Service Systems with Balking Based on Queueing Time

Liqiang Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2007

Approved by,
Advisor: Professor Vidyadhar G. Kulkarni
Reader: Professor Chuanshu Ji
Reader: Professor Jasleen Kaur
Reader: Professor Haipeng Shen
Reader: Professor Serhan Ziya

ABSTRACT

Liqiang Liu: Service Systems with Balking Based on Queueing Time

(Under the direction of Dr. Vidyadhar G. Kulkarni)

We consider service systems with balking based on queueing time, also called queues with wait-based balking. An arriving customer joins the queue and stays until served if and only if the queueing time is no more than some pre-specified threshold at the time of arrival. We assume that the arrival process is a Poisson process.

We begin with the study of the $M/G/1$ system with a deterministic balking threshold. We use level-crossing argument to derive an integral equation for the steady state virtual queueing time (vqt) distribution. We describe a procedure to solve the equation for general distributions and we solve the equation explicitly for several special cases of service time distributions, such as phase type, Erlang, exponential and deterministic service times. We give formulas for several performance criteria of general interest, including average queueing time and balking rate. We illustrate the results with numerical examples.

We then consider the first passage time problem in an $M/PH/1$ setting. We use a fluid model where the buffer content changes at a rate determined by an external stochastic process with finite state space. We derive systems of first-order linear differential equations for both the mean and LST (Laplace-Stieltjes Transform) of the busy period in the fluid model and solve them explicitly. We obtain the mean and LST of the busy period in the $M/PH/1$ queue with wait-based balking as a special limiting case of the fluid model. We illustrate the results with numerical examples.

Finally we extend the method used in the single server case to multi-server case. We consider the vqt process in an $M/G/s$ queue with wait-based balking. We construct a single server system, analyze its operating characteristics, and use it to approximate the multi-server system. The approximation is exact for the $M/M/s$ and $M/G/1$ system. We give both analytical results and numerical examples. We conduct simulation to assess the accuracy of the approximation.

ACKNOWLEDGMENTS

I am sincerely grateful to my advisor, Dr. Vidyadhar G. Kulkarni for his priceless guidance and persistent encouragement. Dr. Kulkarni has earned my respect and admiration as a wise and witty scholar. His enlightening direction and unselfish help over the past three years makes my pursuit of the doctorate a truly rewarding journey.

I would like to express my gratitude to the committee members, Dr. Chuanshu Ji, Dr. Jasleen Kaur, Dr. Haipeng Shen and Dr. Serhan Ziya, for their involvement and inspiring comments. In particular, I am thankful to Dr. Shen for introducing me to the service engineering area which motivates this work, and Dr. Ziya who offered his generous help in my searching of a research topic.

I would also like to thank Dr. David Perry and Dr. Wolfgang Stadje for bringing the first passage time problem to our attention. They had suggested using martingale methods to solve it, which motivated us to seek a numerically easier method presented in this thesis.

I appreciate the flexibility offered by my manager at SAS Institute Inc., Dr. Gehan A. Corea, in accommodating my school schedule.

Last but not least, I would like to thank my parents, Mr. Guichang Liu and Mrs. Huihui Chen, and my wife Liping Cai for their support all the way along.

TABLE OF CONTENTS

List of Figures . . . . . **viii**

1 Introduction **1**

1.1 Overview 1

1.2 Single Sever Queues 4

1.2.1 Steady State Distribution 4

1.2.2 Busy Period 6

1.3 Multi-Server Queues 7

2 $M/G/1$ Queues with Wait-based Balking: Steady State Distributions **10**

2.1 Introduction 10

2.2 An $M/G/1$ Queue with Balking 11

2.3 Equilibrium Distribution of Workload Process 13

2.4 Rational $G^*(s)$ and $M/PH/1$ Queue with Balking 17

2.4.1 $M/PH/1$: Transform Method 18

2.4.2 $M/PH/1$: Differential Equation Approach 21

2.4.3 Special Cases 24

2.5 An Example of Non-rational $G^*(s)$: $M/D/1$ 28

2.6 Numerical Examples 29

2.7 Concluding Remarks 31

3 $M/PH/1$ Queues with Wait-based Balking: Busy Period Analysis **37**

3.1 Introduction 37

3.2 The Fluid Model 38

3.3 First Passage Time: the Fluid Model 40

3.4 A Special Case of the Fluid Model 44

3.5 First Passage Time: the Balking Queueing Model 50

3.6 Special Case: Exponential Service Times 53

3.7 Numerical Results 55

3.8 Concluding Remarks 57

4 Balking and Reneging in $M/G/s$ Systems: Exact Analysis and Approximations **66**

4.1 Introduction 66

4.2 The $M/M/s$ Balking Model 67

4.3 The $M/G/s$ Balking Model 72

4.3.1 Approximation I: $\dot{J} = \ddot{J} = \bar{S}$. 79

4.3.2 Approximation II: $\dot{J} = \bar{S}$, $\ddot{J} = \hat{S}$. 83

4.4 Connection Between Balking and Reneging 83

4.5 Design of Simulation Experiments 86

4.6 Numerical Results 87

4.7 Concluding Remarks 89

Bibliography **93**

List of Figures

2.1 A typical sample path of $W(t)$ 12

2.2 A sample path of $W(t)$ when $\lambda \to \infty$ 16

2.3 $f(x)$ for different service time distributions, $\rho = 0.8, b = 5$ 32

2.4 $f(x)$ for different service time distributions, $\rho = 1, b = 5$ 32

2.5 $f(x)$ for different service time distributions, $\rho = 1.2, b = 5$ 33

2.6 $f(x)$ for different ρ, exponential service time, $b = 5$ 33

2.7 $\bar{W}$ vs. ρ for different service time distributions, $b = 5$ 34

2.8 $\bar{W}$ vs. b for different service time distributions, $\rho = 0.8$ 34

2.9 $\bar{W}$ vs. b for different service time distributions, $\rho = 1$ 35

2.10 $\bar{W}$ vs. b for different service time distributions, $\rho = 1.2$ 35

2.11 c vs. λ for different service time distributions 36

3.1 A Typical Sample Path of $X_r(t)$ 45

3.2 Construction of $W(t)$ 51

3.3 Construction of $X_r(t)$ 51

3.4 Mean Plot, $b = 2, \rho = 0.8$ 58

3.5 Mean Plot, $b = 2, \rho = 1$ 58

3.6 Mean Plot, $b = 2, \rho = 1.2$ 59

3.7 Mean Plot, Exponential Service Time, $b = 2$ 59

3.8 Distribution Plot, $b = 2, \rho = 0.8$ (two sets of three curves for x=1 and x=2) 60

3.9 Distribution Plot, $b = 2, \rho = 1$ (two sets of three curves for x=1 and x=2) 60

3.10 Distribution Plot, $b = 2, \rho = 1.2$ (two sets of three curves for x=1 and x=2) 61

3.11 Distribution Plot, Erlang Service Time, $b = 2, x = 1$ 61

3.12 Distribution Plot, Erlang Service Time, $b = 2, \rho = 0.8$ 62

3.13 Density Plot, $b = 2, \rho = 0.8$, (two sets of three curves for x=1 and x=2) 62

3.14 Density Plot, $b = 2, \rho = 1$ (two sets of three curves for x=1 and x=2) 63

3.15 Density Plot, $b = 2, \rho = 1.2$ (two sets of three curves for x=1 and x=2) 63

3.16 Density Plot, Erlang Service Time, $b = 2, x = 1$ 64

3.17 Density Plot, Erlang Service Time, $b = 2, \rho = 0.8$ 64

3.18 Mean Plot, Exponential Service Time, $\rho = 0.8$ 65

4.1 A sample path of $W(t)$ and $N(t)$ 68

4.2 Long-run Average Queueing Time for All Served Customers 90

4.3 Fraction of Rejected Customers 90

4.4 Relative Errors of Approximations: $s = 3$ 91

4.5 Relative Errors of Approximations: $s = 10$ 91

4.6 Relative Errors of Approximations: $s = 100$ 92

Chapter 1

Introduction

1.1 Overview

A significant aspect in modeling call centers is customer impatience (cf. Koole and Mandelbaum [27], Garnett et al. [15], Whitt [48]). Motivated by analyzing the call center operations, we consider an $M/G/s$ queueing system with impatient customers. The customers arrive according to a Poisson process with rate λ, and request iid (independent and identically distributed) service times with a general distribution. There are $s \geq 1$ servers in the system available to serve the customers. All servers are identical and unit-rate, i.e., each server is capable of processing one unit of service requirement per unit time.

Two common modes in which customers display their impatience are balking and reneging. A call-in customer who cannot be helped immediately by a human server might be told how long a wait he/she faces before an operator is available. Then the customer might hang up (i.e. balk) or decide to hold. This is an example of the balking behavior: a customer refuses to enter the queue if the wait is too long. We call this *wait-based balking* to differentiate it from balking in a more conventional sense, i.e., a customer refuses to enter the queue if the waiting line is too long. On the other hand, a customer who is waiting for an operator might hang up (i.e. renege) before

getting served if the wait in line becomes too long. This is the reneging behavior. Of course, there can be a combination of the two.

Queueing models with balking incorporate the characteristics of the customers' impatience or a specific admission control policy in force at a service system. In a typical queueing model with balking the service requirement of an arriving customer may not be (completely) accepted if the system is "too congested" at the time of its arrival. From the perspective of an arriving customer one natural measurement of system congestion is the queueing time he/she faces to get service started. A no-join decision based on this congestion measurement is the aforementioned wait-based balking. The model considered in this thesis uses a wait-based balking rule. Before stating the balking rule, we define the *virtual queueing time* (vqt) in the system. The vqt at time t in the system, denoted by $W(t)$, is the queueing time (i.e., time spent in the system before commencing service) that would be experienced if a customer joins the system at time t. The process $\{W(t), t \geq 0\}$ is referred as vqt process. We call a queueing system with balking based on the vqt a *wait-based balking queue*. It works as follows. We assume that each customer knows his/her exact queueing time at the time of arrival. A customer arriving at time t joins the system if and only if $W(t-)$ is no more than a pre-specified threshold (possibly random). The balking customers (i.e., customers who do not join) are lost forever. The entering customers wait in an infinite capacity FCFS (first-come, first-served) queue until a server is available and leave when the service completes.

Our justification of such a model proceeds as follows. Although the queueing time information in call centers is not precise, this model incorporates the right characteristics of the customer behavior. On the other hand, there are some systems, for instance, some communication systems, where the information about queueing time is always precisely available. What's more, from the view of control problems, the balking rule can be regarded as a threshold type of customer acceptance/rejection

policy based on the system workload. Such a policy is shown to be optimal under certain conditions by Johansen and Stidham [25]. This threshold type policy generates the model considered here. The usefulness of the wait-based balking model for studying reneging systems will become clear in Chapter 4, where we reveal a close relation between two models.

More complicated models involving balking and reneging behaviors can be found in the research with a focus on the comparison of performances under different operational characteristics presented in systems with customer impatience (cf. [1], [2], [18], [47]). For example, delay information used to make a decision can have different degrees of precision. Guo and Zipkin [18] explore the effects of different level of the information on the system performance by using a utility-based approach. The models resulting from "no information" (all customers use long-term average queueing time, a constant, as an estimate of their delays) and "full information" (customers know their queueing times upon arrival) cases can be analyzed by using the method in this thesis. Explicit results can be obtained in a more general settings than $M/M/1$. If precise delay is not available at the time the customer arrives, then it can be provided as various estimates. Armony, Shimkin and Whitt [2] study the impact of delay announcement based on two estimates, most recent observed delay and steady state delay. The latter belongs to the class of models considered in this thesis. They also show that such models with delay announcements can be well approximated by the corresponding $M/M/s$ models. Armony and Maglaras [1] study a multi-server queueing system with a call-back option, which is modeled as a two-channel system. They propose a scheme to estimate the delay which is asymptotically correct. They find that offering call-back option and delay estimate to induce rational routing and balking behaviors improves system performance substantially. These works typically focus on overloaded regimes and use asymptotic analysis for systems with large number of servers. In principle, if the effective arrival rate depends on delays and does

not depend on the number in the system, then the exact analysis of such models can be done using the methods developed here.

The vqt process is also known as work-in-system, or virtual-delay process, which is introduced by Beneš [6] and Takács [44]. See Heyman and Sobel [19] (pages 383–390) for details. Many system performance measures of the queue can be derived from the steady state distribution of the vqt process. The central topic of this thesis is the steady state analysis of this process in several particular wait-based balking queues.

1.2 Single Sever Queues

Single server wait-based balking queues are studied under a variety of names in the literature: "finite workload capacity", "dams", "workload-dependent arrival rates", or "queues with limited accessibility". See Prabhu [42], Perry et al. [41], Perry and Stadje [39], Bekker et al. [5] and Bekker [4]. Wait-based balking queues also have found applications in the study of clearing models, see Boxma et al. [9].

Notice for single server systems with a FCFS service discipline, the vqt coincides with the workload (work content) of the system, i.e., sum of the service times of all customs in queue and the remaining service time of the customer in service (see page 10 for an alternative definition). Therefore, we use two terms interchangeably in the single server context.

1.2.1 Steady State Distribution

In Chapter 2 we deal with the steady state distribution of the workload process in the $M/G/1$ queue with wait-based balking.

The single server case has been extensively studied in the literature. For example, Hu and Zazanis [23], consider various types of restrictions on workload (i.e. balking rules). They obtain the steady state distribution of the workload process for the sys-

tem with workload restrictions in terms of that of the corresponding queue without restrictions. This requires the unrestricted system to be stable so as to solve the restricted system. But clearly the system with workload restriction is always stable. The inability to solve the restricted system when the corresponding unrestricted system is unstable is inevitable due to the "cut and paste" technique they use, since the basic idea of such a method is to obtain the steady state distribution of the limited access queueing models in terms of known steady state distribution of simpler models with no access constraints.

Level-crossing argument is an appropriate tool to analyze such queueing models. Cohen [11] introduces Level-Crossing Theory (LCT) for regenerative processes of the $GI/G/1$ type. Doshi [13] generalizes the theory to stationary dam process and presents applications of level-crossing analysis to many queueing systems, especially to single-server queues. Gavish and Schweitzer [16] use level-crossing analysis to study an $M/G/1$ system where arrivals are rejected if their queueing plus service times would exceed a fixed amount (note that the service time for each customer is known upon arrival, which is different from the wait-based balking queue we consider here). Hokstad [22] uses the method of level crossing method and computes explicitly the steady state distribution in the $M/D/1$ queue with wait-based balking. Perry and Asmussen [38] deal with the most general type of single server wait-based balking models with several variations. They develop the integral equation for the steady state distribution of the workload. The solution is given in terms of an infinite sum of iterated convolution integrals. The authors give explicit solution for the $M/M/1$ queue, and mention that explicit solutions can be obtained for $M/PH/1$ queues, but do not give the expressions. We have been unable to find the explicit expressions in the literature.

In Chapter 2, we use level-crossing argument to derive an integral equation for the steady state workload distribution. We describe a procedure to solve the equation for

general distributions and we solve the equation explicitly for several special cases of service time distributions, such as phase type, Erlang, exponential and deterministic service times. For the $M/PH/1$ case we show that the integral equation can be reduced to a differential equation with constant coefficients, and hence can be solved by standard methods. We illustrate the results with several numerical examples.In Chapter 2, we use level-crossing argument to derive an integral equation for the steady state workload distribution. We describe a procedure to solve the equation for general distributions and we solve the equation explicitly for several special cases of service time distributions, such as phase type, Erlang, exponential and deterministic service times. For the $M/PH/1$ case we show that the integral equation can be reduced to a differential equation with constant coefficients, and hence can be solved by standard methods. We illustrate the results with several numerical examples.

1.2.2 Busy Period

The *busy period* is the first passage time that the vqt process enters the state 0. There are relatively few papers studying the busy period of the wait-based balking systems. Perry and Asmussen [38] find the LST of the busy period for the $M/M/1$ system via differential equations and martingales. They further extend the results to the case where b is an exponential random variable. Perry et al. [40] give closed-form expression for the LST of the busy period for an $M/G/1$ queue with wait-based balking as a function of the LSTs of certain stopping times. The paper also analyzes the busy periods in $G/M/1$ queues by exploiting the duality between the $M/G/1$ and $G/M/1$ queues.

In Chapter 3 we use an alternative method to analyze the first passage time problem in an $M/PH/1$ setting. The method involves constructing a standard fluid model so that the $M/PH/1$ wait-based balking model is a limiting case of this fluid model. For a general discussion of fluid models we refer the readers to the survey

paper by Kulkarni [30]. Various authors have studied the first passage times in fluid models. See Asmussen and Bladt [3], Chen and Samalam [10], Kulkarni and Narayanan [31], Boxma and Dumas [8] and Kulkarni and Tzenova [32]. In particular, Kulkarni and Tzenova [32] developed a differential equation method which is more suitable for numerical algorithms. We extend the method and use it to solve the first passage time problem arising in wait-based balking queues. In Chapter 3 we begin by considering a fluid model where the buffer content changes at a rate determined by an external stochastic process with finite state space. We derive systems of first-order linear different equations for both the mean and LST (Laplace-Stieltjes Transform) of the busy period in this model and solve explicitly. We obtain the mean and LST of the busy period in the $M/PH/1$ queue with wait-based balking as a special limiting case of this fluid model. We illustrate the results with numerical examples.

1.3 Multi-Server Queues

In Chapter 4 we extend the method used in the single server case to the analysis of the multi-server case.

The reneging version of the model we consider has been studied by Gnedenko and Kovalenko [17] under the name "systems with limited waiting time". They consider exponential service times to obtain a multidimensional Markov process for the number of busy servers and workload in each server. They derive a system of integro-differential equations for the limiting joint distribution and give explicit solution. They give formulas for the loss probability and average queueing time. They also give the limiting distribution of the vqt process. However, as Boots and Tijms [7] noted, the results in [17] are quite technical and not generally applicable. Instead, they give an alternative formula for the loss probability as a function of the tail probability of the stationary vqt process in a corresponding queue with no impatience. They prove that their formula is exact in the $M/M/s$ case and can be used as a

heuristic for the $M/G/s$ case. A severe restriction is that the formula is valid only when the traffic intensity is less than 1, which is not required for the reneging queue to be stable. The method we use in this thesis overcomes the preceding drawbacks and can be easily extended to the general case. Although we are unable to give the joint distribution for the workload and busy servers, we don't lose much since many common performance measures can be derived directly from the limiting distribution of the vqt process.

Even in the absence of the balking behavior the $M/G/s$ queueing system is notorious for its complexity which forbids analytical solutions. Analytical results are available for only a few special cases, while a handful of approximations for the limiting analysis have been proposed in the past decades (cf. Chapter 13, Heyman and Sobel [19]). In this thesis we focus on *system approximations*, i.e. approximations that take the results from an exact analysis of a simpler system as approximations of the true operating characteristics of the original system. Although the approximate methods vary by motivations and the techniques used, it turns out that all results can be viewed as the so-called "systems interpolation", i.e., some mixture of the known analytical results for a few special cases, such as $M/M/s$, $M/E_k/s$, $M/D/s$, and $M/G/\infty$. See Kimura [26] for details. We cannot find any system approximations of the $M/G/s$ queueing system with impatient customers in the literature.

To develop a system approximation for the multi-server system with impatient customers, we borrow a simple idea used by Lee and Longton [33], Takács [44] (page 160), Newell [36] (page 86), Hokstad [21], Nozaki and Ross [37], Tijms et al. [45], and Miyazawa [34]. In brief, the idea is to treat the s-server system as an $M/G/\infty$ system (or $M/G/s-1$ loss system) when some servers are idle and an $M/G/1$ system when all servers are busy. Using such a system decomposition we construct a single server system whose operating characteristics approximate those of the $M/G/s$ queueing system with wait-based balking. The approximation is exact when $G = M$, the

balking threshold is zero or $s = 1$. The exact analysis of the approximate system follows the same line as Chapter 2, where we solve the $s = 1$ case. The approximation is evaluated by comparing performance measures against simulations. The connection between wait-based balking and reneging is discussed in Chapter 4, which reveals that the analysis of the vqt process in this thesis is indeed a treatment to a queue incorporating customer impatience, whether balking or reneging.

Chapter 2

$M/G/1$ Queues with Wait-based Balking: Steady State Distributions

2.1 Introduction

In this chapter, we analyze a single server first-come first-served (FCFS) wait-based balking system that operates as follows: an arriving customer joins the queue only if he/she sees that the workload in the system is no more than a fixed amount b. We assume the customer knows the exact amount of the workload in the system at the time of arrival. We also assume that once a customer decides to join the queue, he/she stays in the system until service completion.

The *workload* at time t is defined as the time it takes the server to empty the system, provided there are no arrivals after t. Since the service discipline is FCFS, from the perspective of an arriving customer, the workload at the time of arrival can be also interpreted as the queueing time (the time he/she spends in the queue until service starts) if he/she chooses to join the queue at all. The balking rule mentioned above is very natural in situations where the customers are unwilling to wait longer

than b unit of time for start of service.

The assumption of FCFS discipline is not necessary if we do not require vqt to be the same as workload. The analysis in this chapter proceeds identically without this assumption since in general, the workload process is invariant under work-conserving service disciplines. Here we assume FCFS in which case the work content also represents the queuing time, and hence the model represents customers balking in face of long waits.

The rest of this chapter is organized as follows. Section 2.2 is a description of the model. In Section 2.3, level-crossing argument is used to derived the integral equation for the steady state distribution of the workload process. Section 2.4 deals with the case when service time distribution has a rational Laplace transform (LT). In particular, if the service time has a phase type distribution, we show a method to reduce the integral equation to a differential equation that can be solved by standard methods and give explicit formulas for probability distribution and selected performance measures. Special cases, namely, exponential, phase type and hyper exponential service time distributions are included. Section 2.5 deals with the $M/D/1$ case. We give several numerical examples in Section 2.6. The chapter is concluded with a comment on other performance measures and balking rules (Section 2.7).

2.2 An $M/G/1$ Queue with Balking

We begin with an $M/G/1$ system with Poisson arrivals with rate λ, and iid service times. Let S be a generic service time random variable, and

$$\Pr\{S > x\} = G(x), \quad E(S) = \tau, \quad \mathrm{Var}(S) = \sigma^2. \tag{2.1}$$

Thus, the traffic intensity is

$$\rho = \lambda\tau. \tag{2.2}$$

Let $\{W(t), t \geq 0\}$ be the workload or vqt process. The balking rule is characterized by a constant $b < \infty$ as follows: A customer arriving at time t joins the system (and stays until service completion) if $W(t-) \leq b$, else he/she leaves and is lost. A typical sample path of $W(t)$ is shown in Figure 2.1. The arriving times are denoted by T_i. Note that the arrival at T_4 balks since $W(T_4-) > b$.

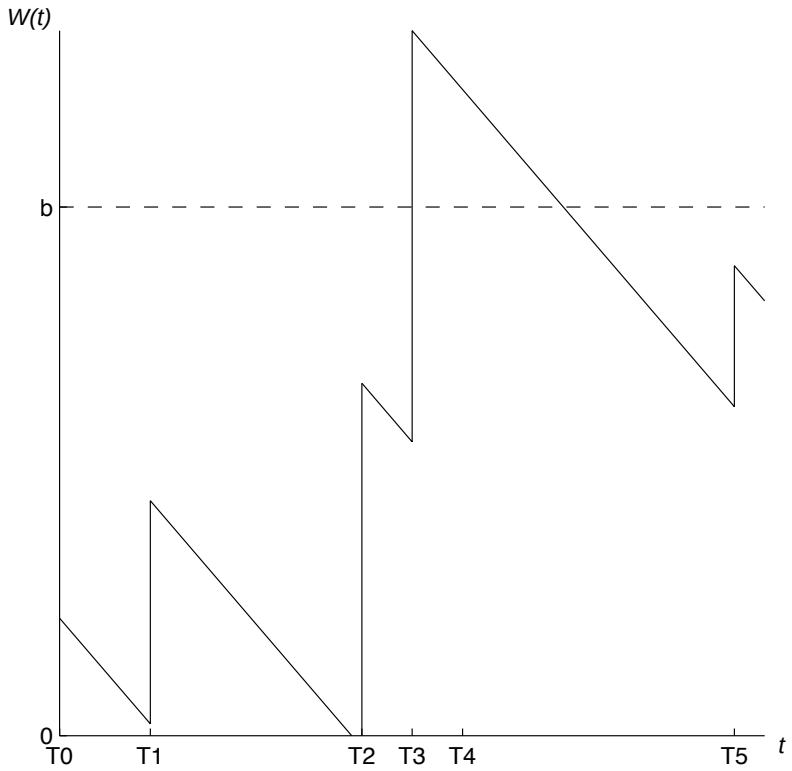

Figure 2.1: A typical sample path of $W(t)$

It is known that the $\{W(t), t \geq 0\}$ process has a limiting distribution if $E(S) < \infty$ (cf. [38]), which we will assume in this thesis. Let

$$F(x) = \lim_{t \to \infty} \Pr\{W(t) \leq x\},$$

$$\bar{W} = \lim_{t \to \infty} E(W(t)).$$

It is clear that the limiting distribution has a mass at 0, $c = F(0)$, and a density $f(x)$

for $x > 0$. We focus on computing c, $f(x)(x \geq 0)$ and $F(x)(x \geq 0)$.

2.3 Equilibrium Distribution of Workload Process

In this section, we derive the equation for $f(x)$ and c in Theorem 1. We describe a procedure to find the solution in Theorem 2. We also discuss several limiting cases of parameters b and ρ.

Theorem 1. *The equilibrium probability density function (pdf) of the workload process of the $M/G/1$ queue with balking satisfies:*

$$f(x) = \lambda \int_0^{x \wedge b} f(u)G(x-u)du + c\lambda G(x), \tag{2.3a}$$

$$\int_0^{\infty} f(x)dx + c = 1, \tag{2.3b}$$

where $x \wedge b = \min(x, b)$.

Proof. We prove this by level-crossing argument. Suppose the process $\{W(t), t \geq 0\}$ is stationary. Then, during interval $(t, t+h)$, the probability that the workload down-crosses level x is:

$$[F(x+h) - F(x)](1 - \lambda h). \tag{2.4}$$

Thus this is also the expected number of down-crossings during $(t, t+h)$.

Similarly, the probability that the workload up-crosses level x is:

$$\int_0^{x \wedge b} f(u)\lambda h \Pr\{S \geq x-u\}du + \Pr\{W(\infty) = 0\}\lambda h \Pr\{S \geq x\}, \tag{2.5}$$

which, using our notations yields

$$\int_0^{x \wedge b} f(u)\lambda h G(x-u)du + c\lambda h G(x). \tag{2.6}$$

Thus this is also the expected number of up-crossings during $(t, t+h)$.

The level-crossing argument implies that (2.4) must equal to (2.6) (cf. [13]). Now divide both side by h and let $h \to 0$. We get Equation (2.3a). Equation (2.3b) is the normalizing equation. □

Notice that the first term in the right hand side of Equation (2.3a) is just the convolution of $f(x)$ and $G(x)$ multiplied by λ, when $x \wedge b$ is replaced by x. Let $f_1(x)$ be the solution to

$$f_1(x) = \lambda \int_0^x f_1(u)G(x-u)du + G(x), \quad x \geq 0. \tag{2.7}$$

Let

$$f_2(x) = \lambda \int_0^b f_1(u)G(x-u)du + G(x), \quad x \geq b \tag{2.8}$$

The solution to Equation (2.3) is given in the following theorem.

Theorem 2. *The solution to* (2.3) *is:*

$$f(x) = \begin{cases} c\lambda f_1(x) & \text{if } 0 \leq x \leq b \\ c\lambda f_2(x) & \text{if } x > b \end{cases} \tag{2.9}$$

where

$$c = \left[\lambda \int_0^b f_1(x)dx + \lambda \int_b^\infty f_2(x)dx + 1\right]^{-1}. \tag{2.10}$$

Proof. The solution is easy to verify by substitution. □

From the above theorem it is clear that a possible procedure to obtain $f(x)$ is to find $f_1(x)$ first, then compute $f_2(x)$ by using Equation (2.8). By the normalizing equation (2.3b), after computing the integral, we are able to compute c. This completes the computation of $f(x)$. Obviously, one main step is to solve Equation (2.7) for $f_1(x)$. One method is to use LT.

Let $G^*(s)$ be the LT of $G(x)$. From (2.7) we get the LT of $f_1(x)$ (assuming its existence):

$$f_1^*(s) = \frac{G^*(s)}{1 - \lambda G^*(s)}. \tag{2.11}$$

To continue our procedure, we need the inverse LT of $f_1^*(s)$. A close form inversion is possible if $G^*(s)$ is rational. However, in this case, there is an alternative method to solve Equation (2.7). We demonstrate these in Section 2.4.

We can instantly obtain several interesting results from Theorem 2 under some limiting values of b and λ. The first case is $b \to 0$. Under this regime, the system reduces to a normal $M/G/1/1$ queue. From Theorem 2, as $b \to 0$,

$$c \to \frac{1}{1+\rho}, \tag{2.12}$$

$$f(x) \to \frac{\lambda}{1+\rho} G(x), \quad x > 0 \tag{2.13}$$

$$\bar{W} \to \frac{\lambda}{2(1+\rho)}(\sigma^2 + \tau^2). \tag{2.14}$$

It is easy to verify that the results above coincide with the results of an $M/G/1/1$ queueing system.

Next consider the limiting regime $\lambda \to \infty$. A sample path of workload is illustrated in Figure 2.2. Obviously,

$$\lim_{\lambda \to \infty} f(x+b) = \lim_{\lambda \to \infty} \lim_{b \to 0} f(x), \quad x \geq 0.$$

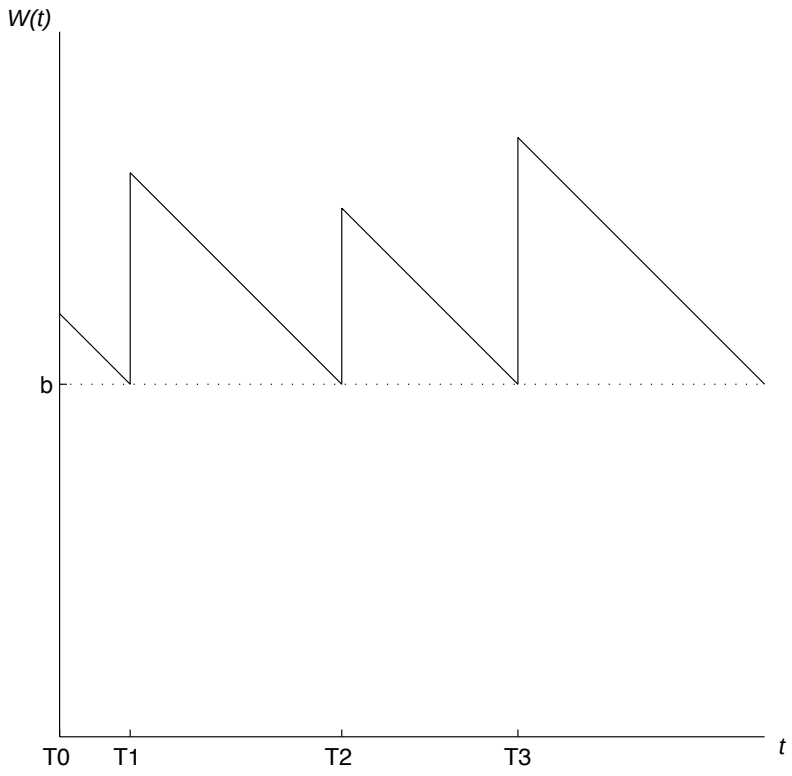

Figure 2.2: A sample path of $W(t)$ when $\lambda \to \infty$

Using the fact above we get

$$f(x) \to 0, \text{ when } 0 \leq x \leq b, \tag{2.15}$$

$$f(x) \to \frac{1}{\tau} G(x-b), \text{ when } x > b, \tag{2.16}$$

$$c \to 0, \tag{2.17}$$

$$\bar{W} \to b + \frac{1}{2\tau}(\sigma^2 + \tau^2). \tag{2.18}$$

When $b \to \infty$, in the limit the system reduces to a normal $M/G/1$ queue, which

is stable for $\rho < 1$. Notice that

$$\int_0^\infty f_1(x)dx = f_1^*(0),$$

$$G^*(0) = \tau,$$

$$\bar{W} = -\left.\frac{df^*(s)}{ds}\right|_{s=0}.$$

From Theorem 2 and Equation (2.11), we get the following limiting results which are consistent with the usual $M/G/1$ results.

$$f(x) \to c\lambda f_1(x), \tag{2.19}$$

$$c \to 1 - \rho, \tag{2.20}$$

$$\bar{W} \to \frac{\lambda(\sigma^2 + \tau^2)}{2(1-\rho)}. \tag{2.21}$$

In Section 2.4 and 2.5 we focus mainly on solving Equation (2.7) for several specific service time distributions by transform or directly.

2.4 Rational $G^*(s)$ and $M/PH/1$ Queue with Balking

Suppose $G^*(s)$ is rational, i.e.,

$$G^*(s) = \frac{N(s)}{D(s)}, \tag{2.22}$$

where $D(s)$ is a p degree polynomial in s, $N(s)$ is a polynomial in s whose degree is less than p. Then

$$f^*(s) = \frac{N(s)}{D(s) - \lambda N(s)}. \tag{2.23}$$

Let θ_i, $i = 1, 2, \cdots, p$, be the roots to

$$D(s) - \lambda N(s) = 0. \tag{2.24}$$

We assume they are distinct. Then from general method of computing inverse LT (cf. [28]), we obtain a closed form expression for $f_1(x)$ as follows:

$$f_1(x) = \sum_{i=1}^{p} A_i e^{\theta_i x}, \tag{2.25}$$

where

$$A_i = \lim_{s \to \theta_i} \frac{(s - \theta_i) G^*(s)}{1 - \lambda G^*(s)}, \ i = 1, 2, \cdots, p. \tag{2.26}$$

Next, as a specific example, we apply our method to solve the $M/PH/1$ case (which has a rational $G^*(s)$). In addition, we give results for Erlang, hyper exponential and exponential distributions as three more special cases of the phase type distribution.

2.4.1 $M/PH/1$: Transform Method

For a common phase type distribution with parameter (α, M) (cf. [29]), the complementary cumulative distribution function (ccdf) $G(x)$ is given by

$$G(x) = \alpha e^{Mx} \vec{1}, \tag{2.27}$$

where $\vec{1}$ is a column vector with all coordinates equal to 1, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ is a non-negative row vector and $\alpha\vec{1} = 1$, and M is an n by n sub-matrix of the generator of an irreducible CTMC (continuous time Markov chain), with the following properties:

- M is invertible;
- M is diagonally dominant with all diagonal elements negative.

It is well known that these properties imply that all eigenvalues of M have negative real part. We need this condition in the computation of $f_2(x)$. In this case the LT of $G(x)$ is

$$G^*(s) = \alpha(sI - M)^{-1}\vec{1}, \tag{2.28}$$

where I is an n by n identity matrix.

Following the procedure described in Section 2.3, we collect our results for the $M/PH/1$ with wait-based balking in Theorem 3.

Theorem 3. *Let the service time distribution be a common phase type distribution with ccdf given by* (2.27)*. The equilibrium pdf of the workload process of the $M/PH/1$ wait-based balking queue with threshold b is:*

$$f(x) = \begin{cases} c\lambda\sum_{i=1}^{n} A_i e^{\theta_i x} & \text{if } 0 \le x \le b \\ c\lambda\alpha\left[e^{Mx} + \sum_{i=1}^{n}\lambda A_i e^{Mx}(\theta_i I - M)^{-1}(e^{(\theta_i I - M)b} - I)\right]\vec{1} & \text{if } x > b \end{cases} \tag{2.29}$$

where θ_i, $i = 1, 2, \cdots, n$, are n distinct roots to Equation (2.24)*. A_i, $i = 1, 2, \cdots, n$, are given by Equation* (2.26)*, and $G^*(s)$ is given by Equation* (2.28)*.*

The probability that the system is empty is:

$$\begin{aligned} c = \Bigg\{ & \sum_{i=1}^{n} \lambda\frac{A_i}{\theta_i}(e^{\theta_i b} - 1) \\ & - \alpha\Big[\sum_{i=1}^{n}\lambda^2 A_i M^{-1} e^{Mb}(\theta_i I - M)^{-1}(e^{(\theta_i I - M)b} - I)\Big]\vec{1} \\ & - \lambda\alpha M^{-1} e^{Mb}\vec{1} + 1\Bigg\}^{-1}. \end{aligned} \tag{2.30}$$

Proof. Use Equation 2.25 to take inverse LT of $f_1^*(s)$ and apply Theorem 2. □

A straight forward computation yields the following corollaries.

Corollary 1 (cdf)**.** *The equilibrium cdf of the workload process of the $M/PH/1$ wait-*

based balking queue with threshold b is:

$$F(x) = \begin{cases} c + c\lambda \sum_{i=1}^{n} \frac{A_i}{\theta_i}(e^{\theta_i x} - 1), & \text{if } 0 \le x \le b, \\ F(b) + c\lambda\alpha M^{-1}(e^{Mx} - e^{Mb}) & \\ \quad \times \left[I + \sum_{i=1}^{n} \lambda A_i (\theta_i I - M)^{-1}(e^{(\theta_i I - M)b} - I) \right] \vec{1}, & \text{if } x > b. \end{cases}$$

Corollary 2 (Mean workload in equilibrium)**.** *The expected value of the workload in steady state of the $M/PH/1$ wait-based balking queue with threshold b is:*

$$\begin{aligned} \bar{W} = b - bc - c\lambda \sum_{i=1}^{n} \frac{A_i}{\theta_i^2}(e^{\theta_i b} - \theta_i b - 1) \\ + c\lambda\alpha M^{-2} e^{Mb} \left[I + \sum_{i=1}^{n} \lambda A_i (\theta_i I - M)^{-1}(e^{(\theta_i I - M)b} - I) \right] \vec{1}. \end{aligned}$$

Remarks:

1. If we write Equation (2.24) as $1 - \lambda\alpha(sI - M)^{-1}\vec{1} = 0$, it is easy to see that when $\rho = 1$, i.e.,$-\lambda\alpha M^{-1}\vec{1} = 1$, then $\theta_1 = 0$ is one of the roots. In this case, we replace the zero-dividing terms which appear in our results by the limits. These terms and the limits are: $\lim_{\theta_1 \to 0}(e^{\theta_1 b} - 1)/\theta_1 = b$ and $\lim_{\theta_1 \to 0}(e^{\theta_1 b} - \theta_1 b - 1)/\theta_1^2 = b^2/2$. Therefore, we do not give the results separately for the case when $\rho = 1$.

2. Computing c needs the integral $\int_b^\infty e^{Mx}$ to converge. This is guaranteed by the aforementioned fact that all eigenvalues of M have negative real part.

3. Notice that

$$E(S) = -\alpha M^{-1}\vec{1} = \tau, \tag{2.31}$$

$$E(S^2) = 2\alpha M^{-2}\vec{1} = \sigma^2 + \tau^2. \tag{2.32}$$

It can be verified that the formulas above for limiting parameters b and ρ are consistent with those given for general service time distribution in Section 2.3. This is also illustrated by numerical examples in Section 2.6.

2.4.2 $M/PH/1$: Differential Equation Approach

In practice it can be hard to compute $G^*(s)$, i.e., specifying the polynomials, for a general phase type distribution. Therefore, we seek an alternative way to solve Equation (2.7) directly. In order to solve for $f(x)$, we first solve the integral equation (2.3a) for the case $x \leq b$ by solving a derived differential equation. Then we compute $f(x)$ for the case $x > b$ by the integral equation. We describe the method in Theorem 4.

First we introduce some notations that will be used in the statement and proof of Theorem 4.

Let $a_i, i = 0, 1, \cdots, n$, be the coefficients of the characteristic polynomial of M, i.e. :

$$\det(xI - M) = \sum_{j=0}^{n} a_j x^j. \tag{2.33}$$

Let

$$P(\theta) = \sum_{i=0}^{n} \left[\alpha \Big(a_i I + \lambda \sum_{j=0}^{i} a_j M^{j-i-1} \Big) \vec{1} \right] \theta^i \tag{2.34}$$

be an n-th order polynomial in θ. We assume $P(\theta)$ has n distinct roots. Note that if $-\alpha M^{-1}\vec{1} = 1/\lambda$ (i.e. traffic density $\rho = -\lambda\alpha M^{-1}\vec{1} = 1$), then $\theta_1 = 0$ is one of the roots.

Let $M_0 = I$, and define $M_j, j \geq 1$ recursively by:

$$M_j = M M_{j-1} + \lambda \alpha M_{j-1} \vec{1} I. \tag{2.35}$$

Let

$$\begin{aligned} m_i &= \alpha M_i \vec{1},\ i = 0, 1, \cdots, n-1 \text{ and} \\ m &= (m_0, m_1, \cdots, m_{n-1})^{\mathrm{T}}. \end{aligned} \tag{2.36}$$

With these notations, we are ready to state the following theorem.

Theorem 4. *The constants θ_i, $i = 1, 2, \cdots, n$, in Theorem 3 are the roots of $P(\theta)$ given in (2.34). The coefficient $A = (A_1, A_2, \cdots, A_n)^{\mathrm{T}}$ is uniquely determined by:*

$$\Theta A = m, \tag{2.37}$$

where Θ is the Vandermonde matrix of $\theta_1, \theta_2, \cdots, \theta_n$, i.e.:

$$\Theta = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \cdots & \theta_n \\ \theta_1^2 & \theta_2^2 & \cdots & \theta_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1^{n-1} & \theta_2^{n-1} & \cdots & \theta_n^{n-1} \end{pmatrix}.$$

Proof. Plugging (2.27) in Equation (2.7) and simplifying, we get:

$$f_1(x) = \lambda \alpha e^{Mx} \left[\int_0^x f_1(u) e^{-Mu} du + I/\lambda \right] \vec{1}. \tag{2.38}$$

Taking repeated derivatives of the equation above with respect to x, we get (using

$f_1^{(i)}(x)$ to denote the i-th order derivative of $f_1(x)$ with respect to x):

$$\begin{aligned}
f_1^{(0)}(x) &= \lambda\alpha\Big[M^0 e^{Mx}\Big(\int_0^x f_1(u)e^{-Mu}du + I/\lambda\Big)\Big]\vec{1},\\
f_1^{(1)}(x) &= \lambda\alpha\Big[M^1 e^{Mx}\Big(\int_0^x f_1(u)e^{-Mu}du + I/\lambda\Big) + f_1^{(0)}(x)I\Big]\vec{1},\\
f_1^{(2)}(x) &= \lambda\alpha\Big[M^2 e^{Mx}\Big(\int_0^x f_1(u)e^{-Mu}du + I/\lambda\Big) + Mf_1^{(0)}(x) + f_1^{(1)}(x)I\Big]\vec{1},\\
&\vdots\\
f_1^{(n)}(x) &= \lambda\alpha\Big[M^n e^{Mx}\Big(\int_0^x f_1(u)e^{-Mu}du + I/\lambda\Big) + \sum_{j=0}^{n-1} M^j f_1^{(n-1-j)}(x)\Big]\vec{1}.
\end{aligned} \tag{2.39}$$

Now multiply the i-th equation above by coefficient a_i in the characteristic polynomial of M as defined in (2.33) and add. We get:

$$\sum_{i=0}^{n} a_i f_1^{(i)}(x) = \lambda\alpha\Big[\sum_{i=0}^{n} a_i M^i e^{Mx}\Big(\int_0^x f_1(u)e^{-Mu}du + I/\lambda\Big) + \sum_{i=0}^{n}\sum_{j=0}^{i-1} a_i M^j f_1^{(i-1-j)}(x)\Big]\vec{1}. \tag{2.40}$$

By Cayley-Hamilton Theorem, we know

$$\sum_{i=0}^{n} a_i M^i = 0. \tag{2.41}$$

Using this and doing algebraic manipulations, Equation (2.40) can be simplified and rewritten as:

$$\sum_{i=0}^{n}\Big[\alpha\Big(a_i I + \lambda\sum_{j=0}^{i} a_j M^{j-i-1}\Big)\vec{1}\Big] f_1^{(i)}(x) = 0. \tag{2.42}$$

This is simply an n-th order differential equation with constant coefficients. Using standard methods of solving such equations (cf. [28]), we get the polynomial (2.34) and the solution $f_1(x) = \sum_{i=1}^{n} A_i e^{\theta_i x}$, where the constants A_i's are to be determined by using the initial conditions.

The initial conditions can be found by plugging $x = 0$ in (2.39):

$$\lim_{x\to 0} f_1^{(j)}(x) = m_j,\ j = 0, 1, \cdots, n-1, \tag{2.43}$$

where m_j are as defined in (2.36). This yields (2.37). Since all θ_i are distinct by assumption, Θ is invertible (cf. [20]) thus A is uniquely determined. □

2.4.3 Special Cases

Next, we consider three special cases of service time distribution: Erlang, hyper-exponential and exponential. These belong to the common phase type distribution category, and have special parameters (α, M). For these cases we can, more or less, simplify the results for a general phase type distribution.

Erlang

An Erlang distribution with parameter (n, μ) is simply a phase type distribution with parameter:

$$\alpha = (1, 0, \cdots, 0)_{1\times n},$$

$$M = \begin{pmatrix} -\mu & \mu & & & \\ & -\mu & \mu & & \\ & & \ddots & \ddots & \\ & & & -\mu & \mu \\ & & & & -\mu \end{pmatrix}_{n\times n}$$

(We display only the non-zero entries of M).

We solve $f_1(x)$ by transform. Using Equation (2.28), $G^*(s)$ can be shown to be

$$G^*(s) = \frac{(s+\mu)^n - \mu^n}{(s+\mu)^n s}. \tag{2.44}$$

Finding the roots to $1 - \lambda G^*(s)$ is equivalent to solving the following equation:

$$\frac{1}{s}[(s-\lambda)(s+\mu)^n + \lambda\mu^n] = 0. \tag{2.45}$$

Note that the left hand side is actually an n degree polynomial in s. It can be proved that there are exactly n distinct roots, $\theta_1, \theta_2, \cdots, \theta_n$. So, writing $f_1^*(s)$ as

$$\begin{aligned} f_1^*(s) &= \frac{G^*(s)}{1-\lambda G^*(s)} \\ &= \frac{\frac{1}{s}[(s+\mu)^n - \mu^n]}{\frac{1}{s}[(s-\lambda)(s+\mu)^n + \lambda\mu^n]} \\ &= \frac{(s-\lambda)^{-1}[\prod_{i=1}^n (s-\theta_i) - \prod_{i=1}^n (\lambda-\theta_i)]}{\prod_{i=1}^n (s-\theta_i)}, \end{aligned} \tag{2.46}$$

then computing A_i by Equation (2.26) yields

$$A_i = \prod_{j\neq i} \frac{\lambda-\theta_j}{\theta_i-\theta_j},\ i = 1, 2, \cdots, n.$$

Next, we compute the matrix exponential explicitly and give the result in Theorem 5. Before that, we introduce two more notations. We denote the well known incomplete gamma function by:

$$\Gamma(n, x) = \int_x^\infty t^{n-1}e^{-t}dt = (n-1)!e^{-x}\sum_{k=0}^{n-1}\frac{x^k}{k!}.$$

Let

$$d_i = \frac{\mu}{\mu+\theta_i},\ i = 1, 2, \cdots, n.$$

Theorem 5. *The equilibrium pdf of the workload process of the $M/E_n/1$ wait-based balking queue with threshold b is:*

$$f(x) = c\lambda\sum_{i=1}^n A_i e^{\theta_i x},\ \textit{if } 0 \le x \le b,$$

and

$$f(x) = c\lambda^2 \sum_{i=1}^{n} \frac{A_i}{\theta_i (n-1)!} \Big\{ d_i^n e^{\theta_i x} \big[\Gamma\big(n, (\mu+\theta_i)x\big) - \Gamma\big(n, (\mu+\theta_i)(x-b)\big) \big] + e^{\theta_i b} \Gamma\big(n, \mu(x-b)\big) - \Gamma(n, \mu x) \Big\} + c\lambda \frac{\Gamma(n, \mu x)}{(n-1)!}, \ \textit{if } x > b.$$

$$\begin{aligned} c^{-1} =& \lambda \sum_{i=1}^{n} \frac{A_i}{\theta_i} (e^{\theta_i b} - 1) + \frac{\lambda}{\mu (n-1)!} [\Gamma(n+1, \mu b) - \mu b \Gamma(n, \mu b)] + 1 \\ &+ \lambda^2 \sum_{i=1}^{n} \frac{A_i}{\mu \theta_i^2 (n-1)!} \{ \mu (1 + \theta_i b) \Gamma(n, \mu b) - \mu d_i^n e^{\theta_i b} \Gamma\big(n, (\mu+\theta_i) b\big) \\ &- \theta_i \Gamma(n+1, \mu b) + \theta_i e^{\theta_i b} n! - \mu e^{\theta_i b} (1 - d_i^n)(n-1)! \} \end{aligned}$$

Hyper-exponential

If the service time S has a hyper-exponential distribution (cf. [29]), then

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T,$$

$$M = \begin{pmatrix} -\mu_1 & & \\ & \ddots & \\ & & -\mu_n \end{pmatrix}.$$

In this case, $f_1(x)$ can be solved either by transform or directly. Here we only show some results by using Theorem 4 to solve $f_1(x)$ directly.

The characteristic polynomial of M is:

$$\sum_{i=0}^{n} a_i x^i = \prod_{i=1}^{n} (x + \mu_i).$$

Then the coefficients a_i can be computed easily. It is possible to simplify $P(\theta)$ in

Equation (2.34) in terms of the moments of S as follows:

$$P(\theta) = \sum_{i=0}^{n} b_i \theta^i,$$

where

$$b_i = a_i + \lambda \sum_{j=1}^{i+1} a_{i+1-j}(-1)^j \frac{E(S^j)}{j!}.$$

Unfortunately, the initial conditions do not simplify, hence we keep the remaining results in terms of m of (2.36) and A of (2.37).

Exponential

Exponential distribution is the most special case of phase type distribution. It is also a special case of Erlang or hyper-exponential distributions. All matrices and vectors we use in a common phase type distribution degenerate to scalars in an exponential distribution. The parameters are simply $\alpha = (1)$ and $M = (-\mu)$. Using either transform of direct method, the formulas are much simplified.

We give the simplified solution in the following theorem and skip the proof.

Theorem 6. *The equilibrium pdf of the workload process of the $M/M/1$ wait-based balking queue with threshold b is:*

$$f(x) = \begin{cases} c\lambda e^{-(\mu-\lambda)x} & \text{if } 0 \le x \le b \\ c\lambda e^{\lambda b} e^{-\mu x} & \text{if } x > b \end{cases},$$

where

$$c = \begin{cases} \frac{1-\rho}{1-\rho^2 e^{-(\mu-\lambda)b}} & \text{if } \rho \neq 1 \\ \frac{1}{2+\mu b} & \text{if } \rho = 1 \end{cases}.$$

2.5 An Example of Non-rational $G^*(s)$: $M/D/1$

As we mentioned before, it can be difficult to find the inverse LT of $f_1^*(s)$ when $G^*(s)$ is not rational. In this case, we try to solve Equation (2.7) directly. Here we give the solution when the service time is deterministic, i.e.,

$$G(x) = \begin{cases} 1 & \text{if } 0 \le x < \tau \\ 0 & \text{if } x \ge \tau \end{cases} \tag{2.47}$$

In this case, the LT of f_1 is given by

$$f_1^*(s) = \frac{1 - e^{s\tau}}{s - \lambda + \lambda e^{-s\tau}}. \tag{2.48}$$

Computing its inverse is intractable. We show how we solve (2.7) directly in this case.

First we partition $[0, +\infty)$ into intervals of length τ: $[0, \tau), [\tau, 2\tau), \cdots$ and denote them as $I_0, I_1, \cdots$ respectively. Since $G(x)$ is 1 when $x \in I_0$ (or $\lfloor \frac{x}{\tau} \rfloor = 0$) and 0 elsewhere, we rewrite Equation (2.7) as follows:

$$\begin{aligned} &\lambda \int_0^x f_1(u)du + 1 = f_1(x), && \text{when } x \in I_0, \\ &\lambda \int_{x-\tau}^{k\tau} f_1(u)du + \lambda \int_{k\tau}^{x} f_1(u)du = f_1(x), && \text{when } x \in I_k, k = 1, 2, \cdots. \end{aligned} \tag{2.49}$$

We solve these equations recursively and obtain $f_1(x)$ for each interval. That is, we solve the first equation and get $f_1(x) = e^{\lambda x}$ when $x \in I_0$. Plugging this in the second equation, we are able solve for $f_1(x)$ when $x \in I_1$, and so on. Suppose

$$f_1(x) = Q_k(x)e^{\lambda(x-k\tau)}, \text{ when } x \in I_k, k = 0, 1, 2, \cdots, \tag{2.50}$$

where $Q_k(x)$ is a polynomial in x and $Q_0(x) = 1$. Substituting (2.50) in (2.49) and

take derivative with respect to x, we get

$$Q_k'(x) = -\lambda Q_{k-1}(x-\tau), \quad k = 1, 2, \cdots . \tag{2.51}$$

Therefore, $Q_k(x)$ can be computed recursively as

$$Q_k(x) = -\lambda \int_0^x Q_{k-1}(u-\tau)du + B_k, \quad k = 1, 2, \cdots . \tag{2.52}$$

The constant B_k can be computed by the fact that $f_1(\tau-) = f_1(\tau+) + 1$ and $f_1(x)$ is continuous at $2\tau, 3\tau, \cdots$. We get

$$B_1 = e^{\rho} + \rho - 1,$$
$$B_k = Q_{k-1}(k\tau)e^{\rho} + \lambda \int_0^{k\tau} Q_{k-1}(u-\tau)du, \quad k = 2, 3, \cdots .$$

In the special case when $b < \tau$, the computation is fairly easy. We give the results here.

$$f(x) = \begin{cases} c\lambda e^{\lambda x} & \text{if } 0 \le x \le b \\ c\lambda e^{\lambda b} & \text{if } b < x < \tau \\ c\lambda(e^{\lambda b} - e^{\lambda(x-\tau)}) & \text{if } \tau \le x < b + \tau \\ 0 & \text{elsewhere} \end{cases} \tag{2.53}$$

In this case the probability that the system is empty is

$$c = \frac{1}{\tau\lambda e^{\lambda b} + 1}. \tag{2.54}$$

2.6 Numerical Examples

In this section, we illustrate our results with several numerical examples. We consider three different service time distributions:

1. Exponential (exp): $\mu = 1$ ($\tau = 1$, $\sigma^2 = 1$);

2. 5-Erlang (erlang): $\mu = 5$ ($\tau = 1$, $\sigma^2 = 0.2$);

3. Hyper-exponential (hyper): $\mu_1 = 4, \mu_2 = 2, \mu_3 = 1, \mu_4 = 0.8, \mu_5 = 0.5$, $\alpha_1 = \cdots = \alpha_5 = 0.2$ ($\tau = 1$, $\sigma^2 = 1.75$).

All of them have mean service time of one. The variances are different, with 5-Erlang the smallest and hyper-exponential the largest.

The first set of figures, Figure 2.3, 2.4, 2.5 illustrate the shapes of $f(x)$ for different service time distributions, with $\rho = 0.8, 1, 1.2$ respectively and $b = 5$. Then, we pick curves for exponential service time from Figure 2.3, 2.4, 2.5 and put them together in Figure 2.6 to show how $f(x)$ is affected by the traffic intensity. As ρ gets larger, the turn at $x = b$ becomes sharper. When $\rho \gg 1$ (in our experiment, $\rho = 10$ is large enough), $f(x)$ is almost 0 when $x < b$. Note that $f(x)$ is a decreasing function of x for $x > b$. However, for $0 < x < b$, the density function can exhibit complex behavior. It may be increasing, decreasing, constant or non-monotonic.

The second set of figures, Figure 2.7, 2.8, 2.9 and 2.10, are about the expected workload in steady state ($\bar{W}$). In Figure 2.7, we fix $b = 5$ and compare $\bar{W}$ for different service time distributions against ρ (since $\tau = 1$, $\rho = \lambda$). When ρ gets larger, $\bar{W}$ clearly converges to the levels as expected (see Equation (2.18)), to be specific, to 6, 5.6 and 6.375 for exponential, Erlang and hyper-exponential distributions, respectively.

In Figure 2.8, 2.9 and 2.10, for each graph, we fix ρ and plot $\bar{W}$ against b for different service time distributions. At $b = 0$, $\bar{W}$ starts from the value we expect (see Equation (2.12) and (2.14)), and then it increases as b increases. When $\rho = 0.8$, the convergence of $\bar{W}$ to the theoretical level (i.e. normal $M/G/1$ queue) is again clearly shown. On the other hand, $\bar{W}$ increases rapidly in b when $\rho \geq 1$ ($\rho = 1, 1.2$).

Finally, Figure 2.11 shows the probability that the system is empty in steady state. As expected, c is decreasing in arrival rate λ.

2.7 Concluding Remarks

It is possible to extend the method used in this chapter to handle more complicated cases, e.g., models with load dependent service rates and vacations, or random balking threshold. We do not cover those in order to emphasize a basic idea by relatively simple models.

Although we focus on the steady state distribution of the workload process in this chapter, typically several other system performance measures are of interest. For example, the steady state rejection rate can be easily computed as $1 - F(b)$, where F is the steady state cdf of the vqt process. Secondly, the expected busy period can be computed from the standard regenerative analysis as $(1 - c)/(\lambda c)$, where c is the probability that the workload is zero in steady state. Similarly, the expected number in the system can be computed by using Little's law. However, the distribution of the busy period would require further analysis. We shall study this topic in Chapter 3. Also, since we have computed the steady state distribution of the workload, the expected values of its functionals are easy to compute.

Two more types of the queue with restricted accessibility, where balking rules are related to workload, are introduced by Perry and Asmussen (cf. [38], Model II and Model III). If the service time is phase type distributed, then our method can be applied to these models as well.

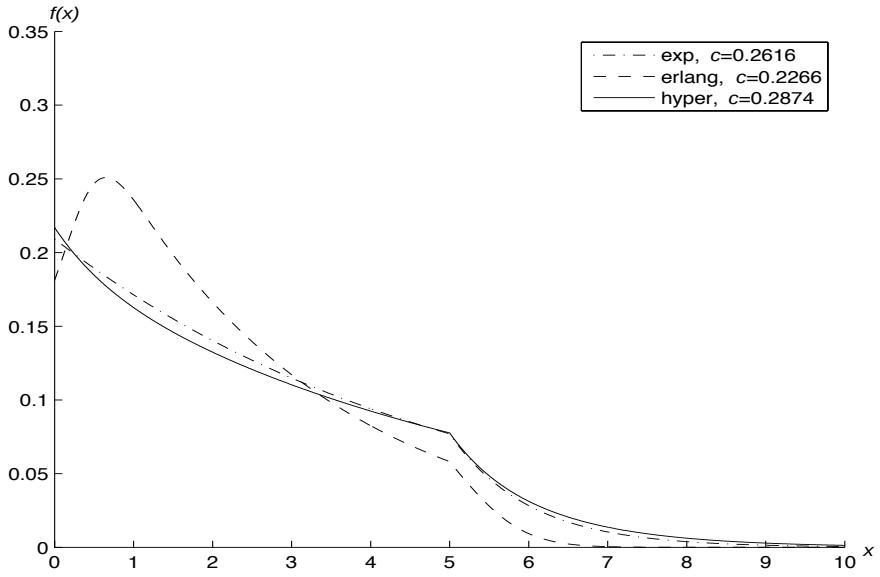

Figure 2.3: $f(x)$ for different service time distributions, $\rho = 0.8, b = 5$

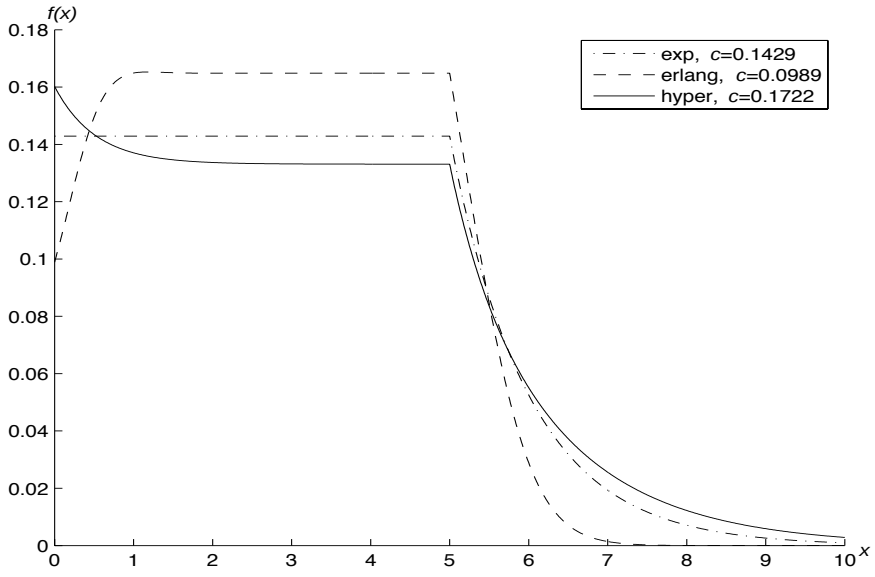

Figure 2.4: $f(x)$ for different service time distributions, $\rho = 1, b = 5$

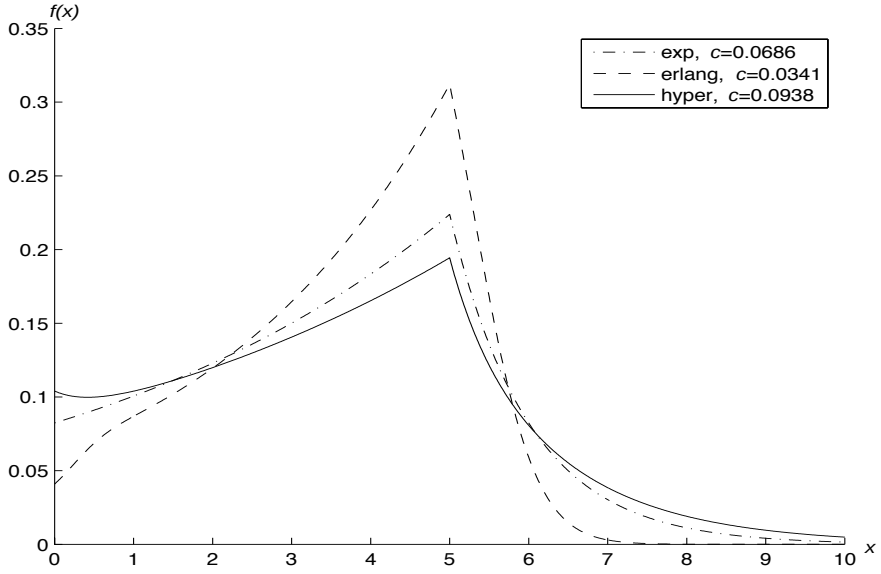

Figure 2.5: $f(x)$ for different service time distributions, $\rho = 1.2, b = 5$

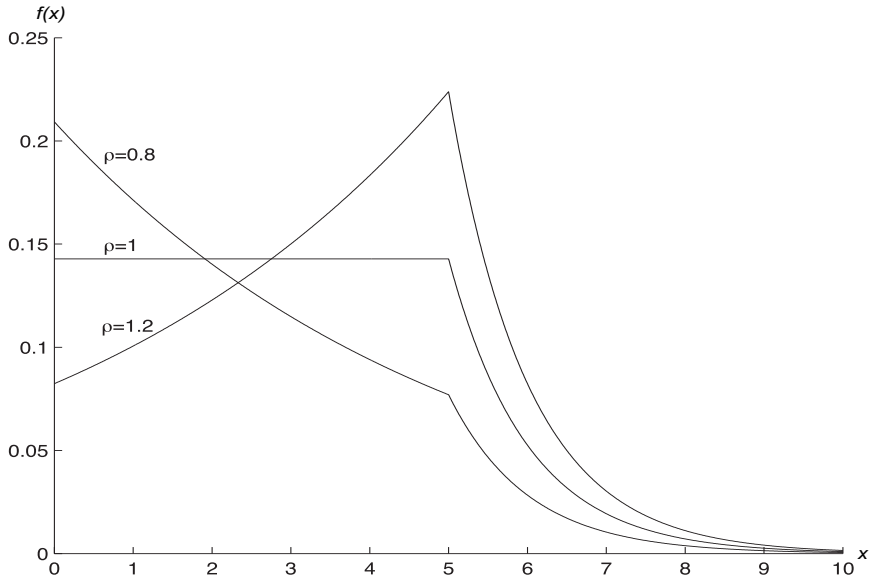

Figure 2.6: $f(x)$ for different ρ, exponential service time, $b = 5$

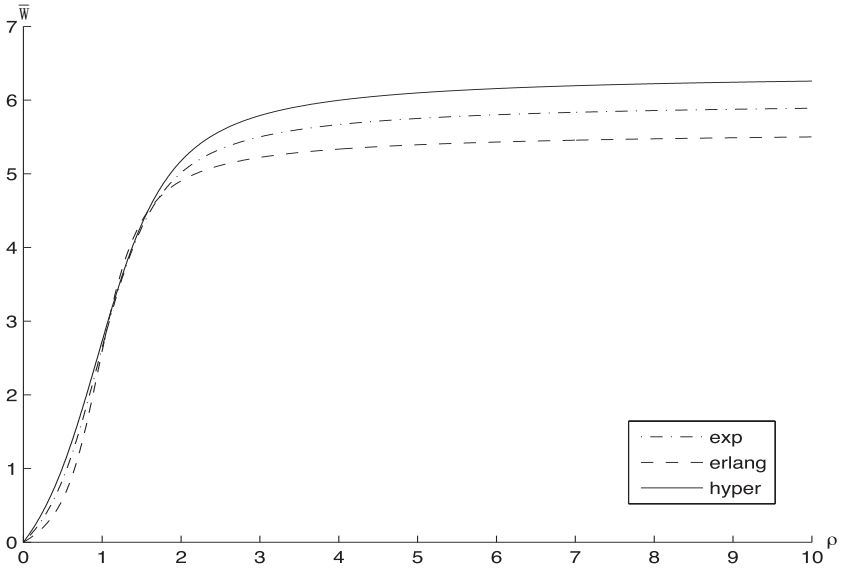

Figure 2.7: $\bar{W}$ vs. ρ for different service time distributions, $b = 5$

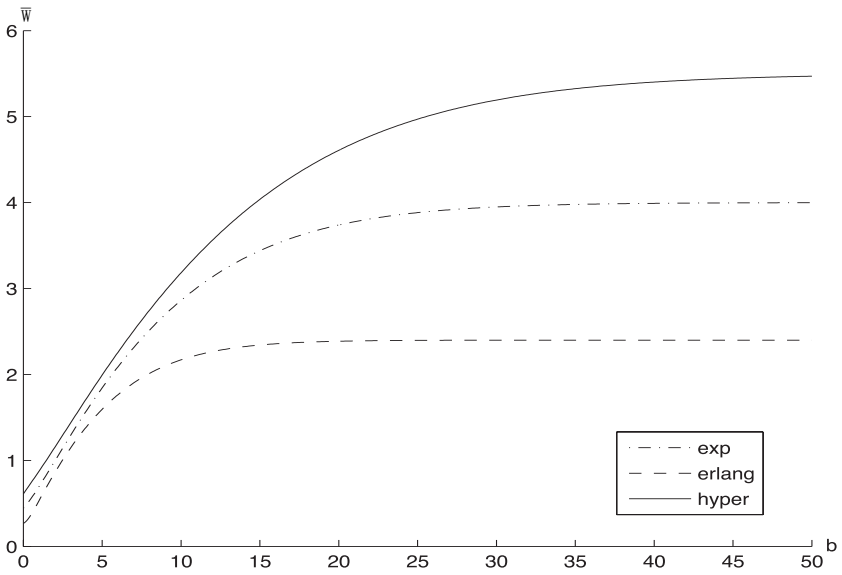

Figure 2.8: $\bar{W}$ vs. b for different service time distributions, $\rho = 0.8$

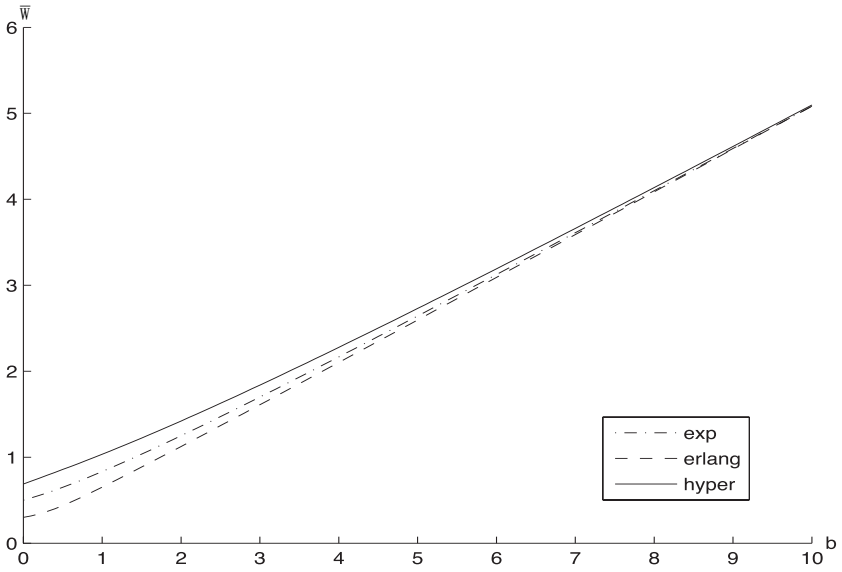

Figure 2.9: $\bar{W}$ vs. b for different service time distributions, $\rho = 1$

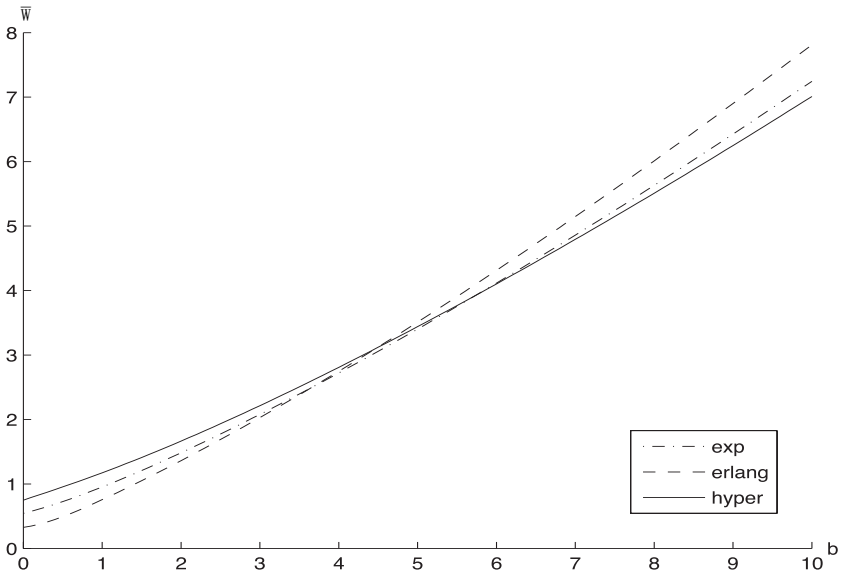

Figure 2.10: $\bar{W}$ vs. b for different service time distributions, $\rho = 1.2$

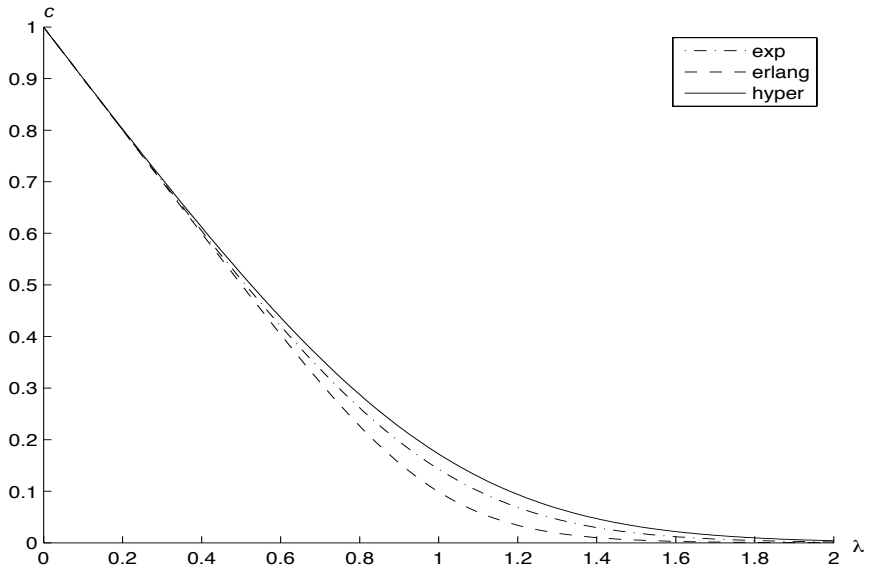

Figure 2.11: c vs. λ for different service time distributions

Chapter 3

$M/PH/1$ Queues with Wait-based Balking: Busy Period Analysis

3.1 Introduction

Consider the $M/PH/1$ wait-based balking queue with a fixed balking threshold b, as described in Chapter 2 (see page 18 for phase type distribution).

In this chapter, we are interested in the first passage time

$$B = \min\{t \geq 0 : W(t) = 0\}.$$

Specifically, we compute the mean

$$m(x) = \mathrm{E}[B|W(0) = x] \tag{3.1}$$

and the LST

$$\psi(s, x) = \mathrm{E}[e^{-sB}|W(0) = x]. \tag{3.2}$$

Our method involves constructing a standard fluid model so that the $M/PH/1$ wait-based balking model is a limiting case of this fluid model. Kulkarni and Tzenova

[32] studied the first passage times in fluid models. They developed a differential equation method which is more suitable for numerical algorithms. We slightly extend the method so that it can be applied to the fluid model we construct, and obtain the mean and LST of the first passage time. The precise formulation of our fluid model is given in Section 3.2.

The rest of the chapter is organized as follows. We formulate the relevant fluid model in Section 3.2 and give a general treatment of the first passage time problem for this fluid model in Section 3.3. In Section 3.4, we consider a special case of the fluid model and give the mean and LST in Theorems 10 and 11. In Section 3.5 we show that the balking model is a limiting case of the fluid model analyzed in Section 3.4. Then we take the limits of the results given in Theorems 10 and 11 and obtain explicit formulas for the mean and LST of the first passage time for the balking model in Theorems 12 and 13. In Section 3.6, we illustrate the usage of these formulas to compute the first passage time in the $M/M/1$ case and verify several known results. We present several numerical examples in Section 3.7.

3.2 The Fluid Model

Consider a fluid model with an infinite capacity buffer where the net flow rate is governed by a stochastic process $\{Z(t), t \geq 0\}$ with a finite state space $\mathcal{S} = \{0, 1, \cdots, n\}$ as follows: if $Z(t) = i$ the buffer level changes at rate r_i. For convenience, we assume that $r_i \neq 0, \forall i \in \mathcal{S}$. It will be clear later that this assumption is not a severe restriction in the context of our application. Let $\mathcal{S}_- = \{i : r_i < 0\}$, $\mathcal{S}_+ = \{i : r_i > 0\}$, $n_- = |\mathcal{S}_-|$ and $n_+ = |\mathcal{S}_+|$.

Let $X(t)$ be the amount of fluid in the buffer at time t. The dynamics of the

buffer content process $X = \{X(t), t \geq 0\}$ is given by:

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \begin{cases} r_i & \text{if } Z(t) = i, X(t) > 0, \\ \max(r_i, 0) & \text{if } Z(t) = i, X(t) = 0. \end{cases}$$

The X process is called a *fluid input-output process driven by the Z process* (cf. Kulkarni [30]). The driving process Z behaves as a CTMC whose infinitesimal generator matrix $Q(x)$ depends on the current buffer level x as follows:

$$Q(x) = \begin{cases} Q & \text{if } 0 \leq x \leq b, \\ \bar{Q} & \text{if } x > b, \end{cases}$$

where $Q = [q_{ij}]$ ($\bar{Q} = [\bar{q}_{ij}]$) is the generator of a CTMC. We assume that Q and $\bar{Q}$ have exactly one irreducible class. Such CTMCs have unique limiting distributions that are independent of their initial states.

A more general version of a such a model is studied by Scheinhardt et al. [43] under the name "feedback fluid queue".

Clearly the joint process $\{(X(t), Z(t)), t \geq 0\}$ is a Markov process which is characterized by the matrices Q, $\bar{Q}$ and $R = \mathrm{diag}(r_0, r_1, \cdots, r_n)$. Let $p = [p_0, p_1, \cdots, p_n]$ be the solution to

$$pQ = 0, \quad p\vec{1} = 1$$

and $\bar{p} = [\bar{p}_0, \bar{p}_1, \cdots, \bar{p}_n]$ be the solution to

$$\bar{p}\bar{Q} = 0, \quad \bar{p}\vec{1} = 1.$$

Let

$$d = \sum_{i \in \mathcal{S}} p_i r_i$$

and

$$\bar{d} = \sum_{i \in \mathcal{S}} \bar{p}_i r_i.$$

It is easy to see that $\bar{d} < 0$ is a sufficient condition for the joint process to be stable (Kulkarni [30]). We assume this condition holds, i.e., the joint process is stable.

3.3 First Passage Time: the Fluid Model

In this section, we compute the mean and LST of the first passage time

$$T = \min\{t \geq 0 : X(t) = 0\}.$$

Let

$$\pi_i(x) = \mathrm{E}[T | Z(0) = i, X(0) = x]$$

and denote

$$\pi(x) = [\pi_0(x), \pi_1(x), \cdots, \pi_n(x)]^{\mathrm{T}}.$$

Theorem 7. *The vector $\pi(x)$ satisfies the following system of differential equations*

$$R\pi'(x) = \begin{cases} -\vec{1} - Q\pi(x), & \text{if } 0 < x < b, \\ -\vec{1} - \bar{Q}\pi(x), & \text{if } x > b, \end{cases} \tag{3.3}$$

with boundary conditions:

$$\begin{aligned} \pi_i(0) &= 0, \ i \in \mathcal{S}_-, \\ \pi(b-) &= \pi(b+). \end{aligned} \tag{3.4}$$

Proof. For $0 < x < b$, consider an infinitesimal time interval $[0, h]$ and use first step analysis to obtain:

$$\pi_i(x) = h + (1 + q_{ii}h)\pi_i(x + r_i h) + \sum_{j \neq i} q_{ij} h \pi_j(x + r_i h), \ \forall i,$$

or

$$\pi_i(x - r_i h) = h + (1 + q_{ii}h)\pi_i(x) + \sum_{j \neq i} q_{ij} h \pi_j(x), \ \forall i.$$

Divide both sides by h and let $h \downarrow 0$. After some algebra, we get:

$$-r_i \pi_i'(x) = 1 + \sum_j q_{ij} \pi_j(x). \tag{3.5}$$

The system of differential equations is obtained by rearranging the terms and writing the $(n+1)$ equations in a matrix form. Same argument goes for the case $x > b$. The boundary conditions are obvious. □

Notice that Equation (3.5) shows that it is possible to eliminate an unknown $\pi_j(x)$ if $r_j = 0$. Hence assuming $r_i \neq 0, \ \forall i \in \mathcal{S}$ is not a severe restriction.

Similar equations are derived and solved in Kulkarni and Tzenova [32] by using well-known techniques. We slightly extend the method that is used in Theorem 3.3 in Kulkarni and Tzenova [32] and apply it to the problem we consider here. First we state the following lemma.

Lemma 1. (From Theorem 11.5, Kulkarni [30]). *Suppose $\bar{Q}$ has exactly one irreducible class. Then there are $n+1$ (possibly repeated) eigenvalues of $-R^{-1}\bar{Q}$. When $\bar{d} < 0$, exactly n_+ have positive real parts, one is zero, and $n_- - 1$ have negative real parts.*

Order the eigenvalues $\bar{\theta}_i$ as follows:

$$Re(\bar{\theta}_0) \leq Re(\bar{\theta}_1) \leq \cdots \leq Re(\bar{\theta}_{n_- - 2}) \leq Re(\bar{\theta}_{n_- - 1}) = 0 < Re(\bar{\theta}_{n_-}) \leq \cdots \leq Re(\bar{\theta}_n).$$

Let $\bar{v}_i$ be the right eigenvector corresponding to eigenvalue $\bar{\theta}_i$.

Similarly, we use the notation θ_i and v_i $(i = 0, 1, ..., n)$, respectively, for the eigenvalues and right eigenvectors of $-R^{-1}Q$. But we do not order θ_i in the same fashion as $\bar{\theta}_i$, and we do not assume $d < 0$.

We give the main result in the following theorem. The proof is similar to that of Theorem 3.3 in Kulkarni and Tzenova [32] and is omitted here.

Theorem 8. *Let I be the $(n+1)\times(n+1)$ identity matrix. The solution to Equation (3.3) and (3.4) is given by:*

$$\pi(x) = \begin{cases} \sum_{j=0}^{n} a_j v_j e^{\theta_j x} - \frac{\vec{1}x}{d} + c, & \text{if } 0 \le x \le b, \\ \sum_{j=0}^{n_- -1} \bar{a}_j \bar{v}_j e^{\bar{\theta}_j x} - \frac{\vec{1}x}{\bar{d}} + \bar{c}, & \text{if } x > b, \end{cases} \tag{3.6}$$

where c is any solution to the linear system

$$Qc = (R/d - I)\vec{1},$$

$\bar{c}$ is any solution to the linear system

$$\bar{Q}\bar{c} = (R/\bar{d} - I)\vec{1},$$

and the coefficients $\{a_j, 0 \le j \le n\}$ and $\{\bar{a}_j : 0 \le j \le n_- - 1\}$ are obtained by using the boundary conditions given in Equation (3.4).

Next, we compute the LST of the first passage time T:

$$\phi_i(s,x) = \mathrm{E}[e^{-sT} | Z(0) = i, X(0) = x].$$

Let

$$\phi(s,x) = [\phi_0(s,x), \phi_1(s,x), \cdots, \phi_n(s,x)]^{\mathrm{T}}.$$

The following theorem is analogous to Theorem 7.

Theorem 9. *The vector $\phi(s,x)$ satisfies the following system of differential equations*

$$R\frac{\mathrm{d}\phi(s,x)}{\mathrm{d}x} = \begin{cases} (sI-Q)\phi(s,x), & \text{if } 0<x<b, \\ (sI-\bar{Q})\phi(s,x), & \text{if } x>b, \end{cases} \tag{3.7}$$

with boundary conditions:

$$\begin{aligned} &\phi_i(s,0) = 1, \ i \in \mathcal{S}_-, \\ &\phi(s,b-) = \phi(s,b+). \end{aligned} \tag{3.8}$$

Proof. For $0<x<b$, consider an infinitesimal time interval $[0,h]$ and use first step analysis to obtain:

$$\phi_i(s,x) = e^{-sh}(1+q_{ii}h)\phi_i(s,x+r_ih) + \sum_{j\neq i} e^{-sh}q_{ij}h\phi_j(s,x+r_ih), \ \forall i,$$

or

$$\phi_i(s,x-r_ih) = e^{-sh}(1+q_{ii}h)\phi_i(s,x) + \sum_{j\neq i} e^{-sh}q_{ij}h\phi_j(s,x), \ \forall i.$$

Divide both sides by h and let $h \downarrow 0$. After some algebra, we get:

$$-r_i\frac{\mathrm{d}\phi_i(s,x)}{\mathrm{d}x} + s\phi_i(s,x) = \sum_j q_{ij}\phi_j(s,x), \ \forall i.$$

The system of differential equations is obtained by rearranging the terms and writing the $(n+1)$ equations in a matrix form. Same argument goes for the case $x>b$. The boundary conditions are obvious. □

It is more complicated to solve Equations (3.7) and (3.8). However, we shall see in the next section that for some special cases, explicit solutions can be obtained.

3.4 A Special Case of the Fluid Model

In this section we consider a special case of the fluid model whose R, Q and $\bar{Q}$ matrices are as given below. As we shall see, this helps us solve the first passage time problem for the balking model we have described at the beginning of this chapter. Let M and α be the parameters of the phase type distribution as defined in Section 2.4.1 and λ be the arrival rate. The fluid model is parameterized by a real number $r > 0$. Let

$$\begin{aligned} r_0 = -1, r_i = r > 0 \quad \text{for} \;\; i = 1, 2, \cdots, n, \\ R = \text{diag}(r_0, r_1, \cdots, r_n), \end{aligned} \tag{3.9}$$

$$Q = \begin{bmatrix} -\lambda & \lambda\alpha \\ -Mr\vec{1} & Mr \end{bmatrix}, \tag{3.10}$$

$$\bar{Q} = \begin{bmatrix} 0 & 0 \\ -Mr\vec{1} & Mr \end{bmatrix}. \tag{3.11}$$

To understand the motivation behind this model consider two cases.

Case 1: The buffer content is no more than b. Then the buffer content increases at rate r as long as the Z process is in the set $\{1, 2, .., n\}$ (we say it is "up"), and it decreases at rate 1 when the Z process is in state 0 (we say it is "down"). The Z process alternates between up and down periods. The down times are iid $\exp(\lambda)$ random variables, and the up times are iid with phase type distribution with parameters α and M.

Case 2: The buffer content is more than b. Then the buffer content increases at rate r as long as the Z process is up, and it decreases at rate 1 when the Z process is down. Once the Z process is down, it stays down until the buffer content drops below b and we switch to case 1 above.

Let $X_r(t)$ be the buffer content at time t in the fluid process described by the

parameters Q, $\bar{Q}$, R given above. A typical sample path of the $\{X_r(t), t \geq 0\}$ process is shown in Figure 3.1.

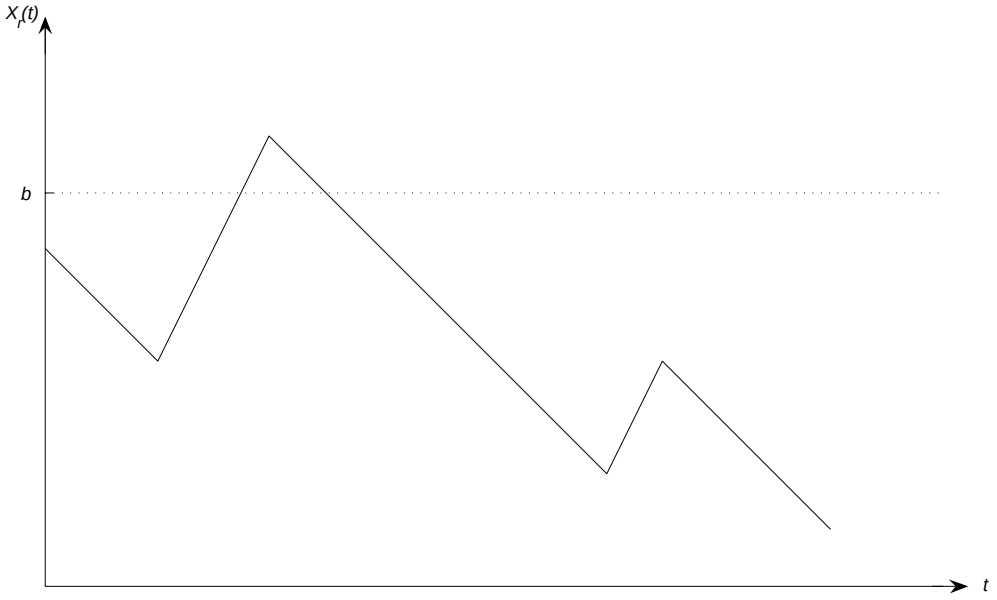

Figure 3.1: A Typical Sample Path of $X_r(t)$

Now define some special notation for this special case as follows:

$$T_r = \min\{t \geq 0 : X_r(t) = 0\},$$

$$\pi_i(x, r) = \mathrm{E}[T_r | Z(0) = i, X(0) = x],$$

$$\pi(x, r) = [\pi_0(x, r), \pi_1(x, r), \cdots, \pi_n(x, r)]^{\mathrm{T}},$$

$$\phi_i(s, x, r) = \mathrm{E}[e^{-sT_r} | Z(0) = i, X(0) = x],$$

$$\phi(s, x, r) = [\phi_0(s, x, r), \phi_1(s, x, r), \cdots, \phi_n(s, x, r)]^{\mathrm{T}},$$

$$A(s, r) = R^{-1}(sI - Q) = \begin{bmatrix} -\lambda - s & \lambda\alpha \\ M\vec{1} & -M + \frac{s}{r}I \end{bmatrix},$$

and

$$\bar{A}(s,r) = R^{-1}(sI - \bar{Q}) = \begin{bmatrix} -s & 0 \\ M\vec{1} & -M + \frac{s}{r}I \end{bmatrix}.$$

First we give an explicit formula for the mean first passage time $\pi(x,r)$ in the following theorem.

Theorem 10. *With R, Q and $\bar{Q}$ as specified in Equation (3.9), (3.10) and (3.11), respectively, the solution to Equation (3.3) and (3.4) is*

$$\pi(x,r) = \begin{cases} e^{A(0,r)x}(c - \int_0^x e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t), & \text{if } 0 \le x \le b, \\ (x-b)\vec{1} + \pi(b,r), & \text{if } x > b, \end{cases}$$

where

$$c = \begin{bmatrix} u_1 \\ \left[M\vec{1}, -M\right]e^{A(0,r)b} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ (1+1/r)\vec{1} + \left[M\vec{1}, -M\right]e^{A(0,r)b}\int_0^b e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t \end{bmatrix},$$

and u_1 is the first row of the identity matrix.

Proof. Substituting Q and $\bar{Q}$ in Equation (3.3), we get:

$$\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x} = -R^{-1}\vec{1} + A(0,r)\pi(x,r), \text{ if } 0 < x < b, \tag{3.12a}$$

$$\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x} = -R^{-1}\vec{1} + \bar{A}(0,r)\pi(x,r), \text{ if } x > b. \tag{3.12b}$$

It is well known (cf. Finizo and Ladas [14]) that the solution to Equation (3.12a) is

$$\pi(x,r) = e^{A(0,r)x}(c - \int_0^x e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t), \ 0 < x < b, \tag{3.13}$$

where c is some constant vector to be determined. Notice that for $x > b$:

$$\pi_i(x,r) = \mathrm{E}[T_r|Z(0) = i, X(0) = x] = \mathrm{E}[x - b + T_r|Z(0) = i, X(0) = b] = x - b + \pi_i(b,r).$$

Then

$$\frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x} = \frac{\mathrm{d}[(x-b)\vec{1} + \pi(b,r)]}{\mathrm{d}x} = \vec{1},\ x > b. \tag{3.14}$$

Consider the value $\lim_{x\downarrow b} \frac{\mathrm{d}\pi(x,r)}{\mathrm{d}x}$. From Equation (3.14) and (3.12b), use the fact $\pi(b-,r) = \pi(b+,r)$, we get:

$$-R^{-1}\vec{1} + \bar{A}(0,r)\pi(b-,r) = \vec{1},$$

which reduces to:

$$[M\vec{1}, -M]\pi(b-,r) = (1+1/r)\vec{1}.$$

Using Equation (3.13) for $\pi(b-,r)$ we get:

$$[M\vec{1}, -M]e^{A(0,r)b}(c - \int_0^b e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t) = (1+1/r)\vec{1}.$$

Since $\pi_0(0,r) = 0$, the first component of c is 0. Writing this condition as an additional row of the equation above, we get:

$$\begin{bmatrix} u_1 \\ \big[M\vec{1}, -M\big]e^{A(0,r)b} \end{bmatrix} c = \begin{bmatrix} 0 \\ (1+1/r)\vec{1} + \big[M\vec{1}, -M\big]e^{A(0,r)b}\int_0^b e^{-A(0,r)t}R^{-1}\vec{1}\mathrm{d}t \end{bmatrix}.$$

□

Remark: Since $A(0,r)$ is singular, it makes the computation of the integral $\int_0^x e^{-A(0,r)t}\mathrm{d}t$ tricky. There are several numerical methods for computing this integral. The method we use in our computations is based on the following observation. Since $\pi(x,r) = \frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}s}|_{s=0}$, we have

$$\int_0^x e^{-A(0,r)t}\mathrm{d}t = \lim_{s\to 0}\int_0^x e^{-A(s,r)t}\mathrm{d}t = \lim_{s\to 0} A^{-1}(s,r)(I - e^{-A(s,r)x}).$$

Next we derive an explicit formula for the LST of the first passage time $\phi(s,x,r)$.

First we need the following lemma.

Lemma 2. *The matrix $\bar{A}(s,r)$ has an eigenvalue $-s$ with geometric multiplicity 1 and the corresponding right eigenvector:*

$$\bar{v}_0(r) = \begin{bmatrix} 1 \\ (M - s(1+1/r)I)^{-1}M\vec{1} \end{bmatrix}.$$

Proof. It is easy to verify that

$$(-sI - \bar{A}(s,r))\bar{v}_0(r) = 0,$$

and the null space of $sI + \bar{A}(s,r)$ has dimension 1. □

The explicit formula for $\phi(s,x,r)$ is given in the following theorem.

Theorem 11. *With R, Q and $\bar{Q}$ as specified in Equation (3.9), (3.10) and (3.11), respectively, the solution to Equation (3.7) and (3.8) is*

$$\phi(s,x,r) = \begin{cases} e^{A(s,r)x}c & \text{if } 0 \le x \le b, \\ e^{-s(x-b)}e^{A(s,r)b}c & \text{if } x > b, \end{cases}$$

where

$$c = ke^{-A(s,r)b}\bar{v}_0(r),$$

k is the scalar such that $u_1 c = 1$.

Proof. Substituting Q and $\bar{Q}$ in Equation (3.3), we get:

$$\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x} = A(s,r)\phi(s,x,r), \text{ if } 0 < x < b, \tag{3.15a}$$

$$\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x} = \bar{A}(s,r)\phi(s,x,r), \text{ if } x > b. \tag{3.15b}$$

It is well known that the solution to Equation (3.15a) is

$$\phi(s,x,r) = e^{A(s,r)x}c,\ 0 < x < b, \tag{3.16}$$

where c is some constant vector to be determined. Notice that for $x > b$

$$\begin{aligned}\phi_i(s,x,r) &= \mathrm{E}[e^{-sT_r}|Z(0)=i, X(0)=x] \\ &= \mathrm{E}[e^{-s(x-b+T_r)}|Z(0)=i, X(0)=b] \\ &= e^{-s(x-b)}\phi_i(s,b,r).\end{aligned}$$

Then

$$\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x} = \frac{\mathrm{d}[e^{-s(x-b)}\phi(s,b,r)]}{\mathrm{d}x} = -se^{-s(x-b)}\phi(s,b,r),\ x > b. \tag{3.17}$$

Consider $\lim_{x\downarrow b}\frac{\mathrm{d}\phi(s,x,r)}{\mathrm{d}x}$. From Equation (3.17) and (3.15b), using the fact $\phi(s,b-,r) = \phi(s,b+,r)$, we get:

$$\bar{A}(s,r)\phi(s,b-,r) = -s\phi(s,b-,r).$$

Using Equation (3.16) for $\phi(s,b-,r)$ we get

$$\bar{A}(s,r)e^{A(s,r)b}c = -se^{A(s,r)b}c,$$

which implies that $e^{A(s,r)b}c$ is a right eigenvector of $\bar{A}(s,r)$ corresponding to the eigenvalue $-s$, i.e.:

$$e^{A(s,r)b}c = k\bar{v}_0(r),$$

due to Lemma 2. Finally, c is completely determined by the boundary condition that $\phi_0(s,0,r) = 1$. □

3.5 First Passage Time: the Balking Queueing Model

In this section, we compute the mean and LST of the first passage time for the balking queuing model. First we give the following construction which shows that the balking model is a limiting case of the fluid model we analyze in the previous section.

We construct the sample path of $\{W(t), t \geq 0\}$ in the balking model and the sample path of $\{X_r(t), t \geq 0\}$ in the fluid model on a common probability space as follows. Without loss of generality, assume the process $\{X_r(t), t \geq 0\}$ start in "down" time.

Let $\{U_i, i \geq 1\}$ be iid random variables with phase type distribution with parameter α and M; $\{D_i, i \geq 1\}$ be iid random variables with exponential distribution with parameter λ. We think of U_i as the service time of the i-th arriving customer (who may or may not balk) and D_i as the inter-arrival time between the i-th and $(i+1)$-st arriving customer. Then the sample path of the $\{W(t), t \geq 0\}$ process in the balking model is completely described by these two sequences and the parameter b. It is shown in Figure 3.2. Note that the second customer (arriving at time $D_1 + D_2$) finds the workload above b and hence balks.

Next we construct a sample path of the buffer content process $\{X_r(t), t \geq 0\}$ by using the same two sequences $\{U_i, i \geq 1\}$ and $\{D_i, i \geq 1\}$. It is shown in Figure 3.3. Here we use D_i as the i-th down time, and U_i/r as the i-th up time. Note that if the i-th down time finishes while the buffer content is above b, we do not use the i-th uptime at all, and immediately start the next down time. This is equivalent to having a null transition in the Z process from state 0 to 0. Such a transition occurs in Figure 3.3 at time $D_1 + \frac{U_1}{r} + D_2$.

Then, clearly,

$$\{X_r(t), t \geq 0\} \stackrel{a.s.}{\to} \{W(t), t \geq 0\} \text{ as } r \to \infty.$$

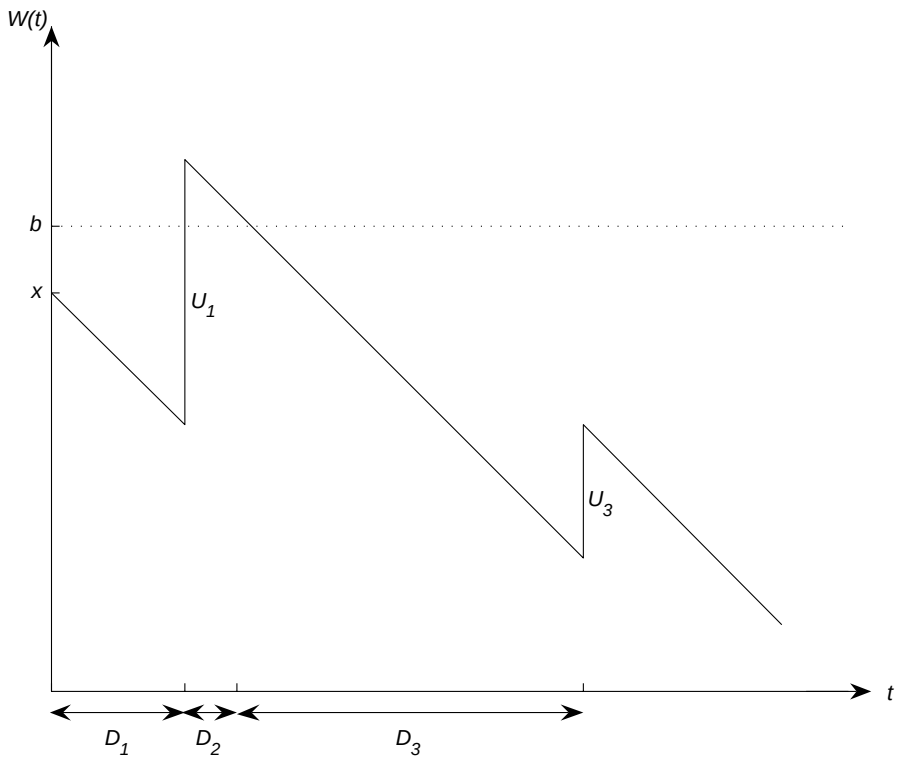

Figure 3.2: Construction of $W(t)$

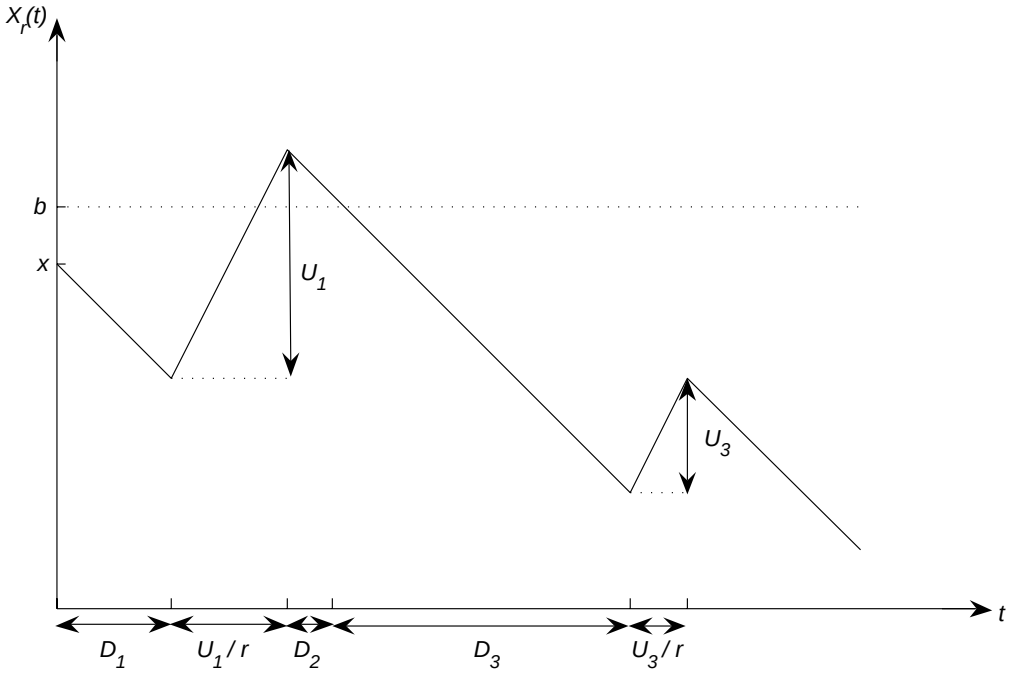

Figure 3.3: Construction of $X_r(t)$

Recall that $B = \min\{t \geq 0 : W(t) = 0\}$ and $T_r = \min\{t \geq 0 : X_r(t) = 0\}$. It follows from the preceding construction that

$$T_r \overset{a.s.}{\to} B$$

as $r \to \infty$.

Let

$$A(s) = \lim_{r\to\infty} A(s,r) = \begin{bmatrix} -\lambda - s & \lambda\alpha \\ M\vec{1} & -M \end{bmatrix},$$

$$\bar{A}(s) = \lim_{r\to\infty} \bar{A}(s,r) = \begin{bmatrix} -s & 0 \\ M\vec{1} & -M \end{bmatrix},$$

and

$$\bar{v}_0 = \lim_{r\to\infty} \bar{v}_0(r) = \begin{bmatrix} 1 \\ (M - sI)^{-1}M\vec{1} \end{bmatrix}.$$

It is clear that if we take the limit of the results given in Theorem 10 and 11, then the first component of the vector is the first passage time for the balking model. We summarize the results in Theorems 12 and 13. The proof is straightforward and hence is omitted.

Theorem 12. *The mean first passage time defined in Equation (3.1) is given by*

$$m(x) = \begin{cases} u_1 e^{A(0)x}(c + \int_0^x e^{-A(0)t}u_1^{\mathrm{T}}\mathrm{d}t), & \text{if } 0 \leq x \leq b, \\ (x - b) + m(b), & \text{if } x > b, \end{cases}$$

where

$$c = \begin{bmatrix} u_1 \\ \left[M\vec{1}, -M\right]e^{A(0)b} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vec{1} - \left[M\vec{1}, -M\right]e^{A(0)b}\int_0^b e^{-A(0)t}u_1^{\mathrm{T}}\mathrm{d}t \end{bmatrix},$$

and u_1 is the first row of the identity matrix.

Theorem 13. *The LST of the first passage time defined in Equation (3.2) is given by*

$$\psi(s,x) = \begin{cases} u_1 e^{A(s)x} c & \text{if } 0 \le x \le b, \\ u_1 e^{-s(x-b)} e^{A(s)b} c & \text{if } x > b, \end{cases}$$

where

$$c = k e^{-A(s)b} \bar{v}_0,$$

k is the scalar such that $u_1 c = 1$.

Remark: Differential equations similar to (3.3) and (3.7) also hold for other first passage times. For example, suppose $T = \min\{t \ge 0 : X(t) = 0 \text{ or } X(t) = b\}$. This case is actually easier since the constant vector in the solution to Equation (3.3) is completely determined by the boundary conditions $\pi_i(0) = 0, i \in \mathcal{S}_-$ and $\pi_i(b) = 0, i \in \mathcal{S}_+$. Similarly, the constant vector in the solution to Equation (3.7) is completely determined by the boundary conditions $\phi_i(s,0) = 1, i \in \mathcal{S}_-$ and $\phi_i(s,b) = 1, i \in \mathcal{S}_+$.

3.6 Special Case: Exponential Service Times

In this section, we illustrate the results of the previous section with exponential service time with rate μ, and verify the known results.

The exponential distribution is simply a phase type distribution with parameters:

$$M = [-\mu], \alpha = [1].$$

Thus we have

$$A(s) = \begin{bmatrix} -s-\lambda & \lambda \\ -\mu & \mu \end{bmatrix}, \bar{A}(s) = \begin{bmatrix} 0 & 0 \\ -\mu & \mu \end{bmatrix}, \bar{v}_0 = \begin{bmatrix} 1 \\ \mu/(\mu+s) \end{bmatrix}.$$

Using Theorem 12, after tedious algebra, we get the following result for the mean first passage times

$$m(x) = \begin{cases} \frac{1}{(\mu-\lambda)^2}[\mu(\mu-\lambda)x - \frac{\lambda^2}{\mu}e^{-(\mu-\lambda)(b-x)} + \frac{\lambda^2}{\mu}e^{-(\mu-\lambda)b}], & \text{if } 0 \le x \le b, \\ (x-b) + m(b), & \text{if } x > b, \end{cases} . \quad (3.18)$$

Equation 3.18 is equivalent to the formula given in Proposition 3.1 in Perry and Asmussen [38].

Using Theorem 13, after simplification, we get the following formula which is consistent with Theorem 3.1 in Perry and Asmussen [38]:

$$\psi(s,x) = \begin{cases} \frac{\gamma_1 e^{\theta_1 x} - \gamma_2 e^{\theta_2 x}}{\gamma_1 - \gamma_2} & \text{if } 0 \le x \le b, \\ \\ e^{-s(x-b)}\psi(s,b) & \text{if } x > b, \end{cases} \quad (3.19)$$

where

$$\theta_1 = \frac{(\mu - s - \lambda) + \sqrt{(s+\lambda-\mu)^2 + 4\mu s}}{2}, \quad (3.20)$$

$$\theta_2 = \frac{(\mu - s - \lambda) - \sqrt{(s+\lambda-\mu)^2 + 4\mu s}}{2}, \quad (3.21)$$

are the eigenvalues of $A(s)$ and

$$\gamma_i = (\mu - \theta_i - \frac{\lambda\mu}{s+\mu})e^{-\theta_i b}, \ i = 1, 2.$$

From Equation (3.18) and (3.19), by taking the limit $b \to \infty$, we obtain the mean and LST of the first passage time for the classical $M/M/1$ queueing model:

$$\lim_{b\to\infty} m(x) = \frac{\mu}{\mu - \lambda}x, \quad (3.22)$$

$$\lim_{b\to\infty} \psi(s,x) = e^{\theta_2 x}. \tag{3.23}$$

Equation (3.23) is exactly Equation (3.4) in Perry and Asmussen [38]. Inversion of the LST given by Equation (3.23) yields identical result given in Theorem 8 in Prabhu [42]. It is also worth noting that

$$-\left.\frac{\mathrm{d}e^{\theta_2 x}}{\mathrm{d}s}\right|_{s=0} = -\left. xe^{\theta_2 x}\frac{\mathrm{d}\theta_2}{\mathrm{d}s}\right|_{s=0} = \frac{\mu}{\mu-\lambda}x,$$

which matches with Equation (3.22).

3.7 Numerical Results

In this section, we illustrate our results numerically. We consider three different phase type service time distributions: exponential, Erlang and Hyper-exponential. The parameters are the same as those used in Section 2.6 on page 29. The balking threshold b is set to be 2.

In addition to the plots of the mean of B, we also include the plots of the cdf and pdf of B which are defined as follows,

$$F(t,x) = \Pr\{B \le t | W(0) = x\},\ t \ge x,$$

and

$$f(t,x) = \frac{\mathrm{d}F(t,x)}{\mathrm{d}t},\ t > x.$$

Recall that $\psi(s,x) = \mathrm{E}[e^{-sB}|W(0)=x]$, then $F(t,x)$ and $f(t,x)$ can be calculated as the inverse Laplace transform of $\psi(s,x)/s$ and $\psi(s,x)$ respectively. Note that the distribution of B has a mass at $t=x$, since

$$\Pr\{B = x | W(0) = x\} = \Pr\{\text{no arrival during } [0,x]\} = e^{-\lambda x}.$$

We make different plots by varying the service time distributions, ρ or x as summarized in Table 3.1.

Mean: $m(x)$	$\rho = 0.8$ $\rho = 1$ $\rho = 1.2$	exp, erlang, hyper	Figure 3.4 Figure 3.5 Figure 3.6
Mean: $m(x)$	$\rho = 0.8, 1, 1.2$	exp	Figure 3.7
Distribution: $F(t, x = 1, 2)$	$\rho = 0.8$ $\rho = 1$ $\rho = 1.2$	exp, erlang, hyper	Figure 3.8 Figure 3.9 Figure 3.10
Distribution: $F(t, x = 1)$	$\rho = 0.8, 1, 1.2$	erlang	Figure 3.11
Distribution: $F(t, x = 1, 2, 3)$	$\rho = 0.8$	erlang	Figure 3.12
Density: $f(t, x = 1, 2)$	$\rho = 0.8$ $\rho = 1$ $\rho = 1.2$	exp, erlang, hyper	Figure 3.13 Figure 3.14 Figure 3.15
Density: $f(t, x = 1)$	$\rho = 0.8, 1, 1.2$	erlang	Figure 3.16
Density: $f(t, x = 1, 2, 3)$	$\rho = 0.8$	erlang	Figure 3.17

Note:

1. Figure 3.8,3.9,3.10 3.13, 3.14 and 3.15 include two sets of three curves for x=1 and x=2.
2. $b = 2$.

Table 3.1: Plots Summary

From Theorem 12, we have $m(x) = (x - b) + m(b)$ when $x > b$. The linearity is illustrated in all mean plots. Also, from Theorem 13, we have $\psi(s, x) = e^{-s(x-b)}\psi(s, b)$ when $x > b$. This is reflected as a shift of the pdf curve as shown in Figure 3.17.

Surprisingly, it is worth noting that as the variance of the service time becomes smaller, the mean $m(x)$ becomes larger. Figure 3.4, 3.5 and 3.6 indicate that $m_{\text{hyper}}(x) < m_{\text{exp}}(x) < m_{\text{erlang}}(x), x > 0$. Moreover, in Figure 3.8,3.9 and 3.10, it is easy to see that $B_{\text{hyper}} <_{st.} B_{\text{exp}} <_{st.} B_{\text{erlang}}$.

Let $m(x, b)$ be the mean first passage time parameterized by the balking threshold b. Figure 3.18 numerically verifies the following obvious identity:

$$m(x^* + x, b) = m(x^*, b) + m(x, b - x^*),\ 0 \le x^* \le b,\ x \ge 0,$$

by using the exponential service time and $x^* = 1,\ b = 3$.

3.8 Concluding Remarks

In this chapter, we have developed an alternative method to study the first passage time problem for the $M/PH/1$ queues with wait-based balking via fluid models. The two models are connected by the construction illustrated in Section 3.5. We used elementary techniques to analyze the first passage time problem for the fluid model and obtained explicit solutions for the balking model. The method can also be applied to the dam model where service requirement is truncated if the complete admission of an arriving customer causes the workload to go beyond a given level.

Although the model we consider in this chapter is a single server one, it incorporates the right characteristics of the customer behavior and can be used in approximations of multi-server models.

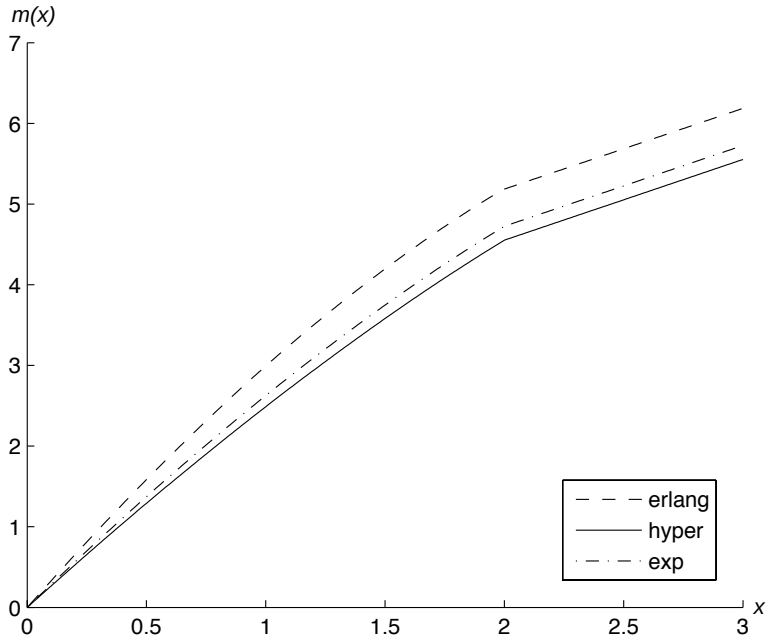

Figure 3.4: Mean Plot, $b = 2, \rho = 0.8$

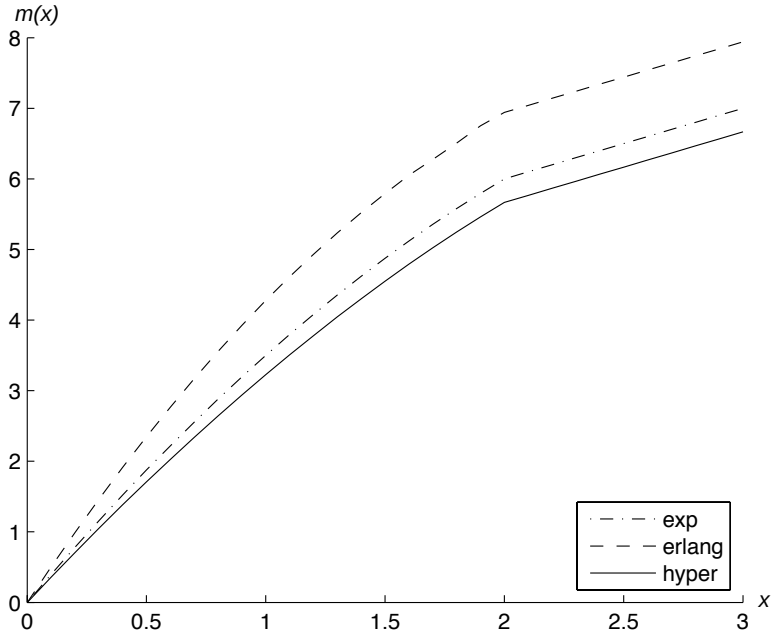

Figure 3.5: Mean Plot, $b = 2, \rho = 1$

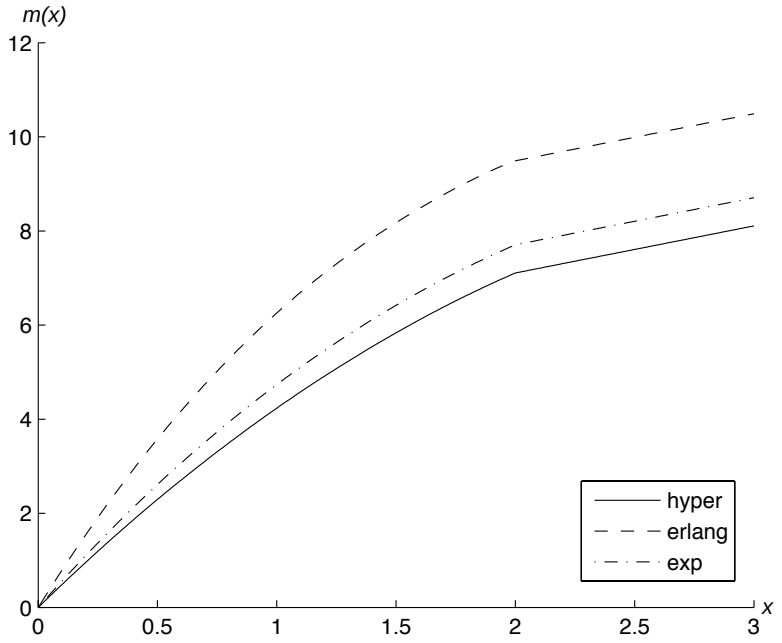

Figure 3.6: Mean Plot, $b = 2, \rho = 1.2$

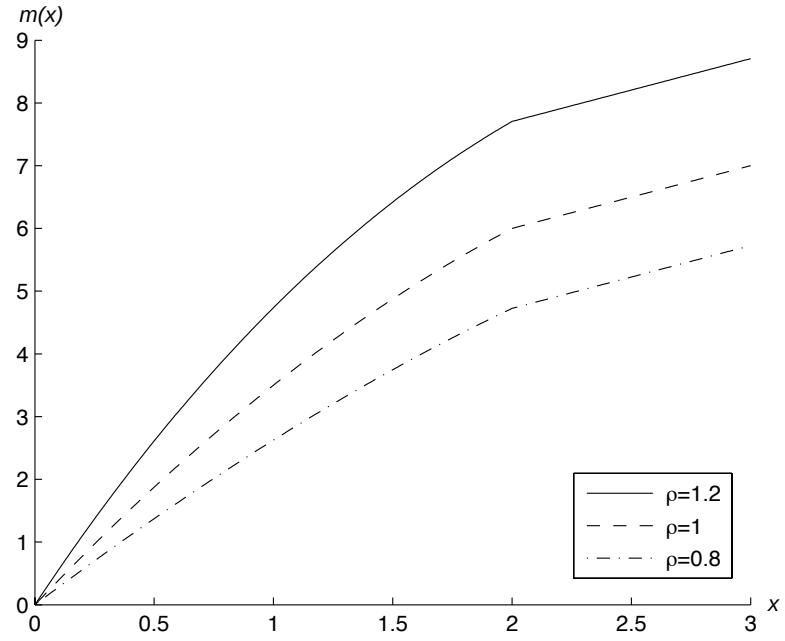

Figure 3.7: Mean Plot, Exponential Service Time, $b = 2$

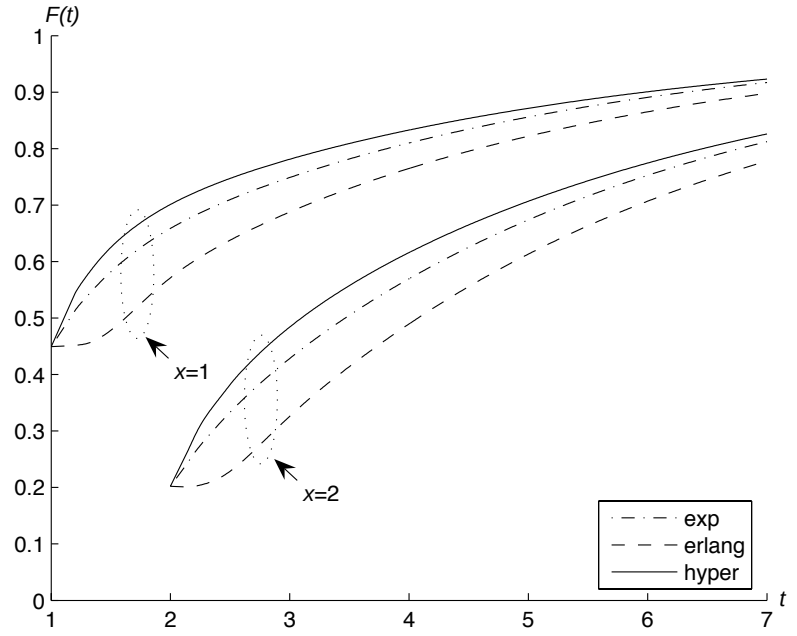

Figure 3.8: Distribution Plot, $b = 2, \rho = 0.8$ (two sets of three curves for x=1 and x=2)

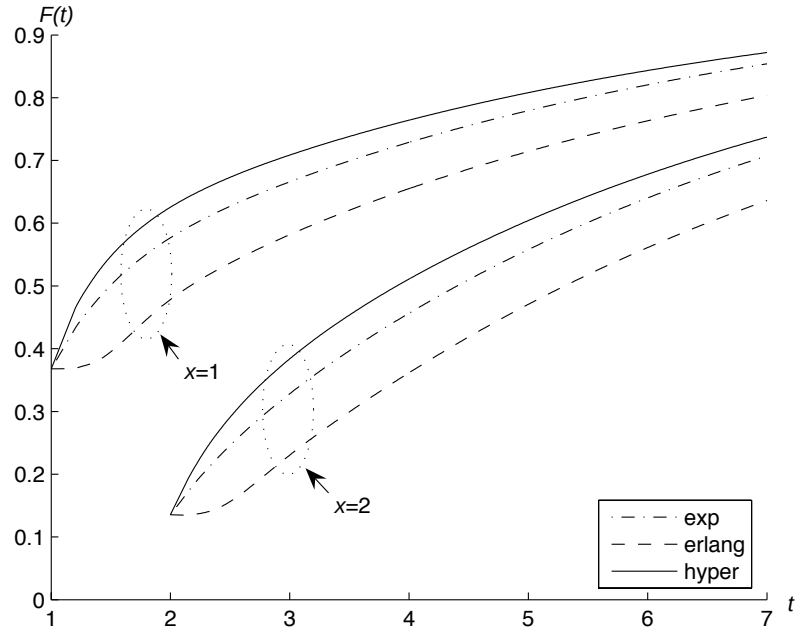

Figure 3.9: Distribution Plot, $b = 2, \rho = 1$ (two sets of three curves for x=1 and x=2)

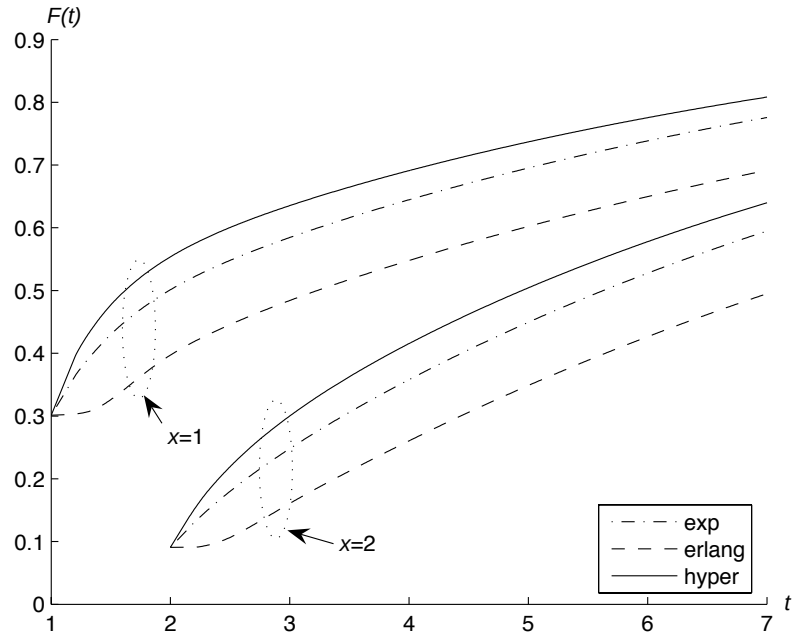

Figure 3.10: Distribution Plot, $b = 2, \rho = 1.2$ (two sets of three curves for x=1 and x=2)

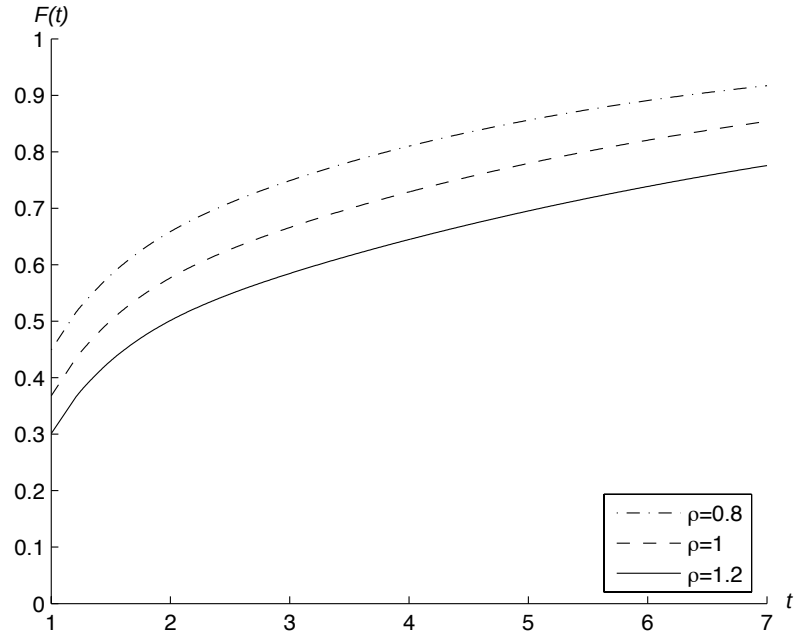

Figure 3.11: Distribution Plot, Erlang Service Time, $b = 2, x = 1$

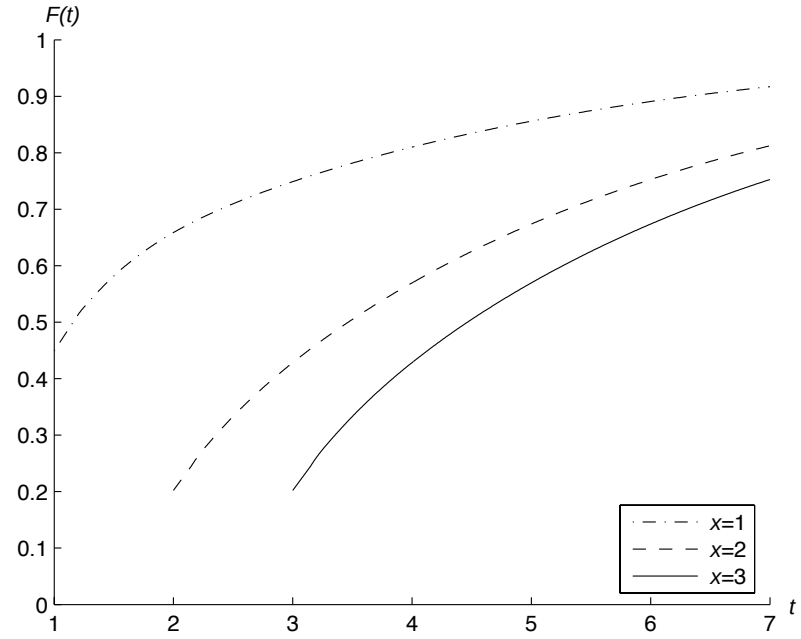

Figure 3.12: Distribution Plot, Erlang Service Time, $b = 2, \rho = 0.8$

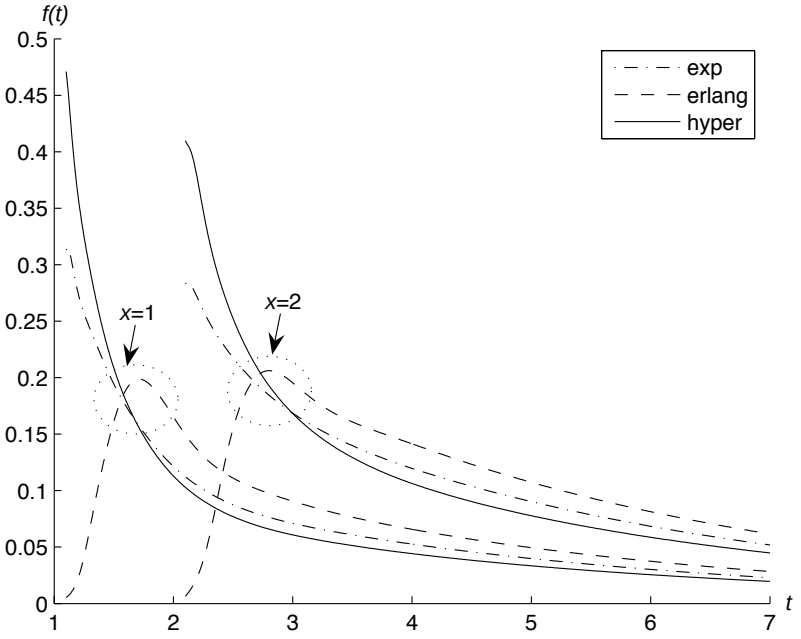

Figure 3.13: Density Plot, $b = 2, \rho = 0.8$, (two sets of three curves for x=1 and x=2)

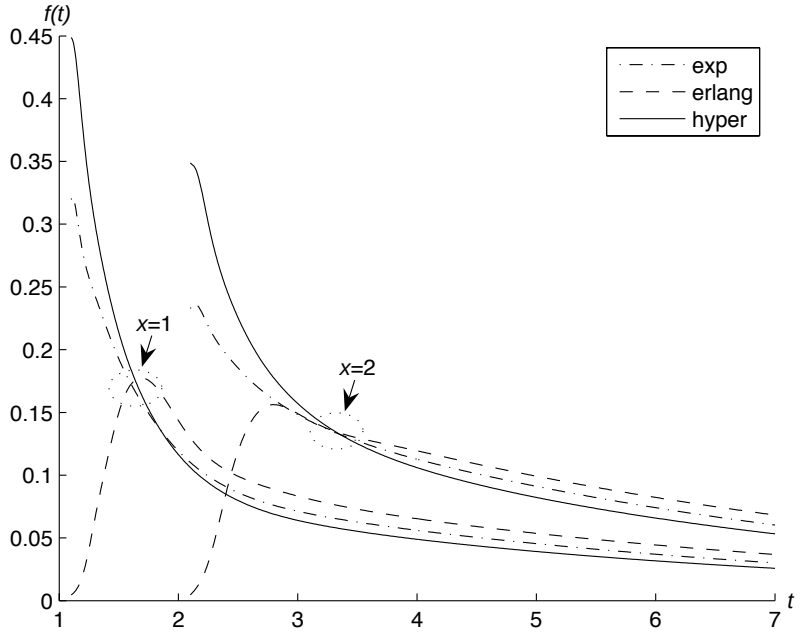

Figure 3.14: Density Plot, $b = 2, \rho = 1$ (two sets of three curves for x=1 and x=2)

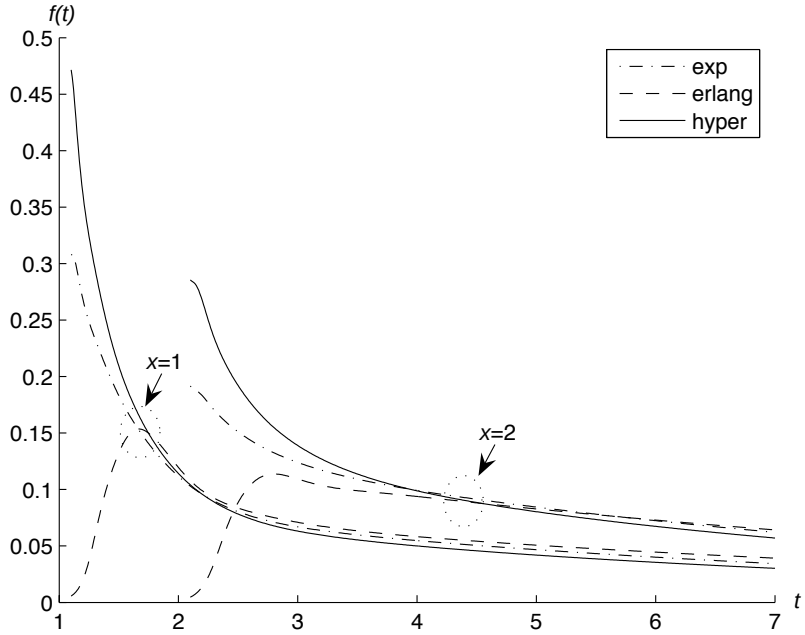

Figure 3.15: Density Plot, $b = 2, \rho = 1.2$ (two sets of three curves for x=1 and x=2)

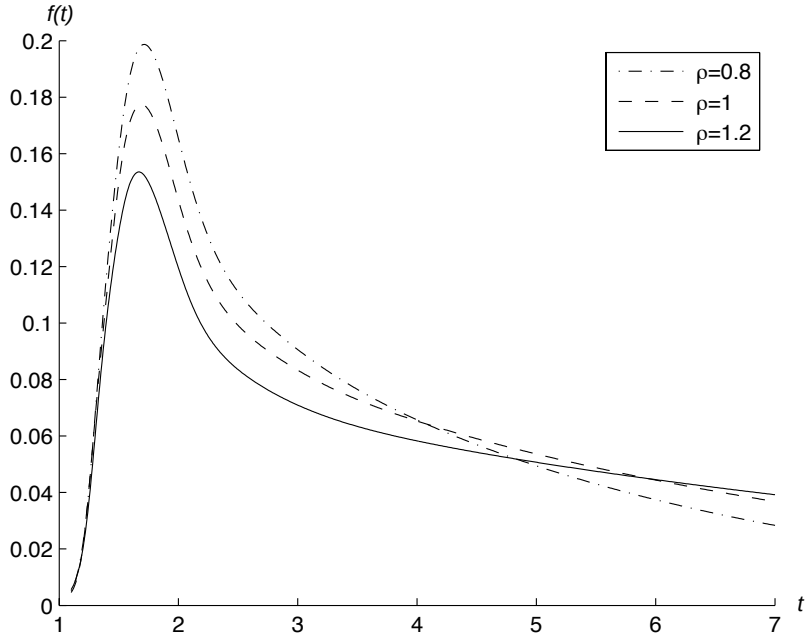

Figure 3.16: Density Plot, Erlang Service Time, $b = 2, x = 1$

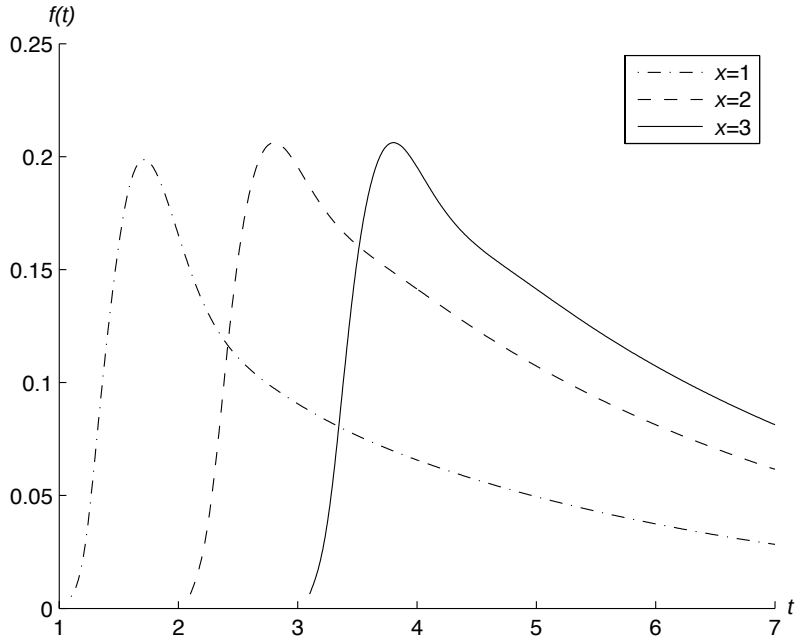

Figure 3.17: Density Plot, Erlang Service Time, $b = 2, \rho = 0.8$

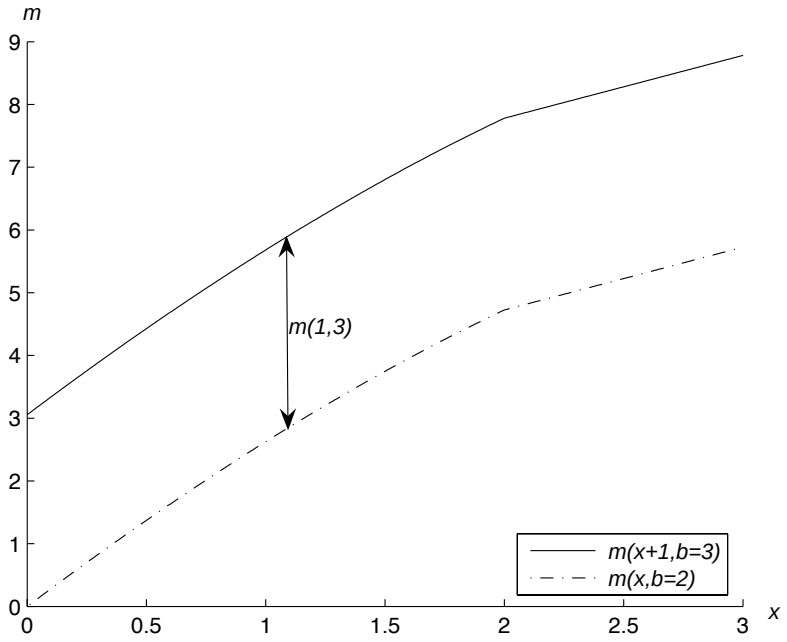

Figure 3.18: Mean Plot, Exponential Service Time, $\rho = 0.8$

Chapter 4

Balking and Reneging in $M/G/s$ Systems: Exact Analysis and Approximations

4.1 Introduction

In this chapter we consider the vqt process in an $M/G/s$ queue with impatient customers. Although the vqt process is introduced in Chapter 1 in the context of a wait-based balking queue, this process also plays an essential role in the analysis of the models with reneging customers. According to the definition of vqt, if the system has customers who eventually will renege without being served, then the service times for those customers are not included in the vqt. The model with reneging customers can be analyzed via a closely related wait-based balking model. Such an idea appears in Tijms [46] (pages 318–322). We shall discuss the connection between balking and reneging in details in Section 4.4. In this chapter we focus on the wait-based balking model and relate it to reneging behavior of impatient customers in terms of the steady state distribution of the vqt process.

An exact analysis of the multi-server queue is too complicated, or in many cases,

intractable. Therefore, we develop a system approximation for the multi-server system with impatient customers. The idea is to treat the s-server system as an $M/G/\infty$ system (or $M/G/s-1$ loss system) when some servers are idle and an $M/G/1$ system when all servers are busy. Since balking and/or reneging can only happen when all servers are busy, we can easily introduce the customer impatience to the $M/G/1$ system that approximates the original system during the period when all servers are busy. Using this idea we construct a single-server system whose operating characteristics approximates those of the $M/G/s$ queueing system with wait-based balking. The approximation is exact when $G = M$, $b = 0$ or $s = 1$. The exact analysis of the approximate system follows the same line as Chapter 2, where we solve the $s = 1$ case. We give both analytical results and numerical examples. We conduct simulation to assess the accuracy of the approximation.

We begin with analyzing the $G = M$ case in Section 4.2. In Section 4.3, we consider general service times and propose an approximate system. We conclude with numerical results (Section 4.6) following a discussion about the design of our simulation experiments (Section 4.5).

4.2 The $M/M/s$ Balking Model

In this section we consider the case where the service requirement has an exponential distribution with mean $1/\mu$. The arrival process is Poisson with rate λ. This is an $M/M/s$ FCFS system with balking based on vqt. At the arrival epoch, an arriving customer joins the queue if and only if he/she observes that the vqt is no more than a fixed amount b.

Let $N(t)$ be the number of customers in the system at time t. The definition of $W(t)$ implies that $W(t) = 0$ if and only if $N(t) \leq s-1$. If the $N = \{N(t),\ t \geq 0\}$ process undergoes a transition from state $s-1$ to s at time t, then there is a jump in the $W = \{W(t),\ t \geq 0\}$ process at the same time. This is illustrated in Figure 4.1

which shows the sample path of a 2-server system. Note that, in the sample path shown there, the customer who arrives at time T_3 balks. At time T_1 (and T_4), the number of customer in the system increases from 1 to 2. At the same time, the vqt jumps from 0 to a positive number. It is clear that the W process finishes a regenerative cycle from T_1 to T_4. It is easy to see that the size of the jumps are iid $\exp(s\mu)$.

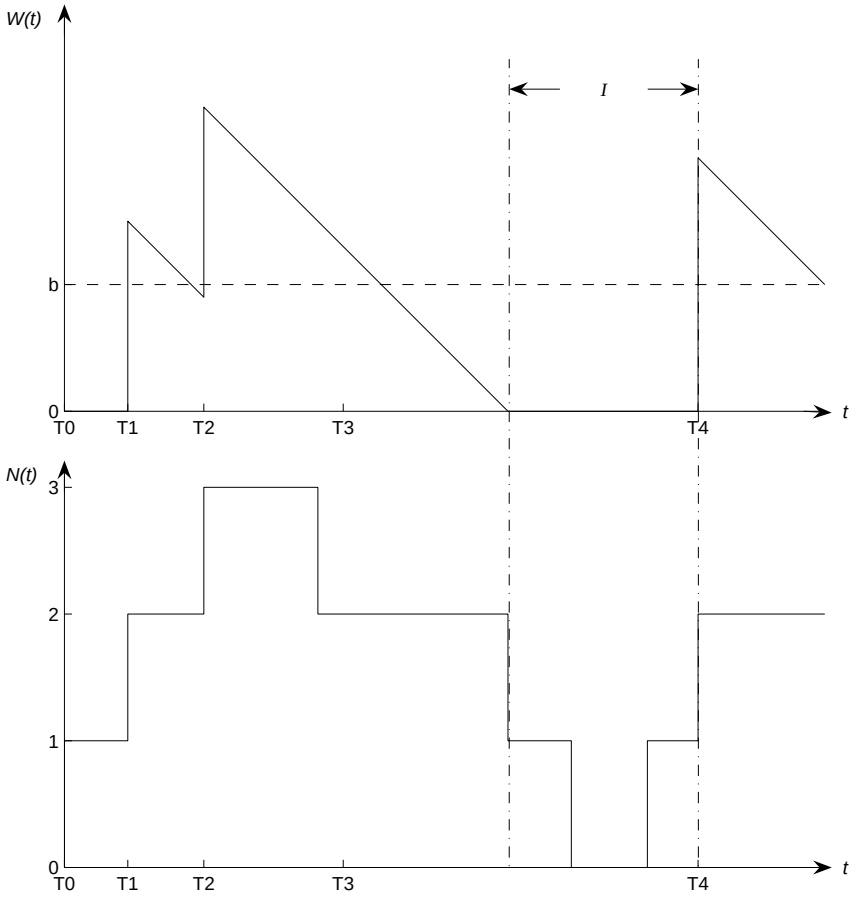

Figure 4.1: A sample path of $W(t)$ and $N(t)$

Let I be a generic random variable representing the idle period defined as the interval of time during which $W(t) = 0$. In other words, let t_1 be the service completion epoch such that $N(t_1-) = s$. Then $I = \min_{t>0}\{t : W(t_1 + t) > 0\}$ or $I = \min_{t>0}\{t : N(t_1 + t) = s\}$.

Theorem 14. *The expected length of I is given by*

$$\mathrm{E}(I) = \frac{1 - p_s}{s\mu p_s}, \tag{4.1}$$

where

$$p_s = \frac{(\lambda/\mu)^s}{s! \sum_{i=0}^{s} \frac{(\lambda/\mu)^i}{i!}}.$$

Proof. Consider a standard $M/M/s/s$ system with arrival rate λ and mean service times $1/\mu$. Obviously, the time between two consecutive periods when the system is full has the same distribution as I. The expected length of each system-full period is clearly $1/(s\mu)$. From the theory of alternating renewal process (ARP), we get

$$\frac{1/(s\mu)}{1/(s\mu) + \mathrm{E}(I)} = p_s,$$

where p_s is the probability that the $M/M/s/s$ system is full (cf. [29]). The identity in the theorem follows. □

Consider the regenerative cycle from T_1 to T_4 as shown in Figure 4.1. The cycle consists of a busy period (where $W(t) > 0$) and an idle period (where $W(t) = 0$). It is easy to see that the evolution of the W process during the busy period is stochastically identical to that in a single-server balking system discussed in Chapter 2 with arrival rate of λ and iid $\exp(s\mu)$ service times. This observation motivates the following single-server system with iid $\exp(s\mu)$ service times. Let $\tilde{W}(t)$ be the vqt and $\tilde{N}(t)$ be the number of customers at time t in this system. The same balking rule applies, i.e., a customer arriving at time t enters if and only if $\tilde{W}(t-) \le b$. Customers arrive as a Poisson process with arrival rate $\tilde{\lambda}(t)$ depending on $\tilde{N}(t)$ as follows:

$$\tilde{\lambda}(t) = \begin{cases} \gamma & \text{if} \quad \tilde{N}(t) = 0 \\ \lambda & \text{otherwise} \end{cases},$$

where γ is defined to be $1/\mathrm{E}(I)$.

Denote the limiting cdf of the W process and the $\tilde{W}$ process as follows.

$$F(x) = \lim_{t\to\infty} \Pr\{W(t) \le x\}, \quad x \ge 0;$$

$$\tilde{F}(x) = \lim_{t\to\infty} \Pr\{\tilde{W}(t) \le x\}, \quad x \ge 0.$$

Let

$$F(0) = c, \quad \tilde{F}(0) = \tilde{c}.$$

Let

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}, \quad \tilde{f}(x) = \frac{\mathrm{d}\tilde{F}(x)}{\mathrm{d}x}, \quad x > 0,$$

be the pdf.

We apply the same method used in Chapter 2 and give the results in Theorem 15 and 16, omitting the proofs. Theorem 15 gives the balance equation and normalizing equation satisfied by $\tilde{f}(x)$. Theorem 16 gives the expression of $\tilde{f}(x)$ explicitly by solving the equations.

Theorem 15. *The equilibrium pdf $\tilde{f}(x)$ of the $\tilde{W}$ process satisfies:*

$$\tilde{f}(x) = \lambda \int_0^{x\wedge b} \tilde{f}(u) e^{-s\mu(x-u)} \mathrm{d}u + \tilde{c}\gamma e^{-s\mu x}, \tag{4.2a}$$

$$\int_0^{\infty} \tilde{f}(x)\mathrm{d}x + \tilde{c} = 1, \tag{4.2b}$$

where $x \wedge b = \min(x, b)$.

Let $\rho = \frac{\lambda}{s\mu}$ be the traffic intensity.

Theorem 16. *The equilibrium pdf of the $\tilde{W}$ process is:*

$$\tilde{f}(x) = \begin{cases} \tilde{c}\gamma e^{-(s\mu-\lambda)x} & \text{if } 0 < x < b \\ \tilde{c}\gamma e^{\lambda b} e^{-s\mu x} & \text{if } x \ge b \end{cases}. \tag{4.3}$$

where

$$\tilde{c} = \begin{cases} [\gamma(\frac{1}{s\mu-\lambda} - e^{-(s\mu-\lambda)b}\frac{\lambda}{(s\mu-\lambda)s\mu}) + 1]^{-1} & \text{if } \rho \neq 1 \\ \frac{\lambda}{\lambda+\gamma+\lambda\gamma b} & \text{if } \rho = 1 \end{cases} . \tag{4.4}$$

The following theorem gives the limiting distribution of the W process via the single-server system we construct.

Theorem 17. *The W process and $\tilde{W}$ process have same limiting distribution, i.e.,*

$$F(x) = \tilde{F}(x), \quad x \geq 0.$$

Proof. Let $B(t) = 1$ if $W(t) > 0$ and $B(t) = 0$ if $W(t) = 0$. Then $B = \{B(t),\ t \geq 0\}$ is an ARP. Define $\tilde{B}(t)$ associated with $\tilde{W}(t)$ in the same fashion, then $\tilde{B} = \{\tilde{B}(t),\ t \geq 0\}$ is also an ARP. Since the expected up and down times in the B process and the $\tilde{B}$ process are the same, we get

$$\lim_{t\to\infty} \Pr\{B(t) = i\} = \lim_{t\to\infty} \Pr\{\tilde{B}(t) = i\}, \ \ i = 0, 1.$$

It is easy to see that the sample paths of the W process and the $\tilde{W}$ process over the busy periods are stochastically identical. Hence we get

$$\lim_{t\to\infty} \Pr\{W(t) \leq x | B(t) = 1\} = \lim_{t\to\infty} \Pr\{\tilde{W}(t) \leq x | \tilde{B}(t) = 1\}.$$

Now

$$\begin{aligned} F(x) &= \lim_{t\to\infty} \Pr\{W(t) \leq x\} \\ &= \lim_{t\to\infty} \Pr\{W(t) \leq x | B(t) = 1\} \Pr\{B(t) = 1\} + \lim_{t\to\infty} \Pr\{B(t) = 0\} \\ &= \lim_{t\to\infty} \Pr\{\tilde{W}(t) \leq x | \tilde{B}(t) = 1\} \Pr\{\tilde{B}(t) = 1\} + \lim_{t\to\infty} \Pr\{\tilde{B}(t) = 0\} \\ &= \lim_{t\to\infty} \Pr\{\tilde{W}(t) \leq x\} \\ &= \tilde{F}(x) \end{aligned}$$

This proves the theorem. □

Other performance measures of interest in the $M/M/s$ balking system are computed directly based on the results above. Let r be the probability that a customer balks or the balking rate. Then

$$r = \int_b^\infty f(x)\mathrm{d}x = c\gamma \frac{e^{-(s\mu-\lambda)b}}{s\mu}.$$

Define the queueing time of any arriving customer to be the time from the arrival epoch to the service starting epoch if he/she joins and 0 if he/she balks. Let w be the long-run average queueing time for all customers. Then

$$w = \int_0^b xf(x)\mathrm{d}x = \begin{cases} \frac{c\gamma[1-(s\mu-\lambda)be^{-(s\mu-\lambda)b}-e^{-(s\mu-\lambda)b}]}{(s\mu-\lambda)^2} & \text{if } \rho \neq 1 \\ \frac{c\gamma b^2}{2} & \text{if } \rho = 1 \end{cases}.$$

It is clear that the long-run average queueing time or the expected queueing time for the entering customers is $w' = w/(1-r)$. It can be verified by straightforward algebra that the results given in this section are consistent with the corresponding ones in [17] and [7] [1]. However, unlike in [7], we do not need to assume that $\rho < 1$, and our results are more explicit than those in [17].

4.3 The $M/G/s$ Balking Model

In this section, we extend the $M/M/s$ balking model in Section 4.2 to an $M/G/s$ balking model. All settings for the $M/M/s$ balking model are unchanged except that we assume the service times are iid with a general distribution with mean $1/\mu$ and ccdf $G(x)$. We use the subscript G in our notations for the general service time case. The definitions correspond to those for the exponential case and are omitted. In order

[1] The formula for $p_0^{(\tau)}$ in [7] consists a term with inverted sign, which we believe is a typo.

to follow the analysis in Section 4.2, we need $\mathrm{E}(I_G)$ and the distribution of the size of the jumps in the W_G process.

To compute $\mathrm{E}(I_G)$ we consider a standard $M/G/s/s$ system with arrival rate λ and mean service time $1/\mu$. At each departure epoch of a customer who leaves behind $s-1$ customers in the system, the remaining service time in the busy servers may not have the same joint distribution as that of the $M/G/s$ system with balking. Ignoring this fact, we use the expected length of the time between two consecutive periods when the $M/G/s/s$ system is full as an approximation of $\mathrm{E}(I_G)$. We know that in equilibrium, an arriving customer to the $M/G/s/s$ system with $s-1$ busy servers sees the remaining service times in the busy servers as having iid distribution with ccdf $G_e(x)$, which is the associated complementary equilibrium distribution of $G(x)$ defined by

$$G_e(x) = \mu \int_x^\infty G(u)\mathrm{d}u.$$

Then the length of the period during which the $M/G/s/s$ system is full is

$$\min\{R_1, R_2, \cdots, R_{s-1}, S\},$$

where $\{R_i, \quad i = 1, 2, \cdots, s-1\}$ are iid random variables with ccdf $G_e(x)$. Notice that

$$\mathrm{d}G_e(x) = -\mu G(x)\mathrm{d}x,$$

then the expected length of this period is

$$\int_0^\infty G_e^{s-1}(x)G(x)\mathrm{d}x = \frac{1}{s\mu}. \tag{4.5}$$

Using the same method in the proof of Theorem 14, we get that the expected

duration of the interval during which the $M/G/s/s$ system is *not* full is given by

$$\frac{1-p_s}{s\mu p_s}. \tag{4.6}$$

This is the same as in the $M/M/s/s$ system. We use this expression as an approximation for $\mathrm{E}(I_G)$.

The distribution of the size of the jumps is more complicated in the system with general service times. Suppose the k-th jump in the W_G process occurs at time T_k. Let J_k be the size of this jump. Unfortunately, $\{J_i, i = 1, 2, \cdots\}$ are neither independent nor identically distributed in general and this makes the model intractable. In the next paragraph we explain the source of this intractability and it can be skipped in first reading without affecting the flow of the material.

The jump size J_k in the W_G process at time T_k is the minimum of this customer's service time and the remaining service times of all other customers in service at time $T_k + W_G(T_k)$ (when this customer begins the service). Thus the distribution of the jump sizes are not even identical in general. Equivalently, let's regard the $M/G/s$ system as s parallel single server queues that operate as follows. Denote the workload at time t in the i-th queue as $W_i'(t)$, $i = 1, 2, \cdots, s$, and let $W'(t) = (W_1'(t), W_2'(t), \cdots, W_s'(t))$ (cf [12]). Every entering customer is routed to the queue with the least workload. Then

$$W_G(t) = \min\{W_1'(t), W_2'(t), \cdots, W_s'(t)\}. \tag{4.7}$$

Suppose the customer who arrives at time T_k with service time S is routed to the i-th server, which has the least workload, then

$$\begin{aligned} J_k &= W_G(T_k+) - W_G(T_k-) \\ &= \min\{W_1'(T_k-), \cdots, W_i'(T_k-) + S, \cdots, W_s'(T_k-)\} - W_i'(T_k-). \end{aligned} \tag{4.8}$$

Clearly the distribution of J_k is determined by the distribution of $W'(T_k)$ and the distribution of S. The dependence of J_k and $W'(T_k)$ causes great complexity in the analysis of the model and makes it intractable.

As an approximation, we assume that $\{J_i : W(T_i-) > 0\}$ are iid with $\dot{J}$ being the generic jump size and $\{J_i : W(T_i-) = 0\}$ are iid with $\ddot{J}$ being the corresponding generic jump size. One principle of choosing the distribution for $\dot{J}$ and $\ddot{J}$ is to preserve the traffic intensity, i.e., $\rho = \lambda/(s\mu)$. That is, keep the mean of $\dot{J}$ and $\ddot{J}$ to be $1/(s\mu)$. We assume $\mathrm{E}(\dot{J}) = \mathrm{E}(\ddot{J}) = 1/(s\mu)$ in the rest of this chapter. We consider two possibilities. The first choice is $\bar{S} = S/s$. It is easy to see that $\mathrm{E}(\bar{S}) = 1/(s\mu)$ in this case. The second choice is $\hat{S} = \min\{R_1, R_2, \cdots, R_{s-1}, S\}$. This is motivated by the renewal-theoretic result that in steady state the remaining services times in the busy servers are independent random variables with common ccdf G_e (cf. [44], page 161). From Equation (4.6), we see that $\mathrm{E}(\hat{S}) = 1/(s\mu)$.

Analogous to the exponential case, we consider the $\tilde{W}_G$ process of the following single-server system. Customers arrive according to a Poisson process with arrival rate $\tilde{\lambda}(t)$ depending on $\tilde{N_G}(t)$ as follows:

$$\tilde{\lambda}(t) = \begin{cases} \gamma & \text{if} \quad \tilde{N_G}(t) = 0 \\ \lambda & \text{otherwise} \end{cases},$$

where γ is defined to be $s\mu p_s/(1-p_s)$, which is the approximation for $1/\mathrm{E}(I_G)$. Let $G_{\dot{J}}(x)$ and $G_{\ddot{J}}(x)$ be the ccdf of $\dot{J}$ and $\ddot{J}$ respectively. Service times of the customers who enter a non-empty system are iid with common ccdf $G_{\dot{J}}(x)$. Service times of the customers who enter an empty system are iid with common ccdf $G_{\ddot{J}}(x)$. A customer arriving at time t enters if and only if $\tilde{W}_G(t-) \leq b$. We use the expression in (4.6) as approximation of $\mathrm{E}(I_G)$. We approximate the distributions of $\{J_i\}$ by $\dot{J}$ and $\ddot{J}$. Moreover, the conditions for Theorem 17 do not hold in general. Therefore, the vqt process of the single server model we construct approximates that of the $M/G/s$

balking model, i.e., $F_G(x) \approx \tilde{F}_G(x),\ x \geq 0$. It is worth noting that the approximation is exact in the following three cases: a) the service times are exponential; or b) the balking threshold b is zero (when the system reduces to an $M/G/s/s$ system) and $\ddot{J} = \hat{S}$; or c) $s = 1$.

The following theorem is the general service time version of Theorem 15. It distinguishes the two appearances of the jump size distribution in the balance equation. The rest discussion in this section follows the same line as Chapter 2 (with modifications and new formulas).

Theorem 18. *The steady state pdf $\tilde{f}_G(x)$ of the $\tilde{W}_G$ process satisfies:*

$$\tilde{f}_G(x) = \lambda \int_0^{x \wedge b} \tilde{f}_G(u) G_{\dot{J}}(x-u) \mathrm{d}u + \tilde{c}_G \gamma G_{\ddot{J}}(x), \tag{4.9a}$$

$$\int_0^{\infty} \tilde{f}_G(x) \mathrm{d}x + \tilde{c}_G = 1, \tag{4.9b}$$

where $x \wedge b = \min(x, b)$.

Notice that the first term in the right hand side of Equation (4.9a) is just the convolution of $\tilde{f}_G(x)$ and $G_{\dot{J}}(x)$ multiplied by λ, when $x \wedge b$ is replaced by x. Let $f_1(x)$ be the solution to

$$f_1(x) = \lambda \int_0^{x} f_1(u) G_{\dot{J}}(x-u) \mathrm{d}u + G_{\ddot{J}}(x), \quad x \geq 0. \tag{4.10}$$

Let

$$f_2(x) = \lambda \int_0^{b} f_1(u) G_{\dot{J}}(x-u) \mathrm{d}u + G_{\ddot{J}}(x), \quad x \geq b \tag{4.11}$$

The solution to Equation (4.9) is given in the following theorem.

Theorem 19. *The solution to* (4.9) *is:*

$$\tilde{f}_G(x) = \begin{cases} \tilde{c}_G \gamma f_1(x) & \text{if } x < b \\ \tilde{c}_G \gamma f_2(x) & \text{if } x \geq b \end{cases} \tag{4.12}$$

where

$$\tilde{c}_G = \left[\gamma \int_0^b f_1(x)\mathrm{d}x + \gamma \int_b^\infty f_2(x)\mathrm{d}x + 1\right]^{-1}. \tag{4.13}$$

Proof. The solution is easy to verify by substitution. □

From the above theorem, it is clear that a possible procedure to obtain $\tilde{f}_G(x)$ is to find $f_1(x)$ first, then compute $f_2(x)$ by using Equation (4.11). By the normalizing equation (4.9b), after computing the integral, we are able to compute $\tilde{c}_G$. This completes the computation of $\tilde{f}_G(x)$. Obviously, one main step is to solve Equation (4.10) for $f_1(x)$. One method is to use LT.

Let $G_j^*(\xi)$ and $G_{\ddot{J}}^*(\xi)$ be the LT of $G_j(x)$ and $G_{\ddot{J}}(x)$ respectively, i.e.,

$$G_j^*(\xi) = \int_0^\infty e^{-\xi x} G_j(x)\mathrm{d}x,$$

$$G_{\ddot{J}}^*(\xi) = \int_0^\infty e^{-\xi x} G_{\ddot{J}}(x)\mathrm{d}x.$$

From (4.10), we get the LT of $f_1(x)$ (assuming its existence):

$$f_1^*(\xi) = \frac{G_{\ddot{J}}^*(\xi)}{1 - \lambda G_j^*(\xi)}. \tag{4.14}$$

To continue our procedure, we need the inverse LT of $f_1^*(\xi)$. A closed form inversion is possible if $f_1^*(\xi)$ is rational.

We can instantly obtain two interesting results from the analysis above when $b \to 0$ and $b \to \infty$. The first case is $b \to 0$. In this case, the system reduces to a normal $M/G/s/s$ model. Our approximation is exact if $\ddot{J} = \hat{S}$. From Theorem 18, as $b \to 0$,

$$\tilde{f}_G(x) \to \tilde{c}_G \gamma G_{\ddot{J}}(x), \quad x \geq 0,$$

$$\tilde{c}_G \to 1 - p_s.$$

Next we compute $\tilde{w}_G$, the long-run average queueing time for all customers and $\tilde{c}_G$,

as $b \to \infty$. The LT is convenient in this case. Using

$$\int_0^\infty f_1(x)\mathrm{d}x = f_1^*(0),$$

$$G_{\dot{J}}^*(0) = \mathrm{E}(\dot{J}), \quad G_{\ddot{J}}^*(0) = \mathrm{E}(\ddot{J}),$$

$$\tilde{w}_G = -\left.\frac{\mathrm{d}\tilde{f}_G^*(\xi)}{\mathrm{d}\xi}\right|_{\xi=0},$$

Theorem 19 and Equation (4.14), we get

$$\tilde{c}_G \to \frac{1-\rho}{\nu+1-\rho}, \tag{4.15}$$

$$\tilde{w}_G \to \frac{\gamma[(1-\rho)\mathrm{E}(\ddot{J}^2) + \rho\mathrm{E}(\dot{J}^2)]}{2(1-\rho)(\nu+1-\rho)}, \tag{4.16}$$

where $\nu = \gamma/(s\mu) = p_s/(1-p_s)$.

When $b \to \infty$, the W_G process becomes the vqt process in a normal $M/G/s$ system and $r_G \to 0$. For several non-exponential distributions (e.g. Erlang-k) of service time, exact table of $E(W_G)$ (and other performance measures) are available and can be compared with $E(\tilde{W}_G)$ to assess the accuracy of our approximation.

Depending on the expression of the service time distribution and the choice of the distributions of $\dot{J}$ and $\ddot{J}$, the solution to Equation (4.9) can be obtained, in most cases, by numerical methods. We have two options for $\dot{J}$ or $\ddot{J}$, namely $\bar{S} = S/s$ and $\hat{S} = \min\{R_1, R_2, \cdots, R_{s-1}, S\}$. Let

$$G_{\bar{S}}(x) = \Pr\{\bar{S} \geq x\} = G(sx),$$

$$G_{\bar{S}}^*(\xi) = \int_0^\infty e^{-\xi x} G_{\bar{S}}(x)\mathrm{d}x = \frac{1}{s}G^*(\xi/s),$$

$$G_{\hat{S}}(x) = \Pr\{\hat{S} \geq x\} = G_e^{s-1}(x)G(x),$$

$$G_{\hat{S}}^*(\xi) = \int_0^\infty e^{-\xi x} G_{\hat{S}}(x)\mathrm{d}x = -\frac{1}{\mu}\int_0^\infty e^{-\xi x} G_e^{s-1}(x)\mathrm{d}[G_e(x)],$$

where $G^*(\xi)$ is the LT for $G(x)$. The expression of $G_{\bar{S}}(x)$ and $G^*_{\bar{S}}(\xi)$ are very easy to obtained once $G(x)$ and $G^*(\xi)$ are specified. But it is hard to compute $G_{\hat{S}}(x)$ and $G^*_{\hat{S}}(\xi)$. Computing $G^*_{\hat{S}}(\xi)$ is hard even for phase-type (except exponential) service times. Since all jumps caused by the customers who see $s-1$ busy servers upon arrival are iid distributed as $\hat{S}$, we introduce the complexity with the hope of improving accuracy. In the following sections, we consider two possibilities in choosing $\dot{J}$ and $\ddot{J}$ and define *Approximation I* and *II* correspondingly.[2]

4.3.1 Approximation I: $\dot{J} = \ddot{J} = \bar{S}$.

The equilibrium distribution of the W process is approximated by the solution to the following system of equations:

$$\tilde{f}_G(x) = \lambda \int_0^{x \wedge b} \tilde{f}_G(u) G_{\bar{S}}(x-u) \mathrm{d}u + \tilde{c}_G \gamma G_{\bar{S}}(x), \tag{4.17a}$$

$$\int_0^{\infty} \tilde{f}_G(x) \mathrm{d}x + \tilde{c}_G = 1, \tag{4.17b}$$

When $b \to \infty$, this approximation becomes a classical $M/G/s$ system approximation which appears in [33]. It tells us that the system behaves as an $M/G/s/s$ system when the vqt is zero. For vqt greater than zero, the system behaves like a busy $M/G/1$ system with service time S/s. See [44], [36], [37] and [19] for details. Equation (4.16) becomes

$$\tilde{w}_G \to \frac{\gamma \mathrm{E}(S^2)}{2s^2(1-\rho)(\nu+1-\rho)}.$$

The most general case where explicit closed form solution to (4.17) is available is the phase-type service time.

[2]The approximation obtained by using $\dot{J} = \ddot{J} = \hat{S}$ is omitted since it is seen to be inferior according to the results of our numerical experiments.

Consider the phase type service time, i.e.,

$$G(x) = \alpha e^{Mx}\vec{1}, \tag{4.18}$$

where the n-dimensional row vector α and the matrix M are the parameter of the phase type distribution, $\vec{1}$ is a column vector with all component be 1. Then,

$$G_{\bar{S}}(x) = \alpha e^{sMx}\vec{1}.$$

The results in Chapter 2 can be easily adapted to the case we consider here. The following is the balance equation used in Chapter 2, Equation 2.3a.

$$f(x) = \lambda \int_0^{x\wedge b} f(u)G(x-u)du + c\lambda G(x),$$

where $G(x) = \alpha e^{Mx}\vec{1}$. The balance equation we have for $\tilde{f}(x)$ in this case is

$$\tilde{f}_G(x) = \lambda \int_0^{x\wedge b} \tilde{f}_G(u)G_{\bar{S}}(x-u)\mathrm{d}u + \tilde{c}_G\gamma G_{\bar{S}}(x),$$

where $G_{\bar{S}}(x) = \alpha e^{sMx}\vec{1}$. We list the equations along to highlight the similarity between them. The parameter M is scaled to sM. The coefficient $c\lambda$ is changed to $\tilde{c}_G\gamma$. We obtain the solution to Equation 4.17 directly from Theorem 4 by scaling and substitution and give it in Theorem 20, skipping the proof.

First we redefined the notations used in the statement of Theorem 4 as follows.

Let $a_0, a_1, \cdots, a_n$ be the coefficients of the characteristic polynomial of the matrix sM, i.e. :

$$\det(xI - sM) = \sum_{j=0}^{n} a_j x^j. \tag{4.19}$$

Let

$$P(\theta) = \sum_{i=0}^{n}\left[\alpha\Big(a_i I + \lambda\sum_{j=0}^{i} a_j (sM)^{j-i-1}\Big)\vec{1}\right]\theta^i \tag{4.20}$$

be an n-th order polynomial in θ. Let $\theta_1, \theta_2, \cdots, \theta_n$ be the roots of $P(\theta)$. We assume they are distinct.

Let Θ be the Vandermonde matrix of $\theta_1, \theta_2, \cdots, \theta_n$, i.e.:

$$\Theta = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \theta_1 & \theta_2 & \cdots & \theta_n \\ \theta_1^2 & \theta_2^2 & \cdots & \theta_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1^{n-1} & \theta_2^{n-1} & \cdots & \theta_n^{n-1} \end{pmatrix}.$$

Since all θ_i are distinct, Θ is invertible (cf. [20]).

Let $M_0 = I$, and define $M_j, j \geq 1$ recursively by:

$$M_j = sMM_{j-1} + \lambda\alpha M_{j-1}\vec{1}I. \tag{4.21}$$

Let

$$\begin{aligned} &m_i = \alpha M_i \vec{1} \qquad (i = 0, 1, \cdots, n-1) \quad \text{and} \\ &m = (m_0, m_1, \cdots, m_{n-1})^{\mathrm{T}}. \end{aligned} \tag{4.22}$$

With these notations, we are ready to state the following theorem.

Theorem 20. *Suppose the service time has a phase type distribution with ccdf given by* (4.18)*, and assume that the polynomial in* (4.20) *has n distinct roots, $\theta_1, \theta_2, \cdots, \theta_n$. The equilibrium pdf of the $\tilde{W}_G$ process is given by*

$$\tilde{f}_G(x) = \begin{cases} \tilde{c}_G\gamma\sum_{i=1}^{n} A_i e^{\theta_i x} & \text{if } 0 < x < b \\ \tilde{c}_G\gamma\alpha\left[e^{sMx} + \sum_{i=1}^{n}\lambda A_i e^{sMx}(\theta_i I - sM)^{-1}(e^{(\theta_i I - sM)b} - I)\right]\vec{1} & \text{if } x \geq b \end{cases} \tag{4.23}$$

where $A = (A_1, A_2, \cdots, A_n)^T$ *is given by:*

$$\Theta A = m. \tag{4.24}$$

The probability that the vqt is zero is

$$\begin{aligned}\tilde{c}_G = \Big\{ & \sum_{i=1}^{n} \gamma \frac{A_i}{\theta_i}(e^{\theta_i b} - 1) \\ & - \alpha \Big[\sum_{i=1}^{n} \lambda\gamma A_i (sM)^{-1} e^{sMb} (\theta_i I - sM)^{-1} (e^{(\theta_i I - sM)b} - I) \Big] \vec{1}. \\ & - \gamma\alpha (sM)^{-1} e^{sMb} \vec{1} + 1 \Big\}^{-1}.\end{aligned} \tag{4.25}$$

We also compute

$$r_G \approx \int_b^\infty \tilde{f}_G(x)\mathrm{d}x = -\tilde{c}_G \gamma\alpha (sM)^{-1} e^{sMb} \Big[I + \sum_{i=1}^{n} \lambda A_i (\theta_i I - sM)^{-1} (e^{(\theta_i I - sM)b} - I) \Big] \vec{1},$$

and

$$w_G \approx \int_0^b x\tilde{f}_G(x)\mathrm{d}x = \tilde{c}_G \gamma \sum_{i=1}^{n} \frac{A_i(1 + \theta_i b e^{\theta_i b} - e^{\theta_i b})}{\theta_i^2}, \quad w'_G = \frac{w_G}{1 - r_G}. \tag{4.26}$$

Remark: We do not give formulas for the case $\rho = 1$. Notice that $\rho = 1$ implies $-\alpha(sM)^{-1}\vec{1} = 1/\lambda$. Then $\theta_1 = 0$ is one of the roots of (4.20). In this case, we replace the zero-dividing terms in the formulas by the corresponding limits. These terms and their limits are: $\lim_{\theta_1 \to 0}(e^{\theta_1 b} - 1)/\theta_1 = b$ and $\lim_{\theta_1 \to 0}(1 + \theta_1 b e^{\theta_1 b} - e^{\theta_1 b})/\theta_1^2 = b^2/2$. Also note that some of our computations require convergence of the integral $\int_b^\infty e^{sMx}$, which is guaranteed by the fact that all eigenvalues of M have negative real part.

4.3.2 Approximation II: $\dot{J} = \bar{S}$, $\ddot{J} = \hat{S}$.

The equilibrium distribution of the W process is approximated by the solution to the following system of equations:

$$\tilde{f}_G(x) = \lambda \int_0^{x \wedge b} \tilde{f}_G(u) G_{\bar{S}}(x-u) \mathrm{d}u + \tilde{c}_G \gamma G_{\hat{S}}(x), \tag{4.27a}$$

$$\int_0^{\infty} \tilde{f}_G(x) \mathrm{d}x + \tilde{c}_G = 1, \tag{4.27b}$$

When $b \to \infty$, Equation (4.16) becomes

$$\tilde{w}_G \to \frac{\gamma[(1-\rho)\mathrm{E}(\hat{S}^2) + \rho \mathrm{E}(\bar{S}^2)]}{2(1-\rho)(\nu + 1 - \rho)}. \tag{4.28}$$

For phase-type service times, neither $\mathrm{E}(\hat{S}^2)$ nor the solution to Equation (4.27) can be obtained analytically. We use numerical methods. Particularly, we use quadrature method in solving Equation (4.27) where numerically solving Volterra Integral Equation (VIE) of the second kind plays a key role (note that the balance equation becomes a VIE of the second kind if $x \wedge b$ is replaced by x). Meanwhile, numerically solving VIE of the second kind alone is a deserving topic in applied mathematics (cf. [35]). A fast and reliable method to solve VIE of the second kind can be found in [24]. Of course, the numerical method used in solving Equation (4.27) can also be used to solve Equation (4.17).

4.4 Connection Between Balking and Reneging

Reneging is also a common phenomenon in queueing systems. As balking, reneging can be viewed as an indication of customers' impatience. In this section, we address the connection between the wait-based balking model and reneging model. We argue that the wait-based balking rule we introduced in Section 4.1 in fact produces the

same vqt process as the one corresponding to a particular reneging rule and vice versa. Consider the following balking rule. An arriving customer joins the queue if and only if he/she observes that the vqt is no more than $B \geq 0$, which can be either deterministic or random. The corresponding reneging rule is as follows. A customer waiting in queue leaves after being kept in queue for B unit of time. Using the same service times $\{S_i\}$ and inter-arrival time $\{U_i\}$, from the definition of vqt, the sample path of the vqt process governed purely by the balking rule is identical to that governed purely by the reneging rule. This is because in both cases, a jump in the vqt process occurs only if the vqt at the arrival epoch is no more than B. Moreover, suppose the vqt process evolves under a mixture of balking and reneging rules with threshold B and $R \geq 0$ (deterministic or random) respectively. This can be converted to a pure balking or pure reneging model with threshold $\min\{B, R\}$. Therefore, balking and reneging can be viewed as two interpretations of the general concept of customer impatience, and the vqt process unifies them. It is worth mentioning that although we illustrate our method by deterministic balking threshold, our method is not restricted to that. A random B is useful for the following reasons. First, we have a random B when converting to a pure balking model if the reneging threshold is random. Second, for pure balking interpretation, if the vqt information is not precisely available to the arriving customer and is provided as an estimate instead, the randomness presented by the estimate can not be eliminated.

The balking model and reneging model do differ in the number of customers and workload in the system. The reneging behavior results in more number of customers and workload in the system than balking behavior. It is possible to derive the relationship between the performance measures of the original (pure reneging, or reneging and balking) system and the vqt-equivalent pure balking model. For example, suppose the arrival is $\mathrm{PP}(\lambda)$. Let $n_{B,R}$ and n_B be the long-run average number of customers in the balking-reneging model and the vqt-equivalent pure balking model respectively.

Let $w_{B,R}$ and w_B be the long-run average workload in the balking-reneging model and the vqt-equivalent pure balking model respectively. Then by PASTA, it can be shown that

$$n_{B,R} = n_B + p_r \lambda \mathrm{E}(R),$$

and

$$w_{B,R} = w_B + p_r \lambda \mathrm{E}(R)\mathrm{E}(S),$$

where

$$p_r = \lim_{t \to \infty} \Pr\{R \leq W_G(t) \leq B\}$$

is the fraction of reneging customers. From the economic point of view, the balking rule saves system resources (waiting room, buffers etc.). For reneging rule, the reneging customers spend time waiting in the queue but do not get the desired service in the end.

The balking interpretation is advantageous in analytical study. For example, the previous sections actually solve the corresponding problems for the reneging model with deterministic threshold, which is new and would be difficult to analyze if we start from the reneging interpretation. The Erlang-A model, which is a pure reneging model with an exponentially distributed threshold, can be easily solved as well with slight modification of Equation (4.2). On the other hand, the reneging interpretation is advantageous in simulation. To simulate the wait-based balking model, one needs to keep track of the vqt, which involves keeping track of the workload in each queue of the equivalent parallel queue (see Equation (4.7)). This can be very expensive, especially when the number of servers is large. In fact, we use the reneging interpretation in our simulation.

4.5 Design of Simulation Experiments

The exact results for the $M/G/s$ balking models are needed to assess the accuracy of the approximations developed in Section 4.3. Since such exact results are lacking except when $G = M$ or $s = 1$, we develop a high precision, high confidence level and high coverage simulation. Our objective is to estimate the performance measures sufficiently accurately to serve as the references against which the approximations can be judged. The simulation aims at constructing a 99% confidence interval whose width is less than 1% of the estimate value and whose actual coverage is high. As an illustration, we use the long-run average queueing time as the performance measure of interest in carrying out the simulation.

The simulation performs independent replications until the desired precision is achieved or the maximum number of replications is reached. At the end of each replication a 99% confidence interval is constructed. If the confidence interval is narrow enough, the simulation stops. Otherwise, the simulation continues to perform another independent replication. A confidence interval for k replications are constructed as follows. We assume the means of the replications $\{X_j : 1 \leq j \leq k\}$ constitute a random sample of size k from a Normal distribution with mean μ and variance σ^2. Then the $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm t_{1-\alpha/2,k-1}\frac{S_k}{\sqrt{k}},$$

where $\bar{X} = \frac{1}{k}\sum_{j=1}^{k} X_j$ is the sample mean, $S_k^2 = \frac{1}{k-1}\sum_{j=1}^{k}(X_j - \bar{X})^2$ is the sample variance and $t_{1-\alpha/2,k-1}$ is the $(1-\alpha/2)$-th quantile of the Student t-distribution with $k-1$ degrees of freedom. For $\alpha = 0.01$, the simulation performs at most 150 replications since more replications does not help decreasing the half length of the confidence interval significantly.

The actual coverage of the confidence interval relies on how precisely the assump-

tion that $\{X_j : 1 \leq j \leq k\}$ constitute a random sample of size k from a normal distribution holds. We assume that the process obeys a Central Limit Theorem (CLT) for dependent processes. In order to achieve a high actual coverage, a replication stops after each server on the average serves 1 million customers. We discard the first 10% observations to ensure that the process reaches steady state and to avoid initialization bias. To ensure the quality of randomness, we use a non-linear additive feedback pseudo-random number generator provided by the standard C library. The period of this pseudo-random number generator is approximately $16(2^{31} - 1)$.

In order to reduce the computation load, we simulate the equivalent reneging model. It is easy to translate or compute the performance measures for the balking model from this reneging model. For example, the reneging fraction of the reneging model is the balking fraction of the balking model; the average queueing time for all customers in the balking model is the total queueing time for all served customers in the reneging model divided by the total number of customers.

Note that the simulation cannot always achieve the preset goal. The relative width of the confidence interval can be higher than 0.01 after 150 replications. According to our experience, this happens when the quantity in estimation is very small (in the range of 10^{-4} to 10^{-9}), for example, the average queueing time of a low traffic intensity (e.g. 0.1) system. Or the replications consistently return zero as the estimate. This happens when the quantity in estimation is extremely small (in the degree of 10^{-12} or smaller), for example, the average queueing time of a low traffic intensity (e.g. 0.1) and large number of servers (e.g. 100) system. In such cases, numerical and/or assumption errors become significant.

4.6 Numerical Results

In this section, we illustrate our numerical results. We consider the $M/PH/s$ model with wait-based balking and three different phase type service time distributions:

exponential, Erlang and Hyper-exponential. The parameters are the same as those used in Section 2.6 on page 29. The balking threshold b is set to be 2. For each of the three service time distributions, we use $s \in \{3, 10, 100\}$ and compute the long-run average queueing time for all served customers (w') and fraction of rejected customers (r) for different values of $\rho \in [0.1, 1.2]$ by using: a) Approximation I (using analytic formulas for the solution to Equation (4.17)), b) Approximation II (using numerical method to solve Equation (4.27)) , and c) simulation methods. Notice that for exponential service time distribution, both Approximation I and Approximation II are exact. This gives us a method to verify the accuracy of simulation which turned out to be satisfactory in our experiments.

Figure 4.2 shows the long-run average queueing time for all served customers as a function of ρ. The w' values for exponential service times are exact. Others are from simulation. It can be seen that for almost all given value of ρ, there is an ordering of the queueing times for different service times according the order of the variances, either $w'_{\text{hyper}} > w'_{\text{exp}} > w'_{\text{erlang}}$ or $w'_{\text{hyper}} < w'_{\text{exp}} < w'_{\text{erlang}}$. The order reverses as ρ increases beyond a critical region. Intuitively, the value w' should converge to $b = 2$ as ρ approaches infinity. This trend is best illustrated by the set of curves for $s = 100$. The more the servers, the faster these curves approach b. Moreover, as s increases, the queueing time becomes more sensitive around the point $\rho = 1$ and the overall difference of the queueing time between these service time distributions diminishes.

Figure 4.3 is the same as Figure 4.2 except that it shows the fraction of rejected customers. We have similar observations as those for w'. Notice that for all given ρ, $r_{\text{hyper}} > r_{\text{exp}} > r_{\text{erlang}}$. The order reversion we observe from Figure 4.2 does not happen here. In addition, the fraction of rejected customers is almost linear in ρ when $\rho \geq 1$.

Figure 4.4, 4.5 and 4.6 display the relative errors in computing the queuing times using the two approximations compared to the simulation results. We truncated the

part where the queueing time is extremely close to 0. In such a case, the relative error of the simulation results to the unknown exact values can be so large that it is not reasonable to use the simulation results to access the accuracy of the approximations in a relative sense. The figures verify the fact that as ρ increases, Approximation I and Approximation II become closer since the coefficient $\tilde{c}_G\gamma$ in Equation (4.9a) approaches 0. Overall, Approximation II is more accurate than Approximation I. It is worth noting that the curves for Approximation II are flatter. This, in some sense, indicates the robustness of Approximation II in comparison to Approximation I. However, this advantage of Approximation II over Approximation I is balanced by the fact that we need to use numerical methods to solve Equation (4.27). Both approximations are, of course, far quicker than the simulation.

4.7 Concluding Remarks

In this chapter we have obtained exact analytical results for the limiting behavior of an $M/M/s$ system with wait-based balking. These results also yield analytical results for the corresponding reneging case, which is more complicated if studied as a reneging system. Using these results we have proposed two approximations for the $M/G/s$ system with wait-based balking. We have done extensive numerical and simulation experiments to conclude that Approximation I is easier to compute than Approximation II, but Approximation II is more accurate than Approximation I over a wide parameter space.

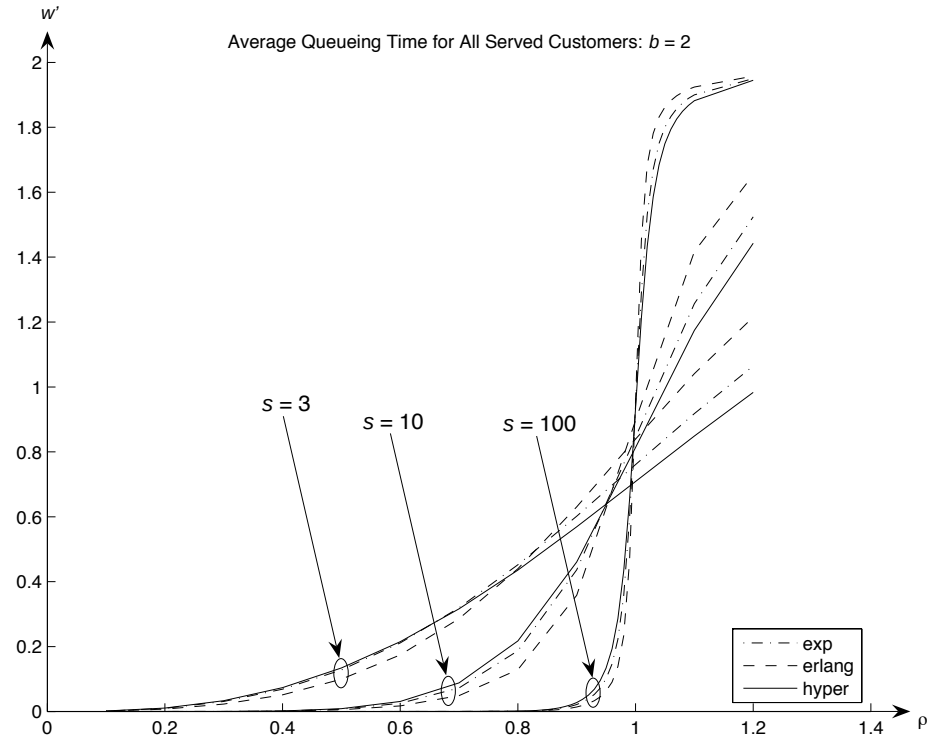

Figure 4.2: Long-run Average Queueing Time for All Served Customers

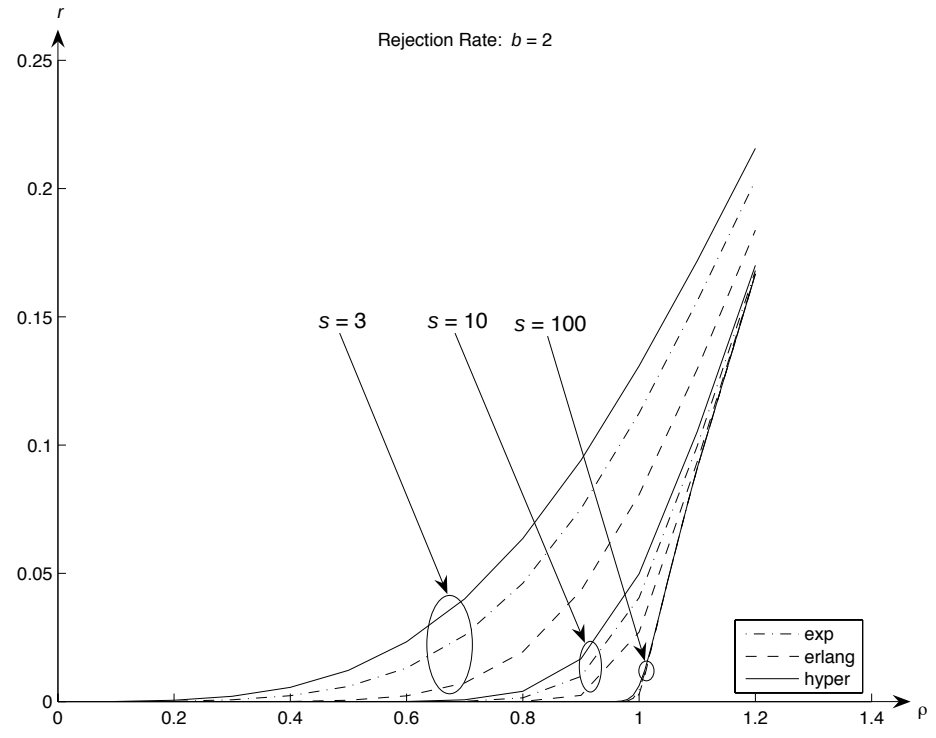

Figure 4.3: Fraction of Rejected Customers

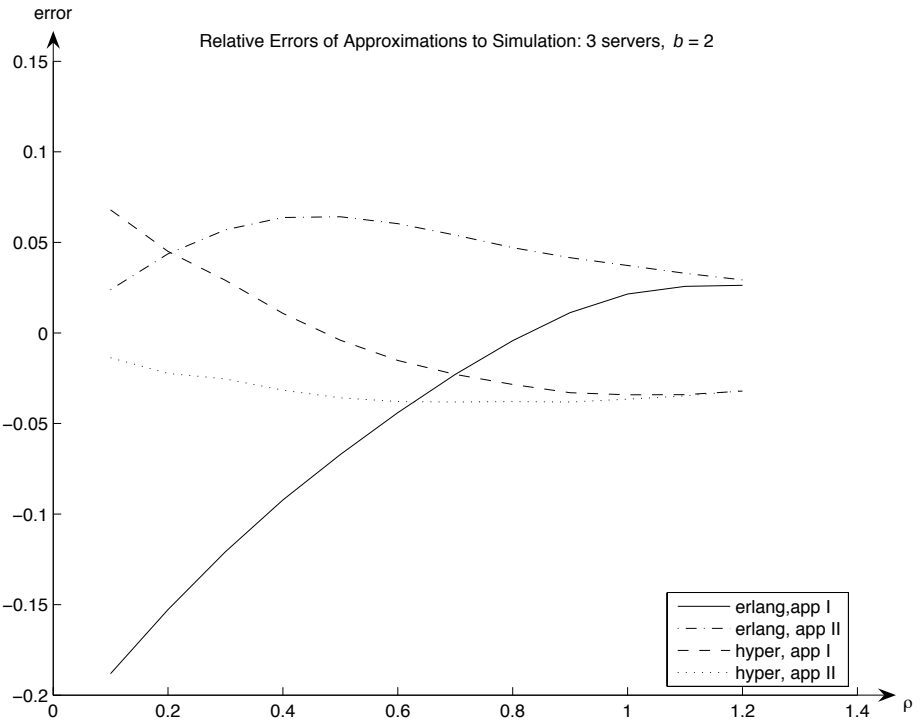

Figure 4.4: Relative Errors of Approximations: $s = 3$

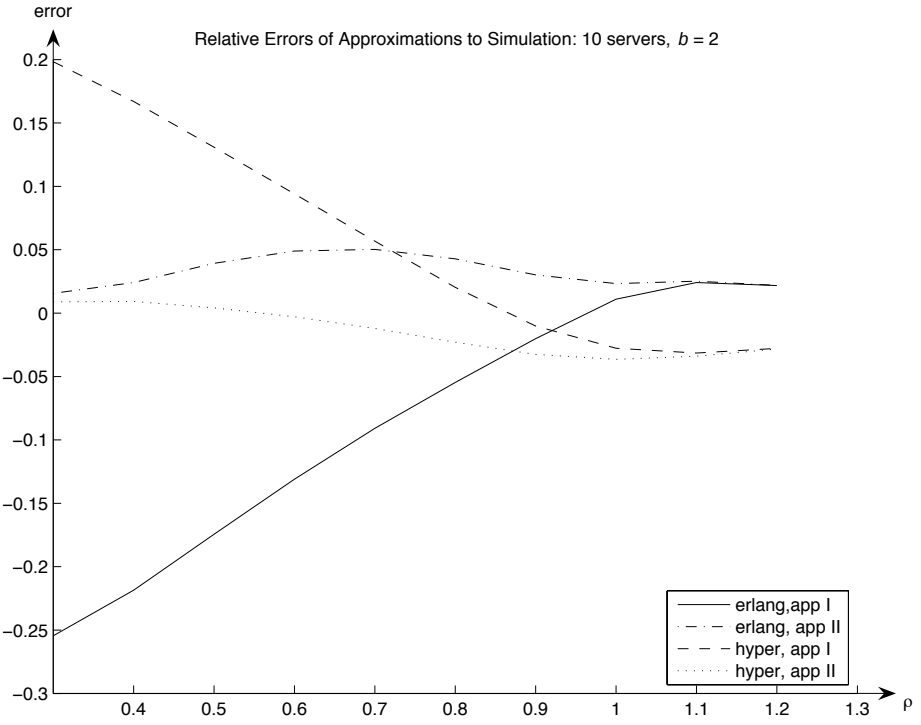

Figure 4.5: Relative Errors of Approximations: $s = 10$

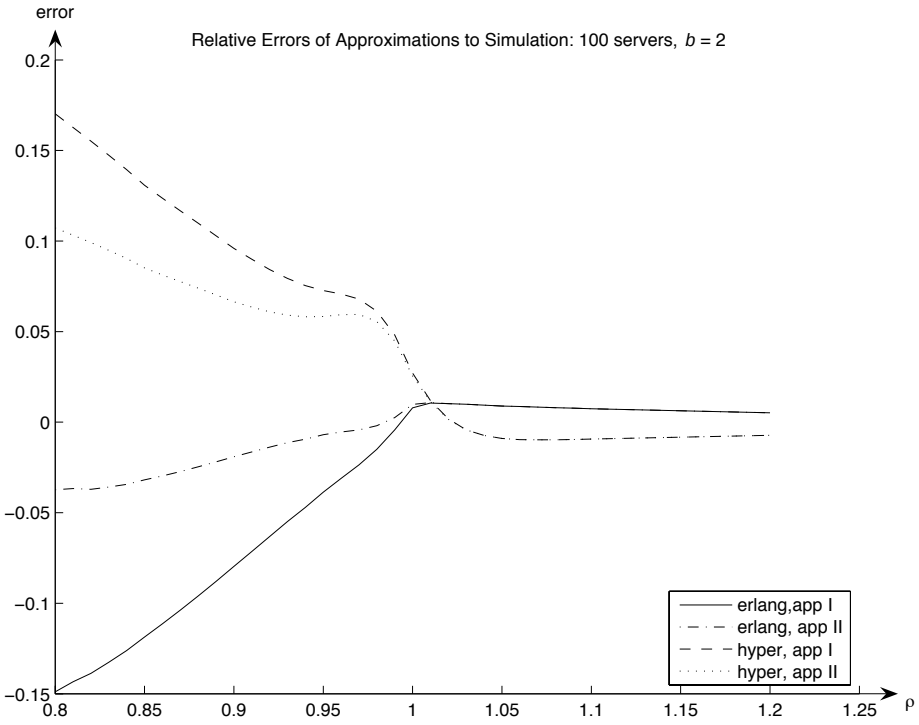

Figure 4.6: Relative Errors of Approximations: $s = 100$

Bibliography

[1] M. Armony and C. Maglaras, *Contact centers with a call-back option and real-time delay information*, Oper. Res. **52** (2004), 527–545.

[2] M. Armony, N. Shimkin and W. Whitt, *The impact of delay announcements in many-server queues with abandonment*, preprint.

[3] S. Asmussen and M. Bladt, *A sample path approach to mean busy periods for Markov-modulated queues and fluids*, Adv. Appl. Prob. **26** (1994), 1117–1121.

[4] R. Bekker, *Finite-buffer queues with workload-dependent service and arrival rates*, Queue Syst. Theory Appl. **50** (2005), 231–253.

[5] R. Bekker, S. C. Borst, O. J. Boxma, and O. Kella, *Queues with workload-dependent arrival and service rates*, Queue Syst. Theory Appl. **46** (2004), 537–556.

[6] Václav E. Beneš, *General stochastic processes in the theory of queues*, Addison-Wesley, Reading, MA, 1963.

[7] N. K. Boots and Henk C. Tijms, *A multiserver queueing system with impatient customers*, Manage. Sci. **45** (1999), 444–448.

[8] O. J. Boxma and V. Dumas, *The busy period in the fluid queue*, Tech. report, CWI Report PNA-R9718, November 1997.

[9] O. J. Boxma, D. Perry, and W. Stadje, *Clearing models for $M/G/1$ queues*, Queue Syst. Theory Appl. **38** (2001), 287–306.

[10] T. M. Chen and V. K. Samalam, *Time dependent behavior of fluid buffer models with Markov input and constant output rates*, SIAM J. Appl. Math. **55** (1995), 784–799.

[11] J. W. Cohen, *On up and down crossings*, J. Appl. Prob. **14** (1977), 405–410.

[12] Daryl J. Daley, *Frontiers in queueing: models and applications in science and engineering*, Probability and stochastics, pp. 35–59, CRC Press, Inc., 1997.

[13] B. Doshi, *Level-crossing analysis of queues*, pp. 3–33, Oxford University Press, New York, 1992.

[14] N. Finizo and G. Ladas, *Ordinary differential equations with modern applications*, Wadsworth, Belmont, CA, 1982.

[15] O. Garnett, A. Mandelbaum, and M. Reiman, *Designing a call center with impatient customers*, Manufacturing & Service Operations Management **4** (2002), 208–227.

[16] B. Gavish and P. J. Schweitzer, *The markovian queue with bounded waiting time*, Management Science **23** (1977), 1349–1357.

[17] B. V. Gnedenko and I. N. Kovalenko, *Introduction to queueing theory*, second edition, pp. 57–63, Birkhäuser, Boston, 1989.

[18] P. Guo and P. Zipkin, *Analysis and comparison of queues with different levels of delay information*, Management Science **53** (2007), 962–970.

[19] Daniel P. Heyman and Matthew J. Sobel, *Stochastic models in operations research: stochastic processes and operating characteristics*, McGraw-Hill, New York, 1982.

[20] K. Hoffman and R. Kunze, *Linear algebra, second edition*, Prentice Hall, New York, 1971.

[21] P. Hokstad, *Approximations for the $M/G/m$ queue*, Oper. Res. **26** (1978), 510–523.

[22] _____, *A single server queue with constant service time and restricted accessibility*, Manage. Sci. **25** (1979), 205–208.

[23] J. Q. Hu and M. A. Zazanis, *A sample path analysis of $M/GI/1$ queues with workload restrictions*, Queue Syst. Theory Appl. **14** (1993), 203–213.

[24] P. W. den Iseger and M. A. J. Smith and R. Dekker, *Computing compound distributions faster!*, Insurance: Mathematics and Economics, **20** (1997), 23–34.

[25] S. G. Johansen and S. Stidham, *Control of arrivals to a stochastic input-output system*, Adv. Appl. Prob. **12** (1980), 972–999.

[26] Toshikazu Kimura, *Approximations for multi-server queues: system interpolations*, Queue Syst. Theory Appl. **17** (1994), 347–382.

[27] G. Koole and A. Mandelbaum, *Queueing models of call centers: an introduction*, Ann. Oper. Res. **113** (2002), 41–59.

[28] E. Kreyszig, *Advanced engineering mathematics, eighth edition*, John Wiley and Sons, New York, 1999.

[29] V. G. Kulkarni, *Modeling and analysis of stochastic systems*, Chapman and Hall, London, 1995.

[30] _____, *Frontiers in queueing: models and applications in science and engineering*, Probability and stochastics, pp. 321–338, CRC Press, Inc., 1997.

[31] V. G. Kulkarni and A. Narayanan, *First passage times in fluid models with application to two priority fluid systems*, IPDS'96, 1996.

[32] V. G. Kulkarni and E. Tzenova, *Mean first passage times in fluid queues*, Oper. Res. Letters **30** (2002), 308–318.

[33] A. M. Lee and P. A. Longton, *Queueing processes associated with airline passenger check-in*, Oper. Res. Quart. **10** (1959), 56–71.

[34] M. Miyazawa, *Approximation for the queue-length distribution of an $M/GI/s$ queue by the basic equations*, J. Appl. Prob. **23** (1986), 443–458.

[35] Arun N. Netravali, *Spline approximation to the solution of the volterra integral equation of the second kind*, Math. Comp. **27** (1973), 99–106.

[36] G. F. Newell, *Approximate stochastic behavior of n-server service systems with large n (lecture notes in economics and math. systems, vol. 87)*, Springer-Verlag, New York, 1973.

[37] A. Nozaki and Sheldon M. Ross, *Approximations in finite-capacity multi-server queues with poisson arrivals*, J. Appl. Prob. **15** (1978), 826–834.

[38] D. Perry and S. Asmussen, *Rejection rules in the $M/G/1$ queue*, Queue Syst. Theory Appl. **19** (1995), 105–130.

[39] D. Perry and W. Stadje, *Duality of dams via mountain processes*, Oper. Res. Letters **31** (2003), 451–458.

[40] D. Perry, W. Stadje, and S. Zacks, *Busy period analysis for $M/G/1$ and $G/M/1$ type queues with restricted accessibility*, Oper. Res. Letters **27** (2000), 163–174.

[41] _____, *The $M/G/1$ queue with finite workload capacity*, Queue Syst. Theory Appl. **39** (2001), 7–22.

[42] N. U. Prabhu, *Stochastic storage processes: queues, insurance risk, dams, and data communication*, second ed., Applications of mathematics, pp. 123–124, Springer-Verlag, New York, 1998.

[43] W. Scheinhardt, N. van Foreest, and M. Mandjes, *Continuous feedback fluid queues*, Oper. Res. Letters **33** (2005), 551–559.

[44] Lajos Takács, *Introduction to the theory of queues*, Oxford University Press, New York, 1962.

[45] Henk C. Tijms, *Approximations for the steady-state probabilities in the $M/G/c$ queue*, Adv. Appl. Prob. **13** (1981), 186–206.

[46] _____, *Stochastic modelling and analysis: a computational approach*, John Wiley and Sons, Chichester, 1986.

[47] W. Whitt, *Improving service by informing customers about anticipated delays*, Management Science, **45** (1999), 192–207.

[48] _____, *Engineering solution of a basic call-center model*, Management Science **51** (2005), 221–235.