A two-queue polling model with two priority levels in the first queue

Boon, M.A.A.; Adan, I.J.B.F.; Boxma, O.J.

Publication date
2008

A Two-Queue Polling Model with Two Priority Levels in the First Queue*

M.A.A. Boon[†] I.J.B.F. Adan[†] O.J. Boxma[†]
marko@win.tue.nl iadan@win.tue.nl boxma@win.tue.nl

May 13, 2008

Abstract

In this paper we consider a single-server cyclic polling system consisting of two queues. Between visits to successive queues, the server is delayed by a random switch-over time. Two types of customers arrive at the first queue: high and low priority customers. For this situation the following service disciplines are considered: gated, globally gated, and exhaustive. We study the cycle time distribution, the waiting times for each customer type, the joint queue length distribution at polling epochs, and the steady-state marginal queue length distributions for each customer type.

Keywords: Polling, priority levels, queue lengths, waiting times

1 Introduction

A polling model is a single-server system in which the server S visits n queues $Q_1, \ldots, Q_n$ in cyclic order. Customers that arrive at Q_i are referred to as type i customers. The special feature of the model considered in the present paper is that, within a customer type, we distinguish high and low priority customers. More specifically, we study a polling system which consists of two queues, Q_1 and Q_2. The first of these queues contains customers of two priority classes, high (H) and low (L). The exhaustive, gated and globally gated service disciplines are studied.

Our motivation to study a polling model with priorities is that scheduling through the introduction of priorities in a polling system can improve the performance of the system significantly without

---

having to purchase additional resources [15]. Priority polling systems can be used to study the Bluetooth and 802.11 protocols, or scheduling policies at routers and I/O subsystems in web servers. Different priority levels may be introduced to provide differentiated Quality-of-Service; e.g. one could give highest priority to jobs with a service requirement below a certain threshold level. Priority polling models also can be used to study traffic intersections where conflicting traffic flows face a green light simultaneously; e.g. traffic which takes a turn may have to give right of way to conflicting traffic that moves straight on, even if the traffic light is green for both traffic flows.

Although there is quite an extensive amount of literature available on polling systems, only very few papers treat priorities in polling models. Most of these papers only provide approximations or focus on pseudo-conservation laws. Wierman, Winands and Boxma [15] have obtained exact mean waiting time results using the Mean Value Analysis (MVA) framework for polling systems, developed in [16]. The MVA framework can only be used to find the first moment of the waiting time distribution for each customer type, and the mean residual cycle time. The main contribution of the present paper is the derivation of Laplace Stieltjes Transforms (LSTs) of the distributions of the marginal waiting times for each customer type; in particular it turns out to be possible to obtain exact expressions for the waiting time distributions of both high and low priority customers at a queue of a polling system. Probability Generating Functions (GFs) are derived for the joint queue length distribution at polling epochs, and for the steady-state marginal queue length distribution of the number of customers at an arbitrary epoch. Although we only consider a polling system with two queues, and two priority classes in Q_1, we believe that the results and the approach can be extended to models with any number of queues and any number of customer classes in each queue for the exhaustive, gated and globally gated service disciplines. This is the topic of a forthcoming paper.

The present paper is structured as follows: Section 2 gathers known results of nonpriority polling models which are relevant for the present study. Sections 3 (gated), 4 (globally gated), and 5 (exhaustive) give new results on the priority polling model. In each of the sections we successively discuss the joint queue length distribution at polling epochs, the cycle time distribution, the marginal queue length distributions and waiting time distributions. The mean waiting times are given at the end of each section.

2   Notation and description of the nonpriority polling model

The model that is considered in this section, is a polling model with two queues (Q_1 and Q_2) without priorities. We consider three service disciplines: gated, globally gated, and exhaustive. The gated service discipline states that during a visit of S to Q_i, S serves only those type i customers who are present at the moment that S arrives at Q_i. All type i customers that arrive during the visit of S to Q_i will be served during the next cycle. A cycle is the time between two successive visit beginnings (or completions) to a queue. The exhaustive service discipline states that when S arrives at Q_i, all type i customers that are present at that polling epoch, and

all type i customers that arrive during this particular visit of S to Q_i, are served until no type i customer is present in the system. We also consider the globally gated service discipline, which is similar to the gated service discipline, except for the fact that the symbolic gate is being set at the beginning of a cycle for *all* queues. This means that during a cycle only those customers will be served that were present at the beginning of that cycle.

Customers of type i arrive at Q_i according to a Poisson process with arrival rate λ_i $(i = 1, 2)$. Service times can follow any distribution, and we assume that a customer's service time is independent of other service times and independent of the arrival processes. The LST of the distribution of the generic service time B_i of type i customers is denoted by $\beta_i(\cdot)$. The fraction of time that the server is serving customers of type i equals $\rho_i := \lambda_i E(B_i)$. Switches of the server from Q_i to Q_{i+1} (all indices modulo 2), require a switch-over time S_i. The LST of this switch-over time distribution is denoted by $\sigma_i(\cdot)$. The fraction of time that the server is working (i.e., not switching) is $\rho := \rho_1 + \rho_2$. We assume that $\rho < 1$, which is a necessary and sufficient condition for the steady state distributions of cycle times, queue lengths and waiting times to exist.

This model has been extensively studied. Takács [14] studied this model, but without switch-over times and only with the exhaustive service discipline. Cooper and Murray [7] analysed this polling system for any number of queues, and for both gated and exhaustive service disciplines. Eisenberg [8] obtained results for a polling system with switch-over times (but only exhaustive service) by relating the GFs of the joint queue length distributions at visit beginnings, visit endings, service beginnings and service endings. Resing [13] was the first to point out the relation between polling systems and Multitype Branching Processes with immigration in each state. His results can be applied to polling models in which each queue satisfies the following property:

Property 2.1 If the server arrives at Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function $h_i(z_1, \ldots, z_n)$, which can be any n-dimensional probability generating function.

We use this property, and the relation to Multitype Branching Processes, to find results for our polling system with two queues, two priorities in the first queue, and gated, globally gated, and exhaustive service discipline. Notice that, unlike the gated and exhaustive service disciplines, the globally gated service discipline does not satisfy Property 2.1. But the results obtained by Resing also hold for a more general class of polling systems, namely those which satisfy the following (weaker) property that is formulated in [1]:

Property 2.2 If there are k_i customers present at Q_i at the beginning (or the end) of a visit to $Q_{\pi(i)}$, with $\pi(i) \in \{1, \ldots, n\}$, then during the course of the visit to Q_i, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function $h_i(z_1, \ldots, z_n)$, which can be any n-dimensional probability generating function.

Globally gated and gated are special cases of the synchronised gated service discipline, which states that only customers in Q_i will be served that were present at the moment that the server

3

reaches the "parent queue" of Q_i: $Q_{\pi(i)}$. For gated service, $\pi(i) = i$, for globally gated service, $\pi(i) = 1$. The synchronised gated service discipline is discussed in [12], but no observation is made that this discipline is a member of the class of polling systems satisfying Property 2.2 which means that results as obtained in [13] can be extended to this model.

Borst and Boxma [2] combined the results of Resing [13] and Eisenberg [8] to find a relation between the GFs of the marginal queue length distribution for polling systems with and without switch-over times, expressed in the Fuhrmann-Cooper queue length decomposition form [9].

2.1 Joint queue length distribution at polling epochs

The probability generating function $h_i(z_1, \ldots, z_n)$ which is mentioned in Property 2.1 depends on the service discipline. In a polling system with two queues and gated service we have $h_i(z_1, z_2) = \beta_i(\lambda_1(1 - z_1) + \lambda_2(1 - z_2))$. For exhaustive service this GF becomes $h_i(z_1, z_2) = \pi_i(\sum_{j \neq i} \lambda_j(1 - z_j))$, where $\pi_i(\cdot)$ is the LST of a busy period (BP) distribution in an $M/G/1$ system with only type i customers, so it is the root of the equation $\pi_i(\omega) = \beta_i(\omega + \lambda_i(1 - \pi_i(\omega)))$. We choose the beginning of a visit to Q_1 as start of a cycle. In order to find the joint queue length distribution at the beginning of a cycle, we have to define the immigration GF and the offspring GF analogous to [13]. The offspring GFs for queues 2 and 1 are given below.

$$f^{(2)}(z_1, z_2) = h_2(z_1, z_2),$$
$$f^{(1)}(z_1, z_2) = h_1(z_1, f^{(2)}(z_1, z_2)).$$

The immigration GFs are:

$$g^{(2)}(z_1, z_2) = \sigma_2(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)),$$
$$g^{(1)}(z_1, z_2) = \sigma_1(\lambda_1(1 - z_1) + \lambda_2(1 - f^{(2)}(z_1, z_2))).$$

The total immigration GF is the product of these two GFs:

$$g(z_1, z_2) = \prod_{i=1}^{2} g^{(i)}(z_1, z_2) = g^{(1)}(z_1, z_2)g^{(2)}(z_1, z_2).$$

We define the GF for the nth generation of offspring recursively:

$$f_n(z_1, z_2) = (f^{(1)}(f_{n-1}(z_1, z_2)), f^{(2)}(f_{n-1}(z_1, z_2))),$$
$$f_0(z_1, z_2) = (z_1, z_2).$$

The joint queue length GF at the beginning of a cycle (starting with a visit to Q_1) is

$$P_1(z_1, z_2) = \prod_{n=0}^{\infty} g(f_n(z_1, z_2)). \tag{2.1}$$

4

Resing [13] proves that this infinite product converges if and only if $\rho < 1$.

We can relate the joint queue length distribution at other polling epochs to $P_1(z_1, z_2)$. We denote the GF of the joint queue length distribution at a visit beginning to Q_i by $V_{b_i}(\cdot)$, so $P_1(\cdot) = V_{b_1}(\cdot)$. The queue length at a visit completion to Q_i is denoted by $V_{c_i}(\cdot)$. The following relations hold:

$$
\begin{aligned}
V_{b_1}(z_1, z_2) &= V_{c_2}(z_1, z_2)\sigma_2(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \\
&= V_{b_2}(z_1, h_2(z_1, z_2))\sigma_2(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \\
&= V_{b_2}(z_1, f^{(2)}(z_1, z_2))g^{(2)}(z_1, z_2), \quad\quad\quad\quad (2.2) \\
V_{b_2}(z_1, z_2) &= V_{c_1}(z_1, z_2)\sigma_1(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \\
&= V_{b_1}(h_1(z_1, z_2), z_2)\sigma_1(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)). \quad\quad (2.3)
\end{aligned}
$$

2.2 Cycle time

The cycle time, starting at a visit *beginning* to Q_1, is the sum of the visit times to Q_1 and Q_2, and the two switch-over times which are independent of the visit times. Since type 2 customers who arrive during the visit to Q_1 or the switch from Q_1 to Q_2 will be served during the visit to Q_2, it can be shown that the LST of the distribution of the cycle time C_1, $\gamma_1(\cdot)$, is related to $P_1(\cdot)$ as follows:

$$
\gamma_1(\omega) = \sigma_1(\omega + \lambda_2(1 - \phi_2(\omega)))\sigma_2(\omega)P_1(\phi_1(\omega + \lambda_2(1 - \phi_2(\omega))), \phi_2(\omega)), \quad (2.4)
$$

where $\phi_i(\cdot)$ is the LST of the distribution of the time that the server spends at Q_i due to the presence of one type i customer there. For gated service $\phi_i(\cdot) = \beta_i(\cdot)$, for exhaustive service $\phi_i(\cdot) = \pi_i(\cdot)$. A proof of (2.4) can be found in [5].

In some cases it is convenient to choose a different starting point for a cycle, for example when analysing a polling system with exhaustive service. If we define C_1 to be the time between two successive visit *completions* to Q_1, the LST of its distribution, $\gamma_1(\cdot)$, is:

$$
\begin{aligned}
\gamma_1(\omega) = {}&\sigma_1(\omega + \lambda_1(1 - \phi_1(\omega)) + \lambda_2(1 - \phi_2(\omega + \lambda_1(1 - \phi_1(\omega))))) \\
&\cdot \sigma_2(\omega + \lambda_1(1 - \phi_1(\omega)))V_{c_1}(\phi_1(\omega), \phi_2(\omega + \lambda_1(1 - \phi_1(\omega)))), \quad (2.5)
\end{aligned}
$$

with $V_{c_1}(z_1, z_2) = P_1(h_1(z_1, z_2), z_2)$.

2.3 Marginal queue lengths and waiting times

We denote the GF of the steady-state marginal queue length distribution of Q_1 at the visit beginning by $\widetilde{V}_{b_1}(z) = V_{b_1}(z, 1)$. Analogously we define $\widetilde{V}_{b_2}(\cdot)$, $\widetilde{V}_{c_1}(\cdot)$, and $\widetilde{V}_{c_2}(\cdot)$. It is shown in [2] that the steady-state marginal queue length of Q_i can be decomposed into two parts: the queue length of the corresponding $M/G/1$ queue with only type i customers, and the queue length at an arbitrary epoch during the intervisit period of Q_i, denoted by $N_{i|I}$. Borst [2] shows that by

virtue of PASTA, $N_{i|I}$ has the same distribution as the number of type i customers seen by an arbitrary type i customer arriving during an intervisit period, which equals

$$E(z^{N_{i|I}}) = \frac{E(z^{N_{i|I_{begin}}}) - E(z^{N_{i|I_{end}}})}{(1-z)(E(N_{i|I_{end}}) - E(N_{i|I_{begin}}))},$$

where $N_{i|I_{begin}}$ is the number of type i customers at the beginning of an intervisit period I_i, and $N_{i|I_{end}}$ is the number of type i customers at the end of I_i. Since the beginning of an intervisit period coincides with the completion of a visit to Q_i, and the end of an intervisit period coincides with the beginning of a visit, we know the GFs for the distributions of these random variables: $\widetilde{V}_{c_i}(\cdot)$ and $\widetilde{V}_{b_i}(\cdot)$. This leads to the following expression for the GF of the steady-state queue length distribution of Q_i at an arbitrary epoch, $E[z^{N_i}]$:

$$E[z^{N_i}] = \frac{(1-\rho_i)(1-z)\beta_i(\lambda_i(1-z))}{\beta_i(\lambda_i(1-z)) - z} \cdot \frac{\widetilde{V}_{c_i}(z) - \widetilde{V}_{b_i}(z)}{(1-z)(E(N_{i|I_{end}}) - E(N_{i|I_{begin}}))}. \qquad (2.6)$$

Keilson and Servi [10] show that the distributional form of Little's law can be used to find the LST of the marginal waiting time distribution: $E(z^{N_i}) = E(e^{-\lambda_i(1-z)(W_i+B_i)})$, hence $E(e^{-\omega W_i}) = E[(1 - \frac{\omega}{\lambda_i})^{N_i}]/\beta_i(\omega)$. This can be substituted into (2.6):

$$E[e^{-\omega W_i}] = \frac{(1-\rho_i)\omega}{\omega - \lambda_i(1-\beta_i(\omega))} \cdot \frac{\widetilde{V}_{c_i}\left(1 - \frac{\omega}{\lambda_i}\right) - \widetilde{V}_{b_i}\left(1 - \frac{\omega}{\lambda_i}\right)}{(E(N_{i|I_{end}}) - E(N_{i|I_{begin}}))\omega/\lambda_i}$$

$$= E[e^{-\omega W_{i|M/G/1}}]E\left[\left(1 - \frac{\omega}{\lambda_i}\right)^{N_{i|I}}\right]. \qquad (2.7)$$

The interpretation of this formula is that the waiting time of a type i customer in a polling model is the sum of two independent random variables: the waiting time of a customer in an $M/G/1$ queue with only type i customers, $W_{i|M/G/1}$, and the remaining intervisit time for a customer that arrives at an arbitrary epoch during the intervisit time of Q_i.

For *gated* service, the number of type i customers at the beginning of a visit to Q_i is exactly the number of type i customers that arrived during the previous cycle, starting at Q_i. In terms of GFs: $\widetilde{V}_{b_i}(z) = \gamma_i(\lambda_i(1-z))$. The number of type i customers at the end of a visit to Q_i are exactly those type i customers that arrived during this visit. In terms of GFs: $\widetilde{V}_{c_i}(z) = \gamma_i(\lambda_i(1 - \beta_i(\lambda_i(1-z))))$. We can rewrite $E(N_{i|I_{end}}) - E(N_{i|I_{begin}})$ as $\lambda_i E(I_i)$, because this is the number of type i customers that arrive during an intervisit time. In Section 2.4 we show that $\lambda_i E(I_i) = \lambda_i(1-\rho_i)E(C)$. Using these expressions we can rewrite Equation (2.7) for gated service to:

$$E[e^{-\omega W_i}] = \frac{(1-\rho_i)\omega}{\omega - \lambda_i(1-\beta_i(\omega))} \cdot \frac{\gamma_i(\lambda_i(1-\beta_i(\omega))) - \gamma_i(\omega)}{(1-\rho_i)\omega E(C)}. \qquad (2.8)$$

For *exhaustive* service, $\widetilde{V}_{c_i}(z) = 1$, because Q_i is empty at the end of a visit of S to Q_i. The number of type i customers at the beginning of a visit to Q_i in an exhaustive polling system is equal to the number of type i customers that arrived during the previous intervisit time of Q_i.

Hence, $\widetilde{V}_{b_i}(z) = \widetilde{I}_i(\lambda_i(1-z))$, where $\widetilde{I}_i(\cdot)$ is the LST of the intervisit time distribution for Q_i. Substitution of $\widetilde{I}_i(\omega) = \widetilde{V}_{b_i}(1 - \frac{\omega}{\lambda_i})$ in (2.7) leads to the following expression for the LST of the steady-state waiting time distribution of a type i customer in an exhaustive polling system:

$$E[e^{-\omega W_i}] = \frac{(1-\rho_i)\omega}{\omega - \lambda_i(1 - \beta_i(\omega))} \cdot \frac{1 - \widetilde{I}_i(\omega)}{\omega E(I_i)}. \tag{2.9}$$

To the best of our knowledge, the following result is new.

Proposition 2.3 Let the cycle time C_i be the time between two successive visit *completions* to Q_i. The LST of the cycle time distribution is given by (2.5). An equivalent expression for $E[e^{-\omega W_i}]$ if Q_i is served exhaustively, is:

$$\begin{aligned} E[e^{-\omega W_i}] &= \frac{1 - \gamma_i(\omega - \lambda_i(1 - \beta_i(\omega)))}{(\omega - \lambda_i(1 - \beta_i(\omega)))E(C)} \\ &= E[e^{-(\omega - \lambda_i(1 - \beta_i(\omega)))C_{i,res}}], \end{aligned} \tag{2.10}$$

where $C_{i,res}$ is the residual length of C_i.

Proof:
The cycle time is the length of an intervisit period I_i plus the length of a visit V_i, which is the time required to serve all type i customers that have arrived during I_i, and their type i descendants. Hence, the following equation holds:

$$\gamma_i(\omega) = \widetilde{I}_i(\omega + \lambda_i(1 - \pi_i(\omega))). \tag{2.11}$$

We use this equation to find the inverse relation:

$$\begin{aligned} \widetilde{I}_i(\omega + \lambda_i(1 - \pi_i(\omega))) &= \gamma_i(\omega) \\ &= \gamma_i(\omega + \lambda_i(1 - \pi_i(\omega)) - \lambda_i(1 - \pi_i(\omega)) \\ &= \gamma_i(\omega + \lambda_i(1 - \pi_i(\omega)) - \lambda_i(1 - \beta_i(\omega + \lambda_i(1 - \pi_i(\omega))))). \end{aligned}$$

If we substitute $s := \omega + \lambda_i(1 - \pi_i(\omega))$, we find

$$\widetilde{I}_i(s) = \gamma_i(s - \lambda_i(1 - \beta_i(s))). \tag{2.12}$$

Substitution of (2.12) into (2.9) gives (2.10). $\qquad\square$

Remark 2.4 We can write (2.11) and (2.12) as follows:

$$\begin{aligned} \gamma_i(\omega) &= \widetilde{I}_i(\psi(\omega)), \\ \widetilde{I}_i(s) &= \gamma_i(\phi(s)), \end{aligned}$$

with $\phi(\cdot)$ the Laplace exponent of the Lévy process $\sum_{i=1}^{N(t)} B_{1,i} - t$, where $N(t)$ is a Poisson process with intensity λ_i, and with $\psi(\omega) = \omega + \lambda_i(1 - \pi_i(\omega))$, which is known to be the inverse of $\phi(\cdot)$.

7

2.4 Moments

The focus of this paper is on LST and GF of distribution functions, not on their moments. Moments can be obtained by differentiation or Taylor series expansion, and are also discussed in [15]. In this subsection we will only mention some results that will be used later.

First we will derive the mean cycle time $E(C)$. Unlike higher moments of the cycle time, the mean does not depend on where the cycle starts: $E(C) = \frac{E(S_1)+E(S_2)}{1-\rho}$. This can easily be seen, because $1 - \rho$ is the fraction of time that the server is not working, but switching. The total switch-over time is $E(S_1) + E(S_2)$.

The expected length of a visit to Q_i is $E(V_i) = \rho_i E(C)$. The mean length of an intervisit period for Q_i is $E(I_i) = (1 - \rho_i)E(C)$. Notice that these expectations do not depend on the service discipline used. The expected number of type i customers at polling moments does depend on the service discipline. For gated service the expected number of type i customers at the beginning of a visit to Q_i is $\lambda_i E(C)$. For exhaustive service this is $\lambda_i E(I_i)$. The expected number of type i customers at the beginning of a visit to Q_{i+1} is $\lambda_i(E(V_i) + E(S_i))$ for gated service, and $\lambda_i E(S_i)$ for exhaustive service.

Moments of the waiting time distribution for a type i customer at an arbitrary epoch can be derived from the LSTs given by (2.8), (2.9) and (2.10). We only present the first moment:

$$\text{Gated:} \qquad E(W_i) = (1 + \rho_i)\frac{E(C_i^2)}{2E(C)}, \qquad (2.13)$$

$$\text{Exhaustive:} \qquad E(W_i) = \frac{E(I_i^2)}{2E(I_i)} + \frac{\rho_i}{1 - \rho_i}\frac{E(B_i^2)}{2E(B_i)},$$

$$= (1 - \rho_i)\frac{E(C_i^2)}{2E(C)}. \qquad (2.14)$$

Notice that the start of C_i is the *beginning* of a visit to Q_i for gated service, and the *end* of a visit for exhaustive service. Equations (2.13) and (2.14) are in agreement with Equations (4.1) and (4.2) in [3]. Although at first sight these might seem nice, closed formulas, it should be noted that the expected residual cycle time and the expected residual intervisit time are not easy to determine, requiring the solution of a large set of equations. MVA is an efficient technique to compute mean waiting times, the mean residual cycle time, and also the mean residual intervisit time. We refer to [16] for an MVA framework for polling models.

3 Gated service

In this section we study the gated service discipline for a polling system with two queues and two priority classes in the first queue: high (H) and low (L) priority customers. All type H and L customers that are present at the moment when the server arrives at Q_1, will be served during the server's visit to Q_1. First all type H customers will be served, then all type L customers. Type

H customers arrive at Q_1 according to a Poisson process with intensity λ_H, and have a service requirement B_H with LST $\beta_H(\cdot)$. Type L customers arrive at Q_1 with intensity λ_L, and have a service requirement B_L with LST $\beta_L(\cdot)$. If we do not distinguish between high and low priority customers, we can still use the results from Section 2 if we regard the system as a polling system with two queues where customers in Q_1 arrive according to a Poisson process with intensity $\lambda_1 := \lambda_H + \lambda_L$ and have service requirement B_1 with LST $\beta_1(\cdot) = \frac{\lambda_H}{\lambda_1}\beta_H(\cdot) + \frac{\lambda_L}{\lambda_1}\beta_L(\cdot)$.

We follow the same approach as in Section 2. First we study the joint queue length distribution at polling epochs, then the cycle time distribution, followed by the marginal queue length distribution and waiting time distribution. The last subsection provides the first moment of these distributions.

3.1 Joint queue length distribution at polling epochs

Equations (2.2) and (2.3) give the GFs of the joint queue length distribution at visit beginnings, $V_{b_i}(z_1, z_2)$. A type 1 customer entering the system is a type H customer with probability λ_H/λ_1, and a type L customer with probability λ_L/λ_1. We can express the GF of the joint queue length distribution in the polling system with priorities, $V_{b_i}(\cdot, \cdot, \cdot)$, in terms of the GF of the joint queue length distribution in the polling system without priorities, $V_{b_i}(\cdot, \cdot)$.

Lemma 3.1

$$V_{b_i}(z_H, z_L, z_2) = V_{b_i}\left(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1}, z_2\right). \tag{3.1}$$

Proof:
Let X_H be the number of high priority customers present in Q_1 at the beginning of a visit to Q_i, $i = 1, 2$. Similarly define X_L to be the number of low priority customers present in Q_1 at the beginning of a visit to Q_i. Let $X_1 = X_H + X_L$. Since the type H/L customers in Q_1 are exactly those H/L customers that arrived since the previous visit beginning at Q_i, we know that

$$P(X_H = i, X_L = k - i | X_1 = k) = \binom{k}{i}\left(\frac{\lambda_H}{\lambda_1}\right)^i \left(\frac{\lambda_L}{\lambda_1}\right)^{k-i}.$$

Hence

$$E[z_H^{X_H} z_L^{X_L} | X_1 = k] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} z_H^i z_L^j P(X_H = i, X_L = j | X_1 = k)$$

$$= \sum_{i=0}^{k}\binom{k}{i}\left(\frac{\lambda_H}{\lambda_1} z_H\right)^i \left(\frac{\lambda_L}{\lambda_1} z_L\right)^{k-i}$$

$$= \left(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1}\right)^k.$$

9

Finally,

$$V_{b_i}(z_H, z_L, z_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left(\frac{\lambda_H z_H + \lambda_L z_L}{\lambda_1} \right)^i z_2^j P(X_1 = i, X_2 = j)$$

$$= V_{b_i} \left(\frac{1}{\lambda_1} (\lambda_H z_H + \lambda_L z_L), z_2 \right).$$

$\square$

3.2   Cycle time

The LST of the cycle time distribution is still given by (2.4) if we define $\lambda_1 := \lambda_H + \lambda_L$ and $\beta_1(\cdot) := \frac{\lambda_H}{\lambda_1} \beta_H(\cdot) + \frac{\lambda_L}{\lambda_1} \beta_L(\cdot)$, because the cycle time does not depend on the order of service.

3.3   Marginal queue lengths and waiting times

We first determine the LST of the waiting time distribution for a type L customer, using the fact that this customer will not be served until the next cycle (starting at Q_1). The time from the start of the cycle until the arrival will be called "past cycle time", denoted by C_{1P}. The residual cycle time will be denoted by C_{1R}. The waiting time of a type L customer is composed of C_{1R}, the service times of all high priority customers that arrived during $C_{1P} + C_{1R}$, and the service times of all low priority customers that have arrived during C_{1P}. Let $N_H(T)$ be the number of high priority customers that have arrived during time interval T, and equivalently define $N_L(T)$.

Theorem 3.2

$$E\left[e^{-\omega W_L}\right] = \frac{\gamma_1(\lambda_H(1 - \beta_H(\omega)) + \lambda_L(1 - \beta_L(\omega))) - \gamma_1(\omega + \lambda_H(1 - \beta_H(\omega)))}{[\omega - \lambda_L(1 - \beta_L(\omega))]E(C)}.$$

Proof:

$$E\left[e^{-\omega W_L}\right] = E\left[e^{-\omega\left(C_{1R}+\sum_{i=1}^{N_H(C_{1P}+C_{1R})} B_{H,i}+\sum_{i=1}^{N_L(C_{1P})} B_{L,i}\right)}\right]$$

$$= \int_{t=0}^{\infty}\int_{u=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} E\left[e^{-\omega\left(u+\sum_{i=1}^{m} B_{H,i}+\sum_{i=1}^{n} B_{L,i}\right)}\right]$$
$$\cdot\, P(N_H(C_{1P}+C_{1R})=m,\,N_L(C_{1P})=n)\,\mathrm{d}P(C_{1P}<t,C_{1R}<u)$$

$$= \int_{t=0}^{\infty}\int_{u=0}^{\infty} e^{-\omega u}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} E\left[e^{-\omega\sum_{i=1}^{m} B_{H,i}}\right] E\left[e^{-\omega\sum_{i=1}^{n} B_{L,i}}\right]$$
$$\cdot\,\frac{(\lambda_H(t+u))^m}{m!}e^{-\lambda_H(t+u)}\frac{(\lambda_L t)^n}{n!}e^{-\lambda_L t}\,\mathrm{d}P(C_{1P}<t,C_{1R}<u)$$

$$= \int_{t=0}^{\infty}\int_{u=0}^{\infty} e^{-t(\lambda_H(1-\beta_H(\omega))+\lambda_L(1-\beta_L(\omega)))}e^{-u(\omega+\lambda_H(1-\beta_H(\omega)))}\,\mathrm{d}P(C_{1P}<t,C_{1R}<u)$$
$$= \frac{\gamma_1(\lambda_H(1-\beta_H(\omega))+\lambda_L(1-\beta_L(\omega)))-\gamma_1(\omega+\lambda_H(1-\beta_H(\omega)))}{[\omega-\lambda_L(1-\beta_L(\omega))]E(C)}. \tag{3.2}$$

For the last step in the derivation of (3.2) we used

$$E[e^{-\omega_P C_{1P}-\omega_R C_{1R}}] = \frac{E[e^{-\omega_P C_1}]-E[e^{-\omega_R C_1}]}{(\omega_R-\omega_P)E(C)},$$

which is obtained in [4]. $\qquad\square$

Remark 3.3 The Fuhrmann-Cooper decomposition [9] still holds for the waiting time of type L customers, because (3.2) can be rewritten to

$$E\left[e^{-\omega W_L}\right] = \frac{(1-\rho_L)\omega}{\omega-\lambda_L(1-\beta_L(\omega))}$$
$$\cdot\,\frac{\gamma_1(\lambda_H(1-\beta_H(\omega))+\lambda_L(1-\beta_L(\omega)))-\gamma_1(\omega+\lambda_H(1-\beta_H(\omega)))}{(1-\rho_L)\omega E(C)}. \tag{3.3}$$

We recognise the first term on the right-hand side of (3.3) as the LST of the waiting time distribution of an $M/G/1$ queue with only type L customers. An interpretation of the second term can be found when regarding the polling system as a polling system with *three* queues (Q_H, Q_L, Q_2) and no switch-over time between Q_H and Q_L. The service discipline of this equivalent system is synchronised gated, which is a more general version of gated. The gates for queues Q_H and Q_L are set simultaneously when the server arrives at Q_H, but the gate for Q_2 is still set when the server arrives at Q_2. In the following paragraphs we show that the second term on the right-hand side of (3.3) can be interpreted as $E[\left(1-\frac{\omega}{\lambda_L}\right)^{N_{L|I}}]$, where $N_{L|I}$ is the number of type L customers at a random epoch during the intervisit period of Q_L.

The expression for the LST of the distribution of the number of type L customers at an arbitrary epoch is determined by first converting the waiting time LST to sojourn time LST, i.e., multiplying expression (3.3) with $\beta_L(\omega)$. Second, we apply the distributional form of Little's law [10] to (3.3). This law can be applied because the required conditions are fulfilled for each customer class (H, L, and 2): the customers enter the system in a Poisson stream, every customer enters the system and leaves the system one at a time in order of arrival, and for any time t the entry process into the system of customers after time t and the time spent in the system by any customer arriving before time t are independent. The result is:

$$E\left[z^{N_L}\right] = \frac{(1 - \rho_L)(1 - z)\beta_L(\lambda_L(1 - z))}{\beta_L(\lambda_L(1 - z)) - z} \cdot \frac{\widetilde{V}_{c_L}(z) - \widetilde{V}_{b_L}(z)}{(1 - z)(E(N_{L|I_{end}}) - E(N_{L|I_{begin}}))}. \qquad (3.4)$$

In this equation $\widetilde{V}_{b_L}(z)$ denotes the GF of the distribution of the number of type L customers at the beginning of a visit to Q_L, and $\widetilde{V}_{c_L}(z)$ denotes the GF at the completion of a visit to Q_L:

$$\begin{aligned}
\widetilde{V}_{b_L}(z) &= V_{b_1}(\beta_H(\lambda_L(1 - z)), z, 1) \\
&= \gamma_1(\lambda_H(1 - \beta_H(\lambda_L(1 - z))) + \lambda_L(1 - z)), \\
\widetilde{V}_{c_L}(z) &= V_{b_1}(\beta_H(\lambda_L(1 - z)), \beta_L(\lambda_L(1 - z)), 1) \\
&= \gamma_1(\lambda_H(1 - \beta_H(\lambda_L(1 - z))) + \lambda_L(1 - \beta_L(\lambda_L(1 - z)))).
\end{aligned}$$

The last term in (3.4) is the GF of the distribution of the number of type L customers at an arbitrary epoch during the intervisit period of Q_L, $E[z^{N_{L|I}}]$. Substitution of $\omega := \lambda_L(1 - z)$ in (3.4), and using $(E(N_{L|I_{end}}) - E(N_{L|I_{begin}})) = \lambda_L E(I_L)$, shows that the last term of (3.3) indeed equals $E[\left(1 - \frac{\omega}{\lambda_L}\right)^{N_{L|I}}]$.

The derivation of the LSTs of W_H and W_2 is similar and leads to the following expressions:

$$E\left[e^{-\omega W_H}\right] = \frac{(1 - \rho_H)\omega}{\omega - \lambda_H(1 - \beta_H(\omega))} \cdot \frac{\gamma_1(\lambda_H(1 - \beta_H(\omega))) - \gamma_1(\omega)}{(1 - \rho_H)\omega E(C)}, \qquad (3.5)$$

$$E\left[e^{-\omega W_2}\right] = \frac{(1 - \rho_2)\omega}{\omega - \lambda_2(1 - \beta_2(\omega))} \cdot \frac{\gamma_2(\lambda_2(1 - \beta_2(\omega))) - \gamma_2(\omega)}{(1 - \rho_2)\omega E(C)}. \qquad (3.6)$$

Remark 3.4 Equations (3.5) and (3.6) are equivalent to the LST of W_i in a nonpriority polling system (2.8), which illustrates that the Fuhrmann-Cooper decomposition also holds for the waiting time distributions of high priority customers in Q_1 and type 2 customers in a polling system with gated service.

Application of the distributional form of Little's law to these expressions results in:

$$E\left[z^{N_H}\right] = \frac{(1 - \rho_H)(1 - z)\beta_H(\lambda_H(1 - z))}{\beta_H(\lambda_H(1 - z)) - z} \cdot \frac{\gamma_1(\lambda_H(1 - \beta_H(\lambda_H(1 - z)))) - \gamma_1(\lambda_H(1 - z))}{\lambda_H(1 - \rho_H)(1 - z)E(C)},$$

$$E\left[z^{N_2}\right] = \frac{(1 - \rho_2)(1 - z)\beta_2(\lambda_2(1 - z))}{\beta_2(\lambda_2(1 - z)) - z} \cdot \frac{\gamma_2(\lambda_2(1 - \beta_2(\lambda_2(1 - z)))) - \gamma_2(\lambda_2(1 - z))}{\lambda_2(1 - \rho_2)(1 - z)E(C)}.$$

Remark 3.5 If the service discipline in Q_2 is not gated, but another branching type service discipline that satisfies Property 2.1, (3.6) should be replaced by the more general expression (2.7).

3.4 Moments

As mentioned in Section 2.4, we do not focus on moments in this paper, and we only mention the mean waiting times of type H and L customers. For a type H customer, it is immediately clear that $E(W_H) = (1 + \rho_H)E(C_{1,res})$. The mean waiting time for a type L customer can be obtained by differentiating (3.2). This results in:

$$E(W_L) = (1 + 2\rho_H + \rho_L)E(C_{1,res}).$$

These formulas can also be obtained using MVA, as shown in [15].

4 Globally gated service

In this section we discuss a polling model with two queues (Q_1, Q_2) and two priority classes (H and L) in Q_1 with globally gated service. For this service discipline, only customers that were present when the server started its visit to Q_1 are served. This feature makes the model exactly the same as a nonpriority polling model with three queues (Q_H, Q_L, Q_2). Although this system does not satisfy Property 2.1, it does satisfy Property 2.2 which implies that we can still follow the same approach as in the previous sections.

4.1 Joint queue length distribution at polling epochs

We define the beginning of a visit to Q_1 as the start of a cycle, since this is the moment that determines which customers will be served during the next visits to the queues. Arriving customers will always be served in the next cycle, so the three offspring GFs are:

$$f^{(i)}(z_H, z_L, z_2) = h_i(z_H, z_L, z_2) = \beta_i(\lambda_H(1-z_H)+\lambda_L(1-z_L)+\lambda_2(1-z_2)), \qquad i = H, L, 2.$$

The two immigration functions are:

$$g^{(i)}(z_H, z_L, z_2) = \sigma_i(\lambda_H(1 - z_H) + \lambda_L(1 - z_L) + \lambda_2(1 - z_2)), \qquad i = 1, 2.$$

Using these definitions, the formula for the GF of the joint queue length distribution at the beginning of a cycle is similar to the one found in Section 2:

$$P_1(z_H, z_L, z_2) = \prod_{n=0}^{\infty} g(f_n(z_H, z_L, z_2)). \tag{4.1}$$

13

Notice that in a system with globally gated service it is possible to express the joint queue length distribution at the beginning of a cycle in terms of the cycle time LST, since all customers that are present at the beginning of a cycle are exactly all of the customers that have arrived during the previous cycle:

$$P_1(z_H, z_L, z_2) = \gamma_1(\lambda_H(1 - z_H) + \lambda_L(1 - z_L) + \lambda_2(1 - z_2)). \tag{4.2}$$

4.2 Cycle time

Since only those customers that are present at the start of a cycle, starting at Q_1, will be served during this cycle, the LST of the cycle time distribution is

$$\gamma_1(\omega) = \sigma_1(\omega)\sigma_2(\omega)P_1(\beta_H(\omega), \beta_L(\omega), \beta_2(\omega)). \tag{4.3}$$

Substitution of (4.2) into this expression gives us the following relation:

$$\gamma_1(\omega) = \sigma_1(\omega)\sigma_2(\omega)\gamma_1(\lambda_H(1 - \beta_H(\omega)) + \lambda_L(1 - \beta_L(\omega)) + \lambda_2(1 - \beta_2(\omega))).$$

Boxma, Levy and Yechiali [4] show that this relation leads to the following expression for the cycle time LST:

$$\gamma_1(\omega) = \prod_{i=0}^{\infty} \sigma(\delta^{(i)}(\omega)),$$

where $\sigma(\cdot) = \sigma_1(\cdot)\sigma_2(\cdot)$, and $\delta^{(i)}(\omega)$ is recursively defined as follows:

$$\delta^{(0)}(\omega) = \omega,$$
$$\delta^{(i)}(\omega) = \delta(\delta^{(i-1)}(\omega)), \qquad i = 1, 2, 3, \ldots,$$
$$\delta(\omega) = \lambda_H(1 - \beta_H(\omega)) + \lambda_L(1 - \beta_L(\omega)) + \lambda_2(1 - \beta_2(\omega)).$$

4.3 Marginal queue lengths and waiting times

For type H and L customers, the expressions for $E(e^{-\omega W_H})$ and $E(e^{-\omega W_L})$ are exactly the same as the ones found in Section 3.3, but with $\gamma_1(\cdot)$ as defined in (4.3).

The expression for $E(e^{-\omega W_2})$ can be obtained with the method used in Section 3.3:

$$E\left[e^{-\omega W_2}\right] = \sigma_1(\omega)\frac{\gamma_1(\sum_{i=H,L,2}\lambda_i(1 - \beta_i(\omega))) - \gamma_1(\omega + \sum_{i=H,L}\lambda_i(1 - \beta_i(\omega)))}{[\omega - \lambda_2(1 - \beta_2(\omega))]E(C)}$$

$$= \sigma_1(\omega) \cdot \frac{(1 - \rho_2)\omega}{\omega - \lambda_2(1 - \beta_2(\omega))}$$
$$\cdot \frac{\gamma_1(\sum_{i=H,L,2}\lambda_i(1 - \beta_i(\omega))) - \gamma_1(\omega + \sum_{i=H,L}\lambda_i(1 - \beta_i(\omega)))}{(1 - \rho_2)\omega E(C)}.$$

14

We can use the distributional form of Little's law to determine the LST of the marginal queue length distribution of Q_2:

$$E\left[z^{N_2}\right] = \sigma_1(\lambda_2(1-z))\frac{(1-\rho_2)(1-z)\beta_2(\lambda_2(1-z))}{\beta_2(\lambda_2(1-z))-z}$$
$$\cdot\frac{\gamma_1(\sum_{i=H,L,2}\lambda_i(1-\beta_i(\lambda_2(1-z)))) - \gamma_1(\lambda_2(1-z)+\sum_{i=H,L}\lambda_i(1-\beta_i(\lambda_2(1-z))))}{\lambda_2(1-\rho_2)(1-z)E(C)}.$$

Remark 4.1 The Fuhrmann-Cooper queue length decomposition also holds for all customer classes in a polling system with globally gated service.

4.4 Moments

The expressions for $E(W_H)$ and $E(W_L)$ from section 3.4 also hold in a globally gated polling system, but with a different mean residual cycle time. We only provide the mean waiting time of type 2 customers:

$$E(W_2) = E(S_1) + (1 + 2\rho_H + 2\rho_L + \rho_2)E(C_{1,res}).$$

5 Exhaustive service

In this section we study the same polling model as in the previous two sections, but the two queues are served exhaustively. The section has the same structure as the other sections, so we start with the derivation of the LST of the joint queue length distribution at polling epochs, followed by the LST of the cycle time distribution. LSTs of the marginal queue length distributions and waiting time distributions are provided in the next subsection. In the last part of the section the mean waiting time of each customer type is studied.

It should be noted that, although we assume that both Q_1 and Q_2 are served exhaustively, a model in which Q_2 is served according to another branching type service discipline, requires only minor adaptations.

5.1 Joint queue length distribution at polling epochs

We can derive the joint queue length distribution at the beginning of a cycle for a polling system with two queues and two priority classes in Q_1, $P_1(z_H, z_L, z_2)$, directly from expression (2.1) for $P_1(z_1, z_2)$. Similar to the proof of Lemma 3.1, we can prove that

$$P_1(z_H, z_L, z_2) = P_1\left(\frac{1}{\lambda_1}(\lambda_H z_H + \lambda_L z_L), z_2\right).$$

The same holds for $V_{b_2}(\cdot, \cdot, \cdot)$ and visit completion epochs $V_{c_i}(\cdot, \cdot, \cdot)$, for $i = 1, 2$.

5.2 Cycle time

For the cycle time starting with a visit to Q_1, (2.4) is still valid by defining $\lambda_1 := \lambda_H + \lambda_L$ and $\beta_1(\cdot) := \frac{\lambda_H}{\lambda_1}\beta_H(\cdot) + \frac{\lambda_L}{\lambda_1}\beta_L(\cdot)$. However, when studying the waiting time of a specific customer type in an exhaustively served queue, it is convenient to consider the *completion* of a visit to Q_1 as the start of a cycle. Hence, in this section the notation C_1, or the LST of its distribution, $\gamma_1(\cdot)$, refers to the cycle time starting at the completion of a visit to Q_1. Equation (2.5) gives the LST of the distribution of C_1, again with the definitions $\lambda_1 := \lambda_H + \lambda_L$ and $\beta_1(\cdot) := \frac{\lambda_H}{\lambda_1}\beta_H(\cdot) + \frac{\lambda_L}{\lambda_1}\beta_L(\cdot)$.

5.3 Marginal queue lengths and waiting times

Analysis of the model with exhaustive service requires a different approach. The key observation, made by Fuhrmann and Cooper [9], is that the polling system from the viewpoint of a type i customer is an $M/G/1$ queue with multiple server vacations. The $M/G/1$ queue with priorities and vacations has been extensively analysed by Kella and Yechiali [11]. We use their approach to find the waiting time LST for type H and L customers. Kella and Yechiali [11] distinguish between systems with single and multiple vacations, and preemptive resume and nonpreemptive service. In the present paper we do not consider preemptive resume, so we only use results from the case labelled as NPMV (nonpreemptive, multiple vacations) in [11]. We consider the system from the viewpoint of a type H and type L customer separately to derive $E[e^{-\omega W_H}]$ and $E[e^{-\omega W_L}]$.

From the viewpoint of a type H customer and as far as waiting times are considered, a polling system is a *nonpriority* single server system with multiple vacations. The vacation can either be the intervisit period I_1, or the service of a type L customer. The LSTs of these two types of vacations are:

$$E[e^{-\omega I_1}] = P_1(1 - \omega/\lambda_1, 1), \tag{5.1}$$

$$E[e^{-\omega B_L}] = \beta_L(\omega).$$

Equation (5.1) follows immediately from the fact that the number of type 1 (i.e. both H and L) customers at the beginning of a visit to Q_1 is the number of type 1 customers that have arrived during the previous intervisit period: $P_1(z, 1) = E[e^{-(\lambda_1(1-z))I_1}]$.

The key observation is that an arrival of a tagged type H customer will always take place within either an I_H cycle, or an L_H cycle. An I_H cycle is a cycle that starts with an intervisit period for Q_1, followed by the service of all type H customers that have arrived during the intervisit period, and ends at the moment that no type H customers are left in the system. Notice that at the start of the intervisit period, no type H customers were present in the system either. An L_H cycle is a similar cycle, but starts with the service of a type L customer. This cycle also ends at the moment that no type H customers are left in the system.

The fraction of time that the system is in an L_H cycle is $\frac{\rho_L}{1-\rho_H}$, because type L customers arrive with intensity λ_L. Each of these customers will start an L_H cycle and the length of an L_H cycle

16

equals $\frac{E(B_L)}{1-\rho_H}$:

$$E(L_H \text{ cycle}) = E(B_L) + \lambda_H E(B_L)E(BP_H)$$

$$= E(B_L) + \lambda_H E(B_L)\frac{E(B_H)}{1 - \rho_H}$$

$$= (1 + \frac{\rho_H}{1 - \rho_H})E(B_L) = \frac{E(B_L)}{1 - \rho_H},$$

where $E(BP_H)$ is the mean length of a busy period of type H customers.

The fraction of time that the system is in an I_H cycle, is $1 - \frac{\rho_L}{1-\rho_H} = \frac{1-\rho_1}{1-\rho_H}$. This result can also be obtained by using the argument that the fraction of time that the system is in an intervisit period is the fraction of time that the server is not serving Q_1, which is equal to $1 - \rho_1$. A cycle which starts with such an intervisit period and stops when all type H customers that arrived during the intervisit period and their type H descendants have been served, has mean length $E(I_1) + \lambda_H E(I_1)E(BP_H) = \frac{E(I_1)}{1-\rho_H}$. This also leads to the conclusion that $\frac{1-\rho_1}{1-\rho_H}$ is the fraction of time that the system is in an I_H cycle. A customer arriving during an I_H cycle views the system as a nonpriority $M/G/1$ queue with multiple server vacations I_1; a customer arriving during an L_H cycle views the system as a nonpriority $M/G/1$ queue with multiple server vacations B_L.

Fuhrmann and Cooper [9] showed that the waiting time of a customer in an $M/G/1$ queue with server vacations is the sum of two independent quantities: the waiting time of a customer in a corresponding $M/G/1$ queue without vacations, and the residual vacation time. Hence, the LST of the waiting time distribution of a type H customer is:

$$E[e^{-\omega W_H}] = \frac{(1 - \rho_H)\omega}{\omega - \lambda_H(1 - \beta_H(\omega))} \cdot \left[\frac{1 - \rho_1}{1 - \rho_H} \cdot \frac{1 - \widetilde{I}_1(\omega)}{\omega E(I_1)} + \frac{\rho_L}{1 - \rho_H} \cdot \frac{1 - \beta_L(\omega)}{\omega E(B_L)}\right]. \quad (5.2)$$

Equation (5.2) is in accordance with the more general equation in Section 4.1 in [11].

Remark 5.1 The term $\frac{1-\widetilde{I}_1(\omega)}{\omega E(I_1)}$ in (5.2), which is the LST of the residual intervisit time distribution, is the only difference between $E[e^{-\omega W_H}]$ and $E[e^{-\omega W_{H|M/G/1}}]$, the LST of the waiting time distribution of a high priority customer in a two priority $M/G/1$ queue without vacations:

$$E[e^{-\omega W_{H|M/G/1}}] = \frac{(1 - \rho_1)\omega + \lambda_L(1 - \beta_L(\omega))}{\omega - \lambda_H(1 - \beta_H(\omega))} \quad (5.3)$$

$$= \frac{(1 - \rho_H)\omega}{\omega - \lambda_H(1 - \beta_H(\omega))} \cdot \left[\frac{1 - \rho_1}{1 - \rho_H} + \frac{\rho_L}{1 - \rho_H} \cdot \frac{1 - \beta_L(\omega)}{\omega E(B_L)}\right].$$

The LST of the distribution of the waiting time of a high priority customer in a two priority $M/G/1$ queue usually appears in a form similar to (5.3) (see e.g. [6], Chapter 3), but as shown above it can be rewritten as the LST of the waiting time distribution of a customer in a nonpriority $M/G/1$ queue, where the server occasionally goes on a vacation. The customers in this system arrive with intensity λ_H and have service requirement LST $\beta_H(\omega)$. With probability $\frac{1-\rho_1}{1-\rho_H}$ the waiting time of a customer is the waiting time in an $M/G/1$ queue with no vacations, and with probability $\frac{\rho_L}{1-\rho_H}$ the waiting time of a customer is the sum of the waiting time in an $M/G/1$ queue and the residual length of a vacation with length B_L.

Remark 5.2 Substitution of (2.12) in (5.2) leads to a different expression for $E[e^{-\omega W_H}]$:

$$E[e^{-\omega W_H}] = \frac{1 - \gamma_1(\omega - \lambda_H(1 - \beta_H(\omega)) - \lambda_L(1 - \beta_L(\omega))) + \lambda_L(1 - \beta_L(\omega))E(C)}{(\omega - \lambda_H(1 - \beta_H(\omega)))E(C)}. \quad (5.4)$$

The concept of cycles is not really needed to model the system from the perspective of a type L customer, because for a type L customer the system merely consists of I_{HL} cycles. An I_{HL} cycle is the same as an I_H cycle, discussed in the previous paragraphs, except that it ends when no type H or L customers are left in the system. So the system can be modelled as a nonpriority $M/G/1$ queue with server vacations. The vacation is the intervisit time I_1, plus the service times of all type H customers that have arrived during that intervisit time and their type H descendants. We will denote this extended intervisit time by I_1^* with LST

$$\widetilde{I}_1^*(\omega) = \widetilde{I}_1(\omega + \lambda_H(1 - \pi_H(\omega))).$$

The mean length of I_1^* equals $E(I_1^*) = \frac{E(I_1)}{1-\rho_H}$.

We also have to take into account that a busy period of type L customers might be interrupted by the arrival of type H customers. Therefore the alternative system that we are considering will not contain regular type L customers, but customers still arriving with arrival rate λ_L, whose service time equals the service time of a type L customer in the original model, plus the service times of all type H customers that arrive during this service time, and all of their type H descendants. The LST of the distribution of this extended service time B_L^* is

$$\beta_L^*(\omega) = \beta_L(\omega + \lambda_H(1 - \pi_H(\omega))).$$

This extended service time is often called *completion time* in the literature. In this alternative system, the mean service time of these customers equals $E(B_L^*) = \frac{E(B_L)}{1-\rho_H}$. The fraction of time that the system is serving these customers is $\rho_L^* = \frac{\rho_L}{1-\rho_H} = 1 - \frac{1-\rho_1}{1-\rho_H}$.

Now we use the results from the $M/G/1$ queue with server vacations (starting with the Fuhrmann-Cooper decomposition) to determine the LST of the waiting time distribution for type L customers:

$$E[e^{-\omega W_L}] = \frac{(1 - \rho_L^*)\omega}{\omega - \lambda_L(1 - \beta_L^*(\omega))} \cdot \frac{1 - \widetilde{I}_1^*(\omega)}{\omega E(I_1^*)}$$

$$= \frac{(1 - \rho_1)(\omega + \lambda_H(1 - \pi_H(\omega)))}{\omega - \lambda_L(1 - \beta_L(\omega + \lambda_H(1 - \pi_H(\omega))))} \cdot \frac{1 - \widetilde{I}_1(\omega + \lambda_H(1 - \pi_H(\omega)))}{(\omega + \lambda_H(1 - \pi_H(\omega)))E(I_1)}. \quad (5.5)$$

The last term of (5.5) is the LST of the distribution of the residual intervisit time, plus the time that it takes to serve all type H customers and their type H descendants that arrive during this residual intervisit time. The first term of (5.5) is the LST of the waiting time distribution of a low-priority customer in an $M/G/1$ queue with two priorities, without vacations (see e.g. (3.76) in [6], Chapter 3).

Remark 5.3 The $M/G/1$ queue with two priorities can be viewed as a nonpriority $M/G/1$ queue with vacations, if we consider the waiting time of type L customers. We only need to rewrite the first term of (5.5):

$$
\begin{aligned}
E[e^{-\omega W_{L|M/G/1}}] &= \frac{(1-\rho_1)(\omega + \lambda_H(1-\pi_H(\omega)))}{\omega - \lambda_L(1-\beta_L(\omega + \lambda_H(1-\pi_H(\omega))))} \\
&= \frac{(1-\rho_L^*)\omega}{\omega - \lambda_L(1-\beta_L^*(\omega))} \cdot \frac{1-\rho_1}{1-\rho_L^*} \cdot \frac{\omega + \lambda_H(1-\pi_H(\omega))}{\omega} \\
&= E[e^{-\omega W_{L|M/G/1}^*}]\left[(1-\rho_H) + \rho_H\frac{1-\pi_H(\omega)}{\omega E(BP_H)}\right],
\end{aligned}
$$

where $E[e^{-\omega W_{L|M/G/1}^*}]$ is the LST of the waiting time distribution of a customer in an $M/G/1$ queue where customers arrive at intensity λ_L and have service requirement LST $\beta_L(\omega + \lambda_H(1-\pi_H(\omega)))$. So with probability $1-\rho_H$ the waiting time of a customer is the waiting time in an $M/G/1$ queue with no vacations, and with probability ρ_H the waiting time of a customer is the sum of the waiting time in an $M/G/1$ queue and the residual length of a vacation, which is a busy period of type H customers.

Remark 5.4 Substitution of (2.12) in (5.5) leads to a different expression for $E[e^{-\omega W_L}]$:

$$
\begin{aligned}
E[e^{-\omega W_L}] &= \frac{1 - \gamma_1(\omega - \lambda_L(1-\beta_L(\omega + \lambda_H(1-\pi_H(\omega)))))}{(\omega - \lambda_L(1-\beta_L(\omega + \lambda_H(1-\pi_H(\omega)))))E(C)} \\
&= E[e^{-(\omega - \lambda_L(1-\beta_L(\omega + \lambda_H(1-\pi_H(\omega)))))C_{1,res}}].
\end{aligned} \tag{5.6}
$$

The waiting time of type 2 customers is not affected at all by the fact that Q_1 contains multiple classes of customers, so (2.9) is still valid for $E(e^{-\omega W_2})$.

We will refrain from mentioning the GFs of the marginal queue length distributions here, because they can be obtained by applying the distributional form of Little's law as we have done before.

5.4 Moments

The mean waiting times for high and low priority customers can be found by differentiation of (5.2) and (5.5):

$$
E(W_H) = \frac{\rho_H E(B_{H,res}) + \rho_L E(B_{L,res})}{1-\rho_H} + \frac{1-\rho_1}{1-\rho_H}E(I_{1,res}),
$$

$$
E(W_L) = \frac{\rho_H E(B_{H,res}) + \rho_L E(B_{L,res})}{(1-\rho_H)(1-\rho_1)} + \frac{1}{1-\rho_H}E(I_{1,res}).
$$

19

Differentiation of (5.4) and (5.6) leads to alternative expressions, that can also be found in [15].

$$E(W_H) = \frac{(1-\rho_1)^2}{1-\rho_H} \frac{E(C_1^2)}{2E(C)},$$

$$E(W_L) = \frac{(1-\rho_1)^2}{(1-\rho_H)(1-\rho_1)} \frac{E(C_1^2)}{2E(C)}$$

$$= \left(1 - \frac{\rho_L}{1-\rho_H}\right) \frac{E(C_1^2)}{2E(C)}.$$

6   Example

Consider a polling system with two queues, and assume exponential service times and switch-over times. Suppose that $\lambda_1 = \frac{6}{10}, \lambda_2 = \frac{2}{10}, E(B_1) = E(B_2) = 1, E(S_1) = E(S_2) = 1$. The workload of this polling system is $\rho = \frac{8}{10}$. This example is extensively discussed in [16] where MVA was used to compute mean waiting times and mean residual cycle times for the gated and exhaustive service disciplines.

In this example we show that the performance of this system can be improved by giving higher priority to jobs with smaller service times. We define a threshold t and divide the jobs into two classes: jobs with a service time less than t receive high priority, the other jobs receive low priority. Figures 1 and 2 show the mean waiting time for type 1 customers in the system without priorities, the mean waiting time for type H and type L customers, and the weighted average of these two, as a function of the threshold t. The figures show that a unique optimal threshold exists that minimises the mean weighted waiting time for customers in Q_1. This value depends on the service discipline used and is discussed in [15]. In this example the optimal threshold is 1 for gated, and 1.38 for exhaustive. Figure 1 confirms that the mean waiting times for type H and L customers in the gated model only differ by a constant value: $E(W_L) - E(W_H) = \rho_1 E(C_{1,res})$. For globally gated service no figure is included, because we again have $E(W_L) - E(W_H) = \rho_1 E(C_{1,res})$. The mean residual cycle time is different from the one in the gated model, but this does not affect the optimal threshold which is still $t = 1$. In the exhaustive model we have the following relation: $E(W_L) - E(W_H) = \frac{\rho_1(1-\rho_1)}{1-\rho_H} E(C_{1,res})$. If we increase threshold t, the fraction of customers in Q_1 that receive high priority grows, and so does their mean service time. This means that ρ_H increases as t increases, so $E(W_L) - E(W_H)$ gets bigger, which can be seen in Figure 2. Notice that $\frac{E(W_H)}{E(W_L)} = 1 - \rho_1$, so it does not depend on t.
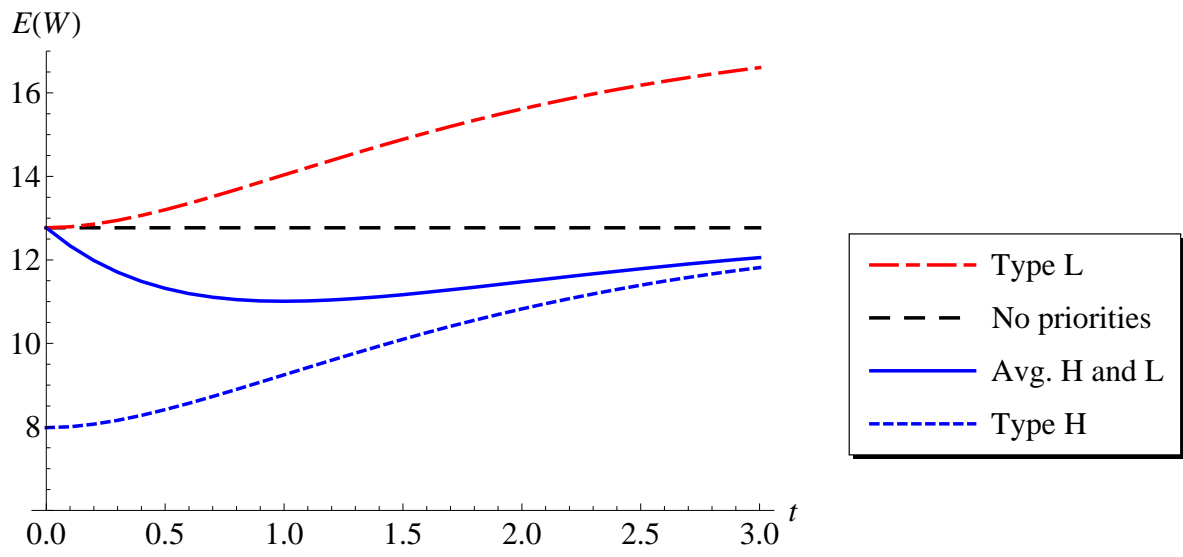
Figure 1: Mean waiting time of customers in Q_1 in the gated polling system, versus threshold t.
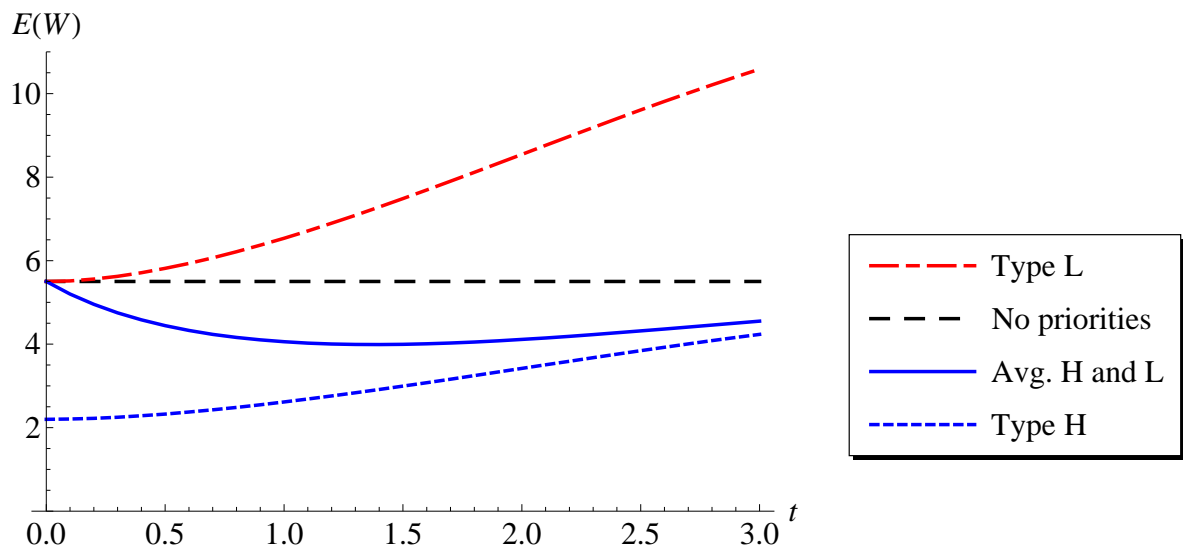

Figure 2: Mean waiting time of customers in Q_1 in the exhaustive polling system, versus threshold t.

References

[1] S. C. Borst. *Polling Systems*, volume 115 of *CWI Tract*. 1996. ISBN 90-6196-467-9.

[2] S. C. Borst and O. J. Boxma. Polling models with and without switchover times. *Operations Research*, 45(4):536 – 543, 1997.

[3] O. J. Boxma. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, 5:185–214, 1989.

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

[5] O. J. Boxma, J. Bruin, and B. H. Fralix. Waiting times in polling systems with various service disciplines. *EURANDOM Report*, 2008.

[6] J. W. Cohen. *The Single Server Queue*. North-Holland, Amsterdam, revised edition, 1982.

[7] R. B. Cooper and G. Murray. Queues served in cyclic order. *The Bell System Technical Journal*, 48(3):675–689, 1969.

[8] M. Eisenberg. Queues with periodic service and changeover time. *Operations Research*, 20(2):440–451, 1972.

[9] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[10] J. Keilson and L. D. Servi. The distributional form of Little's Law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9(4):239–247, 1990.

[11] O. Kella and U. Yechiali. Priorities in $M/G/1$ queue with server vacations. *Naval Research Logistics*, 35:23–34, 1988.

[12] A. Khamisy, E. Altman, and M. Sidi. Polling systems with synchronization constraints. *Annals of Operations Research*, 35:231 – 267, 1992.

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

[14] L. Takács. Two queues attended by a single server. *Operations Research*, 16(3):639–650, 1968.

[15] A. Wierman, E. M. M. Winands, and O. J. Boxma. Scheduling in polling systems. *Performance Evaluation*, 64:1009–1028, 2007.

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