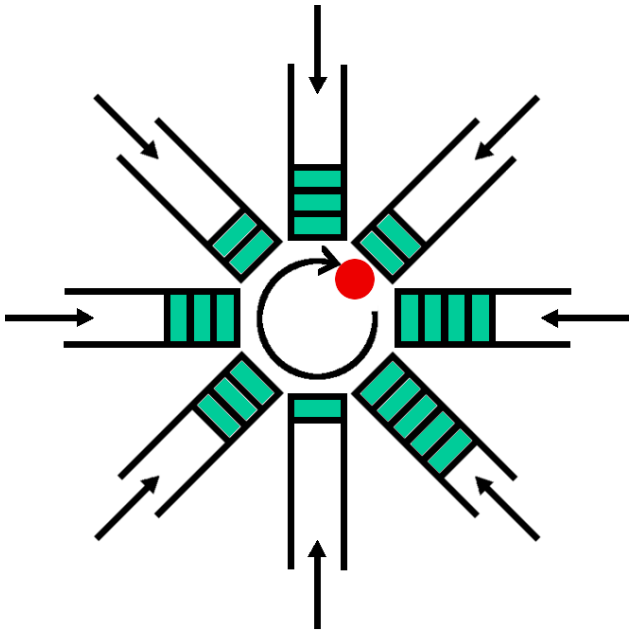
Waiting-time distributions in polling systems with non-FCFS service policies



INTERNSHIP REPORT

Author:
Petra Vis

Supervisor CWI:
prof. dr. R.D. van der Mei

Supervisors VU:
dr. R. Bekker
dr. A. Roubos

Waiting-time distributions in polling systems with non-FCFS service policies

INTERNSHIP REPORT

Author:
Petra Vis

August 2012

Centrum Wiskunde & Informatica

Research Group PNA2
Science Park 123
1098 XG Amsterdam

Supervisor CWI:
prof. dr. R.D. van der Mei

Vrije Universiteit Amsterdam

Faculty of Sciences
De Boelelaan 1081a
1081HV Amsterdam

Supervisors VU:
dr. R. Bekker
dr. A. Roubos

Preface

The end of the BMI (Business Mathematics and Informatics) Master program at the VU University in Amsterdam is marked by an internship carried out at an external business, industry or research facility. The present report contains the results of my internship done at the PNA2 (Probability & Stochastic Networks) research group of the 'Centrum Wiskunde & Informatica' (CWI) in Amsterdam.

I would like to thank my supervisor Rob van der Mei for giving me the opportunity to do the internship at the CWI. I would also like to thank him, Rene Bekker (supervisor VU), Jan-Pieter Dorsman and Erik Winands for their help, advice and support during the internship. Thanks are also due to Alex Roubos (second reader VU). Finally I want to thank my colleagues at CWI for providing a great working atmosphere.

Abstract

Throughout this report, polling systems play a central role. Polling systems are queueing systems consisting of multiple queues, attended by a single server. The server can only serve one queue at a time. Whenever the server moves from one queue to another, a stochastic, non-zero switch-over time is incurred. The server never idles; even when there are no customers waiting in the system, the server keeps moving between queues.

In the literature on these systems, often the First-Come-First-Served service order is assumed. We study polling systems with the following service orders: Last-Come-First-Served, Random Order of Service, Shortest Job First and Processor Sharing. The service discipline in the systems are gated or globally gated. For every service order, the distribution of the waiting time in heavy traffic is derived and used to obtain an approximation that is valid for all loads. This gives fundamental insight in the impact of the local service order.

The main result of the report is the fact that the distribution of the waiting time in polling systems can be approximated by a generalized trapezoidal distribution times a gamma distribution. For FCFS it is already known that the distribution is a uniform times a gamma distribution. For LCFS the waiting time distribution is also a uniform times a gamma distribution, the uniform distribution has different parameters, but the mean is the same. If the service order is ROS, the waiting time distribution is a trapezoidal distribution times a gamma distribution, the mean of the trapezoidal distribution is equal to the means of the uniform distributions found for FCFS and LCFS. Uniform and trapezoidal distributions are special cases of the generalized trapezoidal distribution. In systems with PS and SJF queues, the waiting time distribution is a generalized trapezoidal distribution times a gamma distribution. The probability density function of the generalized trapezoidal distribution depends on the service time distribution.

In general the approximation works best for systems with a load larger than 0.8, a large number of queues or Poisson arrivals. The least accurate performance of the approximation is found when the load of the system lies between 0.3 and 0.5 or when the squared coefficient of variation of the interarrival time distribution is large. In practice these characteristics are uncommon. The just-in-time philosophy dictates that the demand is stable, interarrival time distributions with large SCVs are hardly found. Also, these systems are typically utilized beyond $\rho = 0.5$ to increase productivity. This means that the approximation is applicable in many practical applications.

Contents

Chapter 1

Introduction

1.1 Polling systems

A polling system consists of multiple queues and a single server that visits the queues in a fixed order. While moving from one queue to the next, a switch-over time is incurred. Figure 1.1 gives a graphical representation of a polling system. Typical features of a polling model are discussed below. It is a summary of [3].

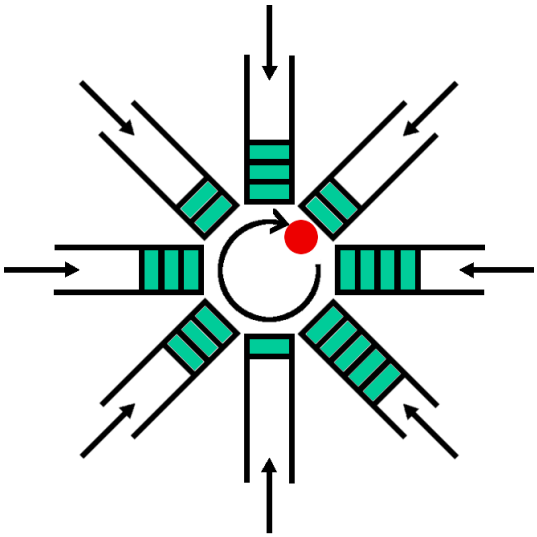


Figure 1.1: Polling system with cyclic service

Arrival processes. The arrival processes of the queues are often assumed to be mutually independent homogeneous Poisson processes. In many cases this is a realistic assumption, e.g., in the case of telephone calls and traffic accidents. Due to the memoryless property of the Poisson process, polling models with Poisson arrivals often allow for a tractable analysis. Yet, in many situations the assumption of Poisson arrivals is unrealistic, e.g., in case of bursty traffic (voice, video) and several production applications.

Buffer size. In most polling models the buffer size is assumed to be infinite.

Service-time distributions. The service times at a queue are typically assumed to be samples from a probability distribution which is characteristic for that queue. The service times

are usually assumed to be mutually independent and independent of the actual state of the system.

Switch-over process. The switch-over times needed by the server to move from one queue to another queue are typically assumed to be samples from a prespecified probability distribution which is characteristic for that pair of queues.

Server routing. A routing scheme determines the order in which the queues are served. The traditional routing mechanism is the cyclic server routing. To model systems in which particular queues are visited more frequently than others, cyclic polling has been extended to periodic polling, in which the server visits the queue periodically according to some service order table of finite length. Other routing schemes are: probabilistic (proceed to queue j with probability p_j), Markovian (proceed from queue i to queue j with probability p_{ij}) and dynamic (e.g. serve longest queue).

Service discipline. The service discipline specifies how many customers are served during one visit of the server to a queue. The most commonly used service disciplines are:

Exhaustive Serve the queue until it is empty.

Gated Serve all customers that were present at "polling instant", i.e., at the moment when the server arrives at the queue.

Globally gated During a cycle, serve all customers that were present in the system at "polling instant" of the first queue.

K-limited Serve at most K customers or until the queue is empty.

Time-limited Serve at most t time units or until the queue is empty.

Queueing discipline. The queueing discipline specifies the order in which the customers present at the same queue are served. The most common queueing discipline is the classical *First-Come-First-Served* (FCFS) discipline. Other service disciplines are *Last-Come-First-Served* (LCFS), *Processor Sharing* (PS), *Random Order of Service* (ROS) and *Shortest Job First* (SJF). Provided the queueing discipline does not depend on the service times, the queue length distribution is independent of the service order, and so are, by Little's law, the mean waiting times. However, the distribution of the waiting time does depend on the queueing discipline.

1.1.1 Applications

Applications of polling systems can be found in communication networks, production systems, traffic and transportation problems and various other fields of engineering. In [3] the main applications in these areas are described. Here we give a short overview of the these applications.

In communication networks different terminals compete for access to a shared medium. If multiple terminals transmit or receive data over the medium at the same time, packet collisions and interference problems may occur. Motivated by this, many medium access control (MAC) protocols have been proposed for different network technologies, in many cases leading to polling models.

An example is a token-ring network. A token-ring network can be characterized as a set of stations connected to a common transmission medium in a ring topology. All messages travel over a fixed route from station to station around the loop. A token can be in two states: occupied or

free. A station with data to transmit reads the free token and changes it to the occupied state before retransmitting it. The occupied token is then incorporated as part of the header of the data transmitted on the ring by the station. Thus, other stations on the ring can read the header, note the occupied token and refrain from transmission. When the token is back at the station that changed it to the occupied state and the station decides to transfer the right for transmission to another station, it changes the token to the free state. The token-ring network allows the transmission of packets in a conflict-free manner. Figure 1.2 illustrates the token-ring network. To model the network as a polling system, the token is represented by the server and the traveling from station to station is represented by the switching of the server. The packets that are send by the stations represent the customers.

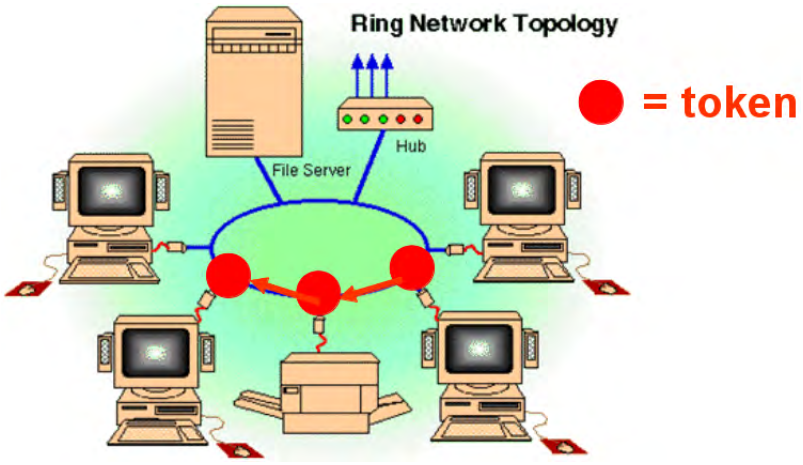


Figure 1.2: A token-ring network

In production systems polling models can be found when a production facility is used to produce multiple different types of products, but only one product at a time. Every product has its own queue where orders arrive. The facility produces the different products in a prespecified order, during the switch from one product to another, a setup time is incurred. Demands, setup times (switch-over times) and production times can be random.

The third important area in which polling systems are frequently applied in practice are traffic and transportation systems. In road traffic a polling system is the natural way to model a situation where queues arise due to the fact that multiple flows of traffic have to share one single lane. The plainest form is a two-way road which is partly blocked because of an accident or road maintenance, but any common traffic intersection also qualifies. Besides road traffic, polling models are employed in transportation systems consisting of driverless electric vehicles that follow a predetermined track. The actual transportation is performed by so-called Automated Guided Vehicles (AGVs). In a conventional AGV system, each vehicle can pick up a load from any station and deliver it to any other station. In order to avoid collisions between vehicles, most systems use the zone blocking concept where the entire system is divided into zones. The control system allows only one vehicle in each zone at a time. Whenever a AGV system consists of a single loop, it can be modeled as a polling system. The vehicle corresponds to the server in the polling system and the stations form the queues of the system. In some cases each station is modeled as two queues, one for the picking-up, and the other one for the drop-off. The inter-station travel times are modeled as switch-over times in the polling model. AGV systems are used in warehouses and container terminals, but also in other application areas like, e.g., health care.

Some examples of other applications of polling models are maintenance, elevators and health

care. For maintenance a polling model is used to describe a patrolling repairman who inspects breakdowns. The repairman is represented by the server, the breakdowns are represented by the customers and the times needed by the repairman to travel from one machine to the next are represented by the switch-over times. For elevators we have the following analogies: server $\leftrightarrow$ elevator; queues $\leftrightarrow$ floors. In health care, there is an example where a medical emergency room is modeled via a multi-server polling model with priority processing [7]. At the emergency room different types of patients arrive and are placed in queues depending on the type of surgical procedure required. Each emergency surgical procedure has a separate queue of infinity buffer capacity. The number of emergency room theaters is limited, and each emergency room theater is modeled by a server in the polling model. Since setting up for surgery procedures takes longer than the actual procedures themselves, the service times within the polling model are much smaller than the switch-over times. Finally, an urgency parameter in terms of average waiting time is assigned to each patient which dictates the existence of (local) priority levels within each queue.

1.1.2 Basic polling models

A basic, standard polling model has Poisson arrival processes and general distributions for the service and switch-over times. The routing scheme of a basic polling system is cyclic and the service policies at the queues are either gated, or exhaustive (mixtures are allowed). For this type of polling systems, there exists an easy to interpret expression for the expected waiting time at a queue. However, to use the formula, the second moments of the cycle time and the residual cycle time need to be calculated, which is not straightforward. A frequently used alternative to find the mean waiting times is the *Descendant set approach*, described in [14].

Boxma and Groenendijk [5] derive a so called *pseudo-conservation law* (PCL) for mean waiting times under various service disciplines. This PCL states that the weighted sum of the waiting times is equal to a constant. Pseudo-conservation laws are useful for obtaining, or testing, approximations for individual mean waiting times. They can also be used to study asymptotics, yielding information about what happens when the number of queues becomes very large or when the offered traffic at a particular queue approaches its stability limit.

There is a sharp distinction between polling models that are "easy" and models that are "complex". It turns out that, a few exceptions aside, an exact analysis is only possible if all service disciplines in the polling system satisfy the following *branching property* described by Fuhrman [13] and Resing [16]:

Property 1. *Suppose at the beginning of a visit at queue i there are k_i customers present at the queue. Then at the end of that visit to the queue, each of these k_i customers have been effectively replaced by some population of customers in an i.i.d. manner.*

The gated and exhaustive service disciplines satisfy this branching property, while the K-limited and time-limited service disciplines generally do not. Borst [4] gives a slight extension of Property 1 that is also held by a globally gated polling system:

Property 2. *Suppose at the beginning of a visit at queue $Q_{\pi(i)}$ there are k_i customers present at Q_i with $\pi(i) \in \{1,\ldots,N\}$. Then at the end of that visit to the queue, each of these k_i customers have been effectively replaced by some population of customers in an i.i.d. manner.*

1.1.3 Generalizations of polling models

A more general polling model is a polling system with renewal arrivals instead of Poisson arrivals. For this type of system, Olsen and Van der Mei [15] derive an approximation for the distribution of the waiting time in heavy traffic, that is, when the load of the system tends to one. Using heavy traffic (HT) limits and light traffic (LT) limits, Boon et al. [2] construct a closed-form approximation for the mean waiting time, that works well for all workloads. Dorsman et al. [11] combine these two approximations, to derive a simple closed-form approximation for the complete waiting time distribution that works well for arbitrary load values. This method is given in Section 2.2.4.

Instead of FCFS, the queueing discipline can be, for example, LCFS, PS, ROS or SJF. For each of these disciplines, Boxma et al. [6] determine the Laplace-Stieltjes transform of the sojourn time at a queue in case of Poisson arrivals and gated and globally gated service disciplines.

Iterative methods exist to find the queue length distribution in case of K-limited [21] or time-limited [12] service disciplines.

1.1.4 Existing methods for polling models

Several numerical approaches have been proposed for computing mean waiting times in general continuous-time polling systems with either exhaustive or gated service.

One such method is the *buffer occupancy method*, first used by Cooper and Murray [9] and Cooper [8]. This method is based on the buffer occupancy variables $X_{i,j}$, which denote the queue length at queue j at a polling instant of queue i, $i, j = 1,2,\ldots,N$. The buffer occupancy method requires the solution of N^3 linear equations with unknowns $\mathbb{E}[X_{i,j}X_{i,k}]$ to compute the mean waiting times in all N stations simultaneously. An advantage of this technique is its applicability to several variations of the basic polling model, that still satisfy the branching property, such as polling systems with other service disciplines than exhaustive and/or (globally) gated service, systems with simultaneous arrivals, or fluid polling systems. However if the assumptions described in Section 1.1.2 are not all valid, one might have to resort to alternative ways to analyze the system. For example, if arrival processes are not Poisson, but batch Markovian arrival processes (BMAPs), a generalization of the buffer occupancy method using Kronecker product notation can be used [17]. Altman and Fiems [1] show how stochastic recursive equations can be used to analyze a polling model with correlated switch-over times. If the service disciplines do not satisfy the branching property, one may have to resort to different techniques or even approximations.

Based on the buffer occupancy method, Konheim, Levy and Srinivasan [14] developed the *descendant set approach*; an iterative technique that computes the mean waiting time at each queue independently of the other queues. The descendant set approach is based on counting the number of descendants of each customer in the system [22].

In [23] an approach is developed to compute the mean waiting times for exhaustive-type or gated-type polling systems in a pure probabilistic manner. They derive a set of N^2 linear equations for these delay figures in case of exhaustive service and $N(N+1)$ linear equations in case of gated service. The equations are derived with help of the following basic queueing results: (i) the PASTA property, i.e., Poisson arrivals see time averages and (ii) Little's Law. The unknowns in the equations are $\mathbb{E}[L_{i,j}]$, the mean queue length at queue i at an arbitrary epoch within a visit time of queue j. The method can be looked upon as a *mean value analysis* (MVA) for general polling

systems with exhaustive or gated service. MVA is known as a powerful tool to determine mean performance measures in all kinds of queueing models.

1.2 About this report

In the previous section several methods to compute the mean waiting time in various polling systems are discussed. The focus of this report lies on the distribution of the waiting time. In Chapter 2 the model is described and notation is introduced. The second section of this chapter summarizes the literature on waiting time distributions, heavy traffic and light traffic limits and approximations for both the mean waiting time and the waiting time distribution in polling systems with a FCFS queueing discipline. Chapter 3 derives approximations of the waiting time distributions in polling systems with gated service for the LCFS, ROS, SJF and PS qeueuing disciplines. For every queueing discipline, the heavy traffic limits are calculated, an approximation for general loads is deduced and finally the performance of the approximations is studied by means of simulation. In Chapter 4, the same is done for polling systems with globally gated service and FCFS, LCFS or ROS queueing disciplines. Finally, Chapter 5 gives a brief conclusion and some suggestions for further research.

Chapter 2

Model description

In this chapter the considered polling systems are specified and notation is introduced. The various methods of deriving expected waiting times and waiting time distributions are stated and explained. These results will be used in later chapters to derive approximations for the waiting time distributions for polling systems with different queueing disciplines.

2.1 Notation

We consider a system of $N \geq 2$ infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server that visits and serves the queues in cyclic order. At every queue, the service discipline is gated or globally gated. Customers arrive at Q_i according to a renewal process with rate $\lambda_i = \frac{1}{\mathbb{E}[A_i]}$, where A_i is the random variable describing the interarrival times of customers at Q_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 Laplace-Stieltjes transform (LST) $B_i^*(\cdot)$ and finite kth moment $\mathbb{E}[B_i^k]$, $k = 1, 2, \ldots$. 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 $\rho = \sum_{i=i}^{N} \rho_i$. A necessary and sufficient condition for stability of the system is $\rho < 1$. Let W_i, with LST $W_i^*(\cdot)$, denote the waiting time of an arbitrary customer at Q_i. T_i, with LST $T_i^*(\cdot)$, denotes the sojourn time of an arbitrary customer at Q_i. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is a random variable S_i with mean $\mathbb{E}[S_i]$ and LST $S_i^*(\cdot)$. Let $S = \sum_{i=1}^{N} S_i$ denote the total switch-over time in a cycle. C_i is a random variable describing the cycle time of the server from Q_i to Q_i. $C_i^*(\cdot)$ denotes the LST of the cycle time at queue i. The mean cycle time does not depend on the queue, so $\mathbb{E}[C_i] = \mathbb{E}[C]$.

A value of a variable x evaluated at $\rho = 1$ is denoted by $\hat{x}$. The residual length of a random variable X is denoted by X^{res} with $\mathbb{E}[X^{res}] = \frac{\mathbb{E}[X^2]}{2\mathbb{E}[X]}$. The squared coefficient of variation (SCV) of a random variable X is denoted by c_X^2. We define $\sigma^2 = \sum_{i=1}^{N} \hat{\lambda}_i (\mathrm{Var}[B_i] + c_{A_i}^2 \mathbb{E}[B_i]^2)$ and $\delta = \sum_{i=1}^{N} \hat{\rho}_i (1 + \hat{\rho}_i)$. When the arrivals are Poisson, σ^2 simplifies to $\sigma^2 = \mathbb{E}[B^2]/\mathbb{E}[B]$ The notation $\rightarrow_d$ means convergence in distribution and $\mathbf{1}_{\{A\}}$ denotes the indicator function on the event A. The subscript a is used when the variable is an approximation, i.e. the variable X is approximated by X_a. When a random variable X is said to have a gamma distribution with shape parameter α and inverse scale parameter μ, its density function is given by $f_X(x) = e^{-\mu x} \mu^\alpha x^{\alpha-1} \mathbf{1}_{\{x \geq 0\}} / \Gamma(\alpha)$, where $\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} \mathrm{d}x$. The LST of X is then given by $\mathbb{E}\left[e^{-sX}\right] = (\mu/(\mu+s))^\alpha$.

2.2 Literature overview

In this section formulas from the literature that are used later in this report are stated and explained.

2.2.1 Poisson arrivals and various queueing disciplines

For a gated and globally gated service discipline and Poisson arrivals, Boxma et al. [6] derive Laplace-Stieltjes transforms of the waiting time distribution for various queueing disciplines.

Gated service discipline

When the queueing discipline is FCFS, we have for the LST of the waiting time at Q_i

$$W^*_{i,FCFS}(s) = \frac{C^*_i(\lambda_i(1-B^*(s))) - C^*_i(s)}{\mathbb{E}[C](s-\lambda_i(1-B^*_i(s)))} \quad (i=1,\ldots,N). \tag{2.1}$$

Note that we need to know the Laplace-Stieltjes transforms of the cycle time and the service time. The LST of the cycle time is derived in [6], but the formula is not easy to use.

The expected waiting time is given by:

$$\mathbb{E}[W_{i,FCFS}] = \mathbb{E}[C^{res}](1+\rho_i). \tag{2.2}$$

When the queueing discipline is LCFS, they have

$$W^*_{i,LCFS}(s) = \frac{1-C^*_i(s+\lambda_i(1-B^*_i(s)))}{\mathbb{E}[C]\,(s+\lambda_i(1-B^*_i(s)))} \quad (i=1,\ldots,N). \tag{2.3}$$

The expected waiting time is equal to the expected waiting of the FCFS system, given in (2.2).

The equation of the LST for a ROS policy is not as nice as the previous two. To compute this LST, first an order mark x is assigned to every customer; this order mark is a realization from a uniform distribution on $[0,1]$. The customers are then served in order of their marks. Because the service policy is gated, this coincides with random order of service. The LST can be computed conditionally on x, yielding

$$W^*_{i,ROS}(s|x) = \frac{C^*_i(\lambda_i x(1-B^*_i(s))) - C^*_i(s+\lambda_i x(1-B^*_i(s)))}{s\,\mathbb{E}[C]} \quad (i=1,\ldots,N). \tag{2.4}$$

The expected waiting time is given in (2.2).

When the queueing discipline is PS, a conditional LST of the waiting time can also be given in terms of the LST of the cycle time. When x is the amount of work that a tagged customer brings into the system, this gives

$$W^*_{i,PS}(s|x) = \frac{C^*_i(\lambda_i(1-\varphi(s,x))) - C^*_i(s+\lambda_i(1-\varphi(s,x)))}{s\,\mathbb{E}[C]} \quad (i=1,\ldots,N). \tag{2.5}$$

where $\varphi(s,x) = \mathbb{E}\left[e^{-s\min(B_i,x)}\right]$, the LST of the minimum of B_i and x. Before we give the unconditional first moment of the sojourn time, let us first consider the finite collection of i.i.d. random variables $\{B_{i,k}\}_{k=1}^n$, where $B_{i,1}$ is equal in distribution to a typical amount of work that is brought

by the customer that visits Q_i. We let $B_{i,k:n}$ denote the kth smallest value among this collection of size n. The first moment of the sojourn time is then given by

$$\mathbb{E}[T_{i,PS}] = \mathbb{E}[B_i] + \mathbb{E}[C^{res}](1 + 2\lambda_i \mathbb{E}[B_{i,1:2}]), \tag{2.6}$$

For a SJF policy, Equation (2.5) also holds, but with $\varphi(s,x) = \mathbb{E}\left[e^{-sB_i \mathbf{1}_{\{B_i \leq x\}}}\right]$. The expected waiting time is

$$\mathbb{E}[W_{i,SJF}] = \mathbb{E}[C^{res}](1 + \lambda_i \mathbb{E}[B_{i,1:2}]). \tag{2.7}$$

Globally gated service discipline

With a locally gated discipline, a customer arriving at queue i has to wait a residual cycle until the server reaches his queue and additionally he has to wait for the customers that are served before him. With a globally gated service discipline, the gate closes at the start of a cycle. With the globally gated discipline, this customer has to wait a residual cycle for the server to reach the first queue and the gate to open. Additionally he has to wait until the customers are served that arrived in the same cycle as he did in the queues before his queue (queues $1, \ldots, i-1$). He also has to wait the switch-over times $S_1, \ldots, S_{i-1}$ and the service times of the customers at his queue that are served before him.

When the queueing discipline is FCFS, the LST of the waiting time distribution at Q_i is given by

$$W^*_{i,FCFS}(s) = \frac{1}{\mathbb{E}[C]} \frac{C_i^*\left(\sum_{j=1}^{i} \lambda_j (1 - B_j^*(s))\right) - C_i^*\left(\sum_{j=1}^{i-1} \lambda_j (1 - B_j^*(s)) + s\right)}{s + \lambda_i (1 - B_i^*(s))} \prod_{j=1}^{i-1} S_j^*(s). \tag{2.8}$$

The expected waiting time is

$$\mathbb{E}[W_{i,FCFS}] = \mathbb{E}[C^{res}] \left(2 \sum_{j=1}^{i-1} \rho_j + \rho_i + 1\right) + \sum_{j=1}^{i-1} \mathbb{E}[S_j]. \tag{2.9}$$

The LST of the waiting time in a LCFS queue is

$$W^*_{i,LCFS}(s) = \frac{C_i^*(\sum_{j=1}^{i-1} \lambda_j (1 - B_j^*(s))) - C_i^*(\sum_{j=1}^{i} \lambda_j (1 - B_j^*(s)) + s)}{(\lambda_i (1 - B_i^*(s)) + s)\, \mathbb{E}[C]} \prod_{j=1}^{i-1} S_j^*(s). \tag{2.10}$$

The expected waiting time is given in (2.9).

When the service order is random, the LST of the waiting time is

$$W^*_{i,ROS}(s) = \frac{1}{\mathbb{E}[C] \lambda_i (1 - B_i^*(s))} \int_{X_i(s)}^{X_{i+1}(s)} \frac{C_i^*(y) - C_i^*(y+s)}{s} \, \mathrm{d}y \prod_{j=1}^{i-1} S_j^*(s), \tag{2.11}$$

with

$$X_i(s) = \sum_{j=1}^{i-1} \lambda_j (1 - B_j^*(s)).$$

The expected waiting time is given in (2.9).

We don't give the LST of the waiting time for systems with SJF or PS queueing disciplines, because they are not used in this report. They can be found in [6].

2.2.2 Heavy traffic

Gated service discpline

When the load of a polling system tends to 1, the system is in heavy traffic. In the heavy traffic limit with FCFS queueing disciplines, a Laplace-Stieltjes transform of the waiting time distribution is known and given in [19] for Poisson arrivals and both gated and exhaustive service disciplines. When $\rho \uparrow 1$, all queues become unstable, therefore the focus lies on the random variable $(1-\rho)W_i$ (referred to as the *scaled* delay at Q_i). Define

$$\tilde{W}_i := \lim_{\rho\uparrow 1}(1-\rho)W_i \quad (i=1,\ldots,N). \tag{2.12}$$

The same holds for the cycle time, so we have for the scaled cycle time

$$\tilde{C}_i := \lim_{\rho\uparrow 1}(1-\rho)C_i \quad (i=1,\ldots,N).$$

For the gated service discipline the closed-form expression for the LST of $\tilde{W}_i$ that is derived in [19] is given by

$$\tilde{W}^*_{i,FCFS}(s) = \frac{1}{\mathbb{E}[S]\,s(1-\hat{\rho}_i)}\left\{\left(\frac{\mu}{\mu+s\hat{\rho}_i}\right)^{\alpha} - \left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\}, \tag{2.13}$$

with

$$\alpha := \frac{\mathbb{E}[S]\,\delta}{\sigma^2}, \quad \mu := \frac{\delta}{\sigma^2}. \tag{2.14}$$

This expression can be found using the fact that in heavy traffic the *scaled* cycle times are gamma distributed with α and μ as given in Equation (2.14). The LST of the *unscaled* cycle time in heavy traffic is then given by

$$\begin{aligned} C^*_i(s) = \mathbb{E}[e^{-s\frac{\tilde{C}_i}{1-\rho}}] &= \int_0^{\infty} f_{\tilde{C}_i}(x)e^{\frac{-sx}{1-\rho}}\,\mathrm{d}x \\ &= \int_0^{\infty} \mu^{\alpha}x^{\alpha-1}e^{-x(\mu+s/(1-\rho)}\,\mathrm{d}x = \left(\frac{\mu(1-\rho)}{\mu(1-\rho)+s}\right)^{\alpha}. \end{aligned} \tag{2.15}$$

So, the parameters of the cycle time distribution C_i in heavy traffic are α and $\mu(1-\rho)$.

The expected cycle time is $\mathbb{E}[C] = \mathbb{E}[S]/(1-\rho)$. Now $\tilde{W}^*_{i,FCFS}(s)$ can be computed using

$$\tilde{W}^*_{i,P}(s) = \lim_{\rho\uparrow 1} W^*_{i,P}(s(1-\rho)), \tag{2.16}$$

where $W^*_{i,P}(s)$ is given in Section 2.2.1 and P denotes the queueing policy.

The parameters above are taken from [15] where σ^2 is used instead of $\mathbb{E}[B]/\mathbb{E}[B^2]$. These parameters also hold in case of renewal arrivals.

In [15] a strong conjecture is provided for the limiting waiting-time distribution. This conjecture holds for a general parameter setting when the load tends to 1. For a FCFS gated policy, this conjecture is as follows.

Conjecture 1.

$$\tilde{W}_i \to_d U\tilde{\mathbf{C}}_i \quad i = 1,\dots N,$$

where U is a uniform$[\hat{\rho}_i, 1]$ 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 (2.14)*.*

In words this means that the distribution of the scaled waiting time goes to a uniform times a gamma distribution in heavy traffic.

$\tilde{\mathbf{C}}_i$ is the length-biased version of $\tilde{C}_i$, a gamma distributed random variable with parameters α and μ as in Equation (2.14). If X is some random variable with probability density function (p.d.f.) $f_X(x)$ and finite expectation $\mathbb{E}[X]$ then the length-biased random variable $\mathbf{X}$ is defined as a random variable with p.d.f. $f_{\mathbf{X}}(x) = xf_X(x)/\mathbb{E}[X]$. It is straightforward to show that if a gamma random variable has parameters α and μ then its length-biased version has parameters $\alpha+1$ and μ.

Note that the parameters of the gamma distribution given in Conjecture 1 coincide with the length-biased variants of the parameters given in Equation (2.14).

To show that Equation (2.13) is indeed the LST of a uniform times a gamma distribution, take the LST of the distribution conditional on the uniform distribution and then uncondition. For a uniform(a,b) distribution times a gamma$(\alpha+1,\mu)$ distribution we have

$$\begin{aligned}
\int_a^b \frac{1}{b-a}\left(\frac{\mu}{\mu+sy}\right)^{\alpha+1} \mathrm{d}y &= \frac{1}{b-a}\int_{\frac{\mu}{\mu+sa}}^{\frac{\mu}{\mu+sb}} u^{\alpha+1}\frac{-\mu}{s}\frac{(sy+\mu)^2}{\mu^2}\,\mathrm{d}u \\
&= \frac{-\mu}{s(b-a)}\int_{\frac{\mu}{\mu+sa}}^{\frac{\mu}{\mu+sb}} u^{\alpha-1}\,\mathrm{d}u \\
&= \frac{-\mu}{s(b-a)}\left[\frac{1}{\alpha}u^{\alpha}\right]_{\frac{\mu}{\mu+sa}}^{\frac{\mu}{\mu+sb}} \\
&= \frac{-\mu}{\alpha s(b-a)}\left\{\left(\frac{\mu}{\mu+sb}\right)^{\alpha} - \left(\frac{\mu}{\mu+sa}\right)^{\alpha}\right\} \\
&= \frac{\mu}{\alpha s(b-a)}\left\{\left(\frac{\mu}{\mu+sa}\right)^{\alpha} - \left(\frac{\mu}{\mu+sb}\right)^{\alpha}\right\}.
\end{aligned}$$

For the first equality integration with substitution is used, the other equalities are basic calculations. Note that $\alpha/\mu = \mathbb{E}[S]$, $a = \hat{\rho}_i$ and $b = 1$, so (2.13) is indeed the LST of a uniform$(\hat{\rho}_i, 1)$ times a gamma$(\alpha+1,\mu)$ distribution.

The fact that the distribution of the scaled waiting time is a uniform times a gamma distribution can be intuitively explained. The gamma distribution represents the residual cycle time, the uniform distribution represents the fraction of the cycle time that a customer has to wait. A lucky customer arrives just before the server starts to serve his queue, his waiting time is a fraction $\hat{\rho}_i$ of the residual cycle time, the amount of work that arrived at Q_i during one residual cycle. An unlucky customer arrives when the server just leaves his queue, or right after the server starts serving his queue, in that case he will be behind the gate. His waiting time is then exactly one residual cycle. The waiting time of an arbitrary customer is a fraction between $\hat{\rho}_i$ and 1, represented by the uniform $(\hat{\rho}_i, 1)$ distribution times a gamma distributed random variable.

Globally gated service discipline

For a globally gated service discipline, the LST of the scaled waiting time in heavy traffic is known in case of FCFS service order with Poisson arrivals. It is derived in [20] and given by

$$\tilde{W}^*_{i,FCFS}(s) = \frac{1}{(1-\hat{\rho}_i)\,\mathbb{E}[S]s}\left\{\left(\frac{\mu}{\mu + s\sum_{j=1}^{i}\hat{\rho}_j}\right)^{\alpha} - \left(\frac{\mu}{\mu + s(1+\sum_{j=1}^{i-1}\hat{\rho}_j)}\right)^{\alpha}\right\}, \tag{2.17}$$

with

$$\alpha := \frac{2\,\mathbb{E}[S]}{\sigma^2} \quad \text{and} \quad \mu := \frac{2}{\sigma^2}. \tag{2.18}$$

These parameters coincide with the parameters given in Equation (2.14) with $\delta = 2$. Again, these parameters are adjusted, so that they also hold for renewal arrivals.

The expected waiting time at Q_i is given by the following expression

$$\mathbb{E}[\tilde{W}_{i,FCFS}] = \frac{1}{2}\left(1 + 2\sum_{j=1}^{i-1}\hat{\rho}_j + \hat{\rho}_i\right)\left(\frac{\sigma^2}{2} + \mathbb{E}[S]\right). \tag{2.19}$$

The form of (2.17) matches the form of Equation (2.13). This suggests that the distribution of the waiting time at queue i has a uniform times a gamma distribution, where the uniform distribution has boundaries $\sum_{j=1}^{i}\hat{\rho}_j$ and $(1+\sum_{j=1}^{i-1}\hat{\rho}_j)$ and the gamma distribution has parameters $\alpha+1$ and μ given in (2.18). The lower bound of the uniform distribution corresponds to a customer arriving at queue i just before the server arrives at queue 1. This customer has to wait for all the customers in queues 1 to i, all customers in queue i are served before him. The upper bound of the uniform distribution corresponds to a customer arriving just after the server arrives at queue 1. This customer has to wait a full residual cycle and for all the customers that are served in queues 1 to $i-1$, he is the first customer in the ith queue.

2.2.3 Light traffic

In light traffic, when the load of the system is close to 0, not much happens. This means that when a customer arrives at a queue in a gated system, he only has to wait until the server starts to serve that queue. The waiting time of a customer in light traffic is therefore equal to the residual switch-over time for a gated service policy. When the system has a globally gated policy, the customer has to wait until the server arrives at queue 1 to open the gates, this time is equal to the residual switch-over time. Then the customer has to wait until the server arrives at his queue, so the sum of the switch-over times from queue 1 to the queue that the customer arrived in. We have:

$$\text{For gated: } \mathbb{E}[W_i] = \mathbb{E}[S^{res}] \qquad i = 1,\ldots,N \tag{2.20}$$

$$\text{For globally gated: } \mathbb{E}[W_i] = \mathbb{E}[S^{res}] + \sum_{j=1}^{i-1}\mathbb{E}[S_j] \qquad i = 1,\ldots,N \tag{2.21}$$

In light traffic, the service order does not matter, the number of customers in the system will be very low and waiting lines never occur.

2.2.4 Interpolation

This section applies to the gated service discipline. These interpolation for systems with a globally gated polling regime, is introduced in Chapter 4.

Boon et al. [2] derived a closed-form approximation for the waiting time using interpolation between LT and HT limits. This approximation works well for general loads and renewal arrivals. The approximation is derived for both gated and exhaustive service policies. The expected waiting time $\mathbb{E}[W_{i,Boon}]$ of a customer at Q_i is approximated by the following function of ρ:

$$\mathbb{E}[W_{i,Boon}] = \frac{K_{0,i} + K_{1,i}\rho + K_{2,i}\rho^2}{1-\rho} \quad i = 1,\ldots,N. \tag{2.22}$$

If all queues receive gated service, the constants are given by:

$$\begin{aligned}
K_{0,i} =& \mathbb{E}[S^{res}],\\
K_{1,i} =& \hat{\rho}_i \left((c_{A_i}^2)^4 \mathbf{1}_{\{c_{A_i}^2 \le 1\}} + 2\frac{c_{A_i}^2}{c_{A_i}^2+1}\mathbf{1}_{\{c_{A_i}^2 > 1\}} - 1 \right) \mathbb{E}[B_i^{res}]\\
&+ \hat{\rho}_i \mathbb{E}[S^{res}] - \frac{1}{\mathbb{E}[S]} \sum_{j=0}^{N-1} \sum_{k=0}^{j} \hat{\rho}_{i+k} \mathrm{Var}[S_{i+j}],\\
K_{2,i} =& \frac{1+\hat{\rho}_i}{2} \left(\frac{\sigma^2}{\delta} + \mathbb{E}[S] \right) - K_{0,i} - K_{1,i}.
\end{aligned}$$

Combining the result above with the heavy traffic limits of the waiting time distribution, [11] provides a closed-form approximation of the waiting time distribution. The focus of that article lies on the case of *exhaustive* service, for which the authors show with use of simulation results that the approximation is highly accurate over a wide range of parameter settings. In case of *gated* service they derive the following approximation for the waiting time distribution:

$$\mathbb{P}[W_{i,FCFS} < x] \approx \mathbb{P}[UI_{i,a} < (1-\rho)x],$$

where U is a uniformly distributed random variable on $[\hat{\rho}_i, 1]$ and $I_{i,a}$ is a gamma distributed variable with parameters

$$\alpha_{ia} = \alpha_a = \frac{\mathbb{E}[S]\,\delta}{\sigma^2} + 1 \quad \text{and} \quad \mu_{ia} = \frac{1+\hat{\rho}_i}{2} \frac{\mathbb{E}[S]\,\delta + \sigma^2}{\sigma^2(1-\rho)\,\mathbb{E}[W_{i,Boon}]}. \tag{2.23}$$

Here $\mathbb{E}[W_{i,Boon}]$ is given in Equation (2.22), α_{ia} and μ_{ia} are uniquely determined by the following three requirements:

1. In HT the approximation must coincide with the diffusion approximation of [15] given in Conjecture 1, i.e.,
$$\frac{\alpha}{\alpha_{ia}} \to 1 \quad \text{and} \quad \frac{\mu_i}{\mu_{ia}} \to 1, \quad \text{when } \rho \uparrow 1.$$
2. The expectation of the approximation coincides with $\mathbb{E}[W_{i,Boon}]$ (Equation (2.22)), Boon's approximation of the expected waiting time.
3. The SCV of the approximating distribution matches the SCV of the HT diffusion approximation of [15], given in Conjecture 1, so that the shape of the refined diffusion approximation matches the shape of the HT diffusion approximation.

Because the SCV of the approximation does not depend on μ_i, we can satisfy requirement 3 by taking $\alpha_{ia} = \alpha$. μ_{ia} is found by solving

$$(1-\rho)\,\mathbb{E}[W_{i,Boon}] = \mathbb{E}[UI_{i,a}] = \mathbb{E}[U]\,\mathbb{E}[I_{i,a}] = \frac{1+\hat{\rho}_i}{2}\frac{\alpha_a}{\mu_{ia}}$$

This automatically satisfies requirements 1 and 2.

Chapter 3

Gated service discipline

The focus of this chapter lies on the gated service discipline. For the different queueing disciplines, LCFS, ROS, SJF and PS, the heavy traffic limits are analyzed. From this analysis, a distribution for the waiting time is derived. Then interpolation is used to find an approximation for the distribution of the waiting time for all loads and renewal arrivals. For every discipline this result will be compared with results that were obtained by simulation to see how well the waiting time distribution is approximated.

3.1 Last come first served

3.1.1 Heavy traffic limits

To find the Laplace-Stieltjes transform of the waiting time distribution in heavy traffic with Poisson arrivals, we use the strategy in Section 2.2.2. If $\tilde{W}_i$ is the scaled delay at Q_i defined as in Equation (2.12), then its LST $\tilde{W}^*_{i,LCFS}$ can be calculated using (2.16). This gives

$$\begin{aligned}
\tilde{W}^*_{i,LCFS}(s) &= \lim_{\rho\uparrow 1} W^*_{i,LCFS}(s(1-\rho)) \\
&= \lim_{\rho\uparrow 1} \frac{1-C_i^*(s(1-\rho)+\lambda_i(1-B_i^*(s(1-\rho))))}{\mathbb{E}[C]\,(s(1-\rho)+\lambda_i(1-B_i^*(s(1-\rho))))} && \text{(use (2.3))} \\
&= \lim_{\rho\uparrow 1} \frac{1-\left(\frac{\mu(1-\rho)}{\mu(1-\rho)+s(1-\rho)+\lambda_i(1-B_i^*(s(1-\rho)))}\right)^{\alpha}}{\frac{\mathbb{E}[S]}{(1-\rho)}\,(s(1-\rho)+\lambda_i(1-B_i^*(s(1-\rho))))} && \text{(use (2.15))} \\
&= \lim_{\rho\uparrow 1} \frac{1-\left(\frac{\mu}{\mu+s+\lambda_i(1-B_i^*(s(1-\rho)))/(1-\rho)}\right)^{\alpha}}{\mathbb{E}[S]\left(s+\frac{\lambda_i(1-B_i^*(s(1-\rho))}{1-\rho}\right)}.
\end{aligned}$$

Here, $-B_i^{*'}(0)$ is the negative derivative of the LST of the service time distribution evaluated at 0, this is equal to the expected service time by definition. Using l'Hôpital's rule and the fact that

$-B_i^{*'}(0) = \mathbb{E}[B_i]$ we see that:

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

Now it is clear that

$$\tilde{W}^*_{i,LCFS}(s) = \frac{1-\left(\frac{\mu}{\mu+s+\hat{\rho}_i s}\right)^{\alpha}}{\mathbb{E}[S](s+\hat{\rho}_i s)}$$
$$= \frac{1}{\mathbb{E}[S]s(\hat{\rho}_i+1)}\left\{1-\left(\frac{\mu}{\mu+s(1+\hat{\rho}_i)}\right)^{\alpha}\right\}, \tag{3.1}$$

where α and μ are given in (2.14).

The form of Equation (3.1) will be explained in Section 3.1.2

Differentiating (3.1) with respect to s and taking $s=0$ gives the expectation of $-\tilde{W}_i$, so we have:

$$\mathbb{E}\left[\tilde{W}_{i,LCFS}\right] = -\lim_{s\uparrow 0}\left[\frac{\left(\frac{\mu}{\mu+s(1+\hat{\rho}_i)}\right)^{\alpha}-1}{\mathbb{E}[S]s^2(\hat{\rho}_i+1)} + \frac{\alpha\left(\frac{\mu}{\mu+s(\hat{\rho}_i+1)}\right)^{\alpha}(1+\hat{\rho}_i)}{\mathbb{E}[S]s(1+\hat{\rho}_i)(\mu+s(1+\hat{\rho}_i))}\right]$$

use l'Hôpital's rule on the first term, rewrite second term:

$$= -\lim_{s\uparrow 0}\left[\frac{-\alpha\left(\frac{\mu}{\mu+s(1+\hat{\rho}_i)}\right)^{\alpha+1}}{\mathbb{E}[S]2s\mu} + \frac{\alpha\left(\frac{\mu}{\mu+s(1+\hat{\rho}_i)}\right)^{\alpha+1}}{\mathbb{E}[S]s\mu}\right]$$

add the two terms and use l'Hôpital's rule again:

$$= \frac{\alpha(\hat{\rho}_i+1)(\alpha+1)}{2\mu^2\mathbb{E}[S]}. \tag{3.2}$$

This same expression will be found for $\mathbb{E}\left[\tilde{W}_{i,FCFS}\right]$ by differentiating -$\tilde{W}^*_{i,FCFS}(s)$ in (2.13) and taking $s=0$. The fact that the expectation of the delay is the same for FCFS and LCFS is intuitively clear, because the amount of work that goes into the system has not changed and the order of service does not depend on the duration of the service required by the customers.

By differentiating (3.1) twice and taking $s=0$ we get the second moment of $\tilde{W}_i$ as

$$\mathbb{E}\left[\tilde{W}^2_{i,LCFS}\right] = \frac{\alpha(\hat{\rho}_i+1)^2(\alpha+1)(\alpha+2)}{3\mu^3\mathbb{E}[S]}. \tag{3.3}$$

From Equations (3.2) and (3.3) and the fact that they were obtained by differentiation, we can see what the kth moment of $\tilde{W}_i$ will be, namely

$$\mathbb{E}\left[\tilde{W}^k_{i,LCFS}\right] = \frac{(\hat{\rho}_i+1)^k\prod_{j=0}^{k}(\alpha+j)}{(k+1)\mu^{k+1}\mathbb{E}[S]}, \quad k\geq 1.$$

3.1.2 General load

In Section 2.2.2 it was stated that the distribution of the waiting time is a uniform $(\hat{\rho}_i, 1)$ distribution times a gamma (α, μ) distributed variable in case of a FCFS service order. We know that the amount of work in the system does not change if we change the queueing discipline to LCFS. This means that the cycle time distribution will also stay the same. From the shape of the LST of the delay given in (3.1) it is suspected that the distribution of the delay is again a uniform "times" a gamma distribution. In this case, the uniformly distributed variable lies between 0 and $(\hat{\rho}_i + 1)$. These boundaries suggest that the waiting time can be close to zero; this is indeed the case when a customer arrives just before the server arrives at his queue. In that case the customer will be taken into service immediately after the server arrives at his queue. A customer that arrives just after the server left his queue, has to wait a full cycle and, when the server arrives at his queue, he also has to wait until all the customers that came in after him are served.

In [18] Van der Mei proves that the boundaries of the uniform distribution are $\hat{\rho}_i$ and 1 for FCFS gated queueing disciplines. Using his method, we give further support by considering a different asymptotic regime that the boundaries for LCFS gated service disciplines are 0 and $(\hat{\rho}_i + 1)$. We look at the behavior of W_i when the switch-over times tend to infinity. The switch-over times are deterministic with length r_i. Define $r = \sum_{i=1}^{N} r_i$, as the total switch-over time per cycle. The waiting times are known to grow without bound when the switch-over times tend to infinity. Therefore the analysis is oriented towards the determination of the distribution of

$$\hat{W}_i := \lim_{r \to \infty} \frac{W_i}{r}.$$

This is referred to as the *asymptotic scaled* waiting time at Q_i.

We need to prove the following theorem

Theorem 1.

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

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

$$\tilde{a}_i = 0, \quad \tilde{b}_i = \frac{\rho_i + 1}{1 - \rho}$$

Proof. If the switch-over times are deterministic and tend to infinity, the distribution of the length of a queue at polling instant of that queue, divided by the total switch-over time per cycle, converges almost surely to a known constant. The cycle time is therefore also deterministic. For the LST of the cycle time we have the LST of a deterministic distributed random variable with expectation $\mathbb{E}[C]$:

$$C^*(s) = e^{-sr/(1-\rho)} = e^{-s\,\mathbb{E}[C]},$$

In [18] it is stated that

$$\hat{W}_i^*(s) = \lim_{r \to \infty} W_i^*(s/r)$$

Using Equation (2.3) in combination with $\mathbb{E}[C] = r/(1-\rho)$ for $W_i^*(s)$ and some basic calculations including L'Hôpital's rule, we get

$$\begin{aligned} \hat{W}_i^*(s) &= \lim_{r \to \infty} \frac{1-\rho}{r} \frac{1 - C^*(s/r + \lambda_i(1 - B_i^*(s/r)))}{s/r + \lambda_i(1 - B_i^*(s/r))} \\ &= \frac{1 - e^{s(1+\hat{\rho}_i)/(1-\rho)}}{s(1+\hat{\rho}_i)/(1-\rho)} \end{aligned}$$

Recall that $\tilde{b}_i = (\rho_i + 1)/(1-\rho)$. This gives

$$\hat{W}_i^*(s) = \frac{1}{\tilde{b}_i s}(1 - e^{-s\tilde{b}_i})$$

This equation is the LST of the uniform distribution on the interval $[0, \tilde{b}_i]$. This shows that the scaled delay is indeed uniformly distributed on the interval interval $[0, \tilde{b}_i]$ when the switch-over times tend to infinity. □

We already have an approximation for the waiting time distribution that works well in case of renewal arrivals with FCFS service and that depends on the fact that the heavy traffic limit has a uniform times a gamma distribution. The gamma distribution does not change and we now know the boundaries of the uniform distribution. This leads to the following approximation for the waiting time distribution in polling systems with renewal arrivals, $\rho < 1$, gated service discipline and LCFS queueing discipline

$$\mathbb{P}[W_i < x] \approx \mathbb{P}[UI_{i,a} < (1-\rho)x],$$

where U is a uniformly distributed random variable on $[0, 1+\hat{\rho}_i]$, and $I_{i,a}$ a gamma distributed random variable with parameters α_a and μ_{ia} given in (2.23). The parameters do not change, because the distribution of the cycle time stays the same and the parameters of the gamma distribution only depend on the uniform distribution through its expectation, which is $(\hat{\rho}_i + 1)/2$ for both uniform distributions.

The kth moments can be calculated using:

$$\begin{aligned}\mathbb{E}[W_{i,LCFS}^k] &= \frac{1}{(1-\rho)^k}\,\mathbb{E}[U^k]\,\mathbb{E}[I_{i,a}^k] \\ &= \frac{1}{(1-\rho)^k}\,\frac{(1+\hat{\rho}_i)^k}{(k+1)}\prod_{j=0}^{k-1}\frac{\alpha_a + j}{\mu_{ia}}.\end{aligned}$$

For the second equality standard properties of the uniform and gamma distribution are used.

3.1.3 Numerical results

To get an insight in the performance of the approximation, it is applied to a test bed of 42 polling systems. For every system, the first three moments are calculated together with a number of percentiles and the cumulative distribution function. These measures are compared to the exact values that were obtained by simulation. The values from the simulation are the averages of a variable number of simulation runs with a length of at least 100,000,000 time units, such that the width of the confidence interval of the the mean waiting time is less than 0.5% of the value of the actual mean.

There are two different polling systems considered, a small system with 3 queues and a large system with 7 queues. For both systems, the service and switch-over time distributions are fixed per queue (by mean and SCV). The ratios between the arrival rates are also fixed, the actual values are determined by the load of the system. For both systems, the interarrival time distributions are exponential (Poisson arrivals), mixed erlang (SCV < 1) or hyper-exponential (SCV > 1), this leads to 6 different systems. These systems will all be tested with 7 different loads, which leads to a total of 42 different systems. The values of the parameters are given in Table 3.1.

Notation	Parameter	$N=3$	$N=7$
$\mathbb{E}[B]$	Mean service times	$\begin{bmatrix}1 & 2 & 3\end{bmatrix}$	$\begin{bmatrix}1 & 2 & 3 & 2 & 1 & 4 & 1\end{bmatrix}$
c_B^2	SCV service times	$\begin{bmatrix}4 & \frac{1}{2} & 1\end{bmatrix}$	$\begin{bmatrix}1 & \frac{1}{2} & \frac{1}{2} & 2 & 4 & 1 & \frac{1}{4}\end{bmatrix}$
λ	Arrival rates	1:3:2	1:1:1:1:5:1:1
$\mathbb{E}[S]$	Mean switch-over times	$\begin{bmatrix}1 & 1 & 3\end{bmatrix}$	$\begin{bmatrix}3 & 1 & 1 & 2 & 3 & 1 & 4\end{bmatrix}$
c_S^2	SCV switch-over times	$\begin{bmatrix}1 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}$
N	Number of queues	$\{3, 7\}$	
c_A^2	SCV interarrival times	$\{1, \frac{1}{4}, 4\}$	
ρ	Load	$\{0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95\}$	

Table 3.1: Parameter values of the test bed

In Figure 3.1 the probability density function of the waiting time is plotted for the first queue of a polling system with 3 queues, a load of 0.8 and Poisson arrivals. The mean service time and mean switch-over time are 1, $c_B^2 = 4$ and $c_S^2 = 1$. The figure shows that the approximation closely resembles the exact waiting time distribution. Only for low waiting times, the simulation values lie higher than the approximation, it is not clear what causes this difference.

The simulation calculates the first 3 moments of the waiting time distribution, 7 percentiles (40, 50,...,90, 95) and a histogram of the waiting time distribution with 512 bins.

Tables 3.2 - 3.11 show the errors of the approximation categorized in bins. The errors of the moments and percentiles are calculated in the following way,

$$\Delta\% = \frac{|a-s|}{s} \times 100\%, \tag{3.4}$$

where a denotes the approximated value and s denotes the exact value. The error in cumulative distribution function is given by the maximum absolute difference between the approximated and exact cumulative distribution of the waiting time:

$$\max_i \{|a_i - s_i|\}, \quad i = 1, \ldots, 512. \tag{3.5}$$

The 512 bins of the simulated histogram are used to calculate 512 points of the cumulative distribution function (s_i) and the same 512 points are calculated using the approximation (a_i). The numbers in the tables are percentages, but for readability the percentage signs are omitted.

Table 3.2 shows that the approximation of the mean waiting time is most accurate for $\rho = 0.95$, with 100% of the errors smaller than 5%. The table also shows that the approximation works well for small values and large values of ρ, the worst case is $\rho = 0.5$. A similar conclusion can be drawn from [2]. This is to be expected, since the same approximation for the mean waiting time is used here.

Table 3.3 shows that the approximation of the second moment of the waiting time distribution follows the same behavior as the approximation of the first moment. The best approximation is achieved in case of a load near the heavy traffic limit. When ρ is near 0.5, the approximation is worse. The influence of the distribution is visible in this table, where the mean of the distribution is exact in both heavy traffic and light traffic, the distribution is only exact in case of heavy traffic. This can be seen from the fact that for low loads, the first moment is approximated correctly (with an error of 5% or less) in 93.33% of the cases while for the same load, the second moment is approximated correctly in only 53.33% of the cases.

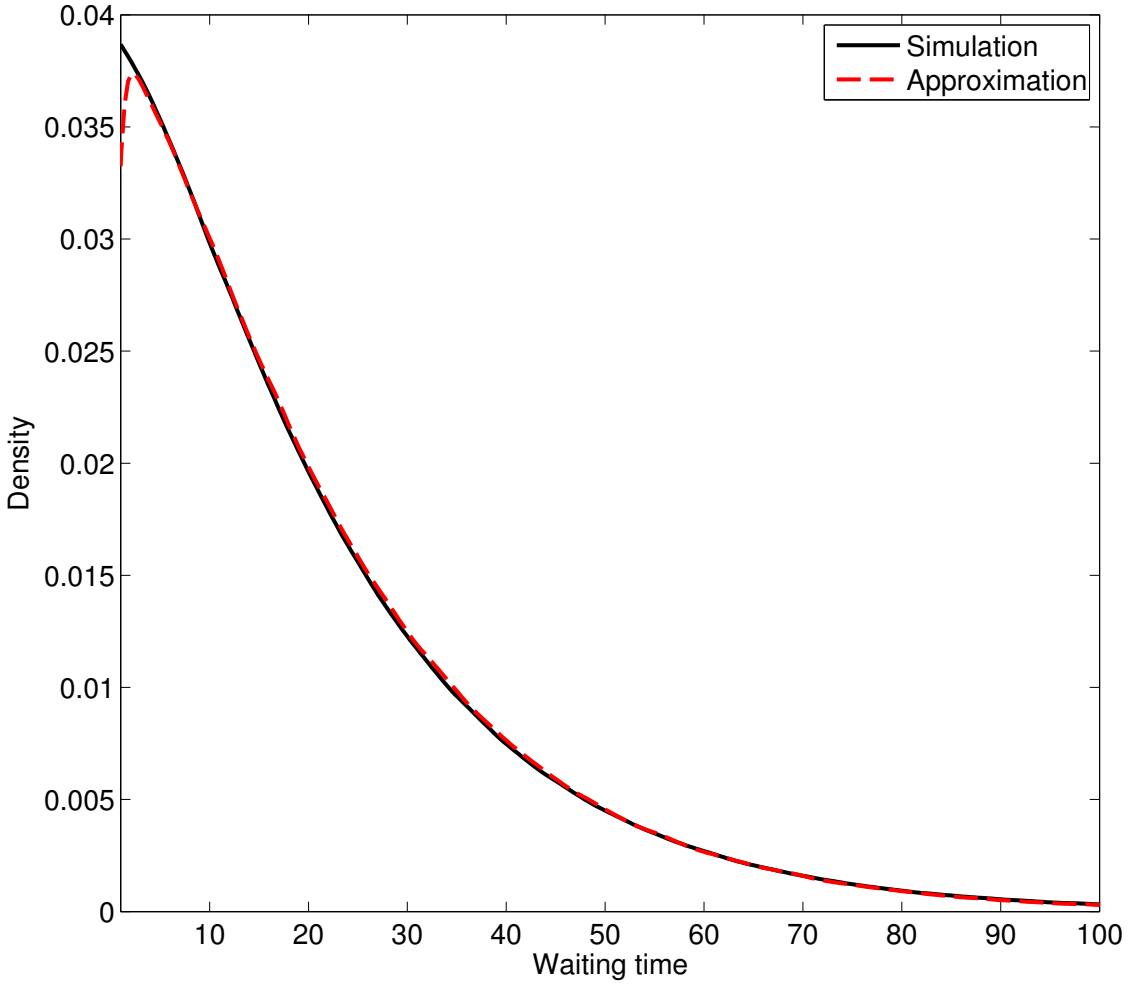


Figure 3.1: Approximated and simulated density function of the waiting time of an arbitrary queue in the example in Section 3.1.3

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	93.33	6.67	0.00	0.00	0.00	0.00
0.3	66.67	23.33	10.00	0.00	0.00	0.00
0.5	60.00	26.67	13.33	0.00	0.00	0.00
0.7	63.33	30.00	6.67	0.00	0.00	0.00
0.8	76.67	20.00	3.33	0.00	0.00	0.00
0.9	93.33	6.67	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table 3.2: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	53.33	20.00	6.67	20.00	0.00	0.00
0.3	50.00	23.33	16.67	10.00	0.00	0.00
0.5	46.67	26.67	23.33	3.33	0.00	0.00
0.7	50.00	33.33	10.00	3.33	3.33	0.00
0.8	73.33	10.00	13.33	3.33	0.00	0.00
0.9	76.67	20.00	3.33	0.00	0.00	0.00
0.95	93.33	6.67	0.00	0.00	0.00	0.00

Table 3.3: Second moment of the waiting time errors categorized in bins of 5%

Table 3.4 shows that the third moment is harder to approximate. The third moment is very large and very variable. When $\rho = 0.5$, only 26.7% of the errors is between 0 and 5%. The case of $\rho = 0.95$ performs reasonably well, 83.3% of the errors is lower than 5% and all errors are lower than 10%.

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	30.00	26.67	10.00	3.33	0.00	30.00
0.3	36.67	16.67	13.33	6.67	13.33	13.33
0.5	26.67	40.00	20.00	3.33	6.67	3.33
0.7	46.67	30.00	10.00	10.00	0.00	3.33
0.8	60.00	20.00	13.33	6.67	0.00	0.00
0.9	73.33	16.67	10.00	0.00	0.00	0.00
0.95	83.33	16.67	0.00	0.00	0.00	0.00

Table 3.4: Third moment of the waiting time errors categorized by rho

Table 3.5 shows that the percentiles are also best approximated if ρ is large, which is also caused by the fact that the distribution is exact in heavy traffic. The light traffic interpolation is visible in the fact that the approximation in case of $\rho = 0.1$ is better than the case where ρ is between 0.3 and 0.7.

In Table 3.6 the maximum absolute differences between the cumulative distributions are shown.

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	68.10	12.86	10.95	7.62	0.48	0.00
0.3	50.95	27.14	13.33	5.24	2.86	0.48
0.5	54.29	25.24	13.33	4.76	1.43	0.95
0.7	66.19	21.43	9.52	2.38	0.48	0.00
0.8	74.76	19.05	6.19	0.00	0.00	0.00
0.9	87.62	11.90	0.48	0.00	0.00	0.00
0.95	99.52	0.48	0.00	0.00	0.00	0.00

Table 3.5: Percentile errors categorized in bins of 0.05

Note that a difference of 0.1 is already a big difference. From the table it can be seen that differences larger than 0.1 do not occur in this test bed. The same relation with ρ that is also found in the other errors can be found here.

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	66.67	33.33	0.00	0.00	0.00	0.00
0.3	56.67	43.33	0.00	0.00	0.00	0.00
0.5	60.00	40.00	0.00	0.00	0.00	0.00
0.7	80.00	20.00	0.00	0.00	0.00	0.00
0.8	90.00	10.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table 3.6: Maximum absolute differences in cdf categorized in bins of 0.05

In the previously discussed tables, all errors are categorized by ρ. From the tables it is clear that the approximation works best for heavily loaded systems and worst for systems with a load around 0.5. The two other parameters that were varied during the tests, are the SCV of the interarrival times and the number of queues.

To determine the influence of the number of queues on the waiting time distribution, all percentile errors and moment errors are categorized in bins and split by number of queues. The results can be found in Table 3.7. From the table it is clear that the approximation works better if N is larger. This is also a property of the expected waiting time derived by Boon et. al. [2].

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	58.41	22.22	11.75	4.44	2.22	0.95
7	74.22	15.92	6.39	2.38	0.27	0.82

Table 3.7: Errors by number of queues categorized in bins of 5%

From Table 3.8 it can be seen that the approximation works very good when the arrivals are according to a Poisson process. When the SCV of the interarrival times is not equal to 1, the

accuracy of the approximation declines. This is especially the case when interarrival times have a large SCV.

	Bins					
$c^2_{A_i}$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	95.24	3.87	0.60	0.30	0.00	0.00
0.25	66.37	22.77	7.44	1.34	1.64	0.45
4	43.01	28.13	17.86	7.74	1.04	2.23

Table 3.8: Errors by SCV of the interarrival times categorized in bins of 5%

In Table 3.9 all percentile errors are categorized in bins and split by percentile. The 80, 90 and 95 percentiles are approximated considerably better than the 40, 50, 60 and 70 percentiles.

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	55.24	19.05	11.90	10.95	1.90	0.95
50	59.52	18.57	15.24	4.76	1.43	0.48
60	65.24	20.00	12.38	0.95	1.43	0.00
70	70.00	20.00	5.24	3.81	0.95	0.00
80	82.86	13.33	3.33	0.48	0.00	0.00
90	82.38	15.71	1.90	0.00	0.00	0.00
95	82.86	11.43	3.81	1.90	0.00	0.00

Table 3.9: Errors in percentiles categorized in bins of 5%

In Table 3.10 the results are split by queue number for the system with 3 queues. From the table, it can be concluded that the waiting time distribution for the third queue is harder to approximate than that for the other two queues. It seems that the higher mean switch-over time between queue 3 and 1 has something to do with this. It is also possible that it is caused by the high mean service time at that queue.

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	64.29	27.14	6.19	1.43	0.48	0.48
2	60.95	19.05	12.38	4.29	2.86	0.48
3	50.00	20.48	16.67	7.62	3.33	1.90

Table 3.10: Errors by queue in the systems with 3 queues categorized in bins of 5%

Table 3.11 indicates that in the 7 queue system, queues 2, 3 and 6 are the hardest queues for the approximation. These are surprisingly enough the queues with the lowest mean switch-over times to the next queues. It are also the queues with the highest mean service time duration, as was also the case with the hardest queue in the 3 queue system. So apparently a higher mean service time leads to a drop in the accuracy of the approximation.

In recap, the approximation works best when $\rho \geq 0.8$, N is large and when arrivals are Poisson. The distribution of the waiting time approximation is least accurate when ρ is 0.3 - 0.5 and when the SCV of the interarrival times is large.

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	82.86	10.48	3.33	1.90	0.48	0.95
2	70.48	20.95	4.76	2.86	0.00	0.95
3	65.71	20.48	9.52	3.33	0.48	0.48
4	73.33	20.95	2.86	1.90	0.00	0.95
5	88.10	5.71	4.29	0.95	0.48	0.48
6	60.95	18.10	15.71	4.29	0.48	0.48
7	78.10	14.76	4.29	1.43	0.00	1.43

Table 3.11: Errors by queue in the systems with 7 queues categorized in bins of 5%

3.2 Random order of service

In this section the random order of service queueing discipline is analyzed.

3.2.1 Heavy traffic limits

The random order of service is represented by ordering marks. Every customer that arrives gets an ordering mark, 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. Using Equation (2.16) for the first equality, (2.4) for the second equality and for the final equality l'Hôpital's rule and basic calculations, we find the following LST of the waiting time conditional on the ordering mark x.

$$\begin{aligned}\tilde{W}^*_{i,ROS}(s|x) &= \lim_{\rho\uparrow 1} W^*_{i,ROS}(s(1-\rho)|x) \\ &= \lim_{\rho\uparrow 1} \frac{C^*_i(\lambda_i x(1-B^*_i(s(1-\rho)))) - C^*_i(s(1-\rho)+\lambda_i x(1-B^*_i(s(1-\rho))))}{s(1-\rho)\,\mathbb{E}[C]} \\ &= \frac{1}{\mathbb{E}[S]s}\left\{\left(\frac{\mu}{\mu+\hat{\rho}_i x s}\right)^{\alpha} - \left(\frac{\mu}{\mu+(1+\hat{\rho}_i x)s}\right)^{\alpha}\right\}, \end{aligned} \tag{3.6}$$

where x is the ordering mark of a tagged customer, taken from a uniform distribution on [0,1].

The unconditional LST can be found by integrating out x with respect to its density function. First calculate, using integration by substitution,

$$\begin{aligned}\int_0^1 \left(\frac{\mu}{\mu+s\hat{\rho}_i x}\right)^{\alpha} \mathrm{d}x &= \int_1^{\frac{\mu}{\mu+s\hat{\rho}_i}} u^{\alpha} \frac{1}{\frac{-s\mu\hat{\rho}_i}{(\mu+\hat{\rho}_i x s)^2}}\,\mathrm{d}u \\ &= \int_1^{\frac{\mu}{\mu+s\hat{\rho}_i}} \frac{-\mu}{\hat{\rho}_i s} u^{\alpha-2}\,\mathrm{d}u \\ &= \left[\frac{-\mu}{s\hat{\rho}_i(\alpha-1)} u^{\alpha-1}\right]_1^{\frac{\mu}{\mu+s\hat{\rho}_i}} \\ &= \frac{-\mu}{s\hat{\rho}_i(\alpha-1)}\left(\left(\frac{\mu}{\mu+s\hat{\rho}_i}\right)^{\alpha-1} - 1\right)\end{aligned}$$

and similarly,

$$\int_0^1 \left(\frac{\mu}{\mu+s(\hat{\rho}_i x+1)}\right)^{\alpha} \mathrm{d}x = \left[\frac{-\mu}{s\hat{\rho}_i(\alpha-1)}u^{\alpha-1}\right]_{\frac{\mu}{\mu+s}}^{\frac{\mu}{\mu+s(1+\hat{\rho}_i)}}$$
$$= \frac{-\mu}{s\hat{\rho}_i(\alpha-1)}\left(\left(\frac{\mu}{\mu+s(\hat{\rho}_i+1)}\right)^{\alpha-1} - \left(\frac{\mu}{\mu+s}\right)^{\alpha-1}\right).$$

Using the above we get

$$\tilde{W}^*_{i,ROS}(s) = \int_0^1 \tilde{W}^*_i(s|x)\,\mathrm{d}x$$
$$= \frac{\mu}{s^2\,\mathbb{E}[S]\hat{\rho}_i(\alpha-1)} \left\{\left(1-\left(\frac{\mu}{\mu+s\hat{\rho}_i}\right)^{\alpha-1}\right)\right.$$
$$\left. -\left(\left(\frac{\mu}{\mu+s}\right)^{\alpha-1} - \left(\frac{\mu}{\mu+s(\hat{\rho}_i+1)}\right)^{\alpha-1}\right)\right\}. \tag{3.7}$$

This expression can be used to find the distribution of the scaled waiting with random service orders. The kth moments of the waiting time can be found by differentiating k times with respect to s and then take $s=0$. The first moment is equal to the first moment with FCFS/LCFS service orders, given in (3.2).

From the form of (3.6) it appears that the distribution of the waiting time, conditional on the order mark, is a uniform times a gamma distribution. The cycle time does not change, so that remains a gamma distribution with parameters $\alpha+1$ and μ as given in (2.14). The boundaries of the uniform distribution are $\hat{\rho}_i x$ and $1+\hat{\rho}_i x$, where x is a uniformly distributed variable on $[0,1]$. This can be interpreted as follows, when a customer arrives just before the server starts serving his queue, he has to wait a fraction of the residual cycle time somewhere between 0 and $\hat{\rho}_i$. On the other hand, when a customer arrives just after the server leaves his queue, he has to wait a full cycle plus a fraction of the residual cycle time somewhere between 0 and $\hat{\rho}_i$.

To find the unconditional distribution of the waiting time, we need to find the unconditional "uniform" distribution $\tilde{U}$.

First we consider a more general case, because it is needed later. We have the following lemma:

Lemma 1. *Suppose we have a conditional random variable $T|x$ that is uniformly distributed on $[a(x),a(x)+1]$. We want to find the unconditional distribution $\tilde{T}$. Here, x is a realization of the random variable X with continious probability density function $f_X(x)$, $x \geq 0$. Suppose that $a(x)$ has the following properties: $a(0)=m$, $a(\infty)=\hat{\rho}_i$ and $a(x)$ is increasing in x. The function $a^{-1}(y)$ is the inverse of $a(x)$. Then, the unconditional distribution of $T|x$, denoted by $\tilde{T}$, has probability density function*

$$f_{\tilde{T}}(y) = \begin{cases} F_X(a^{-1}(y)) & y \in [m,\hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1+m] \\ 1-F_X(a^{-1}(y-1)) & y \in (1+m, 1+\hat{\rho}_i]. \end{cases} \tag{3.8}$$

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

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

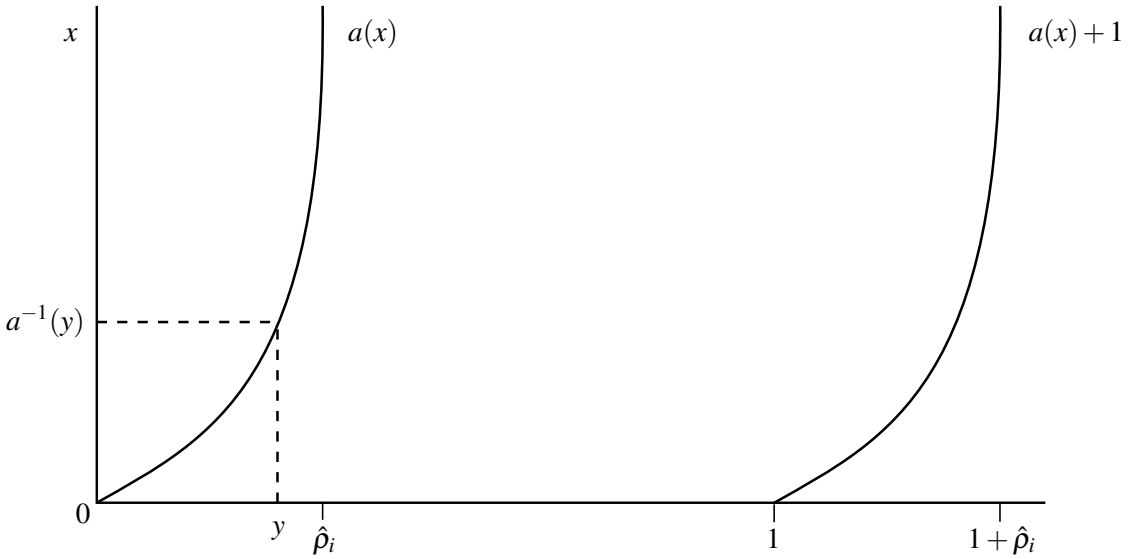


Figure 3.2: Boundaries of the uniform distribution

For a given x, the boundaries of the uniform distribution are known. Figure 3.2 shows an example of all possible boundaries of the uniform distribution, by plotting $a(x)$ and $a(x)+1$with x on the y-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 we pick a $y \in (m, \hat{\rho}_i)$, in that case, the probability mass of $f_{\tilde{T}}(y)$ is obtained from the parts where x is smaller than $a^{-1}(y)$, the inverse function of $a(x)$, in y. This is illustrated in Figure 3.2, if x becomes larger than $a^{-1}(y)$, y is not between the boundaries of the uniform distribution. This gives, for $y \in (m, \hat{\rho}_i)$,

$$f_{\tilde{T}}(y) = \int_{x=0}^{a^{-1}(y)} f_X(x) * f_{T|x}(y)\, \mathrm{d}x = F_X(a^{-1}(y)),$$

the cumulative distribution function of X in the point $a^{-1}(y)$. If $y \in (\hat{\rho}_i, 1+m)$, it lies between the boundaries of the uniform distribution for every x. Hence, we get

$$f_{\tilde{T}}(y) = \int_{x=0}^{\infty} f_X(x) * f_{T|x}(y)\, \mathrm{d}x = 1.$$

Finally, for $y \in (1+m, 1+\hat{\rho}_i)$, we can use the fact that we have the same curve twice, i.e., x needs to be larger than $a^{-1}(y-1)$, so

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

Putting the three equations above together, we get the probability density function of $\tilde{T}$ given in Equation (3.8). Because of the properties of a, the fact that its inverse exists and is unique, and X, this is a continuous function in y. □

We now apply the lemma above to the case of ROS. Then $a(x) = \hat{\rho}_i x$, with $x \in [0,1]$. This function has the desired properties: $a(0) = 0$, $a(1) = \hat{\rho}_i$ and $a(x) \uparrow x$. The cumulative distribution function of X is given by $F_X(x) = x$ and the inverse function of a is $a^{-1}(y) = y/\hat{\rho}_i$. The boundaries of the

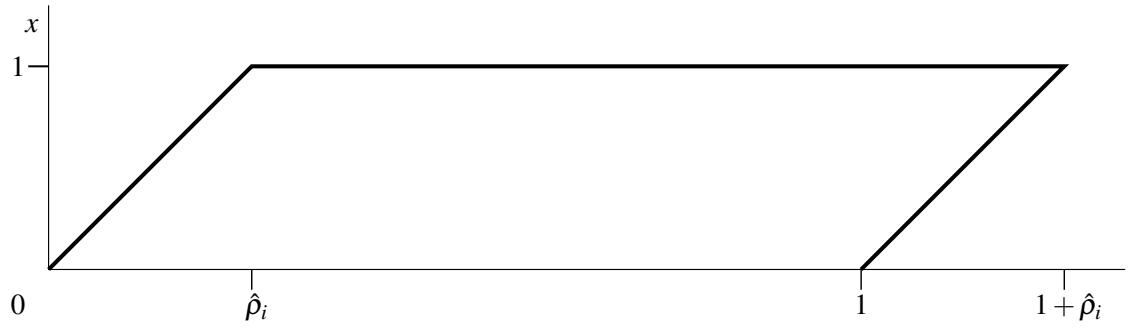


Figure 3.3: Boundaries of the uniform distribution

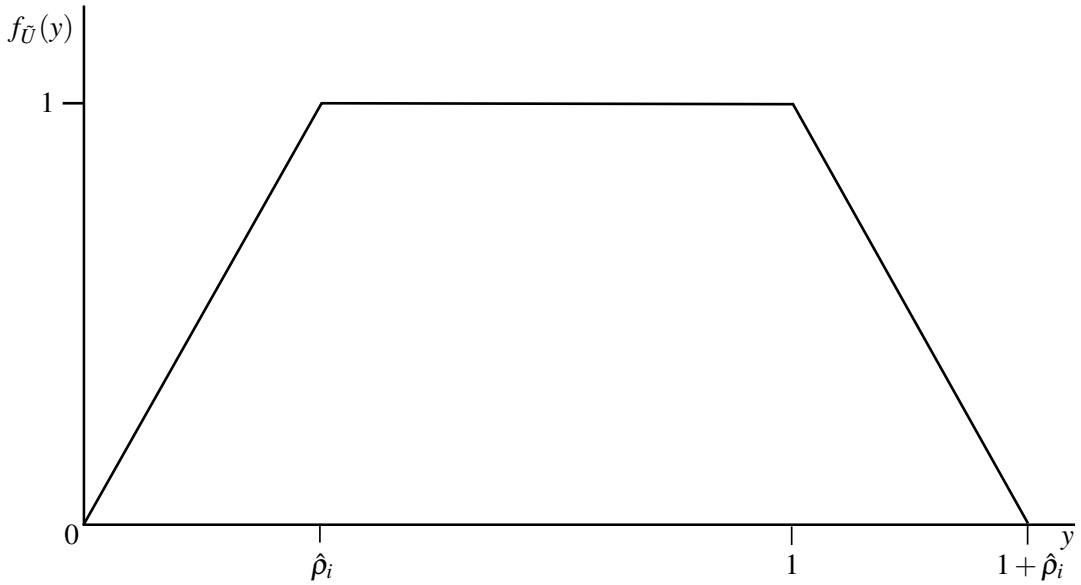


Figure 3.4: Probability density function of $\tilde{U}$

uniform distribution U are graphically represented in Figure 3.3. Equation (3.8) can now be used to find the distribution function of $\tilde{U}$, giving

$$f_{\tilde{U}}(y) = \begin{cases} y/\hat{\rho}_i & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ (1+\hat{\rho}_i - y)/\hat{\rho}_i & y \in (1, 1+\hat{\rho}_i]. \end{cases} \tag{3.9}$$

In [10] this distribution is described as a trapezoidal distribution with parameters $a = 0$, $b = \hat{\rho}_i$, $c = 1$ and $d = 1 + \hat{\rho}_i$. This fact can also be found by noting that the unconditional LST of the waiting time given in Equation (3.7) is the LST of a trapezoidal distribution with these parameters times a gamma distribution with parameters $\alpha + 1$ and μ. Figure 3.4 illustrates the trapezoidal shape of the graph of the probability density function of $\tilde{U}$.

So the uniform distribution is replaced by a trapezoidal distribution. This 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 the moment that the customer is served. The trapezoidal distribution represents the fraction of the residual cycle time that the customer has to wait. Figure 3.3 shows that a fraction between 0 and $\hat{\rho}_i$ is only possible for certain values of x. A fraction close to $\hat{\rho}_i$ lies in more intervals than a fraction close to 0, so the probability that the fraction is around $\hat{\rho}_i$ is higher than the probability that a fraction is close to 0, this probability increases linearly. The

fractions between and including $\hat{\rho}_i$ and 1 lie in every interval, so the probabilities of these fractions are equal. For fractions larger than 1, we see that a fraction close to 1 lies in more intervals than a fraction close to $1+\hat{\rho}_i$. The probabilities of the fractions decrease as the fractions increase. This gives exactly the trapezoidal form of the distribution of the fractions.

3.2.2 General load

This heavy traffic limit leads to the following approximation for the waiting distribution in polling systems with renewal arrivals, $\rho < 1$, gated service discipline and ROS queueing discipline

$$\mathbb{P}[W_i < w] \approx \mathbb{P}[\tilde{U} I_{i,a} < (1-\rho)w],$$

where $\tilde{U}$ is a trapezoidal distributed random variable with probability density function given in Equation (3.9), and $I_{i,a}$ is a gamma distributed random variable with parameters α_a and μ_{ia} given in (2.23). The parameters do not change, because the distribution of the cycle time stays the same and the parameters only depend on the "uniform" distribution through its expectation, which is again $(\hat{\rho}_i+1)/2$.

The kth moments can be calculated using

$$\begin{aligned}
\mathbb{E}[W_{i,ROS}^k] &= \frac{1}{(1-\rho)^k}\,\mathbb{E}[\tilde{U}^k]\,\mathbb{E}[I_{i,a}^k] \\
&= \frac{1}{(1-\rho)^k}\int_{\hat{\rho}_i}^{1+\hat{\rho}_i} u^k f_{\tilde{U}}(u)\,\mathrm{d}u \prod_{j=0}^{k-1}\frac{\alpha_a+j}{\mu_{ia}} \\
&= \frac{1}{(1-\rho)^k}\,\frac{(1+\hat{\rho}_i)^{k+2}-\hat{\rho}_i^{k+2}-1}{\hat{\rho}_i(k+1)(k+2)}\prod_{j=0}^{k-1}\frac{\alpha_a+j}{\mu_{ia}}
\end{aligned}$$

The same result can be obtained by unconditioning the kth moment of the conditional uniform distribution, i.e. $\mathbb{E}[U^k] = \int_0^1 \int_{\hat{\rho}_i x}^{1+\hat{\rho}_i x} u^k \,\mathrm{d}u\,\mathrm{d}x$.

3.2.3 Numerical results

In Section 3.1.3, the approximation for LCFS service discipline was evaluated. The same approach can also be applied to the ROS approximation. The approximation is applied to the same test bed, given in Table 3.1 and the same simulation procedure was used. The results are given in Tables A.1 - A.10 and can be found in Appendix A.1. Figure 3.5 depicts exactly the same as Figure 3.1, however instead of LCFS, the service order is now ROS. It shows the probability density function of the first queue of a 3 queue system with a load of 0.8 and a $c_{A_i}^2$ of 1 for all i.

The results for ROS are comparable with the results for LCFS. From Tables A.1 - A.3 it can be concluded that the first two moments are well approximated for both heavy and light traffic. The third moment is harder to approximate in case of light traffic. The distributions are nicely approximated, especially in heavy traffic. This can be found in Tables A.4 and A.5. The errors in percentile estimation are all between 0 and 5% when ρ is equal to 0.95 and when ρ is 0.9 or 0.95, the maximum differences between the approximated cumulative density functions are smaller than 0.05. Tables A.6 and A.7 show that larger N and Poisson arrivals respectively give more accurate results. The same things could also be concluded in the LCFS case. Again the 90th percentile is approximated best and the same holds for the waiting time distribution in the first queue in the 3

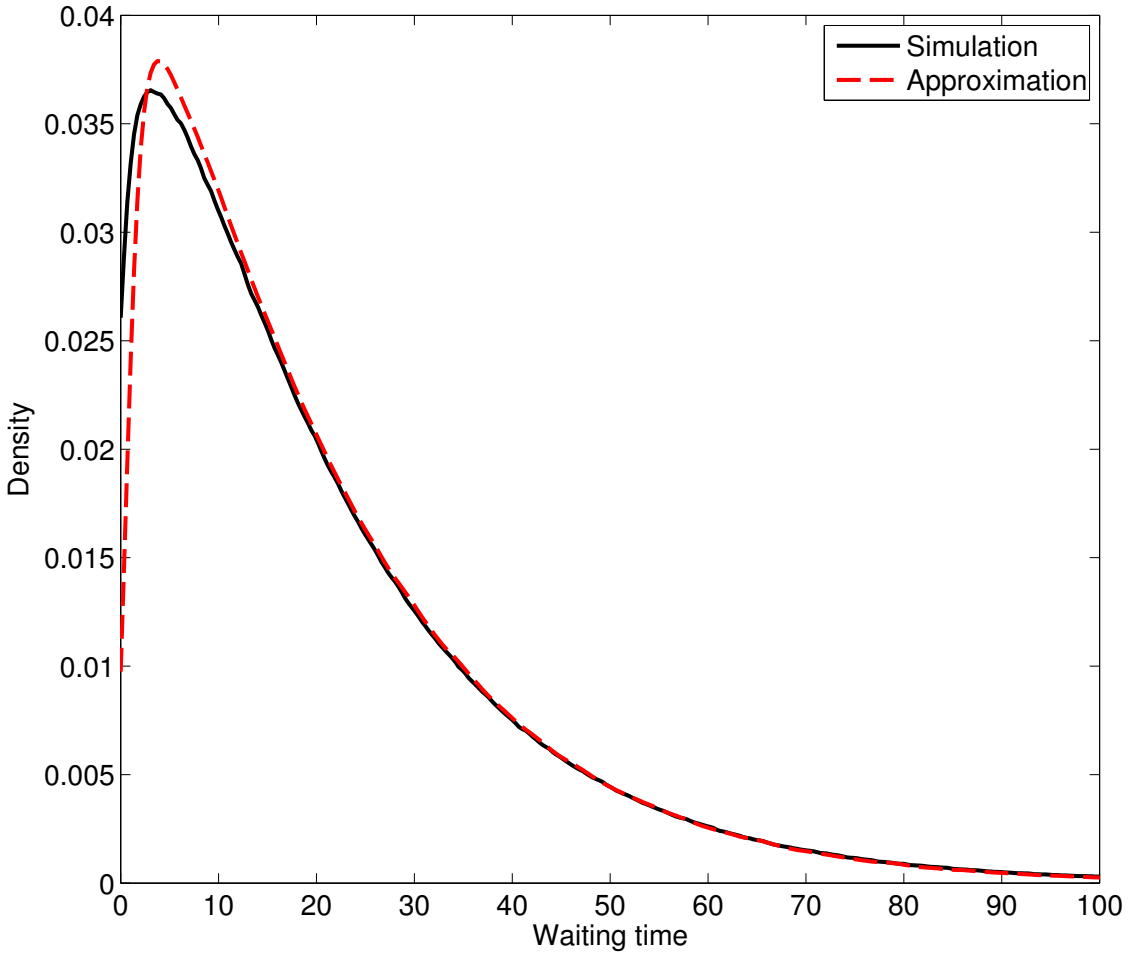


Figure 3.5: Approximated and simulated density function of the waiting time of an arbitrary queue in the example in Section 3.2.3

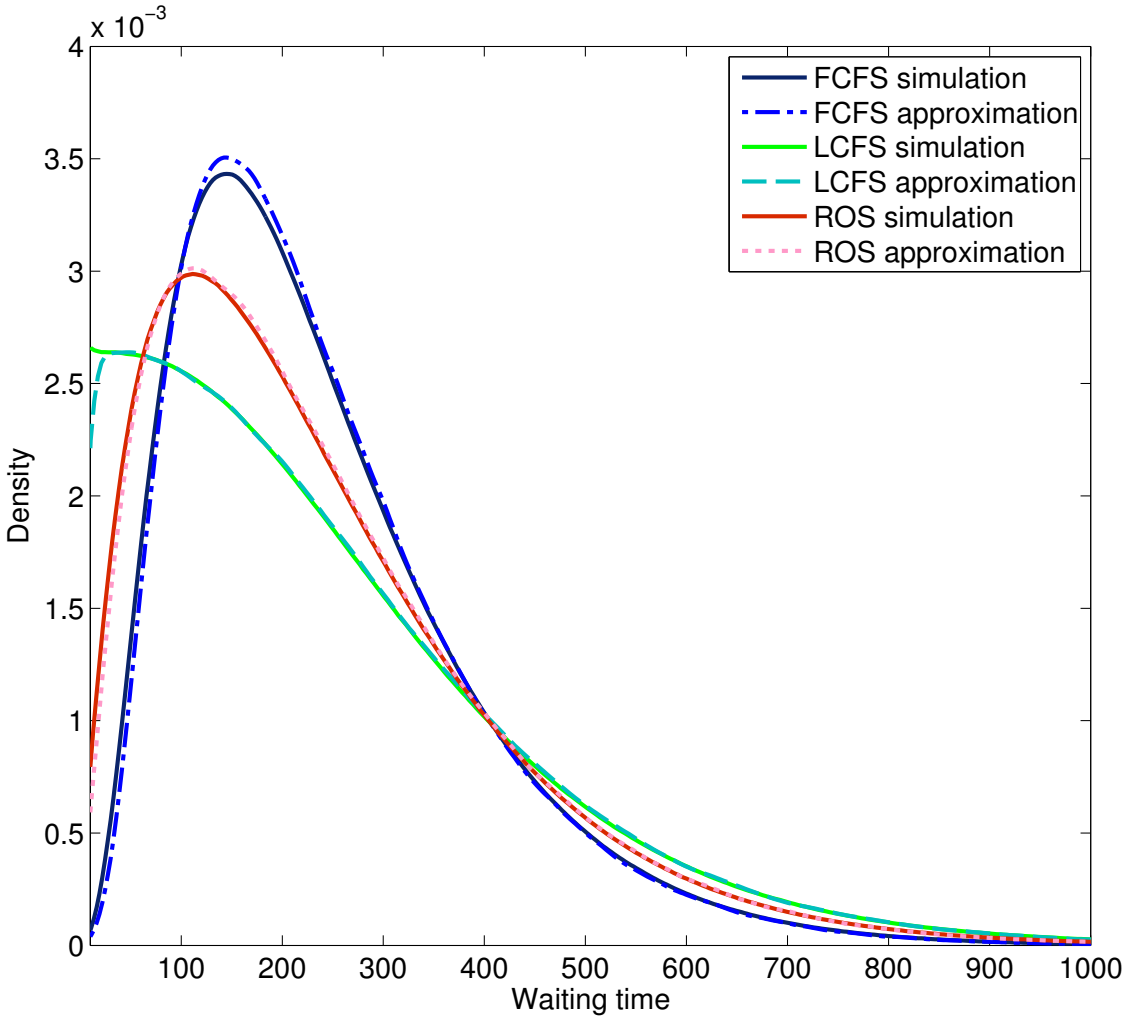


Figure 3.6: Approximated and simulated cumulative distribution function of the waiting time with different service disciplines

queue system and the waiting time distributions of queues 1, 5 and 7 in the 7 queue system. This confirms that the mean service time has a high impact on the accuracy of the approximation.

Figure 3.6 shows three probability density functions of the waiting times with three different service disciplines. They are the waiting time distributions of the fifth queue of a system with $\rho = 0.95$, $c_{A_i}^2 = 1$ and $N = 7$. The variation in waiting times is bigger when the service order is LCFS, this is represented in the figure by the fat tail of the waiting time distribution. The lowest variation in waiting time is encountered with a FCFS queueing discipline.

3.3 Shortest job first

3.3.1 Heavy traffic limits

Now we look at heavy traffic limits for SJF queueing policies. Filling in (2.16) using (2.5), which holds for both the PS and the SJF queueing discipline by adjusting $\varphi(s,x)$, and (2.15), we get for

the LST of the waiting time in a system with a SJF service order

$$\tilde{W}_{i,SJF}(s|x) = \frac{1}{\mathbb{E}[S]s}\left\{\left(\frac{\mu}{\mu+\hat{\lambda}_i\mathbb{E}[B_{i,\varphi(x)}]s}\right)^{\alpha} - \left(\frac{\mu}{\mu+(1+\hat{\lambda}_i\mathbb{E}[B_{i,\varphi(x)}])s}\right)^{\alpha}\right\}, \tag{3.10}$$

with $\alpha+1$ and μ as given in (2.14) and $\mathbb{E}[B_{i,\varphi(x)}] = \mathbb{E}[B_i\mathbf{1}_{\{B_i\leq x\}}]$. If the probability density function of the service time at queue i is given by $f_{B_i}(x)$, this expectation is calculated using

$$\mathbb{E}[B_i\mathbf{1}_{\{B_i\leq x\}}] = \int_0^x y f_{B_i}(y)\,\mathrm{d}y. \tag{3.11}$$

The unconditional expectation is obtained by integrating out x with respect to its density, i.e.,

$$\mathbb{E}[B_{i,\varphi}] = \int_0^\infty f_{B_i}(x)\int_0^x y f_{B_i}(y)\,\mathrm{d}y\,\mathrm{d}x. \tag{3.12}$$

The distribution of the waiting time in heavy traffic not only depends on the first two moments of the service time, as was the case with the other disciplines, but it depends on the complete service time distribution. The LST of the waiting time shows that the distribution of the conditional waiting time is still a uniform times a gamma distribution. We can use Lemma 1, given in Section 3.2.1. Now U is uniformly distributed on $[\hat{\lambda}_i\mathbb{E}[B_{i,\varphi(x)}], 1+\hat{\lambda}_i\mathbb{E}[B_{i,\varphi(x)}]]$, so $a(x) = \hat{\lambda}_i\mathbb{E}[B_i\mathbf{1}_{\{B_i\leq x\}}]$. From (3.11) it is clear that indeed $a(0)=0$, $a(\infty) = \hat{\lambda}_i\mathbb{E}[B_i] = \hat{\rho}_i$ and $a(x)\uparrow x$. Using Equation (3.8) the probability density function of the unconditioned uniform distribution $\tilde{U}$ is

$$f_{\tilde{U}}(y) = \begin{cases} F_{B_i}(a^{-1}(y)) & y\in[0,\hat{\rho}_i) \\ 1 & y\in[\hat{\rho}_i,1] \\ 1-F_{B_i}(a^{-1}(y-1)) & y\in(1,1+\hat{\rho}_i]. \end{cases} \tag{3.13}$$

This distribution is referred to as a generalized trapezoidal distribution.

The distribution of B_i needs to be a continuous distribution, i.e. it should have a density. The expression in (3.13) seems simple at first sight. However, $a(x)$ needs to be calculated as well as its inverse, which does not always provide closed-form results. Note that the function is increasing for $y\in[0,\hat{\rho}_i]$ and decreasing for $y\in[1,1+\hat{\rho}_i]$.

Proposition 1. *The heavy traffic limit of the waiting time distribution in a SJF queueing system is a generalized trapezoidal distribution times a gamma distribution. The probability density function of the trapezoidal distribution is given in Equation* (3.13) *with* $a(x) = \hat{\lambda}_i\mathbb{E}[B_i\mathbf{1}_{\{B_i\leq x\}}]$. *The parameters of the gamma distribution are given in* (2.14).

Because the trapezoidal distribution depends on the service time distribution, below a number of examples with different distributions for B_i are given.

Example: exponential service time distribution

The exponential distribution is a common choice for the service time distribution in queueing systems because it has some nice properties. Suppose B_i is exponentially distributed with parameter b_i, then $\mathbb{E}[B_i] = 1/b_i$ and $f_{B_i}(x) = b_i e^{-b_i x}$. First calculate

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

So $a(x) = \hat{\rho}_i(1 - e^{-b_i x}(1 + b_i x))$, to calculate a^{-1}, solve $a(x) = y$ for x

$$\begin{aligned}
&=> \hat{\rho}_i(1 - e^{-b_i x}(1 + b_i x)) = y \\
&=> 1 - y/\hat{\rho}_i = (1 + b_i x)e^{-b_i x} \\
&=> e^{-1}(1 - y/\hat{\rho}_i) = (1 + b_i x)e^{-(1+b_i x)} \\
&=> -e^{-1}(1 - y/\hat{\rho}_i) = te^t, \qquad t = -(1 + b_i x) \\
&=> t = W_{-1}(-e^{-1}(1 - y/\hat{\rho}_i)) \\
&=> x = -\frac{W_{-1}(-e^{-1}(1 - y/\hat{\rho}_i)) + 1}{b_i} = a^{-1}(y).
\end{aligned}$$

W(y) is the solution of the equation $y = xe^x$ and is known as the Lambert W function. The equation xe^x may have multiple solutions, we need the solutions for real $x \leq -1$, this is denoted $W_{-1}(y)$. It decreases from $W_{-1}(-1/e) = -1$ to $W_{-1}(0^-) = -\infty$.

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

$$f_{\tilde{U}}(y) = \begin{cases} 1 - e^{W_{-1}(-e^{-1}(1-y/\hat{\rho}_i))+1} & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ e^{W_{-1}(-e^{-1}(1-(y-1)/\hat{\rho}_i))+1} & y \in (1, 1 + \hat{\rho}_i]. \end{cases} \tag{3.14}$$

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 3.7, 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.

Example: uniform service time distribution

The uniform distribution is a simple distribution with SCV < 1. This can be used to see the effect of the SCV on the waiting time distribution. Suppose B_i has a uniform distribution with parameters a_i and b_i, then $\mathbb{E}[B_i] = (a_i + b_i)/2$ and $f_{B_i}(x) = 1/(b_i - a_i)$. 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 $a(x) = \hat{\rho}_i \frac{x^2 - a_i^2}{b_i^2 - a_i^2}$, now find $a^{-1}(y)$

$$\begin{aligned}
&\hat{\rho}_i \frac{x^2 - a_i^2}{b_i^2 - a_i^2} = y \\
&=> {}^{y}/_{\hat{\rho}_i} \left(b_i^2 - a_i^2\right) + a_i^2 = x^2 \\
&=> a^{-1}(y) = x = \sqrt{{}^{y}/_{\hat{\rho}_i} \left(b_i^2 - a_i^2\right) + a_i^2}.
\end{aligned}$$

Because $F_{B_i}(x) = \frac{x - a_i}{b_i - a_i}$,

$$f_{\tilde{U}}(y) = \begin{cases} \frac{\sqrt{{}^{y}/_{\hat{\rho}_i}\left(b_i^2 - a_i^2\right) + a_i^2} - a_i}{b_i - a_i} & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ 1 - \frac{\sqrt{{}^{(y-1)}/_{\hat{\rho}_i}\left(b_i^2 - a_i^2\right) + a_i^2} - a_i}{b_i - a_i} & y \in (1, 1 + \hat{\rho}_i]. \end{cases}$$

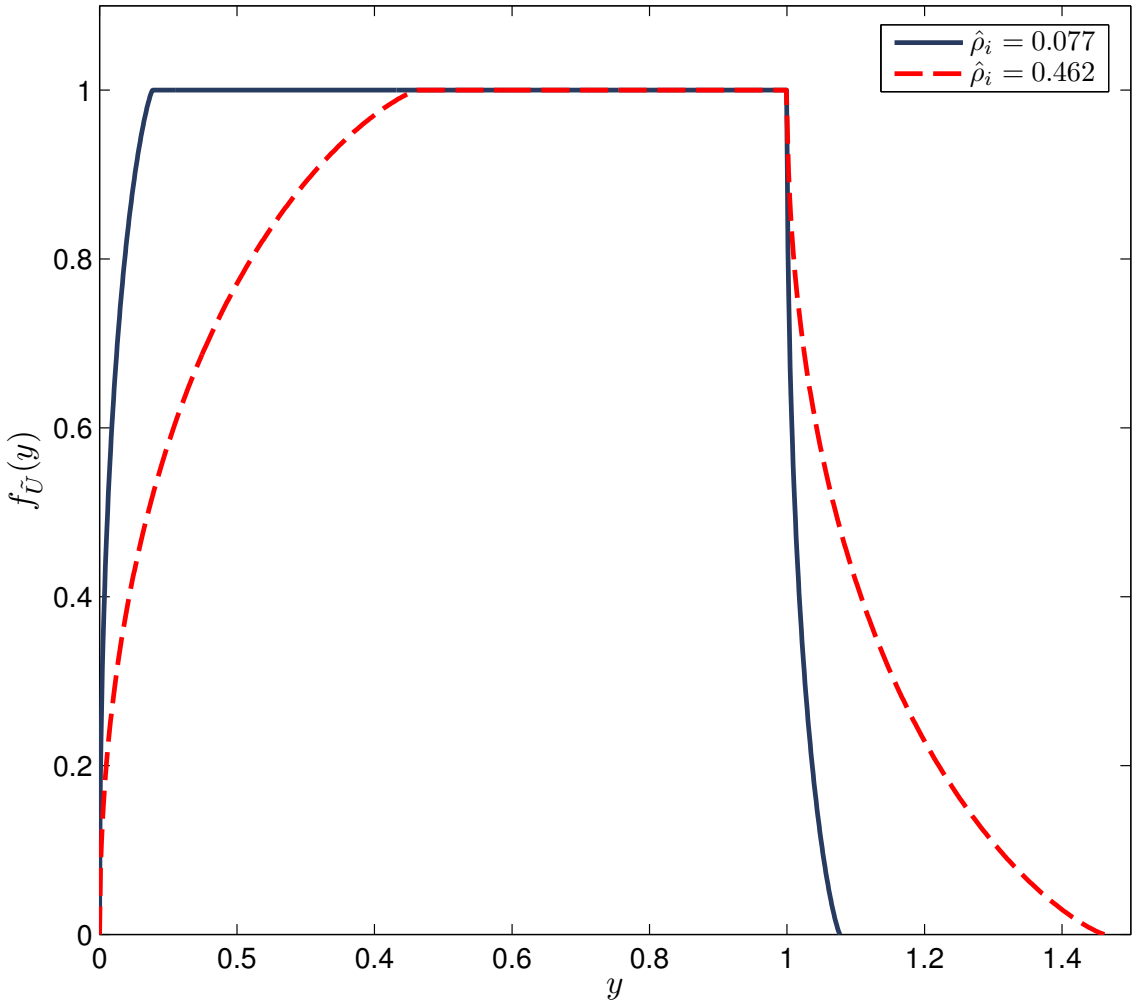


Figure 3.7: Probability density function of $\tilde{U}$ with exponental service times in a SJF polling system

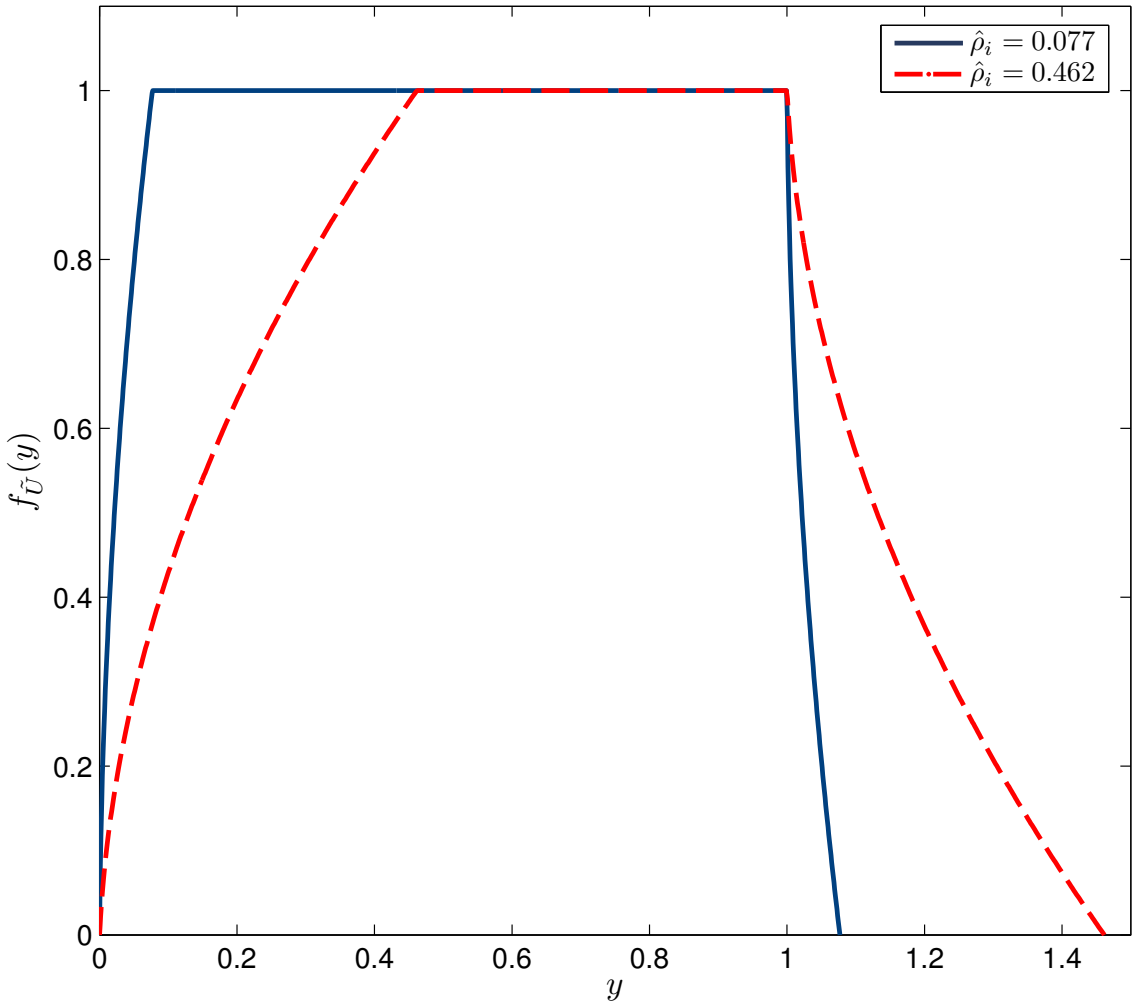


Figure 3.8: Probability density function of $\tilde{U}$ with uniform service times in a SJF polling system

Figure 3.8 depicts the probability density function for two different values of $\hat{\rho}_i$. The SCV of the uniform service time distribution is equal to 0.25. This picture looks similar to Figure 3.7, but the curved lines are slightly less curved.

Example: pareto distributed service times

The pareto distribution is also suitable as example distribution because of its easy form. However, the moments of the distribution may not exist which can be a problem for the validity of the heavy traffic limits. The pareto distribution for the service times gives some interesting properties about the relation between the waiting time distribution and the SCV of the service time distribution. Suppose B_i is pareto distributed with parameters a_i and b_i, then $\mathbb{E}[B_i^k] = \frac{a_i b_i^k}{a_i - k}$, $a_i > k$ are the kth moments. The probability density function is given by $f_{B_i}(x) = a_i b_i^{a_i} x^{-(a_i+1)}$, $x \geq b_i$. Finally the SCV equals $c_{B_i}^2 = \frac{1}{a_i(a_i-2)}$, $a_i > 2$. Note that if $a_i \to \infty$ the SCV equals 0 and if $a \downarrow 2$ the SCV goes to infinity.

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 $F_{B_i}(x) =$

$1-(b_i/x)^{a_i}$, $x \geq b_i$, this gives

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

With this formula it is possible to see what happens if the SCV goes to 0, so if $a_i \to \infty$.

$$\begin{aligned} \lim_{a_i\to\infty} f_{\tilde{U}}(y) &= \begin{cases} \lim_{a_i\to\infty}(1-(1-y/\hat{\rho}_i)^{\frac{-1}{1-a_i}}) & y \in [0,\hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ \lim_{a_i\to\infty}((1-(y-1)/\hat{\rho}_i)^{\frac{-1}{1-a_i}}) & y \in (1, 1+\hat{\rho}_i] \end{cases} \\ &= \begin{cases} y/\hat{\rho}_i & y \in [0,\hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ 1-(y-1)/\hat{\rho}_i & y \in (1, 1+\hat{\rho}_i]. \end{cases} \end{aligned}$$

This equation is equal to Equation (3.9), which means that the variation is not exactly zero, but there are tiny differences in service duration. These differences cause the service order to become random and thus do we find the waiting time distribution that we also found in case of ROS.

We also want to know what happens if the variation is extreme, so take $a_i \downarrow 1$. In that case the second moment and SCV of the service time distribution do not exist, i.e. the second moment and SCV equal infinity. Note that the second moment needs to be finite in order to find the heavy traffic limit of the cycle time distribution, because of the definition of σ^2. For pareto distributed service times $a_i > 2$ is needed.

$$\begin{aligned} \lim_{a\downarrow 1} f_{\tilde{U}}(y) &= \begin{cases} \lim_{a_i\downarrow 1}(1-(1-y/\hat{\rho}_i)^{\frac{-1}{1-a_i}}) & y \in [0,\hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ \lim_{a\downarrow 1}((1-(y-1)/\hat{\rho}_i)^{\frac{-1}{1-a_i}}) & y \in (1, 1+\hat{\rho}_i] \end{cases} \\ &= \begin{cases} 1 & y \in (0,\hat{\rho}_i) \quad \text{(exclude 0 from interval)} \\ 1 & y \in [\hat{\rho}_i, 1] \\ 0 & y \in (1, 1+\hat{\rho}_i]. \end{cases} \end{aligned} \tag{3.15}$$

This is the probability density function of a uniform distribution on the interval $(0,1]$. So the waiting time distribution in heavy traffic is a uniform times a gamma distribution, with the uniform distribution on $(0,1]$. This suggests that the waiting time can be close to 0, which can indeed happen if the customer with the shortest job arrives just before the server starts serving his queue. It also suggests that the waiting time cannot be longer than a residual cycle. This is interesting, since the customer with the longest queue can arrive just after the server started serving his queue and he has to wait $(1+\hat{\rho}_i)$ times the residual cycle. A possible explanation for the fact that this does never happen is the fact that large jobs are much less common than small jobs, they do not influence the waiting time distribution.

Note that this results can be correct, and used to find the waiting time distribution assuming that the cycle time distribution can be found. The cycle time distribution depends on the second moment of the service time distribution, so other methods using a different type of scaling are needed to find it.

Figure 3.9 shows the pdf of $\tilde{U}$ for a system with pareto distributed service times, where the pareto distribution has a SCV equal to 4 and only the first two moments exist. The figure does not differ much from Figures 3.7 and 3.8.

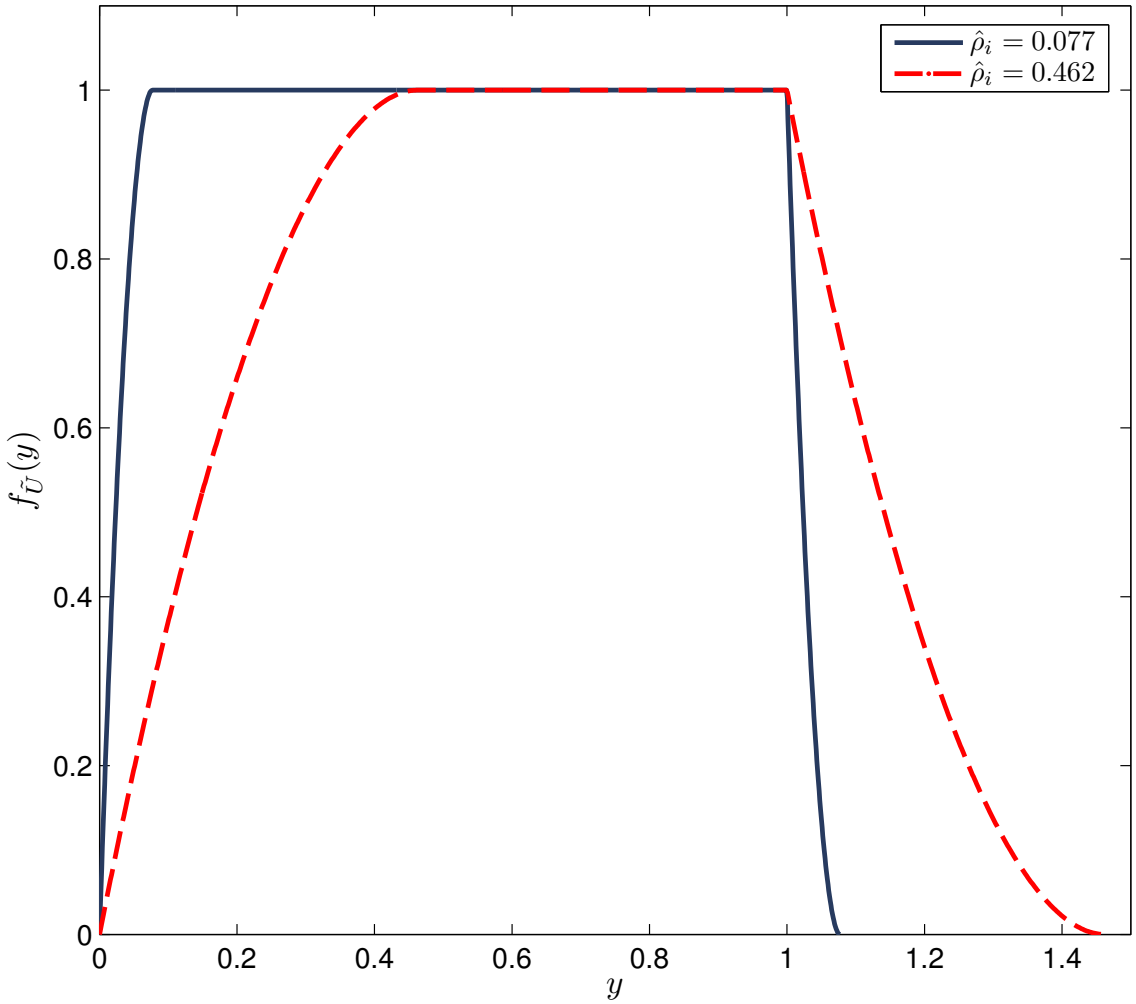


Figure 3.9: Probability density function of $\tilde{U}$ with pareto service times in a SJF polling system

3.3.2 General load

The distribution of the waiting time in a polling model with SJF queueing policy and renewal arrivals, can be approximated by a generalized trapezoidal distribution times a gamma distribution. The probability density function of the trapezoidal distribution depends on the distribution of the service time and is given in Equation (3.13). For the parameters of the gamma distribution, we need to know the mean of the trapezoidal distribution and an approximation for the mean waiting time for all values of ρ. It is not possible to use Boon's approximation for the waiting time, because the service order depends on the service times of the customers. In Equation (2.7) an expression for the mean waiting time in an SJF polling system is given. The problem with this equation is the fact that $\mathbb{E}[C^{res}]$ is unknown. Thankfully, we also have Equation (2.2), which can be approximated by Boon's approximation. This can be used to approximate $\mathbb{E}[C^{res}]$, this gives

$$\mathbb{E}[W_{ia,SJF}] = \frac{\mathbb{E}[W_{i,Boon}]}{1+\rho_i}(1+\lambda_i \mathbb{E}[B_{i,1:2}]) \tag{3.16}$$

as an approximation for the waiting time. $\mathbb{E}[B_{i,1:2}]$ is the expectation of the minimum of two random variables with the same distribution as the service time distribution. For continuous distributions it is equal to $2 * \mathbb{E}[B_{i,\varphi}]$, where $\mathbb{E}[B_{i,\varphi}]$ is given in Equation (3.12), or it can also be calculated using Equation (3.20).

The mean of the generalized trapezoidal distribution can be found using

$$\mathbb{E}[\tilde{U}] = \int_0^{1+\hat{\rho}_i} x * f_{\tilde{U}}(x)\,\mathrm{d}x. \tag{3.17}$$

This integral is hard to calculate, for instance, it cannot be calculated in case of an exponential service distribution. Off course it can always be calculated numerically.

The distribution of the waiting time can be approximated using

$$\mathbb{P}[W_i < x] \approx \mathbb{P}[\tilde{U} I_{i,a} < (1-\rho)x],$$

where $\tilde{U}$ has a generalized trapezoidal distribution with probability density function given in Equation (3.13). $I_{i,a}$ has a gamma distribution with the following parameters

$$\alpha_a = \frac{\mathbb{E}[S]\delta}{\sigma^2} + 1 \quad \text{and} \quad \mu_{ia} = \mathbb{E}[\tilde{U}]\frac{\mathbb{E}[S]+\sigma^2}{\sigma^2(1-\rho)\,\mathbb{E}[W_{ia,SJF}]}. \tag{3.18}$$

These parameters are similar to the parameters given in Equation (2.23). Recall that the parameters of the gamma distribution were found by matching the expectation of a uniform times a gamma distribution with Boon's approximation for the waiting time by adjusting the μ of the gamma distribution. For SJF, we need to match the expectation of a trapezoidal distribution times a gamma distribution with the expectation of the waiting time given in Equation (3.16). We only need to adjust the μ, so we replace Boon's approximation with the approximation for the mean waiting time in a SJF system and we replace the expectation of the uniform distribution that appears in the μ given in (2.23) with $\mathbb{E}[\tilde{U}]$, given in Equation (3.17).

3.3.3 Numerical results

The accuracy of the waiting time approximation for SJF systems, is determined by comparing the approximation with simulation results. The approximation is again tested on the testbed given in

Table 3.1, but for the SCV of the service times, we do not take the values from the table. Because the distribution of the service time is needed for the approximation, we consider systems with uniformly distributed service times and systems with exponential service times. This means that the test is run twice. The parameters of the uniform distribution are chosen in such a way, that the mean service time at each queue is equal to the mean given in Table 3.1 and the SCV is equal to 0.25. The tables with the results are given in Appendix A.2.1 for the uniform service times and in Appendix A.2.2 for the exponential service times.

The results are similar to the LCFS and ROS results. We see again that the approximation for the mean waiting time works well for low and high loads, but the approximation for the waiting time distribution works best for high loads. This can be seen in Tables A.11 - A.15 and A.21 - A.25. Tables A.16 - A.18 and A.26 - A.28 show that for the approximation in SJF systems it is also better to have more queues and Poisson arrivals and the 80th percentile is again approximated the best for both uniform and exponential service times.

Overall the approximation performs better on systems with exponential service times. When the service times have a uniform distribution, the approximation seems to have trouble with the higher moments in systems with low loads. In the polling systems with a load of 0.1, the errors are higher than 25% in 43.3% of the cases, as can be found in Table A.13.

3.4 Processor sharing

3.4.1 Heavy traffic limits

For processor sharing, we look at the sojourn time distribution $T_{i,PS}$. The sojourn time is the time that a customer spends in the system, its expectation is the expected service time plus the expected waiting time. Note that for PS it is not clear what is meant by waiting time. In heavy traffic the waiting time distribution is equal to the sojourn time distribution, because the expected service time is negligible compared to the expected waiting time.

The conditional LST of the waiting time given in Section 2.2.1 is equal for SJF and PS, but with different $\varphi(s,x)$. For the conditional LST of the waiting time in heavy traffic we find Equation (3.10), in heavy traffic this expression is also the LST of the sojourn time. Thus, $\tilde{T}_{i,PS}(s|x) = \tilde{W}_{i,SJF}(s|x)$ in heavy traffic. In this case $\mathbb{E}[B_{i,\varphi(x)}] = \mathbb{E}[\min(B_i,x)]$, this can be calculated as follows

$$\mathbb{E}[\min(B_i,x)] = \int_0^x y f_{B_i}(y)\,\mathrm{d}y + x\int_x^\infty f_{B_i}(y)\,\mathrm{d}y, \tag{3.19}$$

with $f_{B_i}(x)$ the probability density function of the service time distribution, assuming that it exists. The unconditional expectation is given by

$$\mathbb{E}[B_{i,\varphi}] = \int_0^\infty f_{B_i}(x)\,\mathbb{E}[\min(B_i,x)]\,\mathrm{d}x. \tag{3.20}$$

The only difference with the SJF case is the form of φ. The distribution of the sojourn time is in the PS case also a generalized trapezoidal distribution times a gamma distribution. The probability density function of the former is given in Equation (3.13). It depends on the form of φ through $a(x)$, which becomes $a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$. If the random variable B_i can only take values greater

than a certain value $m \geq 0$, for example if B_i has a uniform distribution with lower bound m, then $a(m) = \hat{\lambda}_i m$. This gives for the probability density function of $\tilde{U}$

$$f_{\tilde{U}}(y) = \begin{cases} F_{B_i}(a^{-1}(y)) & y \in [\hat{\lambda}_i m, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i m] \\ 1 - F_{B_i}(a^{-1}(y-1)) & y \in (1 + \hat{\lambda}_i m, 1 + \hat{\rho}_i], \end{cases} \tag{3.21}$$

where m is the minimum value that B_i can take.

Proposition 2. *The heavy traffic limit of the sojourn time distribution in a PS queueing system is a generalized trapezoidal distribution times a gamma distribution. The probability density function of the trapezoidal distribution is given in Equation* (3.21) *with* $a(x) = \hat{\lambda}_i \mathbb{E}[\min(B_i, x)]$ *and* m *the lowest possible value of* B_i*. The parameters of the gamma distribution are given in* (2.14)*.*

Below are some examples with different distributions for the service time. The same distributions were discussed for SJF in Section 3.3.1.

Example: exponential service time distribution

Suppose B_i is exponentially distributed with parameter b_i. Then

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

so $a(x) = \hat{\rho}_i(1 - e^{-b_i x})$. Solve $a(x) = y$ for x to find $a^{-1}(y) = \frac{\ln(1 - y/\hat{\rho}_i)}{-b_i}$. Now use this in Equation (3.13) to find the probability density function of the generalized trapezoidal distribution $\tilde{U}$:

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

This equation is equal to the density function of $\tilde{U}$ with ROS given in (3.9). This means that, if the service time distribution is exponential, the sojourn time distribution at a queue with random service order is equal to the sojourn time distribution at a queue with processor sharing. This is also pointed out in [6]. Note that for small $\hat{\rho}_i$, the distribution of $\tilde{U}$ is close to a uniform distribution, this is also the case for a SJF queue. Figure 3.10 illustrates the trapezoidal shape for large $\hat{\rho}_i$ and the close to uniform shape for small $\hat{\rho}_i$.

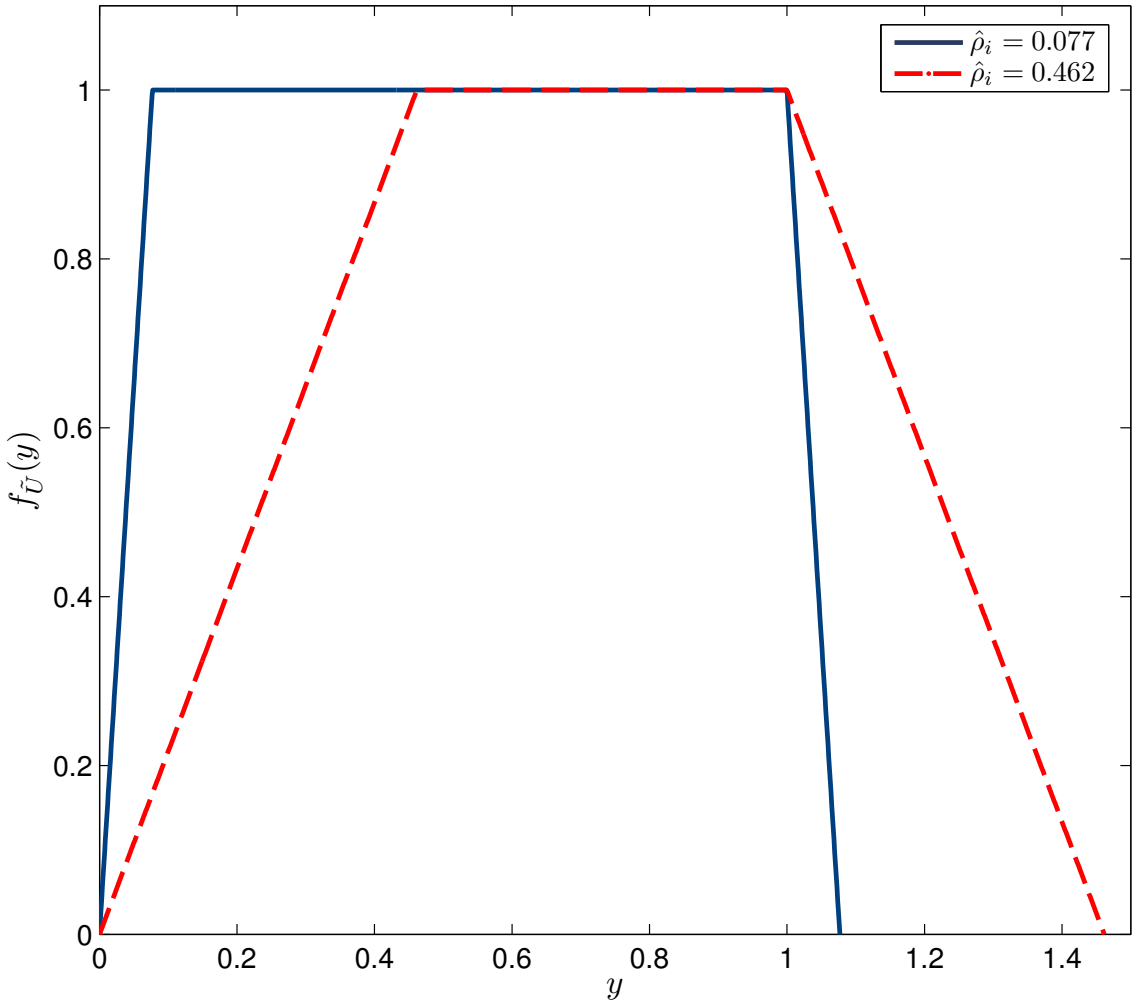


Figure 3.10: Probability density function of $\tilde{U}$ with exponential service times in a PS polling system

Example: uniform service time distribution

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

$$\begin{aligned} a(x) &= \hat{\lambda}_i \mathbb{E}[\min(B_i, x)] = \hat{\lambda}_i \left(\int_{a_i}^{x} y/(b_i - a_i)\, \mathrm{d}y + x \int_{x}^{b_i} 1/(b_i - a_i)\, \mathrm{d}y \right) \\ &= \hat{\lambda}_i \left(\frac{x^2}{2(b_i - a_i)} - \frac{a_i^2}{2(b_i - a_i)} + \frac{2xb_i}{2(b_i - a_i)} - \frac{2x^2}{2(b_i - a_i)} \right) \\ &= \frac{-\hat{\lambda}_i}{2(b_i - a_i)} \left(a_i^2 - 2b_i x + x^2 \right) \\ &= \frac{-\hat{\rho}_i}{b_i^2 - a_i^2} \left(a_i^2 - 2b_i x + x^2 \right). \end{aligned}$$

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

$$\begin{aligned} a^{-1}(y) &= \left(2b_i \pm \sqrt{4b_i^2 - 4(a_i^2 + y/\hat{\rho}_i(b_i^2 - a_i^2))} \right)/2 \\ &= b_i \pm \sqrt{b_i^2 - a_i^2 - yb_i^2/\hat{\rho}_i + ya_i^2/\hat{\rho}_i} \\ &= b_i - \sqrt{(1 - y/\hat{\rho}_i)(b_i^2 - a_i^2)}. \end{aligned}$$

The final equality holds, because $x \in [a_i, b_i]$.

In this case $a(a_i) \neq 0$, but the probability density function needs to start at 0. This means we cannot take $y = 0$, 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

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

Figure 3.11 illustrates the shape of the pdf of $\tilde{U}$, when the $c_{B_i}^2$ of the uniform service time distribution is equal to 0.25 and for two different values of $\hat{\rho}_i$. The direction of the curves in the curved lines is opposite to the direction of the lines in Figure 3.8, for SJF. This indicates that for PS, the mean of $\tilde{U}$ is higher than for SJF, this is also the case for the mean waiting times, see Equations (2.6) for PS and (2.7) for SJF.

Example: pareto distributed service times

If B_i is a pareto distributed random variable with parameters a_i and b_i, then

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

Some basic calculations lead to

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

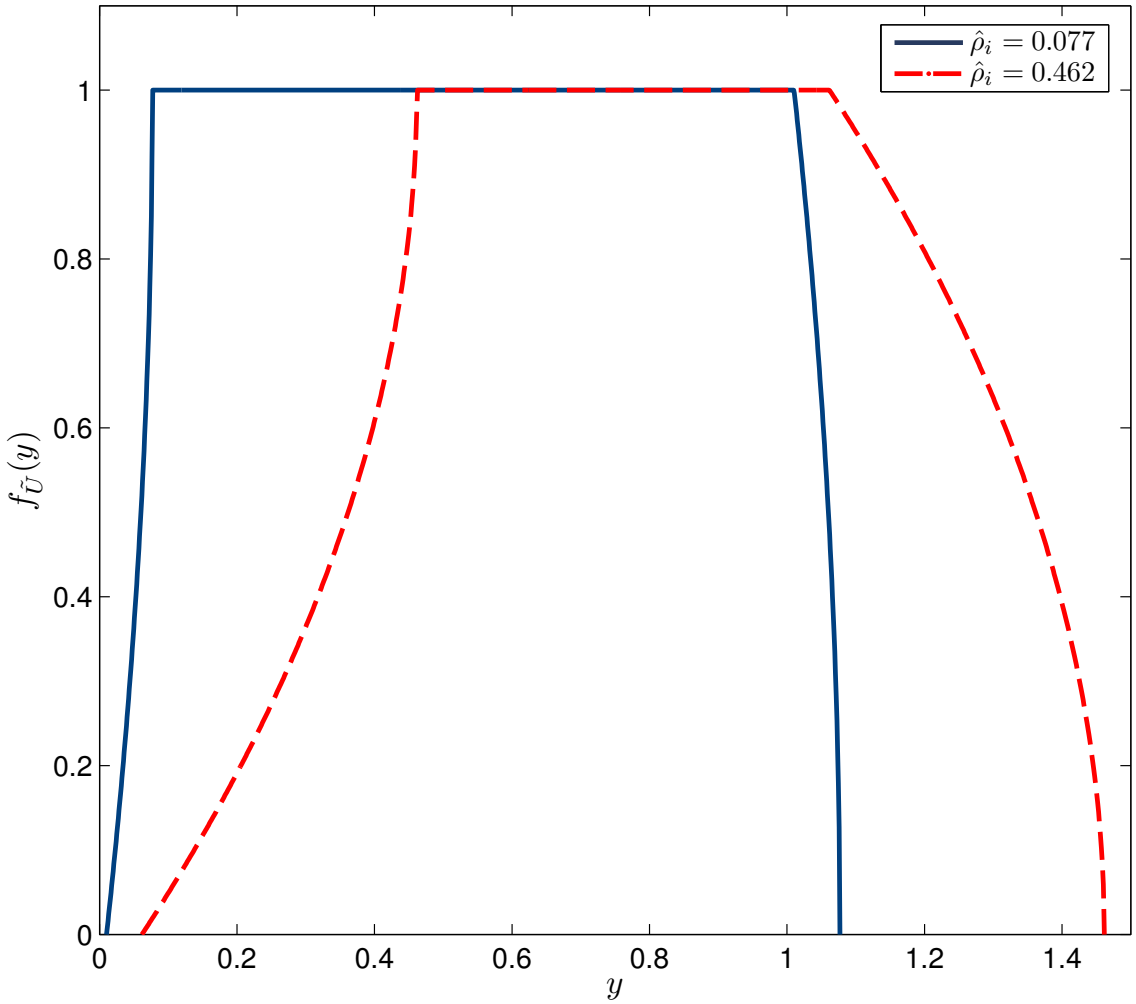


Figure 3.11: Probability density function of $\tilde{U}$ with uniform service times in a PS polling system

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

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

Take $a_i \to \infty$ to see what the behavior of this distribution is, if the SCV of the pareto distribution goes to zero. Note that $\mathbb{E}[B_i] = \frac{a_i b_i}{a_i - 1}$, this gives $b_i = \frac{\mathbb{E}[B_i] a_i}{a_i - 1}$ and $\lim_{a_i \to \infty} b_i = \mathbb{E}[B_i]$. From this it can be concluded that if a_i goes to infinity, $\tilde{U}$ has a uniform distribution on the interval $[\hat{\rho}_i, 1+\hat{\rho}_i]$. This result makes sense, because all customers have practically the same job length. They leave the queue all at the same time, so when they leave all customers are served. The time to serve all customers at a queue is $\hat{\rho}_i$ times the residual cycle time, the time that a lucky customer has to wait when he arrives just before the server arrives at his queue. An unlucky customer has to wait a residual cycle plus the time that it takes to serve him and the other customers at his queue.

The other extreme case to consider is the case where $a_i \downarrow 1$. Recall that in this case, the SCV and the second moment of the pareto distribution are infinity, which can give problems finding the heavy traffic limit of the cycle time distribution. Taking the limit of $a_i \downarrow 1$ we get Equation (3.15), a uniform distribution on the interval $(0,1]$. Here we give the same explanation as we did for SJF, the very rare long jobs do not affect the sojourn time distribution of the short jobs.

Figure 3.12 shows the pdf of $\tilde{U}$ if the pareto service time distribution has a squared coefficient of variation equal to 4, for two different values of $\hat{\rho}_i$.

Example: deterministic or double deterministic service times

With the pareto distribution some interesting limits were explored. However, the pareto distribution is a continuous distribution. We now investigate what will happen if the service times are deterministic, so they have zero variation. Another interesting case with deterministic service times is the invented double deterministic distribution. This distribution equals a small value a_i with probability p_i close to 1, or a large value b_i with probability $1-p_i$. Note that deterministic distributions cannot be used if the queueing policy is SJF, because there is no way to determine which job is the shortest one.

If the service times are deterministic, we have $a(x) = \hat{\lambda}_i \mathbb{E}[\min(B_i, x)] = \hat{\lambda}_i x$, for $x \leq \mathbb{E}[B_i]$. It is easy to see that $a^{-1}(y) = y/\hat{\lambda}_i$. Because

$$F_{B_i}(x) = \begin{cases} 0 & x < \mathbb{E}[B_i] \\ 1 & x \geq \mathbb{E}[B_i], \end{cases}$$

we have,

$$F_{\tilde{U}}(y) = \begin{cases} 1 & y \in [\hat{\rho}_i, 1+\hat{\rho}_i] \\ 0 & \text{otherwise.} \end{cases}$$

This is the same result as was found in the example with the pareto distribution, the same explanation can be given here.

If the service times have a double deterministic distribution, then

$$\mathbb{E}[\min(B_i, x)] = \begin{cases} x & \text{if } x = b_i \\ a_i & \text{if } x = a_i \end{cases} = (1-p_i)x + p_i a_i.$$

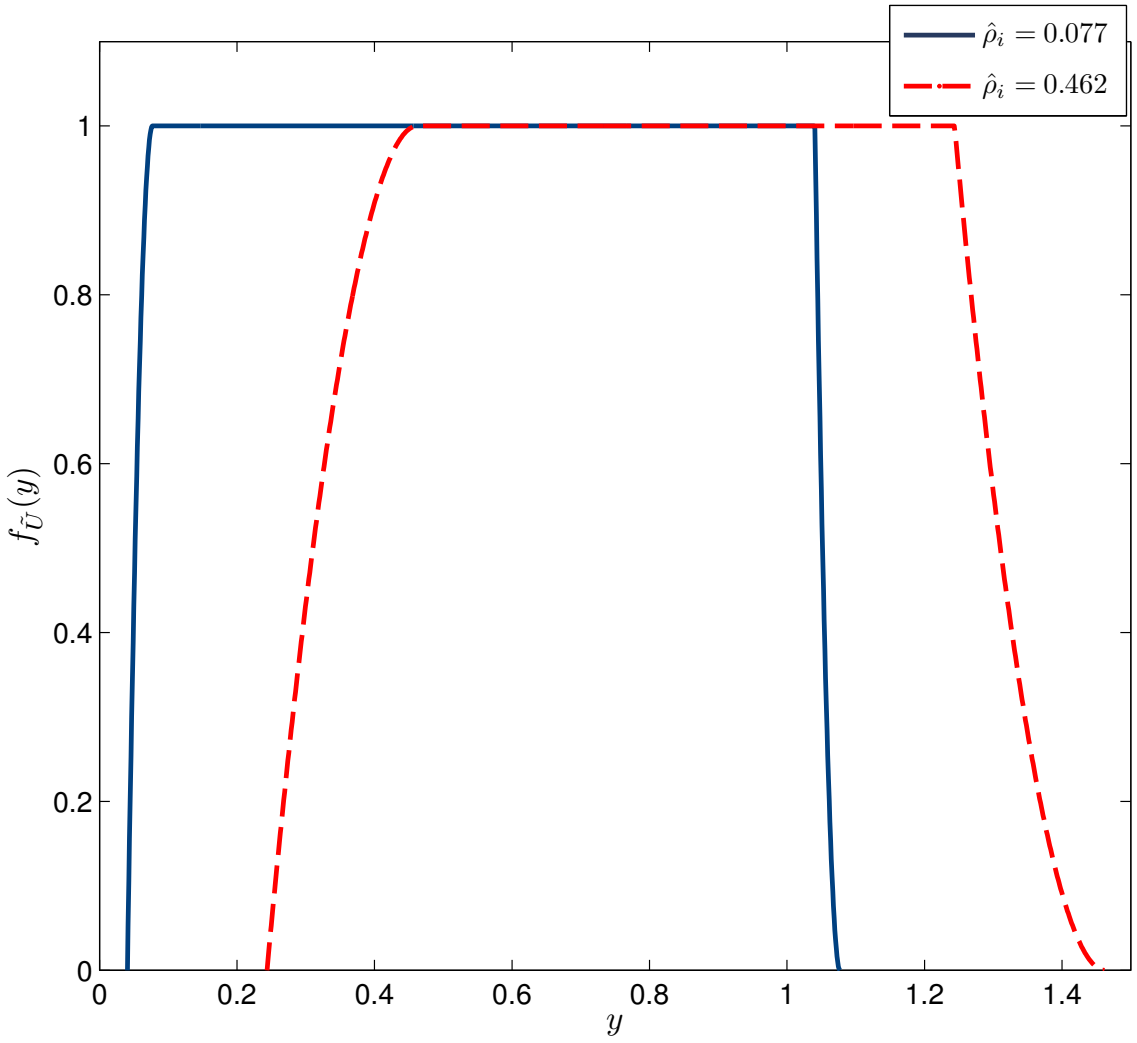


Figure 3.12: Probability density function of $\tilde{U}$ with pareto service times in a PS polling system

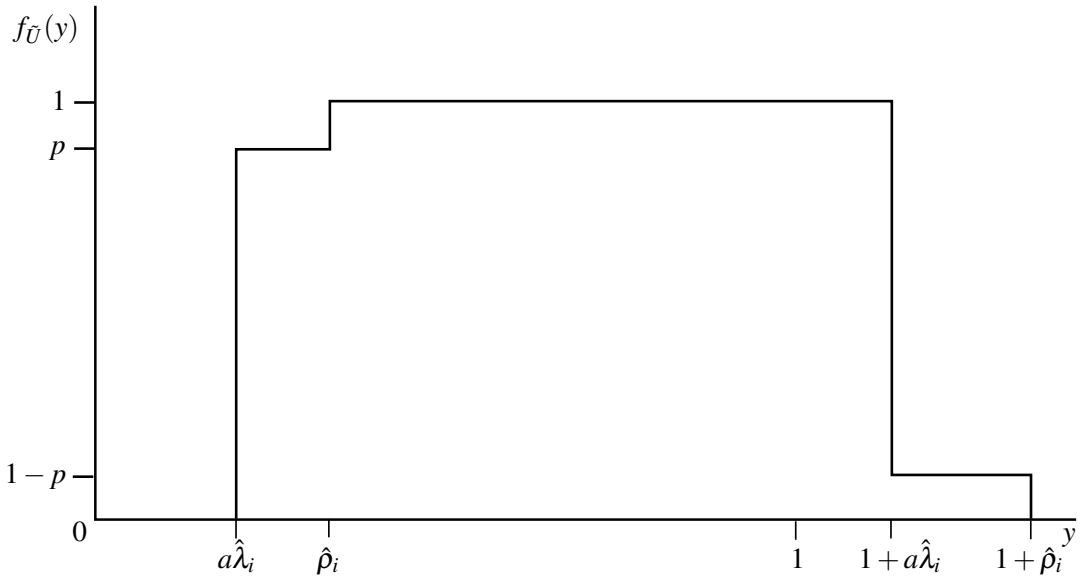


Figure 3.13: Probability density function of $\tilde{U}$ in a polling system with PS queues and double deterministic service times

This gives $a(x) = \hat{\lambda}_i((1-p_i)x + p_i a_i)$ and thus $a^{-1}(y) = \frac{y/\hat{\lambda}_i - p_i a_i}{1-p_i}$. The double deterministic distribution has the following cumulative distribution function

$$F_{B_i}(x) = \begin{cases} 0 & x < a \\ p_i & a_i \le x < b_i \\ 1 & x \ge b_i. \end{cases}$$

Note that $y \ge a(a_i) = \hat{\lambda}_i a_i$, so

$$F_{\tilde{U}}(y) = \begin{cases} p & y \in [a\hat{\lambda}_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1+a\hat{\lambda}_i] \\ 1-p & y \in (1+a\hat{\lambda}_i, 1+\hat{\rho}_i]. \end{cases}$$

Figure 3.13 illustrates the step shape of the distribution.

3.4.2 General load

The sojourn time distribution in a polling system with renewal arrivals and general load $\rho < 1$, with processor sharing can be approximated by a general trapezoidal distribution times a gamma distribution. The expectation of this distribution has to coincide with an approximation of the expectation of the sojourn time. This expectation can be found using Equations (2.2) and (2.6) and Boon's approximation.

$$\mathbb{E}[T_{ia,PS}] = \mathbb{E}[B_i] + \frac{\mathbb{E}[W_{i,Boon}]}{1+\hat{\rho}_i}(1 + 2\lambda_i \mathbb{E}[B_{i,1:2}]). \tag{3.22}$$

$\mathbb{E}[B_{i,1:2}]$ can be calculated using (3.20). For the parameters of the gamma distribution, we can use Equation (3.18), we only have to replace $\mathbb{E}[W_{ia.SJF}]$ with $\mathbb{E}[T_{ia,PS}]$. This leads to the following approximation for the sojourn time distribution

$$\mathbb{P}[T_i < x] \approx \mathbb{P}[\tilde{U} I_{i,a} < (1-\rho)x],$$

where $\tilde{U}$ has a generalized trapezoidal distribution with probability density function given in (3.21). $I_{i,a}$ has a gamma distribution with the following parameters

$$\alpha_a = \frac{\mathbb{E}[S]}{\sigma^2} + 1 \quad \text{and} \quad \mu_{ia} = \mathbb{E}[\tilde{U}] \frac{\mathbb{E}[S] + \sigma^2}{\sigma^2 (1-\rho)\, \mathbb{E}[T_{ia,PS}]}.$$

$\mathbb{E}[\tilde{U}]$ is given in Equation (3.17). The fact that the sojourn time distribution is a generalized trapezoidal distribution times a gamma distribution is a remarkable result. In heavy traffic the sojourn time and the waiting time are identically distributed, since we use the heavy traffic limit, we found this distribution for the sojourn time.

3.4.3 Numerical results

A simulation study is used to gain an insight in the accuracy of the sojourn time approximation. The accuracy test is run on the testbed given Table 3.1 twice. The first time, the SCVs of the service time distributions given in the table are replaced by 0.25 and a uniform distribution is used. The second time, the SCVs of the service time distributions are replaced by 1 and an exponential distribution is used. The tables with the results can be found in Appendix A.3.1 for the uniform service times and in Appendix A.3.2 for the exponential service times.

The results show the same things as we saw with other queueing disciplines. Tables A.31 and A.41 show that the mean sojourn times are approximated the worst for systems with a load of 0.5. From Tables A.32 - A.35 and A.42 - A.45 it can be seen that the approximation of the sojourn time distribution works better for systems with high loads. It is interesting to note that for uniform service time a load of 0.1 seems to be better than a load of 0.3 and for exponential service times this same effect cannot be observed. Tables A.36 - A.38 and A.46 - A.48 display the fact that a large number of queues is better for the performance of the approximation. Interestingly enough, a low SCV of the interarrival times seems to be better when the service time distribution is exponential. For systems with uniform service times, Poisson arrivals are better. In both cases, the 80th percentile is approximated best. Exponential service times are overall better for the approximation.

Chapter 4

Globally gated service discipline

In this chapter we look at polling systems with a globally gated service discipline. For FCFS an approximation for the mean waiting time using interpolation is derived. For LCFS and ROS, the LST of the waiting time distribution in heavy traffic is obtained. The approximation for the waiting time is then plugged into this distribution, to approximate the distribution of the waiting time for all loads. The performance of the approximation is analyzed by comparing the approximation to simulation results.

4.1 First come first served

4.1.1 Mean waiting time

For the globally gated system, we already have the mean waiting times in the light and heavy traffic limits. As in Section 2.2.4 for gated, interpolation between light and heavy traffic can be applied to find an approximation for the mean waiting time for general load $0 < \rho < 1$. We will not use the second term from Equation (2.22), this term matches the derivative of the approximation with the derivative of the light traffic limit. It is hard to find, moreover, for renewal arrivals in a gated system an exact expression cannot be found. So we will use a first order approximation for the waiting time $\mathbb{E}[W_{i,app}]$, which becomes

$$\mathbb{E}[W_{i,app}] = \frac{a_i + b_i\rho}{1-\rho}, \tag{4.1}$$

with

$$a_i = \mathbb{E}[S^{res}] + \sum_{j=1}^{i-1} \mathbb{E}[S_j]$$
$$b_i = \frac{1}{2}\left(1 + 2\sum_{j=1}^{i-1} \hat{\rho}_j + \hat{\rho}_i\right)\left(\frac{\sigma^2}{2} + \mathbb{E}[S]\right) - a_i.$$

When $\rho = 0$ the expected waiting time is equal to a_i, which is the LT limit given in (2.21). In HT, the expected waiting time is equal to the HT limit of the wating given in (2.19). To see this, note that (2.19) is the expectation of the scaled waiting time $(1-\rho)W$.

4.1.2 Waiting time distribution

From the form of Equation (2.17) we can see that the distribution of the waiting time in heavy traffic is a uniform times a gamma distribution. We have the following approximation for the waiting time distribution in polling systems with renewal arrivals, $\rho < 1$, globally gated service discipline and FCFS queueing discipline:

$$\mathbb{P}[W_i < w] \approx \mathbb{P}[UI_{i,a} < (1-\rho)w],$$

where U is a uniformly distributed random variable on $[\sum_{j=1}^{i} \hat{\rho}_j, 1 + \sum_{j=1}^{i-1} \hat{\rho}_j]$ and $I_{i,a}$ is a gamma distributed random variable. The parameters of the gamma distribution are similar to those given in Equation (2.23), the α and μ used for the gated polling system. For the globally gated polling system, the α remains exactly the same, with $\delta = 2$ in this case, this follows from Equation (2.18). μ depends on the expectation of the uniform distribution, so we just need to replace this expectation, this leads to

$$\alpha := \frac{2\,\mathbb{E}[S]}{\sigma^2} + 1 \quad \text{and} \quad \mu_i := \frac{\sum_{j=1}^{i} \hat{\rho}_j + 1 + \sum_{j=1}^{i-1} \hat{\rho}_j}{2} \frac{2\,\mathbb{E}[S] + \sigma^2}{\sigma^2 (1-\rho)\,\mathbb{E}[W_{i,app}]}. \tag{4.2}$$

$\mathbb{E}[W_{i,app}]$ is given in (4.1).

The kth moment is given by

$$\begin{aligned}
\mathbb{E}[W_{i,FCFS}^k] &= \frac{1}{(1-\rho)^k} \frac{\left(1 + \sum_{j=1}^{i-1} \hat{\rho}_j\right)^{k+1} - \left(\sum_{j=1}^{i} \hat{\rho}_j\right)^{k+1}}{\left(1 + \sum_{j=1}^{i-1} \hat{\rho}_j - \sum_{j=1}^{i} \hat{\rho}_j\right)(k+1)} \prod_{j=0}^{k-1} \frac{\alpha + j}{\mu_i} \\
&= \frac{1}{(1-\rho)^k} \frac{\left(1 + \sum_{j=1}^{i-1} \hat{\rho}_j\right)^{k+1} - \left(\sum_{j=1}^{i} \hat{\rho}_j\right)^{k+1}}{(1-\hat{\rho}_i)(k+1)} \prod_{j=0}^{k-1} \frac{\alpha + j}{\mu_i}.
\end{aligned}$$

4.1.3 Numerical results

To evaluate the approximation, it is applied to the test bed given in Table 3.1. The first three moments, and the $40, 50, \ldots, 90$ and 95 percentiles are calculated, as well as the cumulative distribution function. The same measures are also obtained by simulation to compare to the approximated values. The percentual errors are calculated using Equation (3.4) and the absolute maximum difference in cumulative distribution function is calculated using (3.5). The results are given in Tables B.1 - B.10. In Figure 4.1 the approximated and simulated probability density functions are plotted. The service order of the corresponding system is FCFS and the service discipline is globally gated. The load of the system is 0.8, arrivals are Poisson and the number of queues is 7. The plot correspondents to the first queue. The approximated curve nicely follows the simulated curve.

To see the difference between the globally gated and locally gated service discipline, we plot the waiting time distributions of all 7 queues in a locally gated polling system with a load of 0.95, because with higher load the difference is clearer, in one plot (Figure 4.2) and we do the same for the distributions of all 7 queues of a globally gated polling system, also with a load of 0.95 (Figure 4.3). The first figure shows that the waiting time distributions of the queues lie close together and the waiting time distribution of queue 1 is almost exactly equal to the waiting time distribution of queue 7, the same holds for the distributions at queues 2 and 4. This is caused by the fact that $\hat{\rho}_1 = \hat{\rho}_7$ and $\hat{\rho}_2 = \hat{\rho}_4$, so they have the same uniform distribution. The mean waiting

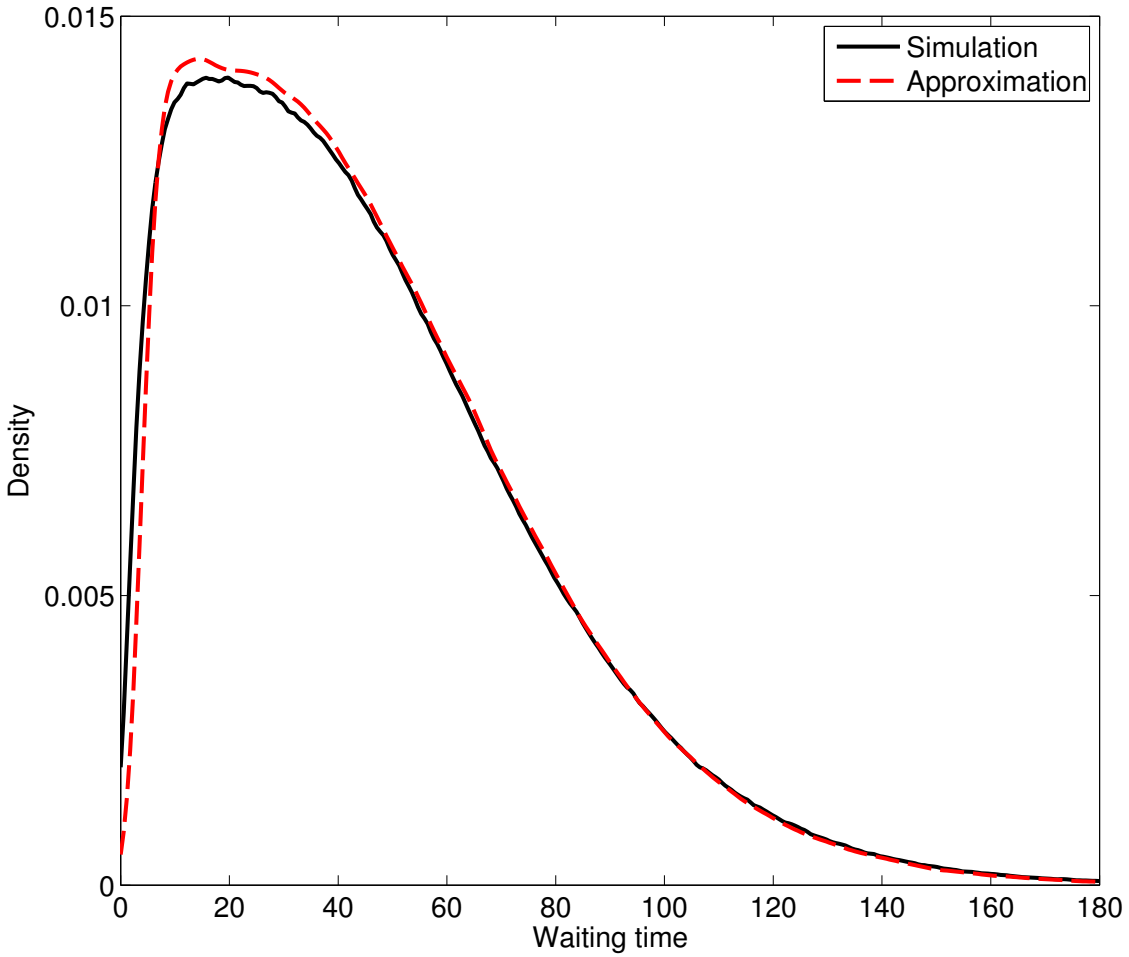


Figure 4.1: Approximated and simulated density function of the waiting time of an arbitrary queue in the example in Section 4.1.3

times at the queues lie close to each other. The second figure shows that with a globally gated service discipline, the higher the number of the queue is, the higher the mean (distribution shifted to the right) and variance (fatter tails) of the distribution of the waiting time at that queue are. The distributions of the first and the last queue are very different from each other.

Tables B.1 - B.3 show that the first three moments are approximated very well for $\rho \geq 0.7$. The first table shows that the mean is nicely approximated for all loads. The other two tables indicate that the approximation has troubles estimating higher moments with lower loads. The approximation of the third moment of the waiting time distribution in a system with load lower than 0.7 is not very accurate. Because the third moment is already hard to approximate, higher moments will be even harder to approximate. It is clear that the distribution is only accurate in heavy traffic.

Tables B.4 and B.5 confirm this conclusion. The first table shows that the approximation of the percentiles gets worse when the load decreases. From the other table we see that the maximum absolute difference between the approximated and simulated cumulative distribution function are larger for smaller ρ.

From Tables B.6 - B.8 we can see that the approximation works better when the number of queues is larger and that the best results are obtained when the arrivals are Poisson. The 80th percentile is approximated the best and lower percentiles are harder to approximate. These results were also

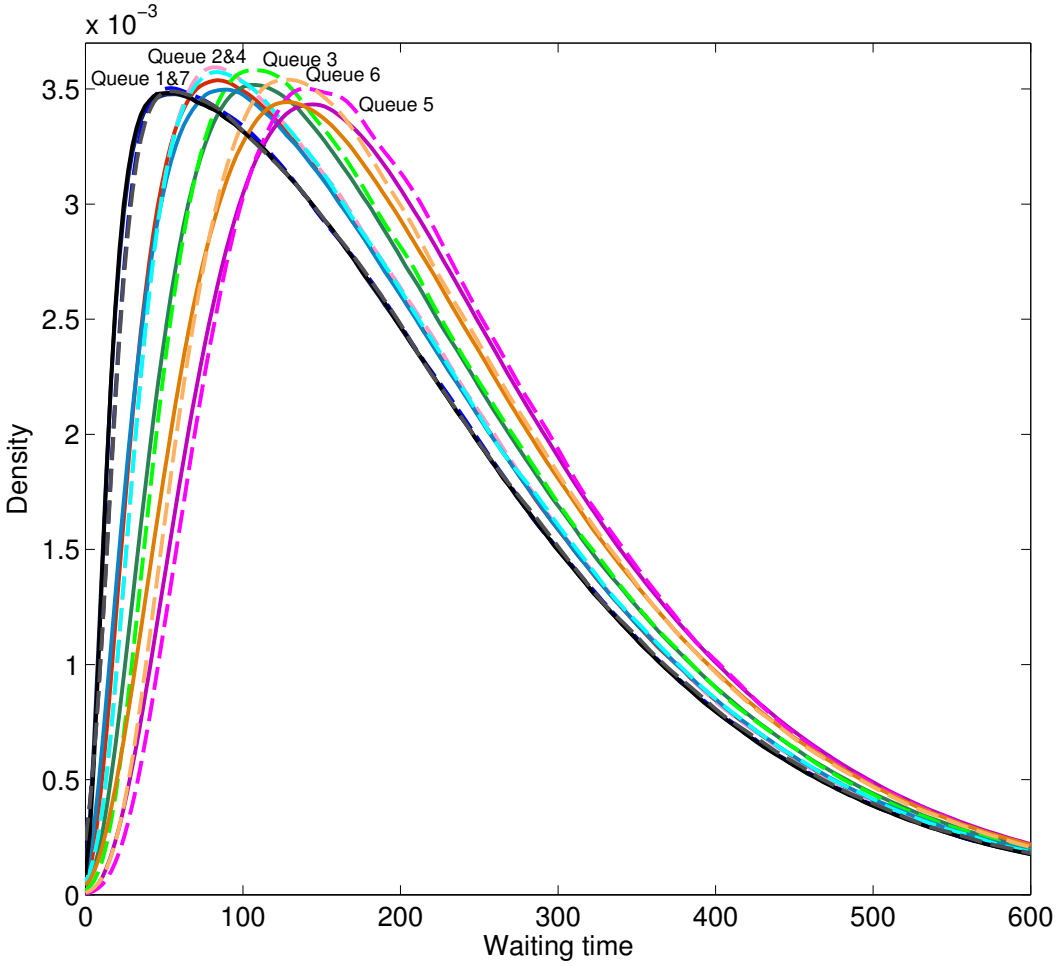


Figure 4.2: Approximated and simulated density functions of the waiting times of all the queues of an arbitrary polling system with locally gated service discipline

found for the locally gated service discipline.

The considered polling systems are asymmetric, different queues can have different accuracies. In the system with a locally gated service discipline, we found that the approximation was less accurate at queues with a higher mean service time. From Tables B.9 and B.10 the same thing cannot be concluded for the globally gated case. Queue 6, the queue with the highest mean service time is approximated the most accurate. Comparing these tables with the same tables for the LCFS and ROS systems , Tables B.19, B.20, B.29 and B.30, we see that the waiting time distribution at the first queue of the 3 queue system is always approximated the best. For the 7 queue system, the best approximated queues are 4 and 7. The worst approximated queues are number 2 in the 3 queue system and 3 in the 7 queue system. The fact that there are queues that are always approximated best and worst, the parameters and the position of the queues do have an influence on the accuracy. It is not clear how the accuracy is affected.

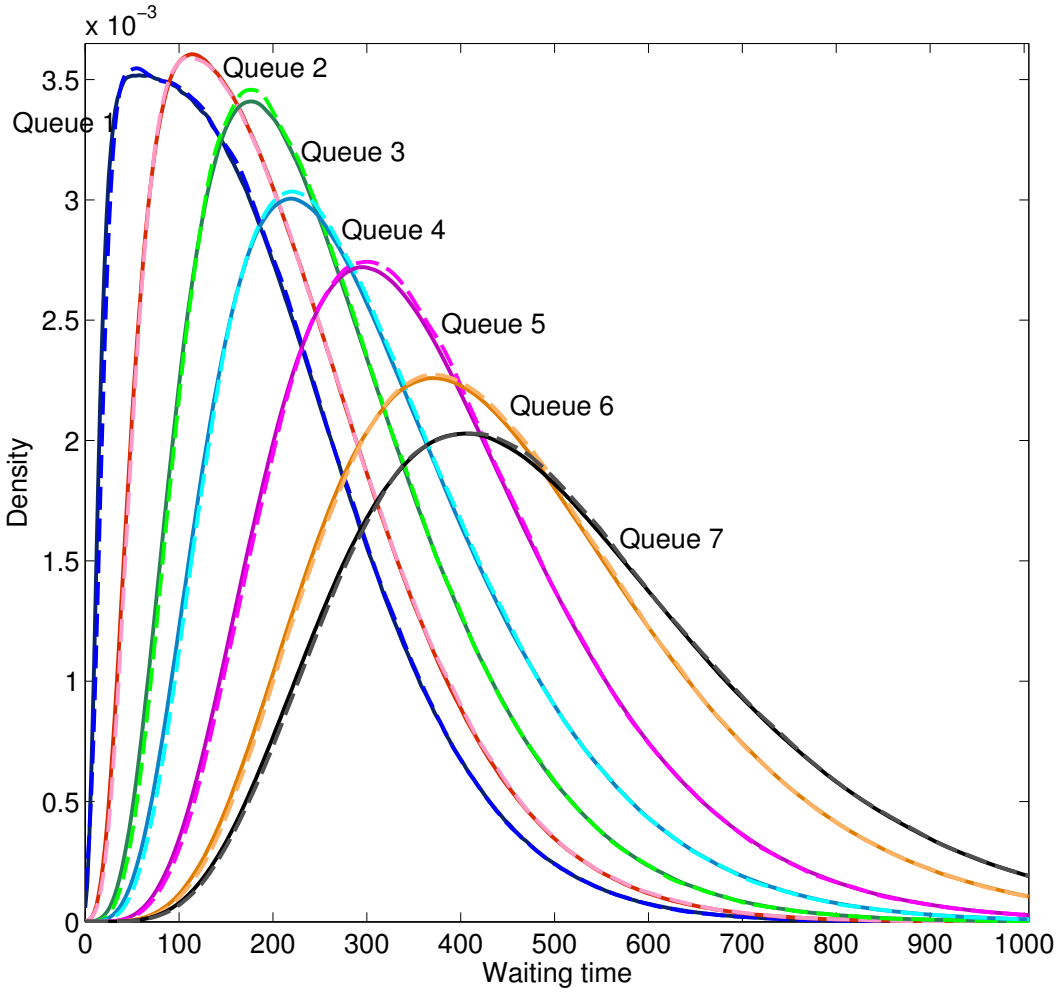


Figure 4.3: Approximated and simulated density functions of the waiting times of all the queues of an arbitrary polling system with globally gated service discipline

4.2 Last come first served

4.2.1 Heavy traffic limits

The LST of the scaled waiting time distribution in a LCFS polling system can be calculated using (2.16) with (2.10) and for the LST of the cycle time (2.15), because in a system with globally gated service the scaled cycle time is still a gamma distribution, but the parameters differ from the parameters in a system with a gated service discipline. This gives, using calculations similar to

those used to find Equation (3.1),

$$\begin{aligned}\tilde{W}^*_{i,LCFS}(s) &= \lim_{\rho\uparrow 1} W^*_{i,LCFS}(s(1-\rho)) \\ &= \lim_{\rho\uparrow 1} \frac{\prod_{j=1}^{i-1} S_i^*(s(1-\rho))}{\mathbb{E}[C](s(1-\rho)-\lambda_i(1-B_i^*(s(1-\rho))))}* \\ &\qquad C_i^*\left(\sum_{j=1}^{i}\lambda_j(1-B_j^*(s(1-\rho)))\right) - C_i^*\left(\sum_{j=1}^{i-1}\lambda_j(1-B_j^*(s(1-\rho)))+s(1-\rho)\right) \\ &= \frac{1}{(1+\hat{\rho}_i)s\,\mathbb{E}[S]}\left\{\left(\frac{\mu}{\mu+s\sum_{j=1}^{i-1}\hat{\rho}_j}\right)^{\alpha} - \left(\frac{\mu}{\mu+s(1+\sum_{j=1}^{i}\hat{\rho}_j)}\right)^{\alpha}\right\}, \end{aligned} \tag{4.3}$$

with α and μ given in (2.18). Differentiating (4.3) and taking $s=0$ gives the following expression for the scaled expected waiting time:

$$\begin{aligned}\mathbb{E}[\tilde{W}_{i,LCFS}] &= -\lim_{s\to 0}\tilde{W}^{*'}_{i,LCFS}(s) \\ &= \frac{\left(\left(1+\sum_{j=1}^{i}\hat{\rho}_j\right)^2 - \left(\sum_{j=1}^{i-1}\hat{\rho}_j\right)^2\right)\alpha(\alpha+1)}{(1+\hat{\rho}_i)2\,\mathbb{E}[S]\mu^2}. \end{aligned} \tag{4.4}$$

Note that this expression can be rewritten as follows

$$\begin{aligned}\mathbb{E}[\tilde{W}_{i,LCFS}] &= \frac{\left(1+2\sum_{j=1}^{i-1}\hat{\rho}_j + \left(\sum_{j=1}^{i}\hat{\rho}_j\right)^2 - \left(\sum_{j=1}^{i-1}\hat{\rho}_j\right)^2\right)\frac{2\mathbb{E}[S]}{\sigma^2}\left(1+\frac{2\mathbb{E}[S]}{\sigma^2}\right)}{(1+\hat{\rho}_i)2\,\mathbb{E}[S]\left(\frac{2}{\sigma^2}\right)^2} \\ &= \frac{\left(1+2\sum_{j=1}^{i}\hat{\rho}_j + 2\hat{\rho}_i\sum_{j=1}^{i-1}\hat{\rho}_j + \hat{\rho}_i^2\right)\left(1+\frac{2\mathbb{E}[S]}{\sigma^2}\right)\frac{\sigma^2}{2}}{2(1+\hat{\rho}_i)} \\ &= \frac{(1+\hat{\rho}_i)\left(1+2\sum_{j=1}^{i-1}\hat{\rho}_j+\hat{\rho}_i\right)\left(\frac{\sigma^2}{2}+\mathbb{E}[S]\right)}{2(1+\hat{\rho}_i)}, \end{aligned}$$

which is equal to (2.19), the expected scaled delay in a FCFS system.

4.2.2 General load

Now that we have the heavy traffic limit of the waiting time distribution, we can use this to find the waiting time distribution for general load and renewal arrivals. From the form of (4.3) we can see that we have again a uniform times a gamma distribution. Thus, we have the following approximation for the waiting time distribution in polling systems with renewal arrivals, $\rho<1$, globally gated service discipline and LCFS queueing discipline:

$$\mathbb{P}[W_i < w] \approx \mathbb{P}[UI_{i,a} < (1-\rho)w],$$

where U is a uniformly distributed random variable on $[\sum_{j=1}^{i-1}\hat{\rho}_j, 1+\sum_{j=1}^{i}\hat{\rho}_j]$ and $I_{i,a}$ is a gamma distributed random variable with parameters α and μ given in equation (4.2).

The kth moment is given by

$$\mathbb{E}[W^k_{i,LCFS}] = \frac{1}{(1-\rho)^k} \frac{\left(1+\sum_{j=1}^{i}\hat{\rho}_j\right)^{k+1} - \left(\sum_{j=1}^{i-1}\hat{\rho}_j\right)^{k+1}}{\left(1+\sum_{j=1}^{i}\hat{\rho}_j - \sum_{j=1}^{i-1}\hat{\rho}_j\right)(k+1)} \prod_{j=0}^{k-1}\frac{\alpha+j}{\mu_i}$$
$$= \frac{1}{(1-\rho)^k} \frac{\left(1+\sum_{j=1}^{i}\hat{\rho}_j\right)^{k+1} - \left(\sum_{j=1}^{i-1}\hat{\rho}_j\right)^{k+1}}{(1+\hat{\rho}_i)(k+1)} \prod_{j=0}^{k-1}\frac{\alpha+j}{\mu_i}.$$

4.2.3 Numerical results

To evaluate the approximation for the LCFS queueing discipline, we use the approach discussed in Section 3.1.3. We apply the approximation to the testbed given in Table 3.1 and compare this to the simulation results using percentual errors and the maximum absolute difference between the approximated and simulated cumulative density function. The results are shown in Tables B.11 - B.20. To visualize how the approximation works, the probability density function of the first queue of the test bed is plotted. The polling system has a load of 0.8, 7 queues and Poisson arrivals. The density function is approximated and plotted together with the simulated density function in Figure 4.4. The figure shows that the approximation closely follows the simulated distribution.

The conclusions that can be drawn from the tables coincide with the FCFS globally gated case in Section 4.1.3. The approximation gives very accurate results for $\rho \geq 0.7$. The mean waiting times are approximated very well for loads higher than 0.7 and also for a load around 0.1. For other loads, the approximation is also good, as can be seen in Table B.11. Tables B.12 and B.13 show that the approximations for second and third moment are also good if the load is 0.7 or higher. For lower loads the approximation of the third moments becomes unreliable.

From Table B.14 it can be concluded that the percentiles are approximated correctly most of the time, especially for higher loads. The maximum absolute difference between the approximated and simulated cumulative density functions are always lower than 0.15 and most of the time they are lower than 0.1, which can be found in Table B.15. If we look at Tables B.16 - B.18, we see from the first table that the approximation works better if the number of queues is higher. From the second table we see that the approximation is the most accurate when arrivals are Poisson and the least accurate if the interarrival times have a higher SCV. The third table shows that the 80th percentile is approximated best, this was also noted in the other results sections and also in [11].

4.3 Random order of service

4.3.1 Heavy traffic limits

In this section we discuss the random order of service queueing policy with glabally gated service policy. Using equations (2.11), (2.16) and (2.15), we can find the LST of the scaled waiting time in HT. The integration is done using integration by substitution and the limits are found using l'Hôpital's rule.

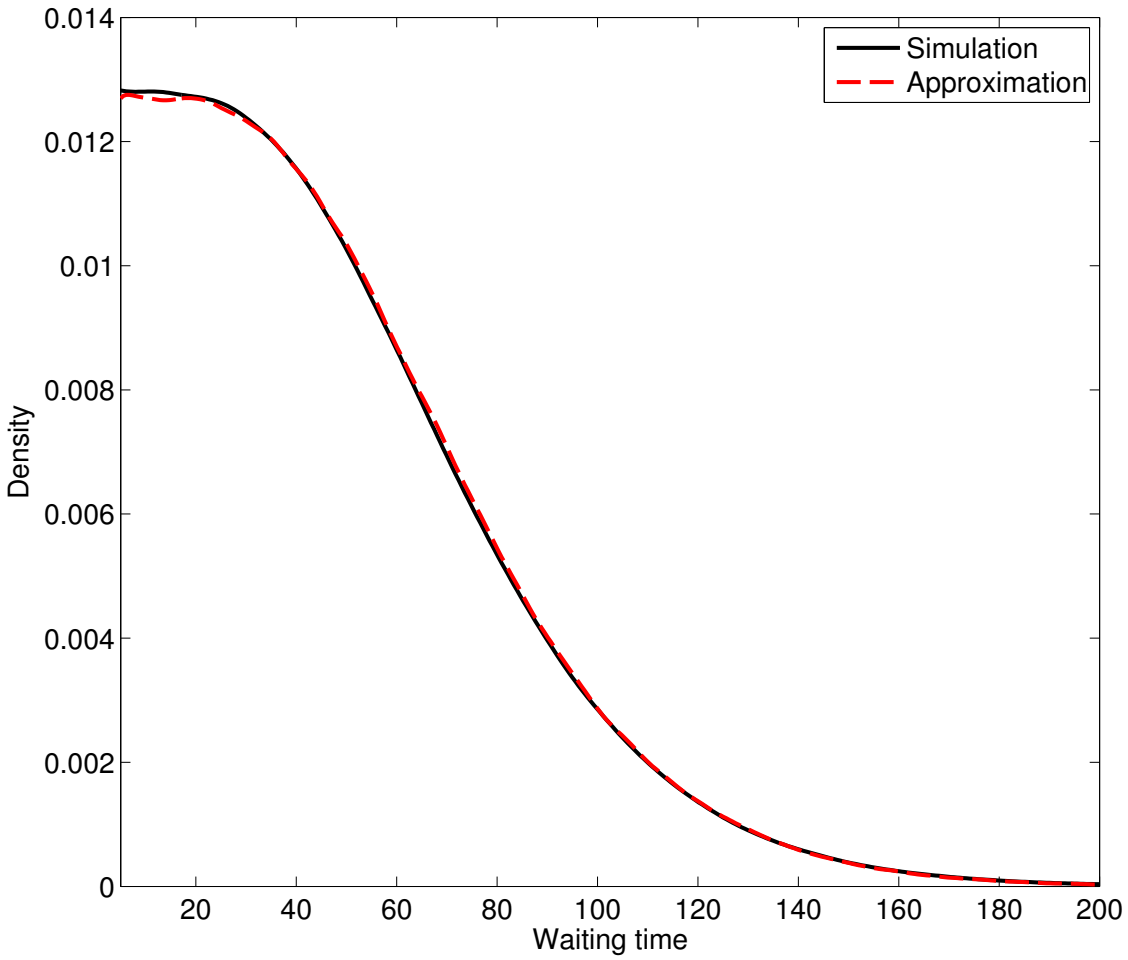


Figure 4.4: Approximated and simulated density function of the waiting time of an arbitrary queue in the example in Section 4.2.3

Recall that $X_i(s) = \sum_{j=1}^{i-1} \lambda_j(1-B_j^*(s))$, thus we have

$$\begin{aligned}
\tilde{W}_{i,ROS}^*(s) &= \lim_{\rho\uparrow 1} W_{i,ROS}^*(s(1-\rho)) \\
&= \lim_{\rho\uparrow 1} \frac{1}{\mathbb{E}[C]\lambda_i(1-B_i^*(s(1-\rho))} \int_{X_i(s(1-\rho))}^{X_{i+1}(s(1-\rho))} \frac{C_i^*(y)+C_i^*(y+s(1-\rho))}{s(1-\rho)}\,\mathrm{d}y \prod_{j=1}^{i-1} S_j^*(s(1-\rho)) \\
&= \lim_{\rho\uparrow 1} \frac{1}{\hat{\rho}_i s^2\,\mathbb{E}[S](1-\rho)} \int_{X_i(s(1-\rho))}^{X_{i+1}(s(1-\rho))} \left[\left(\frac{\mu}{\mu+y/(1-\rho)}\right)^{\alpha} - \left(\frac{\mu}{\mu+y/(1-\rho)+s}\right)^{\alpha}\right] \mathrm{d}y \\
&= \frac{\mu}{(\alpha-1)\hat{\rho}_i s^2\,\mathbb{E}[S]} \left\{ \left(\left(\frac{\mu}{\mu+s\sum_{j=1}^{i-1}\hat{\rho}_j}\right)^{\alpha-1} - \left(\frac{\mu}{\mu+s\sum_{j=1}^{i}\hat{\rho}_j}\right)^{\alpha-1} \right) \right. \\
&\quad \left. - \left(\left(\frac{\mu}{\mu+s(1+\sum_{j=1}^{i-1}\hat{\rho}_j)}\right)^{\alpha-1} - \left(\frac{\mu}{\mu+s(1+\sum_{j=1}^{i}\hat{\rho}_j)}\right)^{\alpha-1} \right) \right\}, \qquad (4.5)
\end{aligned}$$

with α and μ as in (2.18). The final equality is obtained by taking the limit after working out the integral using integration with substitution.

This expression has the same form as Equation (3.7) in Section 3.2.1 with the gated system. Note that the LST we found for gated came from a conditional LST. It is possible to heuristically derive a conditional LST for the globally gated case. Again the customers are marked with an order mark x, a uniform [0,1] distributed random variable. Every customer has to wait a fraction $\sum_{j=1}^{i-1}\hat{\rho}_j + x\hat{\rho}_i$ of the residual cycle, an unlucky customer has to wait one additional residual cycle. This leads to the following conditional LST:

$$\tilde{W}_{i,ROS}^*(s|x) = \frac{1}{\mathbb{E}[S]s} \left\{ \left(\frac{\mu}{\mu+s(\sum_{j=1}^{i-1}\hat{\rho}_j + x\hat{\rho}_i)}\right)^{\alpha} - \left(\frac{\mu}{\mu+s(\sum_{j=1}^{i-1}\hat{\rho}_j + x\hat{\rho}_i+1)}\right)^{\alpha} \right\}. \qquad (4.6)$$

Integrating this expression with respect to the density of the uniformly distributed order mark gives exactly (4.5).

4.3.2 General load

From Equation (4.6) we can see that the conditional waiting time distribution is a uniform times a gamma distribution, with the uniform distribution on $[\sum_{j=1}^{i-1}\hat{\rho}_j + x\hat{\rho}_i, \sum_{j=1}^{i-1}\hat{\rho}_j + x\hat{\rho}_i + 1]$, x is a random variable with a uniform distribution on $[0,1]$. The unconditional waiting time distribution can be found by noting that Equation (4.5) is the LST of a trapezoidal times a gamma distribution as was mentioned in Section 3.2.2, or it is possible to use Lemma 1. The parameters of the trapezoidal distribution are $a = \sum_{j=1}^{i-1}\hat{\rho}_j$, $b = \sum_{j=1}^{i}\hat{\rho}_j$, $c = \sum_{j=1}^{i-1}\hat{\rho}_j + 1$ and $d = \sum_{j=1}^{i}\hat{\rho}_j + 1$. This leads to the following probability density function of the unconditional "uniform" distribution $\tilde{U}_i$

$$f_{\tilde{U}}(y) = \begin{cases} \frac{y-\sum_{j=1}^{i-1}\hat{\rho}_j}{\hat{\rho}_i} & y \in [\sum_{j=1}^{i-1}\hat{\rho}_j, \sum_{j=1}^{i}\hat{\rho}_j) \\ 1 & y \in [\sum_{j=1}^{i}\hat{\rho}_j, \sum_{j=1}^{i-1}\hat{\rho}_j + 1] \\ \frac{\sum_{j=1}^{i}\hat{\rho}_j + 1 - y}{\hat{\rho}_i} & y \in (\sum_{j=1}^{i-1}\hat{\rho}_j + 1, \sum_{j=1}^{i}\hat{\rho}_j + 1]. \end{cases} \qquad (4.7)$$

The following approximation for the waiting time distribution in polling systems with renewal arrivals, $\rho < 1$, globally gated service discipline and ROS queueing discipline can be used:

$$\mathbb{P}[W_i < w] \approx \mathbb{P}[\tilde{U}_i I_{i,a} < (1-\rho)w],$$

where $\tilde{U}_i$ is a trapezoidal distributed random variable with pdf as in Equation (4.7) and $I_{i,a}$ is a gamma distributed random variable with parameters α and μ_i given in Equation (4.2).

The kth moment is calculated using:

$$\mathbb{E}[W_{i,ROS}^k] = \frac{(\sum_{j=1}^{i} \hat{\rho}_j + 1)^{k+2} - (\sum_{j=1}^{i} \hat{\rho}_j)^{k+2} - (\sum_{j=1}^{i-1} \hat{\rho}_j + 1)^{k+2} + (\sum_{j=1}^{i-1} \hat{\rho}_j)^{k+2}}{(1-\rho)^k \hat{\rho}_i (k+1)(k+2)} \prod_{j=0}^{k-1} \frac{\alpha + j}{\mu_i}$$

4.3.3 Numerical results

Figure 4.5 depicts the simulated and approximated probability density function of the first queue of the test bed. The system has a load of 0.8, Poisson arrivals and 7 queues. The approximated and the simulated density functions lie on top of each other.

The results of the accuracy test for the ROS approximation can be found in Tables B.21 - B.30. They are very similar to the results of the LCFS approximation, given in Section 4.2.3. The following conclusions can be drawn for the FCFS, LCFS and ROS approximation.

Parameters for which the approximation works well

- $\rho \geq 0.7$
- N large
- $c_{A_i}^2 = 1$

Measures that are harder to approximate

- Second moment
- Third moment (and higher moments)
- Lower percentiles

In Figure 4.6 the three approximations for globally gated polling systems are plotted, together with the simulation. The probability density functions in the figure belong to the first queue of a polling system with a load of 0.95, Poisson arrivals and 7 queues. The figure shows that for this system the approximation is very accurate, for all three service orders. The effect of the service order on the distribution is clear from the image, the distribution of the LFCS system has the fattest tail and thus the highest variance. The variance of the FCFS system's waiting time distribution has the lowest variance. Not surprisingly, this was also the case for the locally gated polling systems.

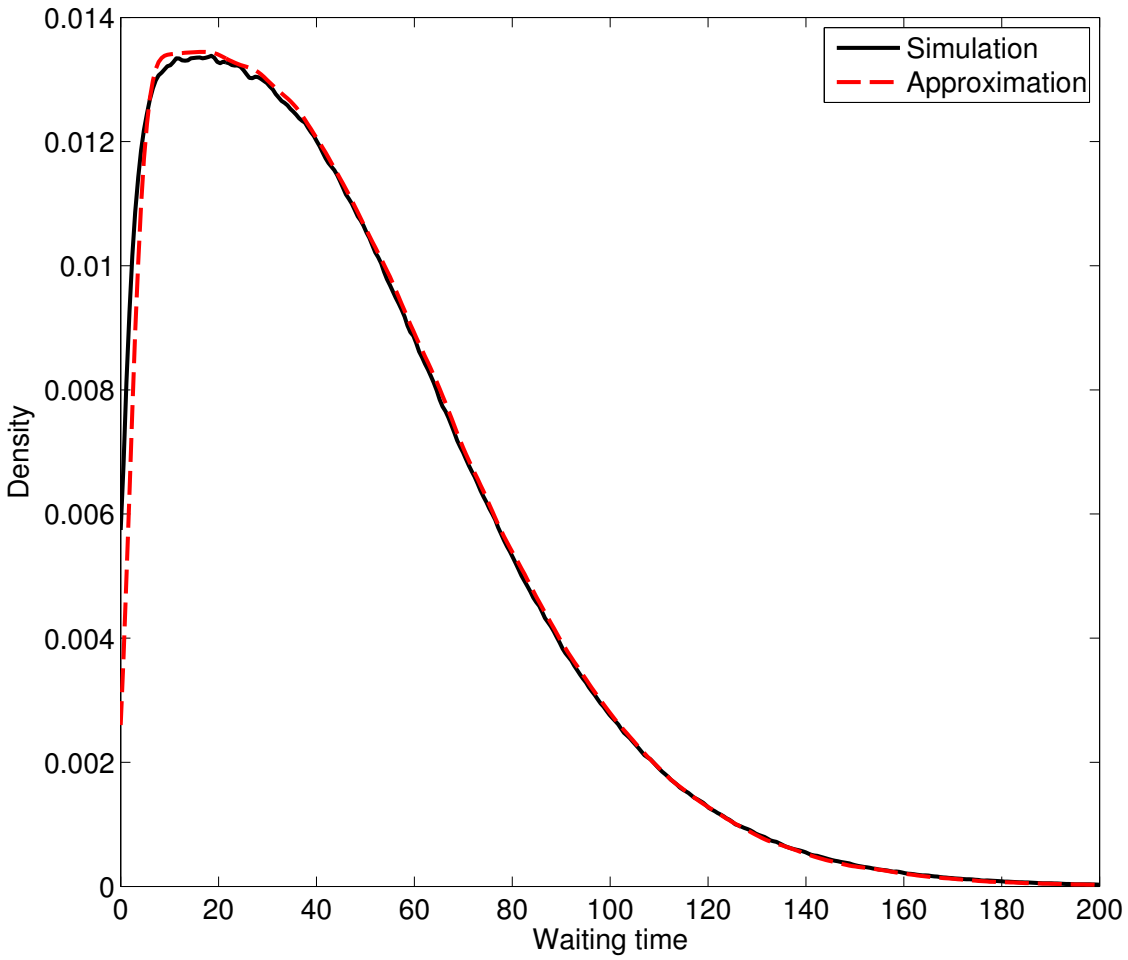


Figure 4.5: Approximated and simulated density function of the waiting time of an arbitrary queue in the example in Section 4.3.3

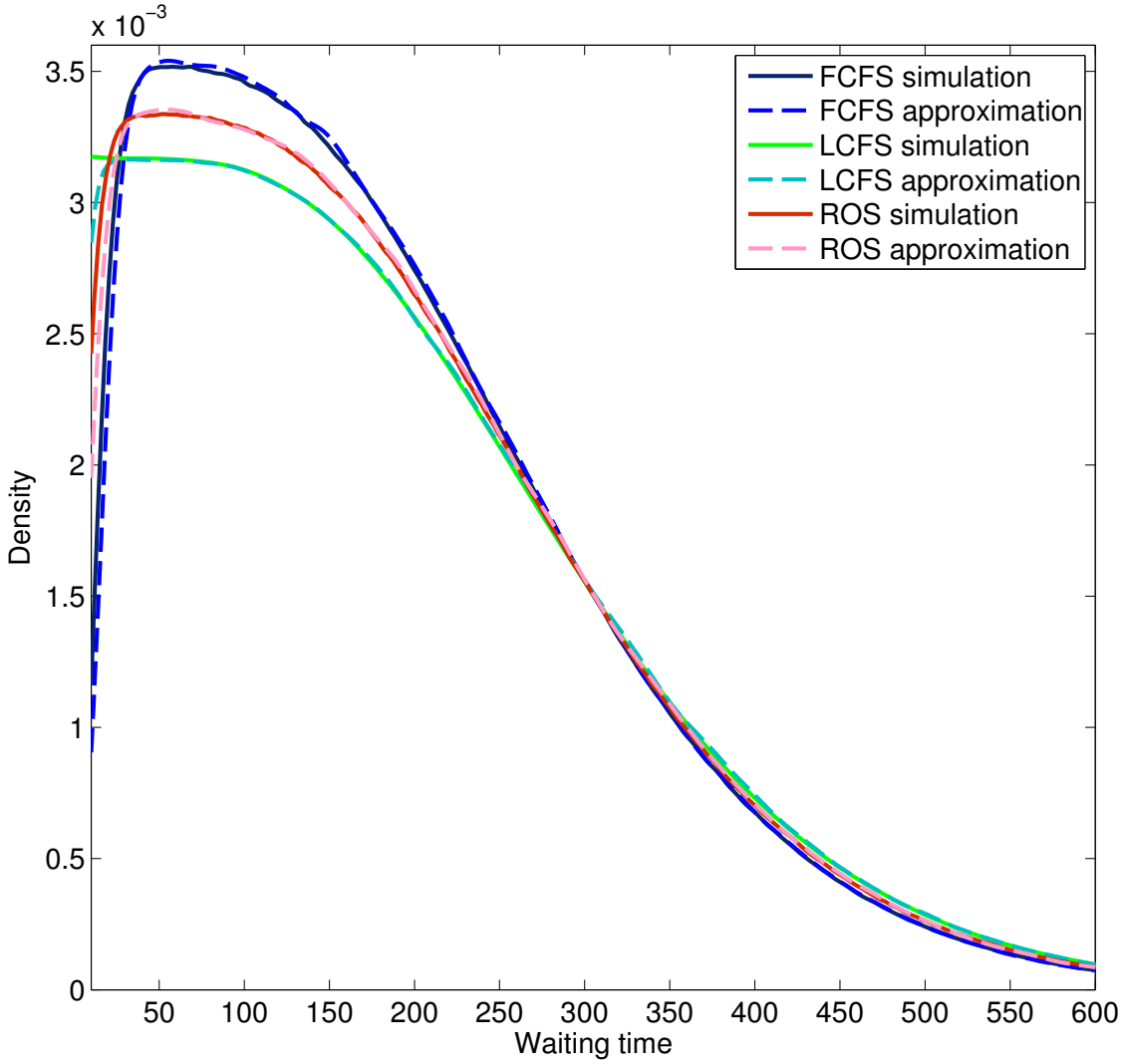


Figure 4.6: Approximated and simulated density functions of the first queue for the different service orders in a globally gated polling system

Chapter 5

Conclusion

For both gated and globally gated service disciplines approximations are derived for the distribution of the waiting time in polling systems with different service orders. The form of the approximations is simple and gives fundamental insight in the impact of the local service order. The approximations can also be used to accurately analyze various practical situations that can be modeled using a polling system.

The worst cases for the approximation are a low load or a large SCV of the interarrival times. In practice these characteristics are uncommon. For example, in production systems settings like $c_{A_i}^2 = 4$ are hardly found due to the just-in-time philosophy, which dictates that the demand is stable. Also, these systems are typically utilized beyond $\rho = 0.5$ to increase productivity.

For globally gated SJF and PS, the approximation should also be possible to find using the method in this report. Another topic of further research is finding approximations for the waiting time distributions in systems with exhaustive service. Finally the approximations can be used for optimization of polling systems.

Bibliography

[1] E. Altman and D. Fiems. Expected waiting time in symmetric polling systems with correlated walking times. *Queueing Systems*, 56(3):241–253, 2007.

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

[3] 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.

[4] S. Borst. *Polling Systems.* Ph.d. thesis, CWI, Amsterdam, The Netherlands, 1994.

[5] O.J. Boxma and W.P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability*, pages 949–964, 1987.

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

[7] M. Cicin-Sain, C.E.M. Pearce, and J. Sunde. On the application of a polling model with non-zero walk times and priority processing to a medical emergency-room environment. In *Information Technology Interfaces, 2001. ITI 2001. Proceedings of the 23rd International Conference on*, pages 49–56. IEEE, 2001.

[8] R.B. Cooper. Queues served in cyclic order: waiting times. *Bell Syst. Tech. J*, 49(3):399–413, 1970.

[9] RB Cooper and G. Murray. Queues served in cyclic order. *Bell Syst. Tech. J*, 48(3):675–689, 1969.

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

[11] 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.

[12] I. Frigui and A.S. Alfa. Analysis of a time-limited polling system. *Computer communications*, 21(6):558–571, 1998.

[13] S.W. Fuhrmann. Performance analysis of a class of cyclic schedules. *Bell laboratories technical memorandom*, pages 81–59531–1, 1981.

[14] A.G. Konheim, H. Levy, and M.M. Srinivasan. Descendant set: an efficient approach for the analysis of polling systems. *IEEE Transactions on Communications*, 42(234):1245–1253, 1994.

[15] 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.

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

[17] Z. Saffer and M. Telek. Unified analysis of BMAP/G/1 cyclic polling models. *Queueing Systems*, 64(1):69–102, 2010.

[18] R.D. Van der Mei. Delay in polling systems with large switch-over times. *Journal of applied probability*, 36(1):232–243, 1999.

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

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

[21] M. Van Vuuren and E.M.M. Winands. Iterative approximation of k-limited polling systems. *Queueing Systems*, 55(3):161–178, 2007.

[22] E.M.M. Winands. *Polling, Production and Priorities*. Ph.d. thesis, Eindhoven University of Technology, The Netherlands, 2007.

[23] 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.

Appendix A

Tabels results gated

A.1 ROS

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	93.33	6.67	0.00	0.00	0.00	0.00
0.3	66.67	23.33	10.00	0.00	0.00	0.00
0.5	60.00	26.67	13.33	0.00	0.00	0.00
0.7	63.33	30.00	6.67	0.00	0.00	0.00
0.8	76.67	20.00	3.33	0.00	0.00	0.00
0.9	93.33	6.67	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.1: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	70.00	10.00	16.67	3.33	0.00	0.00
0.3	46.67	36.67	16.67	0.00	0.00	0.00
0.5	50.00	36.67	6.67	3.33	3.33	0.00
0.7	66.67	16.67	6.67	6.67	3.33	0.00
0.8	70.00	13.33	13.33	3.33	0.00	0.00
0.9	76.67	20.00	3.33	0.00	0.00	0.00
0.95	93.33	6.67	0.00	0.00	0.00	0.00

Table A.2: Second moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	23.33	26.67	10.00	3.33	16.67	20.00
0.3	40.00	16.67	13.33	13.33	3.33	13.33
0.5	26.67	40.00	16.67	3.33	10.00	3.33
0.7	36.67	40.00	6.67	10.00	3.33	3.33
0.8	53.33	26.67	6.67	10.00	3.33	0.00
0.9	66.67	26.67	6.67	0.00	0.00	0.00
0.95	86.67	13.33	0.00	0.00	0.00	0.00

Table A.3: Third moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	63.81	19.52	11.43	4.29	0.48	0.48
0.3	55.71	26.19	12.86	2.86	0.95	1.43
0.5	57.14	24.76	13.81	2.86	0.95	0.48
0.7	67.62	20.48	10.95	0.95	0.00	0.00
0.8	74.76	19.52	5.71	0.00	0.00	0.00
0.9	88.10	11.90	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.4: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	53.33	46.67	0.00	0.00	0.00	0.00
0.3	50.00	43.33	6.67	0.00	0.00	0.00
0.5	63.33	33.33	3.33	0.00	0.00	0.00
0.7	76.67	23.33	0.00	0.00	0.00	0.00
0.8	86.67	13.33	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.5: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	56.19	25.08	11.90	3.33	2.06	1.43
7	76.12	15.78	5.92	1.29	0.34	0.54

Table A.6: Errors by number of queues categorized in bins of 5%

	Bins					
$c_{A_i}^2$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	91.52	6.10	1.49	0.45	0.30	0.15
0.25	71.43	20.39	5.21	1.19	1.04	0.74
4	43.75	30.65	18.30	4.32	1.34	1.64

Table A.7: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	51.90	21.43	15.71	8.10	0.95	1.90
50	59.52	21.43	16.19	1.43	0.95	0.48
60	66.67	21.90	10.00	0.95	0.48	0.00
70	72.86	20.00	5.71	1.43	0.00	0.00
80	86.19	11.90	1.90	0.00	0.00	0.00
90	86.67	12.38	0.95	0.00	0.00	0.00
95	80.95	12.86	5.71	0.48	0.00	0.00

Table A.8: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	64.29	28.10	5.24	1.90	0.00	0.48
2	57.62	23.81	10.48	5.24	2.38	0.48
3	46.67	23.33	20.00	2.86	3.81	3.33

Table A.9: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	86.19	8.10	2.38	1.90	0.48	0.95
2	75.24	18.57	4.29	0.95	0.00	0.95
3	68.10	20.95	8.57	1.90	0.00	0.48
4	76.19	18.57	3.81	0.48	0.48	0.48
5	84.76	10.48	4.29	0.00	0.48	0.00
6	62.38	19.52	14.29	3.33	0.48	0.00
7	80.00	14.29	3.81	0.48	0.48	0.95

Table A.10: Errors by queue in the systems with 7 queues categorized in bins of 5%

A.2 SJF

A.2.1 Uniform service times

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	93.33	6.67	0.00	0.00	0.00	0.00
0.3	73.33	20.00	6.67	0.00	0.00	0.00
0.5	66.67	26.67	6.67	0.00	0.00	0.00
0.7	70.00	23.33	6.67	0.00	0.00	0.00
0.8	76.67	20.00	3.33	0.00	0.00	0.00
0.9	90.00	10.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.11: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	20.00	50.00	26.67	3.33	0.00	0.00
0.3	63.33	26.67	6.67	3.33	0.00	0.00
0.5	63.33	30.00	6.67	0.00	0.00	0.00
0.7	73.33	16.67	6.67	3.33	0.00	0.00
0.8	70.00	16.67	13.33	0.00	0.00	0.00
0.9	80.00	20.00	0.00	0.00	0.00	0.00
0.95	96.67	3.33	0.00	0.00	0.00	0.00

Table A.12: Second moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	0.00	6.67	13.33	10.00	26.67	43.33
0.3	3.33	26.67	20.00	23.33	3.33	23.33
0.5	43.33	30.00	6.67	6.67	6.67	6.67
0.7	56.67	30.00	10.00	3.33	0.00	0.00
0.8	73.33	10.00	16.67	0.00	0.00	0.00
0.9	80.00	13.33	6.67	0.00	0.00	0.00
0.95	93.33	6.67	0.00	0.00	0.00	0.00

Table A.13: Third moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	43.33	29.05	22.86	2.86	0.95	0.95
0.3	49.52	20.95	18.57	8.10	1.43	1.43
0.5	56.19	21.43	14.29	5.24	1.90	0.95
0.7	65.24	22.38	8.57	3.81	0.00	0.00
0.8	72.86	20.95	5.71	0.48	0.00	0.00
0.9	87.62	11.90	0.48	0.00	0.00	0.00
0.95	98.57	1.43	0.00	0.00	0.00	0.00

Table A.14: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	43.33	50.00	6.67	0.00	0.00	0.00
0.3	36.67	56.67	6.67	0.00	0.00	0.00
0.5	50.00	46.67	3.33	0.00	0.00	0.00
0.7	76.67	23.33	0.00	0.00	0.00	0.00
0.8	86.67	13.33	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.15: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	58.73	21.59	10.79	5.08	1.75	2.06
7	70.75	17.07	8.64	1.84	0.61	1.09

Table A.16: Errors by number of queues categorized in bins of 5%

	Bins					
$c^2_{A_i}$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	91.37	6.40	1.49	0.60	0.15	0.00
0.25	62.50	20.24	10.42	3.27	1.93	1.64
4	42.41	31.25	17.86	4.91	0.89	2.68

Table A.17: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	47.14	16.67	18.57	13.33	1.90	2.38
50	52.86	16.19	23.81	4.76	1.43	0.95
60	58.10	28.10	11.43	1.43	0.95	0.00
70	69.05	21.43	4.76	3.81	0.95	0.00
80	89.52	9.52	0.95	0.00	0.00	0.00
90	81.43	18.57	0.00	0.00	0.00	0.00
95	72.86	16.19	10.48	0.48	0.00	0.00

Table A.18: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	63.81	23.33	7.14	4.29	0.00	1.43
2	60.95	20.00	10.95	3.81	2.38	1.90
3	51.43	21.43	14.29	7.14	2.86	2.86

Table A.19: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	77.62	13.33	7.14	0.00	0.48	1.43
2	70.00	19.05	8.10	1.43	0.48	0.95
3	64.76	21.90	9.52	2.38	0.48	0.95
4	72.38	16.19	8.57	0.95	0.95	0.95
5	75.71	13.81	6.19	2.38	0.48	1.43
6	59.52	21.90	11.90	5.24	0.95	0.48
7	75.24	13.33	9.05	0.48	0.48	1.43

Table A.20: Errors by queue in the systems with 7 queues categorized in bins of 5%

A.2.2 Exponential servcie times

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	100.00	0.00	0.00	0.00	0.00	0.00
0.3	80.00	20.00	0.00	0.00	0.00	0.00
0.5	76.67	23.33	0.00	0.00	0.00	0.00
0.7	80.00	20.00	0.00	0.00	0.00	0.00
0.8	86.67	13.33	0.00	0.00	0.00	0.00
0.9	96.67	3.33	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.21: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	66.67	6.67	16.67	10.00	0.00	0.00
0.3	63.33	20.00	16.67	0.00	0.00	0.00
0.5	70.00	20.00	10.00	0.00	0.00	0.00
0.7	73.33	16.67	10.00	0.00	0.00	0.00
0.8	76.67	16.67	6.67	0.00	0.00	0.00
0.9	80.00	20.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.22: Second moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	30.00	20.00	16.67	0.00	3.33	30.00
0.3	56.67	13.33	3.33	3.33	0.00	23.33
0.5	56.67	23.33	3.33	6.67	6.67	3.33
0.7	63.33	20.00	13.33	3.33	0.00	0.00
0.8	73.33	10.00	16.67	0.00	0.00	0.00
0.9	76.67	20.00	3.33	0.00	0.00	0.00
0.95	93.33	6.67	0.00	0.00	0.00	0.00

Table A.23: Third moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	67.14	17.14	10.48	5.24	0.00	0.00
0.3	60.95	22.86	11.43	4.29	0.48	0.00
0.5	65.24	22.38	10.95	1.43	0.00	0.00
0.7	77.62	17.62	4.76	0.00	0.00	0.00
0.8	82.38	15.71	1.90	0.00	0.00	0.00
0.9	94.76	5.24	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.24: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	63.33	36.67	0.00	0.00	0.00	0.00
0.3	60.00	40.00	0.00	0.00	0.00	0.00
0.5	70.00	30.00	0.00	0.00	0.00	0.00
0.7	90.00	10.00	0.00	0.00	0.00	0.00
0.8	96.67	3.33	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.25: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	79.37	20.63	0.00	0.00	0.00	0.00
7	84.35	15.65	0.00	0.00	0.00	0.00

Table A.26: Errors by number of queues categorized in bins of 5%

	Bins					
$c_{A_i}^2$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	100.00	0.00	0.00	0.00	0.00	0.00
0.25	94.29	5.71	0.00	0.00	0.00	0.00
4	54.29	45.71	0.00	0.00	0.00	0.00

Table A.27: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	59.05	19.52	10.95	10.00	0.48	0.00
50	67.62	15.71	16.19	0.48	0.00	0.00
60	71.90	22.38	5.71	0.00	0.00	0.00
70	80.00	14.76	3.81	1.43	0.00	0.00
80	92.86	7.14	0.00	0.00	0.00	0.00
90	88.57	11.43	0.00	0.00	0.00	0.00
95	86.67	7.62	5.24	0.48	0.00	0.00

Table A.28: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	74.29	18.57	4.76	1.43	0.00	0.95
2	73.33	17.62	7.14	1.43	0.00	0.48
3	65.24	20.95	10.95	1.90	0.95	0.00

Table A.29: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	86.67	7.62	2.38	1.90	0.48	0.95
2	80.00	13.33	4.76	0.95	0.00	0.95
3	74.76	17.14	5.71	1.43	0.00	0.95
4	82.38	11.43	3.81	1.43	0.00	0.95
5	86.19	8.57	2.86	0.95	0.48	0.95
6	69.52	18.57	9.52	1.90	0.00	0.48
7	84.29	9.05	4.29	0.95	0.00	1.43

Table A.30: Errors by queue in the systems with 7 queues categorized in bins of 5%

A.3 PS

A.3.1 Uniform service times

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	96.67	3.33	0.00	0.00	0.00	0.00
0.3	66.67	23.33	10.00	0.00	0.00	0.00
0.5	56.67	26.67	13.33	3.33	0.00	0.00
0.7	56.67	30.00	13.33	0.00	0.00	0.00
0.8	70.00	23.33	6.67	0.00	0.00	0.00
0.9	83.33	16.67	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.31: Mean sojourn time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	40.00	20.00	10.00	13.33	10.00	6.67
0.3	36.67	23.33	20.00	13.33	0.00	6.67
0.5	50.00	20.00	16.67	10.00	0.00	3.33
0.7	56.67	20.00	16.67	6.67	0.00	0.00
0.8	53.33	26.67	20.00	0.00	0.00	0.00
0.9	73.33	26.67	0.00	0.00	0.00	0.00
0.95	93.33	6.67	0.00	0.00	0.00	0.00

Table A.32: Second moment of the sojourn time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	16.67	16.67	16.67	3.33	10.00	36.67
0.3	10.00	23.33	6.67	13.33	13.33	33.33
0.5	20.00	20.00	10.00	16.67	10.00	23.33
0.7	46.67	23.33	23.33	0.00	3.33	3.33
0.8	46.67	36.67	10.00	6.67	0.00	0.00
0.9	73.33	20.00	6.67	0.00	0.00	0.00
0.95	90.00	10.00	0.00	0.00	0.00	0.00

Table A.33: Third moment of the sojourn time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	43.81	35.24	8.57	7.14	3.33	1.90
0.3	43.33	20.48	20.48	12.86	1.90	0.95
0.5	47.62	22.38	20.00	7.14	2.38	0.48
0.7	58.57	26.19	10.48	3.33	1.43	0.00
0.8	68.10	22.38	7.14	2.38	0.00	0.00
0.9	86.19	12.38	1.43	0.00	0.00	0.00
0.95	96.19	3.81	0.00	0.00	0.00	0.00

Table A.34: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	46.67	30.00	13.33	10.00	0.00	0.00
0.3	33.33	40.00	20.00	6.67	0.00	0.00
0.5	33.33	43.33	20.00	3.33	0.00	0.00
0.7	63.33	30.00	6.67	0.00	0.00	0.00
0.8	73.33	26.67	0.00	0.00	0.00	0.00
0.9	90.00	10.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.35: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	50.79	30.16	9.52	9.52	0.00	0.00
7	68.03	23.81	8.16	0.00	0.00	0.00

Table A.36: Errors by number of queues categorized in bins of 5%

	Bins					
$c^2_{A_i}$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	90.00	10.00	0.00	0.00	0.00	0.00
0.25	67.14	30.00	2.86	0.00	0.00	0.00
4	31.43	37.14	22.86	8.57	0.00	0.00

Table A.37: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	44.29	22.38	14.76	11.90	4.29	2.38
50	49.52	20.48	18.57	10.00	1.43	0.00
60	53.33	29.52	14.29	2.86	0.00	0.00
70	63.81	22.38	7.14	2.38	1.90	2.38
80	77.62	18.57	2.86	0.95	0.00	0.00
90	76.19	15.71	5.71	1.90	0.48	0.00
95	75.24	12.38	4.29	4.29	2.86	0.95

Table A.38: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	48.57	31.90	9.05	4.29	2.38	3.81
2	51.90	25.24	10.95	6.67	3.81	1.43
3	42.38	22.38	16.67	10.00	2.38	6.19

Table A.39: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	77.62	11.90	6.19	2.86	0.00	1.43
2	67.62	19.52	7.62	2.86	0.95	1.43
3	60.95	20.95	10.48	4.29	2.38	0.95
4	66.67	20.48	8.57	2.38	0.48	1.43
5	76.19	13.81	6.19	2.38	0.95	0.48
6	55.24	20.95	14.29	6.19	2.38	0.95
7	73.33	15.24	6.67	3.33	0.00	1.43

Table A.40: Errors by queue in the systems with 7 queues categorized in bins of 5%

A.3.2 Exponential service times

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	100.00	0.00	0.00	0.00	0.00	0.00
0.3	76.67	23.33	0.00	0.00	0.00	0.00
0.5	66.67	33.33	0.00	0.00	0.00	0.00
0.7	70.00	23.33	6.67	0.00	0.00	0.00
0.8	80.00	20.00	0.00	0.00	0.00	0.00
0.9	93.33	6.67	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.41: Mean sojourn time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	13.33	33.33	20.00	3.33	23.33	6.67
0.3	13.33	33.33	20.00	26.67	3.33	3.33
0.5	40.00	30.00	23.33	3.33	3.33	0.00
0.7	50.00	36.67	10.00	3.33	0.00	0.00
0.8	63.33	26.67	10.00	0.00	0.00	0.00
0.9	80.00	20.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.42: Second moment of the sojourn time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	10.00	6.67	6.67	10.00	16.67	50.00
0.3	10.00	6.67	3.33	20.00	3.33	56.67
0.5	6.67	23.33	16.67	20.00	16.67	16.67
0.7	30.00	36.67	26.67	3.33	0.00	3.33
0.8	36.67	50.00	10.00	3.33	0.00	0.00
0.9	70.00	30.00	0.00	0.00	0.00	0.00
0.95	96.67	3.33	0.00	0.00	0.00	0.00

Table A.43: Third moment of the sojourn time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	51.43	24.76	10.00	5.71	7.14	0.95
0.3	40.48	35.71	13.33	7.62	2.86	0.00
0.5	49.05	36.19	9.52	3.33	1.90	0.00
0.7	76.67	14.29	6.19	2.86	0.00	0.00
0.8	80.48	13.81	4.76	0.95	0.00	0.00
0.9	90.00	9.52	0.48	0.00	0.00	0.00
0.95	98.57	1.43	0.00	0.00	0.00	0.00

Table A.44: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	40.00	36.67	23.33	0.00	0.00	0.00
0.3	56.67	20.00	23.33	0.00	0.00	0.00
0.5	63.33	23.33	13.33	0.00	0.00	0.00
0.7	76.67	23.33	0.00	0.00	0.00	0.00
0.8	80.00	20.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table A.45: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	65.08	22.22	12.70	0.00	0.00	0.00
7	77.55	15.65	6.80	0.00	0.00	0.00

Table A.46: Errors by number of queues categorized in bins of 5%

	Bins					
$c^2_{A_i}$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	84.29	15.71	0.00	0.00	0.00	0.00
0.25	98.57	1.43	0.00	0.00	0.00	0.00
4	38.57	35.71	25.71	0.00	0.00	0.00

Table A.47: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	57.14	20.00	6.67	7.62	8.10	0.48
50	61.90	20.00	9.05	9.05	0.00	0.00
60	69.05	22.38	8.57	0.00	0.00	0.00
70	75.71	16.67	1.90	1.90	3.33	0.48
80	82.38	16.67	0.95	0.00	0.00	0.00
90	68.57	21.43	9.05	0.95	0.00	0.00
95	66.67	18.10	7.62	3.33	3.81	0.48

Table A.48: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	52.38	27.14	8.10	5.24	1.90	5.24
2	59.05	24.76	8.10	2.38	3.33	2.38
3	49.52	24.76	15.24	4.29	2.38	3.81

Table A.49: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	80.95	10.00	4.29	1.90	1.43	1.43
2	68.10	20.00	4.29	3.81	2.86	0.95
3	60.95	24.29	5.24	4.76	3.33	1.43
4	67.14	20.48	5.24	3.81	1.90	1.43
5	85.71	8.10	3.81	1.43	0.00	0.95
6	58.57	25.24	7.14	4.29	3.33	1.43
7	76.67	14.29	4.76	1.90	0.95	1.43

Table A.50: Errors by queue in the systems with 7 queues categorized in bins of 5%

Appendix B

Tabels results globally gated

B.1 FCFS

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	93.33	6.67	0.00	0.00	0.00	0.00
0.3	86.67	10.00	3.33	0.00	0.00	0.00
0.5	86.67	13.33	0.00	0.00	0.00	0.00
0.7	96.67	3.33	0.00	0.00	0.00	0.00
0.8	100.00	0.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.1: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	43.33	26.67	20.00	10.00	0.00	0.00
0.3	50.00	30.00	13.33	3.33	3.33	0.00
0.5	66.67	26.67	6.67	0.00	0.00	0.00
0.7	96.67	3.33	0.00	0.00	0.00	0.00
0.8	93.33	6.67	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.2: Second moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	10.00	20.00	13.33	13.33	6.67	36.67
0.3	23.33	16.67	10.00	13.33	16.67	20.00
0.5	26.67	36.67	13.33	13.33	3.33	6.67
0.7	73.33	20.00	6.67	0.00	0.00	0.00
0.8	83.33	16.67	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.3: Third moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	65.24	24.76	7.14	2.38	0.48	0.00
0.3	69.05	22.38	5.71	1.43	0.95	0.48
0.5	77.14	19.52	2.38	0.48	0.48	0.00
0.7	91.43	7.62	0.95	0.00	0.00	0.00
0.8	95.71	4.29	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.4: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	56.67	36.67	6.67	0.00	0.00	0.00
0.3	56.67	36.67	3.33	3.33	0.00	0.00
0.5	66.67	26.67	6.67	0.00	0.00	0.00
0.7	83.33	16.67	0.00	0.00	0.00	0.00
0.8	96.67	3.33	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.5: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	73.17	16.51	4.76	3.02	1.11	1.43
7	87.41	8.98	2.04	0.41	0.41	0.75

Table B.6: Errors by number of queues categorized in bins of 5%

	Bins					
$c_{A_i}^2$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	91.07	6.70	1.04	0.89	0.15	0.15
0.25	80.06	13.69	3.57	1.49	0.74	0.45
4	74.85	15.92	4.46	1.34	1.04	2.38

Table B.7: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	70.00	23.81	2.38	1.90	1.43	0.48
50	84.76	11.43	2.38	0.95	0.48	0.00
60	93.81	4.76	1.43	0.00	0.00	0.00
70	94.76	5.24	0.00	0.00	0.00	0.00
80	95.71	4.29	0.00	0.00	0.00	0.00
90	82.38	14.76	2.86	0.00	0.00	0.00
95	76.67	15.24	6.67	1.43	0.00	0.00

Table B.8: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	76.19	13.81	4.76	2.86	0.48	1.90
2	70.00	19.05	4.76	2.38	2.38	1.43
3	73.33	16.67	4.76	3.81	0.48	0.95

Table B.9: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	85.24	10.95	1.90	0.48	0.95	0.48
2	87.14	8.57	2.86	0.00	0.48	0.95
3	83.81	13.81	1.43	0.00	0.48	0.48
4	89.52	6.67	2.86	0.00	0.00	0.95
5	86.67	10.00	1.90	0.48	0.48	0.48
6	90.00	7.14	1.90	0.00	0.48	0.48
7	89.52	5.71	1.43	1.90	0.00	1.43

Table B.10: Errors by queue in the systems with 7 queues categorized in bins of 5%

B.2 LCFS

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	93.33	6.67	0.00	0.00	0.00	0.00
0.3	86.67	10.00	3.33	0.00	0.00	0.00
0.5	86.67	13.33	0.00	0.00	0.00	0.00
0.7	100.00	0.00	0.00	0.00	0.00	0.00
0.8	100.00	0.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.11: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	46.67	20.00	6.67	13.33	10.00	3.33
0.3	46.67	6.67	26.67	13.33	3.33	3.33
0.5	56.67	26.67	13.33	3.33	0.00	0.00
0.7	90.00	10.00	0.00	0.00	0.00	0.00
0.8	100.00	0.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.12: Second moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	26.67	26.67	10.00	0.00	3.33	33.33
0.3	30.00	16.67	13.33	6.67	0.00	33.33
0.5	43.33	16.67	10.00	10.00	10.00	10.00
0.7	73.33	26.67	0.00	0.00	0.00	0.00
0.8	90.00	10.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.13: Third moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	66.67	20.48	7.62	4.76	0.48	0.00
0.3	66.19	20.00	10.95	2.86	0.00	0.00
0.5	72.38	20.95	5.71	0.95	0.00	0.00
0.7	88.57	10.95	0.48	0.00	0.00	0.00
0.8	92.38	7.62	0.00	0.00	0.00	0.00
0.9	99.52	0.48	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.14: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	63.33	33.33	3.33	0.00	0.00	0.00
0.3	70.00	30.00	0.00	0.00	0.00	0.00
0.5	80.00	20.00	0.00	0.00	0.00	0.00
0.7	93.33	6.67	0.00	0.00	0.00	0.00
0.8	100.00	0.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.15: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	74.29	15.56	5.87	1.90	0.95	1.43
7	85.92	8.71	2.72	1.36	0.20	1.09

Table B.16: Errors by number of queues categorized in bins of 5%

	Bins					
$c_{A_i}^2$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	96.13	3.27	0.60	0.00	0.00	0.00
0.25	79.02	15.03	4.32	1.49	0.15	0.00
4	68.75	16.37	6.70	3.27	1.19	3.72

Table B.17: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	74.29	16.67	7.62	1.43	0.00	0.00
50	79.05	17.62	2.38	0.95	0.00	0.00
60	86.67	11.43	1.43	0.48	0.00	0.00
70	88.10	7.62	3.33	0.95	0.00	0.00
80	88.10	10.48	1.43	0.00	0.00	0.00
90	83.33	10.00	6.19	0.48	0.00	0.00
95	82.38	8.10	4.29	4.76	0.48	0.00

Table B.18: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	76.67	15.24	4.29	0.95	0.95	1.90
2	69.52	17.62	6.19	3.81	1.43	1.43
3	76.67	13.81	7.14	0.95	0.48	0.95

Table B.19: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	88.10	8.10	2.38	0.00	0.48	0.95
2	73.81	17.14	4.29	2.86	0.48	1.43
3	78.10	14.76	4.29	1.90	0.00	0.95
4	92.86	3.33	1.90	0.95	0.00	0.95
5	89.52	5.71	2.38	0.95	0.48	0.95
6	88.57	7.14	2.38	0.95	0.00	0.95
7	90.48	4.76	1.43	1.90	0.00	1.43

Table B.20: Errors by queue in the systems with 7 queues categorized in bins of 5%

B.3 ROS

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	93.33	6.67	0.00	0.00	0.00	0.00
0.3	86.67	10.00	3.33	0.00	0.00	0.00
0.5	86.67	13.33	0.00	0.00	0.00	0.00
0.7	96.67	3.33	0.00	0.00	0.00	0.00
0.8	100.00	0.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.21: Mean waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	56.67	6.67	20.00	16.67	0.00	0.00
0.3	43.33	23.33	20.00	10.00	0.00	3.33
0.5	60.00	33.33	3.33	3.33	0.00	0.00
0.7	100.00	0.00	0.00	0.00	0.00	0.00
0.8	96.67	3.33	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.22: Second moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
0.1	26.67	26.67	6.67	3.33	0.00	36.67
0.3	23.33	33.33	10.00	0.00	0.00	33.33
0.5	50.00	13.33	13.33	10.00	3.33	10.00
0.7	76.67	23.33	0.00	0.00	0.00	0.00
0.8	96.67	3.33	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.23: Third moment of the waiting time errors categorized in bins of 5%

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	72.86	16.19	9.52	1.43	0.00	0.00
0.3	71.90	16.67	8.57	2.38	0.48	0.00
0.5	78.10	18.10	2.86	0.95	0.00	0.00
0.7	90.00	10.00	0.00	0.00	0.00	0.00
0.8	92.86	7.14	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.24: Percentile errors categorized in bins of 0.05

	Bins					
ρ	0-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	0.25+
0.1	66.67	33.33	0.00	0.00	0.00	0.00
0.3	70.00	30.00	0.00	0.00	0.00	0.00
0.5	83.33	16.67	0.00	0.00	0.00	0.00
0.7	90.00	10.00	0.00	0.00	0.00	0.00
0.8	100.00	0.00	0.00	0.00	0.00	0.00
0.9	100.00	0.00	0.00	0.00	0.00	0.00
0.95	100.00	0.00	0.00	0.00	0.00	0.00

Table B.25: Maximum absolute differences in cdf categorized in bins of 0.05

	Bins					
N	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
3	76.98	14.13	5.56	1.75	0.16	1.43
7	88.16	7.69	2.18	0.82	0.07	1.09

Table B.26: Errors by number of queues categorized in bins of 5%

	Bins					
$c_{A_i}^2$	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	96.73	2.53	0.74	0.00	0.00	0.00
0.25	83.18	12.50	3.13	0.89	0.15	0.15
4	71.43	16.07	6.25	2.53	0.15	3.57

Table B.27: Errors by SCV of the interarrival times categorized in bins of 5%

	Bins					
Percentile	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
40	74.29	18.57	5.24	1.43	0.48	0.00
50	79.52	17.62	1.90	0.95	0.00	0.00
60	90.00	8.10	1.90	0.00	0.00	0.00
70	91.90	6.67	1.43	0.00	0.00	0.00
80	94.76	5.24	0.00	0.00	0.00	0.00
90	87.62	8.10	4.29	0.00	0.00	0.00
95	85.24	5.71	6.67	2.38	0.00	0.00

Table B.28: Errors in percentiles categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	77.62	13.33	5.71	0.95	0.00	2.38
2	76.19	14.76	5.24	2.38	0.48	0.95
3	77.14	14.29	5.71	1.90	0.00	0.95

Table B.29: Errors by queue in the system with 3 queues categorized in bins of 5%

	Bins					
Queue	0-5%	5-10%	10-15%	15-20%	20-25%	25%+
1	86.19	10.00	1.90	0.95	0.00	0.95
2	84.29	10.00	2.38	1.90	0.00	1.43
3	83.33	12.38	3.33	0.00	0.00	0.95
4	92.86	3.81	1.90	0.48	0.00	0.95
5	91.43	4.76	2.38	0.00	0.48	0.95
6	88.57	8.10	1.90	0.48	0.00	0.95
7	90.48	4.76	1.43	1.90	0.00	1.43

Table B.30: Errors by queue in the systems with 7 queues categorized in bins of 5%