Heavy-traffic limits for Polling Models with Exhaustive Service and non-FCFS Service Order Policies

P. Vis[a,b], R. Bekker[a] and R.D. van der Mei[b,a]

[a]VU University Amsterdam, Amsterdam, Netherlands
[b]Centre for Mathematics and Computer Science, Amsterdam, Netherlands

August 26, 2014

Abstract

We study cyclic polling models with exhaustive service at each queue under a variety of non-FCFS local service orders, namely Last-Come-First-Served (LCFS) with and without preemption, Random-Order-of-Service (ROS), Processor Sharing (PS), the multi-class priority scheduling with and without preemption, Shortest-Job-First (SJF) and the Shortest Remaining Processing Time (SRPT) 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 in 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. The accuracy of the approximations is evaluated by simulations.

Keywords: Polling system, service discipline, waiting-time distribution, heavy traffic
2010 Mathematics Subject Classification: 60K25, 90B22

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; we refer to [32] for an overview of the state-of-the-art. 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, and (3) the order in which customers at the same queue are served. The vast majority of papers in the literature is focused on the first two decisions. In the current 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- and switch-over time distributions with exhaustive service at all queues. We consider the following local service disciplines: LCFS (with and without preemption), ROS, local PS, the multi-class priority scheduling (with and without preemption), SJF and SRPT; see Table 2 for a brief description. In doing so, we derive new, exact expressions for the waiting-time distributions. We use these expressions to derive exact expressions for 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 FCFS: examples are Bluetooth and 802.11 protocols, scheduling policies at routers, and I/O subsystems in web servers [17; 31]. 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. [33] use the Mean Value Analysis (MVA, [34]) framework to derive the mean delay at each of the queues for various scheduling disciplines such as FCFS, LCFS, Foreground-Background (FB), PS, SJF and fixed priorities. Boxma et al. [10] obtain the waiting-time distribution in cyclic (globally-)gated polling systems for various local service orders. Bekker et al. [4] derive 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 the current 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 (see also Section 11 for more detailed discussion). Nonetheless, recent progress for exhaustive models has been made. Boon et al. [7] study 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 contains customers of two priority classes. In [6] these results are 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] derive 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 study Poisson-driven cyclic polling models with general service-time and switch-over time distributions, and with exhaustive service at all queues (see Section 11 for a relaxation of that assumption). For this model, we consider the following seven scheduling policies that determine the local order in which the customers at a given queue are served: FCFS (which is used as a benchmark), LCFS (with and without preemption), ROS, PS, the multi-class priority scheduling (with and without preemption), SJF and SRPT. For these models, we derive new, exact expressions for the waiting-time distributions in terms of the intervisit time distributions for stable systems. Subsequently, we use these expressions to derive the asymptotic waiting-time distributions for each of the local order policies under HT assumptions (i.e., when the load approaches 1). We show that in all 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 in 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.

The remainder of the paper is organized as follows. In Section 2, the model is described and the notation required is introduced. In Section 3, we present preliminary results, including the HT asymptotics for FCFS that serve as a benchmark. The waiting-time distributions and HT asymptotics for LCFS, ROS, PS, multi-class priority queues, and SJF and SRPT are derived in Sections 4–8, respectively. The results are summarized in Section 9. Furthermore, Section 10 proposes a simple approximation for the waiting-time distributions and present numerical results to evaluate the accuracy of the approximations. Finally, Section 11 contains a number of concluding remarks and addresses several topics for further research.

2 Notation and model description

In this section we introduce the notation and give a description of the model. To start, Table 1 gives useful notation with respect to a one-dimensional absolutely-continuous random variable X.

The model is as follows. We consider a system of $N \geq 2$ infinite-buffer queues, $Q_1, \dots, Q_N$, and a single server that visits and serves the queues in cyclic order. At each queue, the

$f_X(\cdot)$	Probability density function (pdf) of X
$F_X(\cdot)$	Cumulative distribution function (cdf) of X
$X^*(\cdot)$	Laplace-Stieltjes transform (LST) of X, i.e., $X^*(s) = \mathbb{E}[e^{-sX}]$
$\mathbb{E}[X]$	Expected value of X
$\mathbb{E}[X^k]$	kth moment of X
c_X^2	Squared coefficient of variation (SCV) of X
X^{res}	Residual length of X with $\mathbb{E}[X^{res}] = \frac{\mathbb{E}[X^2]}{2\,\mathbb{E}[X]}$ and LST $\mathbb{E}[e^{-sX^{res}}] = \frac{1-\mathbb{E}[e^{-sX}]}{s\,\mathbb{E}[X]}$
$\mathbf{X}$	Length-biased version of X with $f_{\mathbf{X}}(x) = \frac{x f_X(x)}{\mathbb{E}[X]}$

Table 1: Notation with respect to a random variable X.

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 $\{N_i(t),\ t \in \mathbb{R}\}$ 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 random variable 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 is equal to $\rho = \sum_{i=1}^N \rho_i$. A necessary and sufficient condition for stability of the system is $\rho < 1$. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is a random variable S_i. We let $S = \sum_{i=1}^N S_i$ denote the total switch-over time in a cycle. The random variable 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 2. We only consider work-conserving policies. For policy $P \in$ {FCFS, LCFS, LCFS-PR, ROS, PS, NPRIOR, NPRIOR-PR, SJF, 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 a FCFS basis.

In this paper we mainly focus on heavy-traffic (HT) limits, i.e., the limiting behavior as ρ approaches 1. The HT limits, denoted $\rho \uparrow 1$, taken in this paper are defined such that the arrival rates are increased, while keeping both the service-time and switch-over time distributions and the ratios between the arrival rates fixed. The notation $\to_d$ 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

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* non-preemptively 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.

Table 2: A brief description of the scheduling policies discussed in this paper.

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 random variables $\tilde{W}_i := (1-\rho)W_i$ 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 random variable with shape and rate parameters α and μ, respectively. Moreover, we denote by $U[a,b]$, with $a < b$, a random variable that is uniformly distributed over the interval $[a,b]$. For later reference, note that the LST of the random variable $U[a,b]\Gamma(\alpha+1,\mu)$, where $U[a,b]$ and $\Gamma(\alpha+1,\mu)$ are independent, is given by

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

In Sections 3 to 8 we derive expressions for the LSTs of the waiting-time distributions for the scheduling disciplines shown in Table 2.

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. 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 to [29] for rigorous proofs of these results.

3.1 Cycle and intervisit times

To start, let us consider the distribution of the cycle time C_i. Recall that C_i is defined as the time between two successive *departures* of the server from Q_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, stating that the (scaled) cycle times $\tilde{C}_i := (1-\rho)C_i$ converge to a gamma distribution with known parameters.

Property 1 (Convergence of cycle times). *For $i = 1, \dots, N$,*

$$\tilde{C}_i \to_d \tilde{\Gamma}, \tag{2}$$

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

$$\alpha := \frac{\mathbb{E}[S]\delta}{\sigma^2}, \qquad \mu := \frac{\delta}{\sigma^2}, \tag{3}$$

with

$$\sigma^2 := \frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}, \quad \textit{and} \quad \delta := \sum_{i=1}^{N} \hat{\rho}_i(1-\hat{\rho}_i). \tag{4}$$

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]):

$$\mathbb{E}[I_i] = (1-\rho_i)\,\mathbb{E}[C_i], \quad \text{and} \quad \mathbb{E}[e^{-(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))I_i}] = \mathbb{E}[e^{-sC_i}]. \tag{5}$$

Here ξ_i is the busy period of a regular M/G/1 queue with arrival rate λ_i and service time B_i. The (scaled) intervisit times $\tilde{I}_i := (1-\rho)I_i$ converge (in distribution) to a gamma distribution with known parameters as stated in the property below.

Property 2 (Convergence of intervisit times). *For $i = 1, \dots, N$, as $\rho \uparrow 1$,*

$$\tilde{I}_i \to_d \tilde{\Gamma}_i, \tag{6}$$

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

$$\alpha := \frac{\mathbb{E}[S]\delta}{\sigma^2}, \qquad \mu_i := \frac{\delta}{(1-\hat{\rho}_i)\sigma^2}, \tag{7}$$

where δ and σ^2 are given in Equation (4).

In the sequel, we repeatedly use Properties 1 and 2 to derive expressions for the asymptotic scaled waiting-time distributions associated with each of the service disciplines considered herein. For each policy we use a two-step approach:

(a) we derive an expression for the LST of the limiting distribution of the waiting times in terms of the cycle- and/or intervisit-time distribution;

(b) we combine this expression with Property 1 or 2 to obtain an expression for the LST of the waiting-time distribution in HT and interpret the resulting LST.

To conclude, we add intuition for the distribution using the Heavy Traffic Averaging Principle (HTAP).

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. [26]):

Property 3 (Waiting times in terms of intervisit times). *For $Re(s) > 0$ and $\rho < 1$,*

$$W_i^*(s) = \frac{(1-\rho_i)s}{s - \lambda_i(1 - B_i^*(s))} \frac{1 - I_i^*(s)}{s\,\mathbb{E}[I_i]} \quad (i \in FCFS). \tag{8}$$

Next, as step (b), combining Properties 2 and 3, the expression for $\mathbb{E}[C_i]$, and taking limits we obtain: For $Re(s) > 0$,

$$\tilde{W}_i^*(s) := \lim_{\rho \uparrow 1} W_i^*(s(1-\rho)) = \frac{1}{(1-\hat{\rho}_i)\,\mathbb{E}[S]s} \left\{ 1 - \left(\frac{\mu_i}{\mu_i + s} \right)^{\alpha} \right\} \quad (i \in FCFS). \tag{9}$$

Using (1), this leads to the following characterization of the limiting behavior of the scaled waiting-time distribution derived in [30]

Property 4 (Convergence of the waiting times). *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d U_i \tilde{\mathbf{I}}_i \quad (i \in FCFS), \tag{10}$$

where U_i is a uniformly distributed random variable on $[0,1]$, and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i, where α and μ_i are given in Equation (7).

Note that $\tilde{\mathbf{I}}_i$ is the *length-biased* counterpart of $\tilde{I}_i$, a gamma distributed random variable with parameters α and μ_i as in Equation (7). It is well known that if a gamma random variable has parameters α and μ_i, then its length-biased version has parameters $\alpha + 1$ and μ_i.

Remark 1 (Intuition by the Heavy Traffic Averaging Principle). Property 4 states that the limiting behavior of W_i is of the form $U_{FCFS}\Gamma$, where U_{FCFS} 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 ([12; 13; 21]). Loosely speaking, the HTAP principle 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.

Figure 1 gives 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 in 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 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. We refer to [5, p. 34-39] for an intuitive explanation based on this picture.

Here, we opt for a more direct analysis of the fluid model. Let the uniform random variable U on [0,1] denote the fraction of the cycle c that has elapsed at the arrival epoch of this

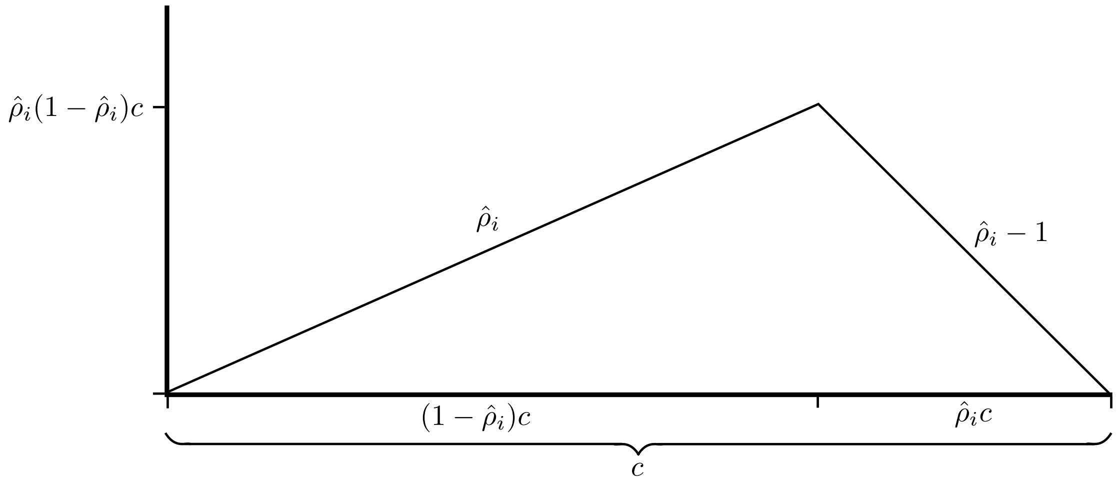


Figure 1: Fluid limits in heavy traffic; the amount of fluid in Q_i is plotted over the course of a cycle.

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 equals $\hat{\rho}_i(1-U)c$. Hence, the waiting time equals $(1-U)c - \hat{\rho}_i(1-U)c = (1-U)(1-\hat{\rho}_i)c$. 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 $\widetilde{W}_i$ is uniformly distributed on $[0,1]I_i$. This interpretation gives much insight in the heavy-traffic asymptotics.

4 Last-Come-First-Served

In this section we consider the LCFS service discipline. In Subsection 4.1 we derive the results for LCFS without preemption and in Subsection 4.2 we look at queues with LCFS preemptive resume (LCFS-PR) service. In both subsections, we first provide a derivation of the LST of W_i for all $\rho < 1$, giving insight in the terms contributing to the delay. Then we study the behavior of 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 $T_{i,0} \leq 0 \leq T_{i,1}$. 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 $\mathbb{R}^2$.

4.1 Non-Preemptive LCFS

Now 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 zero. We have to distinguish between the case where T arrives during an intervisit time, and the case where T arrives during a visit time.

Case I: the tagged customer arrives during an intervisit time
In this case, T has to wait for the server to start serving queue i; this is a residual inter-

visit time. In addition, T has to wait for all customers that arrived after him during the residual intervisit time and for the busy periods they generate. We have, for $i \in LCFS$,

$$W_i \ (\textit{given } T \text{ arrives during intervisit time}) = I_i^{res} + \sum_{T_{i,k} \in (0, I_i^{res})} \xi_{i,k}. \tag{11}$$

Conditioning on I_i^{res} and the number of arrivals during I_i^{res} (as in [10]), we have for $Re(s) > 0$,

$$\begin{aligned}
&\mathbb{E}[e^{-sW_i} | \text{arrival during intervisit time}] \\
&\quad = \int_{t=0}^{\infty} e^{-st} \sum_{n=0}^{\infty} e^{-\lambda_i t} \frac{(\lambda_i t)^n}{n!} \mathbb{E}[e^{-s\xi_i}]^n \, \mathrm{d}\,\mathbb{P}(I_i^{res} \le t) && (12) \\
&\quad = \int_{t=0}^{\infty} e^{-t(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))} \, \mathrm{d}\,\mathbb{P}(I_i^{res} \le t) \\
&\quad = \frac{1 - \mathbb{E}[e^{-(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}])I_i)}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[I_i]} \\
&\quad = \frac{1 - \mathbb{E}[e^{-sC_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1 - \rho_i)} \qquad (i \in LCFS), && (13)
\end{aligned}$$

where for the final step we use Equation (5).

Case II: the tagged customer arrives during a visit time
Note that T now arrives during the service of another customer. Hence, he has to wait for a residual service duration. In addition, he has to wait for the duration of the busy periods generated by the customers that arrived during the residual service time, as they are served before the tagged customer. Hence, we have for $i \in LCFS$,

$$W_i \ (\textit{given} \text{ arrival during visit time}) = B_i^{res} + \sum_{T_{i,k} \in (0, B_i^{res})} \xi_{i,k}. \tag{14}$$

Using the similarity between (11) and (14), we immediately see that, for $i \in LCFS$,

$$\begin{aligned}
\mathbb{E}[e^{-sW_i} | \text{arrival during visit time}] &= \frac{1 - \mathbb{E}[e^{-(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}])B_i)}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[B_i]} \\
&= \frac{1 - \mathbb{E}[e^{-s\xi_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[B_i]},
\end{aligned}$$

where the second equality follows from the well known functional equation satisfied by the LST of the busy period of an $M/G/1$ queue (see e.g., [28, p. 354]). Note that the probability that an arrival occurs during a visit time is equal to ρ_i. This leads to the following proposition.

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

$$\begin{aligned}
W_i^*(s) = \rho_i &\frac{1 - \mathbb{E}[e^{-s\xi_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[B_i]} \\
&+ (1 - \rho_i) \frac{1 - \mathbb{E}[e^{-sC_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1 - \rho_i)} \qquad (i \in LCFS). && (15)
\end{aligned}$$

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., [28, p. 357]). Also note that Equation (15) was found in [24], where intervisit periods are replaced with rest periods.

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

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

$$\tilde{W}_i \to_d \begin{cases} 0 & w.p. \quad \hat{\rho}_i \\ U_i \tilde{\mathbf{C}}_i & w.p. \quad 1-\hat{\rho}_i \end{cases} \qquad (i \in LCFS),$$

where U_i is a uniformly distributed random variable on the interval $[0,1]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (3).

Proof. Combining Proposition 1 with Property 1 gives the following expressions for the LST of the (scaled) waiting-time distribution. For $i \in LCFS$, $Re(s) > 0$,

$$\begin{aligned} \tilde{W}_i^*(s) &= \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) \\ &= \lim_{\rho\uparrow 1}\Bigg(\rho_i \frac{1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]}{(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))\,\mathbb{E}[B_i]} \\ &\qquad + (1-\rho_i)\frac{1-\mathbb{E}[e^{-s(1-\rho)C_i}]}{(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))\,\mathbb{E}[C](1-\rho_i)}\Bigg). \end{aligned} \tag{16}$$

Let us first consider the first term on the right-hand side of the final equation:

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

In the second equality, we use l'Hôpital's rule on both the numerator and the denominator, and the fact that the derivative of $\mathbb{E}[e^{-s(1-\rho)\xi_i}]$ at $s(1-\rho)=0$ is equal to $-\mathbb{E}[\xi_i]$. 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 (16):

$$\begin{aligned} &\lim_{\rho\uparrow 1}(1-\rho_i)\frac{1-\mathbb{E}[e^{-s(1-\rho)C_i}]}{\mathbb{E}[C](1-\rho_i)(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))} \\ &= \lim_{\rho\uparrow 1}(1-\rho_i)\frac{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}}{\mathbb{E}[S](1-\rho_i)(s+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}])/(1-\rho))} \\ &= (1-\hat{\rho}_i)\frac{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}}{\mathbb{E}[S](1-\hat{\rho}_i)s(1+\lambda_i\,\mathbb{E}[\xi_i])} \\ &= (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\}. \end{aligned} \tag{17}$$

Combining the above gives

$$\tilde{W}_i^*(s) = \hat{\rho}_i + (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\} \quad (i \in LCFS), \tag{18}$$

where α and μ are given in (3). Note that (18) corresponds to the LST of a random variable that is equal to 0 with probability $\hat{\rho}_i$ and to a uniform random variable on $[0,1]$ times a gamma distribution with probability $1-\hat{\rho}_i$. This completes the proof. □

Remark 2 (Intuition via Heavy Traffic Averaging Principle). The mixed distribution can be intuitively explained with the Heavy Traffic Averaging Principle (HTAP) and a fluid model, see Figure 1. With probability $\hat{\rho}_i$ a particle arrives during V_i. In this case the scaled waiting time is negligible in HT, since the residual service time and the busy periods generated by customers arriving during this time, do not scale with ρ. With probability $(1-\hat{\rho}_i)$ a particle arrives during I_i. Let the uniform random variable U_I denote the fraction of I_i that has elapsed at the arrival epoch of this particle. This arriving particle has to wait for the remaining intervisit time $(1-U_I)I_i$, in addition it has to wait for the busy periods generated by particles that arrived during that time for duration $\hat{\rho}_i(1-U_I)I_i/(1-\hat{\rho}_i)$, the amount of work built up during the remaining intervisit time divided by the rate at which the queue is emptied. Adding the two terms and noting that $(1-U_I)$ is in distribution equal to U_I we get for the scaled waiting time of a particle arriving during an intervisit time: $W_i^{(I)} = U_I I_i/(1-\hat{\rho}_i) = U_I c$. Now we can use the HTAP and the results from [29] to find the distribution of c and arrive at the result given in Theorem 1.

4.2 LCFS with Preemptive Resume

The analysis of LCFS-PR service is largely similar to the non-preemptive LCFS case. 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. This gives for Case I (see Section 4.1): For $i \in LCFS\text{-}PR$,

$$W_i \ (\textit{given } T \text{ arrives during intervisit time}) = \sum_{T_{i,k}\in(0,B_i)} \xi_{i,k} + I_i^{res} + \sum_{T_{i,k}\in(0,I_i^{res})} \xi_{i,k}. \tag{19}$$

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. We have in Case II: For $i \in LCFS\text{-}PR$,

$$W_i \ (\textit{given } T \text{ arrives during visit time}) = \sum_{T_{i,k}\in(0,B_i)} \xi_{i,k}. \tag{20}$$

Due to the preemptive nature of the discipline, the first term of (19) is equal to (20), the waiting time in Case II, so we calculate the LST of the waiting time of Case II first. Conditioning on the service time and the number of arrivals therein yields: For

$i \in I_{LCFS\text{-}PR}$,

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

The last two terms of (19) are equal to the waiting time of non-preemptive LCFS given in (11). We use the corresponding LST given in (13) to arrive at (21): For $i \in LCFS\text{-}PR$, $Re(s) > 0$,

$$\begin{aligned}&\mathbb{E}[e^{-sW_i}|T \text{ arrives during intervisit time}] =\\ &\qquad B_i^*(\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\frac{1-\mathbb{E}[e^{-sC_i}]}{(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1-\rho_i)}.\end{aligned} \tag{21}$$

Combining the two cases leads to the following expression for the LST of the waiting time at Q_i in terms of the cycle time.

Proposition 2. *For $\rho < 1$, $i \in LCFS\text{-}PR$, $Re(s) > 0$,*

$$\begin{aligned}W_i^*(s) &= B_i^*(\lambda_i(1-\mathbb{E}[e^{-s\xi_i}])) \qquad (22)\\ &\quad\times\left(\rho_i + (1-\rho_i)\frac{1-\mathbb{E}[e^{-sC_i}]}{(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1-\rho_i)}\right).\end{aligned}$$

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

Theorem 2. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d \begin{cases} 0 & w.p.\ \ \hat{\rho}_i \\ U_i\tilde{\mathbf{C}}_i & w.p.\ \ 1-\hat{\rho}_i \end{cases} \qquad (i \in LCFS\text{-}PR),$$

where U_i is a uniformly distributed random variable on the interval $[0,1]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (3).

Proof. Using Equation (17) and the fact that for $Re(s) > 0$ it holds that $\lim_{\rho\uparrow 1} B_i^*(\lambda_i(1-\mathbb{E}[e^{-s\xi_i}])) = 1$, we immediately see that the LST of $\tilde{W}_i$ in HT is given by

$$\begin{aligned}\tilde{W}_i^*(s) &:= \lim_{\rho\uparrow 1} W_i^*(s(1-\rho))\\ &= \hat{\rho}_i + (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\} \quad (i \in LCFS\text{-}PR),\end{aligned} \tag{23}$$

with α and μ given in (3). □

Note that the HT scaled waiting-time distribution (23) for $i \in LCFS\text{-}PR$ is equal to the HT scaled waiting-time distribution (18) for $i \in LCFS$. This holds because the busy periods generated by customers arriving during service of the tagged customer do not scale with ρ.

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 3. *For $\rho < 1$, $i \in ROS$, $Re(s) > 0$,*

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

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

$$K(x,s) := \exp\left(-\int\limits_{y=x}^{1} \frac{1}{y - B_i^*(s+\lambda_i-\lambda_i y)}\,\mathrm{d}y\right). \tag{24}$$

Proof. The derivation proceeds along the lines of Kingman [20]. Define the waiting time of a tagged 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 $u = I_i^{res}$, or during a visit period, yielding $u = B_i^{res}$.

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 $Q(z, X)$, with $X \in \{\mathbf{B}_i, \mathbf{I}_i\}$. From Kawasaki et al. [18] we have for an arrival during a visit period:

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

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

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

Kingman [20] (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 $\mathbf{B}_i$ is in service: For $Re(s) > 0$, $i \in ROS$,

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

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

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

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

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

Now, using the fact that $\mathbb{E}\left[e^{-\phi\mathbf{B}_i}/\mathbf{B}_i\right] = \frac{B_i^*[\phi]}{\mathbb{E}[B_i]}$ (see [20]), we have for $Re(s) > 0$, $i \in ROS$,

$$\mathbb{E}[\mathbb{E}[e^{-sw}|\mathbf{B}_i]] = \frac{1-\rho_i}{s\lambda_i\,\mathbb{E}[I_i]}\int\limits_{\xi_i^*(s)}^{1} \frac{1-I_i^*(\lambda_i-\lambda_i x)}{B_i^*(\lambda_i-\lambda_i x)-x}\,\frac{B_i^*(\lambda_i(1-x))-B_i^*(s+\lambda_i(1-x))}{\mathbb{E}[B_i]}\,\mathrm{d}K(x,s).$$

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

Next, we turn to the heavy-traffic limit. Before we state our result, we define Y as a random variable with pdf and cdf

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

The r.v. 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 is served before the arriving customer, see Remarks 4 and 5.

The next theorem gives the HT limit of the distribution of $\tilde W_i$ in terms of Y.

Theorem 3. *For $\rho \uparrow 1$,*

$$\tilde W_i \to_d \begin{cases} U_i^f\tilde{\mathbf{C}} & w.p.\ \hat\rho_i \\ U_i^g\tilde{\mathbf{C}} & w.p.\ 1-\hat\rho_i \end{cases} \qquad (i \in ROS),$$

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 3. Noting that $\frac{\mathrm{d}K(x,s)}{\mathrm{d}x} = \frac{K(x,s)}{x-B_i^*(s+\lambda_i(1-x))}$, we get

$$\begin{aligned}W_i^*(s) = \frac{1-\rho_i}{s\,\mathbb{E}[I_i]}\Bigg(&\int\limits_{x=\xi_i^*(s)}^{1} K(x,s)(1-I_i^*(\lambda_i-\lambda_i x)) \\ &\times\left(\frac{1}{B_i^*(\lambda_i(1-x))-x}+\frac{1}{x-B_i^*(s+\lambda_i(1-x))}\right)\mathrm{d}x \\ &+\int\limits_{x=\xi_i^*(s)}^{1} K(x,s)\left(I_i^*(\lambda_i(1-x))-I_i^*(s+\lambda_i(1-x))\right)\frac{1}{x-B_i^*(s+\lambda_i(1-x))}\,\mathrm{d}x\Bigg)\end{aligned}$$

In line with Takagi and Kudoh [27] we take $y = \frac{1-x}{1-\xi_i^*(s)}$; this gives $x = 1 - y(1-\xi_i^*(s))$ and $\mathrm{d}x = -(1-\xi_i^*(s))\,\mathrm{d}y$, yielding

$$\begin{aligned}W_i^*(s) = \frac{1-\rho_i}{s\,\mathbb{E}[I_i]}\Bigg(&\int\limits_{y=0}^{1} K\big(1-y(1-\xi_i^*(s)),s\big)\big(1-I_i^*(y\lambda_i(1-\xi_i^*(s)))\big)\\ &\times\left(\frac{1-\xi_i^*(s)}{B_i^*(y\lambda_i(1-\xi_i^*(s)))-1+y(1-\xi_i^*(s))}+\frac{1-\xi_i^*(s)}{1-y(1-\xi_i^*(s))-B_i^*(s+y\lambda_i(1-\xi_i^*(s)))}\right)\mathrm{d}y\\ &+\int\limits_{y=0}^{1} K\big(1-y(1-\xi_i^*(s)),s\big)\big(I_i^*(y\lambda_i(1-\xi_i^*(s)))-I_i^*(s+y\lambda_i(1-\xi_i^*(s)))\big)\\ &\times\left(\frac{1-\xi_i^*(s)}{1-y(1-\xi_i^*(s))-B_i^*(s+y\lambda_i(1-\xi_i^*(s)))}\right)\mathrm{d}y\Bigg).\end{aligned}$$

We now take heavy-traffic limits for the terms separately. We start with the most involved term, $K(x,s)$. Using the substitution $t = \frac{1-y}{1-x}$ in (24), we may write

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

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

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

In the second step we use the fact that $\mathbb{E}[\xi_i] = \frac{\mathbb{E}[B_i]}{1-\hat{\rho}_i}$. 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 get

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

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

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

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 $[0, y\hat{\rho}_i]$ times 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 $[y\hat{\rho}_i, 1-\hat{\rho}_i+y\hat{\rho}_i]$ times a gamma distribution with the same parameters. This completes the proof. □

Remark 3. The expressions for U_i^f and U_i^g in Theorem 3 can be rewritten more explicitly, similar to those in Theorem 5, see also Remark 8.

Remark 4 (HTAP). The HT limit states that conditional on $Y = y$, the scaled waiting-time distribution is a uniform times a gamma distribution with probability $\hat{\rho}_i$ and another uniform times a gamma distribution with probability $1 - \hat{\rho}_i$. Here, y is a tag representing the fraction of work from the work present and arriving until the server's departure from the queue that is served before the tagged customer in a fluid model. See Remark 5 below for a more intuitive derivation of the tag-distribution $F_Y(\cdot)$.
With probability $1 - \hat{\rho}_i$ a particle arrives during an intervisit time of length $c(1 - \hat{\rho}_i)$. If U_I is the fraction of the intervisit time that has elapsed at the arrival epoch of a tagged particle, it first has to wait $(1 - U_I)c(1 - \hat{\rho}_i)$ until Q_i is visited. The total work present upon arrival plus the amount of work arriving until the server's departure from Q_i equals the total workload arriving during a cycle and is $\hat{\rho}_i c$. Given the tag $Y = y$, the total scaled waiting time equals $((1 - U_I)(1 - \hat{\rho}_i) + y\hat{\rho}_i)c$, corresponding to a uniform distribution on $[y\hat{\rho}_i, 1 - \hat{\rho}_i + y\hat{\rho}_i]$. With probability $\hat{\rho}_i$ a particle arrives during a visit time of length $\hat{\rho}_i c$. If U_V is the fraction of the intervisit time that remains, the amount work present upon arrival in addition to the remaining amount of work arriving equals $U_V\hat{\rho}_i c$. Given a tag $Y = y$, the scaled waiting time is $yU_V\hat{\rho}_i c$, which is a uniform distribution on $[0, y\hat{\rho}_i]$ times c. Theorem 3 thus follows intuitively from HTAP.

Remark 5 (Intuition for tag-distribution Y). We provide an intuitive explanation for the distribution of Y using a fluid model for the number of customers or particles. Assume the tagged customer arrives during a visit time, say at time 0, finding x particles present. The queue length is decreasing at rate $1 - \hat{\rho}_i$, i.e. at time t the queue length $L_i(t) = x - (1 - \hat{\rho}_i)t$, until the queue is empty at time $x/(1 - \hat{\rho}_i)$. Observe that with $L_i(t)$ particles present, the probability for service selection is $1/L_i(t)$. Let $\bar{F}(t)$ be the probability that the tagged customer has not been taken into service at time t. Since there are continuously options for service selection in the fluid model, $\bar{F}(t)$ satisfies the following first-order differential equation (DE), for $0 < t < x/(1-\hat{\rho}_i)$,

$$-\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(t) = \bar{F}(t) \times \frac{1}{L_i(t)}.$$

Solving the above DE with boundary condition $\bar{F}(0) = 1$ and using the fluid version of $L_i(t)$, we have, for $0 < t < x/(1-\hat{\rho}_i)$,

$$\bar{F}(t) = \exp\left(\int \frac{1}{x-(1-\hat{\rho}_i)t}\,\mathrm{d}t\right) = \left(1 - t\frac{1-\hat{\rho}_i}{x}\right)^{\frac{1}{1-\hat{\rho}_i}}.$$

Finally, the queue being empty at time $x/(1-\hat{\rho}_i)$ implies that also $x/(1-\hat{\rho}_i)$ particles have been served since time 0. When at least a fraction y of those has been served before the tagged customer is taken into service, then we look for

$$\bar{F}\left(y \times \frac{x}{1-\hat{\rho}_i}\right) = (1-y)^{\frac{1}{1-\hat{\rho}_i}}.$$

This coincides with one minus the cdf of Y.

6 Processor sharing

In a processor sharing (PS) queue, all customers present at the queue that is 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 time, see also Section 11. We extend the work done in [1], where they derive the heavy-traffic limit of the LST of the scaled waiting time conditional on the service requirement. In Subsection 6.1, we give the conditional scaled waiting-time distribution. In Subsection 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 heavy-traffic limit of the conditional waiting time $W_i|x$:

Theorem 4. *For $\rho \uparrow 1$, $x \geq 0$,*

$$\tilde{W}_i|x \to_d \begin{cases} U^f_{i,x}\tilde{\mathbf{I}}_i & w.p.\ \hat{\rho}_i \\ U^g_{i,x}\tilde{\mathbf{I}}_i & w.p.\ 1-\hat{\rho}_i \end{cases} \qquad (i \in PS),$$

where $U^f_{i,x} = U[0,\omega(x)]$, $U^g_{i,x} = U[\omega(x),\omega(x)+1]$ and $\tilde{\mathbf{I}}_i \sim \Gamma(\alpha+1,\mu_i)$. The parameters α and μ_i can be found in Equation (7), and $\omega(x) = \frac{\hat{\rho}_i}{1-\hat{\rho}_i}(1-e^{-b_i x(1-\hat{\rho}_i)})$.

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

$$\begin{aligned}\tilde{W}^*_i(s|x) = {} & \frac{\hat{\rho}_i}{s\omega(x)\,\mathbb{E}[S](1-\hat{\rho}_i)}\left\{1 - \left(\frac{\mu_i}{\mu_i + s\omega(x)}\right)^{\alpha}\right\} \\ & + \frac{1-\hat{\rho}_i}{s\,\mathbb{E}[S](1-\hat{\rho}_i)}\left\{\left(\frac{\mu_i}{\mu_i+s\omega(x)}\right)^{\alpha} - \left(\frac{\mu_i}{\mu_i+s(\omega(x)+1)}\right)^{\alpha}\right\}. \qquad (25)\end{aligned}$$

From this LST we see that the distribution of the conditional waiting time is a uniform $[0, \omega(x)]$ times 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 $[\omega(x), \omega(x)+1]$ times a gamma distribution with parameters $\alpha+1$ and μ_i. This completes the proof. □

Remark 6 (HTAP). Theorem 4 states that the conditional waiting-time distribution is a uniform times a gamma distribution with probability $\hat{\rho}_i$ and another uniform times a gamma distribution with probability $1-\hat{\rho}_i$. This can be intuitively explained with a fluid model. In the fluid model $\omega(x)c(1-\hat{\rho}_i)$ is the scaled waiting time of a particle, with service requirement x, arriving at the start of a visit period. With probability $1-\hat{\rho}_i$ a particle arrives during an intervisit period of length $c(1-\hat{\rho}_i)$. If U_I is the fraction of the intervisit time that has elapsed at the arrival epoch of a tagged particle, then the scaled waiting time of this particle is the remaining intervisit time $(1-U_I)c(1-\hat{\rho}_i)$ plus $\omega(x)c(1-\hat{\rho}_i)$. Using the HTAP gives a uniform distribution on $[\omega(x), \omega(x)+1]$ times a gamma distribution with parameters $\alpha+1$ and μ_i. A particle arriving during a visit period has to wait an amount of time that is uniformly distributed between 0 (arrive at the end of the visit time) and $\omega(x)c(1-\hat{\rho}_i)$ (arrive at the start of the visit time). Using the HTAP now gives a uniform distribution on $[0, \omega(x)]$ times a gamma distribution with parameters $\alpha+1$ and μ_i.

Remark 7 (Intuition for $\omega(x)$). The sojourn time of a tagged customer with service time x from the start of the visit time $(\omega(x)I_i)$ can be intuitively explained with a fluid model. As long as the tagged customer is present, the amount of service received during $(0,t)$ is $B(t) = \int_0^t 1/L(u)\,\mathrm{d}u$ with $L(u)$ the number of customers at time u. During the visit time, we have in a fluid model $L(t) = L(0) - (1-\hat{\rho}_i)b_i t$. Hence,

$$B(t) = \int\limits_{u=0}^{t} \frac{1}{L(0) - (1-\hat{\rho}_i)b_i u}\,\mathrm{d}u = -\frac{1}{(1-\hat{\rho}_i)b_i}\left(\ln(L(0) - (1-\hat{\rho}_i)b_i t) - \ln L(0)\right).$$

To obtain the time until service completion, we solve $B(t) = x$ for t. Moreover, using that $L(0) = \hat{\lambda}_i c(1-\hat{\rho}_i)$ in the fluid model, yields

$$\omega(x) \times I_i = \frac{\hat{\lambda}_i}{(1-\hat{\rho}_i)b_i}\left(1 - e^{-x(1-\hat{\rho}_i)b_i}\right) \times c(1-\hat{\rho}_i).$$

The result follows from $\hat{\rho}_i = \hat{\lambda}_i / b_i$.

6.2 Unconditional waiting-time distribution in heavy traffic

In the previous section we derived the heavy-traffic 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 random variable, denoted $T|x$, with pdf $f_{T|x}(y)$, cdf $F_{T|x}(y)$, and $y \in [a(x), b(x)]$, with $a(x) < b(x) \quad \forall x$. We want to find the unconditional distribution $\tilde{T}$. Here, x is a realization of a random variable X with support $x \in [x_{min}, x_{max}]$. We have the following lemma.

Lemma 1. *Assume that the conditional random variable $T|x$ has density $f_{T|x}(y)$ and distribution function $F_{T|x}(y)$, with support $y \in [a(x), b(x)]$. Suppose $a(x)$ and $b(x)$ are both increasing in x and $a(x) < b(x) \quad \forall x$. Let $a^{-1}(\cdot)$ be the inverse of $a(\cdot)$ and $b^{-1}(\cdot)$*

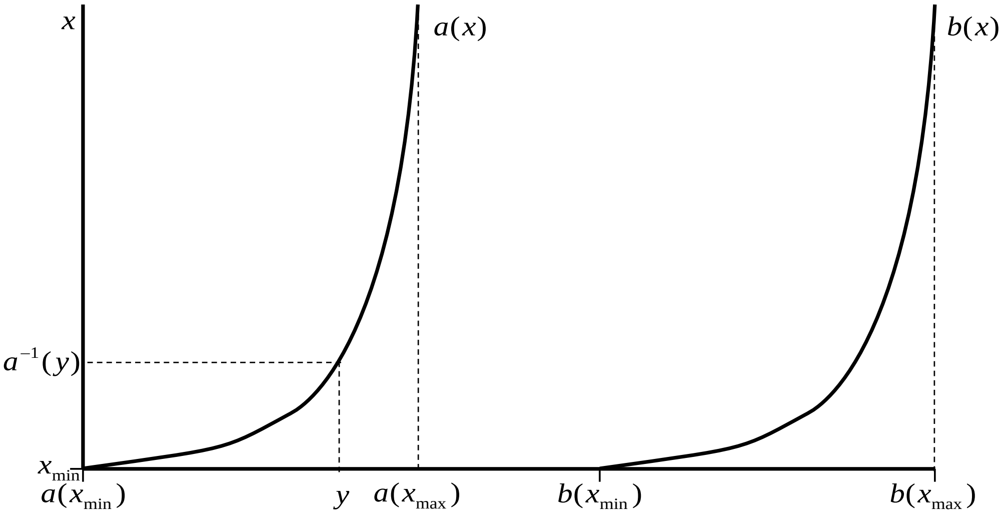


Figure 2: Boundaries of the conditional distribution.

be the inverse of $b(\cdot)$. *Then, the unconditional distribution of* $T|x$, *denoted by* $\tilde{T}$, *has probability density function, for* $a(x_{max}) \le b(x_{min})$,

$$f_{\tilde{T}}(y) = \begin{cases} \int_{x=x_{min}}^{a^{-1}(y)} f_{T|x}(y) f_X(x) \,\mathrm{d}x & y \in [a(x_{min}), a(x_{max})] \\ \int_{x=x_{min}}^{x_{max}} f_{T|x}(y) f_X(x) \,\mathrm{d}x & y \in [a(x_{max}), b(x_{min})] \\ \int_{x=b^{-1}(y)}^{x_{max}} f_{T|x}(y) f_X(x) \,\mathrm{d}x & y \in [b(x_{min}), b(x_{max})], \end{cases} \tag{26}$$

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

$$f_{\tilde{T}}(y) = \begin{cases} \int_{x=x_{min}}^{a^{-1}(y)} f_{T|x}(y) f_X(x) \,\mathrm{d}x & y \in [a(x_{min}), b(x_{min})] \\ \int_{x=b^{-1}(y)}^{a^{-1}(y)} f_{T|x}(y) f_X(x) \,\mathrm{d}x & y \in [b(x_{min}), a(x_{max})] \\ \int_{x=b^{-1}(y)}^{x_{max}} f_{T|x}(y) f_X(x) \,\mathrm{d}x & y \in [a(x_{max}), b(x_{max})]. \end{cases} \tag{27}$$

Proof. First consider the case that $a(x_{max}) \le b(x_{min})$. Figure 2 shows an example of the boundaries of the conditional distribution, by plotting $a(x)$ and $b(x)$ with x on the vertical axis. The possible values of $T|x$ 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 probability density function $f_{\tilde{T}}(y)$ is obtained from the parts where x is smaller than $a^{-1}(y)$. This gives

$$f_{\tilde{T}}(y) = \int\limits_{x=x_{min}}^{a^{-1}(y)} f_{T|x}(y) f_X(x) \,\mathrm{d}x. \tag{28}$$

If $y \in [a(x_{max}), b(x_{min})]$ then y is between the boundaries of the conditional distribution for every $x \in [x_{min}, x_{max}]$. Hence, we get

$$f_{\tilde{T}}(y) = \int\limits_{x=x_{min}}^{x_{max}} f_{T|x}(y) f_X(x) \,\mathrm{d}x. \tag{29}$$

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 $b^{-1}(y)$. This gives

$$f_{\tilde{T}}(y) = \int_{x=b^{-1}(y)}^{x_{max}} f_{T|x}(y) f_X(x) \,\mathrm{d}x. \tag{30}$$

The case $a(x_{max}) > b(x_{min})$ is similar. It may be checked $f_{\tilde{T}}(\cdot)$ is a density function. This completes the proof. □

Note that the distribution in Equation (26) 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 [14], we refer to (26) as a *generalized trapezoidal distribution.*

We now apply Lemma 1 to the case $i \in PS$, in which case we have two conditional distributions, $U^f_{i,x}$ and $U^g_{i,x}$. We need to find the unconditional versions of both uniform distributions.

Theorem 5. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d \begin{cases} \tilde{U}^f_i \tilde{\mathbf{I}}_i & \textit{w.p. } \hat{\rho}_i \\ \tilde{U}^g_i \tilde{\mathbf{I}}_i & \textit{w.p. } 1-\hat{\rho}_i \end{cases} \qquad (i \in PS),$$

where $\tilde{U}^f_i$ has a generalized trapezoidal distribution with pdf

$$f_{\tilde{U}_i}(y) = \frac{1}{\hat{\rho}_i} Beta_{1-y(1-\hat{\rho}_i)/\hat{\rho}_i}\left(1 + \frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0\right) \quad y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i)], \tag{31}$$

where $Beta_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$. $\tilde{U}^g_i$ has a generalized trapezoidal distribution with pdf, for $\hat{\rho}_i \le \frac{1}{2}$,

$$g_{\tilde{U}_i}(y) = \begin{cases} 1 - \left(1 - \frac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i)) \\ 1 & y \in [\hat{\rho}_i/(1-\hat{\rho}_i), 1] \\ \left(1 - \frac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in (1, \hat{\rho}_i/(1-\hat{\rho}_i)+1], \end{cases} \tag{32}$$

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

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

and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i. The parameters α and μ_i can be found in Equation (7)*.*

Proof. Let $f_{U_{i,x}}(\cdot)$ and $g_{U_{i,x}}(\cdot)$ be the densities of $U^f_{i,x}$ and $U^g_{i,x}$, respectively. First consider $f_{U_{i,x}}(y) = \frac{1}{\omega(x)}$ 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 of (26) 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 1, this gives

$$\begin{aligned}
f_{\tilde{U}_i}(y) &= \int\limits_{x=\omega^{-1}(y)}^{\infty} f_{B_i}(x) f_{U_{i,x}}(y)\,\mathrm{d}x \\
&= \int\limits_{x=\frac{\ln(1-y(1-\hat{\rho}_i)/\hat{\rho}_i)}{-b_i(1-\hat{\rho}_i)}}^{\infty} b_i e^{-b_i x} \frac{1-\hat{\rho}_i}{\hat{\rho}_i} \left(1-e^{-b_i x(1-\hat{\rho}_i)}\right)^{-1} \mathrm{d}x \\
&= \int\limits_{t=1-y(1-\hat{\rho}_i)/\hat{\rho}_i}^{0} b_i \frac{1-\hat{\rho}_i}{\hat{\rho}_i} (1-t)^{-1} \frac{1}{-b_i(1-\hat{\rho}_i)} t^{\frac{\hat{\rho}_i}{1-\hat{\rho}_i}}\,\mathrm{d}t \\
&= \int\limits_{t=0}^{1-y(1-\hat{\rho}_i)/\hat{\rho}_i} \frac{1}{\hat{\rho}_i} (1-t)^{-1} t^{\frac{\hat{\rho}_i}{1-\hat{\rho}_i}}\,\mathrm{d}t \\
&= \frac{1}{\hat{\rho}_i} \mathrm{Beta}_{1-y(1-\hat{\rho}_i)/\hat{\rho}_i} \left(1 + \frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0\right).
\end{aligned}$$

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

Now we turn to the second term involving $U^g_{i,x}$. Note that $g_{U_{i,x}}(y) = 1$ for $y \in [\omega(x), \omega(x)+1]$. To apply Lemma 1, observe that for $\hat{\rho}_i/(1-\hat{\rho}_i) \leq 1$ it holds that $a(x_{max}) \leq b(x_{min})$. First assume that $\hat{\rho}_i/(1-\hat{\rho}_i) \leq 1$, implying $\hat{\rho}_i < 1/2$. 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 $\omega^{-1}(y)$. This gives for the pdf of $\tilde{U}^g_i$

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

Substituting $F_{B_i}(x) = 1 - e^{-b_i x}$ and the inverse of $\omega(\cdot)$ gives Equation (32). The case $\hat{\rho}_i > 1/2$ implies $a(x_{max}) > b(x_{min})$ and is similar, completing the proof. □

Remark 8 (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 (Equation (3) of [9]) also hold for the polling models under consideration. More specifically, consider a tagged customer T_i arriving at Q_i when the server visits Q_i. Then, the sojourn-time distribution of T_i for PS, given n_i customers at Q_i upon arrival, is identical to the waiting-time distribution of T_i 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 and the one customer vanish, explaining the equivalence between Theorems 5 and 3 (see Remark 3).

7 n-class priority queues

In this section we look at n-class priority queues. Each customer is assigned to a priority index k, $1 \leq k \leq n$, where customers with a low priority index are served before customers with higher priority indices. Within each class the service order is FCFS. In Subsection 7.1, the focus lies on the non-preemptive n-class priority regime. We will later use this discipline to find the waiting-time distribution in shortest job first (SJF) queues, by letting the number of priority classes go to infinity. In [19], Kella and Yechiali study the M/G/1 queue with single and multiple server vacations under both the preemptive and non-preemptive 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 also study the preemptive n-class priority regime in Subsection 7.2.

7.1 Non-preemptive n-class priority queues

Here, we are interested in the non-preemptive n-class priority regime. We now introduce our notation and terminology based on [19], as this turns out to be useful and provide intuition for this and the next section. We replace vacation times with intervisit times and add the subscript i to every queue-dependent variable: $\lambda_{i,k}$ is the arrival rate of class-k customers and $B_{i,k}$ is 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 $\rho_{i,b} = \lambda_{i,b}\,\mathbb{E}[B_{i,b}]$. $\xi_{i,a}$ denotes 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 $B_{i,a}$. Consequently, the LST of $\xi_{i,a}$ and its mean are given by: For $Re(s) > 0$,

$$\xi^*_{i,a}(s) = B^*_{i,a}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)), \quad \mathbb{E}[\xi_{i,a}] = \mathbb{E}[B_{i,a}]/(1-\rho_{i,a}). \tag{33}$$

For this model, Kella and Yechiali [19] derive the following LST for the waiting-time distribution $W_{i,k}$ of a class-k customer in Q_i: For $Re(s) > 0$, $k = 1, \dots, n$,

$$\begin{aligned} W^*_{i,k}(s) = {} & \frac{(1-\rho_i)(1 - I^*_i(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s)} \\ & + \frac{\rho_{i,b}(1 - B^*_{i,b}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[B_{i,b}](\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s)} \quad (i \in NPRIOR). \end{aligned} \tag{34}$$

The first term of (34) 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 heavy-traffic limit of the distribution of $W_{i,k}$.

Theorem 6. *For $\rho \uparrow 1$, $k = 1, \dots, n$,*

$$\tilde{W}_{i,k} \to_d \begin{cases} 0 & \text{w.p. } \frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \\ U_i\tilde{\mathbf{I}}_i & \text{w.p. } \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \end{cases} \quad (i \in NPRIOR),$$

where U_i is a uniformly distributed random variable that lies between 0 and $\frac{1}{1-\hat{\rho}_{i,a}}$ and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i. The parameters α and μ_i are given in (7).

Proof. Combining Equation (34) and Property 2, we get for the LST of the (scaled) waiting time of a class-k customer: for $Re(s) > 0$, $k = 1, \dots, n$, $i \in NPRIOR$:

$$\begin{aligned}
\tilde{W}^*_{i,k}(s) &= \lim_{\rho \uparrow 1} W^*_{i,k}(s(1-\rho)) \\
&= \lim_{\rho \uparrow 1} \Bigg[\frac{(1-\rho_i)\left(1 - \left(\frac{\mu_i}{\mu_i + s + \lambda_{i,a}(1-\xi^*_{i,a}(s(1-\rho)))/(1-\rho)}\right)^{\alpha}\right)}{\mathbb{E}[I_i](\lambda_{i,k} B^*_{i,k}(s(1-\rho) + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s(1-\rho))) - \lambda_{i,k} + s(1-\rho))} \\
&\qquad + \frac{\rho_{i,b}(1 - B^*_{i,b}(s(1-\rho) + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s(1-\rho))))}{\mathbb{E}[B_{i,b}](\lambda_{i,k} B^*_{i,k}(s(1-\rho) + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s(1-\rho))) - \lambda_{i,k} + s(1-\rho))} \Bigg] \\
&= \frac{(1-\hat{\rho}_i)\left(1 - \left(\frac{\mu_i}{\mu_i + s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}\right)^{\alpha}\right)}{\mathbb{E}[S](1-\hat{\rho}_i)s(1-\hat{\rho}_{i,k}(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}]))} + \frac{\hat{\rho}_{i,b}(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}{1-\hat{\rho}_{i,k}(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])} \\
&= \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \frac{1}{\mathbb{E}[S]s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])(1-\hat{\rho}_i)} \left\{1 - \left(\frac{\mu_i}{\mu_i + s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}\right)^{\alpha}\right\} \\
&\qquad + \frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \\
&= \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \frac{1}{\mathbb{E}[S]s(1-\hat{\rho}_i)/(1-\hat{\rho}_{i,a})} \left\{1 - \left(\frac{\mu_i}{\mu_i + s/(1-\hat{\rho}_{i,a})}\right)^{\alpha}\right\} \\
&\qquad + \frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}.
\end{aligned} \tag{35}$$

The third equality was found using l'Hôpital's rule and some basic calculations. After some rewriting we arrive at the fourth equation, and writing out $\mathbb{E}[\xi_{i,a}]$ using (33) leads to the final equation. Recognizing this as the LST of a random variable that is equal to zero with probability $\frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ and a uniform times a gamma distribution with probability $\frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ completes the proof. □

Remark 9 (HTAP). We can use the fluid model to give some intuition for the asymptotic waiting-time distribution, which corresponds to a uniform times a gamma distribution in addition to a probability mass at zero. In the fluid model, we only consider class a and class k particles, as the impact of class b is negligible in HT. Figure 3 gives a graphical representation of the fluid model; the workload of class a and k particles in Q_i is plotted over the course of a cycle of length c. The considered particles arrive at the queue with rate $\hat{\rho}_{i,a} + \hat{\rho}_{i,k}$ and during a visit time they are served with rate 1 until the queue is empty. The cycle is divided in three parts: the first part is the intervisit time I_i with length $(1-\hat{\rho}_i)c$. The second part is the duration between a polling instant and the first time since the start of the cycle for which no class a and k particles are present. This part has length $\frac{(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})(1-\hat{\rho}_i)c}{1-(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})}$. In this part only class a and k particles are served. The last part is the part where class b particles are served, interrupted by classes a and k, having length

$$c - (1-\hat{\rho}_i)c - \frac{(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})(1-\hat{\rho}_i)c}{1-(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})} = \frac{\hat{\rho}_{i,b}c}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}.$$

Now, first consider the atom in zero. With probability $\frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ a class-k particle arrives during the last part of the cycle where hardly any class a or k particles are present. In this case the scaled waiting time of the particle is negligible in HT, since the residual service time of the particle in service and the busy periods generated by class-a customers arriving during this remaining service time do not scale with ρ.

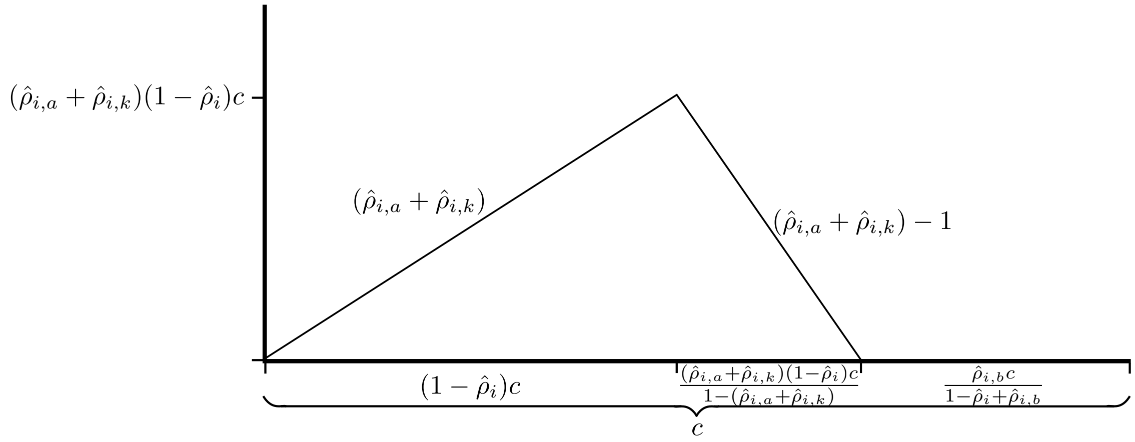


Figure 3: Fluid limits in heavy traffic. The workload of class a and k particles in Q_i is plotted over the course of a cycle.

Second, with probability $\frac{1-\hat\rho_i}{1-\hat\rho_i+\hat\rho_{i,b}}$ a particle arrives during the first or second part of the cycle. Let the uniform random variable U_i denote the fraction of the length of the first two parts of the cycle together that has elapsed at the arrival epoch of the tagged arriving particle. Similar to FCFS, the scaled waiting time of this particle is the remaining duration $(1-U_i)\frac{(1-\hat\rho_i)c}{1-\hat\rho_i+\hat\rho_{i,b}}$ minus the time required to serve the class-k work (or extended service time) that arrives during the first two parts of the cycle, but after the tagged particle. Due to class-a interruptions, the extended service time of class k is $\mathbb{E}[B_{i,k}]/(1-\hat\rho_{i,a})$. Hence, the scaled waiting time is $(1-U_i)\frac{(1-\hat\rho_i)c}{1-\hat\rho_i+\hat\rho_{i,b}}(1-\frac{\hat\rho_{i,k}}{1-\hat\rho_{i,a}}) = (1-U_i)\frac{(1-\hat\rho_i)c}{1-\hat\rho_{i,a}}$. Since $I_i = (1-\hat\rho_i)c$, this term corresponds to a uniform distribution on $[0, \frac{1}{1-\hat\rho_{i,a}}]I_i$, explaining the result for non-negligible waiting times.

7.2 Preemptive n-class priority queues

Similar to the previous section, the results of [19] 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 the current 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

$$R^*_{i,k} = B^*_{i,k}(s+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s)). \tag{36}$$

For a class-k customer in Q_i the LST of waiting time in queue is: For $Re(s) > 0$, $k = 1,\dots,n$,

$$\begin{aligned} W^{(q),*}_{i,k}(s) = & \frac{(1-\rho_i)(1-I^*_i(s+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B^*_{i,k}(s+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s))-\lambda_{i,k}+s)} \\ & + \frac{\rho_{i,b}(\lambda_{i,a}(1-\xi^*_{i,a}(s))+s)}{\lambda_{i,k}B^*_{i,k}(s+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s))-\lambda_{i,k}+s} \quad (i \in NPRIOR\text{-}PR). \end{aligned} \tag{37}$$

For n-class priority queues, the waiting-time distribution in heavy traffic is equal to the case of non-preemptive 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 we get using (37): For $Re(s) > 0$, $i \in NPRIOR - PR$, $k = 1 \dots, n$,

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

which is equal to (35) from the non-preemptive case. As before, α and μ_i are given in (7). From (36) it follows directly that the residence time can be neglected in heavy traffic.

8 Shortest-Job-First and SRPT

The Shortest-Job-First (SJF) service discipline can be thought of as a non-preemptive 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 [25], 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 heavy-traffic limit derived in the previous section we can immediately derive the heavy-traffic limit of the waiting-time distribution for SJF. In Subsection 8.1 we give the scaled waiting-time distribution conditional on the service requirement. In Subsection 8.2 we give the unconditional scaled waiting-time distribution. SRPT and preemptive SJF are discussed in Subsection 8.3.

8.1 Conditional waiting-time distribution in heavy traffic

To go from Equation (35) to SJF we let the service time of the customer determine its priority. Note that we can apply 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 $\rho_i(x) = \lambda_i\,\mathbb{E}[B_i \mathbb{1}_{\{B_i<x\}}]$ 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 down the conditional LST using (35): For $Re(s) > 0$, $x > 0$,

$$\begin{aligned}\tilde{W}_i^*(s|x) = {}& \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} \frac{1}{\mathbb{E}[S]s(1-\hat{\rho}_i)/(1-\hat{\rho}_i(x))} \left\{1-\left(\frac{\mu_i}{\mu_i+s/(1-\hat{\rho}_i(x))}\right)^{\alpha}\right\} \\ & + \frac{\hat{\rho}_i-\hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} \quad (i \in SJF). \end{aligned} \tag{38}$$

This result gives rise to the following theorem.

Theorem 7. *For $\rho \uparrow 1$,*

$$\tilde{W}_{i,x} \to_d \begin{cases} 0 & \text{w.p. } \frac{\hat{\rho}_i-\hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} \\ U_{i,x}\tilde{\mathbf{I}}_i & \text{w.p. } \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} \end{cases} \quad (i \in SJF). \tag{39}$$

$U_{i,x}$ is a random variable with a uniform distribution on $[0, \frac{1}{1-\hat{\rho}_i(x)}]$ and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ_i as given in (7).

Proof. The results follows directly from (38). □

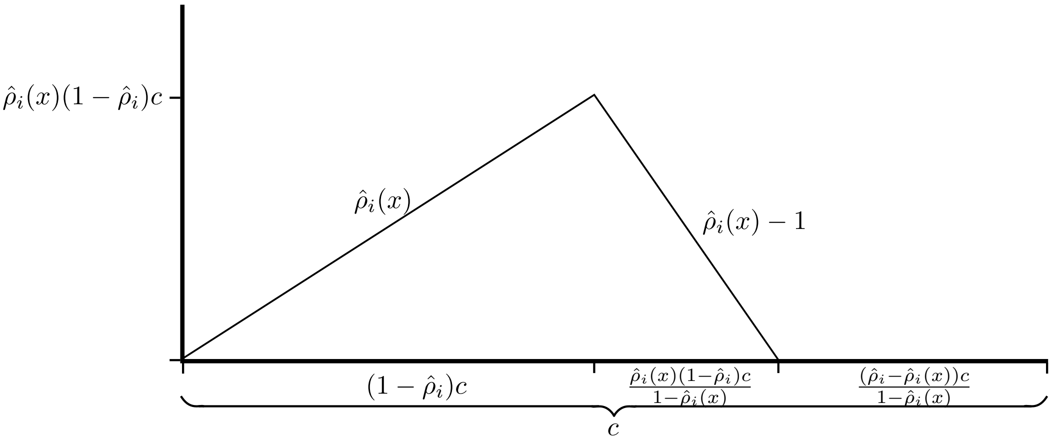

$\hat{\rho}_i(x)(1-\hat{\rho}_i)c$
$\hat{\rho}_i(x)$
$\hat{\rho}_i(x)-1$
$(1-\hat{\rho}_i)c$
$\frac{\hat{\rho}_i(x)(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)}$
$\frac{(\hat{\rho}_i-\hat{\rho}_i(x))c}{1-\hat{\rho}_i(x)}$
c

Figure 4: Fluid limits in heavy traffic. The amount of type a workload in Q_i is plotted over the course of a cycle.

Remark 10 (HTAP). The intuition for the asymptotic waiting-time distribution is similar to the n-class priority queue, but slightly simpler. For the fluid model, we only consider particles that are served before a particle with service requirement x, i.e., type-a particles. Figure 4 gives a graphical representation of the fluid model; on the horizontal axis the course of a cycle with length c is plotted. On the vertical axis the workload of type-a particles in Q_i is plotted. The cycle is divided in three parts; the first part is the intervisit time I_i with length $(1-\hat{\rho}_i)c$. The second part is the first part of the visit time where type-a particles are being served; it starts at polling instant of Q_i and ends the first moment since the start of the cycle that no type-a particles are present. This part has length $\frac{\hat{\rho}_i(x)(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)}$. The last part is the part where the other particles are served and has length

$$c-(1-\hat{\rho}_i)c-\frac{\hat{\rho}_i(x)(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)}=\frac{c(\hat{\rho}_i-\hat{\rho}_i(x))}{1-\hat{\rho}_i(x)}.$$

With probability $\frac{\hat{\rho}_i-\hat{\rho}_i(x)}{1-\hat{\rho}_i(x)}$ a particle with service requirement x arrives during the last part of the cycle where hardly any type-a particles are present. Again, the scaled waiting time in HT is negligible in this case, since the remaining service duration of the particle in service and the type-a busy periods generated by type-a particles arriving during this remaining duration do not scale with ρ. With probability $\frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)}$ a particle arrives during the duration of the first two parts together. Let the uniform random variable U_i denote the fraction of combined length of the first two parts that has elapsed at the arrival epoch of the arriving particle. This particle is served at the start of the third part of the cycle, so the waiting time of this particle is the remaining duration of the first two parts $(1-U_i)\frac{(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)}$. Using $I_i=(1-\hat{\rho}_i)c$, it follows that the scaled waiting time is now uniformly distributed on $[0,\frac{1}{1-\hat{\rho}_i(x)}]I_i$.

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

$$\tilde{W}_i \to_d \tilde{U}_i\tilde{\mathbf{I}}_i \quad (i \in SJF), \tag{40}$$

where $\tilde{U}_i$ *has probability density function*

$$f_{\tilde{U}_i}(y) = \begin{cases} 1-\hat{\rho}_i & y \in [0,1] \\ (1-\hat{\rho}_i)\left(1 - F_{B_I}\left(\hat{\rho}_i^{-1}\left(\frac{y-1}{y}\right)\right)\right) & y \in \left(1, \frac{1}{1-\hat{\rho}_i}\right], \end{cases} \tag{41}$$

with a point mass at zero of

$$\int\limits_0^\infty \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} f_{B_i}(x)\,\mathrm{d}x. \tag{42}$$

$\tilde{\mathbf{I}}_i$ *has a gamma distribution with parameters* $\alpha+1$ *and* μ_i *as given in* (7).

Proof. Note that the conditional waiting-time distribution in (39) can be written as a gamma distribution times a uniform distribution with a point mass at zero; 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 1. The cumulative distribution function of the conditional "uniform" distribution is given by

$$F_{U_{i,x}}(y) = \begin{cases} 0, & y < 0, \\ \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} + \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} y(1-\hat{\rho}_i(x)), & 0 \leq y \leq \frac{1}{1-\hat{\rho}_i(x)}, \\ 1, & y > \frac{1}{1-\hat{\rho}_i(x)}. \end{cases}$$

The probability density function of $U_{i,x}$ is given by $f_{U_{i,x}}(y) = 1-\hat{\rho}_i$, for $y \in [0, \frac{1}{1-\hat{\rho}_i(x)}]$, thus we have $a(x) = 0$ and $b(x) = \frac{1}{1-\hat{\rho}_i(x)}$. Recall that $\hat{\rho}_i(x) = \hat{\lambda}_i\,\mathbb{E}[B_i \mathbb{1}_{\{B_i < x\}}]$ 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 $y \leq 1$, we find

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

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

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

Combining the results above we see that $\tilde{U}_i$ has probability mass (42) in zero, and density (41). This completes the proof. □

8.3 SRPT and preemptive SJF

In this subsection 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; 16]. Similar to SJF, the waiting-time distribution for preemptive SJF follows directly from the preemptive n-class priority queue of Subsection 7.2.

The analysis of SRPT does not follow directly from the results of Kella and Yechiali [19]. 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 [19]. In particular, letting class-a represent customers with service times smaller than x, $\xi_{i,a}^*(s)$ is defined by

$$\xi_{i,a}^*(s) = \frac{1}{F_{B_i}(x)} \int_0^x \exp\left(-t(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s))\right) f_{B_i}(t)\,\mathrm{d}t, \tag{43}$$

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 4. *For $\rho < 1$, $i \in SRPT$, $Re(s) > 0$,*

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

Proof. We start with the multi-class case, where class-k is the class under consideration having service times in $(x - \epsilon, x]$, for $\epsilon > 0$ small, and classes a and b have priority index lower and higher than k, respectively. That is, the service times of class-a is smaller than $x - \epsilon$ and of class-b is larger than x. Applying the idea of Schrage and Miller [25], 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 [19], 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 [19, p.28], we have the cycle probabilities

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

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

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

Using the probabilities above in Equations (7a) and (8) of [19], we obtain, for $Re(s) > 0$,

$$\begin{aligned} W_{i,k}^{(q),*}(s) = {} & \frac{(1-\rho_i)(1 - I_i^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B_{i,k}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) - \lambda_{i,k} + s)} \\ & + \frac{\Pi_{i,0}(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) + \lambda_{i,b}\left(1 - \exp\left(-x(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s))\right)\right)}{\lambda_{i,k}B_{i,k}^*(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s)) - \lambda_{i,k} + s}. \end{aligned} \tag{44}$$

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

As in Subsection 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 [25].

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

Theorem 9. *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 Subsection 7.2 that the heavy-traffic limit is identical to the one for SJF. Using Proposition 4, it follows that $\lim_{\rho\uparrow 1} W_{i,x}^{(q),*}(s(1-\rho))$ equals the right-hand side of (38). Using (36) 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. □

9 Summary of the results

In this section we give a summary of the most important results obtained in this paper. The main result of the paper is the fact that the scaled waiting-time distribution can always be characterized as a product of two distributions. The first distribution is a service-order specific distribution, the second distribution is a gamma distribution. The gamma distribution is a scaled length-biased intervisit-time distribution or cycle-time distribution; the most intuitive representation for the second depends on the scheduling policy. Due to the fact that for exhaustive service at queue i it holds that $C_i^*(s) = I_i^*(s + \lambda_i(1 - xi_i^*(s)))$, see also (5), we can rewrite the second (gamma) distribution as the scaled length-biased intervisit-time distribution for all scheduling policies.

Let Θ_i denote the service-order specific distribution; the probability density functions for the different service policies are then given in Table 3. In Figure 5 we plot the pdf $f_{\Theta_i}(x)$ of Θ_i (Figure 5a) and also the cumulative distribution functions $F_{\Theta_i}(x)$ (Figure 5b). 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 5a 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 zero can be observed from Figure 5b. In addition, these cdfs allow us to see the impact of scheduling policy. For instance, SJF is here superior to ROS and PS.

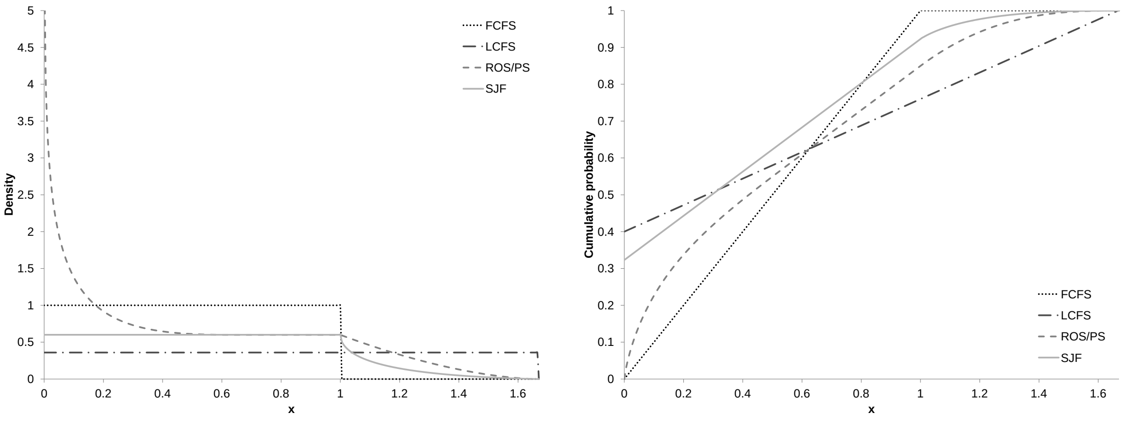
10 Numerical results

In this section we illustrate the results by calculating moments and tail probabilities of the waiting-time distribution for different service disciplines by simulations. Moreover, we use the heavy-traffic 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 $i = 1, \ldots, N$,

$$\mathbb{P}(W_i \leq x) \approx \mathbb{P}(\Theta_i \Gamma_i \leq (1-\rho)x). \tag{45}$$

Service order	pdf of Θ_i
FCFS	$f_{\Theta_i}(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$
LCFS/LCFS-PR	$f_{\Theta_i}(x) = (1-\hat{\rho}_i) \begin{cases} 1-\hat{\rho}_i & x \in [0, \frac{1}{1-\hat{\rho}_i}] \\ 0 & \text{otherwise} \end{cases}$ with a point mass of $\hat{\rho}_i$ in zero
ROS/PS	$f_{\Theta_i}(x) = \hat{\rho}_i \begin{cases} \frac{1}{\hat{\rho}_i}\mathrm{Beta}_{1-x(1-\hat{\rho}_i)/\hat{\rho}_i}(1+\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0) & x \in [0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}] \\ 0 & \text{otherwise} \end{cases}$ $+\mathbb{1}_{\{\hat{\rho}_i \leq 1/2\}}(1-\hat{\rho}_i) \begin{cases} 1-g(x) & x \in [0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}) \\ 1 & x \in [\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 1] \\ g(x-1) & x \in (1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}+1] \\ 0 & \text{otherwise} \end{cases}$ $+\mathbb{1}_{\{\hat{\rho}_i > 1/2\}}(1-\hat{\rho}_i) \begin{cases} 1-g(x) & x \in [0, 1) \\ g(x-1)-g(x) & x \in [1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}] \\ g(x-1) & x \in (\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}+1], \\ 0 & \text{otherwise} \end{cases}$ where $g(x) = \left(1 - \frac{x(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}}$
NPRIOR/	$f_{\Theta_{i,k}}(x) = \dfrac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \begin{cases} 1-\hat{\rho}_{i,a} & x \in \left[0, \frac{1}{1-\hat{\rho}_{i,a}}\right] \\ 0 & \text{otherwise} \end{cases}$
NPRIOR-PR	with a point mass of $\dfrac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ in zero
SJF/SRPT	$f_{\Theta_i}(x) = \begin{cases} 1-\hat{\rho}_i & x \in [0,1] \\ (1-\hat{\rho}_i)\left(1-F_{B_i}\left(\hat{\rho}_i^{-1}\left(\frac{x-1}{x}\right)\right)\right) & x \in \left(1, \frac{1}{1-\hat{\rho}_i}\right] \\ 0 & \text{otherwise} \end{cases}$ with a point mass of $\displaystyle\int_0^\infty \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} f_{B_i}(x)\,\mathrm{d}x$ in zero

Table 3: The probability density functions of the service-order specific distributions.



(a) Probability density functions

(b) Cumulative distribution functions

Figure 5: Shapes of the service order specific distributions.

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

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

See Section 11 and references therein for a discussion on convergence of moments.

We consider a polling model with $N = 3$ queues and all queues receive exhaustive service. Service times and switch-over times are exponentially distributed. The mean service durations at queue 1, 2, and 3 equal 2, 3, and 1 respectively. The mean switch-over times are given by $\mathbb{E}[S_1] = \mathbb{E}[S_3] = 1$ and $\mathbb{E}[S_2] = 3$. Arrivals are Poisson and the arrival rates at the different queues are chosen such that the ratios between the arrival rates are 3:2:1, while the total load of the system is varied. Note that the system is rather asymmetric and that the ratios between the loads of the queues are 6:6:1. We apply the approximation to a system with a load of 0.95 and let the service order be ROS, PS and SJF. We plot the approximated and simulated cumulative distributions of the waiting time at the first queue. Figure 6 shows that the approximation follows the simulation closely. ROS and PS are plotted together, since the distributions are equal. Note that for the SJF service discipline the approximation shows a point mass at zero, this effect does not show up as clearly in the simulation. This is caused by the fact that the point mass at zero only occurs if the load is very close to 1.

To illustrate the differences between the various scheduling policies we plot the approximated cumulative distribution functions of the scaled waiting times at the first queue of the system described above. In Figure 7 we clearly see a point mass at zero if the service discipline is LCFS or SJF. The line of SJF always lies above the line of PS; as the service-time distribution is exponential, this indicates that for exponential service times SJF is a better policy than PS. Table 4 shows the simulated and approximated values of the mean waiting times at Q_1 and their relative absolute differences defined as

$$\Delta\% := 100\% \times \frac{|\text{app} - \text{sim}|}{\text{sim}}$$

for different values of ρ and for different scheduling policies considered in this paper. The mean waiting times are equal for FCFS, LCFS and ROS and also for PS if the service-time distribution is exponential. In Table 5, the results for the standard deviations of

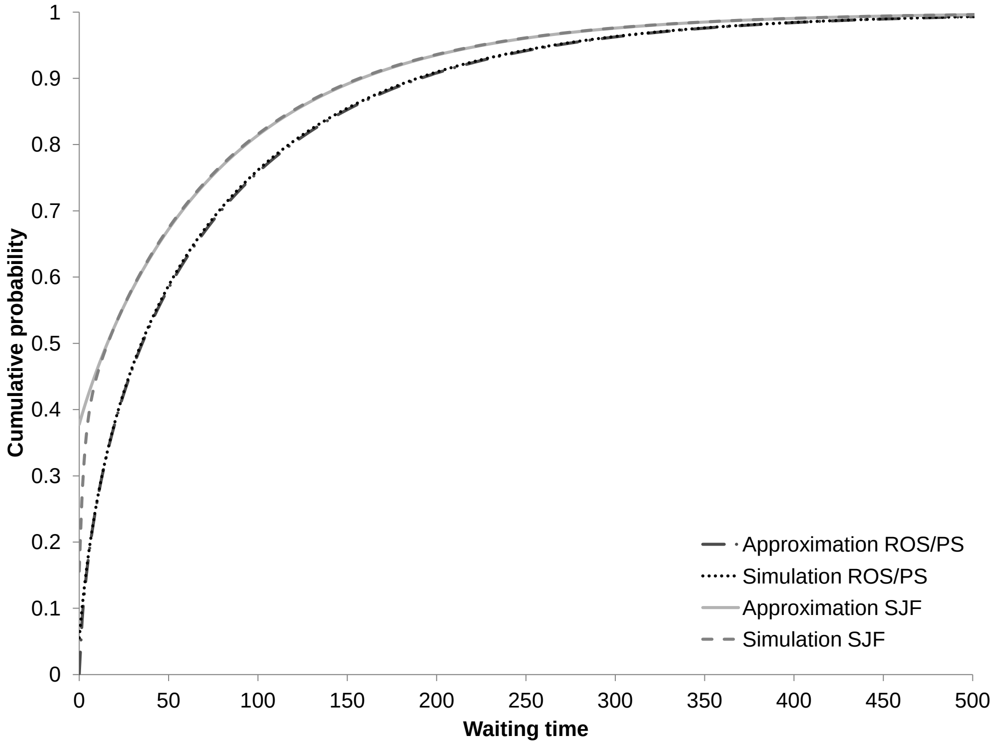


Figure 6: Approximated and simulated cumulative distribution functions of the waiting-time distribution in a system with a load of 0.95.

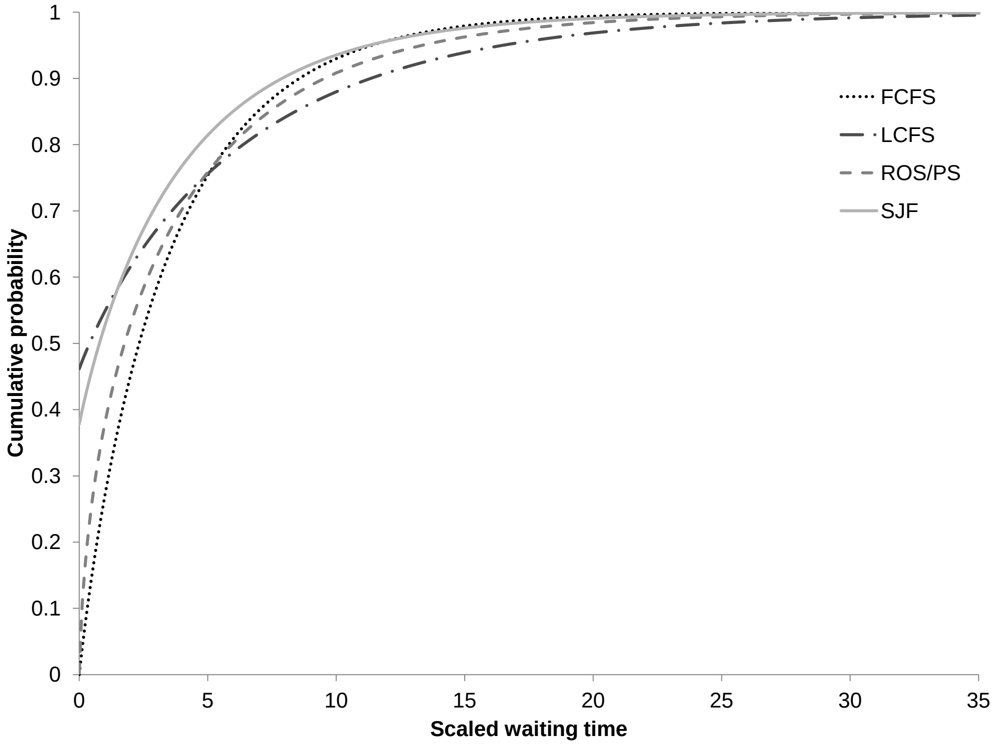


Figure 7: Cumulative distribution functions of the scaled waiting times at the first queue for different service orders.

the waiting times at Q_1 are given. Both tables show that the relative differences decrease to 0 if ρ increases to 1. It is interesting to note that for lower values of ρ, the error in the standard deviation is quite high, especially if the service order is LCFS. This can be explained by the fact that in HT the waiting time is equal to zero if an arrival occurs during a visit period. For lower loads this effect does not occur, busy periods will influence the waiting time. The numerical approximations can be improved using an interpolation with light-traffic limits, as carried out in [4; 15].

	FCFS/LCFS/ROS/PS			SJF		
ρ	sim	app	$\Delta\%$	sim	app	$\Delta\%$
0.7	12.25	12.02	1.89	10.11	8.88	12.19
0.8	18.43	18.03	2.15	14.69	13.31	9.34
0.9	36.68	36.07	1.66	28.16	26.63	5.45
0.95	72.79	72.13	0.90	54.86	53.26	2.92
0.98	180.91	180.33	0.32	134.82	133.14	1.25
0.99	361.32	360.66	0.18	267.83	266.29	0.57

Table 4: Simulated value, approximated value and delta of the mean waiting time for different service disciplines and loads.

	LCFS			ROS/PS			SJF		
ρ	sim	app	$\Delta\%$	sim	app	$\Delta\%$	sim	app	$\Delta\%$
0.7	17.86	20.92	17.10	15.10	16.22	7.37	13.35	14.30	7.12
0.8	28.53	31.38	9.99	23.46	24.33	3.69	20.64	21.44	3.91
0.9	60.02	62.76	4.56	48.02	48.65	1.31	42.16	42.89	1.72
0.95	122.64	125.52	2.35	96.72	97.30	0.60	85.01	85.78	0.90
0.98	310.73	313.80	0.99	242.49	243.26	0.31	213.70	214.44	0.35
0.99	624.20	627.59	0.54	485.78	486.51	0.15	427.88	428.88	0.24

Table 5: Simulated value, approximated value and delta of the standard deviation of the waiting time for different service disciplines and loads.

11 Discussion and Concluding Remarks

In this paper we assume that all queues receive exhaustive service, which is an important extension of the results obtained for similar models but with gated service at all queues [4]. We emphasize that the exhaustive service case is more complicated than the gated case, despite the fact that both the exhaustive and the gated service disciplines satisfy the well-known branching structure identified in [23]. The complexity lies in the fact that for exhaustive service the local service order of the customers during a visit period V_i of the server to a given queue i cannot be determined at the polling instant marking the beginning of V_i; for gated the service order is determined at the beginning of V_i. As a consequence, newly arriving customers at queue i during V_i may change the local service order and the sharing of server capacity among the customers served during V_i, and hence affect the waiting-time and sojourn-time distributions in a complex manner. For example, this complexity manifests itself in the case of PS service and multiple vacations, where analytic results on (conditional) sojourn times, conditioned on the number of customers at the beginning of a service period, are only known under the assumption of exponential service times (see [11]). Even for multiple vacation models, extension of such results to the case of general service times is complicated, because of the complex relation between the number of customers in the system and the remaining amounts of per-customer service times.

The assumption that all queues are served exhaustively can easily be relaxed to the general setting where a subset of the queues receive gated service (or some other branching-type service policy). More specifically, for general mixtures of exhaustive and gated service, let G be the set of indices i for which Q_i receives gated service, and $E := \{1, \dots, N\} \backslash G$ the subset of queues that receive exhaustive service. Then the results presented above still hold; the only difference is that the parameter δ in (4) should be replaced by

$$\delta_{mixture} := 1 - \sum_{i \in E} \hat{\rho}_i^2 + \sum_{i \in G} \hat{\rho}_i^2. \tag{46}$$

In the present paper it is assumed that the arrival processes at the queues are Poisson. This assumption can easily be relaxed to renewal arrivals. Following a well-established line of argumentation (see [12; 13; 22]), one may conjecture that results presented in Section 3 to 8 are still valid when σ^2 defined in (4) is replaced by

$$\sigma^2_{renewal} := \sum_{i=1}^{N} \hat{\lambda}_i \left(Var[B_i] + \hat{\rho}_i^2 Var[\hat{A}_i] \right), \tag{47}$$

where the random variable A_i denotes the interarrival times at Q_i with $\hat{A}_i$ being the limiting case $\rho \uparrow 1$.

Finally, we address a number of topics for further research. First, the heavy-traffic results proven in this paper demonstrate convergence *in distribution* by demonstrating point-wise convergence of the LST's to their limiting regimes, and application of Levy's Continuity Theorem. An interesting question is whether the results can be extended to other types of convergence, and under what assumptions. For example, convergence in distribution does not necessarily imply moment-wise convergence; the latter requires the finiteness of higher moments of the service times and switch-over times. We refer to [30] (Section 3.3) for more detailed discussion about moment-wise convergence. In the case of PS service at queue i we made the additional assumption that the service times are exponentially distributed. Under this assumption, we proved the correctness of Theorem 4 by using the results in [1] (Section 5) which, in turn, rely on the classical results by Coffmann et al. [11] for the M/M/1 PS queue (without vacations). It is an open question how the results for PS can be extended to the case of generally distributed service times.

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

References

[1] U. Ayesta, O. J. Boxma, and I. M. Verloop. Sojourn times in a processor sharing queue with multiple vacations. *Queueing Systems*, 71(1-2):53–78, 2012.

[2] K. R. Baker. Sequencing rules and due-date assignments in a job shop. *Management science*, 30(9):1093–1104, 1984.

[3] N. Bansal and D. Gamarnik. Handling load with less stress. *Queueing Systems*, 54 (1):45–54, 2006.

[4] R. Bekker, P. Vis, J. L. Dorsman, R. D. Van der Mei, and E. M. M. Winands. The impact of scheduling policies on the waiting-time distributions in polling systems. *To appear in Queueing Systems*, 2014.

[5] M. A. A. Boon. *Polling Models, From Theory to Traffic Intersections*. Ph.d. thesis, Eindhoven University of Technology, The Netherlands, 2011.

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

[7] M. A. A. Boon, I. J. B. F. Adan, and O. J. Boxma. A two-queue polling model with two priority levels in the first queue. *Discrete Event Dynamic Systems*, 20(4): 511–536, 2010.

[8] M. A. A. Boon, R. D. Van der Mei, and E. M. M. Winands. Applications of polling systems. *Surveys in Operations Research and Management Science*, 16(2):67–82, 2011.

[9] S. C. Borst, O. J. Boxma, J. A. Morrison, and R. Núnez Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, 31(4):254–262, 2003.

[10] O. J. Boxma, J. Bruin, and B. Fralix. Sojourn times in polling systems with various service disciplines. *Performance Evaluation*, 66(11):621–639, 2009.

[11] E. G. Coffman, R. R. Muntz, and H. Trotter. Waiting-time distributions for processor-sharing systems. *Journal of the ACM*, 17:123–130, 1970.

[12] E. G. Coffman, A. A. Puhalskii, and M. I. Reiman. Polling systems with zero switchover times: a heavy-traffic averaging principle. *The Annals of Applied Probability*, 5(3):681–719, 1995.

[13] E. G. Coffman, A. A. Puhalskii, and M. I. Reiman. Polling systems in heavy traffic: A bessel process limit. *Mathematics of Operations Research*, 23(2):257–304, 1998.

[14] J. R. Dorp and S. Kotz. Generalized trapezoidal distributions. *Metrika*, 58(1):85–97, 2003.

[15] J. L. Dorsman, R. D. Van der Mei, and E. M. M. Winands. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *Stochastic Models*, 27(2):318–332, 2011.

[16] M. Harchol-Balter. Queueing disciplines. *Wiley Encyclopedia of Operations Research and Management Science*, 2009.

[17] R. Hariharan, W. K. Ehrlich, P. K. Reeser, and R. D. Van der Mei. Performance of web servers in a distributed computing environment. *Teletraffic Engineering in the Internet Era*, pages 137–148, 2001.

[18] N. Kawasaki, H. Takagi, Y. Takahashi, S. j. Hong, and T. Hasegawa. Waiting time analysis of $M^X/G/1$ queues with/without vacations under random order of service discipline. *Journal of the Operations Research Society of Japan*, 43(4):455–468, 2000.

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

[20] J. F. C. Kingman. On queues in which customers are served in random order. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(1):79–91, 1962.

[21] T. L. Olsen and R. D. Van der Mei. Polling systems with periodic server routeing in heavy traffic: distribution of the delay. *Journal of Applied Probability*, 40(2):305–326, 2003.

[22] T. L. Olsen and R. D. Van der Mei. Polling systems with periodic server routing in heavy traffic: renewal arrivals. *Operations Research Letters*, 33(1):17–25, 2005.

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

[24] M. Scholl and L. Kleinrock. On the M/G/1 queue with rest periods and certain service-independent queueing disciplines. *Operations Research*, 31(4):705–719, 1983.

[25] L. E. Schrage and L. W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14(4):670–684, 1966.

[26] H. Takagi. *Analysis of Polling Systems.* MIT press, 1986.

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

[28] H. C. Tijms. *A First Course in Stochastic Models.* Wiley, 2003.

[29] R. D. Van der Mei. Distribution of the delay in polling systems in heavy traffic. *Performance Evaluation*, 38(2):133–148, 1999.

[30] R. D. Van der Mei. Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Systems*, 57(1):29–46, 2007.

[31] R. D. van der Mei, R. Hariharan, and P. K. Reeser. Web server performance modeling. *Telecommunication Systems*, 16(3-4):361–378, 2001.

[32] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.

[33] A. Wierman, E. M. M. Winands, and O. J. Boxma. Scheduling in polling systems. *Performance Evaluation*, 64(9):1009–1028, 2007.

[34] E. M. M. Winands, I. J. B. F. Adan, and G. J. Van Houtum. Mean value analysis for polling systems. *Queueing Systems*, 54(1):35–44, 2006.