\mathrm{E}[S_1]}{2(1-\rho)}\left(2\rho^2\right)$, and for exhaustive services the last term vanishes (cf. (5)).

5. MEAN VALUE ANALYSIS

In this section we aim to find the mean steady state waiting times $\mathrm{E}[W_i]$, using *Mean Value Analysis* (MVA), as developed by Winands, Adan, and Van Houtum [16]. First we briefly outline the main idea of MVA for purely exhaustive or purely gated service, and then we adapt it to suit the G/E policy.

5.1 MVA for polling systems

For polling system with purely exhaustive or purely gated service, a system of N^2, respectively $N(N+1)$ linear equations is derived by making use of PASTA and Little's Law. The unknowns are the conditional mean queue lengths $\mathrm{E}[L_{ij}]$, i.e., the expected number of type i customers during a visit time at Q_j and a switchover time. In [16] the MVA method is also extended to the case with mixed exhaustive and gated service, i.e., some queues always receive gated service and the others exhaustive. Below we sketch the main ideas of MVA; we will provide more details in the extension to the G/E service policy.

First, by making use of PASTA, an equation is derived for the mean waiting time $\mathrm{E}[W_i]$ of a type i customer. This is the mean time it takes to serve all type i customers that are already waiting in the queue on the arrival of the (tagged) type i customer, given by $\mathrm{E}[L_i]$, plus a mean residual service time of a type i customer with probability ρ_i, plus the mean time it takes before the server starts working on Q_i again, denoted by $\mathrm{E}[T_i]$, and which depends on the service disciplines at the queues. So, for $i = 1, \ldots, N$,

$$\mathrm{E}[W_i] = \mathrm{E}[L_i]\,\mathrm{E}[B_i] + \rho_i \mathrm{E}[R_{B_i}] + \mathrm{E}[T_i].$$

Second, Little's Law gives, for $i = 1, \ldots, N$,

$$\mathrm{E}[L_i] = \lambda_i\, \mathrm{E}[W_i].$$

Hence, it remains to derive $\mathrm{E}[T_i]$. For this purpose, we define the following (service) periods. For (purely) exhaustive service, period i is the switchover time to Q_i plus the visit time to Q_i; for gated service, period i is the visit time to Q_i plus the switchover time to Q_{i+1}. Clearly, in case of exhaustive service, Q_i is empty at the end of period i, while in case of gated service, there are no customers behind the gate of Q_i at the start of period i. For both cases, denote by q_j the fraction of the time the system is in period j and let $\mathrm{E}[L_{ij}]$ be the expected length of Q_i during period j. Then the mean number of type i customers waiting in the queue, $\mathrm{E}[L_i]$, is a weighted average of $\mathrm{E}[L_{ij}]$, so for $i = 1, \ldots, N$:

$$\mathrm{E}[L_i] = \sum_{j=1}^{N} q_j\, \mathrm{E}[L_{ij}].$$

Further, let the interval (i, j) consist of the periods $i, i+1, \ldots, i+j-1$ and denote by $\mathrm{E}[R_{ij}]$ the expected residual

time of interval (i,j). Clearly, $\mathrm{E}[T_i] = \mathrm{E}[R_{i+1,N-1}]$ in case of exhaustive service, and $\mathrm{E}[T_i] = \mathrm{E}[R_{i,N}]$ for gated service. The final and crucial step of MVA is the derivation of a system of linear equations relating $\mathrm{E}[R_{ij}]$ and $\mathrm{E}[L_{ij}]$.

5.2 MVA for G/E

In this section we consider the G/E policy. Recall that a cycle is given by a visit to all of the queues twice, once served gated and once exhaustive, and all switchover times incurred. We number the switchover times and visit times to the N queues from 1 to $4N$, starting with the switchover to Q_1 served gated. This gives

S_1	Q_{1_G}	S_2	Q_{2_G}	$\ldots$	S_N	Q_{N_G}
1	2	3	4	$\ldots$	$2N-1$	$2N$

S_1	Q_{1_E}	S_2	Q_{2_E}	$\ldots$	S_N	Q_{N_E}
$2N+1$	$2N+2$	$2N+3$	$2N+4$	$\ldots$	$4N-1$	$4N$

Now we define period j, $j = 1,\ldots,4N$, as either the switch-over time or visit time numbered correspondingly. By q_j we denote the fraction of time the system is in period j, $j = 1,\ldots,4N$. Let interval (i,j) again consist of the periods $i, i+1, \ldots, i+j-1$. For the mean cycle length we have $\mathrm{E}[C] = 2\mathrm{E}[S]/(1-\rho)$, see (1), and the mean visit times to the queues $\mathrm{E}[V_{i_G}]$ and $\mathrm{E}[V_{i_E}]$ are derived in Section 3. Recall that the switchover times to Q_i served gated and served exhaustively are probabilistically identical.

5.2.1 System of equations

We derive a system of equations in order to determine $\mathrm{E}[W_i]$. Let $\mathrm{E}[W_{i_G}]$ denote the mean waiting time of a type i customer receiving gated service, and $\mathrm{E}[W_{i_E}]$ denote the mean waiting time for one receiving exhaustive service. For these we have, for $i = 1,\ldots,N$,

$$\mathrm{E}[W_i] = q_{G,i}\,\mathrm{E}[W_{i_G}] + q_{E,i}\,\mathrm{E}[W_{i_E}],$$

where $q_{G,i}$ denotes the fraction of type i customers that will receive gated service, and $q_{E,i}$ the fraction that will receive exhaustive service. Clearly $q_{E,i} = 1 - q_{G,i}$.

Let $\mathrm{E}[L_{i_G}]$ and $\mathrm{E}[L_{i_E}]$ be the mean number of waiting type i customers in the system that will receive gated, respectively exhaustive service. Then we have, for $i = 1,\ldots,N$,

$$\mathrm{E}[L_i] = \mathrm{E}[L_{i_G}] + \mathrm{E}[L_{i_E}].$$

There are type i customers in the system waiting for exhaustive service during the periods $2i,\ldots,2N+2i$, and type i customers waiting for gated service during the periods $2N+2i+1,\ldots,4N,1,\ldots,2i$. These two intervals are almost complementary; only during period $2i$, which is the visit to Q_{i_G}, both types of customers can be simultaneously present in the system: the ones that will receive (gated) service during this period and the ones that have arrived in this period, but who have to wait until the next (exhaustive) service to the queue. Hence we obtain, by PASTA, the following expressions, $i = 1,\ldots,N$,

$$q_{G,i} = \sum_{j=2N+2i+1}^{2i-1} q_j, \quad q_{E,i} = \sum_{j=2i}^{2N+2i} q_j, \tag{12}$$

where the summation for $q_{G,i}$ should be understood to be cyclical, i.e., over all $j \in \{2N+2i+1,\ldots,4N,1,\ldots,2i-1\}$.

Denote by λ_{i_G} and λ_{i_E} the arrival rates of type i customers that will be served gated, respectively exhaustively. So, for $i = 1,\ldots,N$,

$$\lambda_{i_G} = \lambda_i\, q_{G,i}, \quad \lambda_{i_E} = \lambda_i\, q_{E,i}.$$

Little's Law gives the following relations, for $i = 1,\ldots,N$,

$$\mathrm{E}[L_{i_G}] = \lambda_{i_G}\mathrm{E}[W_{i_G}], \quad \mathrm{E}[L_{i_E}] = \lambda_{i_E}\mathrm{E}[W_{i_E}].$$

Recall that $\mathrm{E}[L_{ij}]$ denotes the mean number of type i customers waiting in the queue during period j, for $i = 1,\ldots,N$ and $j = 1,\ldots,4N$. Hence, we have, for $i = 1,\ldots,N$,

$$\mathrm{E}[L_i] = \sum_{j=1}^{4N} q_j\,\mathrm{E}[L_{ij}].$$

During a gated service to Q_i, which is period $2i$, we distinguish between type i customers that will still receive service during this period (the ones behind the gate), and type i customers that have to wait until the next visit of the server to the queue. These last ones are those type i customers that arrived *during* this period, and so, as the service discipline is gated, will not receive service any more in this period (but have to wait before the gate). By $\bar{L}_{i,2i}$ we denote the ones that will receive service, and by $\tilde{L}_{i,2i}$ the ones that have to wait. We have $L_{i,2i} = \bar{L}_{i,2i} + \tilde{L}_{i,2i}$. Analogously to (12), this gives, for $i = 1,\ldots,N$,

$$\mathrm{E}[L_{i_G}] = q_{2i}\mathrm{E}[\bar{L}_{i,2i}] + \sum_{j=2N+2i+1}^{2i-1} q_j\,\mathrm{E}[L_{ij}],$$

$$\mathrm{E}[L_{i_E}] = q_{2i}\mathrm{E}[\tilde{L}_{i,2i}] + \sum_{j=2i+1}^{2N+2i} q_j\,\mathrm{E}[L_{ij}],$$

where the summation for $\mathrm{E}[L_{i_G}]$ should again be understood to be cyclical.

By making use of the PASTA property we obtain for the mean waiting time of a gated type i customer,

$$\mathrm{E}[W_{i_G}] = \frac{\mathrm{E}[L_{i_G}] - q_{2i}\mathrm{E}[\bar{L}_{i,2i}]}{q_{G,i}}\mathrm{E}[B_i] + \mathrm{E}[R_{2N+2i+1,2N-1}],$$

which can be interpreted as follows. A type i customer that will receive gated service, has to arrive in the periods $2N+2i+1,\ldots,4N,1,\ldots,2i-1$, consisting of $2N-1$ periods, and in which the system is a fraction $q_{G,i}$ of the time. Arriving customers have to wait for the services of all customers already present in the queue, and for the time it takes before the server starts working on Q_i again. The latter time has mean duration $\mathrm{E}[R_{2N+2i+1,2N-1}]$, since R_{ij} denotes the residual time of the periods $i, i+1,\ldots,i+j-1$. On arrival of a type i customer, there are on average $\sum_{j=2N+2i+1}^{2i-1} q_j\,\mathrm{E}[L_{ij}] = \mathrm{E}[L_{i_G}] - q_{2i}\mathrm{E}[\bar{L}_{i,2i}]$ type i customers already present in the system, all having mean service time $\mathrm{E}[B_i]$.

For the exhaustive type i customers we similarly find, for $i = 1,\ldots,N$:

$$\mathrm{E}[W_{i_E}] = \frac{\mathrm{E}[L_{i_E}]}{q_{E,i}}\,\mathrm{E}[B_i] + \frac{q_{2i} + \ldots + q_{2i+2N-1}}{q_{E,i}}\,\mathrm{E}[R_{2i,2N-1}] + \frac{q_{2i+2N}}{q_{E,i}}\,\mathrm{E}[R_{B_i}].$$

This gives a system of equations for $\mathrm{E}[W_i]$, $\mathrm{E}[W_{i_G}]$, $\mathrm{E}[W_{i_E}]$, $\mathrm{E}[L_i]$, $\mathrm{E}[L_{i_G}]$, $\mathrm{E}[L_{i_E}]$ which can be solved, provided $\mathrm{E}[R_{ij}]$ and $\mathrm{E}[L_{ij}]$ are known. The required equations for $\mathrm{E}[R_{ij}]$ and $\mathrm{E}[L_{ij}]$ are derived in the next section.

5.2.2 Residual periods and conditional queue lengths

We now derive a set of equations relating $\mathrm{E}[R_{ij}]$ and $\mathrm{E}[L_{ij}]$. At the end of an exhaustive service to Q_i, this queue is empty. From this moment on, the number of type i customers in the system increases at rate λ_i. As the residual duration of a period is in distribution equal to the amount of time already elapsed, so are their means. Hence

$$\lambda_i \,\mathrm{E}[R_{2N+2i+1,j}] = \frac{\sum_{k=1}^{j} q_{2N+2i+k}\mathrm{E}[L_{i,2N+2i+k}]}{\sum_{l=1}^{j} q_{2N+2i+l}},$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, 2N-1$.

The same idea applies to gated service. But now, at the beginning of a gated service to Q_i, there are no type i customers behind the gate, and from this moment on their number starts to increase at rate λ_i. This gives

$$\lambda_i \,\mathrm{E}[R_{2i,j}] = \frac{q_{2i}\mathrm{E}[\tilde{L}_{i,21}] + \sum_{k=1}^{j-1} q_{2i+k}\mathrm{E}[L_{i,2i+k}]}{\sum_{l=0}^{j-1} q_{2i+l}},$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, 2N$, where an empty sum equals zero.

Another way to express $\mathrm{E}[R_{ij}]$ is to determine the expected residual duration of an (i,j) interval based on the mean queue lengths at an arbitrary epoch during this interval. First we consider the case $j = 1$. For i odd, a residual $(i,1)$ period is just a residual switchover time, so

$$\mathrm{E}[R_{i,1}] = \mathrm{E}[R_{S_i}],\ \mathrm{E}[R_{2N+i,1}] = \mathrm{E}[R_{S_i}], \quad i = 1, 3, \ldots, 2N-1.$$

The residual time of a visit time to Q_i served gated satisfies

$$\mathrm{E}[R_{2i,1}] = \mathrm{E}[R_{B_i}] + \mathrm{E}[\bar{L}_{i,2i}]\mathrm{E}[B_i],$$

for $i = 1, \ldots, N$. First we have to wait for the residual service time of the customer in service, and then for the service of the $\bar{L}_{i,2i}$ customers in front of the gate. In case Q_i is served exhaustively, we have to wait for the busy periods induced by the customers present, yielding for $i = 1, \ldots, N$:

$$\mathrm{E}[R_{2N+2i,1}] = \frac{\mathrm{E}[R_{B_i}]}{1-\rho_i} + \mathrm{E}[L_{i,2N+2i}]\frac{\mathrm{E}[B_i]}{1-\rho_i}.$$

We now consider $j = 2$, in which case it is convenient to introduce $q_{i,j}$ defined as the sum of $q_i, \ldots, q_{i+j-1}$. With probability $q_{i+1}/q_{i,2}$ the residual $(i,2)$ period is equal to the residual $(i+1,1)$ period. With probability $q_i/q_{i,2}$ it is equal to the residual $(i,1)$ period plus either a switchover time (if i is even) or plus the busy period incurred by the number of customers present in the system and that of those arriving during the residual $(i,1)$ period (if i is odd). This yields, for $i = 1, \ldots, N$,

$$\begin{aligned}
\mathrm{E}[R_{2i-1,2}] &= \frac{q_{2i-1}}{q_{2i-1,2}} \\
&\qquad \big(\mathrm{E}[R_{S_i}](1+\rho_i) + \mathrm{E}[L_{i,2i-1}]\mathrm{E}[B_i]\big) \\
&\qquad + \frac{q_{2i}}{q_{2i-1,2}}\mathrm{E}[R_{2i,1}], \\
\mathrm{E}[R_{2i,2}] &= \frac{q_{2i}}{q_{2i,2}}\left(\mathrm{E}[R_{2i,1}] + \mathrm{E}[S_{i+1}]\right) \\
&\qquad + \frac{q_{2i+1}}{q_{2i,2}}\mathrm{E}[R_{2i+1,1}],
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[R_{2N+2i-1,2}] &= \frac{q_{2N+2i-1}}{q_{2N+2i-1,2}} \\
&\qquad \left(\frac{\mathrm{E}[R_{S_i}]}{1-\rho_i} + \mathrm{E}[L_{i,2N+2i-1}]\frac{\mathrm{E}[B_i]}{1-\rho_i}\right) \\
&\qquad + \frac{q_{2N+2i}}{q_{2N+2i-1,2}}\mathrm{E}[R_{2N+2i,1}], \\
\mathrm{E}[R_{2N+2i,2}] &= \frac{q_{2N+2i}}{q_{2N+2i,2}}\left(\mathrm{E}[R_{2N+2i,1}] + \mathrm{E}[S_{i+1}]\right) \\
&\qquad + \frac{q_{2N+2i+1}}{q_{2N+2i,2}}\mathrm{E}[R_{2N+2i+1,1}],
\end{aligned}$$

where $2N + 2N + 1$ is assumed to equal 1 as the system is cyclic. This can be readily extended to $j > 2$: with probability $q_{i+1,j-1}/q_{i,j}$ the residual (i,j) period is equal to the residual $(i+1, j-1)$ period, and otherwise, it is equal to the residual $(i,1)$ period plus an $(i+1, j-1)$ period, the mean length of which is determined by the mean queue lengths during period i. The resulting expressions are rather lengthy, and therefore omitted. We thus obtain sufficiently many equations to determine the unknowns $\mathrm{E}[R_{ij}]$ and $\mathrm{E}[L_{ij}]$.

Comparison of gated and exhaustive strategies

We compare gated and exhaustive strategies for a system with two queues where $\lambda_1 = 0.6$, $\lambda_2 = 0.2$, and $\mathrm{E}[S_i] = 1$, $\mathrm{E}[R_{S_i}] = 1$, $\mathrm{E}[B_i] = 1$, $\mathrm{E}[R_{B_i}] = 1$, for $i = 1, 2$. This is the same example as in [16], in which, using MVA, mean waiting times are derived in case both queues are served purely gated, purely exhaustively and mixed gated and exhaustively. The performance of these strategies is shown in Table 1, together with the results for the G/E strategy, starting the cycle either at Q_1 or Q_2; as the G/E strategy is not symmetric, this leads to different mean waiting times, although the weighted sum (i.e., the mean amount of work in the system) is the same.

From Table 1 we see that the weighted sum of the expected waiting times is minimal, when both queues are served exhaustively, as is to be expected, since in this case the server does not unnecessarily switch to the other queue when there is still work at the current queue. Serving both queues gated gives the highest weighted sum of mean waiting times, but this strategy is more fair to the queues, as the difference in the mean waiting times is smaller. The weighted sum of the mean waiting times for the two strategies, where one queue is served exhaustively and the other gated, is bigger than for purely exhaustive, and these strategies are less fair. On the other hand, the G/E strategy, starting the cycle at Q_1, is more fair than purely exhaustive, and the weighted sum of the mean waiting times is only a little more. Starting the G/E strategy at Q_2 does not lead to more fairness.

The given strategies try to achieve fairness in waiting times at the expense of (slightly) higher waiting times. The price paid, however, seems to be far less than that in the case of the two-stage gated service discipline [15]; this strategy achieves more fairness than purely gated service, but the mean waiting times increase by roughly an expected cycle time.

Based on the intuition as explained in the introduction, we initially expected that the G/E discipline would lead to small differences in mean waiting times, and hence more fairness. This clearly turns out not to be the case, probably because the mean visit times at the queues differ quite a lot: in this example we have $\{\mathrm{E}[V_{1_G}], \mathrm{E}[V_{2_G}], \mathrm{E}[V_{1_E}], \mathrm{E}[V_{2_E}]\} =$

Table 1: Comparison of the mean waiting times for a polling system with 2 queues.

Strategy					$\rho_1 E[W_1]+$	$\lvert E[W_1]-$
Q_1	Q_2	Q_1	Q_2	$E[W_1]$	$E[W_2]$	$\rho_2 E[W_2]$	$E[W_2]\rvert$
E	E	E	E	5.60	11.20	5.6	5.60
G	G	G	G	12.77	9.69	9.6	3.08
E	G	E	G	5.04	14.88	6.0	9.84
G	E	G	E	6.64	13.12	9.2	6.48
G	G	E	E	6.96	12.11	6.6	5.15
E	G	G	E	6.84	12.47	6.6	5.63

$\{3, 1, 9, 3\}$. Hence, roughly three quarter of the time the system is in the exhaustive part of the cycle, which explains why the mean waiting times of the exhaustive case dominate the ones for the G/E case. Further research should provide more insight in the potential of achieving fairness by making cycles of one *or more* gated visits to the queues, followed by one or more cycles of exhaustive visits.

6. MULTITYPE BRANCHING PROCESSES AND SMART CUSTOMERS

In Resing [13] it is shown that for polling systems with gated or exhaustive service disciplines, the joint queue length process at the beginning of a visit time to a fixed queue, is a *Multitype Branching Process* (MTBP). This gives expressions for the generating functions of the joint queue length distributions at these times.

We combine this with a concept given in Boxma [3], namely *Smart Customers* for polling systems. In this case, the arrival rates at the queues are dependent on the queue the server is working on. Refining this idea, we use it to model a polling system with the Gated/Exhaustive discipline. For this case we derive the generating functions of the queue lengths.

6.1 Multitype Branching Processes for Polling Systems

In [13] it is derived that, when the service disciplines at the queues satisfy the so-called Branching Property, then the queue length processes at the beginning of a visit time to a fixed queue, e.g. Q_1, form a MTBP with immigration. It is given that both the gated service discipline and the exhaustive service discipline satisfy this property. Then [13, Property 1], if the server arrives at Q_i and it finds k_i customers there, during the visit of the server each of these k_i customers is replaced in an i.i.d. way by a random population, having probability generating function (pgf) $h_i(z_1, z_2, \ldots, z_N)$. For the gated service discipline this pgf is given by

$$(G) \qquad h_i(z_1, z_2, \ldots, z_N) = \beta_i\left(\sum_{j=1}^{N} \lambda_j(1 - z_j)\right), \qquad (13)$$

where $\beta_i(.)$ is the Laplace Stieltjes Transform (LST) of service time distribution of type i customers. For the exhaustive service discipline the pgf is given by

$$(E) \qquad h_i(z_1, z_2, \ldots, z_N) = \theta_i\left(\sum_{j\neq i} \lambda_j(1 - z_j)\right), \qquad (14)$$

where $\theta_i(.)$ is the LST of a busy period in an M/G/1 queue.

Now [13, Theorem 2.2] states that, for a cyclic polling model where the service disciplines at each queue Q_i satisfy the Branching Property with pgf $h_i(z_1, z_2, \ldots, z_N)$, the numbers of customers in Q_1 at successive time points that the server reaches Q_1 constitute a MTBP with immigration in each state, where the offspring pgf's $f^{(i)}(z_1, z_2, \ldots, z_N)$ are given by

$$f^{(i)}(z_1, z_2, \ldots, z_N) = h_i\big(z_1, \ldots, z_i, f^{(i+1)}(z_1, z_2, \ldots, z_N), \ldots, f^{(N)}(z_1, z_2, \ldots, z_N)\big) \qquad (15)$$

and the immigration pgf $g(z)$ is given by

$$g(z_1, z_2, \ldots, z_N) = \prod_{i=1}^{N} \sigma_{i+1}\Big(\sum_{k=1}^{i} \lambda_k(1 - z_k) + \sum_{k=i+1}^{N} \lambda_k(1 - f^{(k)}(z_1, z_2, \ldots, z_N))\Big), \qquad (16)$$

where $\sigma_i(\cdot)$ is the LST of the switchover time distribution when switching to Q_i, and index $N+1$ should be read as 1. Then the pgf $P(z_1, z_2, \ldots, z_N)$ of the stationary distribution $\pi(j_1, j_2, \ldots, j_N)$ of the number of customers present in the system at the moment that the server starts working on Q_1, is given by

$$P(z_1, z_2, \ldots, z_N) = \prod_{n=0}^{\infty} g(f_n(z_1, z_2, \ldots, z_N)), \qquad (17)$$

where $f_n(z_1, z_2, \ldots, z_N)$ is recursively defined as:

$$f_0(z_1, z_2, \ldots, z_N) = (z_1, z_2, \ldots, z_N),$$
$$f_n(z_1, z_2, \ldots, z_N) = (f^{(1)}(f_{n-1}(z_1, z_2, \ldots, z_N)), \ldots, f^{(N)}(f_{n-1}(z_1, z_2, \ldots, z_N))), \ n \geq 1.$$

6.2 Smart customers

For modeling the G/E policy as a MTBP, we make use of the so-called concept of *Smart Customers*, see [3]. This idea is applicable for general polling systems and gives that the arrival rate at a queue depends on the position of the server. The arrival process of customers at Q_i when the server is working at Q_j is Poisson with rate λ_{ij}, and is Poisson with rate μ_{ij} when the server is switching from Q_{j-1} to Q_j.

In order to easily distinguish between the visits to a given queue in the gated service part of the cycle or in the exhaustive one, we number the queues as if there were $2N$ queues:

$$Q_1, Q_2, \ldots, Q_N, Q_{N+1}, \ldots, Q_{2N}.$$

For $i = 1, 2, \ldots, N$ we have that Q_i represents a gated visit to Q_i, and for $i = N+1, N+2, \ldots, 2N$ we have that Q_{N+i} represents an exhaustive visit to Q_i. A cycle of the server is given by:

$$S_1 - Q_1 - S_2 - Q_2 - \ldots - S_N - Q_N - S_1 - Q_{N+1} - S_2 - Q_{N+2} - \ldots - S_N - Q_{2N}.$$

The important observation here is that Q_i and Q_{N+i} are actually the same queue. When a type i customer arrives it should be directed to the appropriate queue. This can be achieved by a proper choice for the λ_{ij} and μ_{ij}.

First we look at the gated part of the cycle. When the server had not been working on Q_i yet, we direct arriving type i customers to queue Q_i. If the server has already

served Q_i, then arriving type i customers is directed to the queue served exhaustively, so to Q_{N+i}. If the server is working at Q_i, we have that arriving type i customers are not served in this service interval anymore, as the service discipline is gated. They will be served when the server is for the first time back at this queue, and so they are also directed to Q_{N+i}.

For the exhaustive part of the cycle we have almost the same reasoning. If the server has not yet been working on Q_{N+i}, arriving type i customers are directed to this queue. If the server has already served at Q_{N+i}, they are sent to Q_i. But when the server is working at Q_{N+i}, newly arriving customers are in this case served in this service interval, as the service discipline is exhaustive. So they are sent to Q_{N+i}.

We can summarize the above as follows. Let the set $J_i = \{i+1, \ldots, N+i\}$, then

$$\lambda_{ij} = \begin{cases} \lambda_i & \text{for } i = 1, \ldots, N,\ j \notin J_i \cup \{i\}, \\ \lambda_i & \text{for } i = N+1, \ldots, 2N,\ j \in J_i \cup \{i\}, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

and

$$\mu_{ij} = \begin{cases} \lambda_i & \text{for } i = 1, \ldots, N,\ j \notin J_i, \\ \lambda_i & \text{for } i = N+1, \ldots, 2N,\ j \in J_i, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

6.3 Multitype Branching Processes for Smart Customers

As we now have smart customers, we have to make small adjustments to (13), (14) and (16). Replacing λ_j by λ_{ji} in the first two gives:

$$(G) \qquad h_i(z_1, z_2, \ldots, z_N) = \beta_i \left(\sum_{j=1}^{N} \lambda_{ji}(1 - z_j) \right), \tag{20}$$

$$(E) \qquad h_i(z_1, z_2, \ldots, z_N) = \theta_i \left(\sum_{j \neq i} \lambda_{ji}(1 - z_j) \right), \tag{21}$$

and replacing λ_k by μ_{ki} in the third one gives:

$$g(z_1, z_2, \ldots, z_N) = \prod_{i=1}^{N} \sigma_i \Big(\sum_{k=1}^{i} \mu_{ki}(1 - z_k) + \sum_{k=i+1}^{N} \mu_{ki}(1 - f^{(k)}(z_1, z_2, \ldots, z_N)) \Big). \tag{22}$$

As the service disciplines are either gated or exhaustive, they do satisfy the Branching Property, with pgf $h_i(z_1, z_2, \ldots, z_n)$ at Q_i, and immigration process $g(z_1, z_2, \ldots, z_N)$.

In order to analyze the G/E discipline, we can now follow the same procedure as in Section 6.1, with $2N$ queues and the λ_{ij} and μ_{ij} as given in (18) and (19). Using the h_i of (20) and (21) we can by (15) calculate the offspring pgf's $f^{(i)}(z_1, z_2, \ldots, z_{2N})$ for $i = 1, \ldots, 2N$. In combination with $g(z_1, z_2, \ldots, z_{2N})$ of (22) the pgf of the stationary distribution $\pi(j_1, j_2, \ldots, j_{2N})$ follows by (17). This is the pgf of the number of customers present in the system at the moment that the server starts working on Q_1 according to the gated discipline. By renumbering the queues, we can in the same way find expressions for the moment that the server starts working on Q_i, $i = 2, \ldots, 2N$, i.e. to Q_i, $i = 1, \ldots, N$ served either gated or exhaustively.

6.4 Sojourn time distribution

Let D_i denote the steady-state sojourn time in Q_i. Using the distributional form of Little's law (cf. [10]), and introducing N_i^d, for the steady-state number of customers in Q_i immediately after a departure from Q_i, we have

$$\mathrm{E}[\mathrm{e}^{-\lambda_i(1-z)D_i}] = \mathrm{E}[z^{N_i^d}].$$

Indeed, those who have arrived during the sojourn time D_i of the tagged customer K are exactly those who are left behind by K.

Further introducing N_i and N_i^a, the steady-state numbers of customers in Q_i at an arbitrary epoch and just before an arrival at Q_i, it is easily seen that

$$N_i^d \stackrel{d}{=} N_i^a \stackrel{d}{=} N_i,$$

where $\stackrel{d}{=}$ denotes equality in distribution. The first equality follows since, for any j, there are just as many upcrossings in Q_i from j to $j+1$ customers as downcrossings from $j+1$ to j customers, and the second equality is due to PASTA. Hence

$$\mathrm{E}[\mathrm{e}^{-\lambda_i(1-z)D_i}] = \mathrm{E}[z^{N_i}]. \tag{23}$$

Now apply the Fuhrmann-Cooper decomposition [9] to N_i, yielding that

$$N_i \stackrel{d}{=} N_i^{M/G/1} + N_{i|I}, \tag{24}$$

in which $N_i^{M/G/1}$ and $N_{i|I}$ are independent; $N_i^{M/G/1}$ is the steady-state number of customers in the $M/G/1$ queue Q_i in isolation (i.e., the ordinary $M/G/1$ queue with arrival rate λ_i and service time distribution $B_i(\cdot)$); $N_{i|I}$ is the steady-state number of customers in Q_i at an arbitrary epoch in an intervisit period for Q_i.

We conclude from (23) and (24) that

$$\begin{aligned} \mathrm{E}[\mathrm{e}^{-\lambda_i(1-z)D_i}] &= \mathrm{E}[z^{N_i^{M/G/1}}]\mathrm{E}[z^{N_{i|I}}] \\ &= \frac{(1-\rho_i)(1-z)\beta_i(\lambda_i(1-z))}{\beta_i(\lambda_i(1-z)) - z}\mathrm{E}[z^{N_{i|I}}], \end{aligned}$$

the last step following from the well-known $M/G/1$ queue length result (cf. [6, p.238]).

We next determine $\mathrm{E}[z^{N_{i|I}}]$. Consider the intervisit periods I_i^E following an exhaustive visit to Q_i, and subsequently the intervisit periods I_i^G following a gated visit to Q_i. Q_i is empty at the beginning of each I_i^E, and subsequently grows to a random number N_i^E. During I_i^G, the number of customers in Q_i grows from some random number N_i^g to some random number N_i^G. N_i^G is also the number of customers at the beginning of an E-visit to Q_i. Hence [2]

$$\begin{aligned} \mathrm{E}[z^{N_{i|I}}] = &\frac{\mathrm{E}I_i^E}{\mathrm{E}I_i^E + \mathrm{E}I_i^G} \frac{1 - \mathrm{E}[z^{N_i^E}]}{(1-z)\mathrm{E}N_i^E} \\ &+ \frac{\mathrm{E}I_i^G}{\mathrm{E}I_i^E + \mathrm{E}I_i^G} \frac{\mathrm{E}[z^{N_i^g}] - \mathrm{E}[z^{N_i^G}]}{(1-z)(\mathrm{E}N_i^G - \mathrm{E}N_i^g)}. \end{aligned}$$

It remains to determine $\mathrm{E}[z^{N_i^E}]$, $\mathrm{E}[z^{N_i^g}]$ and $\mathrm{E}[z^{N_i^G}]$. N_i^E is the number of customers at the beginning of a G-visit to Q_i, so it has pgf $F_i^G(1, \ldots, 1, z, 1, \ldots, 1) = P(1, \ldots, 1, z, 1, \ldots, 1)$ with z occurring at the i-th position; the pgf P is given in (17), but with the adaptation for smart customers as outlined in Subsection 6.2. N_i^G is the number of customers at the beginning of an E-visit to Q_i, and hence has pgf

$F_i^E(1,\ldots,1,z,1,\ldots,1) = P(1,\ldots,1,z,1,\ldots,1)$ with z occurring at the $(N+i)$-th position; the pgf P is as above. N_i^g equals the number of arrivals to Q_i during a G-visit to Q_i. Hence

$$\mathrm{E}[z^{N_i^g}] = F_i^G(1,\ldots,1,\beta_i(\lambda_i(1-z)),1,\ldots,1).$$

7. CONCLUSION AND DISCUSSION

In this work we introduced the Gated/Exhaustive service discipline for polling systems. We derived a Pseudo Conservation Law for these systems, for an arbitrary number of queues. We adapted the Mean Value Analysis for polling systems of [16] to suit these models, and as an example we compared the mean waiting times for a number of mixes of gated and exhaustive strategies, for the case of 2 queues.

Using Smart Customers we were able to model the Gated/Exhaustive discipline as an 'ordinary' polling system with twice the number of queues. The concept of Multitype Branching Processes enabled us to derive the transform of the number of customers in the system at the moment the server starts working on a given queue.

One of the original aims of our study was to investigate whether the Gated/Exhaustive discipline leads to almost identical mean waiting times at all queues. It turns out that this is not the case. A possible topic for further study is to devise polling systems that do lead to better equalized mean waiting times. One could, e.g., think of the following:

(i) Other mixes of gated and exhaustive services, e.g., gated and exhaustive cycles in a ratio of $k_G : k_E$ for some k_G and k_E to be determined. We could vary the order in which the cycles are applied, e.g. $G-E-G-E-G$ repetitively, or we could take different ratios for each of the queues.

(ii) A mixed strategy of exhaustive and gated services, where the one chosen depends on a coin flip. There are more ways to do this. One way is to flip a coin at the beginning of a cycle and let this determine whether we do the entire round gated services or exhaustive services. Another way would be to decide this at each queue separately, at the moment the server arrives. For both cases, we could also let these probabilities depend on whether a gated or an exhaustive service was previously applied to the cycle respectively the queue, in that way letting the order of strategies become a Markov chain.

(iii) The fractional gated policy or fractional exhaustive policy [11]. In these strategies, for each of the customers it is decided whether or not it will be served during this visit of the server to the queue.

The above mentioned fractions and probabilities could be chosen in such a way as to equalize the mean waiting times, by minimizing the difference between the largest and the smallest mean waiting time; they could also be chosen such that they optimize some other performance measure.

Another topic for further research is to investigate whether it is possible to adapt the Mean Value Analysis to the case of smart customers. This will involve a more complicated system of equations, where it is the question if we can make the necessary changes to each of the equations. If we manage to do this, this would give another way to derive the mean waiting times in the Gated/Exhaustive models.

Finally, as suggested by a referee, it would be natural to apply the Gated/Exhaustive policy to non-cyclic polling systems, like systems with fixed polling tables [1, 5] (e.g. Q_1, Q_2, Q_1, Q_3 repetitively). Such a model fits into the framework of branching type models [13] and that of smart customers.

8. ACKNOWLEDGMENTS

The authors would like to thank anonymous referees for interesting suggestions for further work.

9. REFERENCES

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

[2] S. Borst. *Polling Systems*. PhD Thesis, 1995.

[3] O. 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.

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

[5] O. Boxma, W. Groenendijk, and J. Weststrate. A pseudoconservation law for service systems with a polling table. *IEEE Transactions on Communications*, 38(10):1865–1870, 1990.

[6] J. Cohen. *The Single Server Queue*. North-Holland, Amsterdam; 2nd. ed., 1982.

[7] D. Everitt. Simple Approximations for Token Rings. *IEEE Transactions on Communications*, 34(7):719–721, 1986.

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

[9] S. Fuhrmann and R. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 33:1117–1129, 1985.

[10] J. Keilson and L. Servi. The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9:239–247, 1990.

[11] H. Levy. Optimization of polling systems: the fractional exhaustive service method. *Report, Tel-Aviv University*, 1988.

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

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

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

[15] R. van der Mei and J. Resing. Polling models with two-stage gated service: Fairness versus efficiency. In L. Mason, T. Drwiega, and J. Yan, editors, *Managing Traffic Performance in Converged Networks, Proceedings ITC-20*, pages 544–555. LNCS 4516, Springer, Berlin, 2007.

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