Adv. Appl. Prob. **47,** 989–1014 (2015)
Printed in Northern Ireland

HEAVY-TRAFFIC LIMITS FOR POLLING MODELS WITH EXHAUSTIVE SERVICE AND NON-FCFS SERVICE ORDER POLICIES

P. VIS,* ** *VU University Amsterdam and Centre for Mathematics and Computer Science*

R. BEKKER,* *** *VU University Amsterdam*

R. D. VAN DER MEI,* ** *VU University Amsterdam and Centre for Mathematics and Computer Science*

Abstract

We study cyclic polling models with exhaustive service at each queue under a variety of non-FCFS (first-come–first-served) local service orders, namely last-come–first-served with and without preemption, random-order-of-service, processor sharing, the multi-class priority scheduling with and without preemption, shortest-job-first, and the shortest remaining processing time policy. For each of these policies, we first express the waiting-time distributions in terms of intervisit-time distributions. Next, we use these expressions to derive the asymptotic waiting-time distributions under heavy-traffic assumptions, i.e. when the system tends to saturate. The results show that in all cases the asymptotic waiting-time distribution at queue i is fully characterized and of the form $\Gamma\Theta_i$, with Γ and Θ_i independent, and where Γ is gamma distributed with known parameters (and the same for all scheduling policies). We derive the distribution of the random variable Θ_i which explicitly expresses the impact of the local service order on the asymptotic waiting-time distribution. The results provide new fundamental insight into the impact of the local scheduling policy on the performance of a general class of polling models. The asymptotic results suggest simple closed-form approximations for the complete waiting-time distributions for stable systems with arbitrary load values.

Keywords: Polling system; service discipline; waiting-time distribution; heavy traffic

2010 Mathematics Subject Classification: Primary 60K25
Secondary 90B22; 68M20

1. Introduction

Polling systems are multi-queue systems in which a single server visits the queues in some order to serve customers. Polling models find many applications in areas like computer-communication systems, production systems, manufacturing systems, inventory systems, and robotics; see [8] for an extensive overview. Motivated by their wide applicability, polling models have been extensively studied over the past few decades; see [30] for an overview of

Received 11 February 2014; revision received 8 October 2014.
* Postal address: VU Amsterdam, Department of Mathematics, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
** Postal address: Centre for Mathematics and Computer Science, Stochastics group, Science Park 123, 1098 XG Amsterdam, The Netherlands.
*** Email address: r.bekker@vu.nl

TABLE 1: A brief description of the scheduling policies discussed in this paper.

FCFS	*First-come–first-served* serves jobs in the order of arrival.
LCFS	*Last-come–first-served* serves the job that arrived most recently, without preemption.
LCFS-PR	*Last-come–first-served with preemptive resume* serves the job that arrived most recently preempting the job currently in service.
ROS	*Random-order-of-service* randomly selects a job from the jobs that are waiting.
PS	*Processor sharing* serves all jobs simultaneously at the same rate.
NPRIOR	*n-class priority regime* serves jobs within the highest priority class first, continuing with other priority classes as long as no jobs with higher priority are present. Jobs within the same priority class are served in the order of arrival.
NPRIOR-PR	*n-class priority regime with preemptive resume* serves jobs with higher priority first, preempting jobs with lower priority which are already in service, jobs within the same priority class are served FCFS.
SJF	*Shortest-job-first* nonpreemptively serves the job in the system with the smallest original service time.
SRPT	*Shortest-remaining-processing-time* preemptively serves the job with the shortest remaining processing time.

the state-of-the-art models. For operating a polling system, design choices have to be made about

1. the order in which the server visits the queues;
2. which customers are served during a visit of the server to a queue;
3. the order in which customers at the same queue are served.

The vast majority of papers in the literature are focused on the first two decisions. In this paper we address the third decision, by investigating the influence of the local service order policy on the waiting-time distributions of the customers at each of the queues. To this end, we study Poisson-driven cyclic polling systems with general service time and switchover time distributions with exhaustive service at all queues; see [29, Section 11] for a slight relaxation. We consider the following local service disciplines: last-come–first-served (LCFS) with and without preemption, random-order-of-service (ROS), local processor sharing (PS), the multi-class priority scheduling with and without preemption, shortest-job-first (SJF), and shortest remaining processing time (SRPT); see Table 1 for a brief description. In doing so, we derive new, exact expressions for the Laplace–Stieltjes transform of the waiting-time distributions. We use these expressions to find the asymptotic waiting-time distributions under heavy-traffic (HT) assumptions, i.e. when the load approaches 1.

The motivation for studying the impact of the local service order on the waiting-time performance is two-fold. First, in many real-life applications the local service order is not first-come–first-served (FCFS); examples are Bluetooth® and 802.11 protocols, scheduling policies at routers, and I/O subsystems in web servers [13], [27]. In these cases the workloads are known to have high variability and priority-based scheduling could therefore be beneficial; other examples are in the domain of production-inventory control, where local scheduling proved its worth [2]. Second, gaining fundamental understanding of the implications of the choice of the local service order on the waiting-time performance of polling systems is of queueing-theoretical interest.

There are several good reasons for studying HT asymptotics. First, it is the most important and challenging regime from a practical point of view, because the proper operation of the system is particularly critical when the system is heavily loaded. Optimizing the local service order policy is, therefore, an effective mechanism for improving system performance without purchasing additional resources. Second, an attractive feature of HT asymptotics is that in many cases they lead to strikingly simple expressions for the performance measures of interest. This remarkable simplicity of the HT asymptotics leads to structural insights into the dependence of the performance measures on the system parameters and gives fundamental understanding of the behavior of the system in general. Third, HT asymptotics form an excellent basis for developing simple, accurate approximations of the performance measures (distributions, moments, tail probabilities) for stable systems.

In the literature, many papers focus either on the analysis of polling systems or on scheduling policies for single-queue systems, but the combination of the two has received very little attention. More precisely, almost all theoretical studies of scheduling policies are performed in single-queue settings such as the M/G/1 and GI/G/1 queue with only a few exceptions studying the effect of local scheduling in multi-queue polling systems. For cyclic polling systems with gated and exhaustive service, Wierman *et al.* [31] used the mean value analysis framework [32] to derive the mean delay at each of the queues for various scheduling disciplines such as FCFS, LCFS, foreground–background, PS, SJF, and fixed priorities. Boxma *et al.* [10] obtained the waiting-time distribution in cyclic (globally-)gated polling systems for various local service orders. Bekker *et al.* [4] derived HT limits of the waiting-time distributions in cyclic polling models with gated and globally-gated service for the LCFS, ROS, PS, and SJF local service orders. In this paper we extend the results to the case of exhaustive service at each of the queues, which is fundamentally more complicated than the gated and globally-gated case (as also stated in [10]). The additional complexity of the exhaustive-service model compared to the (globally-)gated model is that customers that arrive during a visit of the server at a queue may intervene with the customers that were present at the beginning of that visit period. Nonetheless, recent progress for exhaustive models has been made. Boon *et al.* [7] studied the waiting-time distribution in a two-queue polling model with either the exhaustive, gated, or globally-gated service discipline, where the first of these two queues contained customers of two priority classes. In [6] these results were generalized to a polling model with N queues and K_i priority levels in queue i. Moreover, for the case of exponential service times at each queue, Ayesta *et al.* [1] derived the sojourn-time distribution in polling systems with exhaustive service and where the local scheduling policy is PS. For a general service requirement distribution, the analysis is restricted to the mean sojourn time.

In this paper we show that for all considered cases the asymptotic waiting-time distribution at queue i can be expressed as the product of two independent random variables Γ and Θ_i, where Γ is gamma-distributed with known parameters that are independent of the scheduling policy. Moreover, we derive the distribution of the random variable Θ_i, which expresses the impact of the local service order on the asymptotic waiting-time distribution. The results are exact and give a full characterization of the limiting behavior of the system, and as such provide new fundamental insight into the influence of the local scheduling policy on the waiting-time performance of polling models. As a by-product, the HT limits suggest simple closed-form approximations for the complete waiting-time distributions for stable systems with arbitrary load values strictly less than 1. The accuracy of the approximations is evaluated by several numerical examples, which can be found in [29].

2. Notation and model description

In this section we introduce the notation and give a description of the model. To start, we denote by $f_X(\cdot)$, $F_X(\cdot)$, $X^*(\cdot)$, and $\mathbb{E}[X]$ the probability density function (PDF), cumulative distribution function (CDF), Laplace–Stieltjes transform (LST), and expected value, respectively, of a one-dimensional absolutely-continuous random variable (RV) X.

The model is as follows. We consider a system of $N \geq 2$ infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server that visits and serves the queues in cyclic order. At each queue, the service discipline is exhaustive; that is, the server proceeds to the next queue when the queue is empty. Customers arrive at Q_i according to a Poisson process $\{N_i(t),\ t \in \mathrm{Re}\}$ with rate λ_i. These customers are referred to as type-i customers. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a RV B_i. The kth moment of the service time of an arbitrary customer is denoted by $\mathbb{E}[B^k] = \sum_{i=1}^{N} \lambda_i \mathbb{E}[B_i^k]/\Lambda$, $k = 1, 2, \ldots$. The load offered to Q_i is $\rho_i = \lambda_i \mathbb{E}[B_i]$ and the total load offered to the system can be expressed as $\rho = \sum_{i=1}^{N} \rho_i$. A necessary and sufficient condition for stability of the system is $\rho < 1$. The switchover time required by the server to proceed from Q_i to Q_{i+1} is a RV S_i. We let $S = \sum_{i=1}^{N} S_i$ denote the total switchover time in a cycle. The RV C_i describes the cycle time of the server, defined as the time between two successive departures of the server from Q_i. The mean cycle time is known to be the same for all queues, and is given by $\mathbb{E}[C_i] = \mathbb{E}[C] = \mathbb{E}[S]/(1-\rho)$. Denote by V_i the visit time at Q_i, defined as the time elapsed between a polling instant at Q_i (i.e. the moment the server arrives at the queue) and the server's successive departure from Q_i. Denote by I_i the intervisit time of Q_i, defined as the time elapsed between a departure of the server from Q_i and the successive polling instant at Q_i. Note that $C_i = I_i + V_i$ for $i = 1, \ldots, N$.

The *local service order policy* of a queue determines the order in which the customers are served during a visit period of the server at that queue. Throughout this paper we consider the local service order policies given in Table 1. We only consider work-conserving policies. For policy $P \in \{\mathrm{FCFS}, \mathrm{LCFS}, \mathrm{LCFS\text{-}PR}, \mathrm{ROS}, \mathrm{PS}, \mathrm{NPRIOR}, \mathrm{NPRIOR\text{-}PR}, \mathrm{SJF}, \mathrm{SRPT}\}$, we denote $i \in P$ if Q_i receives scheduling policy P; for example, $FCFS$ is the (index) set of queues that are served on an FCFS basis.

In this paper we mainly focus on HT limits, i.e. the limiting behavior as ρ approaches 1. The HT limits, denoted by $\rho \uparrow 1$, taken in this paper are defined such that the arrival rates are increased, while keeping both the service-time and switchover time distributions and the ratios between the arrival rates fixed. The notation '$\xrightarrow{\mathrm{D}}$' means convergence in distribution. For each variable x that is a function of ρ, we denote its value *evaluated at* $\rho = 1$ by $\hat{x}$.

Let T_i denote the sojourn time of an arbitrary customer at Q_i, defined as the time between the moment of arrival of a customer and the moment at which the customer departs from the system. The waiting time W_i of an arbitrary customer at Q_i is defined as the sojourn time minus the service requirement. When $\rho \uparrow 1$, all queues become unstable; therefore, the focus lies on the limiting distribution for $\rho \uparrow 1$ of the RVs $\tilde{W}_i := (1-\rho)W_i$ and $\tilde{T}_i := (1-\rho)T_i$, referred to as the *scaled* waiting times and sojourn times at Q_i, respectively. We denote by $\Gamma(\alpha, \mu)$ a gamma-distributed RV with shape and rate parameters α and μ, respectively. Moreover, we denote by $U[a, b]$, with $a < b$, a RV that is uniformly distributed over the interval $[a, b]$. For later reference, note that the LST of the RV $U[a, b]\Gamma(\alpha+1, \mu)$, where $U[a, b]$ and $\Gamma(\alpha+1, \mu)$ are independent, is given by

$$\mathbb{E}[\mathrm{e}^{-sU[a,b]\Gamma(\alpha+1,\mu)}] = \frac{\mu}{\alpha s(b-a)} \left\{ \left(\frac{\mu}{\mu + sa} \right)^{\alpha} - \left(\frac{\mu}{\mu + sb} \right)^{\alpha} \right\}, \qquad \mathrm{Re}(s) > 0. \quad (2.1)$$

3. Preliminaries and method outline

In this section we formulate a number of known preliminary results that serve as a reference for the remaining sections. Moreover, we provide an outline of the method that we apply here to FCFS. In Section 3.1 we give expressions for the asymptotic distributions of the cycle and intervisit times under HT assumptions. In Section 3.2 we use these results to give an expression for the LST of the waiting-time distribution for the case of FCFS service. We refer the reader to [24] for rigorous proofs of these results. In Section 3.3 we provide intuition for the main result based on a heavy traffic averaging principle. Such intuition can be useful for interpreting the waiting-time distributions for the other service disciplines as well.

3.1. Cycle and intervisit times

To start, let us consider the distribution of the cycle time C_i. A simple but important observation is that the distribution of C_i does not depend on the local scheduling policy, provided that the policy is work-conserving. This means that we can use the results for the cycle times and also for the intervisit times throughout the rest of this paper. The following result gives a characterization of the limiting behavior of the scaled cycle-time distributions.

Proposition 3.1. (Convergence of cycle times.) *Define* $\tilde{C}_i := (1-\rho)C_i$, *then for* $i = 1, \ldots, N$, *as* $\rho \uparrow 1$,

$$\tilde{C}_i \xrightarrow{\mathrm{D}} \tilde{\Gamma},$$

where $\tilde{\Gamma}$ *has a gamma distribution with parameters*

$$\alpha := \frac{\mathbb{E}[S]\delta}{\sigma^2}, \qquad \mu := \frac{\delta}{\sigma^2}, \quad \textit{with} \quad \sigma^2 := \frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}, \qquad \delta := \sum_{i=1}^{N} \hat{\rho}_i(1-\hat{\rho}_i). \tag{3.1}$$

Note that the distribution of the cycle time C_i is related to the intervisit time I_i in the following way (see, e.g. [5]):

$$\mathbb{E}[I_i] = (1-\rho_i)\mathbb{E}[C_i], \qquad \mathbb{E}[\exp(-(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}]))I_i)] = \mathbb{E}[\mathrm{e}^{-sC_i}], \tag{3.2}$$

where ξ_i is the busy period of a regular M/G/1 queue with arrival rate λ_i and service time B_i. Note that (3.2) can easily be obtained by simple branching arguments. The next proposition characterizes the scaled intervisit times.

Proposition 3.2. (Convergence of intervisit times.) *Define* $\tilde{I}_i := (1-\rho)I_i$, *then for* $i = 1, \ldots, N$, *as* $\rho \uparrow 1$,

$$\tilde{I}_i \xrightarrow{\mathrm{D}} \tilde{\Gamma}_i,$$

where $\tilde{\Gamma}_i$ *has a gamma distribution with parameters*

$$\alpha := \frac{\mathbb{E}[S]\delta}{\sigma^2}, \qquad \mu_i := \frac{\delta}{(1-\hat{\rho}_i)\sigma^2}, \tag{3.3}$$

where δ *and* σ^2 *are given in (3.1).*

In what follows, we repeatedly use Propositions 3.1 and 3.2 to derive expressions for the asymptotic scaled waiting-time distribution associated with each of the service disciplines

considered herein. For each policy we use a two-step approach:

(a) derive an expression for the LST of the steady-state waiting-time distribution in terms of the cycle- or intervisit-time distribution;

(b) we combine this expression with Proposition 3.1 and/or Proposition 3.2 to obtain an expression for the LST of the waiting-time distribution in HT and interpret the resulting LST.

3.2. First-come–first-served

Here we illustrate the two-step approach described above for FCFS service. Regarding the first step, the following result gives an expression for the LST of the waiting time W_i in terms of the distribution of the intervisit time I_i (cf. [21]).

Proposition 3.3. (Waiting times in terms of intervisit times.) *For* $\mathrm{Re}(s) > 0$ *and* $\rho < 1$,

$$W_i^*(s) = \frac{(1-\rho_i)s}{s-\lambda_i(1-B_i^*(s))}\frac{1-I_i^*(s)}{s\mathbb{E}[I_i]}, \qquad i \in FCFS.$$

Next, as outlined in step (b), combining Propositions 3.2 and 3.3, the expression for $\mathbb{E}[C_i]$, and taking limits, we obtain, for $\mathrm{Re}(s) > 0$,

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) = \frac{1}{(1-\hat{\rho}_i)\mathbb{E}[S]s}\left\{1-\left(\frac{\mu_i}{\mu_i+s}\right)^{\alpha}\right\}, \qquad i \in FCFS.$$

Using (2.1), this leads to the following characterization of the limiting behavior of the scaled waiting-time distribution derived in [26].

Proposition 3.4. (Convergence of the waiting times.) *For* $\rho \uparrow 1$,

$$\tilde{W}_i \xrightarrow{\mathrm{D}} U_i \tilde{\boldsymbol{I}}_i, \qquad i \in FCFS,$$

where U_i *is a uniformly distributed RV on* $[0,1]$*, and* $\tilde{\boldsymbol{I}}_i$ *has a gamma distribution with parameters* $\alpha+1$ *and* μ_i*, where* α *and* μ_i *are given in (3.3).*

Note that $\tilde{\boldsymbol{I}}_i$ is the *length-biased* counterpart of $\tilde{I}_i$, a gamma distributed RV with parameters α and μ_i as in (3.3). Given the arrival of a particular customer, the cycle time is known to be biased and is given as the sum of the age and residual cycle time at the moment of arrival; see, e.g. [10] and the references therein. It is well known that if a gamma RV has parameters α and μ_i, then its length-biased version has parameters $\alpha+1$ and μ_i.

3.3. Intuition by the heavy traffic averaging principle

Proposition 3.4 states that the limiting behavior of W_i is of the form $U\Gamma$, where U is uniformly distributed on the interval [0, 1]. An intuitive explanation for this follows from the heavy traffic averaging principle (HTAP) combined with a fluid model; see [11], [12], [18]. Loosely speaking, the HTAP states that the work in each queue is emptied and refilled at a rate that is much faster than the rate at which the total workload is changing. This implies that the total workload can be considered as a constant during the course of a cycle, while the loads of the individual queues fluctuate like a fluid model.

In Figure 1 we provide a graphical representation of the fluid model. On the horizontal axis, the course of a cycle with fixed length c is plotted. The cycle is divided into two parts, the intervisit time I_i with length $(1-\hat{\rho}_i)c$ and the visit time V_i with length $\hat{\rho}_i c$. On the vertical axis,

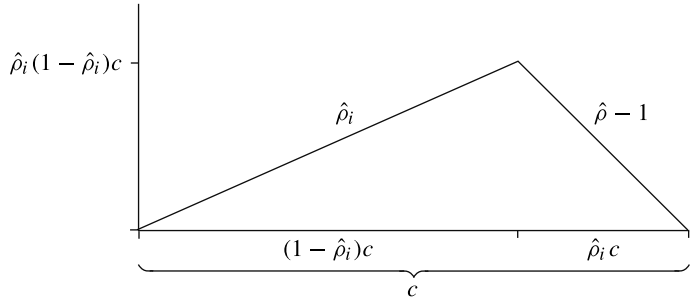


FIGURE 1: Fluid limits in heavy traffic; the amount of fluid in Q_i is plotted over the course of a cycle.

the workload in Q_i is plotted. The cycle starts at the completion of a visit to Q_i. Throughout the cycle, work arrives with intensity 1 and a fraction $\hat{\rho}_i$ is directed to Q_i. During the visit time V_i work flows out of Q_i with rate 1 until the queue is empty.

Let the uniform RV U on [0,1] denote the fraction of the cycle c that has elapsed at the arrival epoch of an arbitrary particle. The particle has to wait for the remaining length of the cycle $(1-U)c$ except for the amount of work that arrives at Q_i during the cycle after the arrival of the particle. As work to Q_i arrives at rate $\hat{\rho}_i$, the latter can be expressed as $\hat{\rho}_i(1-U)c$. Hence, the waiting time can be expressed as $(1-U)c - \hat{\rho}_i(1-U)c = (1-U)(1-\hat{\rho}_i)c$. Using the fact that $U[0,1]$ is in distribution equal to $1-U[0,1]$ and $I_i = (1-\hat{\rho}_i)c$, we conclude that $\tilde{W}_i$ is uniformly distributed on $[0,1]I_i$. This interpretation gives much insight into the HT asymptotics.

For the other service disciplines, HTAP and the cycle time, represented by Γ, are unaffected. To interpret the results in the light of HTAP, we need to study the fluid model for each discipline. For compactness of this paper the details are omitted. We refer the reader to [29] for the HTAP intuitions for the other service disciplines.

4. Last-come–first-served

In this section we consider the LCFS service discipline. In Section 4.1 we derive the results for LCFS without preemption and in Section 4.2 we look at queues with LCFS preemptive resume (LCFS-PR) service. In both sections we first provide a derivation of the LST of W_i for all $\rho < 1$, giving insight into the terms contributing to the delay. Then we study the behavior of $\tilde{W}_i$ in the HT regime. Since we are interested in deriving the waiting-time distributions of customers that arrive in steady state, it is convenient to define stationary versions of the arrival processes on the entire real line. Hence, each arrival process N_i consists of points $\{T_{i,n}\}_{n\in\mathbb{Z}}$, where $T_{i,0} \le 0 \le T_{i,1}$. Associated with each point is the busy period $\xi_{i,n}$ generated by the arriving customer. The points $(T_{i,n}, \xi_{i,n})$ define a marked Poisson process on Re^2.

4.1. Nonpreemptive LCFS

The LST for LCFS with rest periods was found in [19]; replacing the rest periods with intervisit times and adding the subscript i to the queue-dependent variables yields the following proposition. We refer the reader to [29] for a more direct derivation.

Proposition 4.1. *For $\rho < 1$, $\mathrm{Re}(s) > 0$,*

$$W_i^*(s) = \rho_i \frac{1 - \mathbb{E}[\mathrm{e}^{-s\xi_i}]}{(s + \lambda_i(1 - \mathbb{E}[\mathrm{e}^{-s\xi_i}]))\mathbb{E}[B_i]} + (1 - \rho_i)\frac{1 - \mathbb{E}[\mathrm{e}^{-sC_i}]}{(s + \lambda_i(1 - \mathbb{E}[\mathrm{e}^{-s\xi_i}]))\mathbb{E}[C](1 - \rho_i)}, \qquad i \in LCFS. \tag{4.1}$$

Note that the first term appears in the LST of the waiting time in an M/G/1 queue with LCFS service order; see, e.g. [23, p. 357].

The following result gives an expression for the asymptotic waiting-time distribution for LCFS service in heavy traffic.

Theorem 4.1. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \xrightarrow{\mathrm{D}} \begin{cases} 0 & \textit{with probability (w.p.) } \hat{\rho}_i, \\ U_i \tilde{\boldsymbol{C}}_i & \textit{w.p. } 1 - \hat{\rho}_i, \end{cases} \qquad i \in LCFS,$$

where U_i is a uniformly distributed RV on the interval $[0, 1]$ and $\tilde{\boldsymbol{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in (3.1).

Proof. Combining Proposition 4.1 with Proposition 3.1 gives the following expressions for the LST of the (scaled) waiting-time distribution. For $i \in LCFS$, $\mathrm{Re}(s) > 0$,

$$\begin{aligned} \tilde{W}_i^*(s) &= \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) \\ &= \lim_{\rho\uparrow 1} \left(\rho_i \frac{1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]}{(s(1-\rho) + \lambda_i(1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]))\mathbb{E}[B_i]} \right. \\ &\qquad \left. + (1 - \rho_i)\frac{1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)C_i}]}{(s(1-\rho) + \lambda_i(1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]))\mathbb{E}[C](1 - \rho_i)} \right). \end{aligned} \tag{4.2}$$

Let us initially consider the first term on the right-hand side of (4.2):

$$\begin{aligned} &\lim_{\rho\uparrow 1} \rho_i \frac{1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]}{(s(1-\rho) + \lambda_i(1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]))\mathbb{E}[B_i]} \\ &\quad = \lim_{\rho\uparrow 1} \rho_i \frac{(1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}])/(1-\rho)}{s\mathbb{E}[B_i] + \rho_i((1 - \mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}])/(1-\rho))} \\ &\quad = \hat{\rho}_i \frac{\mathbb{E}[\xi_i]s}{\mathbb{E}[B_i]s + \hat{\rho}_i\mathbb{E}[\xi_i]s} \\ &\quad = \hat{\rho}_i. \end{aligned}$$

In the second equality, we use l'Hôpital's rule and the fact that the derivative of $\mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]$ at $s(1-\rho) = 0$ can be expressed as $-\mathbb{E}[\xi_i]$. For the third equality, we apply the well-known result $\mathbb{E}[\xi_i] = \mathbb{E}[B_i]/(1 - \rho_i)$.

Now consider the second term on the right-hand side of (4.2):

$$\begin{aligned}
&\lim_{\rho\uparrow 1}(1-\rho_i)\frac{1-\mathbb{E}[\mathrm{e}^{-s(1-\rho)C_i}]}{\mathbb{E}[C](1-\rho_i)(s(1-\rho)+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}]))}\\
&\quad=\lim_{\rho\uparrow 1}(1-\rho_i)\frac{1-(\mu/(\mu+s))^{\alpha}}{\mathbb{E}[S](1-\rho_i)(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s(1-\rho)\xi_i}])/(1-\rho))}\\
&\quad=(1-\hat{\rho}_i)\frac{1-(\mu/(\mu+s))^{\alpha}}{\mathbb{E}[S](1-\hat{\rho}_i)s(1+\lambda_i\mathbb{E}[\xi_i])}\\
&\quad=(1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\}. \qquad (4.3)
\end{aligned}$$

Combining the above, we have

$$\tilde{W}_i^*(s)=\hat{\rho}_i+(1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\},\qquad i\in LCFS, \qquad (4.4)$$

where α and μ are given in (3.1). Note that (4.4) corresponds to the LST of a RV that is equal to 0 with probability $\hat{\rho}_i$ and to a uniform RV on [0, 1] multiplied by a gamma distribution with probability $1-\hat{\rho}_i$. This completes the proof.

4.2. LCFS with preemptive resume

We derive the LST of the waiting time of a tagged customer T that arrives at queue i in steady state. Without loss of generality, we assume that T arrives at time 0. We have to distinguish between the case where T arrives during an intervisit time (case 1), and the case where T arrives during a visit time (case 2).

Case 1. When an arrival occurs during an intervisit time, the waiting time of the customer consists of the busy periods generated by the customers arriving during the service of the tagged customer, the residual intervisit time and the busy periods generated by the customers arriving during the residual intervisit time. For $i\in LCFS\text{-}PR$,

$$W_i\ (\textit{given } T \text{ arrives during intervisit time})=\sum_{T_{i,k}\in(0,B_i)}\xi_{i,k}+I_i^{\mathrm{res}}+\sum_{T_{i,k}\in(0,I_i^{\mathrm{res}})}\xi_{i,k}. \qquad (4.5)$$

Case 2. When the arrival occurs during a visit period, the waiting time of T consists of the busy period generated by customers arriving during the service of the tagged customer. For $i\in LCFS\text{-}PR$,

$$W_i\ (\textit{given } T \text{ arrives during visit time})=\sum_{T_{i,k}\in(0,B_i)}\xi_{i,k}. \qquad (4.6)$$

Due to the preemptive nature of the discipline, the first term in (4.5) is equal to (4.6), the waiting time in case 2, so we calculate the LST of the waiting time of case 2 first. Conditioning on the service time and the number of arrivals therein (as in [10]) yields, for $i\in LCFS\text{-}PR$,

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-sW_i}\mid T \text{ arrives during visit time}]&=\mathbb{E}[\mathrm{e}^{-s(\sum_{T_{i,k}\in(0,B_i)}\xi_i)}]\\
&=\int_{t=0}^{\infty}\sum_{n=0}^{\infty}\mathrm{e}^{-\lambda_i t}\frac{(\lambda_i t)^n}{n!}\mathbb{E}[\mathrm{e}^{-s\xi_i}]^n\,\mathrm{d}\mathbb{P}(B_i\le t)\\
&=\int_{t=0}^{\infty}\exp(-t(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])))\,\mathrm{d}\mathbb{P}(B_i\le t)\\
&=B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])).
\end{aligned}$$

For the last two terms in (4.5) we condition on I_i^{res} and the number of arrivals during I_i^{res}, yielding for $\mathrm{Re}(s) > 0$, $i \in LCFS\text{-}PR$,

$$\begin{aligned}
&\mathbb{E}[\mathrm{e}^{-sW_i} \mid T \text{ arrives during intervisit time}] \\
&\quad = B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])) \int_{t=0}^{\infty} \mathrm{e}^{-st} \sum_{n=0}^{\infty} \mathrm{e}^{-\lambda_i t} \frac{(\lambda_i t)^n}{n!} \mathbb{E}[\mathrm{e}^{-s\xi_i}]^n \,\mathrm{d}\mathbb{P}(I_i^{\mathrm{res}} \le t) \\
&\quad = B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])) \int_{t=0}^{\infty} \exp(-t(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}]))) \,\mathrm{d}\mathbb{P}(I_i^{\mathrm{res}} \le t) \\
&\quad = B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])) \frac{1-\mathbb{E}[\exp(-(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])I_i))]}{(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}]))\mathbb{E}[I_i]} \\
&\quad = B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])) \frac{1-\mathbb{E}[\mathrm{e}^{-sC_i}]}{(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}]))\mathbb{E}[C](1-\rho_i)},
\end{aligned}$$

where the last equality follows from (3.2).

We combine the two cases, using the fact that the probability that an arrival occurs during a visit time is equal to ρ_i. This leads to the following expression for the LST of the waiting time at Q_i in terms of the cycle time.

Proposition 4.2. *For $\rho < 1$, $\mathrm{Re}(s) > 0$,*

$$\begin{aligned}
W_i^*(s) &= B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])) \\
&\quad \times \left(\rho_i + (1-\rho_i)\frac{1-\mathbb{E}[\mathrm{e}^{-sC_i}]}{(s+\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}]))\mathbb{E}[C](1-\rho_i)}\right), \qquad i \in LCFS\text{-}PR.
\end{aligned}$$

The next result gives the HT limit of the distribution of $\tilde{W}_i$.

Theorem 4.2. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \xrightarrow{\mathrm{D}} \begin{cases} 0 & \text{w.p. } \hat{\rho}_i, \\ U_i \tilde{\boldsymbol{C}}_i & \text{w.p. } 1-, \end{cases} \qquad i \in LCFS\text{-}PR,$$

where U_i is a uniformly distributed RV on the interval $[0,1]$ and $\tilde{\boldsymbol{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in (3.1).

Proof. Using (4.3) and the fact that $\lim_{\rho\uparrow 1} B_i^*(\lambda_i(1-\mathbb{E}[\mathrm{e}^{-s\xi_i}])) = 1$ for $\mathrm{Re}(s) > 0$, we immediately see that the LST of $\tilde{W}_i$, $i \in LCFS\text{-}PR$, in HT is given by

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) = \hat{\rho}_i + (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1 - \left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\} \tag{4.7}$$

with α and μ given in (3.1).

Note that the HT scaled waiting-time distribution (4.7) for $i \in LCFS\text{-}PR$ is equal to the HT scaled waiting-time distribution (4.4) for $i \in LCFS$. This is true because the busy periods generated by customers arriving during service of the tagged customer do not scale with ρ. A similar argument explains the probability mass in zero of the scaled waiting-time distribution.

5. Random-order-of-service

In this section we first derive the LST of the scaled waiting-time distribution for ROS in terms of the intervisit times. Then we use this result to obtain the waiting-time distribution in heavy traffic.

Proposition 5.1. *For $\rho < 1, \mathrm{Re}(s) > 0$,*

$$W_i^*(s) = \frac{1-\rho_i}{s\mathbb{E}[I_i]} \times \left(\int_{x=\xi_i^*(s)}^1 \frac{1 - I_i^*(\lambda_i - \lambda_i x)}{B_i^*(\lambda_i - \lambda_i x) - x} (B_i^*(\lambda_i(1-x)) - B_i^*(s + \lambda_i(1-x)))\, \mathrm{d}K(x,s) + \int_{x=\xi_i^*(s)}^1 (I_i^*(\lambda_i(1-x)) - I_i^*(s+\lambda_i(1-x)))\, \mathrm{d}K(x,s) \right), \qquad i \in ROS$$

with $\xi_i^*(s) = B_i^*(s + \lambda_i(1 - \xi_i^*(s)))$, the LST of a busy period at queue i with a dedicated server, and

$$K(x,s) := \exp\left(- \int_{y=x}^1 \frac{1}{y - B_i^*(s + \lambda_i - \lambda_i y)}\, \mathrm{d}y \right). \tag{5.1}$$

Proof. The derivation proceeds along the lines of Kingman [17]. Define the waiting time of a tagged type-i customer T as $w = u + v$. Here, u is the time between the arrival instant of T and the time the server begins working on a new type-i customer, and v is the time from that moment until T is taken into service. A customer may arrive during an intervisit period of Q_i, in which case $u = I_i^{\mathrm{res}}$, or during a visit period, yielding $u = B_i^{\mathrm{res}}$.

For v, we first consider the transform of the number of customers at moments when the server is able to take a customer from queue i into service, denoted as $Q(z, X)$, for $|z| < 1$ and with $X \in \{\boldsymbol{B}_i, \boldsymbol{I}_i\}$. From Kawasaki *et al.* [15] for an arrival during a visit period, we have

$$Q(z, \boldsymbol{B}_i) = \frac{(1-\rho_i)(1 - I_i^*(\lambda_i - \lambda_i z))\mathrm{e}^{-\lambda_i(1-z)\boldsymbol{B}_i}}{\lambda_i \mathbb{E}[I_i](B_i^*(\lambda_i - \lambda_i z) - z)}.$$

If the customer arrives during an intervisit period we have, for $|z| < 1$, $i \in ROS$,

$$Q(z, \boldsymbol{I}_i) = \mathrm{e}^{-\lambda_i(1-z)\boldsymbol{I}_i}.$$

Kingman [17, Theorem 2] provides the LST of v given the number of customers present. Combining this theorem with the equations above, we obtain the LST of v for an arrival during a visit period while a customer of size $\boldsymbol{B}_i$ is in service: For $\mathrm{Re}(s) > 0$, $i \in ROS$,

$$\mathbb{E}[\mathrm{e}^{-sv} \mid \boldsymbol{B}_i \text{ and arrival during visit period}] = \int_{\xi_i^*(s)}^1 \frac{(1-\rho_i)(1 - I_i^*(\lambda_i - \lambda_i x))\mathrm{e}^{-\lambda_i(1-x)\boldsymbol{B}_i}}{\lambda_i \mathbb{E}[I_i](B_i^*(\lambda_i - \lambda_i x) - x)}\, \mathrm{d}K(x,s).$$

Similarly, we have for a customer arriving during an intervisit period of length $\boldsymbol{I}_i$, for $\mathrm{Re}(s) > 0$, $i \in ROS$,

$$\mathbb{E}[\mathrm{e}^{-sv} \mid \boldsymbol{I}_i \text{ and arrival during intervisit period}] = \int_{\xi_i^*(s)}^1 \mathrm{e}^{-\lambda_i(1-x)\boldsymbol{I}_i}\, \mathrm{d}K(x,s).$$

Note that given $\boldsymbol{B}_i$ or $\boldsymbol{I}_i$, u and v are independent. For an arrival during a visit while a customer of size $\boldsymbol{B}_i$ is in service, we obtain, for $\mathrm{Re}(s) > 0$, $i \in ROS$,

$$\begin{aligned}
&\mathbb{E}[\mathrm{e}^{-sw} \mid \boldsymbol{B}_i] \\
&\quad = \mathbb{E}[\mathrm{e}^{-sB_i^{\mathrm{res}}} \mid \boldsymbol{B}_i]\mathbb{E}[\mathrm{e}^{-sv} \mid \boldsymbol{B}_i] \\
&\quad = \frac{1-\mathrm{e}^{-s\boldsymbol{B}_i}}{s\boldsymbol{B}_i} \int_{\xi_i^*(s)}^1 \frac{(1-\rho_i)(1-I_i^*(\lambda_i-\lambda_i x))\mathrm{e}^{-\lambda_i(1-x)\boldsymbol{B}_i}}{\lambda_i \mathbb{E}[I_i](B_i^*(\lambda_i-\lambda_i x)-x)}\,\mathrm{d}K(x,s) \\
&\quad = \frac{1-\rho_i}{s\lambda_i\mathbb{E}[I_i]} \int_{\xi_i^*(s)}^1 \frac{1-I_i^*(\lambda_i-\lambda_i x)}{B_i^*(\lambda_i-\lambda_i x)-x} \frac{\mathrm{e}^{-\lambda_i(1-x)\boldsymbol{B}_i}-\mathrm{e}^{-(s+\lambda_i(1-x))\boldsymbol{B}_i}}{\boldsymbol{B}_i}\,\mathrm{d}K(x,s).
\end{aligned}$$

Now, using the fact that $\mathbb{E}[\mathrm{e}^{-\phi\boldsymbol{B}_i}/\boldsymbol{B}_i] = B_i^*[\phi]/\mathbb{E}[B_i]$ (see [17]), we have, for $\mathrm{Re}(s) > 0$, $i \in ROS$,

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mathrm{e}^{-sw} \mid \boldsymbol{B}_i]] \\
&\quad = \frac{1-\rho_i}{s\lambda_i\mathbb{E}[I_i]} \int_{\xi_i^*(s)}^1 \frac{1-I_i^*(\lambda_i-\lambda_i x)}{B_i^*(\lambda_i-\lambda_i x)-x} \frac{B_i^*(\lambda_i(1-x))-B_i^*(s+\lambda_i(1-x))}{\mathbb{E}[B_i]}\,\mathrm{d}K(x,s).
\end{aligned}$$

Again it holds that a customer arrives with probability ρ_i during a visit period. Hence, $W_i^*(s) = \rho_i\mathbb{E}[\mathbb{E}[\mathrm{e}^{-sw} \mid B_i]] + (1-\rho_i)\mathbb{E}[\mathbb{E}[\mathrm{e}^{-sw} \mid I_i]]$. Using similar arguments for the final term in addition to some rewriting, we obtain the result.

Next, we turn to the HT limit. Before we state our result, we define Y as a RV with PDF and CDF

$$f_Y(y) = \frac{(1-y)^{\hat{\rho}_i/(1-\hat{\rho}_i)}}{(1-\hat{\rho}_i)}, \qquad F_Y(y) = 1-(1-y)^{1/(1-\hat{\rho}_i)}, \quad y \in [0,1].$$

The RV Y is to be interpreted as the fraction of customers, including both present customers and those arriving until the server's departure from the queue, that are served before the arriving customer (see [29] for an interpretation in terms of a fluid model).

The next theorem gives the HT limit of the distribution of $\tilde{W}_i$ in terms of Y.

Theorem 5.1. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \xrightarrow{\mathrm{D}} \begin{cases} U_i^f \tilde{\boldsymbol{C}} & \text{w.p. } \hat{\rho}_i, \\ U_i^g \tilde{\boldsymbol{C}} & \text{w.p. } 1-\hat{\rho}_i, \end{cases} \qquad i \in ROS,$$

where U_i^f has a uniform distribution on the interval $[0, Y\hat{\rho}_i]$ and U_i^g has a uniform distribution on $[Y\hat{\rho}_i, 1-\hat{\rho}_i+Y\hat{\rho}_i]$.

Proof. First we rewrite the LST of the waiting time given in Proposition 5.1. Noting that

$$\frac{\mathrm{d}K(x,s)}{\mathrm{d}x} = \frac{K(x,s)}{x-B_i^*(s+\lambda_i(1-x))},$$

we obtain

$$W_i^*(s) = \frac{1-\rho_i}{s\mathbb{E}[I_i]}\bigg(\int_{x=\xi_i^*(s)}^1 K(x,s)(1-I_i^*(\lambda_i-\lambda_i x)) \times\bigg(\frac{1}{B_i^*(\lambda_i(1-x))-x}+\frac{1}{x-B_i^*(s+\lambda_i(1-x))}\bigg)\,\mathrm{d}x + \int_{x=\xi_i^*(s)}^1 K(x,s)(I_i^*(\lambda_i(1-x))-I_i^*(s+\lambda_i(1-x))) \times\frac{1}{x-B_i^*(s+\lambda_i(1-x))}\,\mathrm{d}x\bigg).$$

In line with Takagi and Kudoh [22] we take $y=(1-x)/(1-\xi_i^*(s))$; this gives $x=1-y(1-\xi_i^*(s))$ and $\mathrm{d}x=-(1-\xi_i^*(s))\,\mathrm{d}y$, yielding

$$W_i^*(s) = \frac{1-\rho_i}{s\mathbb{E}[I_i]}\bigg(\int_{y=0}^1 K(1-y(1-\xi_i^*(s)),s)(1-I_i^*(y\lambda_i(1-\xi_i^*(s)))) \times\bigg(\frac{1-\xi_i^*(s)}{B_i^*(y\lambda_i(1-\xi_i^*(s)))-1+y(1-\xi_i^*(s))} + \frac{1-\xi_i^*(s)}{1-y(1-\xi_i^*(s))-B_i^*(s+y\lambda_i(1-\xi_i^*(s)))}\bigg)\,\mathrm{d}y + \int_{y=0}^1 K(1-y(1-\xi_i^*(s)),s) \times(I_i^*(y\lambda_i(1-\xi_i^*(s)))-I_i^*(s+y\lambda_i(1-\xi_i^*(s)))) \times\bigg(\frac{1-\xi_i^*(s)}{1-y(1-\xi_i^*(s))-B_i^*(s+y\lambda_i(1-\xi_i^*(s)))}\bigg)\,\mathrm{d}y\bigg).$$

We now take HT limits for the terms separately. We start with the most involved term, $K(x,s)$. Using the substitution $t=(1-y)/(1-x)$ in (5.1), we write

$$K(x,s)=\exp\bigg(-\int_{t=0}^1\frac{1-x}{1-t(1-x)-B_i^*(s+\lambda_i t(1-x))}\,\mathrm{d}t\bigg).$$

Taking the HT limit of $K(1-y(1-\xi_i^*(s)),s)$, we obtain, using l'Hôpital's rule and some rewriting,

$$\begin{aligned}&\lim_{\rho\uparrow 1}K(1-y(1-\xi_i^*(s(1-\rho))),s(1-\rho))\\&\quad=\exp\bigg(-\int_{t=0}^1\frac{y\mathbb{E}[\xi_i]}{-\mathbb{E}[\xi_i]ty+\mathbb{E}[B_i](1+\lambda_i ty\mathbb{E}[\xi_i])}\,\mathrm{d}t\bigg)\\&\quad=\exp\bigg(-\frac{y}{1-\hat\rho_i}\int_{t=0}^1\frac{1}{1-ty}\,\mathrm{d}t\bigg)\\&\quad=\exp\bigg(\frac{1}{1-\hat\rho_i}\ln(1-y)\bigg)\\&\quad=(1-y)^{1/(1-\hat\rho_i)}.\end{aligned}$$

In the second step we use the fact that $\mathbb{E}[\xi_i] = \mathbb{E}[B_i]/(1-\hat{\rho}_i)$. The HT limits for the other terms can be determined using l'Hôpital's rule in addition to some rewriting and the expression for $\mathbb{E}[\xi_i]$ above. In particular, we obtain

$$\lim_{\rho\uparrow 1} I_i^*(y\lambda_i(1-\xi_i^*(s(1-\rho)))) = \tilde{I}_i^*\left(\frac{y\hat{\rho}_i s}{1-\hat{\rho}_i}\right),$$

$$\lim_{\rho\uparrow 1} I_i^*(s(1-\rho)+y\lambda_i(1-\xi_i^*(s(1-\rho)))) = \tilde{I}_i^*\left(\frac{s(1-\hat{\rho}_i+y\hat{\rho}_i)}{1-\hat{\rho}_i}\right),$$

$$\lim_{\rho\uparrow 1} \frac{1-\xi_i^*(s(1-\rho))}{B_i^*(y\lambda_i(1-\xi_i^*(s(1-\rho)))) - 1 + y(1-\xi_i^*(s(1-\rho)))} = \frac{1}{y(1-\hat{\rho}_i)},$$

$$\lim_{\rho\uparrow 1} \frac{1-\xi_i^*(s(1-\rho))}{1-y(1-\xi_i^*(s(1-\rho))) - B_i^*(s(1-\rho)+y\lambda_i(1-\xi_i^*(s(1-\rho))))} = \frac{1}{(1-y)(1-\hat{\rho}_i)}.$$

Moreover, we have $\tilde{I}_i^*(cs/(1-\hat{\rho}_i)) = \tilde{C}_i^*(cs) = (\mu/(\mu+cs))^\alpha$ for fixed $c>0$. Combining the above, after some rewriting, we obtain

$$\begin{aligned}
\tilde{W}_i^*(s) &= \frac{1-\hat{\rho}_i}{s\mathbb{E}[S](1-\hat{\rho}_i)} \\
&\quad\times\left(\int_{y=0}^1\left(1-\tilde{I}_i^*\left(\frac{y\hat{\rho}_i s}{1-\hat{\rho}_i}\right)\right)\frac{(1-y)^{1/(1-\hat{\rho}_i)}}{y(1-y)(1-\hat{\rho}_i)}\,\mathrm{d}y\right. \\
&\qquad\left.+\int_{y=0}^1\left(\tilde{I}_i^*\left(\frac{y\hat{\rho}_i s}{1-\hat{\rho}_i}\right)-\tilde{I}_i^*\left(\frac{s(1-\hat{\rho}_i+y\hat{\rho}_i)}{1-\hat{\rho}_i}\right)\right)\frac{(1-y)^{1/(1-\hat{\rho}_i)}}{(1-y)(1-\hat{\rho}_i)}\,\mathrm{d}y\right) \\
&= \hat{\rho}_i\int_{y=0}^1\frac{1}{s\mathbb{E}[S]y\hat{\rho}_i}\left\{1-\left(\frac{\mu}{\mu+y\hat{\rho}_i s}\right)^\alpha\right\}\frac{(1-y)^{\hat{\rho}_i/(1-\hat{\rho}_i)}}{(1-\hat{\rho}_i)}\,\mathrm{d}y \\
&\quad+(1-\hat{\rho}_i)\int_{y=0}^1\frac{1}{s\mathbb{E}[S](1-\hat{\rho}_i)}\left\{\left(\frac{\mu}{\mu+y\hat{\rho}_i s}\right)^\alpha-\left(\frac{\mu}{\mu+s(1-\hat{\rho}_i+y\hat{\rho}_i)}\right)^\alpha\right\} \\
&\qquad\times\frac{(1-y)^{\hat{\rho}_i/(1-\hat{\rho}_i)}}{(1-\hat{\rho}_i)}\,\mathrm{d}y.
\end{aligned}$$

This LST corresponds to a mixture of two distributions. With probability $\hat{\rho}_i$ and conditioning on $Y=y$, it is the LST of a uniform $[0, y\hat{\rho}_i]$ multiplied by a gamma distribution with parameters $\alpha+1$ and μ; with probability $1-\hat{\rho}_i$ and conditioning on $Y=y$, it is the LST of a uniform $[y\hat{\rho}_i, 1-\hat{\rho}_i+y\hat{\rho}_i]$ multiplied by a gamma distribution with the same parameters. This completes the proof.

Remark 5.1. The expressions for U_i^f and U_i^g in Theorem 5.1 can be rewritten more explicitly, similar to those in Theorem 6.2; see also Remark 6.1.

6. Processor sharing

In a PS queue all customers present at the queue receiving service are served simultaneously and at the same rate. We note that the waiting time W_i (to be interpreted as the delay) is thus defined as the sojourn time minus the service requirement. In this section we will only consider the case of exponentially distributed service times. We extend the work of [1], where the authors derived the HT limit of the LST of the scaled waiting time conditional on the service requirement. In Section 6.1 we give the conditional scaled waiting-time distribution. In Section 6.2 we derive the unconditional scaled waiting-time distribution.

6.1. Conditional waiting-time distribution in heavy traffic

Let customers in Q_i have exponentially distributed service requirements with rate b_i. Let x be the required service duration of a tagged customer. Then we have the following theorem for the HT limit of the conditional waiting time $W_i \mid x$.

Theorem 6.1. *For $\rho \uparrow 1$, $x \geq 0$,*

$$\tilde{W}_i \mid x \xrightarrow{\mathrm{D}} \begin{cases} U_{i,x}^f \tilde{\boldsymbol{I}}_i & \text{w.p. } \hat{\rho}_i, \\ U_{i,x}^g \tilde{\boldsymbol{I}}_i & \text{w.p. } 1-\hat{\rho}_i, \end{cases} \qquad i \in PS,$$

where $U_{i,x}^f = U[0, \omega(x)]$, $U_{i,x}^g = U[\omega(x), \omega(x)+1]$, and $\tilde{\boldsymbol{I}}_i \sim \Gamma(\alpha+1, \mu_i)$. The parameters α and μ_i can be found in (3.3), and $\omega(x) = \hat{\rho}_i/(1-\hat{\rho}_i)(1-\mathrm{e}^{-b_i x(1-\hat{\rho}_i)})$.

Proof. The authors of [1] derive the LST of the scaled conditional waiting time in heavy traffic. For $\rho \uparrow 1$, $x \geq 0$, $i \in PS$,

$$\begin{aligned} \tilde{W}_i^*(s \mid x) = {} & \frac{\hat{\rho}_i}{s\omega(x)\mathbb{E}[S](1-\hat{\rho}_i)} \left\{ 1 - \left(\frac{\mu_i}{\mu_i + s\omega(x)} \right)^{\alpha} \right\} \\ & + \frac{1-\hat{\rho}_i}{s\mathbb{E}[S](1-\hat{\rho}_i)} \left\{ \left(\frac{\mu_i}{\mu_i + s\omega(x)} \right)^{\alpha} - \left(\frac{\mu_i}{\mu_i + s(\omega(x)+1)} \right)^{\alpha} \right\}. \end{aligned}$$

From this LST we see that the distribution of the conditional waiting time is a uniform $[0, \omega(x)]$ multiplied by a gamma distribution with parameters $\alpha + 1$ and μ_i with probability $\hat{\rho}_i$. With probability $1 - \hat{\rho}_i$, the conditional waiting time has a uniform $[\omega(x), \omega(x) + 1]$ multiplied by a gamma distribution with parameters $\alpha + 1$ and μ_i. This completes the proof.

We note that $\omega(x) I_i$ can be interpreted as the sojourn time of a tagged customer with service time x from the start of the visit period; see [29].

6.2. Unconditional waiting-time distribution in heavy traffic

In the previous section we derived the HT limit of the waiting-time distribution conditional on the service requirement. To obtain the *unconditional* waiting-time distribution, we first consider a more general setting that also covers 'unconditioning' for SJF. Suppose we have a conditional RV, denoted $T \mid x$, where x is a realization of a RV X with support $x \in [x_{\min}, x_{\max}]$. We have the following lemma.

Lemma 6.1. *Assume that the conditional RV $T \mid x$ has density $f_{T \mid x}(y)$ and distribution function $F_{T \mid x}(y)$ with support $y \in [a(x), b(x)]$. Suppose that $a(x)$ and $b(x)$ are both increasing in x and $a(x) < b(x)$ for all x. Let $a^{-1}(\cdot)$ be the inverse of $a(\cdot)$ and $b^{-1}(\cdot)$ be the inverse of $b(\cdot)$. Then, the unconditional distribution of $T \mid x$, denoted by $\tilde{T}$, has PDF, for $a(x_{\max}) \leq b(x_{\min})$,*

$$f_{\tilde{T}}(y) = \begin{cases} \displaystyle\int_{x=x_{\min}}^{a^{-1}(y)} f_{T \mid x}(y) f_X(x) \, \mathrm{d}x, & y \in [a(x_{\min}), a(x_{\max})], \\ \displaystyle\int_{x=x_{\min}}^{x_{\max}} f_{T \mid x}(y) f_X(x) \, \mathrm{d}x, & y \in [a(x_{\max}), b(x_{\min})], \\ \displaystyle\int_{x=b^{-1}(y)}^{x_{\max}} f_{T \mid x}(y) f_X(x) \, \mathrm{d}x, & y \in [b(x_{\min}), b(x_{\max})] \end{cases} \tag{6.1}$$

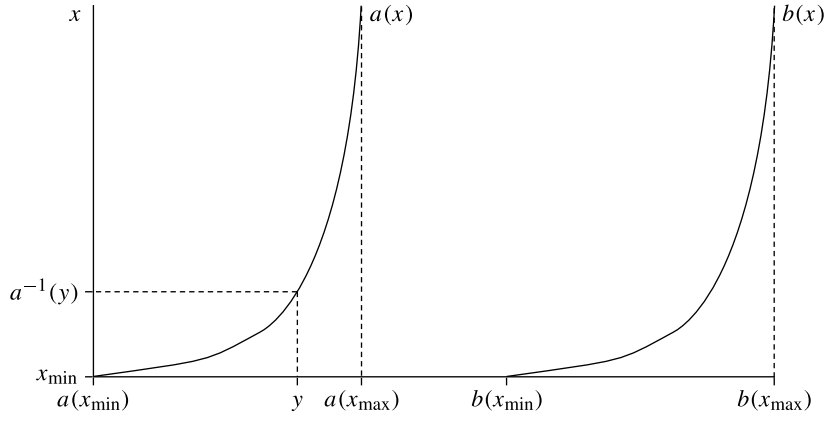


FIGURE 2: Boundaries of the conditional distribution.

and, for $a(x_{\max}) > b(x_{\min})$,

$$f_{\tilde{T}}(y) = \begin{cases} \displaystyle\int_{x=x_{\min}}^{a^{-1}(y)} f_{T\,|\,x}(y) f_X(x)\,\mathrm{d}x, & y \in [a(x_{\min}), b(x_{\min})], \\ \displaystyle\int_{x=b^{-1}(y)}^{a^{-1}(y)} f_{T\,|\,x}(y) f_X(x)\,\mathrm{d}x, & y \in [b(x_{\min}), a(x_{\max})], \\ \displaystyle\int_{x=b^{-1}(y)}^{x_{\max}} f_{T\,|\,x}(y) f_X(x)\,\mathrm{d}x, & y \in [a(x_{\max}), b(x_{\max})]. \end{cases}$$

Proof. First consider the case that $a(x_{\max}) \leq b(x_{\min})$. In Figure 2 we show an example of the boundaries of the conditional distribution, plotting $a(x)$ and $b(x)$ with x on the vertical axis. The possible values of $T \mid x$ then lie between the two lines. To find $f_{\tilde{T}}(y)$, we need to integrate out x with respect to its density function. First, take $y \in [a(x_{\min}), a(x_{\max})]$, in which case the PDF $f_{\tilde{T}}(y)$ is obtained from the parts where x is smaller than $a^{-1}(y)$. We have

$$f_{\tilde{T}}(y) = \int_{x=x_{\min}}^{a^{-1}(y)} f_{T\,|\,x}(y) f_X(x)\,\mathrm{d}x.$$

If $y \in [a(x_{\max}), b(x_{\min})]$ then y is between the boundaries of the conditional distribution for every $x \in [x_{\min}, x_{\max}]$. This yields the second case of (6.1). Finally, for $y \in [b(x_{\min}), b(x_{\max})]$, $f_{\tilde{T}}(y)$ can now be obtained from the parts where x is larger than $b^{-1}(y)$. The $a(x_{\max}) > b(x_{\min})$ case is similar. It may be checked that $f_{\tilde{T}}(\cdot)$ is a density function. This completes the proof.

Note that the distribution in (6.1) is continuous, increasing on $[a(x_{\min}), a(x_{\max})]$, constant on $[a(x_{\max}), b(x_{\min})]$, and decreasing on $[b(x_{\min}), b(x_{\max})]$, which closely resembles the traditional trapezoidal distribution. In line with [28], we refer to (6.1) as a *generalized trapezoidal distribution*.

We now apply Lemma 6.1 to the $i \in PS$ case, in which we have two conditional distributions, $U_{i,x}^{f}$ and $U_{i,x}^{g}$. We need to find the unconditional versions of both uniform distributions.

Theorem 6.2. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \xrightarrow{\mathrm{D}} \begin{cases} \tilde{U}_i^f \tilde{\boldsymbol{I}}_i & \text{w.p. } \hat{\rho}_i, \\ \tilde{U}_i^g \tilde{\boldsymbol{I}}_i & \text{w.p. } 1-\hat{\rho}_i, \end{cases} \qquad i \in PS,$$

where $\tilde{U}_i^f$ has a generalized trapezoidal distribution with PDF

$$f_{\tilde{U}_i}(y) = \frac{1}{\hat{\rho}_i} \mathrm{beta}_{1-y(1-\hat{\rho}_i)/\hat{\rho}_i}\left(1+\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0\right), \qquad y \in \left[0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}\right],$$

where $\mathrm{beta}_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$. *Then $\tilde{U}_i^g$ has a generalized trapezoidal distribution with PDF, for $\hat{\rho}_i \le \frac{1}{2}$,*

$$g_{\tilde{U}_i}(y) = \begin{cases} 1-\left(1-\dfrac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{1/(1-\hat{\rho}_i)}, & y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i)), \\ 1, & y \in [\hat{\rho}_i/(1-\hat{\rho}_i), 1], \\ \left(1-\dfrac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{1/(1-\hat{\rho}_i)}, & y \in (1, \hat{\rho}_i/(1-\hat{\rho}_i)+1] \end{cases} \tag{6.2}$$

and, for $\hat{\rho}_i > \frac{1}{2}$,

$$g_{\tilde{U}_i}(y) = \begin{cases} 1-\left(1-\dfrac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{1/(1-\hat{\rho}_i)}, & y \in [0,1), \\ \left(1-\dfrac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{1/(1-\hat{\rho}_i)} & \\ \quad -\left(1-\frac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{1/(1-\hat{\rho}_i)}, & y \in [1, \hat{\rho}_i/(1-\hat{\rho}_i)], \\ \left(1-\dfrac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{1/(1-\hat{\rho}_i)}, & y \in (\hat{\rho}_i/(1-\hat{\rho}_i), \hat{\rho}_i/(1-\hat{\rho}_i)+1] \end{cases}$$

and $\tilde{\boldsymbol{I}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ_i. The parameters α and μ_i can be found in (3.3).

Proof. Let $f_{U_{i,x}}(\cdot)$ and $g_{U_{i,x}}(\cdot)$ be the densities of $U_{i,x}^f$ and $U_{i,x}^g$, respectively. First consider $f_{U_{i,x}}(y) = 1/\omega(x)$ for $y \in [0, \omega(x)]$; thus, $a(x) = 0$ and $b(x) = \omega(x)$. Here, x is the service requirement, a realization of an exponential distribution, so $x \in [0,\infty)$. Since $\omega(0) = 0$ and $\omega(\infty) = \hat{\rho}_i/(1-\hat{\rho}_i)$, we only have to find the final term in (6.1) and consider the interval $[0, \hat{\rho}_i/(1-\hat{\rho}_i)]$. For a fixed y, the inverse function of ω is $\omega^{-1}(y) = \ln(1-y(1-\hat{\rho}_i)/\hat{\rho}_i)/(-b_i(1-\hat{\rho}_i))$. By Lemma 6.1, we have

$$\begin{aligned} f_{\tilde{U}_i}(y) &= \int_{x=\omega^{-1}(y)}^{\infty} f_{B_i}(x) f_{U_{i,x}}(y)\,\mathrm{d}x \\ &= \int_{x=\ln(1-y(1-\hat{\rho}_i)/\hat{\rho}_i)/-b_i(1-\hat{\rho}_i)}^{\infty} b_i \mathrm{e}^{-b_i x} \frac{1-\hat{\rho}_i}{\hat{\rho}_i} (1-\mathrm{e}^{-b_i x(1-\hat{\rho}_i)})^{-1}\,\mathrm{d}x \end{aligned}$$

$$= \int_{t=1-y(1-\hat{\rho}_i)/\hat{\rho}_i}^{0} b_i \frac{1-\hat{\rho}_i}{\hat{\rho}_i}(1-t)^{-1} \frac{1}{-b_i(1-\hat{\rho}_i)} t^{\hat{\rho}_i/(1-\hat{\rho}_i)}\,\mathrm{d}t$$
$$= \int_{t=0}^{1-y(1-\hat{\rho}_i)/\hat{\rho}_i} \frac{1}{\hat{\rho}_i}(1-t)^{-1} t^{\hat{\rho}_i/(1-\hat{\rho}_i)}\,\mathrm{d}t$$
$$= \frac{1}{\hat{\rho}_i}\mathrm{beta}_{1-y(1-\hat{\rho}_i)/\hat{\rho}_i}\left(1+\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0\right).$$

The third equality is obtained by taking $t = \mathrm{e}^{-b_i x(1-\hat{\rho}_i)}$. This leads to an incomplete beta function.

Now we turn to the second term involving $U_{i,x}^g$. Note that $g_{U_{i,x}}(y) = 1$ for $y \in [\omega(x), \omega(x)+1]$. To apply Lemma 6.1, observe that for $\hat{\rho}_i/(1-\hat{\rho}_i) \le 1$ it holds that $a(x_{\max}) \le b(x_{\min})$. First assume that $\hat{\rho}_i/(1-\hat{\rho}_i) \le 1$, implying $\hat{\rho}_i < \frac{1}{2}$. For a fixed $y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i))$, x needs to be smaller than $\omega^{-1}(y)$. If $y \in [\hat{\rho}_i/(1-\hat{\rho}_i), 1]$ it lies between the boundaries of the uniform distribution for all x, and if $y \in (1, \hat{\rho}_i/(1-\hat{\rho}_i)+1]$, then x needs to be larger than $\omega^{-1}(y)$. Then, for the PDF of $\tilde{U}_i^g$,

$$g_{\tilde{U}_i}(y) = \begin{cases} F_{B_i}(\omega^{-1}(y)), & y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i)), \\ 1, & y \in [\hat{\rho}_i/(1-\hat{\rho}_i), 1], \\ 1 - F_{B_i}(\omega^{-1}(y-1)), & y \in (1, \hat{\rho}_i/(1-\hat{\rho}_i)+1]. \end{cases}$$

Substituting $F_{B_i}(x) = 1 - \mathrm{e}^{-b_i x}$ and the inverse of $\omega(\cdot)$ into the above gives (6.2). The $\hat{\rho}_i > \frac{1}{2}$ case implies that $a(x_{\max}) > b(x_{\min})$ and is similar, completing the proof.

Remark 6.1. (*PS and ROS.*) For regular GI/M/1 queues, the relation between PS and ROS has been characterized by Borst *et al.* [9]. It is easily seen that the sample path relations [9, Equation (3)] also hold for the polling models under consideration. More specifically, consider a tagged customer T arriving at Q_i when the server visits Q_i. Then, the sojourn-time distribution of T for PS, given n_i customers at Q_i upon arrival, is identical to the waiting-time distribution of T for ROS, given n_i waiting customers at Q_i upon arrival in addition to the one in service. Under HT scalings, the differences between waiting and sojourn times vanish, explaining the equivalence between Theorems 6.2 and 5.1 (see Remark 5.1).

7. *n*-class priority queues

In this section we look at n-class priority queues. Each customer is assigned to a priority index k, $1 \le k \le n$, where customers with a low priority index are served before customers with higher priority indices. Within each class the service order is FCFS. In Section7.1 the focus lies on the nonpreemptive n-class priority regime. We will later use this discipline to find the waiting-time distribution in the SJF case. In [16], Kella and Yechiali studied the M/G/1 queue with single and multiple server vacations under both the preemptive and nonpreemptive priority regimes. The M/G/1 queue with multiple vacations is similar to a polling model, since we express the waiting times in cycle times and we can replace vacations by intervisit times. This relation has also been used in [6] to analyze multi-class polling models. We study the preemptive n-class priority regime in Section 7.2.

7.1. Nonpreemptive *n*-class priority queues

For the nonpreemptive n-class priority regime, we introduce notation and terminology based on [16], as this turns out to be useful and provide intuition for this and the next section.

Let $\lambda_{i,k}$ be the arrival rate of class-k customers and $B_{i,k}$ be the service duration of class-k customers. Class-a customers are the customers with priority index lower than k, i.e. they are served before class-k customers. They have arrival rate $\lambda_{i,a} = \sum_{j=1}^{k-1}\lambda_{i,j}$ and service duration $B_{i,a}$. Class-b customers are customers with priority index higher than k, their arrival rate is $\lambda_{i,b} = \sum_{j=k+1}^{n}\lambda_{i,j}$ and their service duration is $B_{i,b}$. We have $\rho_{i,a} = \lambda_{i,a}\mathbb{E}[B_{i,a}]$ and $\rho_{i,b} = \lambda_{i,b}\mathbb{E}[B_{i,b}]$. Let $\xi_{i,a}$ denote the length of time from a moment a class-a customer enters service and no other class-a customers are present, until the first moment when there are no class-a customers in the queue. Clearly, $\xi_{i,a}$ is the duration of a busy period in a standard M/G/1 queue with arrival rate $\lambda_{i,a}$ and service times $B_{i,a}$. From [16], we have the following LST for the waiting-time distribution $W_{i,k}$ of a class-k customer in Q_i. For $\mathrm{Re}(s) > 0, k = 1, \ldots, n$,

$$W_{i,k}^*(s) = \frac{(1-\rho_i)(1 - I_i^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B_{i,k}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) - \lambda_{i,k} + s)} + \frac{\rho_{i,b}(1 - B_{i,b}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)))}{\mathbb{E}[B_{i,b}](\lambda_{i,k}B_{i,k}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) - \lambda_{i,k} + s)}, \qquad i \in NPRIOR. \tag{7.1}$$

The first term in (7.1) corresponds to the waiting time of class-k customers in Q_i that arrive during the time from the start of the intervisit time until the moment a class-b customer at Q_i is taken into service. The second term corresponds to the waiting time of class-k customers that arrive during the time from the moment the first class-b customer is taken into service until the end of the cycle. Note that this expression was also derived in [6].

The following theorem gives the HT limit of the distribution of $W_{i,k}$.

Theorem 7.1. *For $\rho \uparrow 1, k = 1, \ldots, n$,*

$$\tilde{W}_{i,k} \xrightarrow{\mathrm{D}} \begin{cases} 0 & \text{w.p. } \hat{\rho}_{i,b}/(1 - \hat{\rho}_i + \hat{\rho}_{i,b}), \\ U_i\tilde{\boldsymbol{I}}_i & \text{w.p. } (1 - \hat{\rho}_i)/(1 - \hat{\rho}_i + \hat{\rho}_{i,b}), \end{cases} \qquad i \in NPRIOR,$$

where U_i is a uniformly distributed RV that lies between 0 and $1/(1 - \hat{\rho}_{i,a})$ and $\tilde{\boldsymbol{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i, where α and μ_i are given in (3.3).

Proof. The scaled waiting-time distribution is obtained by combining (7.1) with Proposition 3.2 and using l'Hôpital's rule (see [29]). Specifically, for $\mathrm{Re}(s) > 0, k = 1, \ldots, n, i \in NPRIOR$,

$$\begin{aligned} \tilde{W}_{i,k}^*(s) &= \lim_{\rho\uparrow 1} W_{i,k}^*(s(1-\rho)) \\ &= \frac{1 - \hat{\rho}_i}{1 - \hat{\rho}_i + \hat{\rho}_{i,b}} \frac{1}{\mathbb{E}[S]s(1 - \hat{\rho}_i)/(1 - \hat{\rho}_{i,a})} \left\{1 - \left(\frac{\mu_i}{\mu_i + s/(1 - \hat{\rho}_{i,a})}\right)^{\alpha}\right\} \\ &\quad + \frac{\hat{\rho}_{i,b}}{1 - \hat{\rho}_i + \hat{\rho}_{i,b}}. \end{aligned} \tag{7.2}$$

Recognizing this as the LST of a RV that is equal to 0 with probability $\hat{\rho}_{i,b}/(1 - \hat{\rho}_i + \hat{\rho}_{i,b})$ and a uniform multiplied by a gamma distribution with probability $(1 - \hat{\rho}_i)/(1 - \hat{\rho}_i + \hat{\rho}_{i,b})$ completes the proof.

7.2. Preemptive n-class priority queues

Similar to the previous section, the results of [16] also allow the derivation of the LST of the time until service in a polling system where different priority classes are served with preemptive priority. Let $W_i^{(q)}$ denote the time until a customer first receives service, or the waiting time in queue. We observe that this is not equal to the waiting time as defined in this paper (i.e. sojourn time minus service time) due to service preemptions. For class k, the LST of the time from the start until the end of service $R_{i,k}$, often referred to as the *residence time*, is

$$R_{i,k}^*=B_{i,k}^*(s+\lambda_{i,a}-\lambda_{i,a}\xi_{i,a}^*(s)). \tag{7.3}$$

For a class-k customer in Q_i the LST of the waiting time in queue , for $\mathrm{Re}(s)>0$, $k=1,\ldots,n$, is

$$W_{i,k}^{(q),*}(s)=\frac{(1-\rho_i)(1-I_i^*(s+\lambda_{i,a}-\lambda_{i,a}\xi_{i,a}^*(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B_{i,k}^*(s+\lambda_{i,a}-\lambda_{i,a}\xi_{i,a}^*(s))-\lambda_{i,k}+s)} + \frac{\rho_{i,b}(\lambda_{i,a}(1-\xi_{i,a}^*(s))+s)}{\lambda_{i,k}B_{i,k}^*(s+\lambda_{i,a}-\lambda_{i,a}\xi_{i,a}^*(s))-\lambda_{i,k}+s}, \qquad i\in NPRIOR\text{-}PR. \tag{7.4}$$

For n-class priority queues, the waiting-time distribution in heavy traffic is equal to the case of nonpreemptive priority queues. For the scaled waiting time in queue $W_{i,k}^{(q)}$ of a class-k customer in Q_i with preemptive priority service using (7.4), for $\mathrm{Re}(s)>0, i\in NPRIOR\text{-}PR, k=1\ldots,n$, we obtain

$$\tilde{W}_{i,k}^{(q),*}(s)=\frac{(1-\hat{\rho}_i)(1-(\mu_i/(\mu_i+s(1+\hat{\lambda}_{i,a}\mathbb{E}[\xi_{i,a}])))^{\alpha})}{\mathbb{E}[S](1-\hat{\rho}_i)s(1-\hat{\rho}_{i,k}(1+\hat{\lambda}_{i,a}\mathbb{E}[\xi_{i,a}]))} + \frac{\hat{\rho}_{i,b}(1+\hat{\lambda}_{i,a}\mathbb{E}[\xi_{i,a}])}{1-\hat{\rho}_{i,k}(1+\hat{\lambda}_{i,a}\mathbb{E}[\xi_{i,a}])},$$

which is equal to (7.2) for the nonpreemptive case. As before, α and μ_i are given in (3.3). From (7.3) it follows directly that the residence time can be neglected in heavy traffic.

8. Shortest-job-first and SRPT

The SJF service discipline can be thought of as a nonpreemptive priority queue with different priority classes. It may be interpreted as the continuous equivalent to having an infinite number of priority classes, where the priority classes correspond to job sizes. Alternatively, in Schrage and Miller [20], for the waiting time conditional on the service requirement x, a 3-class priority queue is used where the second class consists of customers of size x. From the HT limit derived in the previous section, we can immediately derive the HT limit of the waiting-time distribution for SJF. The conditional and unconditional scaled waiting-time distributions are given in Sections 8.1 and 8.2, respectively. The SRPT and preemptive SJF are discussed in Section 8.3.

8.1. Conditional waiting-time distribution in heavy traffic

To go from (7.2) to SJF we let the service time of the customer determine its priority. Note that we can apply the model of Section 7.1 if the distribution is discrete. In this section we assume that the service-time distribution has a density. First we derive the LST of the waiting time conditional on x, the service duration required by a tagged customer.

Define $\rho_i(x) = \lambda_i \mathbb{E}[B_i 1_{\{B_i<x\}}]$, which is the continuous equivalent of $\rho_{i,a}$. Because the service-time distribution is continuous, we have $\rho_i - \rho_{i,b} = \rho_{i,a}$. We can now write the conditional LST using (7.2). For $\mathrm{Re}(s) > 0$, $x > 0$,

$$\tilde{W}_i^*(s \mid x) = \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} \frac{1}{\mathbb{E}[S]s(1-\hat{\rho}_i)/(1-\hat{\rho}_i(x))} \left\{1 - \left(\frac{\mu_i}{\mu_i + s/(1-\hat{\rho}_i(x))}\right)^{\alpha}\right\} + \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)}, \qquad i \in SJF. \tag{8.1}$$

This result gives rise to the following theorem.

Theorem 8.1. *For $\rho \uparrow 1$,*

$$\tilde{W}_{i,x} \xrightarrow{\mathrm{D}} \begin{cases} 0 & \text{w.p. } (\hat{\rho}_i - \hat{\rho}_i(x))/(1-\hat{\rho}_i(x)), \\ U_{i,x}\tilde{\boldsymbol{I}}_i & \text{w.p. } (1-\hat{\rho}_i)/(1-\hat{\rho}_i(x)), \end{cases} \qquad i \in SJF. \tag{8.2}$$

Then $U_{i,x}$ is a RV with a uniform distribution on $[0, 1/(1-\hat{\rho}_i(x))]$ and $\tilde{\boldsymbol{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i as given in (3.3).

Proof. The results follow directly from (8.1).

8.2. Unconditional waiting-time distribution in heavy traffic

For the unconditional waiting-time distribution in heavy traffic, we have the following theorem. Let $\hat{\rho}_i^{-1}(y)$ denote the inverse function of $\hat{\rho}_i(x)$.

Theorem 8.2. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \xrightarrow{\mathrm{D}} \tilde{U}_i \tilde{\boldsymbol{I}}_i, \qquad i \in SJF,$$

where $\tilde{U}_i$ has PDF

$$f_{\tilde{U}_i}(y) = \begin{cases} 1-\hat{\rho}_i, & y \in [0,1], \\ (1-\hat{\rho}_i)\left(1 - F_{B_I}\left(\hat{\rho}_i^{-1}\left(\dfrac{y-1}{y}\right)\right)\right), & y \in (1, 1/(1-\hat{\rho}_i)] \end{cases} \tag{8.3}$$

with a point mass at 0 of

$$\int_0^{\infty} \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} f_{B_i}(x)\,\mathrm{d}x, \tag{8.4}$$

and $\tilde{\boldsymbol{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i as given in (3.3).

Proof. Note that the conditional waiting-time distribution in (8.2) can be written as a gamma distribution multiplied by a uniform distribution with a point mass at 0; we refer to the latter as 'uniform' distribution. To find the unconditional distribution of the waiting time, we need to find the unconditional 'uniform' distribution $\tilde{U}_i$ using Lemma 6.1. Note that, for $y \in [0, 1/(1-\hat{\rho}_i(x))]$, the CDF of this conditional 'uniform' distribution is

$$F_{U_{i,x}}(y) = \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} + \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} y(1-\hat{\rho}_i(x)).$$

The PDF of $U_{i,x}$ is given by

$$f_{U_{i,x}}(y) = 1-\hat{\rho}_i \quad \text{for } y \in \left[0, \frac{1}{1-\hat{\rho}_i(x)}\right];$$

thus, we have $a(x) = 0$ and $b(x) = 1/(1-\hat{\rho}_i(x))$. Recall that $\hat{\rho}_i(x) = \hat{\lambda}_i \mathbb{E}[B_i 1_{\{B_i<x\}}]$ and note that $\hat{\rho}_i(x_{\min}) = 0$ and $\hat{\rho}_i(x_{\max}) = \hat{\rho}_i$; $b(x)$ thus increases from 1 to $1/(1-\hat{\rho}_i)$. If $y \le 1$, we have

$$f_{\tilde{U}_i}(y) = \int_{x=0}^{\infty} f_{B_i}(x) * f_{U_{i,x}}(y)\,\mathrm{d}x = 1-\hat{\rho}_i, \qquad y \in [0,1].$$

When $y > 1$, $U_{i,x}$ only has probability mass for $x > \hat{\rho}_i^{-1}((y-1)/y)$. We obtain

$$\begin{aligned} f_{\tilde{U}_i}(y) &= \int_{x=\hat{\rho}_i^{-1}((y-1)/y)}^{\infty} f_{B_i}(x) * f_{U_{i,x}}(y)\,\mathrm{d}x \\ &= (1-\hat{\rho}_i)\left(1 - F_{B_i}\left(\hat{\rho}_i^{-1}\left(\frac{y-1}{y}\right)\right)\right), \qquad y \in \left(1, \frac{1}{1-\hat{\rho}_i}\right]. \end{aligned}$$

Combining the results above we see that $\tilde{U}_i$ has probability mass (8.4) in 0, and density (8.3). This completes the proof.

8.3. The SRPT and preemptive SJF

In this section we consider preemptive size-based scheduling policies. The most common is SRPT, where the customer with the smallest *remaining* service time is preemptively taken into service. A less well-known policy is preemptive SJF, where the customer is preemptively taken into service with the smallest *original* service time. The latter policy also has some desirable properties; see, e.g. [3] and [14]. Similar to SJF, the waiting-time distribution for preemptive SJF follows directly from the preemptive n-class priority queue of Section 7.2.

The analysis of SRPT does not follow directly from the results of [16]. Below, we use their framework to derive the LST of the waiting time in queue $W_{i,x}^{(q)}$ for a customer with service time x. We utilize the notation introduced in Section 7 and adopt the terminology of [16]. In particular, letting class a represent customers with service times smaller than x, $\xi_{i,a}^*(s)$ is defined by

$$\xi_{i,a}^*(s) = \frac{1}{F_{B_i}(x)} \int_0^x \exp(-t(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)))\, f_{B_i}(t)\,\mathrm{d}t$$

with $\lambda_{i,a} = \lambda_i F_{B_i}(x)$, i.e. $\xi_{i,a}^*(s)$ is a type-a busy period. Similarly, let class b represent customers with service times larger than x and $\lambda_{i,b} = \lambda_i(1-F_{B_i}(x))$.

Proposition 8.1. *For $\rho < 1$, $i \in SRPT$, $\mathrm{Re}(s) > 0$,*

$$W_i^{(q),*}(s) = \frac{1-\rho_i}{s\mathbb{E}[I_i]}(1 - I_i^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s))) + \frac{\rho_i - \rho_i(x) - \lambda_{i,b}x}{s}(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) + \frac{\lambda_{i,b}}{s}(1 - \exp(-x(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)))).$$

Proof. We start with the multi-class case, where class k is the class under consideration having service times in $(x-\varepsilon, x]$, for $\varepsilon > 0$ small, and classes a and b have priority indexes lower and higher than k, respectively. That is, the service times of class a are smaller than $x-\varepsilon$ and the service times of class b are larger than x. Applying the idea of [20], customers of size larger than x only affect class k as soon as their remaining service times become x. Specifically, class b initiates a delay cycle, as defined in [16], when their remaining service time is x. In the terminology of Kella and Yechiali, we thus have $T_{i,a,k}$ cycles for $T_i = I_i$, $B_{i,a}$, $B_{i,k}$, but now also for $T = x$. Since the LST of the waiting time given the cycle during which the customer arrives is known, it remains to specify the probabilities that the system is in a specific delay cycle. In line with [16, p. 28], we have the cycle probabilities

$$\Pi_{i,0} := \mathbb{P}(\text{no delay}) = \rho_{i,b} - \lambda_{i,b}x = \rho_i - \rho_{i,a} - \rho_{i,k} - \lambda_{i,b}x,$$

$$\mathbb{P}(B_{i,a}\text{ cycle}) = \frac{\Pi_{i,0}\rho_{i,a}}{1-\rho_{i,a}-\rho_{i,k}}, \qquad \mathbb{P}(B_{i,k}\text{ cycle}) = \frac{\Pi_{i,0}\rho_{i,k}}{1-\rho_{i,a}-\rho_{i,k}},$$

$$\mathbb{P}(I_i\text{ cycle}) = \frac{1-\rho_i}{1-\rho_{i,a}-\rho_{i,k}}, \qquad \mathbb{P}(x\text{ cycle}) = \frac{\lambda_{i,b}x}{1-\rho_{i,a}-\rho_{i,k}}.$$

Using the probabilities above in [16, Equations (7a) and (8)], we obtain, for $\mathrm{Re}(s) > 0$,

$$W_{i,k}^{(q),*}(s) = \frac{(1-\rho_i)(1 - I_i^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B_{i,k}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) - \lambda_{i,k} + s)} + \frac{\Pi_{i,0}(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) + \lambda_{i,b}(1 - \exp(-x(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s))))}{\lambda_{i,k}B_{i,k}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) - \lambda_{i,k} + s}.$$

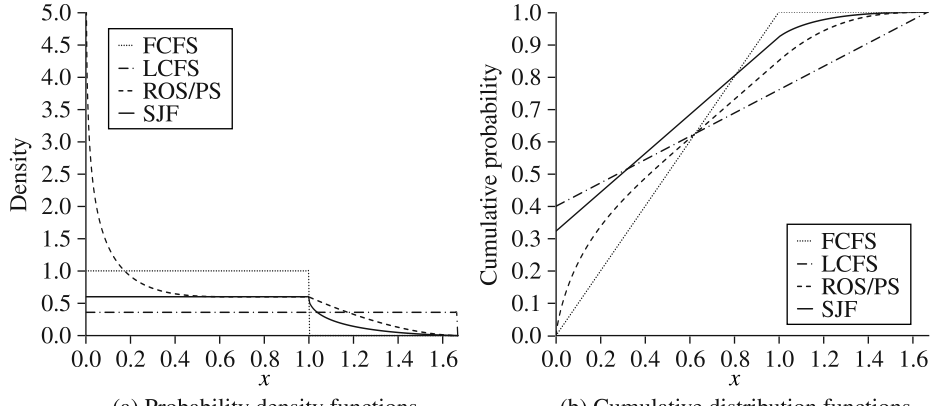
Letting $\varepsilon \downarrow 0$ and substituting $\Pi_{i,0}$, we obtain the result.

As in Section 7.2, $W_{i,x}^{(q)}$ is the waiting time in queue before the customer is first taken into service; this is not the same as the waiting time defined in this paper. We note that the residence time is identical to the residence time in a regular SRPT queue; see [20].

For LCFS and multi-class priority queues, the HT limits for the nonpreemptive and preemptive policies are identical. The same holds for SJF, preemptive SJF, and SRPT as represented by the following theorem.

Theorem 8.3. *For $\rho \uparrow 1$, the scaled waiting times $\tilde{W}_i$ follow the same probability distribution for SJF, preemptive SJF, and SRPT.*

Proof. Consider the conditional scaled waiting time $\tilde{W}_{i,x}(s)$. For preemptive SJF it can be directly observed from Section 7.2 that the HT limit is identical to the one for SJF. Using Proposition 8.1, it follows that $\lim_{\rho\uparrow 1} W_{i,x}^{(q),*}(s(1-\rho))$ equals the right-hand side of (8.1). Using (7.3) as an upper bound for the residence time, it is evident that the additional delay during the service does not contribute to the HT limit.



(a) Probability density functions (b) Cumulative distribution functions

FIGURE 3: Shapes of the service-order specific distributions.

9. Illustration and numerical results

In this section we illustrate the results and indicate how to use them for numerical approximations. We refer to [29] for additional results. Due to exhaustive service taking place at queue i it holds that $C_i^*(s) = I_i^*(s + \lambda_i(1 - \xi_i^*(s)))$, cf. (3.2), we can rewrite the second (gamma) distribution in terms of the scaled length-biased intervisit time distribution for all scheduling policies; thus, obtaining $\tilde{W}_i = \Theta_i \tilde{\boldsymbol{I}}_i$. In Figure 3 we plot the PDFs $f_{\Theta_i}(x)$ of Θ_i (Figure 3(a)) and also the CDFs $F_{\Theta_i}(x)$ (Figure 3(b)). We choose $\hat{\rho}_i = 0.4$. For FCFS, LCFS, ROS, and NPRIOR, the HT limit only depends on the service time distribution through its first moment. This is not the case for PS, SJF, and SRPT. In the figures we took exponential service times for PS and SJF. Figure 3(a) nicely shows how Θ_i behaves; for LCFS and FCFS it is like a uniform distribution, for SJF it is a type of generalized trapezoidal distribution, whereas it slightly deviates from this for ROS and PS. The atoms in 0 can be observed from Figure 3(b). In addition, these CDFs allow us to see the impact of the scheduling policy. For instance, SJF is here superior to ROS and PS.

We use the HT limits as the basis for approximations for the waiting-time distributions for stable systems, i.e. with $\rho < 1$. To this end, the asymptotic results suggest the following approximation for the waiting-time distribution for $\rho < 1$. For $i = 1, \ldots, N$,

$$\mathbb{P}(W_i \leq x) \approx \mathbb{P}(\Theta_i \Gamma_i \leq (1 - \rho)x).$$

The moments of the waiting-time distribution can be approximated using

$$\mathbb{E}[W_i^k] \approx \frac{\mathbb{E}[\Theta_i^k]\mathbb{E}[\Gamma_i^k]}{(1 - \rho)^k}.$$

We refer the reader to [25] for an elaboration on the accuracy of the approximation.

Acknowledgements

The authors wish to thank Jan-Pieter Dorsman and Erik Winands for interesting discussions and for reviewing preliminary versions of this paper.

References

[1] Ayesta, U., Boxma, O. J. and Verloop, I. M. (2012). Sojourn times in a processor sharing queue with multiple vacations. *Queueing Systems* **71,** 53–78.

[2] Baker, K. R. (1984). Sequencing rules and due-date assignments in a job shop. *Manag. Sci.* **30,** 1093–1104.

[3] Bansal, N. and Gamarnik, D. (2006). Handling load with less stress. *Queueing Systems* **54,** 45–54.

[4] Bekker, R. *et al.* (2015). The impact of scheduling policies on the waiting-time distributions in polling systems. *Queueing Systems* **79,** 145–172.

[5] Boon, M. A. A. (2011). Polling models: from theory to traffic intersections. Doctoral thesis, Eindhoven University of Technology.

[6] Boon, M. A. A., Adan, I. J. B. F. and Boxma, O. J. (2010). A polling model with multiple priority levels. *Performance Evaluation* **67,** 468–484.

[7] Boon, M. A. A., Adan, I. J. B. F. and Boxma, O. J. (2010). A two-queue polling model with two priority levels in the first queue. *Discrete Event Dynamic Systems* **20,** 511–536.

[8] Boon, M. A. A., van der Mei, R. D. and Winands, E. M. M. (2011). Applications of polling systems. *Surveys Operat. Res. Manag. Sci.* **16,** 67–82.

[9] Borst, S. C., Boxma, O. J., Morrison, J. A. and Núñez Queija, R. (2003). The equivalence between processor sharing and service in random order. *Operat. Res. Lett.* **31,** 254–262.

[10] Boxma, O., Bruin, J. and Fralix, B. (2009). Sojourn times in polling systems with various service disciplines. *Performance Evaluation* **66,** 621–639.

[11] Coffman, E. G., Jr, Puhalskii, A. A. and Reiman, M. I. (1995). Polling systems with zero switchover times: a heavy-traffic averaging principle. *Ann. Appl. Prob.* **5,** 681–719.

[12] Coffman, E. G., Jr, Puhalskii, A. A. and Reiman, M. I. (1998). Polling systems in heavy traffic: a Bessel process limit. *Math. Operat. Res.* **23,** 257–304.

[13] Ehrlich, W. K., Hariharan, R., Reeser, P. K. and van der Mei, R. D. (2001). Performance of web servers in a distributed computing environment. In *Teletraffic Engineering in the Internet Era*, Elsevier, Amsterdam, pp. 137–148.

[14] Harchol-Balter, M. (2009). Queueing Disciplines. In *Wiley Encyclopedia of Operations Research and Management Science*, John Wiley, New York, 13pp.

[15] Kawasaki, N. *et al.* (2000). Waiting time analysis of $M^X/G/1$ queues with/without vacations under random order of service discipline. *J. Operat. Res. Soc. Japan* **43,** 455–468.

[16] Kella, O. and Yechiali, U. (1988). Priorities in *M/G/*1 queue with server vacations. *Naval Res. Logistics* **35,** 23–34.

[17] Kingman, J. F. C. (1962). On queues in which customers are served in random order. *Proc. Cambridge Phil. Soc.* **58,** 79–91.

[18] Olsen, T. L. and van der Mei, R. D. (2003). Polling systems with periodic server routeing in heavy traffic: distribution of the delay. *J. Appl. Prob.* **40,** 305–326.

[19] Scholl, M. and Kleinrock, L. (1983). On the $M/G/1$ queue with rest periods and certain service-independent queueing disciplines. *Operat. Res.* **31,** 705–719.

[20] Schrage, L. E. and Miller, L. W. (1966). The queue $M/G/1$ with the shortest remaining processing time discipline. *Operat. Res.* **14,** 670–684.

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

[22] Takagi, H. and Kudoh, S. (1997). Symbolic higher-order moments of the waiting time in an $M/G/1$ queue with random order of service. *Commun. Statist. Stoch. Models* **13,** 167–179.

[23] Tijms, H. C. (2003). *A First Course in Stochastic Models*. John Wiley, Chichester.

[24] Van der Mei, R. D. (1999). Distribution of the delay in polling systems in heavy traffic. *Performance Evaluation* **38,** 133–148.

[25] Van der Mei, R. D. (2000). Polling systems with switch-over times under heavy load: moments of the delay. *Queueing Systems* **36,** 381–404.

[26] Van der Mei, R. D. (2007). Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Systems* **57,** 29–46.

[27] Van der Mei, R. D., Hariharan, R. and Reeser, P. K. (2001). Web server performance modeling. *Telecommun. Systems* **16,** 361–378.

[28] Van Dorp, J. R. and Kotz, S. (2003). Generalized trapezoidal distributions. *Metrika* **58,** 85–97.

[29] Vis, P., Bekker, R. and van der Mei, R. D. (2014). Heavy-traffic limits for polling models with exhaustive service and non-FCFS service order policies. *Tech. Rep.* ST-1401, CWI.

[30] VISHNEVSKII, V. M. AND SEMENOVA, O. V. (2006). Mathematical methods to study the polling systems. *Automation Remote Control* **67,** 173–220.
[31] WIERMAN, A., WINANDS, E. M. M. AND BOXMA, O. J. (2007). Scheduling in polling systems. *Performance Evaluation* **64,** 1009–1028.
[32] WINANDS, E. M. M., ADAN, I. J. B. F. AND VAN HOUTUM, G. J. (2006). Mean value analysis for polling systems. *Queueing Systems* **54,** 35–44.