Waiting times in polling systems with various service disciplines

Onno Boxma, Josine Bruin and Brian Fralix
EURANDOM and Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

30th May 2008

Abstract

We consider a polling system of N queues $Q_1, \ldots, Q_N$, cyclically visited by a single server. Customers arrive at these queues according to independent Poisson processes, requiring generally distributed service times. When the server visits Q_i, $i = 1, \ldots, N$, it serves a number of customers according to a certain visit discipline. This discipline is assumed to belong to the class of branching-type disciplines, which includes gated and exhaustive service. The special feature of our study is that, within each queue, we do not restrict ourselves to service in order of arrival (FCFS); we are interested in the effect of different service disciplines, like Last-Come-First-Served, Processor Sharing, Random Order of Service, and Shortest Job First. After a discussion of the joint distribution of the numbers of customers at each queue at visit epochs of the server to a particular queue, we determine the Laplace-Stieltjes transform of the cycle-time distribution, viz., the time between two successive visits of the server to, say, Q_1. This yields the transform of the joint distribution of past and residual cycle time, w.r.t. the arrival of a tagged customer at Q_1. Subsequently concentrating on the case of gated service at Q_1, we use that cycle-time result to determine the (Laplace-Stieltjes transform of the) waiting-time distribution at Q_1.

Next to locally gated visit disciplines, we also consider the globally gated discipline. Again, we consider various non-FCFS service disciplines at the queues, and we determine the (Laplace-Stieltjes transform of the) waiting-time distribution at an arbitrary queue.

1 Introduction

We consider a polling system of N queues $Q_1, \ldots, Q_N$, cyclically visited by a single server. Customers arrive at these queues according to independent Poisson processes, requiring generally distributed service times. Polling systems find many applications in manufacturing, computer-communications, road traffic, maintenance and several other fields, and hence they have been extensively studied. We refer to Takagi [15] and Vishnevskii & Semenova [17] for two surveys. In polling system design several decisions need to be made, for instance one needs to decide on (i) the order of service of the queues, (ii) the number of customers to be served in a queue during a server visit, and (iii) the order of service of the customers within each queue. Regarding (i), a fixed cyclic order is usually assumed, but random polling orders and polling tables have also been studied. With regard to (ii), many visit disciplines have been considered. Well-known visit disciplines are the exhaustive discipline (the server serves the queue until it has become empty), the gated discipline (when the server arrives at a queue to find K customers, it serves exactly those K customers, and no more), and the 1-limited discipline (the server serves just one customer, assuming at least one is present). Hardly any attention has been given to (iii). It is almost invariably assumed that the order of service within each queue is FCFS (First-Come-First-Served). However, in [18] several other service disciplines are considered, like PS (Processor Sharing), ROS (Random Order of Service), LCFS (Last-Come-First-Served), SJF (Shortest Job First), and fixed priorities. Using the recently developed MVA (*Mean*

Value Analysis) approach for polling systems [20], the mean waiting times at the various queues are obtained, for the case of cyclic polling and either the exhaustive or the gated visit discipline at each queue. It is demonstrated in [18] that one can quite easily determine the mean waiting times in this case, and that the effect of the service order may be rather profound, in particular in the case of exhaustive service.

The present paper builds upon [18]. Our goal is to determine the waiting-time *distributions* at the various queues of a cyclic polling system, for several service disciplines. This allows us to study the effect of different service disciplines. We assume the visit disciplines at the various queues belong to the class of branching-type disciplines [11], which includes gated and exhaustive service but which does not include, e.g., 1-limited service. However, we restrict the determination of the waiting-time distribution at some queue to the case that the visit discipline at that particular queue is gated. In a future study we intend to tackle the more difficult problem of deriving the waiting-time distribution at a queue with exhaustive service.

Next to locally gated visit disciplines, we also consider the globally gated discipline, which operates as follows: When the server arrives at Q_1, a gate is closed for *all* queues simultaneously. In the next cycle, the server serves exactly those customers who are located before the gate, i.e., those who were already present when the server arrived at Q_1. Again, we consider various non-FCFS service disciplines at the queues, and we determine the LST (Laplace-Stieltjes transform) of the waiting-time distribution at an arbitrary queue.

The motivation for our work is partly theoretical: we would like to obtain a better insight into the effect of service orders in polling systems, and we would like to develop mathematical tools to accomplish this. However, we are also motivated by the fact that there are many real-world examples where scheduling customer service in a non-FCFS manner would be beneficial. For example, polling models are being used to study the 802.11 and Bluetooth protocols, and scheduling policies at routers and I/O systems in web servers. In such applications, often featuring high workload variability, it may be advantageous to give non-FCFS service. Another example is provided by the *Stochastic Economic Lot Scheduling Problem* (SELSP; see [19] for a survey). In SELSP, a single machine produces multiple standardized products, with setup times between the production of different products. Again, scheduling within the queues is natural and often necessary.

Our approach is as follows. In the case of a branching-type visit discipline at all the queues, Resing [11] has obtained the joint distribution of the number of customers at each queue at visit epochs of the server to a particular queue. His result is easily seen to remain valid when the service order at a queue is not FCFS. Using this queue-length result, we determine the LST of the cycle-time distribution, viz., the time between two successive visits of the server to, say, Q_1. This yields the transform of the joint distribution of past and residual cycle time, w.r.t. the arrival of a tagged customer at Q_1. Finally, we use that cycle-time result to determine the (LST of the) waiting-time distribution at Q_1. Differentiation of this transform gives waiting-time moments, generalizing the mean waiting-time results recently obtained via Mean Value Analysis in [18].

The paper is organized as follows. Section 2 contains a model description. In Section 3 we study the cycle time in the cyclic polling system with a branching-type visit discipline at each queue. These results are then used in Section 4, which contains an analysis of the waiting time distribution in a gated queue, for various service orders like FCFS, LCFS, PS, ROS and SJF. We then show, in Section 5, how our ideas can be applied to polling systems that are served in a globally gated fashion. Finally, Section 6 contains some concluding remarks and mentions topics for further research.

2 Model Description

A single server visits N queues $Q_1, \ldots, Q_N$ in cyclic order. Customers arrive at these queues according to independent Poisson processes $\{N_i(t),\ t \geq 0\}$ with arrival rate λ_i at Q_i, $i = 1, \ldots, N$. The service requirements of customers at Q_i, to be called type-i customers, are i.i.d. (independent, identically distributed) random variables, with distribution $B_i(\cdot)$ and LST $\beta_i(\cdot)$, $i = 1, \ldots, N$; B_i will denote a generic service time at Q_i. The switch-over times of the server from Q_i to Q_{i+1} (Q_{N+1} denoting Q_1) have distribution $S_i(\cdot)$ and LST $\sigma_i(\cdot)$, $i = 1, \ldots, N$; S_i will denote a generic switch-over

time from Q_i. The server even switches among queues when all queues are empty. All interarrival times, service times and switch-over times are assumed to be independent.

When the server visits Q_i, it serves a number of customers according to a certain *visit discipline*. We first concentrate on polling disciplines that belong to the class of branching-type disciplines, as introduced in Resing [11]. This class is characterized by the fact that 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 $h(z_1, \ldots, z_N)$, which may be any N-dimensional probability generating function.*

Important examples of branching-type disciplines are *Exhaustive service* (the server visits a queue until it has emptied the queue) and *Gated service* (during a visit to a queue, the server serves exactly those customers who were present at the beginning of that visit). 1-limited service (the server serves just one customer during a visit, if there is at least one customer present at the beginning of the visit) does *not* belong to the class of branching-type disciplines. Borst [6] gives a slight extension of Property 2.1 that is also held by a globally gated polling system:

Property 2.2 *If there are k_i customers present in Q_i at the beginning of a visit to $Q_{\pi(i)}$ with $\pi(i) \in \{1, ..., N\}$, 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 $h_i(z_1, z_2, ..., z_N)$, which may be any N-dimensional probability generating function.*

When we begin to discuss globally gated visit disciplines, as introduced in Boxma et al. [8], it will be clear that Property 2.2 is satisfied. Under this discipline the server, in a cycle starting at Q_1, only serves the customers that are present at a polling instant at Q_1.

Resing [11] has shown that, if Property 2.1 holds at each queue, the joint queue-length process at polling instants of a fixed queue is a so-called multi-type branching process (MTBP) with immigration. The theory of MTBP (see Athreya and Ney [3] or Resing [11]) now leads to an expression for the generating function of the joint queue length process at polling instants.

For a given visit discipline, we still have to specify the *service discipline* during the visit to a queue. The special feature of the present paper is that we do not restrict ourselves to service in order of arrival (FCFS); we are interested in the effect of different service disciplines (like LCFS, PS, ROS, SJF) on the waiting times of customers.

Define $\rho_i := \lambda_i \mathbb{E}B_i$ the traffic intensity at Q_i, and denote by $\rho := \sum_{i=1}^{N} \rho_i$ the total traffic intensity. We restrict ourselves to the case $\rho < 1$. For the class of polling systems discussed in this paper, this condition guarantees that the vectors of queue lengths at polling epochs and at arbitrary epochs have steady-state distributions.

3 The Cycle-Time in the Branching-Type Polling Model

In this section we determine the LST of the cycle time C for Q_1, i.e., the time between two successive visits of the server to Q_1. In Theorem 3.1 we compute the LST of the conditional cycle time, given the numbers of customers present at all buffers in the polling system at the beginning of the cycle. By unconditioning, the cycle time transform is obtained (Corollary 3.1). But first we present some results from Resing [11], which will be used in the sequel.

In Section 2 we mentioned the class of branching-type visit disciplines (Resing [11]); see Property 2.1. We assume that each queue in our polling system satisfies this property, with generating function $h_i(z_1, \ldots, z_N)$ at Q_i, $i = 1, \ldots, N$. For gated service at Q_i,

$$h_i(z_1, \ldots, z_N) = \beta_i(\sum_{j=1}^{N} \lambda_j(1 - z_j)). \tag{1}$$

For exhaustive service at Q_i, with $\pi_i(\cdot)$ denoting the LST of the busy period of $M/G/1$ queue Q_i in isolation, i.e., an $M/G/1$ queue with arrival rate λ_i and service time distribution $B_i(\cdot)$:

$$h_i(z_1,\ldots,z_N) = \pi_i(\sum_{j\neq i}\lambda_j(1-z_j)). \tag{2}$$

Resing [11] has proven the following. Let $P(z_1,\ldots,z_N)$ denote the GF of the steady-state joint distribution of the numbers of customers $X_1,\ldots,X_N$ in $Q_1,\ldots,Q_N$ at an arbitrary visit beginning of the server at Q_1. Then

$$P(z_1,\ldots,z_N) = \prod_{n=0}^{\infty} g(f_n(z_1,\ldots,z_N)). \tag{3}$$

The functions $f_n(z_1,\ldots,z_N)$ are defined inductively by

$$\begin{aligned} f_0(z_1,\ldots,z_N) &= (z_1,\ldots,z_N), \\ f_n(z_1,\ldots,z_N) &= (f^{(1)}(f_{n-1}(z_1,\ldots,z_N)),\ldots,f^{(N)}(f_{n-1}(z_1,\ldots,z_N))), \end{aligned} \tag{4}$$

where the *off-spring* GFs $f^{(i)}(z_1,\ldots,z_N)$, $i=1,\ldots,N$, are given by

$$f^{(i)}(z_1,\ldots,z_N) = h_i(z_1,\ldots,z_i,f^{(i+1)}(z_1,\ldots,z_N),\ldots,f^{(N)}(z_1,\ldots,z_N)). \tag{5}$$

The *immigration* GF $g(z_1,\ldots,z_N)$ is given by

$$g(z_1,\ldots,z_N) = \prod_{i=1}^{N}\sigma_i(\sum_{k=1}^{i}\lambda_k(1-z_k) + \sum_{k=i+1}^{N}\lambda_k(1-f^{(k)}(z_1,\ldots,z_N))). \tag{6}$$

Let us now turn to the cycle time. Denoting the visit time (time spent in a queue by the server) of Q_i by V_i, $i=1,\ldots,N$, we have

$$C = \sum_{k=1}^{N}(V_k+S_k). \tag{7}$$

Let $\theta_i(\omega)$ represent the LST of the time that the server spends at Q_i due to the presence of one customer there. In the case of gated service, $\theta_i(\omega)=\beta_i(\omega)$, the service time LST; in the case of exhaustive service, $\theta_i(\omega)=\pi_i(\omega)$, the busy-period LST. We also need to introduce the following functions: $\psi_i(\omega)=\omega+\lambda_i(1-\theta_i(\omega))$, $i=1,\ldots,N$, and $\psi_{i,N}(\omega)=\psi_{i+1}(\psi_{i+2}(...(\psi_N(\omega))))$, $i=1,\ldots,N$; here $\psi_{N,N}(\omega)=\omega$.

Theorem 3.1 *The LST of the cycle time C, conditional on the numbers of customers in all queues at the beginning of the cycle, is given by:*

$$\mathbb{E}(\mathrm{e}^{-\omega C}|X_i=m_i, 1\leq i\leq N) = \prod_{i=1}^{N}\sigma_i(\psi_{i,N}(\omega))\theta_i^{m_i}(\psi_{i,N}(\omega)). \tag{8}$$

Proof.
In the formulas below, the condition "$m_1,\ldots,m_k$" denotes $X_1=m_1,\ldots,X_k=m_k$.

$$\begin{aligned} & \mathbb{E}(\mathrm{e}^{-\omega C}|m_1,\ldots,m_N) = \mathbb{E}(\mathrm{e}^{-\omega\sum_{j=1}^{N}(V_j+S_j)}|m_1,\ldots,m_N) \\ = \;& \sigma_N(\omega)\theta_N^{m_N}(\omega)\mathbb{E}(\mathrm{e}^{-(\omega+\lambda_N(1-\theta_N(\omega)))\sum_{j=1}^{N-1}(V_j+S_j)}|m_1,\ldots,m_{N-1}) \\ = \;& \sigma_N(\omega)\theta_N^{m_N}(\omega)\sigma_{N-1}(\psi_N(\omega))\theta_{N-1}^{m_{N-1}}(\psi_N(\omega))\mathbb{E}(\mathrm{e}^{-\psi_{N-1}(\psi_N(\omega))\sum_{j=1}^{N-2}(V_j+S_j)}|m_1,\ldots,m_{N-2}). \end{aligned} \tag{9}$$

Repeating the above iteration procedure finally yields the statement of the theorem.

Deconditioning immediately gives the cycle time LST for Q_1:

Corollary 3.1

$$\mathbb{E}(\mathrm{e}^{-\omega C}) = \prod_{i=1}^{N} \sigma_i(\psi_{i,N}(\omega))\ P(\theta_1(\psi_{1,N}(\omega)), \ldots, \theta_i(\psi_{i,N}(\omega)), \ldots, \theta_N(\psi_{N,N}(\omega))). \tag{10}$$

A similar type of expression can be given for polling systems that satisfy Property 2.2. Notice that the state of the polling system at the embedded instants when the server begins its visit at Q_1 forms a MTBP, either with or without immigration, depending on the setup times. Unlike Resing's property, however, this is not true at any of the other visit epochs in a globally gated visit discipline. Even so, in the globally gated case one can still easily compute the steady-state distribution at any other epoch by knowing the steady-state distribution at Q_1.

3.1 The Biased Cycle Length

Throughout this paper, we will be interested in the distribution of various components of the steady-state cycle time, given that a particular type of customer arrived during such a cycle. Knowing that such an arrival occurred will bias the length of the cycle, and this must be accounted for.

Throughout this paper, our cycles will always begin at the moment the server begins to work on jobs present at Q_1. Assuming that a customer arrives to Q_1 during a cycle, let C^*, C^p and C^r denote the total biased cycle length, the amount of time between the beginning of the cycle and the arrival of the tagged customer to Q_1, and the amount of time between the arrival of such a tagged customer and the end of the cycle, respectively. Clearly $C^* = C^p + C^r$, and the tagged customer will not be served until the next cycle begins. C^* will also be referred to throughout parts of the paper as the cycle time of the tagged customer. When we look at the globally gated case, we will assume that all gates in the system are synchronized with the gate at Q_1, and so this same choice of cycles will be appropriate when we are interested in the sojourn time distribution of customers that arrive at Q_i, for $1 \leq i \leq N$.

Our goal is to now relate their distributions to the distribution of C, which is the steady-state unbiased cycle-length. It is known that, conditional on C^*, the distribution of C^p is uniform on $[0, C^*]$. Furthermore, it is also known in the literature (see, for example, [16]) that

$$\mathrm{d}P(C^* \leq x) = \frac{x\mathrm{d}P(C \leq x)}{\mathbb{E}(C)}. \tag{11}$$

From this result, it is then immediately clear that

$$\mathbb{E}(C^*) = \frac{\mathbb{E}(C^2)}{\mathbb{E}(C)},$$

$$\mathbb{E}(C^p) = \mathbb{E}(C^r) = \frac{\mathbb{E}(C^2)}{2\mathbb{E}(C)}.$$

Moreover, we can use (11) to compute the joint LST of C^p and C^r:

$$\int_{t=0}^{\infty}\int_{u=0}^{\infty} e^{-at}e^{-bu}\mathrm{d}P(C^p \leq t; C^r \leq u) \quad = \quad \frac{\mathbb{E}[e^{-aC}] - \mathbb{E}[e^{-bC}]}{(b-a)\mathbb{E}(C)}. \tag{12}$$

The derivation of these last results is known, and also beyond the scope of the paper so a discussion of their derivation has been omitted. To appease the interested reader, we will mention that these results can be derived through the use of Palm theory, which can be used to capture the biases that are mentioned above. The Palm framework allows us to work with the fact that, under the Palm measure induced by the point process consisting of the times at which a cycle begins, the sequence of cycle lengths formed in the stationary version of this polling system forms a stationary sequence, but does not form an i.i.d. sequence. If this were true, we could instead have made use of well-known results from renewal theory: for instance, the reader may recognize that $\mathbb{E}\left(C^r\right)$ has the same form as the first moment of the stationary residual lifetime from a renewal process. References

on Palm theory are numerous: examples of more recent references include [5] and [13] (both focus on applications in queueing), along with [16].

Throughout our analysis, we will also make use of what is known in the literature as the stationary-excess operator R (see, for instance, Abate and Whitt [1]), which is defined in the following way: for a given nonnegative random variable X,

$$P(R_X \leq t) = \frac{1}{\mathbb{E}(X)} \int_0^t P(X > s)\mathrm{d}s, \qquad t \geq 0.$$

We will also be applying this operator multiple times to a given random variable, and to denote this we will use the abbreviation $R_{X,n}$, where $R_{X,0} = X$, $R_{X,1} = R_X$, and for any $n \geq 0$, $R_{X,n+1} = R_{R_{X,n}}$.

The reader should note that for cycle times, R_C and C^r will both be used throughout various parts of the paper, even though they both have the same distribution. The former will typically be used within computations, while the latter will exclusively be used to represent a particular residual cycle time observed by a tagged customer.

Now we are ready to state the following lemma, which will prove to be useful while computing the first and second moments of many of the types of sojourn times considered in this paper.

Lemma 3.1 *For $a, b \geq 0$, $a \neq b$,*

$$\frac{\mathbb{E}(e^{-aC}) - \mathbb{E}(e^{-bC})}{(b-a)\mathbb{E}(C)} = \sum_{k=0}^{n} (-a)^k \frac{\mathbb{E}(R_C^k)}{k!} \mathbb{E}(e^{-bR_{C,k+1}}) + (-a)^{n+1} \frac{\mathbb{E}(R_C^{n+1})}{(n+1)!} \left[\frac{\mathbb{E}(e^{-aR_{C,n+1}}) - \mathbb{E}(e^{-bR_{C,n+1}})}{(b-a)\mathbb{E}(R_{C,n+1})} \right]. \quad (13)$$

Proof The LST of R_C is known, and can be found in, for instance, [1]:

$$\mathbb{E}(e^{-\omega R_C}) = \frac{1 - \mathbb{E}(e^{-\omega C})}{\omega \mathbb{E}(C)}. \quad (14)$$

Equation (13) then follows from (12) and (14):

$$\begin{aligned} \frac{\mathbb{E}(e^{-aC}) - \mathbb{E}(e^{-bC})}{(b-a)\mathbb{E}(C)} &= \frac{1 - \mathbb{E}(e^{-bC})}{(b-a)\mathbb{E}(C)} - \frac{1 - \mathbb{E}(e^{-aC})}{(b-a)\mathbb{E}(C)} \\ &= \frac{b}{b-a}\mathbb{E}(e^{-bR_C}) - \frac{a}{b-a}\mathbb{E}(e^{-aR_C}) \\ &= \mathbb{E}(e^{-bR_C}) - a\mathbb{E}(R_C)\left[\frac{\mathbb{E}(e^{-aR_C}) - \mathbb{E}(e^{-bR_C})}{(b-a)\mathbb{E}(R_C)} \right]. \end{aligned}$$

At this point we begin to see a pattern: for any $n \geq 1$,

$$\begin{aligned} \frac{\mathbb{E}(e^{-aC}) - \mathbb{E}(e^{-bC})}{(b-a)\mathbb{E}(C)} &= \sum_{k=0}^{n} (-a)^k \left[\prod_{j=1}^{k} \mathbb{E}(R_{C,j}) \right] \mathbb{E}(e^{-bR_{C,k+1}}) \\ &+ (-a)^{n+1} \left[\prod_{j=1}^{n+1} \mathbb{E}(R_{C,j}) \right] \left[\frac{\mathbb{E}(e^{-aR_{C,n+1}}) - \mathbb{E}(e^{-bR_{C,n+1}})}{(b-a)\mathbb{E}(R_{C,n+1})} \right] \quad (15) \end{aligned}$$

where products of the form $\prod_{j=1}^{0}$ will be understood to equal 1.

The proof will be complete once we compute each of the products found in (15). Notice that

$$\begin{aligned}\mathbb{E}(e^{-\omega C}) &= 1-\omega\mathbb{E}(C)\mathbb{E}(e^{-\omega R_C}) \\ &= 1-\omega\mathbb{E}(C)+\omega^2\mathbb{E}(C)\mathbb{E}(R_C)\mathbb{E}(e^{-\omega R_{C,2}}) \\ &= \ldots = \sum_{k=0}^{\infty}(-\omega)^k\prod_{m=0}^{k-1}\mathbb{E}(R_{C,m}).\end{aligned}$$

Therefore

$$\prod_{m=0}^{k-1}\mathbb{E}(R_{C,m}) = \frac{\mathbb{E}(C^k)}{k!},$$

so for each $k \geq 1$,

$$\prod_{m=1}^{k}\mathbb{E}(R_{C,m}) = \frac{\mathbb{E}(C^{k+1})}{(k+1)!\mathbb{E}(C)} = \frac{\mathbb{E}(R_C^k)}{k!}.$$

This proves (13). $\diamondsuit$

4 Sojourn times at a gated queue

In this section we will be interested in the sojourn time distribution of a tagged customer that visits a gated queue, at a time when the system is in steady-state. We will first derive the LST of the sojourn time for policies that are nonanticipating, i.e. those that are not influenced by the service times of customers in the queue. For anticipating policies, it will be more useful to derive the conditional LST of the sojourn time, given the amount of service brought to the buffer by the tagged customer.

4.1 Nonanticipating policies

4.1.1 First-Come-First-Served

We will begin this section by computing the LST of the sojourn time of a tagged customer that visits a queue, whose customers are served in accordance to a FCFS scheduling policy. If $B_{i,k}$ denotes the service time of the k^{th} customer that arrives to Q_i during the cycle time of the tagged customer and $N_i(t)$ denotes the number of type-i customers that arrive during a time interval of length t, then clearly

$$T_{FCFS} = C^r + \sum_{k=1}^{N_1(C^p)+1} B_{1,k}.$$

Here T_{FCFS} represents the sojourn time of a tagged customer that arrives to Q_1 while the system is in equilibrium. Notice that we have suppressed the fact that we're referring to Q_1 in our notation for T_{FCFS}, and we will continue to do so throughout the rest of this section. The reason why we will follow this practice is because, for gated systems, the gate at Q_1 only moves at the moment the server begins working there. This allows us to conclude that the waiting time distribution has the same form for all other Q_i that operate under a gated scheme; the only difference would involve considering cycles that begin at the moment the server begins working at Q_i instead of Q_1.

After conditioning on the past and residual cycle lengths, we see that

$$\begin{aligned}
\mathbb{E}(e^{-\omega T_{FCFS}}) &= \mathbb{E}(e^{-\omega(C^r+\sum_{k=1}^{N_1(C^p)+1} B_{1,k})}) \\
&= \int_0^\infty \int_0^\infty e^{-\omega u} \sum_{n=0}^{\infty} \beta_1(\omega)^{n+1} \frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!} \mathrm{d}P(C^p \leq t, C^r \leq u) \\
&= \beta_1(\omega) \int_0^\infty \int_0^\infty e^{-\omega u} e^{-\lambda_1(1-\beta_1(\omega))t} \mathrm{d}P(C^p \leq t, C^r \leq u) \\
&= \beta_1(\omega) \left[\frac{\mathbb{E}(e^{-\lambda_1(1-\beta_1(\omega))C}) - \mathbb{E}(e^{-\omega C})}{\mathbb{E}(C)(\omega - \lambda_1(1-\beta_1(\omega)))} \right] \\
&= \beta_1(\omega) \mathbb{E}(e^{-\omega D_{FCFS}}), \qquad (16)
\end{aligned}$$

where D_{FCFS} denotes the delay of the tagged customer.

The first moment of T_{FCFS} is well-known, and can be found in many places throughout the polling literature (see, for instance, [7] or [14]):

$$\mathbb{E}(T_{FCFS}) = \mathbb{E}(B_1) + \mathbb{E}(C^r)(1+\rho_1).$$

We will now show how to efficiently use (16) to compute both the first and second moment of T_{FCFS}. By applying Lemma 3.1 to (16), we see that for each $n \geq 1$, when $\omega \downarrow 0$,

$$\mathbb{E}(e^{-\omega D_{FCFS}}) = \sum_{k=0}^{n} (-1)^k (\lambda_1(1-\beta_1(\omega))^k \frac{\mathbb{E}(R_C^k)}{k!} \mathbb{E}(e^{-\omega R_{C,k+1}}) + \mathcal{O}(\omega^{n+1}).$$

Due to the fact that

$$\lambda_1(1-\beta_1(\omega)) = \rho_1\omega - \lambda_1 \frac{\mathbb{E}(B_1^2)}{2}\omega^2 + \mathcal{O}(\omega^3), \qquad \omega \downarrow 0$$

we find that the LST of D_{FCFS} can also be expressed in the following way: as $\omega \downarrow 0$,

$$\begin{aligned}
\mathbb{E}(e^{-\omega D_{FCFS}}) &= \mathbb{E}(e^{-\omega R_C}) - \lambda_1(1-\beta_1(\omega))\mathbb{E}(R_C)\mathbb{E}(e^{-\omega R_{C,2}}) + (\lambda_1(1-\beta_1(\omega)))^2 \frac{\mathbb{E}(R_C^2)}{2}\mathbb{E}(e^{-\omega R_{C,3}}) + \mathcal{O}(\omega^3) \\
&= 1 - \mathbb{E}(R_C)(1+\rho_1)\omega + \left[\lambda_1 \frac{\mathbb{E}(B_1^2)}{2}\mathbb{E}(R_C) + \frac{\mathbb{E}(R_C^2)}{2}\left[1+\rho_1+\rho_1^2\right] \right] \omega^2 + \mathcal{O}(\omega^3).
\end{aligned}$$

Thus,

$$\mathbb{E}(D_{FCFS}^2) = \lambda_1 \mathbb{E}(B_1^2)\mathbb{E}(R_C) + \mathbb{E}(R_C^2)\left(1+\rho_1+\rho_1^2\right),$$

which also implies that

$$\mathbb{E}(T_{FCFS}^2) = \lambda_1 \mathbb{E}(B_1^2)\mathbb{E}(R_C) + \mathbb{E}(R_C^2)\left[1+\rho_1+\rho_1^2\right] + 2\mathbb{E}(B_1)\mathbb{E}(R_C)(1+\rho_1) + \mathbb{E}(B_1^2).$$

4.1.2 Last-Come-First-Served

The LST of the sojourn time of a tagged customer under the Last-Come-First-Served (LCFS) discipline has a form that is similar to the LST of T_{FCFS}. Under LCFS, all of the workload that arrives to Q_1 after the tagged customer, yet during the cycle time of the tagged customer, will be processed before him, and so

$$T_{LCFS} = C^r + \sum_{k=1}^{N_1(C^r)+1} B_{1,k}.$$

In this case, $B_{1,k}$ denotes the amount of work brought to Q_1 by the k^{th} customer that arrives during C^r. By performing a similar calculation as above, we see that the LST of T_{LCFS} is just

$$\begin{aligned}
\mathbb{E}(e^{-\omega T_{LCFS}}) &= \mathbb{E}(e^{-\omega(C^r+\sum_{k=1}^{N_1(C^r)+1} B_{1,k})}) \\
&= \int_0^\infty e^{-\omega t} \sum_{n=0}^{\infty} \beta_1(\omega)^{n+1} \frac{(\lambda_1 t)^n e^{-\lambda_1 t}}{n!} \mathrm{d}P(C^r \leq t) \\
&= \beta_1(\omega)\mathbb{E}(e^{-(\omega+\lambda_1(1-\beta_1(\omega)))C^r}) \\
&= \beta_1(\omega)\left[\frac{1-\mathbb{E}(e^{-(\omega+\lambda_1(1-\beta_1(\omega)))C})}{\mathbb{E}(C)(\omega+\lambda_1(1-\beta_1(\omega)))}\right] \\
&= \beta_1(\omega)\mathbb{E}(e^{-\omega D_{LCFS}}). \qquad (17)
\end{aligned}$$

The form of this LST makes it easy to compute moments without resorting to either differentiation or the use of Lemma 3.1. Clearly,

$$\omega + \lambda_1(1-\beta_1(\omega)) = (1+\rho_1)\omega - \lambda_1 \frac{\mathbb{E}(B_1^2)\omega^2}{2} + \mathcal{O}(\omega^3), \qquad \omega \downarrow 0$$

and this simple fact will allow us to rewrite the LST of D_{LCFS} in the following way:

$$\begin{aligned}
\mathbb{E}(e^{-\omega D_{LCFS}}) &= \sum_{n=1}^{\infty}(-1)^{n-1}((1+\rho_1)\omega - \lambda_1\frac{\mathbb{E}(B_1^2)}{2}\omega^2)^{n-1}\frac{\mathbb{E}(C^n)}{n!\mathbb{E}(C)} \\
&= 1 - \left((1+\rho_1)\omega - \frac{\lambda_1\mathbb{E}(B_1^2)\omega^2}{2}\right)\mathbb{E}(R_C) + \left((1+\rho_1)\omega - \frac{\lambda_1\mathbb{E}(B_1^2)\omega^2}{2}\right)^2\frac{\mathbb{E}(R_C^2)}{2} + \mathcal{O}(\omega^3) \\
&= 1 - (1+\rho_1)\mathbb{E}(R_C)\omega + \left[\lambda_1\frac{\mathbb{E}(B_1^2)\mathbb{E}(R_C)}{2} + (1+\rho_1)^2\frac{\mathbb{E}(R_C^2)}{2}\right]\omega^2 + \mathcal{O}(\omega^3), \qquad \omega \downarrow 0.
\end{aligned}$$

Hence, the first two moments of this random variable are just

$$\mathbb{E}(D_{LCFS}) = (1+\rho_1)\mathbb{E}(R_C)$$

and

$$\mathbb{E}(D_{LCFS}^2) = \lambda_1\mathbb{E}(B_1^2)\mathbb{E}(R_C) + (1+\rho_1)^2\mathbb{E}(R_C^2).$$

From this, we can now compute the first and second moments of the sojourn time:

$$\mathbb{E}(T_{LCFS}) = \mathbb{E}(B_1) + \mathbb{E}(R_C)(1+\rho_1) = \mathbb{E}(T_{FCFS}),$$

and

$$\begin{aligned}
\mathbb{E}(T_{LCFS}^2) &= \lambda_1\mathbb{E}(B_1^2)\mathbb{E}(R_C) + (1+\rho_1)^2\mathbb{E}(R_C^2) + 2\mathbb{E}(B_1)\mathbb{E}(R_C)(1+\rho_1) + \mathbb{E}(B_1^2) \\
&= \mathbb{E}(T_{FCFS}^2) + \rho_1\mathbb{E}(R_C^2).
\end{aligned}$$

We see that the second moment of T_{LCFS} is larger than the one of T_{FCFS}, which proves that the sojourn time under LCFS is actually more variable than its FCFS counterpart.

4.1.3 Random Order of Service

The final nonanticipating policy that we will analyze in this paper is known as the Random Order of Service (ROS) policy. Unfortunately, the LST of the sojourn time under this policy isn't as nice as the previous cases, as the reader will see from the derivation below. To compute the LST, let's imagine that the server immediately creates an order in which he will serve the customers currently waiting in the buffer, and let U denote the position of the tagged customer within this ordering. Clearly U is a uniform random variable on $1, 2, ..., N_1(C^p) + N_1(C^r) + 1$. Therefore,

$$\begin{aligned}
\mathbb{E}(e^{-\omega T_{ROS}}) &= \mathbb{E}(e^{-\omega(C^r + \sum_{k=1}^{U} B_{1,k})}) \\
&= \int_0^\infty \int_0^\infty e^{-\omega u} \mathbb{E}(e^{-\omega \sum_{k=1}^{U} B_{1,k}} | C^p = t, C^r = u) \mathrm{d}P(C^p \leq t, C^r \leq u) \\
&= \beta_1(\omega) \int_0^\infty \int_0^\infty e^{-\omega u} \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{\beta_1(\omega)^k}{n+1} \frac{(\lambda_1(t+u))^n e^{-\lambda_1(t+u)}}{n!} \mathrm{d}P(C^p \leq t, C^r \leq u).
\end{aligned}$$

Notice that

$$(1 - \beta_1(\omega)) \sum_{k=0}^{n} \frac{\beta_1(\omega)^k}{n+1} = \frac{1 - \beta_1(\omega)^{n+1}}{n+1} = \int_{\beta_1(\omega)}^{1} x^n \mathrm{d}x$$

and so

$$\begin{aligned}
\mathbb{E}(e^{-\omega T_{ROS}}) &= \frac{\beta_1(\omega)}{1 - \beta_1(\omega)} \int_0^\infty \int_0^\infty \int_{\beta_1(\omega)}^{1} e^{-\omega u} e^{-\lambda_1(1-x)(t+u)} \mathrm{d}x \mathrm{d}P(C^p \leq t, C^r \leq u) \quad (18) \\
&= \frac{\beta_1(\omega)}{\mathbb{E}(C)(1 - \beta_1(\omega))} \int_{\beta_1(\omega)}^{1} \frac{(\mathbb{E}(e^{-\lambda_1(1-x)C}) - \mathbb{E}(e^{-(\omega + \lambda_1(1-x))C}))}{\omega} \mathrm{d}x. \quad (19)
\end{aligned}$$

What is nice about this expression is the appearance of the ω term in the denominator. We will show in Section 5 that this form allows us to manipulate the transform in a different manner, in order to compute higher moments of the sojourn time.

Again, the mean sojourn time in this case is the same as in the other cases, i.e.

$$\mathbb{E}(T_{ROS}) = \mathbb{E}(B_1) + \mathbb{E}(R_C)(1 + \rho_1).$$

The second moment is just

$$\begin{aligned}
\mathbb{E}(T_{ROS}^2) &= \lambda_1 \mathbb{E}(B_1^2) \mathbb{E}(R_C) + 2\mathbb{E}(B_1)\mathbb{E}(R_C)(1 + \rho_1) + \frac{\mathbb{E}(R_C^2)}{2}\left(2 + 3\rho_1 + 2\rho_1^2\right) + \mathbb{E}(B_1^2) \\
&= E[T_{FCFS}^2] + \frac{\rho_1}{2}\mathbb{E}(R_C^2).
\end{aligned}$$

Clearly $\mathbb{E}(T_{LCFS}^2) > \mathbb{E}(T_{ROS}^2) > \mathbb{E}(T_{FCFS}^2)$. Indeed, our analysis has shown that FCFS seems to perform the best among all of the nonanticipating policies considered here. One cannot help but wonder if a stronger ordering relationship can be established among the distributions of these sojourn times; this is a question that we plan to investigate in a future paper.

4.2 Anticipating policies

Now we will be interested in analyzing policies that use information about the size of the jobs in the system in order to decide how and when various jobs are served. The two policies that we consider in this section are the Shortest Job First (SJF) and the Processor Sharing (PS) policies. Suppose that when the server arrives to Q_1, it orders the jobs in increasing order, i.e., let $B_{1,(k)}$ denote the k^{th}-smallest job in Q_1, where $1 \leq k \leq N_1(C^p) + N_1(C^r) + 1$. Then it is clear that

$$T_{PS} = C^r + \sum_{k=1}^{U} (N_1(C^p) + N_1(C^r) + 1 - k + 1)(B_{1,(k)} - B_{1,(k-1)}) \tag{20}$$

and

$$T_{SJF} = C^r + \sum_{k=1}^{U} B_{1,(k)}. \tag{21}$$

Here we use the convention that $B_{1,(0)} = 0$ with probability one.

Unfortunately, working with order statistics is often a cumbersome task, and so we will not be able to explicitly use these expressions when we compute the LST of the sojourn time, for a general service time distribution. The reader may notice, however, that if the services are exponentially distributed, then T_{PS} is equal in distribution to T_{ROS}. This follows from the following simple property of exponential random variables (see, for instance, page 19 of [9]):

Proposition 4.1 *Let X_1, X_2, ..., X_n denote a collection of n independent and identically distributed exponential random variables with rate α. If $X_{(k)}$ denotes the k^{th}-smallest random variable among the population, then the n variables $X_{(k)} - X_{(k-1)}$, $1 \leq k \leq n$ (set $X_{(0)} = 0$) are independent and $X_{(k)} - X_{(k-1)}$ is exponentially distributed with rate $(n-k+1)\alpha$.*

Even in this case, however, the distribution of T_{SJF} is still difficult to handle. To get around this dilemma, we will need to condition on the service time of the tagged customer.

4.2.1 Conditioning on the service time

Suppose that a tagged customer arrives to Q_1 with an amount of work x. Then the sojourn time of the customer depends on three things: the remaining amount of time it takes for the server to reach Q_1, and the amount of work brought by customers that arrived before, and after, the tagged customer to Q_1. For a given scheduling policy Γ, let $T_\Gamma(x)$ denote the sojourn time of a tagged customer, conditional on the amount of work it brings to the system. Under many policies, this random variable can be written in the following way:

$$T_\Gamma(x) = x + C^r + \sum_{k=1}^{N_1(C^p)} g_1(B_{1,k}, x) + \sum_{m=1}^{N_1(C^r)} g_2(B_{1,m}, x). \tag{22}$$

Here $g_i : [0,\infty) \times [0,\infty) \to \mathbb{R}$, $i = 1, 2$ are functions that capture how the tagged customer's sojourn time is affected by customers that arrive before and after him, respectively. For example, if Γ represents the FCFS policy, $g_1(y, x) = y$ and $g_2(y, x) = 0$, since all customers arriving ahead of the tagged customer will be served first, and no customer arriving afterward will affect the sojourn time. The reader should of course keep in mind that g_i could depend on x as well (such as when analyzing the SJF case), which is why we allow g_i to depend on x.

Modeling the sojourn times in this manner will allow us to easily compute the LST of $T_\Gamma(x)$. For $\omega \geq 0$, we find that

$$\begin{aligned}
\mathbb{E}(e^{-\omega T_\Gamma(x)}) &= \mathbb{E}(e^{-\omega(C^r + \sum_{k=1}^{N_1(C^p)} g_1(B_{1,k},x) + \sum_{m=1}^{N_1(C^r)} g_2(B_{1,m},x))}) \\
&= \int_0^\infty \int_0^\infty e^{-\omega(x+v+\sum_{k=1}^{N_1(u)} g_1(B_{1,k},x) + \sum_{k=1}^{N_1(v)} g_2(B_{1,k},x))} \mathrm{d}P(C^p \leq u, C^r \leq v) \\
&= e^{-\omega x} \int_0^\infty \int_0^\infty e^{-\omega v} e^{-\lambda_1(1-\phi_1(\omega,x))u} e^{-\lambda_1(1-\phi_2(\omega,x))v} \mathrm{d}P(C^p \leq u, C^r \leq v),
\end{aligned}$$

where $\phi_i(\omega, x) = \mathbb{E}(e^{-\omega g_i(B_1,x)})$, for $i = 1, 2$, and $B_{\phi,i}$ denotes a random variable with LST ϕ_i. Therefore,

$$\mathbb{E}(e^{-\omega T_\Gamma(x)}) = e^{-\omega x} \frac{\mathbb{E}(e^{-\lambda_1(1-\phi_1(\omega,x))C}) - \mathbb{E}(e^{-(\omega+\lambda_1(1-\phi_2(\omega,x)))C})}{\mathbb{E}(C)(\omega + \lambda_1(\phi_1(\omega, x) - \phi_2(\omega, x)))}. \quad (23)$$

Showing that the SJF policy fits within this framework is simple: just set $g_1(y, x) = g_2(y, x) = y\mathbf{1}(y \leq x)$. This follows from the fact that all, and only all, jobs present that are of a size smaller than x will be served before the tagged customer.

The PS discipline can also be modeled in this manner. Suppose we choose $g_1(y, x) = g_2(y, x) = \min(y, x)$. Then

$$\begin{aligned}
& x + C^r + \sum_{k=1}^{N_1(C^p)} \min(B_{1,k}, x) + \sum_{m=1}^{N_1(C^r)} \min(B_{1,m}, x) \\
= \; & C^r + \sum_{k=1}^{U-1} B_{1,(k)} + (N_1(C^p) + N_1(C^r) + 1 - U + 1)x \\
= \; & C^r + \sum_{k=1}^{U-1} \sum_{l=1}^{k} (B_{1,(l)} - B_{1,(l-1)}) \\
+ \; & (N_1(C^p) + N_1(C^r) + 1 - U + 1)(x - B_{1,(U-1)}) \\
+ \; & \sum_{l=1}^{U-1} (N_1(C^p) + N_1(C^r) + 1 - U + 1)(B_{1,(l)} - B_{1,(l-1)}) \\
= \; & C^r + \sum_{l=1}^{U-1} \sum_{k=l}^{U-1} (B_{1,(l)} - B_{1,(l-1)}) \\
+ \; & (N_1(C^p) + N_1(C^r) + 1 - U + 1)(x - B_{1,(U-1)}) \\
+ \; & \sum_{l=1}^{U-1} (N_1(C^p) + N_1(C^r) + 1 - U + 1)(B_{1,(l)} - B_{1,(l-1)}) \\
= \; & C^r + \sum_{l=1}^{U-1} (U - 1 - l + 1)(B_{1,(l)} - B_{1,(l-1)}) \\
+ \; & (N_1(C^p) + N_1(C^r) + 1 - U + 1)(x - B_{1,(U-1)}) \\
+ \; & \sum_{l=1}^{U-1} (N_1(C^p) + N_1(C^r) + 1 - U + 1)(B_{1,(l)} - B_{1,(l-1)}) \\
= \; & C^r + \sum_{l=1}^{U-1} (N_1(C^p) + N_1(C^r) + 1 - l + 1)(B_{1,(l)} - B_{1,(l-1)}) \\
+ \; & (x - B_{1,(U-1)})(N_1(C^p) + N_1(C^r) + 1 - U + 1)
\end{aligned}$$

and so it follows from (20) that this choice of g_1 and g_2 correctly models the sojourn time of the tagged customer under PS. In the next two subsections, we will treat both the PS and the SJF cases in further detail.

4.2.2 Processor Sharing

Now we are ready to analyze the sojourn time of a tagged customer at Q_1, which utilizes the processor-sharing rule when providing everyone with service.

It should be noted that similar models, for the single-queue case, have been studied in the literature before. Rege and Sengupta [10], for instance, derive various performance measures for what is known as a gated $M/M/1$ queue, which operates as follows: the server provides service to

at most $m \geq 1$ customers, in a processor-sharing fashion. Once a group has been served, the server then begins serving the next (up to) m waiting customers, and so on. The works of Avi-Itzhak and Halfin [4] and Rietman and Resing [12] focus on various extensions of this model. In particular, [4] considers a gated $M/G/1$ queue, and they consider not only the processor-sharing discipline, but other "conservative" scheduling disciplines, which include FCFS, LCFS, and ROS. They also analyze the same type of model in [12], but they go a step further by deriving the joint distribution of both the amount of time a customer spends on both sides of the gate, and the number of customers on both sides of the gate.

We will now begin our calculation of the conditional LST of the sojourn time under PS. From (23), we see that

$$\mathbb{E}(e^{-\omega T_{PS}(x)}) = e^{-\omega x}\frac{\mathbb{E}(e^{-\lambda_1(1-\phi(\omega,x))C}) - \mathbb{E}(e^{-(\omega+\lambda_1(1-\phi(\omega,x)))C})}{\omega\mathbb{E}(C)}. \tag{24}$$

where $\phi(\omega, x) = \mathbb{E}(e^{-\omega\min(B_k,x)})$. This expression is nice, in that it is given in terms of the LST of the cycle time. To find the unconditional LST of T_{PS}, we only need to integrate with respect to the service time distribution, however in many cases this transform will not be tractable.

Let us assume for now that B_1 is exponential with parameter μ_1. Then

$$\mathbb{E}(e^{-\omega T_{PS}}) = \frac{1}{\omega\mathbb{E}(C)}\int_0^\infty e^{-\omega x}\left[\mathbb{E}(e^{-\lambda_1\frac{\omega}{\mu_1+\omega}(1-e^{-(\mu_1+\omega)x})C}) - \mathbb{E}(e^{-(\omega+\frac{\lambda_1\omega}{\mu_1+\omega}(1-e^{-(\mu_1+\omega)x})C)})\right]\mu_1 e^{-\mu_1 x}\mathrm{d}x.$$

Fortunately this integral can be simplified. First of all,

$$\begin{aligned}
\int_0^\infty e^{-\omega x}\mathbb{E}(e^{-\frac{\lambda_1\omega C}{\mu_1+\omega}(1-e^{-(\mu_1+\omega)x})})\mu_1 e^{-\mu_1 x}\mathrm{d}x &= \mathbb{E}\left(e^{-\frac{\lambda_1\omega C}{\mu_1+\omega}}\int_0^\infty e^{\frac{\lambda_1\omega C}{\mu_1+\omega}e^{-(\mu_1+\omega)x}}\mu_1 e^{-(\mu_1+\omega)x}\mathrm{d}x\right)\\
&= \mathbb{E}\left(e^{-\frac{\lambda_1\omega C}{\mu_1+\omega}}\frac{\mu_1}{\lambda_1\omega C}\int_{-\frac{\lambda_1\omega C}{\mu_1+\omega}}^0 e^{-u}\mathrm{d}u\right)\\
&= \mathbb{E}\left(\frac{1}{\rho_1\omega C}(1-e^{-\frac{\lambda_1\omega C}{\mu_1+\omega}})\right).
\end{aligned}$$

Now that we have performed this calculation, it is easy to see that

$$\int_0^\infty e^{-\omega x}\mathbb{E}(e^{-(\omega+\frac{\lambda_1\omega C}{\mu_1+\omega}(1-e^{-(\mu_1+\omega)x}))})\mu_1 e^{-\mu_1 x}\mathrm{d}x = \mathbb{E}\left(\frac{e^{-\omega C}}{\rho_1\omega C}(1-e^{-\frac{\lambda_1\omega C}{\mu_1+\omega}})\right),$$

and so

$$\begin{aligned}
\mathbb{E}[e^{-\omega T_{PS}}] &= \frac{1}{\rho_1\omega^2\mathbb{E}(C)}\mathbb{E}\left(\frac{(1-e^{-\lambda_1\frac{\omega}{\mu_1+\omega}C})(1-e^{-\omega C})}{C}\right)\\
&= \frac{1}{\rho_1\omega^2\mathbb{E}(C)}\mathbb{E}\left(\int_0^{\frac{\lambda_1\omega}{\mu_1+\omega}}(1-e^{-\omega C})e^{-Cy}\mathrm{d}y\right)\\
&= \frac{1}{\rho_1\omega^2\mathbb{E}(C)}\int_0^{\frac{\lambda_1\omega}{\mu_1+\omega}}(f(y)-f(y+\omega))\mathrm{d}y
\end{aligned}$$

where

$$f(y) = \mathbb{E}(e^{-yC}).$$

The reason why we choose to represent this transform as an integral is because it will help us obtain higher moments of the sojourn time. The approach involves a couple of Taylor series approximations, and it will be given in Section 5.

Throughout the rest of this subsection, we will be interested in how much we can conclude about the distribution of the sojourn time, without assuming that the service times are exponential. An

application of Wald's equality can be used to compute the first moment of $T_{PS}(x)$ by using (22), and it can also be found in [18].

In this case,

$$\mathbb{E}(T_{PS}(x)) = x + \mathbb{E}(C^r)(1 + 2\rho_1(x))$$

where $\rho_1(x) = \lambda_1 \mathbb{E}[\min(B_1, x)] = \lambda_1 \mathbb{E}[B_\phi]$. One can easily check that this result also agrees with the first moment calculation found in [4], where they essentially look at the special case of a polling system with zero setup times, and only one buffer.

At first glance, the LST of $T_{PS}(x)$ doesn't look like a nice function to differentiate, but it's still not too difficult to make use of it in order to compute the first and second moment. By applying Lemma 3.1 to (24), we find that

$$\mathbb{E}(e^{-\omega D_{PS}(x)}) = \sum_{k=0}^{n} (-(1-\phi(\omega)))^k \lambda_1^k \frac{\mathbb{E}(R_C^k)}{k!} \mathbb{E}(e^{-(\omega+\lambda_1(1-\phi(\omega)))R_{C,k+1}}) + \mathcal{O}(\omega^{n+1}) \qquad \omega \downarrow 0.$$

Furthermore, since

$$1 - \phi(\omega) = \mathbb{E}(B_\phi)\omega - \frac{\mathbb{E}(B_\phi^2)}{2}\omega^2 + \mathcal{O}(\omega^3), \qquad \omega \downarrow 0$$

and

$$\omega + \lambda_1(1-\phi(\omega)) = (1+\rho_1(x))\omega - \lambda_1 \frac{\mathbb{E}(B_\phi^2)}{2}\omega^2 + \mathcal{O}(\omega^3), \qquad \omega \downarrow 0,$$

we have for $\omega \downarrow 0$:

$$\mathbb{E}(e^{-\omega D_{PS}(x)}) = 1 - \mathbb{E}(R_C)(1+2\rho_1(x))\omega + \left[\lambda_1 \mathbb{E}(B_\phi^2)\mathbb{E}(R_C) + \frac{\mathbb{E}(R_C^2)}{2}(1 + 3\rho_1(x) + 3\rho_1^2(x))\right]\omega^2 + \mathcal{O}(\omega^3).$$

This expression shows that the first and second moments of the conditional delay are just

$$\mathbb{E}(D_{PS}(x)) = \mathbb{E}(R_C)(1 + 2\rho_1(x))$$

and

$$\mathbb{E}(D_{PS}^2(x)) = 2\lambda_1 \mathbb{E}(B_\phi^2)\mathbb{E}(R_C) + \mathbb{E}(R_C^2)\left(1 + 3\rho_1(x) + 3\rho_1^2(x)\right).$$

After unconditioning, we find that

$$\mathbb{E}(D_{PS}) = \mathbb{E}(R_C)(1 + 2\lambda_1 \mathbb{E}(M_{1,2}))$$

and

$$\begin{aligned}
\mathbb{E}(D_{PS}^2) &= 2\lambda_1 \mathbb{E}(M_{1,2}^2)\mathbb{E}(R_C) + \mathbb{E}(R_C^2)\left[1 + 3\int_0^\infty \rho_1(x)\mathrm{d}B_1(x) + 3\int_0^\infty \rho_1^2(x)\mathrm{d}B_1(x)\right] \\
&= 2\lambda_1 \mathbb{E}(M_{1,2}^2)\mathbb{E}(R_C) + \mathbb{E}(R_C^2)\left[1 + 3\lambda_1 \mathbb{E}(M_{1,2}) + 3P(B_1 > \max(R_{B_1}(1), R_{B_1}(2))\rho_1^2\right] \\
&= 2\lambda_1 \mathbb{E}(M_{1,2}^2)\mathbb{E}(R_C) + \mathbb{E}(R_C^2)\left[1 + 3P(B_1 > R_{B_1}(1))\rho_1 + 3P(B_1 > \max(R_{B_1}(1), R_{B_1}(2))\rho_1^2\right]
\end{aligned}$$

where $R_{B_1}(1)$ and $R_{B_1}(2)$ are independent residual versions of B_1, which are also independent of B_1. Furthermore, $M_{1,2} = \min(B_{1,1}, B_{1,2})$ and $M_{1,3} = \min(B_{1,1}, B_{1,2}, B_{1,3})$. To see the second equality, notice that

$$\begin{aligned}
\mathbb{E}(\min(B_1, x)) &= \int_0^\infty P(\min(B_1, x) > u)\mathrm{d}u \\
&= \int_0^x \overline{B}_1(y)\mathrm{d}B_1(y) \\
&=: \mathbb{E}(B_1)P(R_{B_1}(1) \le x)
\end{aligned}$$

and so

$$\begin{aligned}\int_0^\infty \rho_1^2(x)\mathrm{d}B_1(x) &= \lambda_1^2\int_0^\infty \mathbb{E}(\min(B_1,x))^2\mathrm{d}B_1(x)\\ &= \rho_1^2\int_0^\infty P(R_{B_1}(1)\le x)^2\mathrm{d}B_1(x)\\ &= \rho_1^2 P(B_1>\max(R_{B_1}(1),R_{B_2}(2))).\end{aligned}$$

From here, we can easily compute the first and second moments of the sojourn time:

$$\mathbb{E}(T_{PS}) = \mathbb{E}(B_1)+\mathbb{E}(R_C)(1+2P(B_1>R_{B_1}(1))\rho_1)$$

and

$$\begin{aligned}\mathbb{E}(T_{PS}^2) = 2\lambda_1\mathbb{E}(M_{1,2}^2)\mathbb{E}(R_C) &+ \mathbb{E}(R_C^2)\left[1+3P(B_1>R_{B_1}(1))\rho_1+3P(B_1>\max(R_{B_1}(1),R_{B_1}(2))\rho_1^2\right]\\ &+ 2\mathbb{E}(B_1)\mathbb{E}(R_C)(1+2P(B_1>R_{B_1}(1))\rho_1)+\mathbb{E}(B_1^2).\end{aligned}$$

At this point, our results show that serving customers according to a Processor-Sharing rule may or may not be more efficient, in terms of mean and variance, than serving according to a FCFS rule. The type of policy chosen should depend on the parameters of the polling system.

It may be of interest to find all values x where $\mathbb{E}(T_{PS}(x))\le\mathbb{E}(T_{FCFS}(x))$, and where $\mathbb{E}(T_{PS}(x))\ge\mathbb{E}(T_{FCFS}(x))$. If we assume that the distribution of B_1 is absolutely continuous (i.e. has a density), an application of the dominated convergence theorem shows that the set of points where $\mathbb{E}(T_{PS}(x))\ge\mathbb{E}(T_{FCFS}(x))$ is of the form $[x_{PS},\infty)$, where x_{PS} is the solution to the equation

$$\mathbb{E}(\min(B_1,x)) = \mathbb{E}(B_1)/2.$$

After some simple manipulations, we see that x_{PS} satisfies

$$\int_0^{x_{PS}}\overline{B_1}(t)\mathrm{dt} = \mathbb{E}(\mathrm{B}_1)/2,$$

with $\overline{B_1}(t)=P(B_1>t)$. This implies that x_{PS} is the median of the residual service time distribution. Notice that if B_1 is exponential, then this is just the median of an exponential distribution, and so we can conclude that in this case, half of all customers that arrive to the system will experience a shorter expected sojourn time if the system operates under FCFS, and the other half will experience a shorter expected sojourn time under PS.

4.2.3 Shortest Job First

Now we will present the LST for the sojourn time of a tagged customer that visits Q_1 under the Shortest Job First policy. Due to the fact that $g_1=g_2$ under this policy as well,

$$\mathbb{E}(e^{-\omega T_{SJF}(x)}) = e^{-\omega x}\frac{\mathbb{E}(e^{-\lambda_1(1-\phi(\omega))C})-\mathbb{E}(e^{-(\omega+\lambda_1(1-\phi(\omega)))C})}{\omega\mathbb{E}(C)},$$

but in this case $\phi(\omega)=\mathbb{E}(e^{-\omega B_1\mathbf{1}_{(B_1\le x)}})$. At this point, we can manipulate the transform for this sojourn time in precisely the same manner as was done for the processor-sharing case given above, because we never made explicit use of the form of ϕ. Therefore, the first and second moment of $D_{SJF}(x)$ are as follows:

$$\mathbb{E}(D_{SJF}(x)) = (1+2\rho_1(x))\mathbb{E}(R_C)$$

and

$$\mathbb{E}(D_{SJF}(x)^2) = 2\lambda_1\mathbb{E}(B_\phi^2)\mathbb{E}(R_C)+\mathbb{E}(R_C^2)\left(1+3\rho_1(x)+3\rho_1^2(x)\right),$$

however in this case $\rho_1(x) = \lambda_1 \mathbb{E}(B_1 \mathbf{1}(B_1 \leq x))$. The unconditional moments can also be computed, as in the PS case. First of all,

$$\begin{aligned}\int_0^\infty \mathbb{E}(B_1^2 \mathbf{1}(B_1 \leq x)) \mathrm{d}B_1(x) &= \int_0^\infty \int_0^x t^2 \mathrm{d}B_1(t) \mathrm{d}B_1(x) = \int_0^\infty \int_t^\infty t^2 \mathrm{d}B_1(x) \mathrm{d}B_1(t) \\ &= \frac{1}{2} \int_0^\infty t^2 2\overline{B}_1(t) \mathrm{d}B_1(t) = \frac{\mathbb{E}(M_{1,1}^2)}{2}. \end{aligned} \tag{25}$$

There is only one other integral that needs to be computed, and so

$$\int_0^\infty \mathbb{E}(B_1 \mathbf{1}(B_1 \leq x))^2 \mathrm{d}B_1(x) = \int_0^\infty \left[\mathbb{E}(B_1) P(R_{B_1}(1) \leq x) - x\overline{B}_1(x)\right]^2 \mathrm{d}B_1(x). \tag{26}$$

We already know that

$$\int_0^\infty P(R_{B_1}(1) \leq x)^2 \mathrm{d}B_1(x) = P(B_1 > \max(R_{B_1}(1), R_{B_1}(2))) \tag{27}$$

and

$$\int_0^\infty x^2 \overline{B}_1^2(x) \mathrm{d}B_1(x) = \frac{\mathbb{E}(M_{1,3}^2)}{3}, \tag{28}$$

so

$$\begin{aligned}\int_0^\infty x P(R_{B_1}(1) \leq x) \overline{B}_1(x) \mathrm{d}B_1(x) &= \mathbb{E}(B_1) \int_0^\infty P(R_{B_1}(1) \leq x) \overline{B}_1(x) \mathrm{d}B_1^*(x) &(29)\\ &= \mathbb{E}(B_1) P(R_{B_1}(1) < C_{B_1}(1) < B_1) &(30)\end{aligned}$$

where $C_{B_1}(1)$ is a biased service time, independent of $R_{B_1}(1)$ and B_1. Therefore, inserting (27), (28) and (29) into (26) gives

$$\begin{aligned}\int_0^\infty \rho_1^2(x) \mathrm{d}B_1(x) &= \mathbb{E}(B_1)^2 P(B_1 > \max(R_{B_1}(1), R_{B_1}(2))) \\ &- 2\mathbb{E}(B_1)^2 P(R_{B_1}(1) < C_{B_1}(1) < B_1) + \frac{\mathbb{E}(M_{1,3}^2)}{3}.\end{aligned}$$

Using this along with (25) gives

$$\begin{aligned}\mathbb{E}(D_{SJF}^2) &= \lambda_1 \mathbb{E}(R_C) \mathbb{E}(M_{1,2}^2) + \mathbb{E}(R_C^2)(1 + \lambda_1 \frac{3}{2} \mathbb{E}(M_{1,2})) \\ &+ 3\mathbb{E}(R_C^2) \left(\rho_1^2 P(B_1 > \max(R_{B_1}(1), R_{B_1}(2))) - 2\rho_1^2 P(R_{B_1}(1) < C_{B_1}(1) < B_1) + \lambda_1^2 \frac{\mathbb{E}(M_{1,3}^2)}{3} \right).\end{aligned}$$

5 A globally gated polling regime

In this section, we compute the LST of the sojourn time $T_{\Gamma,i}$ of an arbitrary type-i customer in a globally gated polling system that serves customers at Q_i according to policy Γ. In such a polling system, the server serves only the customers who are present at the start of the cycle, i.e. a gate is placed behind every queue just before the server polls the first queue. This polling regime is not of a branching type visit discipline, but it satisfies Property 2.2 which allows us to decompose $T_{\Gamma,i}$ into the sum of four parts which only depend on the total and the residual length (C^* and C^r) of the cycle in which a tagged customer arrives. If the number of this cycle is n, these four parts are defined by:

1. C^r, the residual cycle length of cycle n,

2. the service times of all customers of type $j = 1, \ldots, i-1$ that arrive during C^p and C^r of cycle n,
3. R_i, the time interval between the polling epoch of Q_i in cycle $n+1$ and the departure of the tagged customer,
4. the switch-over times $S_1, \ldots, S_{i-1}$.

The LST of the total cycle time is derived in [8] and satisfies

$$\gamma(\omega) = \mathbb{E}(e^{-\omega C}) = \prod_{i=1}^{\infty} \sigma(\delta^{(i)}(\omega)),$$

with

$$\begin{aligned} \sigma(\omega) &= \mathbb{E}(e^{-\omega \sum_{i=1}^{N} S_i}) = \prod_{j=1}^{N} \sigma_j(\omega), \\ \delta^{(0)}(\omega) &= \omega, \\ \delta^{(i)}(\omega) &= \delta(\delta^{(i-1)}(\omega)), \\ \delta(\omega) &= \sum_{j=1}^{N} \lambda_j (1 - \beta_j(\omega)). \end{aligned}$$

In the same paper, the LST of the waiting time in Q_i with a FCFS service discipline is derived. This result will be discussed in the following section.

5.1 Nonanticipating policies

5.1.1 First-Come-First-Served (FCFS)

In [8], the LST of the waiting time in Q_i of a globally gated system with a FCFS service discipline is given:

$$\mathbb{E}(e^{-\omega T_{FCFS,i}}) = \frac{1}{\mathbb{E}C} \frac{\gamma\left(\sum_{j=1}^{i} \lambda_j (1 - \beta_j(\omega))\right) - \gamma\left(\sum_{j=1}^{i-1} \lambda_j (1 - \beta_j(\omega)) + \omega\right)}{\omega - \lambda_i (1 - \beta_i(\omega))} \prod_{j=1}^{i-1} \sigma_j(\omega).$$

The first and second moment of T_i in FCFS can be derived with Taylor series approximations in the numerator and the denominator. We find

$$\mathbb{E}(T_{FCFS,i}) = \mathbb{E}(B_i) + \mathbb{E}(R_C) \left(2 \sum_{j=1}^{i-1} \rho_j + \rho_i + 1 \right) + \sum_{j=1}^{i-1} \mathbb{E}(S_j) \tag{31}$$

and

$$\begin{aligned} \mathbb{E}(T_{FCFS,i}^2) &= \mathbb{E}(R_C) \left[\lambda_i \mathbb{E}(B_i^2) + 2(\rho_i + 1)\mathbb{E}(B_i) + 2 \sum_{j=1}^{i-1} \lambda_j \mathbb{E}(B_j^2) + 4 \sum_{j=1}^{i-1} \rho_j \mathbb{E}(B_i) \right] \\ &+ \mathbb{E}(R_C^2) \left(3 \left(\sum_{j=1}^{i-1} \rho_j \right)^2 + \rho_i(\rho_i + 1) + 1 + 3 \sum_{j=1}^{i-1} \rho_j(\rho_i + 1) \right) \\ &+ \mathbb{E}(R_C) \left(4 \sum_{j=1}^{i-1} \rho_j + 2\rho_i + 2 \right) \sum_{j=1}^{i-1} \mathbb{E}(S_j) \\ &+ \mathbb{E}(B_i^2) + \mathbb{E}\left(\left(\sum_{j=1}^{i-1} S_j \right)^2 \right) + 2\mathbb{E}(B_i) \sum_{j=1}^{i-1} \mathbb{E}(S_j), \end{aligned} \tag{32}$$

where R_C is equal in distribution to C^r, as pointed out in Section 3.1.

5.1.2 Last-Come-First-Served (LCFS)

In the LCFS policy, R_i consists only of the service times of the customers who arrive during the residual cycle and the service time of the tagged customer. So we get

$$\begin{aligned}\mathbb{E}(e^{-\omega(T_{LCFS,i}-\sum_{j=1}^{i-1}S_j)}) &= \int_{t=0}^{\infty}\int_{u=0}^{\infty}\sum_{k_i=0}^{\infty}e^{-\lambda_i u}\frac{(\lambda_i u)^{k_i}}{k_i!}e^{-\omega u} \\ &\times \prod_{j=1}^{i-1}e^{-\lambda_j(1-\beta_j(\omega))(t+u)}\mathbb{E}(e^{-\omega R_i}|k_i \text{ arrivals in } C^r)\mathrm{d}P(C^p\le t;C^r\le u).\end{aligned}$$

Clearly, $\mathbb{E}(e^{-\omega R_i}|k_i \text{ arrivals in } C^r)=\beta_i^{k_i+1}(\omega)$, the LST of the sum of k_i+1 service times. So

$$\begin{aligned}\mathbb{E}(e^{-\omega(T_{LCFS,i}-\sum_{j=1}^{i-1}S_j)}) &= \beta_i(\omega)\int_{t=0}^{\infty}\int_{u=0}^{\infty}\sum_{k_i=0}^{\infty}e^{-\lambda_i u}\frac{(\lambda_i u\beta_i(\omega))^{k_i}}{k_i!}e^{-\omega u} \\ &\times e^{-\sum_{j=1}^{i-1}\lambda_j(1-\beta_j(\omega))(t+u)}\mathrm{d}P(C^p\le t;C^r\le u) \\ &= \beta_i(\omega)\int_{t=0}^{\infty}\int_{u=0}^{\infty}e^{-\lambda_i(1-\beta_i(\omega))u}e^{-\omega u} \\ &\times e^{-\sum_{j=1}^{i-1}\lambda_j(1-\beta_j(\omega))(t+u)}\mathrm{d}P(C^p\le t;C^r\le u).\end{aligned}$$

Using (12), we get:

$$\mathbb{E}(e^{-\omega(T_{LCFS,i}-\sum_{j=1}^{i-1}S_j)}) = \beta_i(\omega)\frac{\gamma\left(X_i(\omega)\right)-\gamma\left(X_{i+1}(\omega)+\omega\right)}{(\lambda_i(1-\beta_i(\omega))+\omega)\mathbb{E}C}.$$

The first moment of $T_{LCFS,i}$ is exactly the same as in (31):

$$\mathbb{E}(T_{LCFS,i}) = \mathbb{E}(B_i)+\mathbb{E}(R_C)\left(2\sum_{j=1}^{i-1}\rho_j+\rho_i+1\right)+\sum_{j=1}^{i-1}\mathbb{E}(S_j).$$

The second moment, however, is larger than $\mathbb{E}\left(T_{FCFS,i}^2\right)$:

$$\begin{aligned}\mathbb{E}\left(T_{LCFS,i}^2\right) &= \mathbb{E}(R_C)\left[\lambda_i\mathbb{E}(B_i^2)+2(\rho_i+1)\mathbb{E}(B_i)+2\sum_{j=1}^{i-1}\lambda_j\mathbb{E}(B_j^2)+4\sum_{j=1}^{i-1}\rho_j\mathbb{E}(B_i)\right] \\ &+ \mathbb{E}(R_C^2)\left[3\left(\sum_{j=1}^{i-1}\rho_j\right)^2+(\rho_i+1)^2+3\sum_{j=1}^{i-1}\rho_j(\rho_i+1)\right] \\ &+ \mathbb{E}(R_C)\left(4\sum_{j=1}^{i-1}\rho_j+2\rho_i+2\right)\sum_{j=1}^{i-1}\mathbb{E}(S_j) \\ &+ \mathbb{E}(B_i^2)+\mathbb{E}\left(\left(\sum_{j=1}^{i-1}S_j\right)^2\right)+2\mathbb{E}(B_i)\sum_{j=1}^{i-1}\mathbb{E}(S_j). \qquad (33)\end{aligned}$$

This should not come as a surprise, based on what we have previously seen in the gated section.

5.1.3 Random order of Service (ROS)

For generally distributed service times and a ROS discipline, we derive the LST of the sojourn time of a random customer. The time between the polling epoch of Q_i and the departure of a tagged type-i customer (R_i) depends on the total number of type-i customers that arrived in cycle n, say k_i. The order of service is random, so R_i is the sum of l_i service times, with l_i randomly chosen from $\{1, \ldots, k_i\}$.

Because the switch-over times are independent of C^r, R_i and the service times of all other customers that arrive during C^*, we can focus on just these three parts of the sojourn time of a tagged customer, $T_{ROS,i} - \sum_{j=1}^{i-1} S_j$. Because each of these parts only depends on C^* and/or C^r, we condition on the residual cycle length C^r and the preceding cycle length C^p ($C^* = C^p + C^r$):

$$\begin{aligned}\mathbb{E}(e^{-\omega(T_{ROS,i}-\sum_{j=1}^{i-1} S_j)}) &= \int_{t=0}^{\infty}\int_{u=0}^{\infty}\sum_{k_i=0}^{\infty} e^{-\lambda_i(t+u)}\frac{(\lambda_i(t+u))^{k_i}}{k_i!}e^{-\omega u} \\ &\times \prod_{j=1}^{i-1} e^{-\lambda_j(1-\beta_j(\omega))(t+u)}\mathbb{E}(e^{-\omega R_i}|k_i \text{ others})\mathrm{d}P(C^p \le t; C^r \le u).\end{aligned}$$

Using the result in (12), we get

$$\mathbb{E}(e^{-\omega(T_{ROS,i}-\sum_{j=1}^{i-1} S_j)}) = \frac{\beta_i(\omega)}{\mathbb{E}(C)(1-\beta_i(\omega))}\frac{1}{\lambda_i}\int_{X_i(\omega)}^{X_{i+1}(\omega)}\frac{\gamma(y)-\gamma(y+\omega)}{\omega}\mathrm{d}y, \tag{34}$$

with

$$X_i(\omega) = \sum_{j=1}^{i-1}\lambda_j(1-\beta_j(\omega)). \tag{35}$$

For the first and second moment of $T_{ROS,i}$, we differentiate (34) by using a Taylor series development in ω and find:

$$\mathbb{E}(T_{ROS,i}) = \mathbb{E}(B_i) + \frac{\mathbb{E}(C^2)}{2\mathbb{E}(C)}\left(2\sum_{j=1}^{i-1}\rho_j + \rho_i + 1\right) + \sum_{j=1}^{i-1}\mathbb{E}(S_j). \tag{36}$$

Indeed, the mean sojourn time consists of the mean service time of the tagged customer, the mean residual cycle time, the mean work arriving at $Q_1, \ldots, Q_{i-1}$ during the past and residual cycle time $\left(2 \times \frac{\mathbb{E}(C^2)}{2\mathbb{E}(C)}\right)$, half of the average work arriving at Q_i during the past and residual cycle time and the mean switch over times $\mathbb{E}(S_1), \ldots, \mathbb{E}(S_{i-1})$. Furthermore, the first moment is again exactly the same as in (31).

For the second moment, we find

$$\begin{aligned}\mathbb{E}(T_{ROS,i}^2) &= \mathbb{E}(R_C)\left(\lambda_i\mathbb{E}(B_i^2) + 2(\rho_i+1)\mathbb{E}(B_i) + 2\sum_{j=1}^{i-1}\lambda_j\mathbb{E}(B_j^2) + 4\sum_{j=1}^{i-1}\rho_j\mathbb{E}(B_i)\right) \\ &+ \mathbb{E}(R_C^2)\left[3\left(\sum_{j=1}^{i-1}\rho_j\right)^2 + \rho_i^2 + \frac{3}{2}\rho_i + 1 + 3\sum_{j=1}^{i-1}\rho_j(\rho_i+1)\right] \\ &+ \mathbb{E}(R_C)\left(4\sum_{j=1}^{i-1}\rho_j + 2\rho_i + 2\right)\sum_{j=1}^{i-1}\mathbb{E}(S_j) \\ &+ \mathbb{E}(B_i^2) + \mathbb{E}\left(\left(\sum_{j=1}^{i-1}S_j\right)^2\right) + 2\mathbb{E}(B_i)\sum_{j=1}^{i-1}\mathbb{E}(S_j).\end{aligned} \tag{37}$$

Note that the mean sojourn time of a type-i customer can be larger than the mean sojourn time of a type-$(i+1)$ customer, because $\mathbb{E}(T_{ROS,i+1}) - \mathbb{E}(T_{ROS,i}) \leq 0$ if

$$\mathbb{E}(B_i) \geq \mathbb{E}(B_{i+1}) + \frac{\mathbb{E}(C^2)}{2\mathbb{E}(C)} \left[\lambda_{i+1}\mathbb{E}(B_{i+1}) + \lambda_i \mathbb{E}(B_i)\right] + \mathbb{E}(S_i).$$

Furthermore, notice that, as is true in the gated case and in [4], $\mathbb{E}(T^2_{FCFS,i}) < \mathbb{E}(T^2_{ROS,i}) < \mathbb{E}(T^2_{LCFS,i})$. The differences are as follows:

$$\mathbb{E}(T^2_{LCFS,i}) - \mathbb{E}(T^2_{ROS,i}) = \mathbb{E}(T^2_{ROS,i}) - \mathbb{E}(T^2_{FCFS,i}) = \frac{\mathbb{E}(R_C^2)\rho_i}{2}.$$

5.2 Anticipating policies

5.2.1 Processor sharing (PS)

The derivation of the LST of the waiting time in the case of the PS service discipline is different from the one in ROS, because the waiting time now heavily depends on the required service time of the tagged customer. However, for exponentially(μ) distributed service times, the analysis is the same as for ROS, because of Proposition 4.1.

Now suppose that the service times are generally distributed. To work with this, it will again be to our advantage to condition on the amount of service brought to Q_i by a tagged customer during steady-state. If such a customer brings an amount of work x to Q_i, then its sojourn time minus x is just

$$D_{PS,i}(x) = C^r + \sum_{j=1}^{i-1}(V_j + S_j) + \sum_{m=1}^{N_i(C^p)} \min(B_{i,m}, x) + \sum_{n=1}^{N_i(C^r)} \min(B_{i,n}, x).$$

Again, because the switch-over times are independent of all other quantities present in our representation of $D_{PS,i}(x)$, we will focus on computing the LST of $D_{PS,i}(x) - \sum_{j=1}^{i-1} S_j$. If we let $\phi_i(\omega, x)$ denote the LST of $\min(B_i, x)$, then

$$\begin{aligned}
\mathbb{E}(e^{-\omega(D_{PS,i}(x) - \sum_{j=1}^{i-1} S_j)}) &= \int_{t=0}^{\infty}\int_{u=0}^{\infty} e^{-\omega u} e^{-\lambda_i(1-\phi_i(\omega,x))(t+u)} \prod_{j=1}^{i-1} e^{-\lambda_j(1-\beta_j(\omega))(t+u)} \mathrm{d}P(C^p \leq t, C^r \leq u) \\
&= \frac{\gamma(X_i(\omega) + \lambda_i(1-\phi_i(\omega,x))) - \gamma(X_i(\omega) + \lambda_i(1-\phi_i(\omega,x)) + \omega)}{\omega \mathbb{E}[C]},
\end{aligned}$$

where the second equality follows from (12).

By applying Lemma 3.1, it follows that

$$\mathbb{E}(D_{PS,i}(x) - \sum_{j=1}^{i-1} S_j) = \mathbb{E}(R_C)\left[1 + 2\sum_{j=1}^{i-1}\rho_j + 2\lambda_i \mathbb{E}(\min(B_i, x))\right] \tag{38}$$

and

$$\begin{aligned}
\mathbb{E}((D_{PS,i}(x) - \sum_{j=1}^{i-1} S_j)^2) &= \mathbb{E}(R_C)\left[2\sum_{j=1}^{i-1}\lambda_j \mathbb{E}(B_j^2) + 2\lambda_i \mathbb{E}(\min(B_i, x)^2)\right] \\
&+ \mathbb{E}(R_C^2)\left[1 + 3\sum_{j=1}^{i-1}\rho_j + 3\lambda_i \mathbb{E}(\min(B_i, x))\right. \\
&+ \left. 3\left(\sum_{j=1}^{i-1}\rho_j + \lambda_i \mathbb{E}(\min(B_i, x))\right)^2\right].
\end{aligned} \tag{39}$$

After unconditioning with respect to x, we conclude that

$$\mathbb{E}(D_{PS,i} - \sum_{j=1}^{i-1} S_j) = \mathbb{E}(R_C) \left[1 + 2 \sum_{j=1}^{i-1} \rho_j + 2\lambda_i \mathbb{E}(M_{i,2}) \right] \tag{40}$$

and

$$\begin{aligned} \mathbb{E}((D_{PS,i} - \sum_{j=1}^{i-1} S_j)^2) &= \mathbb{E}(R_C) \left[2 \sum_{j=1}^{i-1} \lambda_j \mathbb{E}(B_j^2) + 2\lambda_i \mathbb{E}(M_{i,2}^2) \right] \\ &+ \mathbb{E}(R_C^2) \left[1 + 3 \sum_{j=1}^{i-1} \rho_j + 3\rho_i P(B_i > R_{B_i}(1)) + 3 \left(\sum_{j=1}^{i-1} \rho_j \right)^2 \right. \\ &+ 6 \sum_{j=1}^{i-1} \rho_j \rho_i P(B_i > R_{B_i}(1)) \\ &+ \left. 3\rho_i^2 P(B_i > \max(R_{B_i}(1), R_{B_i}(2))) \right]. \end{aligned} \tag{41}$$

Finally, after combining the switch-over times and the service time of the tagged customer with (40) and (41), we get

$$\mathbb{E}(T_{PS,i}) = \mathbb{E}(B_i) + \sum_{j=1}^{i-1} \mathbb{E}(S_j) + \mathbb{E}(R_C) \left[1 + 2 \sum_{j=1}^{i-1} \rho_j + 2\lambda_i \mathbb{E}(M_{i,2}) \right] \tag{42}$$

and

$$\begin{aligned} \mathbb{E}(T_{PS,i}^2) &= \mathbb{E}(R_C) \left[2 \sum_{j=1}^{i-1} \lambda_j \mathbb{E}(B_j^2) + 2\lambda_i \mathbb{E}(M_{i,2}^2) \right] \\ &+ \mathbb{E}(R_C^2) \left[1 + 3 \sum_{j=1}^{i-1} \rho_j + 3\rho_i P(B_i > R_{B_i}(1)) + 3 \left(\sum_{j=1}^{i-1} \rho_j \right)^2 \right. \\ &+ \left. 6 \sum_{j=1}^{i-1} \rho_j \rho_i P(B_i > R_{B_i}(1)) + 3\rho_i^2 P(B_i > \max(R_{B_i}(1), R_{B_i}(2))) \right] \\ &+ 2 \left(\mathbb{E}(B_i) + \sum_{j=1}^{i-1} \mathbb{E}(S_j) \right) \mathbb{E}(R_C) \left[1 + 2 \sum_{j=1}^{i-1} \rho_j + 2\lambda_i \mathbb{E}(M_{i,2}) \right] \\ &+ \mathbb{E}(B_i^2) + 2\mathbb{E}(B_i) \sum_{j=1}^{i-1} \mathbb{E}(S_j) + \mathbb{E}((\sum_{j=1}^{i-1} S_j)^2). \end{aligned} \tag{43}$$

5.2.2 Shortest Job First

Now we will compute the first and second moments of the sojourn time under the SJF policy. In this case it is clear that, conditional on the service time of the tagged customer being x,

$$D_{SJF,i}(x) = C^r + \sum_{j=1}^{i-1} (V_j + S_j) + \sum_{m=1}^{N_i(C^p)} B_{i,m} \mathbf{1}(B_{i,m} \leq x) + \sum_{n=1}^{N_i(C^r)} B_{i,n} \mathbf{1}(B_{i,n} \leq x).$$

If we mimic the above derivation of the LST of the conditional delay for the PS case, we see that

$$\mathbb{E}(e^{-\omega(D_{SJF,i}(x) - \sum_{j-1}^{i-1} S_j)}) = \frac{\gamma(X_i(\omega) + \lambda_i(1 - \phi_i(\omega, x))) - \gamma(X_i(\omega) + \lambda_i(1 - \phi_i(\omega, x)) + \omega)}{\omega \mathbb{E}(C)}$$

where in this case $\phi_i(\omega, x)$ is the LST of $B_i\mathbf{1}(B_i \leq x)$.

Just as before, we get

$$\mathbb{E}(D_{SJF,i}(x) - \sum_{j=1}^{i-1} S_j) = \mathbb{E}(R_C)\left[1 + 2\sum_{j=1}^{i-1}\rho_j + 2\lambda_i\mathbb{E}(B_i\mathbf{1}(B_i \leq x))\right] \tag{44}$$

and

$$\begin{aligned}\mathbb{E}((D_{SJF,i}(x) - \sum_{j=1}^{i-1} S_j)^2) &= \mathbb{E}(R_C)\left[2\sum_{j=1}^{i-1}\lambda_j\mathbb{E}(B_j^2) + 2\lambda_i\mathbb{E}(B_i^2\mathbf{1}(B_i \leq x))\right] \\ &+ \mathbb{E}(R_C^2)\left[1 + 3\sum_{j=1}^{i-1}\rho_j + 3\lambda_i\mathbb{E}(B_i\mathbf{1}(B_i \leq x))\right. \\ &+ \left. 3\left(\sum_{j=1}^{i-1}\rho_j + \lambda_i\mathbb{E}(B_i\mathbf{1}(B_i \leq x))\right)^2\right]. \end{aligned} \tag{45}$$

After unconditioning with respect to x, we conclude that

$$\mathbb{E}(D_{SJF,i} - \sum_{j=1}^{i-1} S_j) = \mathbb{E}(R_C)\left[1 + 2\sum_{j=1}^{i-1}\rho_j + \lambda_i\mathbb{E}(M_{i,2})\right] \tag{46}$$

and

$$\begin{aligned}\mathbb{E}((D_{SJF,i} - \sum_{j=1}^{i-1} S_j)^2) &= \mathbb{E}(R_C)\left[2\sum_{j=1}^{i-1}\lambda_j\mathbb{E}(B_j^2) + \lambda_i\mathbb{E}(M_{i,2}^2)\right] \\ &+ \mathbb{E}(R_C^2)\left[1 + 3\sum_{j=1}^{i-1}\rho_j + \frac{3}{2}\rho_i P(B_i > R_{B_i}(1)) + 3\left(\sum_{j=1}^{i-1}\rho_j\right)^2\right. \\ &+ 3\left[\mathbb{E}(B_1)^2 P(B_1 > \max(R_{B_1}(1), R_{B_1}(2))) - 2\mathbb{E}(B_1)^2 P(R_{B_1}(1) < C_{B_1}(1) < B_1)\right] \\ &+ \left. 3\sum_{j=1}^{i-1}\rho_j\rho_i P(B_i > R_{B_i}(1)) + \mathbb{E}(M_{1,3}^2)\right], \end{aligned} \tag{47}$$

with $M_{i,2}$ and $M_{i,3}$ as defined in Section 4.2.2. Therefore, the first and second moments of the sojourn time are as follows:

$$\mathbb{E}(T_{SJF,i}) = \mathbb{E}(B_i) + \sum_{j=1}^{i-1}\mathbb{E}(S_j) + \mathbb{E}(R_C)\left[1 + 2\sum_{j=1}^{i-1}\rho_j + \lambda_i\mathbb{E}(M_{i,2})\right] \tag{48}$$

and

$$\begin{aligned}\mathbb{E}(T_{SJF,i}^2) &= \mathbb{E}(B_i^2) + 2\mathbb{E}(B_i)\sum_{j=1}^{i-1}E(S_j) + \mathbb{E}\left(\left(\sum_{j=1}^{i-1} S_j\right)^2\right) \\ &+ 2\left(\mathbb{E}(B_i)\sum_{j=1}^{i-1}\mathbb{E}(S_j)\right)\left(\mathbb{E}(B_i) + \sum_{j=1}^{i-1}\mathbb{E}(S_j) + \mathbb{E}(R_C)\left[1 + 2\sum_{j=1}^{i-1}\rho_j + \lambda_i\mathbb{E}(M_{i,2})\right]\right) \end{aligned}$$

$$
\begin{aligned}
&+ \quad \mathbb{E}(R_C)\left[2\sum_{j=1}^{i-1}\lambda_j\mathbb{E}(B_j^2)+\lambda_i\mathbb{E}(M_{i,2}^2)\right] \\
&+ \quad \mathbb{E}(R_C^2)\left[1+3\sum_{j=1}^{i-1}\rho_j+\frac{3}{2}\rho_i P(B_i>R_{B_i}(1))+3\left(\sum_{j=1}^{i-1}\rho_j\right)^2\right. \\
&+ \quad 3\left[\mathbb{E}(B_1)^2P(B_1>\max(R_{B_1}(1),R_{B_1}(2)))-2\mathbb{E}(B_1)^2P(R_{B_1}(1)<C_{B_1}(1)<B_1)\right] \\
&+ \quad \left.3\sum_{j=1}^{i-1}\rho_j\rho_i P(B_i>R_{B_i}(1))+\mathbb{E}(M_{1,3}^2)\right]. \qquad (49)
\end{aligned}
$$

6 Conclusion

We have obtained the (LST of the) waiting time distribution in a gated queue of a cyclic polling system, for various service orders within that queue. The first two moments of the waiting time also have been obtained, allowing us to study the impact of the service order.

The gated visit discipline turns out to be very tractable, thanks to the fact that the waiting times of the customers who are being served during a visit are not affected by later arrivals which take place in that visit period. We expect exhaustive service to be more complicated. This is a topic for our further research. Presently the case of fixed priorities within a queue of a polling system also receives attention in our group; cf. [2]. One could also investigate whether or not there exists a sort of ordering among the distributions of the sojourn times considered here. We are also interested in understanding exactly how heavy-tailed service times may influence the tail behavior of the sojourn time, under each of the various disciplines studied in this paper.

Acknowledgments

This research was funded by the Dutch BRICKS project and was conducted within the framework of the European Network of Excellence Euro-FGI.

References

[1] J. Abate and W. Whitt (1996). An operational calculus for probability distributions via Laplace transforms. *Adv. Appl. Prob.* **28**, 75-113.

[2] M.A.A. Boon, I.J.B.F. Adan and O.J. Boxma (2008). A two-queue polling model with two priority levels in the first queue. *EURANDOM report.*

[3] K. Athreya and P. Ney (1972). *Branching Processes.* Springer, Berlin.

[4] B. Avi-Itzhak and S. Halfin (1989). Response times in gated $M/G/1$ queues: the processor-sharing case. *Queueing Systems* **4**, 263-279.

[5] F. Baccelli and P. Brémaud (2003). *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences.* Springer, New York.

[6] S. Borst (1994). Polling Systems. Ph.D. Thesis, CWI, Amsterdam, The Netherlands.

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

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

[9] W. Feller (1971). *An Introduction to Probability Theory and its Applications, Volume 2.* Wiley.

[10] K.M. Rege and B. Sengupta (1989). A single server queue with gated processor-sharing discipline. *Queueing Systems* **4**, 249-261.

[11] J.A.C. Resing (1993). Polling systems and multitype branching processes. *Queueing Systems* **13**, 413-426.

[12] R.Rietman and J.A.C. Resing (2004). An $M/G/1$ queueing model with gated random order of service. *Queueing Systems* **48**, 89-102.

[13] R. Serfozo (1999). *Introduction to Stochastic Networks.* Springer, New York

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

[15] H. Takagi (2000). Analysis and application of polling models. In: G. Haring, C. Lindemann and M. Reiser (eds.). *Performance Evaluation: Origins and Directions.* LNCS 1769 (Springer, Berlin), pp. 423-442.

[16] H. Thörisson (2000). *Coupling, Stationarity and Regeneration.* Springer, New York.

[17] V. Vishnevskii and O. Semenova (2006). Mathematical methods to study the polling systems. *Automation and Remote Control* **67**, 173-220.

[18] A. Wierman, E.M.M. Winands and O.J. Boxma (2007). Scheduling in polling systems. *Performance Evaluation* **64**, 1009-1028.

[19] E.M.M. Winands, I.J.B.F. Adan and G.-J. van Houtum (2005). The stochastic economic lot scheduling problem: a survey. BETA Report WP-133, BETA Research School for Operations Management and Logistics.

[20] E.M.M. Winands, I.J.B.F. Adan and G.-J. van Houtum (2006). Mean value analysis for polling systems. *Queueing Systems* **54**, 45-54.