The Impact of Scheduling Policies on the Waiting-time Distributions in Polling Systems

R. Bekker[a], P. Vis[b,a], J.L. Dorsman[c,b], R.D. van der Mei[b,a], and E.M.M. Winands[d]

[a]VU University Amsterdam, Amsterdam, Netherlands
[b]Centre for Mathematics and Computer Science, Amsterdam, Netherlands
[c]Eindhoven University of Technology, Eindhoven, Netherlands
[d]University of Amsterdam, Netherlands

March 21, 2014

Abstract

We consider polling models consisting of a single server that visits the queues in a cyclic order. In the vast majority of papers that have appeared on polling models, it is assumed that at each of the individual queues the customers are served on a First-Come-First-Served (FCFS) basis. In this paper we study polling models where the local scheduling policy is not FCFS, but instead, is varied as Last-Come-First-Served (LCFS), Random Order of Service (ROS), Processor Sharing (PS) and Shortest-Job-First (SJF). The service policies are assumed to be either gated or globally gated. The main result of the paper is the derivation of asymptotic closed-form expressions for the Laplace-Stieltjes transform (LST) of the scaled waiting-time and sojourn-time distributions under heavy-traffic assumptions. For FCFS service the asymptotic sojourn-time distribution is known to be of the form $U\Gamma$, where U and Γ are uniformly and gamma distributed with known parameters. In this paper we show that the asymptotic sojourn-time distribution (1) for LCFS is also of the form $U\Gamma$, (2) for ROS is of the form $\tilde{U}\Gamma$ where $\tilde{U}$ has a *trapezoidal distribution*, and (3) for PS and SJF is of the form $\tilde{U}^*\Gamma$ where $\tilde{U}^*$ has a *generalized trapezoidal distribution*. These results are rather intriguing and lead to new fundamental insight in the impact of the local scheduling policy on the performance of polling models. As a by-product the heavy-traffic results suggest simple closed-form approximations for the complete waiting-time and sojourn-time distributions for stable systems with arbitrary load values. The accuracy of the approximations is evaluated by simulations.

1 Introduction

A polling system is a multi-queue single-server system in which the server visits the queues in some order to process requests pending at the queues. Polling systems were first introduced in the late 1950s by Mack et al. [14, 13] to model a patrolling repairman. Polling models find a wealth of applications in areas like computer-communication systems and production-inventory systems. For extensive surveys on polling systems and their applications, we refer to [5, 12, 19, 20, 21, 22, 29]. Within a polling system there are a number of design decisions that have to be made, i.e.,

1. The order in which to serve the queues.
2. How many customers to serve during each visit to a queue.
3. The order in which customers within each queue are served.

For the first two decisions a wide variety of policies has been proposed and analyzed. The focus of the current paper is on the third decision. Namely, we investigate the influence of the effective local service order on the waiting times of the customers. As a result, we will limit discussion to the most common configurations for the first two decisions: cyclic service order and (globally-)gated service.

It might be natural to assume that the impact of such local scheduling is small, because it only impacts the system performance locally, leaving the amount of time spent outside the targeted queue unaffected. However, [30] illustrates that the impact on the system performance from scheduling within a queue of a polling system

The research of J.L. Dorsman is funded in the framework of the STAR-project "Multilayered queueing systems" by the Netherlands Organization for Scientific Research (NWO).

can be significant. In many application areas of polling models, such as Bluetooth and 802.11 protocols, scheduling policies at routers and I/O subsystems in web servers, the workloads are known to have high variability and priority-based scheduling could therefore be beneficial. Outside of computer-communication systems, local scheduling proved its worth in the domain of production-inventory control. Our goal is to explore the impact of the local scheduling in polling systems under *heavy traffic* (HT) conditions.

The motivation for studying the HT regime is twofold. First of all, it is the most important and challenging regime from a practical scheduling point of view, i.e. the proper operation of the system is particularly critical when the system is heavily loaded. Optimizing the local scheduling 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. As a result, HT asymptotics form an excellent basis for developing simple accurate approximations for the performance measures (distributions, moments, tail probabilities) for stable systems. These closed-form approximations allow for back-of-the-envelope calculations.

Although an enormous number of papers on both polling systems and scheduling policies have appeared, the combination of the two has received very little attention. That is, almost all theoretical studies of scheduling policies are performed in single-queue settings such as the M/G/1 and G/G/1 queue with only a few exceptions studying the effect of local scheduling in multi-queue polling systems. By using the *Mean Value Analysis* (MVA) framework for polling systems [32], Wierman et al. [30] have derived the mean delay in cyclic exhaustive and gated polling systems for various scheduling disciplines such as *First-Come-First-Served* (FCFS), *Last-Come-First-Served* (LCFS), *Foreground-Background* (FB), *Processor Sharing* (PS), *Shortest-Job-First* (SJF) and fixed priorities. Building upon these results, Boxma et al. [7] have obtained the waiting-time distribution in cyclic (globally-)gated polling systems for various service orders. As indicated by [7], the derivation of the waiting-time distribution in *exhaustive* polling systems is much more intricate. Recently, interesting progress in this direction has, however, been made. Boon et al. [4] have studied the waiting-time distribution in a two-queue polling model with either the exhaustive, gated or globally-gated service discipline. The first of these two queues contains customers of two priority classes. In [3] these results are generalized to a polling model with N queues and K_i priority levels in queue i. Moreover, Ayesta et al. [2] 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.

Concluding we can say that there is a limited amount of research on local scheduling in polling systems, which is in our opinion due to the following factors:

1. One might believe that the added value of local scheduling in a polling system is small, because it does not impact the time spent on the other queues in the system.
2. The analysis of polling systems is difficult even in the simple case of FCFS scheduling.

The HT analysis for scheduling policies in polling systems remained elusive at all; the present study aims to fill that gap.

In this paper we study Poisson-driven cyclic polling models with general service-time and switch-over time distributions, and with two types of service policies: (1) models with gated service at each queue, and (2) models with globally-gated service. For both types of service policies, we consider the following five scheduling policies that determine the local order in which the customers at a given queue are served: FCFS, LCFS, ROS, PS and SJF, where ROS denotes *Random Order of Service.*

For each of these models we derive exact closed-form expressions for the *Laplace-Stieltjes Transform* (LST) of the (scaled) waiting-time and sojourn-time distributions under HT assumptions. Note that it was shown in [25] that the asymptotic cycle-time distributions converge to a gamma distribution with known parameters. Using this result, for FCFS service it is shown in [15] that the asymptotic sojourn-time distribution is a product of the random variables U and Γ, where U and Γ are uniformly and gamma distributed. In this paper, we unify and extend this result by presenting rigorous proofs showing that the asymptotic sojourn-time distribution is (1) for LCFS also of the form $U\Gamma$, (2) for ROS of the form $\tilde{U}\Gamma$ where $\tilde{U}$ has a *trapezoidal distribution*, and (3) for PS and SJF of the form $\tilde{U}^*\Gamma$ where $\tilde{U}^*$ has a *generalized trapezoidal distribution.*

We can conclude that the main result of the paper is that we rigorously prove closed-form expressions for the (scaled) waiting-time distribution under HT assumptions, explicitly quantifying the impact of the local scheduling policies on the waiting-time performance. The main surprising observation made is that the asymptotic distribution of the waiting time for all of these priority-based scheduling disciplines can be described by the product of a generalized trapezoidal distribution and a gamma distribution, both with known parameters. We would like to stress the unearthed dichotomy between the known HT results on FCFS polling models and our novel asymptotic results for other scheduling disciplines. That is, for standard FCFS and LCFS scheduling the uniform distribution is prevalent in the HT distribution, whereas the various alternative scheduling disciplines lead to the appearance of the (generalized) trapezoidal distribution in the asymptotic distributions.

These results are rather intriguing and provide new fundamental insight in the impact of the local scheduling policy on the performance of polling models. Our results lead not only to unification but also to extension of the literature studying scheduling policies, polling systems and HT asymptotics. As a by-product the HT results suggest simple closed-form approximations for the complete waiting-time and sojourn-time distributions for stable systems with arbitrary load values and *general renewal arrival processes.* Numerical results show that these approximations perform well for a wide range of parameter combinations.

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 derive the HT asymptotics for the model with gated service at each queue under various local scheduling policies. Section 4 presents similar results for the case of globally gated service. Furthermore, Section 5 proposes a simple approximation for the sojourn-time distributions for arbitrary load values and present numerical results to evaluate the accuracy of the approximations. Section 6 contains a number of concluding remarks.

2 Model and notation

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. Customers arrive at Q_i according to a Poisson process 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, with LST $B_i^*(\cdot)$ and finite k-th moment $b_i^{(k)} = \mathbb{E}[B_i^k]$, $k = 1, 2$. The k-th moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \mathbb{E}[B^k] = \sum_{i=1}^{N} \lambda_i \, \mathbb{E}[B_i^k]/\Lambda$, $k = 1, 2$. 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 $\rho = \sum_{i=i}^{N} \rho_i$. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is an independent random variable S_i with mean $r_i := \mathbb{E}\left[S_i\right]$. Let $S = \sum_{i=1}^{N} S_i$ denote the total switch-over time in a cycle and let $r := \mathbb{E}\left[S\right]$ denote its mean. A necessary and sufficient condition for stability of the system is $\rho < 1$. C_i denotes the cycle time at queue i, defined as the time between two successive arrivals of the server at queue i; it is well known that $\mathbb{E}\left[C_i\right] = r/(1-\rho)$ for each i (c.f. [19, Equation (5.39b)]).

The *service policy* determines *which* customers are served during a visit of the server to a queue. In this paper we assume two variants: (1) the model with gated service at each of the queues, and (2) the globally-gated model. For gated service, all customers are served that were present at polling instant, i.e., at the moment when the server arrives at the queue. For globally gated, during a cycle, all customers are served that were present at polling instant of the first queue. The *local scheduling policy* determines the *order* in which the customers are served during a visit period at a queue. We consider the following five local scheduling policies: FCFS, LCFS, ROS, PS or SJF. For policy $P \in \{\text{FCFS}, \text{LCFS}, \text{ROS}, \text{PS}, \text{SJF}\}$, we denote $i \in I_P$ if Q_i receives scheduling policy P. For example, I_{FCFS} is the (index) set of queues i that are served on a FCFS basis, I_{LCFS} is the (index) set of queues i that are served on a LCFS basis, and so on.

In this paper we study heavy-traffic limits, i.e., the limiting behavior as ρ approaches 1. The heavy-traffic limits, denoted $\rho \uparrow 1$, taken in this paper are such that the arrival rates are increased, while keeping both the service-time distributions and the ratios between the arrival rates fixed. Light-traffic limits, denoted $\rho \downarrow 0$, are defined similarly. 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 $\hat{x}$. In particular we have $\hat{\rho}_i = \frac{\rho_i}{\rho}$ and $\hat{\lambda}_i = \frac{\hat{\rho}_i}{\mathbb{E}[B_i]}$.

Let W_i denote the waiting time of an arbitrary customer at Q_i, defined as the time between the arrival of a customer and the moment at which he enters service. The sojourn time of an arbitrary customer at Q_i, represented by T_i, is defined as the time between the arrival of a customer and the moment at which he departs from the system. The LSTs of W_i and T_i are denoted by $W_i^*(s)$ and $T_i^*(s)$, respectively. When $\rho \uparrow 1$, all queues become unstable, therefore the focus lies on the random variables $(1-\rho)W_i$ and $(1-\rho)T_i$ as $\rho \uparrow 1$, referred to as the *scaled* waiting times and sojourn times at Q_i, respectively.

Throughout, a key role is played by the gamma distribution and the uniform distribution. A non-negative continuous random variable $\Gamma(\alpha, \mu)$ is said to have a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\mu > 0$ if it has the probability density function

$$f_\Gamma(x) = \frac{\mu^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\mu x} \quad (x > 0), \quad \text{with} \quad \Gamma(\alpha) := \int_0^\infty t^{\alpha-1} e^{-t} \mathrm{d}t \tag{1}$$

and LST

$$\Gamma^*(s) = \left(\frac{\mu}{\mu + s}\right)^\alpha \quad (Re(s) > 0). \tag{2}$$

Note that in the definition of the gamma distribution μ is a scaling parameter, and that $\Gamma(\alpha, \mu)$ has the same distribution as $\mu^{-1}\Gamma(\alpha, 1)$. Moreover, we denote by $U[a, b]$ $(a < b)$ a random variable that is uniformly distributed over the interval $[a, b]$. For later reference note, that the LST of the random variable $\Gamma(\alpha + 1, \mu)U[a, b]$, where $\Gamma(\alpha + 1, \mu)$ and $U[a, b]$ are independent, is given by

$$\mathbb{E}\left[e^{-sU[a,b]\Gamma(\alpha+1,\mu)}\right] = \frac{\mu}{\alpha s(b-a)} \left\{ \left(\frac{\mu}{\mu + sa}\right)^\alpha - \left(\frac{\mu}{\mu + sb}\right)^\alpha \right\} \quad (Re(s) > 0). \tag{3}$$

3 Analysis of models with gated service

In this section we consider the case of gated service at all queues. In Subsection 3.1 we review some known preliminary results for FCFS disciplines to be used for later reference. In Subsections 3.2 – 3.5 we use the results in Subsection 3.1 to derive heavy-traffic limits for LCFS, ROS, PS and SJF, respectively. In Section 5 we use these results to propose and validate approximations for the distributions of the waiting times and sojourn times for arbitrary load values and renewal arrivals.

It is easy to see that for FCFS, LCFS and ROS service the sojourn time is simply the convolution of the waiting time and the service time, i.e., for $Re(s) \geq 0$,

$$T_i^*(s) = W_i^*(s) B_i^*(s), \qquad (i \in I_{FCFS}, I_{LCFS}, I_{ROS}). \tag{4}$$

For this reason, in Subsections 3.1–3.3 we focus on the waiting-time distributions. The results for the sojourn-time distributions then follow directly from (4). Note that for $i \in I_{PS}$ and $i \in I_{SJF}$ relation (4) is generally not true, because in those cases the waiting times and the service times are not independent. Relation (4) is used for the approximation in Section 5, since sojourn times and waiting times are equal in HT.

To start, let us consider the distribution of the cycle time C_i, defined as the time between two successive arrivals of the server at queue i. A simple but important observation is that the distribution of C_i is independent of the local scheduling policy (i.e., FCFS, LCFS, ROS, PS and SJF). To this end, recall that the *service policy* (e.g., gated or globally gated) determines *which* customers are served during a visit V of the server to a queue, and that the local *scheduling policies* determine the *order* in which these customers are served during V. For this reason the cycle-time distributions are the same for all local scheduling policies under consideration, provided that they are work-conserving.

The following result gives a characterization for the limiting behavior of the cycle-time distributions, stating that the (scaled) cycle times $(1 - \rho)C_i$ converge to a gamma distribution with known parameters (proven in [25, 26]).

Property 1. (Convergence of the cycle times). *For the model with gated service at each queue we have, for $i = 1, \ldots, N$,*

$$(1 - \rho)C_i \to_d \tilde{\Gamma},$$

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

$$\alpha := \frac{r\delta}{\sigma^2}, \quad \mu := \frac{\delta}{\sigma^2}, \tag{5}$$

with

$$\sigma^2 := \frac{b^{(2)}}{b^{(1)}}, \quad \textit{and} \quad \delta := \sum_{i=1}^N \hat{\rho}_i(1 + \hat{\rho}_i). \tag{6}$$

3.1 First-Come-First-Served

In this section we review several known results for the case of FCFS service at queue i. In Subsections 3.2–3.5 these results are used to derive new results for LCFS, ROS, PS and SJF, respectively. For FCFS service, the following result gives an expression for the LST of the waiting time W_i in terms of the distribution of the cycle time C_i (proven in [7]):

Property 2. (Cycle-time expression for the waiting times) *For the gated service model, we have for* $Re(s) > 0$ *and* $\rho < 1$,

$$W_i^*(s) = \frac{C_i^*(\lambda_i(1 - B_i^*(s))) - C_i^*(s)}{\mathbb{E}\left[C_i\right](s - \lambda_i(1 - B_i^*(s)))} \qquad (i \in I_{FCFS}). \tag{7}$$

The following result, which was shown in [25], characterizes the limiting behavior of the waiting-time distribution in heavy traffic.

Property 3. (Convergence of the waiting times) *For the gated service model, we have for* $\rho \uparrow 1$,

$$(1 - \rho)W_i \to_d U_i\tilde{\mathbf{C}}_i \quad (i \in I_{FCFS}), \tag{8}$$

where U_i *is uniformly distributed on the interval* $[\hat{\rho}_i, 1]$, *and where* $\tilde{\mathbf{C}}_i$ *has a gamma distribution with parameters* $\alpha + 1$ *and* μ, *where* α *and* μ *are given in Equation* (5). *The random variables* U_i *and* $\tilde{\mathbf{C}}_i$ *are independent.*

Note that here $\tilde{\mathbf{C}}_i$ is the *length-biased* version of $\tilde{C}_i$, a gamma-distributed random variable with parameters α and μ as in Equation (5). It is well known that if a gamma random variable has parameters α and μ then its length-biased version has parameters $\alpha + 1$ and μ. The following result gives an expression for the higher moments of the waiting times in heavy traffic (proven in [27, 28]):

Property 4. (Convergence of moments of the waiting time) *For* $k = 1, 2, \ldots$,

$$\omega_i^{(k)} := \lim_{\rho \uparrow 1} (1 - \rho)^k \, \mathbb{E}\left[W_i^k\right] = \frac{1 - \hat{\rho}_i^{k+1}}{1 - \hat{\rho}_i} \frac{\prod_{j=1}^{k}(\alpha + j)}{(k + 1)\mu^k} \quad (i \in I_{FCFS}), \tag{9}$$

assuming that the $(k+1)$*-st moments of the service-time distributions and the* k*-th moments of the switch-over time distributions are finite.*

3.2 Last-Come-First-Served

The LST for the waiting-time distribution for the LCFS service is expressed in terms of the cycle-time distributions as follows (cf. [7]): For $Re(s) > 0$ and $\rho < 1$,

$$W_i^*(s) = \frac{1 - C_i^*(s + \lambda_i(1 - B_i^*(s)))}{\mathbb{E}\left[C_i\right](s + \lambda_i(1 - B_i^*(s)))} \quad (i \in I_{LCFS}). \tag{10}$$

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

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

$$(1 - \rho)W_i \to_d U_i\tilde{\mathbf{C}}_i \qquad (i \in I_{LCFS}), \tag{11}$$

where U_i *is uniformly distributed on the interval* $[0, 1 + \hat{\rho}_i]$ *and* $\tilde{\mathbf{C}}_i$ *has a gamma distribution with parameters* $\alpha + 1$ *and* μ, *where* α *and* μ *are given in Equation* (5). *The random variables* U_i *and* $\tilde{\mathbf{C}}_i$ *are independent.*

Proof. Take $i \in I_{LCFS}$. Then combining (10) with Property 1 gives the following expressions for the LST of the (scaled) waiting-time distribution. For $i \in I_{LCFS}$, $Re(s) > 0$, we have

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) = \lim_{\rho\uparrow 1} \frac{1 - C_i^*(s(1-\rho) + \lambda_i(1 - B_i^*(s(1-\rho))))}{\mathbb{E}\left[C\right]\left(s(1-\rho) + \lambda_i(1 - B_i^*(s(1-\rho)))\right)} \tag{12}$$

$$= \lim_{\rho\uparrow 1} \frac{1 - \left(\frac{\mu(1-\rho)}{\mu(1-\rho)+s(1-\rho)+\lambda_i(1-B_i^*(s(1-\rho)))}\right)^{\alpha}}{\frac{r}{(1-\rho)}\left(s(1-\rho) + \lambda_i(1 - B_i^*(s(1-\rho)))\right)} \tag{13}$$

$$= \lim_{\rho\uparrow 1} \frac{1 - \left(\frac{\mu}{\mu+s+\lambda_i(1-B_i^*(s(1-\rho)))/(1-\rho)}\right)^{\alpha}}{r\left(s + \frac{\lambda_i(1-B_i^*(s(1-\rho))}{1-\rho}\right)}. \tag{14}$$

Using l'Hôpital's rule and the fact that $-B_i^{*'}(0) = \mathbb{E}[B_i]$ we see that:

$$\lim_{\rho\uparrow 1} \frac{\lambda_i(1 - B_i^*(s(1-\rho)))}{1-\rho} = \lim_{\rho\uparrow 1} \frac{0 - \lambda_i B_i^{*'}(s(1-\rho))s}{1} = \hat{\rho}_i s,$$

which immediately implies that, for $Re(s) > 0$,

$$\tilde{W}_i^*(s) = \frac{1 - \left(\frac{\mu}{\mu+s+\hat{\rho}_i s}\right)^{\alpha}}{r(s + \hat{\rho}_i s)} = \frac{1}{rs(1+\hat{\rho}_i)}\left\{1 - \left(\frac{\mu}{\mu + s(1+\hat{\rho}_i)}\right)^{\alpha}\right\} \quad (i \in I_{LCFS}), \tag{15}$$

where α and μ are given in (5). Using (3) and $\mu/\alpha = 1/r$, it now follows that (15) corresponds to the LST of a uniform random variable on $[0, 1 + \hat{\rho}_i]$ times a gamma distribution. Application of Levy's Continuity Theorem (c.f. [31]) completes the proof. □

Using Theorem 1, it is easily verified that the moments of the asymptotic delay distribution are given by the following expression.

Corollary 1. (Moments of the asymptotic delay) *For $k = 1, 2, \ldots,$*

$$\omega_i^{(k)} := \lim_{\rho\uparrow 1} (1-\rho)^k\, \mathbb{E}[W_i^k] = \frac{(1+\hat{\rho}_i)^k \prod_{j=1}^{k}(\alpha + j)}{(k+1)\mu^k} \quad (i \in I_{LCFS}), \tag{16}$$

where α and μ are defined in Equation (5), assuming that the $(k+1)$-st moments of the service-time distributions and the k-th moments of the switch-over time distributions are finite.

We end this section with a number of remarks.

Remark 1. (Comparison between FCFS and LCFS case using the heavy-traffic averaging principle) Property 3 and Theorem 1 reveal an interesting difference in the waiting-time distributions between the FCFS case and the LCFS case. More precisely, for the FCFS case the limiting behavior of W_i is of the form $U_{FCFS}\Gamma$, where U_{FCFS} is uniformly distributed on the interval $[\hat{\rho}_i, 1]$, whereas for the LCFS case the limiting distribution of W_i is of the form $U_{LCFS}\Gamma$, where U_{LCFS} is uniformly distributed on the interval $[0, 1 + \hat{\rho}_i]$, with the *same* gamma distribution. To provide an intuitive explanation for this, we use the insights that can be obtained by the so-called Heavy-Traffic Averaging Principle (HTAP), see e.g. [8, 9] and [15, 16]. Loosely speaking, HTAP for polling models means that the total scaled workload may be considered as a constant during a cycle, whereas the workloads of the individual queues change much faster according to deterministic trajectories, or a fluid model. Due to the HTAP, we let the constant c denote the cycle length. Let us first consider the fluid model for FCFS. Note that the waiting time consists of two parts. First, a customer has to wait for the residual cycle time, which is $(1-U)c$ for U uniformly distributed on $[0, 1]$. Second, a customer has to wait for all customers that have arrived before him during the course of the ongoing cycle. Hence, this equals $\hat{\rho}_i U c$. The total waiting time in the fluid model then equals $(1 - U + \hat{\rho}_i U)c$, which has a uniform distribution on $[\hat{\rho}_i c, c]$. Using that the cycle time follows a gamma distribution explains the shape of the waiting-time distribution in heavy traffic. For LCFS, as for FCFS, an arriving customer still has to wait for the residual cycle length $(1-U)c$, with U a uniform random variable on $[0, 1]$. In addition, the arriving customer has to wait for all customers that arrived after him during the same cycle, which is of length $\hat{\rho}_i(1-U)c$. Hence, the waiting time in the fluid model is $(1+\hat{\rho}_i)(1-U)c$, which is a uniform distribution on $[0, (1+\hat{\rho}_i)c]$. This interpretation gives much insight in the heavy-traffic asymptotics.

Remark 2. (Alignment with asymptotics with large switch-over times) Further support can be given for the distribution in Theorem 1 by considering a different asymptotic regime as in [24]. Let the switch-over times be deterministic with length r_i. We consider the behavior of W_i when the switch-over times tend to infinity. Because the waiting times are known to grow without bound when the switch-over times increase to infinity, the analysis is oriented towards the limiting distribution of $\frac{W_i}{r}$ as $r \to \infty$. Using similar techniques as in [24], it may be shown that

$$\frac{W_i}{r} \to_d \hat{W}_i \quad (r \to \infty), \tag{17}$$

where $\hat{W}_i$ is uniformly distributed over the interval $[\tilde{a}_i, \tilde{b}_i]$, with

$$\tilde{a}_i = \frac{\rho_i}{1-\rho}, \quad \tilde{b}_i = \frac{1}{1-\rho} \text{ for } i \in I_{FCFS}, \quad \text{and } \tilde{a}_i = 0, \quad \tilde{b}_i = \frac{\rho_i + 1}{1-\rho} \text{ for } i \in I_{LCFS}. \tag{18}$$

Note that the uniform distribution is the same as in the HT regime. However, in HT the cycle times follow a gamma distribution whereas here the cycle times become deterministic as the switch-over times grow large.

3.3 Random Order of Service

In this section we derive heavy-traffic limits for the Random Order of Service (ROS) local scheduling policy. ROS is represented by ordering marks. Each customer that arrives gets an ordering mark x, a realization from a uniform distribution on $[0, 1]$. When the server arrives at the queue, the gate closes and the customers before the gate are served in order of their marks. It is convenient to condition with respect to x and then uncondition. Let $W_i(x)$ be the waiting time of a customer in queue i with ordering mark x, with $i \in I_{ROS}$, and let $W_i^*(s|x)$ be the corresponding LST. The following result was shown in [7]: for $Re(s) > 0$, $0 < x < 1$ and $\rho < 1$,

$$W_i^*(s|x) = \frac{C_i^*(\lambda_i x(1 - B_i^*(s))) - C_i^*(s + \lambda_i x(1 - B_i^*(s)))}{s\,\mathbb{E}[C_i]} \quad (i \in I_{ROS}). \tag{19}$$

The next result gives the heavy-traffic limit of the distribution of $W_i(x)$.

Theorem 2. (Conditional waiting time) *For $\rho \uparrow 1$, $0 < x < 1$,*

$$(1-\rho)W_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i \quad (i \in I_{ROS}), \tag{20}$$

where $U_i(x)$ is uniformly distributed over the interval $[\hat{\rho}_i x, 1 + \hat{\rho}_i x]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (5). The random variables $U_i(x)$ and $\tilde{\mathbf{C}}_i$ are independent.

Proof. Combining (19) and Property 1 and using l'Hôpital's rule, we find the following LST of the waiting time conditional on the ordering mark x: for $Re(s) > 0$, $0 < x < 1$,

$$\begin{aligned}\tilde{W}_i^*(s|x) &= \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)|x) \\ &= \lim_{\rho\uparrow 1} \frac{C_i^*(\lambda_i x(1 - B_i^*(s(1-\rho)))) - C_i^*(s(1-\rho) + \lambda_i x(1 - B_i^*(s(1-\rho))))}{s(1-\rho)\,\mathbb{E}[C_i]} \\ &= \frac{1}{rs}\left\{\left(\frac{\mu}{\mu + \hat{\rho}_i x s}\right)^{\alpha} - \left(\frac{\mu}{\mu + (1+\hat{\rho}_i x)s}\right)^{\alpha}\right\} \quad (i \in I_{ROS}).\end{aligned} \tag{21}$$

Applying Levy's Continuity Theorem completes the proof. □

To obtain the *unconditional* distribution of the waiting time, we first consider a more general setting that also covers 'unconditioning' for PS and SJF. For this, let $a(\cdot)$ be a continuous and strictly increasing function on some interval $\mathcal{X} = [x_{min}, x_{max}]$, where we allow x_{max} to be infinite. Suppose we have a conditional random variable, denoted $T|x$, that is uniformly distributed on the interval $[a(x),\, a(x) + 1]$. We want to find the unconditional distribution $\tilde{T}$. Here, x is a realization of the random variable X with support $\mathcal{X} \subseteq \mathbb{R}^+$ having distribution function $F_X(\cdot)$ and density $f_X(\cdot)$. We have the following lemma.

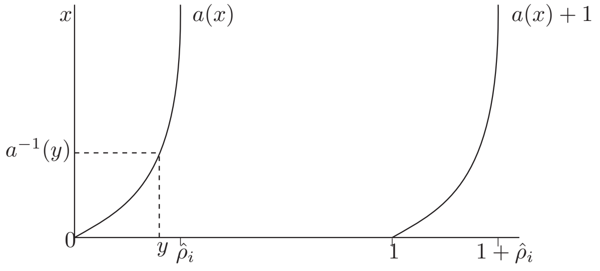

Figure 1: Boundaries of the uniform distribution

Lemma 1. *Assume that the conditional random variable $T|x$ is uniformly distributed on $[a(x),\, a(x)+1]$, where $x \in \mathcal{X} = [x_{min}, x_{max}]$. Suppose that $a(x_{min}) = m$, $a(x_{max}) = \hat{\rho}_i$ and $a(x)$ is continuous and strictly increasing in x, such that $a(\cdot)$ has an inverse denoted by $a^{-1}(\cdot)$. Then, the unconditional distribution of $T|x$, denoted by $\tilde{T}$, has probability density function*

$$f_{\tilde{T}}(y) = \begin{cases} F_X(a^{-1}(y)) & y \in [m, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1+m] \\ 1 - F_X(a^{-1}(y-1)) & y \in (1+m, 1+\hat{\rho}_i]. \end{cases} \tag{22}$$

Proof. Note that $T|x$ has the following probability density function

$$f_{T|x}(y) = \begin{cases} 1 & y \in (a(x), a(x)+1) \\ 0 & \text{Otherwise} \end{cases} \qquad \forall\, x \in \mathcal{X}.$$

Figure 1 shows an example of the boundaries of the uniform distribution, by plotting $a(x)$ and $a(x)+1$ 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 $y \in (m, \hat{\rho}_i)$, in which case the probability density $f_{\tilde{T}}(y)$ is obtained from the parts where x is smaller than $a^{-1}(y)$. This gives, for $y \in (m, \hat{\rho}_i)$,

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

If $y \in (\hat{\rho}_i, 1+m)$ then y is between the boundaries of the uniform distribution for every $x \in \mathcal{X}$. Hence, we get

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

Finally, for $y \in (1+m, 1+\hat{\rho}_i)$, we can use that the boundaries are described by similar curves, i.e., x needs to be larger than $a^{-1}(y-1)$, so

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

Finally, it follows from the properties of $a(\cdot)$ that $f_{\tilde{T}}(\cdot)$ is a density function. This completes the proof. □

Remark 3. For convenience it is assumed in Lemma 1 that the underlying random variable X has a density. For, e.g., PS it can be of interest to consider the case that X is a discrete random variable. This is directly related to the properties of $a(\cdot)$, i.e., that $a(\cdot)$ is continuous and strictly increasing. It is not difficult to modify Lemma 1 to the case of discrete random variables by either redefining the inverse of $a(\cdot)$ as $a^{-1}(y) = \sup\{x \in \mathcal{X} : a(x) \leq y\}$, or by extending the function $a(\cdot)$ from the range of X to an interval $[x_{min}, x_{max}]$, such that $a(\cdot)$ is continuous and strictly increasing.

Note that the density function in (22) is continuous, increasing on $[m, \hat{\rho}_i)$, constant on $[\hat{\rho}_i, 1+m]$ and decreasing on $(1+m, 1+\hat{\rho}_i]$, which closely resembles the traditional trapezoidal distribution. In line with [10], we refer to (22) as a *generalized trapezoidal distribution* consisting of stages of growth, stability, and decay, i.e..

For further references, it is of interest to determine the mean of this generalized trapezoidal distribution. There are different ways to represent this mean, for instance,

$$\mathbb{E}[\tilde{T}] = \int_m^{1+\hat{\rho}_i} x f_{\tilde{T}}(x)\, \mathrm{d}x = \frac{1}{2} + \hat{\rho}_i - \int_m^{\hat{\rho}_i} F_X(a^{-1}(y))\, \mathrm{d}y = \frac{1}{2} + \int_{u \in \mathcal{X}} a(u) f_X(u)\, \mathrm{d}u,$$

where the second step follows after some rewriting. Substituting $y = a(u)$, the third step follows after partial integration. In Subsections 3.4 and 3.5, the mean of the generalized distribution $\mathbb{E}[\tilde{T}]$ is specified for PS and SJF.

We now apply the lemma above to the case $i \in I_{ROS}$, in which case $a(x) = \hat{\rho}_i x$, with $x \in [0, 1]$. The asymptotic scaled unconditional delay is presented in the following theorem.

Theorem 3. (Unconditional waiting time) *For $\rho \uparrow 1$,*

$$(1-\rho)W_i \to_d \tilde{U}_i^* \tilde{\mathbf{C}}_i \quad (i \in I_{ROS}),$$

where $\tilde{U}_i^*$ has a trapezoidal distribution with density function

$$f_{\tilde{U}_i^*}(y) := \begin{cases} y/\hat{\rho}_i & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ (1 + \hat{\rho}_i - y)/\hat{\rho}_i & y \in (1, 1 + \hat{\rho}_i]. \end{cases} \tag{23}$$

and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (5)*. The random variables U_i^* and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Take $i \in I_{ROS}$. Then Theorem 2 implies that $a(x) = \hat{\rho}_i x$ and $\mathcal{X} = [0, 1]$. Note that this function has the desired properties: $a(0) = 0$, $a(1) = \hat{\rho}_i$ and $a(x)$ continuous and strictly increasing in $x \in \mathcal{X}$. The cumulative distribution function of X is given by $F_X(x) = x$ and the inverse function of $a(\cdot)$ is $a^{-1}(y) = y/\hat{\rho}_i$. The use of Lemma 1 now yields the result. □

Remark 4. (HTAP) Interestingly, Theorem 3 shows that the uniform distribution that appears in the heavy-traffic limit for FCFS and LCFS is replaced by a trapezoidal distribution for ROS. The shape of this distribution can be explained by the fact that the waiting time of a customer does not only depend on the time that the customer enters the system, but also on an independent random mechanism that determines the moment that the customer is served. More specifically, exploiting the HTAP we let the constant c denote the cycle length again and consider the fluid model for the conditional waiting time of a customer with mark x. An arriving customer has to wait for the residual cycle length $(1 - U)c$, with U uniformly distributed on $[0, 1]$, and the time required to serve the customers that arrived during the same cycle and have a mark smaller than x, i.e., $\hat{\rho}_i xc$. Clearly, the conditional waiting time in the fluid model is uniformly distributed on $[\hat{\rho}_i xc, (1 + \hat{\rho}_i)xc]$. Since x is an arbitrary order mark, the unconditional waiting time in the fluid model is $(U_1 + U_2)c$, with U_1 and U_2 independent uniform distribution on the intervals $[0, 1]$ and $[0, \hat{\rho}_i]$, respectively. Note that such a convolution gives rise to a trapezoidal distribution as obtained in Theorem 3.

Remark 5. (First moments of waiting times) Observe that it follows from Property 4, Corollary 1 and Theorem 3 that the first moments of the asymptotic waiting-time distributions for the FCFS, LCFS and the ROS scheduling disciplines are the same. This is in line with the observation in [7] that the mean waiting times for these disciplines coincide for a general value of $\rho < 1$. To this end, it is easy to see that $\mathbb{E}[W_i] = (1 + \rho_i)\frac{\mathbb{E}[C_i^2]}{2\,\mathbb{E}[C_i]}$ and that the cycle-time distributions are independent of the local scheduling policy.

3.4 Processor sharing

When the scheduling discipline is PS, the LST of the *conditional* sojourn time (denoted $T_i^*(s|x)$) can also be expressed in terms of the LST of the cycle time. When x is the amount of work that a tagged customer brings into the system, it holds that (c.f. [7]), for $\rho < 1$, $Re(s) > 0$, $x > 0$,

$$T_i^*(s|x) = e^{-sx}\frac{C_i^*(\lambda_i(1 - \varphi(s, x))) - C_i^*(s + \lambda_i(1 - \varphi(s, x)))}{s\,\mathbb{E}[C_i]} \quad (i \in I_{PS}), \tag{24}$$

where $\varphi(s, x) = \mathbb{E}\left[e^{-s\min(B_i,x)}\right]$, the LST of the minimum of B_i and x. The next theorem gives an expression for the asymptotic distribution of the conditional sojourn time $T_i(x)$.

Theorem 4. (Conditional sojourn time) *For $\rho \uparrow 1$, $x > 0$,*

$$(1-\rho)T_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i \quad (i \in I_{PS}), \tag{25}$$

where $U_i(x)$ is uniformly distributed over the interval $[\hat{\lambda}_i\,\mathbb{E}[\min(B_i, x)], 1 + \hat{\lambda}_i\,\mathbb{E}[\min(B_i, x)]]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (5)*. The random variables $U_i(x)$ and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Combining (24) with Property 1, we obtain, for $Re(s) > 0$ and $x > 0$,

$$\tilde{T}_i(s|x) := \lim_{\rho\uparrow 1} T_i^*(s(1-\rho)|x)$$
$$= \frac{1}{rs}\left\{\left(\frac{\mu}{\mu + \hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]s}\right)^{\alpha} - \left(\frac{\mu}{\mu + (1+\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)])s}\right)^{\alpha}\right\}, \tag{26}$$

with $\alpha + 1$ and μ as given in (5). An application of Levy's Continuity Theorem yields the result. □

We now proceed with the unconditional sojourn time. For notational convenience, we assume here that the service-time distributions are absolutely continuous (see however Remark 3).

Theorem 5. (Unconditional sojourn time) *For $\rho \uparrow 1$,*

$$(1-\rho)T_i \to_d U_i^* \tilde{\mathbf{C}}_i \quad (i \in I_{PS}),$$

where U_i^* is a type of generalized trapezoidal distribution as characterized in Equation (27) *with $a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$ and x_{min} the lowest possible value of B_i. The random variables U_i^* and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Take $a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$ such that $U_i(x)$ is uniformly distributed on $[a(x), a(x)+1]$, see Theorem 4. Clearly, $a(x_{min}) = \hat{\lambda}_i x_{min}$, $a(x_{max}) = \hat{\lambda}_i\,\mathbb{E}[B_i] = \hat{\rho}_i$ and $a(x)$ is continuous and strictly increasing. Using Lemma 1, we obtain the unconditioned distribution

$$f_{U_i^*}(y) = \begin{cases} F_{B_i}(a^{-1}(y)) & y \in [\hat{\lambda}_i x_{min}, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i x_{min}] \\ 1 - F_{B_i}(a^{-1}(y-1)) & y \in (1 + \hat{\lambda}_i x_{min}, 1 + \hat{\rho}_i], \end{cases} \tag{27}$$

where x_{min} is the minimum value that B_i can take. This completes the proof. □

Remark 6. (HTAP) Theorem 5 shows that the conditional sojourn time in heavy traffic still is a uniform times a gamma distribution. This can again be intuitively explained from the HTAP. Now, in a cycle of length c, arriving customers have to wait for the residual cycle length $(1-U)c$ and their departure is delayed by all traffic in queue i that arrives during the same cycle and has been served before the tagged customer leaves. The latter equals $\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]c$ in the fluid model. The distribution of the unconditional sojourn time not only depends on the first two moments of the service time, but depends on the complete service-time distribution. In particular, the curve $F_{B_i}(a^{-1}(y))$, with $y \in [\hat{\lambda}_i x_{min}, \hat{\rho}_i)$, can be interpreted as the fluid model of the cumulative number of departures from queue i from the moment that the gate opens at queue i. To interpret this, note that in the fluid model $a(x)$ represents the amount of work served since the gate is open to a customer with service requirement x, and $a^{-1}(\cdot)$ can thus be seen as the time to accumulate such an amount of service. Hence, $F_{B_i}(a^{-1}(y))$ counts the number of customers for which $a^{-1}(y)$ is sufficient to leave.

Remark 7. (Deterministic service times) In most queueing models, high variability leads generally to longer waiting times. However, for the polling model under consideration, note that Theorem 4 implies that for deterministic service times, the waiting time in heavy traffic is also a uniform times a gamma distribution. Here, the boundaries of the uniform distribution are $\hat{\rho}_i$ and $1 + \hat{\rho}_i$. We note that this is the worst possible case for U_i^* among all service-time distributions, in the sense that it has the largest tail $\mathbb{P}(U_i^* > x)$ for all x. This is caused by the fact that all customers are served simultaneously and, in the end, they all jointly leave.

Below we give some examples of the type of generalized trapezoidal distribution U_i^* for some specific service-time distributions. Together with the gamma distribution, representing the cycle time, this fully specifies the scaled sojourn time in heavy traffic.

Exponential service times

Suppose B_i is exponentially distributed with parameter b_i. Then

$$\mathbb{E}[\min(B_i,x)] = \int_0^x y b_i e^{-b_i y}\,\mathrm{d}y + x e^{-b_i x} = \frac{1}{b_i}(1 - e^{-b_i x}),$$

so $a(x) = \hat{\rho}_i(1 - e^{-b_i x})$. Solve $a(x) = y$ for x to find $a^{-1}(y) = \ln(1 - y/\hat{\rho}_i)/(-b_i)$. Now substituting this in Equation (27) it follows after some simplification that the generalized trapezoidal distribution U_i^* coincides with the density function of U_i^* for ROS given in (23). This means that for the case of exponential service-time distributions, the sojourn-time distributions for ROS and PS coincide.

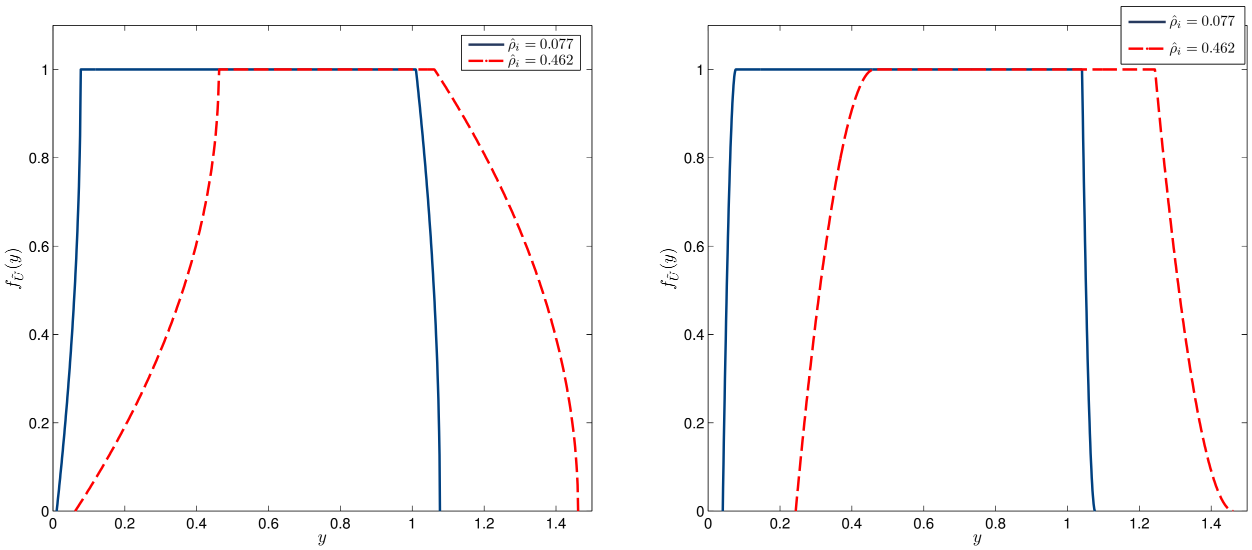

Figure 2: Probability density function of U_i^* with uniform (left) and Pareto (right) service times in a PS polling system

Uniform service times

Suppose B_i is a uniformly distributed random variable on the interval $[a_i, b_i]$. Then

$$a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i, x)] = \hat{\lambda}_i \left(\int_{a_i}^{x} y/(b_i - a_i)\,\mathrm{d}y + x \int_{x}^{b_i} 1/(b_i - a_i)\,\mathrm{d}y \right)$$

$$= \frac{-\hat{\lambda}_i}{2(b_i - a_i)} \left(a_i^2 - 2b_i x + x^2\right) = \frac{-\hat{\rho}_i}{b_i^2 - a_i^2} \left(a_i^2 - 2b_i x + x^2\right).$$

Now $a^{-1}(y)$ can be found using the quadratic formula:

$$a^{-1}(y) = \left(2b_i \pm \sqrt{4b_i^2 - 4(a_i^2 + y/\hat{\rho}_i(b_i^2 - a_i^2))}\right)/2 = b_i - \sqrt{(1 - y/\hat{\rho}_i)(b_i^2 - a_i^2)},$$

where the final equality follows from $x \in [a_i, b_i]$.

In this case $\mathcal{X} = [a_i, b_i]$, which means that the minimum value for y is $a(a_i) = \hat{\lambda}_i a_i$. On the other side of the boundaries of the conditional uniform distribution, y needs to be greater than $1 + \hat{\lambda}_i a_i$, using this we get

$$f_{U_i^*}(y) = \begin{cases} 1 - \frac{\sqrt{(1-y/\hat{\rho}_i)(b_i^2 - a_i^2)}}{b_i - a_i} & y \in [\hat{\lambda}_i a_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i a_i] \\ \frac{\sqrt{(1-(y-1)/\hat{\rho}_i)(b_i^2 - a_i^2)}}{b_i - a_i} & y \in (1 + \hat{\lambda}_i a_i, \hat{\rho}_i + 1]. \end{cases}$$

Figure 2 illustrates the shape of the pdf of U_i^*, when the $c_{B_i}^2$ of the uniform service-time distribution is equal to 0.25 and for two different values of $\hat{\rho}_i$.

Pareto service times

Assume that the service time has a Pareto distribution with parameters a_i and b_i, i.e. we assume that the density of the service time, for $x \geq b_i$, is

$$f_{B_i}(x) = a_i b_i^{a_i} x^{-(a_i+1)}.$$

We assume that $a_i > 2$ such that the second moment is finite. In line with [9, 17] this is sufficient for the HT limit to hold.

Now, we have

$$a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i, x)] = \hat{\lambda}_i \left(\frac{a_i b_i}{a_i - 1} \left(1 - b_i^{a_i - 1} x^{1 - a_i}\right) + b_i^{a_i} x^{1 - a_i} \right) = \hat{\rho}_i \left(1 - b_i^{a_i - 1} x^{1 - a_i} a_i^{-1}\right).$$

Some basic calculations lead to

$$a^{-1}(y) = b_i(a_i(1 - y/\hat{\rho}_i))^{\frac{1}{1-a_i}}.$$

Here y needs to be larger than $a(b_i) = \hat{\rho}_i(1 - a_i^{-1}) = \hat{\lambda}_i b_i$. We have

$$f_{U_i^*}(y) = \begin{cases} 1 - (a_i(1 - y/\hat{\rho}_i))^{\frac{-a_i}{1-a_i}} & y \in [\hat{\lambda}_i b_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i b_i] \\ (a_i(1 - (y-1)/\hat{\rho}_i))^{\frac{-a_i}{1-a_i}} & y \in (1 + \hat{\lambda}_i b_i, 1 + \hat{\rho}_i]. \end{cases}$$

Figure 2 shows the pdf of U_i^* if the Pareto service-time distribution has a squared coefficient of variation equal to 4, for two different values of $\hat{\rho}_i$.

For the special case in which $a_i \to \infty$ the squared coefficient of variation (SCV) of the Pareto distribution goes to zero. In that case, it can be seen that U_i^* has a uniform distribution on the interval $[\hat{\rho}_i, 1 + \hat{\rho}_i]$, which is in line with the case of deterministic service times.

Discrete service times
Using Remark 3, Theorem 5 still applies by extending the range of $a(\cdot)$ (or modifying the inverse $a^{-1}(\cdot)$). An interesting example is when the service time has probability mass at two points. Assume that B_i equals a small value a_i with probability p_i, or a large value b_i with probability $1 - p_i$. Now, letting $x \in [a_i, b_i]$, we have $\mathbb{E}[\min(B_i, x)] = (1 - p_i)x + p_i a_i$, giving $a(x) = \hat{\lambda}_i((1 - p_i)x + p_i a_i)$. With $x \in [a_i, b_i]$, we note that $a(x)$ is thus continuous and strictly increasing. Hence, we obtain

$$f_{U_i^*}(y) = \begin{cases} p_i & y \in [a_i\hat{\lambda}_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + a_i\hat{\lambda}_i] \\ 1 - p_i & y \in (1 + a_i\hat{\lambda}_i, 1 + \hat{\rho}_i]. \end{cases}$$

3.5 Shortest-job-first

For the SJF policy, it is convenient to condition on x, the amount of work that a tagged customer brings into the system. For SJF, the service-time distribution is assumed to be absolutely continuous. The following results gives an expression for the LST of the conditional sojourn time $T_i^*(x)$ in terms of the cycle-time distributions (cf. [7]): for $\rho < 1$, $Re(s) > 0$, $x > 0$,

$$T_i^*(s|x) = e^{-sx}\frac{C_i^*(\lambda_i(1 - \varphi(s,x))) - C_i^*(s + \lambda_i(1 - \varphi(s,x)))}{s\,\mathbb{E}[C_i]} \quad (i \in I_{SJF}), \tag{28}$$

where $\varphi(s,x) := \mathbb{E}\left[e^{-sB_i\mathbf{1}_{\{B_i \leq x\}}}\right]$. This leads to the following theorem for the limiting distribution of the conditional sojourn time $T_i(x)$.

Theorem 6. (Conditional sojourn time) *For $\rho \uparrow 1$, $x > 0$,*

$$(1 - \rho)T_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i \quad (i \in I_{SJF}),$$

where $U_i(x)$ is a uniform$[\hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i \leq x\}}], 1 + \hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i \leq x\}}]]$ random variable and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (5). The random variables $U_i(x)$ and $\tilde{\mathbf{C}}_i$ are independent.

Proof. The result follows directly by combining Equation (28) and Property 1 along lines similar to those in the proof of Theorem 4. □

The unconditional sojourn time is presented in the following theorem.

Theorem 7. (Unconditional sojourn time) *For $\rho \uparrow 1$,*

$$(1 - \rho)T_i \to_d U_i^*\tilde{\mathbf{C}}_i \quad (i \in I_{SJF}),$$

where U_i^* is a generalized trapezoidal distribution as characterized in Equation (29) with $a(x) = \hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i \leq x\}}]$ and $\tilde{\mathbf{C}}_i$ as given in Theorem 6. The random variables U_i^* and $\tilde{\mathbf{C}}_i$ are independent.

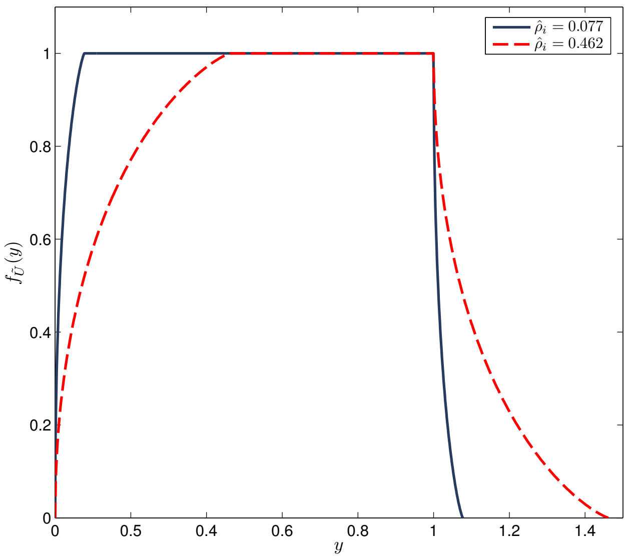

Figure 3: Probability density function of $\tilde{U}$ with exponential service times in a SJF polling system

Proof. Take $a(x) = \hat{\lambda}_i \, \mathbb{E}[B_i \mathbf{1}_{\{B_i \le x\}}]$ such that $U_i(x)$ is uniformly distributed on $[a(x), a(x)+1]$, see Theorem 6. Clearly, $a(x_{min}) = 0$, $a(x_{max}) = \hat{\rho}_i$ and $a(x)$ is continuous and strictly increasing. Using Lemma 1, we obtain the unconditioned distribution

$$f_{U_i^*}(y) = \begin{cases} F_{B_i}(a^{-1}(y)) & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ 1 - F_{B_i}(a^{-1}(y-1)) & y \in (1, 1+\hat{\rho}_i]. \end{cases} \tag{29}$$

This completes the proof. □

Note that, similar to the PS case, the trapezoidal distribution U_i^* depends on the complete service-time distribution. Below, we present some special cases.

Exponential service times

Suppose B_i is exponentially distributed with parameter b_i. First calculate

$$\mathbb{E}[B_i \mathbf{1}_{\{B_i \le x\}}] = \int_0^x y b_i e^{-b_i y} \, \mathrm{d}y = \frac{1}{b_i}(1 - e^{-b_i x}(1 + b_i x)).$$

Hence, $a(x) = \hat{\rho}_i(1 - e^{-b_i x}(1 + b_i x))$. To determine $a^{-1}(y)$, we solve $a(x) = y$ for x and, after some rewriting, obtain the following equation

$$-e^{-1}(1 - y/\hat{\rho}_i) = te^t, \tag{30}$$

where $t = -(1 + b_i x)$. We thus need the solution of (30), which is known to be given in terms of the Lambert W function. Observe that the equation te^t may have multiple solutions, but we need the solutions for real $t \le -1$, denoted by $W_{-1}(\cdot)$. This function decreases from $W_{-1}(-1/e) = -1$ to $W_{-1}(0^-) = -\infty$. From the above we derive $a^{-1}(y) = -(W_{-1}(-e^{-1}(1 - y/\hat{\rho}_i)) + 1)/b_i$.

Since $F_{B_i}(x) = 1 - e^{-b_i x}$, the probability density function $f_{U_i^*}(y)$ of the generalized trapezoidal distribution U_i^* becomes

$$f_{U_i^*}(y) = \begin{cases} 1 - e^{W_{-1}(-e^{-1}(1 - y/\hat{\rho}_i)) + 1} & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ e^{W_{-1}(-e^{-1}(1 - (y-1)/\hat{\rho}_i)) + 1} & y \in (1, 1+\hat{\rho}_i]. \end{cases} \tag{31}$$

The form of this distribution only depends on $\hat{\rho}_i$, this means that it only depends on the ratio between the mean interarrival time and the mean service time. In Figure 3, the probability density function is plotted for two different values of $\hat{\rho}_i$. The figure shows that for small $\hat{\rho}_i$, the distribution is close to a uniform distribution. When $\hat{\rho}_i$ increases, the distribution gets more skewed to the right.

Uniform service times
Suppose B_i has a uniform distribution with parameters a_i and b_i. We have

$$\mathbb{E}[B_i \mathbf{1}_{\{B_i \leq x\}}] = \int_{a_i}^{x} \frac{u}{b_i - a_i}\, \mathrm{d}u = \frac{x^2 - a_i^2}{2(b_i - a_i)} = \mathbb{E}[B_i] \frac{x^2 - a_i^2}{b_i^2 - a_i^2}, \quad \text{for } a_i \leq x \leq b_i.$$

This gives $a(x) = \hat{\rho}_i \frac{x^2 - a_i^2}{b_i^2 - a_i^2}$. Some basic calculus yields the inverse of $a(\cdot)$: $a^{-1}(y) = \sqrt{y\,(b_i^2 - a_i^2)\,/\hat{\rho}_i + a_i^2}$. Because $F_{B_i}(x) = (x - a_i)/(b_i - a_i)$,

$$f_{U_i^*}(y) = \begin{cases} \frac{\sqrt{y/\left(b_i^2 - a_i^2\right)/\hat{\rho}_i + a_i^2} - a_i}{b_i - a_i} & y \in [\hat{\lambda}_i a_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i a_i] \\ 1 - \frac{\sqrt{(y-1)\left(b_i^2 - a_i^2\right)/\hat{\rho}_i + a_i^2} - a_i}{b_i - a_i} & y \in (1 + \hat{\lambda}_i a_i, 1 + \hat{\rho}_i]. \end{cases}$$

Pareto service times
Suppose B_i is Pareto distributed with parameters $a_i > 2$ and b_i. Note that $a_i > 2$ ensures that the second moment is finite, such that the HT limit exists (see e.g. [9, 17]). It is easy to show that $a(x) = \hat{\rho}_i(1 - b_i^{a_i - 1} x^{1 - a_i})$ and $a^{-1}(y) = b_i(1 - y/\hat{\rho}_i)^{\frac{1}{1 - a_i}}$. Using that $F_{B_i}(x) = 1 - (b_i/x)^{a_i},\ x \geq b_i$ gives

$$f_{U_i^*}(y) = \begin{cases} 1 - (1 - y/\hat{\rho}_i)^{\frac{-1}{1 - a_i}} & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ (1 - (y-1)/\hat{\rho}_i)^{\frac{-1}{1 - a_i}} & y \in (1, 1 + \hat{\rho}_i]. \end{cases}$$

4 Results for models with globally gated service

In this section we consider the case of a globally gated service. Recall that (without loss of generality) we assume that the global gate closes at successive polling instants at Q_1 (see for example [7] for a description of the globally gated model). As in Section 3, we analyze LCFS, ROS, PS and SJF in addition to FCFS. Since the derivations for globally gated are largely similar to the case gated service at all queues, we only present the final results and omit the proofs.

The following result (proven in [26]) gives an asymptotic expression for the distribution of the cycle times C_i, defined in Section 2. Note again that the cycle times do not depend on the local scheduling policy.

Property 5. (Convergence of cycle times for globally gated service discipline). *For the globally-gated system we have: for $i = 1, \ldots, N$,*

$$(1 - \rho) C_i \to_d \tilde{\Gamma},$$

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

$$\alpha := \frac{2r}{\sigma^2}, \quad \mu := \frac{2}{\sigma^2}. \tag{32}$$

with σ^2 given by (6).

Following the same line of reasoning as in Section 3, we obtain the waiting-time distributions for all considered scheduling disciplines for globally gated service in heavy traffic. For convenience, we define $\mathrm{P}_i := \sum_{j=1}^{i} \hat{\rho}_j$ for $i = 1, \ldots N$ and by convention $\mathrm{P}_0 := 0$.

Theorem 8. *For globally gated service and $\rho \uparrow 1$, the following properties hold:*

(i) For $i \in I_{FCFS}, I_{LCFS}$,

$$(1 - \rho) W_i \to_d U_i \tilde{\mathbf{C}}_i,$$

where U_i is a uniform$[\mathrm{P}_i, 1 + \mathrm{P}_{i-1}]$ random variable if $i \in I_{FCFS}$ and U_i is a uniform$[\mathrm{P}_{i-1}, 1 + \mathrm{P}_i]$ random variable if $i \in I_{LCFS}$.

(ii) For $i \in I_{ROS}$,

$$(1-\rho)W_i \to_d \tilde{U}_i^* \tilde{\mathbf{C}}_i,$$

where $\tilde{U}_i^*$ has a trapezoidal distribution with probability density function

$$f_{\tilde{U}_i^*}(y) = \begin{cases} (y - \mathrm{P}_{i-1})/\hat{\rho}_i & y \in [\mathrm{P}_{i-1}, \mathrm{P}_i) \\ 1 & y \in [\mathrm{P}_i, \mathrm{P}_{i-1}+1] \\ (\mathrm{P}_i + 1 - y)/\hat{\rho}_i & y \in (\mathrm{P}_{i-1}+1, \mathrm{P}_i + 1]. \end{cases}$$

(iii) For $i \in I_{PS}, I_{SJF}$,

$$(1-\rho)T_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i,$$

where $U_i(x)$ is a uniform$[\mathrm{P}_i + \hat{\lambda}_i\kappa_{i,x}, 1 + \mathrm{P}_i + \hat{\lambda}_i\kappa_{i,x}]$ random variable, with $\kappa_{i,x} := \mathbb{E}[\min\{B_i, x\}]$ for $i \in I_{PS}$ and $\kappa_{i,x} := \mathbb{E}[B_i \mathbf{1}_{\{B_i \leq x\}}]$ for $i \in I_{SJF}$.

In all cases, $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (32) and it is independent of the uniform distributions.

In the above theorem we only presented the conditional waiting times for PS and SJF. Using Lemma 1, this results in a generalized trapezoidal times a gamma distribution for the unconditional waiting time, as in Subsections 3.4 and 3.5. Finally, also the intuitive interpretation of the heavy-traffic limit using HTAP is directly in line with that of Section 3.

Remark 8. (Renewal arrival processes) For the model under consideration with Poisson arrivals, Theorems 1–8 give the asymptotic waiting-time and sojourn-time distributions for the LCFS, ROS, SJF and PS scheduling disciplines. Following the well-established line of argumentation found in [8, 9, 16], we conjecture that similar results hold for renewal arrival processes. In particular, in [16] a strong conjecture is given that in heavy traffic the same results for the scaled cycle-time and waiting-time distributions for FCFS hold as those in Properties 1 and 3, respectively, but where the parameter σ^2 is now replaced by

$$\sigma^2_{renewal} = \sum_{i=1}^{N} \hat{\lambda}_i (\mathrm{Var}[B_i] + c^2_{A_i}\, \mathbb{E}[B_i]^2). \tag{33}$$

Here A_i represents the interarrival times between arriving customers at queue i and $c^2_{A_i}$ is its squared coefficient of variation. Note for the special case of Poisson arrivals, the expression for $\sigma^2_{renewal}$ coincides with σ^2. Based on the HTAP, we derive the following conjecture for renewal arrivals.

Conjecture 1. *For independent renewal arrival processes, Theorems 1–8 are also valid when σ^2 defined in (6) is replaced by $\sigma^2_{renewal}$ defined in (33).*

In the next section, we use this conjecture to derive and validate approximations for the waiting-time and sojourn-time distributions for renewal arrivals.

5 Closed-form approximations for system with arbitrary load

In Sections 3 and 4, we have derived heavy-traffic limits for the (scaled) waiting-time and sojourn-time distributions under several scheduling disciplines. These results not only give valuable insights into polling models operating under a critical load, but are also useful in the study of polling models that are arbitrary loaded (i.e. $\rho < 1$). Below, we describe how the results derived in this paper can be used to obtain closed-form approximations for the waiting-time and sojourn-time distributions in polling models with renewal arrivals and arbitrary load.

For systems with FCFS service at all queues, Boon et al. [6] derive a closed-form approximation, denoted by $\mathbb{E}[W_i^{(app)}]$, for the *mean* waiting time by interpolating between known light-traffic and heavy-traffic limits. Based on this approximation, Dorsman et al. [11] propose to approximate the *complete* waiting-time distribution by, for $x > 0$,

$$\mathbb{P}(W_i < x) \approx \mathbb{P}(U_i C_i < (1-\rho)x) \quad (i \in I_{FCFS}), \tag{34}$$

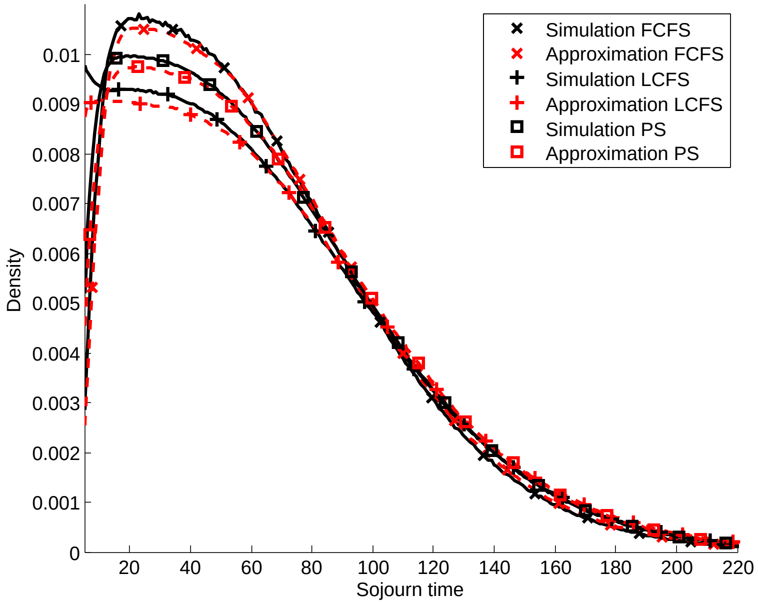

Figure 4: Simulated and approximated sojourn-time distributions at queue 1 for the gated model with $\rho = 0.95$ for FCFS, LCFS and PS.

where U_i is uniformly $[\hat{\rho}_i, 1]$ distributed (as defined in Property 1) and where C_i is gamma-distributed with shape parameter $\alpha + 1$ and scale parameter

$$\mu_i^{(app)} := \frac{1 + \hat{\rho}_i}{1 - \rho} \frac{r\delta + \sigma^2_{renewal}}{2\sigma^2_{renewal}\, \mathbb{E}[W_i^{(app)}]^2}, \tag{35}$$

where α, δ and $\sigma^2_{renewal}$ are defined in Property 1 and (33).

To develop an approximation for the other scheduling disciplines under consideration, recall that the cycle-time distribution is insensitive to the scheduling discipline. Based on this observation, for $i \in I_{LCFS}$, we approximate the waiting-time distribution by (34), where the distribution of C_i is kept the same, but with U_i uniformly distributed on $[0, 1 + \hat{\rho}_i]$ (cf. Theorem 1). Likewise, for $i \in I_{ROS}$, the waiting-time distribution can be approximated by (34) with U_i replaced by a trapezoidal distribution defined in Theorem 3. Approximations for the sojourn-time distributions can be obtained by using (4). For PS and SJF, approximations can be obtained directly for the sojourn-time distributions, using Theorems 5 and 7, respectively. For the case of globally gated service, waiting-time distributions are approximated in a similar way using the results in Section 4; details are omitted here.

Throughout this section we will show numerical results based on simulations to illustrate the usefulness and accuracy of the closed-form approximations. We consider a three-queue polling model with gated service at each queue and with the following parameters. The service times at queues 1, 2 and 3 are uniformly distributed with means 1, 2, 3, respectively, and with squared coefficient of variation 1/4. The switch-over time distributions are exponentially distributed with means $r_1 = r_2 = 1$ and $r_3 = 3$. The arrival processes at each of the queues are renewal and mutually independent. The ratios between the arrival rates are 1:3:2, and interarrival-time distributions are uniformly distributed with squared coefficient of variation 1/4. Note that the system is rather asymmetric and the ratios between the per-queue load values are 1:6:6.

To illustrate the fact that the approximation of the distribution is accurate in heavy traffic, Figure 4 plots the simulated and approximated density functions of the sojourn-time distributions at Q_1 for FCFS, LCFS and PS service (at all queues) for a heavily loaded system with $\rho = 0.95$. As expected, the approximations closely follow the simulations. Figure 4 also illustrates that the differences between the different scheduling disciplines are significant and are well-captured by the asymptotic results.

To proceed, Figure 5 shows the simulated and approximated probability density functions for the per-queue sojourn-time distributions for the model with LCFS service at each queue, for a heavily loaded system with $\rho = 0.95$. Figure 6 shows the results for the same model but with globally-gated service. The results in Figure 5 and 6 illustrate the fact that the per-queue sojourn-time distributions are well-captured by the approximations for heavy-traffic scenarios (as they should).

Next, we assess the accuracy of the approximations for the complete range of load values. To this end, Table 1 shows the simulated and approximated values of the mean sojourn times at Q_1 and their relative absolute

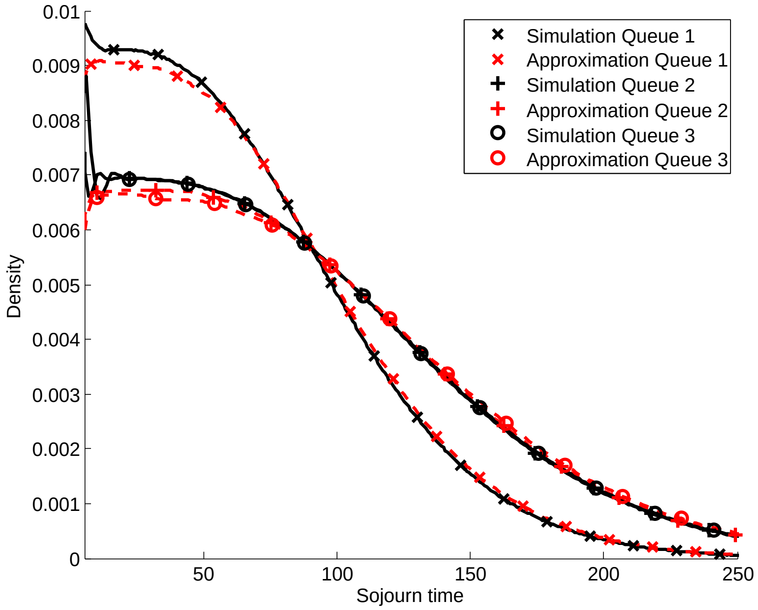

Figure 5: Simulated and approximated per-queue sojourn-time distributions for the gated model with $\rho = 0.95$ and LCFS service.

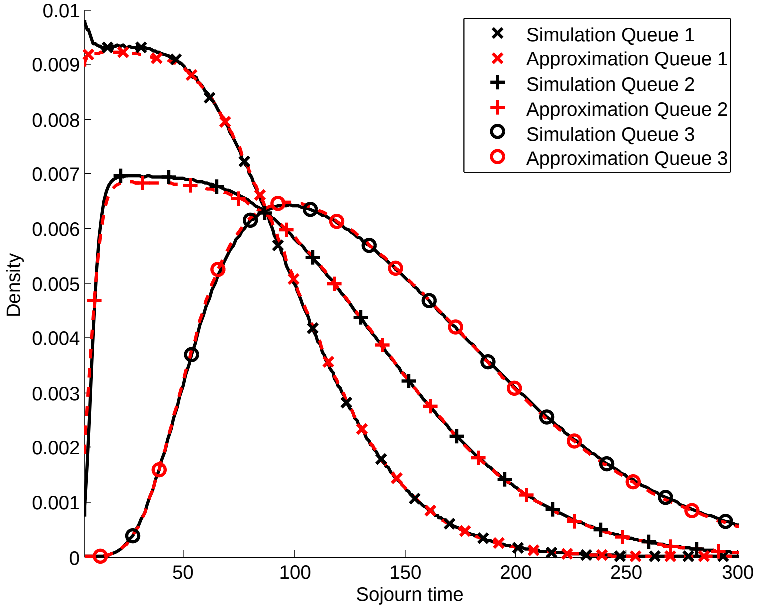

Figure 6: Simulated and approximated per-queue sojourn-time distributions for the globally-gated model with $\rho = 0.95$ and LCFS service.

difference defined as

$$\Delta\% = 100\% \times \frac{|\text{App} - \text{Sim}|}{\text{Sim}}$$

for different values of ρ and for all the scheduling disciplines considered in this paper. Recall that the mean sojourn-times are the same for FCFS, LCFS and ROS service, but may differ for PS and SJF service. Table 2 shows the results for the standard deviations of the sojourn times at Q_1. In Table 1 we see that the approximation of the mean sojourn time is most accurate for lightly and heavily loaded systems. This is due to the fact that, by construction, the approximations are asymptotically exact in the limiting cases of $\rho \downarrow 0$ and $\rho \uparrow 1$. For moderately loaded systems, the error is highest, but it still is no more than a few percent. Table 2 shows that the results for the standard deviations are accurate for heavily loaded systems, but may become less accurate for low-to-medium loaded systems. This is probably caused by the fact that the approximation for the second (and higher) moments of the sojourn times in (34)-(35) is asymptotically exact for $\rho \uparrow 1$, but not for $\rho \downarrow 0$ (as opposed to the first moments, which are asymptotically exact for $\rho \downarrow 0$

	FCFS/LCFS/ROS			PS			SJF		
ρ	Sim	App	Δ%	Sim	App	Δ%	Sim	App	Δ%
0.1	4.92E00	4.97E00	1.1	4.92E00	4.99E00	1.3	4.92E00	4.97E00	0.9
0.3	5.75E00	6.02E00	4.7	5.75E00	6.07E00	5.6	5.75E00	5.99E00	4.2
0.5	7.29E00	7.87E00	7.9	7.30E00	7.98E00	9.2	7.28E00	7.80E00	7.1
0.7	1.13E01	1.21E01	6.9	1.14E01	1.23E01	8.0	1.13E01	1.19E01	6.1
0.8	1.65E01	1.74E01	5.0	1.68E01	1.78E01	5.8	1.64E01	1.71E01	4.4
0.9	3.22E01	3.31E01	2.6	3.24E01	3.40E01	5.0	3.18E01	3.25E01	2.2
0.95	6.37E01	6.45E01	1.3	6.53E01	6.63E01	1.5	6.26E01	6.33E01	1.1
0.98	1.58E02	1.59E02	0.4	1.62E02	1.63E02	0.6	1.55E02	1.56E02	0.5

Table 1: Mean sojourn times for different scheduling disciplines.

by construction).

In summary, the numerical results (1) illustrate the validity of the asymptotic results, and (2) demonstrate that the sojourn-time approximations nicely capture the impact of the local scheduling policies on the sojourn-time distributions and are accurate over the whole range of load values.

6 Concluding remarks

In this paper, we have studied the impact of scheduling within queues on the waiting-time and sojourn-time distributions in polling systems. We have presented the first HT analysis of polling models where the local scheduling policy is not FCFS, but instead, is varied as LCFS, ROS, PS and SJF. The main contribution of the paper is the derivation of asymptotic closed-form expressions for the LST of the scaled waiting-time and sojourn-time distributions under HT conditions. The results raise a number of remarks and challenging open questions for further research, on which we would like to elaborate in the current section.

In this paper we have assumed that *all* the queues in the polling system follow the (globally) gated service discipline. However, this assumption can easily be relaxed; that is, we only have to assume that the specific queue for which we derive the waiting-time distribution is served according to the gated service discipline (see, also, [7]). For all the other queues, we only have to postulate that the service discipline belongs to the broad class of local branching-type disciplines [18], which includes gated and exhaustive service as special cases.

Furthermore, as [7] argues, the analysis of exhaustive polling systems is more complicated because the waiting times of the customers who are served during a visit are affected by later arrivals which take place during that visit period (which is obviously not the case for gated systems). Extension of the results to a broader class of service disciplines is a challenging topic for further research.

Finally, an interesting question is a generic optimization of the system's performance with respect to the choice of the local scheduling disciplines. With respect to mean sojourn times, it follows from [30] that SJF is optimal. For non-anticipating scheduling disciplines, the results in [1] suggest that the optimal discipline for minimizing mean sojourn times belongs to the family of multilevel PS disciplines. Optimization results beyond the mean, e.g. in terms of tails of sojourn times, is still open. In this context, it is worthwhile to note that the sojourn-time distribution at a given queue does not depend on the choice of the local scheduling discipline at any other queue. This implies that the sojourn-time distribution at a queue only depends on the choice of the local service order at that same queue. Therefore, the results presented in this paper provide a good starting point for tackling this type of optimization problem.

	FCFS			ROS			SJF		
ρ	Sim	App	Δ%	Sim	App	Δ%	Sim	App	Δ%
0.1	3.44E00	2.63E00	23.4	3.44E00	2.78E00	19.1	3.44E00	2.81E00	18.1
0.3	3.93E00	3.30E00	15.9	3.93E00	3.49E00	11.2	3.93E00	3.52E00	10.5
0.5	4.89E00	4.49E00	8.1	4.94E00	4.75E00	3.8	4.93E00	4.77E00	3.1
0.7	7.49E00	7.24E00	3.4	7.71E00	7.66E00	0.6	7.68E00	7.65E00	0.3
0.8	1.08E01	1.07E01	1.8	1.13E01	1.13E01	0.1	1.12E01	1.13E01	0.3
0.9	2.11E01	2.09E01	0.9	2.21E01	2.21E01	0.1	2.19E01	2.19E01	0.0
0.95	4.14E01	4.13E01	0.3	4.37E01	4.37E01	0.1	4.34E01	4.35E01	0.1
0.98	1.03E02	1.03E02	0.3	1.09E02	1.09E02	0.0	1.08E02	1.08E02	0.3

Table 2: Standard deviations of the sojourn times for FCFS, ROS and SJF.

References

[1] Aalto, S., Ayesta, U. and Righter, R. (2011). Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences* **25**, 269–288.

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

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

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

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

[6] Boon, M.A.A., Winands, E.M.M., Adan, I.J.B.F., and Van Wijk, A.C.C. (2010). Closed-form waiting time approximations for polling systems. *Performance Evaluation* **68**, 290–306.

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

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

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

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

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

[12] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Transactions on Communications* **38**, 1750–1760.

[13] Mack, C., Murphy, T. and Webb, N. (1957). The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants. *Journal of the Royal Statistical Society Series B* **19**, 166–172.

[14] Mack, C., (1957). The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable. *Journal of the Royal Statistical Society Series B* **19**, 173–178.

[15] Olsen, T.L. and Van der Mei, R.D. (2003). Periodic polling systems in heavy-traffic: distribution of the delay. *Journal of Applied Probability* **40**, 305–326.

[16] Olsen, T.L. and Van der Mei, R.D. (2005). Periodic polling systems in heavy-traffic: renewal arrivals. *Operations Research Letters* **33**, 17–25.

[17] Olvera-Cravioto, M., Blanchet, J. and Glynn, P. (2011). On the transition from heavy traffic to heavy tails for the M/G/1 queue: the regularly varying case. *Annals of Applied Probability* **21**, 645–668.

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

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

[20] Takagi, H. (1990). Queueing analysis of polling models: an update. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 267–318.

[21] Takagi, H. (1991). Application of polling models to computer networks. *Computer Networks and ISDN Systems* **22**, 193-211.

[22] Takagi, H. (1997). Queueing analysis of polling models: progress in 1990-1994. In: *Frontiers in Queueing: Models and Applications in Science and Technology*, ed. J.H. Dshalalow (CRC Press, Boca Raton, FL), 119–146.

[23] Tijms, H.C. (2003). *Stochastic Models: an Algorithmic Approach.* Wiley.

[24] Van der Mei, R.D. (1999). Delay in polling systems with large switch-over times. *Journal of Applied Probability* **36**, 232–243.

[25] Van der Mei, R.D. (1999). Distribution of the delay in polling systems in heavy traffic. *Performance Evaluation* **31**, 163–182.

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

[27] Van der Mei, R.D. (1999). Polling systems in heavy traffic: higher moments of the delay. *Queueing Systems* **31**, 265–294.

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

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

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

[31] Williams, D. (1991). *Probability with Martingales.* Cambridge University Press.

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