On Priority Queues with Impatient Customers: Stationary and Time-Varying Analysis

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE IN
OPERATIONS RESEARCH AND SYSTEMS ANALYSIS

Lubov Rozenshmidt

Submitted to the Senate of the Technion – Israel Institute of Technology

Iyyar, 5768 Haifa May, 2008

This research thesis was done under the supervision of professor Avishai Mandelbaum in the Faculty of Industrial Engineering. I would like to express him my gratitude for his valuable guidance and support throughout every stage of the work, for sharing his vast knowledge and experience.

Thanks to my Mom and Dad, without whom I would not be where I am today, for their continuous support and belief in me. Words are not enough to express all my love and appreciation to you both.

I am thankful to my wonderful boyfriend, Arie, for all his encouragement, and for willing to put up with the long hours that this work required.

My deepest gratitude go to my uncle Professor Yehuda Charit and his wife Lena for showing me the meaning of the word ”Family”.

I thank Doctor Sergey Zeltyn for many helpful discussions that led to this thesis.

Special thanks go to Shimrit Maman, whose help is greatly appreciated.

Finally, the generous financial help of Technion is gratefully acknowledged.

Contents

List of Tables

List of Figures

Abstract

We consider Markovian non-preemptive and preemptive priority queueing systems with impatient customers, under the assumption that service rates and abandonment rates are equal across customer types. For such systems in steady-state, we develop an algorithm for calculating the expected waiting time of any type.

We then assume that the number of servers is large, formally taking this number to infinity, which enables an asymptotic analysis in three operational regimes: an Efficiency-Driven (ED) regime, in which the focus is on servers' utilization (efficiency), a Quality-Driven (QD) regime, where the focus is on the quality of the service and a Quality-and-Efficiency-Driven (QED) regime, where efficiency is carefully balanced against service quality.

Our asymptotic analysis provides simplified expressions for some operational measures, e.g., the expected waiting time of any type. But, as importantly, it yields structural insight. For example, assuming that the offered load of the lowest priority is non-negligible, we show that preemption and non-preemption are essentially equivalent, as far as average waiting times are concerned. Moreover, the delayed customers of all classes other than the lowest priority wait, on average, the same time as in a queue without abandonment. (In other words, their service level is high enough to render their abandonment negligible.)

Stationary behavior turns out to provide important insights on the behavior of time-varying queues. Specifically, under an appropriate scale and time-varying staffing, the performance of time-varying versions of the above-mentioned systems is in fact stable in time. Moreover, this stable performance matches remarkably well the performance of naturally-corresponding steady-state systems. We demonstrate all that via simulating queues in which the arrival-rates are taken from four real call-centers.

More specifically, we first simulate call-center environments with a single customer type whose arrivals are described by an empirical function. After that, we present simulation results of queues with two customer types that arrive according to analytical functions. The conclusion of these experiments, as mentioned, is that in many cases, and under appropriate staffing, stationary models can be used to properly predict the performance of time-varying queues.

In our last chapter, the traditional heavy-traffic approximation for expected waiting time, based on the first two moments of the service-time distribution, is considered. We check this approximation by simulating queues with different service-time distributions and conclude that in the QED regime, such two-moment approximations are inaccurate.

List of Symbols

$\stackrel{d}{=}$	Equal in distribution	14
α	Halfin-Whitt / Garnett function	11
β	Implied service grade under the QED regime	11
Θ	Convergence rate, namely, $\lim_{N\to\infty} X_N = a$, $-\infty < a < \infty$, at rate $\Theta(\phi(N))$ if $\lim_{N\to\infty} \frac{X_N - a}{\phi(N)} = 1$.	28
λ	Total arrival rate	4
λ_k	Arrival rate of type k	33
$\lambda_{1\to k}$	$\lambda_{1\to k} = \sum_{i=1}^{k} \lambda_i$, Arrival rate of the first k types	34
μ	The service rate common to all customer types	5
π	$\pi_j \triangleq \lim_{t\to\infty} P(Q(t) = j),\ j \geq 0$. Steady-State distribution	5
ρ	Offered load per server	18
ρ_k	$\frac{\lambda_k}{N\mu}$. Fraction of time spent on customers of type *k* in $M/M/N$; Offered load per server for type k in $M/M/N+M$.	35
σ_k	$\sum_{i=1}^{k} \rho_i$. Fraction of time spent on customers of the first k types in $M/M/N$ Offered load per server for the first type k in $M/M/N+M$	35
$\Phi(\cdot)$	The standard normal distribution function	24
$\phi(\cdot)$	The standard normal density function	24
$E_{1,N}(\lambda)$	Erlang-B Formula. The blocking probability	6
$E_{2,N}(\lambda)$	Erlang-C Formula. The delay probability in $M/M/N$.	5
$E_{\lambda_k}(X)$	Expected value of X in a queue with arrival rate λ_k	33
$E(X^{1\to k})$	Expected value of X in a queue with arrival rate $\lambda_{1\to k}$	34
$E_{np}(X^k)$	Expected value of X for type k under the non-preemptive priority	33
$E_{pr}(X^k)$	Expected value of X for type k under the preemptive priority	33
$E_{pr}(X^{1\to k})$	Expected value of X for the first k types under the non-preemptive priority	33
$E_{np}(X^{1\to k})$	Expected value of X for the first k types under the preemptive priority	33

$Geom_0(\cdot)$	Geometric distribution starting from zero	19
$h(\cdot)$	Hazard rate of the standard normal distribution	11
K	Number of customer types	32
$L(t)$	Total number of customers in queue at time t	5
$L(\infty)$, L	Number of customers in steady state (when exists)	5
L_-	Number of customers in the system where at least one server is idle	17
L_+	Number of customers in the system where all the servers are busy	17
N	Number of servers	4
O	$a_N = O(b_N)$ if $\exists\, c < \infty :\ \lim_{N\to\infty} a_N/b_N \leq c$	48
$P_{\lambda_k}(X)$	Probability of X in a queue with arrival rate λ_k	34
$P(X^{1\to k})$	Probability of X in a queue with arrival rate $\lambda_{1\to k}$	34
$P_{np}(X^k)$	Probability of X for type k under non-preemptive priority	33
$P_{pr}(X^k)$	Probability of X for the type k under preemptive priority	33
$P_{pr}(X^{1\to k})$	Expected value of X for the first k types under non-preemptive priority	33
$P_{np}(X^{1\to k})$	Expected value of X for the first k types under preemptive priority	33
$Poisson(\lambda)$	Poisson process with λ	5
R	Offered load (often $R = \lambda/\mu$ in Markovian queues)	4
$T_{N-1,N}$	The expected duration of an idle period.	18
$T_{N,N-1}$	The expected duration of a busy period.	18
$W_q(i)$	Waiting time of the i-th arrival	5
$W_q(\infty)$, W_q	Waiting time in steady state (when exists)	5
W_q^k	Waiting time of type k customers in steady-state	33

List of Acronyms

ED	Efficiency Driven	10
FCFS	First Come First Served	66
ISA	Iterative Staffing Algorithm	55
PASTA	Poisson Arrivals See Time Averages	6
PSA	Pointwise Stationary Approximation	64
QD	Quality Driven	10
QoS	Quality of Service	10
QED	Quality and Efficiency Driven	10
cdf	Cumulative Distribution Function	64

Chapter 1

Introduction

Telephone call centers have turned into a widespread and highly preferred means for many organizations to contact their customers. These organizations cover both the public and the private sectors. For some of them, such as cellular companies, call centers are in fact their most important contact channel with their customers. It is no wonder, thus, that the call center world is expanding dramatically and is becoming a vital part of our service-driven society. In concert with this state of affairs, call centers have also become a significant object for academic research; this is amply testified by the growing literature cited in [24] for example, some of which is surveyed in [10].

The majority of the operating costs of a call center are salaries for its staff. Overstaffing leads to undesirable high costs and understaffing results in long waits, dissatisfied customers and overworked and frustrated telephone agents. Many call centers use a ”1-800” service, in which case waiting costs become part of their operational costs. Additionally, the costs of dissatisfied customers could be very significant, especially accounting for the fact that some abandon (during a particular call, or actually opt for the competition).

The call center environment is very complex. One of the main complexity factors is the need to cater to varying types of customers by agents with varying skills. One common approach is to cross-train the agents and then serve customers according to pre-assigned priorities. Naturally, staffing decisions must account for all this complexity, yet the challenge is to develop staffing rules that are simple and insightful enough for implementation. For example, the ”square-root safety staffing” rule is one, as will be surveyed below.

Many mathematical models have been developed for the complex environment of call centers (see [10]). Their main advantage is their simplicity of use, as well as the theoretical insights that they can often provide. Their main weakness is their limited modelling scope, which is restricted by our state-of-the-art analytical capabilities. Another weakness is the fact that some analytical background is required in order to apply these models comfortably. The latter could explain the wide gap between needs and prevalence: indeed, the most commonly-used model in support of call center staffing is the overly-simplistic M/M/N

queue (known as Erlang-C in call center circles): its severe assumptions are time-homogenous Poisson arrivals, exponential service times, i.i.d. customers and i.i.d. servers. A significant practical improvement, that is still alarmingly simple, is the M/M/N+M (Erlang-A), which accommodates customers' impatience, and its M/M/N+G generalization. (Readers are referred to [11] and [41] for more details on the latter two models.) A major goal of present-day call center research is extending the modelling scope of mathematical models - see [10] for a survey of the main directions of this research.

An alternative to mathematical models is offered by computer simulation. Simulation models, if created and used correctly, cope in principle with any level of model complexity, taking into account any small detail one wishes to consider. But simulation has significant limitations as well. For one, it is expensive/cumbersome to develop, maintain and run. Moreover, even with today's powerful computers, it can take many hours to run. Hence, the insight that simulation can provide is limited relative to theoretical models, when the latter are applicable.

To overcome the weaknesses of mathematical models and simulation, a research trend has recently emerged in which the two have been combined to nurture each other, and this is the approach adopted here. (An example of such research is [18].) More precisely, we will follow the approach in [9], which combined theoretical models with simulation in order to develop dynamic staffing rules, in the face of time-varying demand. The models in [9] are restricted to iid customers and iid servers. In this thesis, we extend [9] to cover some of the models in [15], namely allow customers that are of multiple classes.

1.1 The Structure of the Thesis

This work is organized in the following way. Chapter 3 introduces the analytical technique which later allows to represent the Delay Probability of Markovian queues in terms of their Busy and Idle periods. Then (Section 3.4) we analyze the Delay Probability under different operational regimes and obtain both its limit and its convergence rate.

Chapter 4 opens the discussion of priority queues by representing the known results for the Erlang-C queue under preemptive and non-preemptive priority disciplines. Later, in Chapter 5 we expand the same approach to the Erlang-A queue. We present the calculation of the expected waiting time of any customer type under preemptive priority and develop an algorithm for the calculation of this measure under non-preemptive priority.

Chapters 6 and 7 are devoted to the time-varying environment. In Chapter 6 we simulate four different call-centers using empirical data for the average service rate, average customers patience and the dynamics of arrivals during the day. We check the performance of each call center when the staffing level is determined by the *square-root safety-staffing* rule. In addition, in this chapter we check the impact of what is known as the *time-lag*, by comparing the results of call-center staffing using *Pointwise Stationary Approximation (PSA)* with the results of staffing using *Lagged Pointwise Stationary Approximation (Lagged PSA)*.

The analysis of a call center with two customer types under non-preemptive priority is presented in Chapter 7. This chapter is a generalization of some results discussed in [9]. The main finding presented in this chapter is the possibility to calculate many performance measures by using appropriate stationary single-type and two-types models.

In Chapter 8, we compare the $M/G/N+M$ system with a corresponding $M/M/N+M$, emphasizing the effect of the service-time distribution on system performance. The discussion follows the simulations results presented in [32] for $M/G/N$ queues. The purpose of this chapter is to check whether the impact of the service-time distributions on performance is similar to that described in [32]. According to conventional heavy-traffic theory, the expected waiting time in any $M/G/N$ queue under heavy traffic can be well approximated using the Kchinchine-Pollazcek formula. Schwartz [32] shows that this approximation is not good under the QED regime. We check the Kchinchine-Pollazcek formula analogue for the $M/G/N+M$ queues, developed by Ward in [33], and demonstrate that, here as well, in the QED regime there are significant differences among different service-time distributions with the same first two moments. Consequently, with and without abandonment, traditional heavy-traffic two-moment approximations are inaccurate in the QED regime.

Chapter 2

Literature Review and Theoretical Background

This work has several different directions. Thus, in order to achieve a more focused presentation, some of the next chapters begin with the review of the relevant theoretical background. This chapter presents some exact and asymptotic results which are common for all directions of our research.

2.1 Markovian N-Server Queues (Birth & Death)

In this section the two most common models that are used for call centers modeling and staffing are presented. The first model is Erlang-C: first developed around 1910 by Erlang [6], it has served until recently as the "working-horse" of call center staffing. Its main deficiency is that it ignores customers impatience, which is remedied by the second model, namely Erlang-A. Impatience leads to the phenomenon of customers abandonment, and, already around 1940, Palm [29] developed Erlang-A in order to capture it. We will be using Erlang-A to motivate three operating regimes for medium-to-large call centers: one which emphasizes service quality, another that focuses on operational efficiency, and the third, which is the main subject of the present thesis, carefully balances these two goals of quality and efficiency.

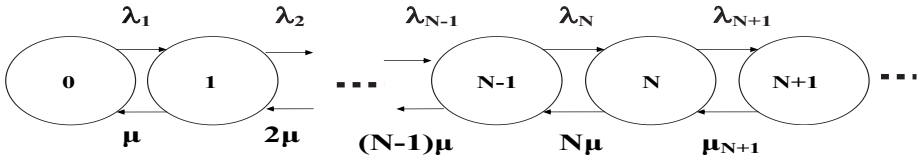
Many queues can be presented as a Birth & Death process. A general form of the transition-rate diagram for queues with N statistically identical independent servers is shown in Figure 2.1. Here

- λ_i is the arrival rate of customers at state i.
- μ_i is the service rate at state i. Note that $\mu_i = i \cdot \mu$ for any $i \leq N$. The service rate in the states $N+1,\ N+2\ldots$ is determined by the model specifics.

Define the following:

- $R = \lambda \times E(S) = \frac{\lambda}{\mu}$ is the Offered Load.

Figure 2.1: General Birth & Death Transition-Rate Diagram



- $L(t)$ is the total number of customers in the queue at time t.
- $W_q(k)$ is the waiting time of the k-th arrival.
- $L(\infty)$, $W_q(\infty)$ is the number of customers and waiting time in steady state (when exists).

Changing the parameters of the general model in Figure 2.1, we can get different special cases of queues. For instance, if in this diagram we set $\lambda_i \equiv \lambda$ for any $i = 1, 2, \ldots, N$, $\lambda_i \equiv 0$ otherwise, and $\mu_i \equiv 0$ for any $i \geq N+1$, we obtain the transition-rate diagram of the Erlang-B queue ($M/M/N/N$); and if $\lambda_i \equiv \lambda$ for any $i = 1, 2, \ldots$ and $\mu_i = N\mu$ for any $i \geq N+1$, we obtain the Erlang-C diagram ($M/M/N$). In the coming subsections we shall present these and some other models in more details.

2.1.1 Erlang-C

The classical $M/M/N$ (*Erlang-C*) queueing model is characterized by Poisson arrivals at rate λ, iid exponential service times with an expected duration $1/\mu$, and N servers working independently in parallel.

Formally speaking, customers' arrivals to the queueing system are described by $Poisson(\lambda)$ process. Individual service time are i.i.d. $exp(\mu)$ random variables. In addition, the processes of arrivals and service are independent.

Erlang-C is ergodic if and only if its traffic intensity $\rho = \lambda/(N\mu) < 1$; ρ is then the servers' utilization, namely the long-run fraction of time that a server is busy.

Let us recall that $L(t)$ is the total number of customers in $M/M/N$ at time t. Then $L = \{L(t), \ \ t \geq 0\}$ is a Markov Birth-and-Death process with the transition-rate diagram in Figure 2.2.

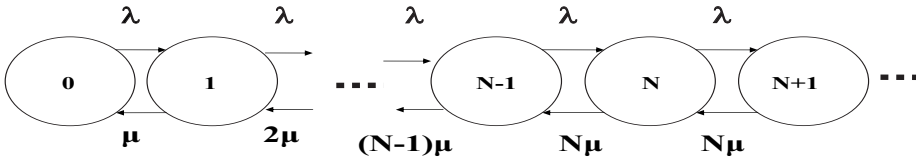
As usual, we denote the limiting-distribution vector of L by π:

$$\pi_j \triangleq \lim_{t\to\infty} P(L(t) = j), \qquad j \geq 0.$$

Solution of the following steady-state equations yields the probabilities π_j of being at any state j during steady-state:

$$\begin{cases} \lambda\pi_j = (j+1)\cdot\mu\pi_{j+1}, & 0 \leq j \leq N-1 \\ \lambda\pi_j = N\mu\pi_{j+1}, \quad j \geq N. \end{cases} \tag{2.1}$$

Figure 2.2: Erlang-C – Transition-Rate Diagram



The probability that in steady-state all the servers are busy is given by $\sum_{j=N}^{\infty} \pi_j$, the stationary probability of being in one of the states $\{N, N+1, \ldots\}$. This probability is sometimes referred to as the Erlang-C formula. It is denoted $E_{2,N}(\lambda)$ and is given by

$$E_{2,N}(\lambda) = \frac{(\lambda/\mu)^N}{N!(1-\rho)} \Big[\sum_{j=0}^{N-1} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^N}{N!(1-\rho)} \Big]^{-1}. \tag{2.2}$$

The Poison distribution of arrivals has an important and useful consequence, known as PASTA (Poison Arrivals See Time Averages): it implies that the probability $E_{2,N}$ is in fact also the probability that a customer is delayed in the queue (as opposed to being served immediately upon arrival).

2.1.2 Erlang-B

Another widely-used model is the $M/M/N/N$ or *Erlang-B* queue. In this model customers are not allowed to wait and when all N servers are busy, arriving customers leave immediately. Fitting the general diagram in Figure 2.1 to this case, we set $\lambda_k = 0$ for any $k \geq N$.

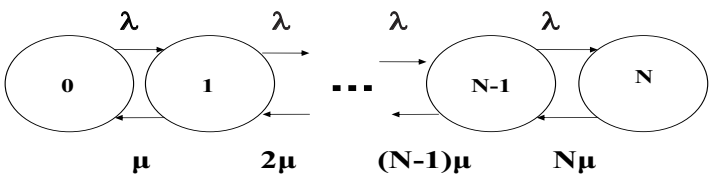
Again, we denote by $L(t)$ the total number of customers in $M/M/N/N$ at time t. Then $L = \{L(t),\ t \geq 0\}$ is a Markov Birth-and-Death process with the transition-rate diagram in Figure 2.3.

Erlang-B is always ergodic, and its steady-state distribution is given by

$$\pi_j = \frac{R^j}{j!} \Big/ \sum_{n=0}^{N} \frac{R^n}{n!}, \qquad 0 \leq j \leq N. \tag{2.3}$$

Here $R = \frac{\lambda}{\mu}$ is the *Offered Load*.

Figure 2.3: Erlang-B – Transition-Rate Diagram



This model is often used to calculate, by PASTA, the Loss Probability, which is denoted in the literature by $E_{1,N}$:

$$\pi_N \equiv E_{1,N} = \frac{R^N}{N!} \Big/ \sum_{n=0}^{N} \frac{R^n}{n!}. \tag{2.4}$$

The Delay Probability or Erlang-C formula (2.2) can be represented in terms of the Loss Probability as follows:

$$E_{2,N} = \Big[1 + \frac{1-\rho}{\rho E_{1,N-1}}\Big]^{-1}. \tag{2.5}$$

2.1.3 Erlang-A

Trying to make the M/M/N model more realistic and useful for modeling call-centers, the following assumption is added: each customer has limited patience, that is, as the waiting time in the queue increases the customer may abandon. We assume that patience is distributed exponentially with mean $1/\theta$. This model is referred to as Erlang-A (A for Abandonment). The Erlang-A model is fully characterized by the following four parameters:

- λ - Poisson arrival rate ($\lambda > 0$);
- μ - individual service rate ($\mu > 0$);
- N - number of agents ($N = 1, 2, \ldots$);
- θ - individual abandonment rate ($\theta > 0$).

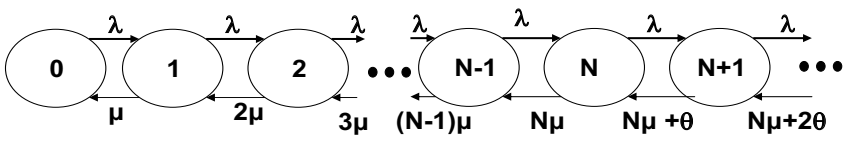
Erlang-A was first analyzed by Palm [29]. Here we give a short summary of some of its properties that are useful for this thesis.

Let us denote by $L(t)$ the total number of customers in $M/M/N+M$ at time t. Then $L = \{L(t),\, t \geq 0\}$ is a Markov Birth-and-Death process with the transition-rate diagram in Figure 2.4.

It can be shown that the limiting distribution of Erlang-A always exists (see, for example, [41]). In accordance with the ergodic theorem, it is equal to the stationary distribution, which is calculated by solving the following steady-state equations:

$$\begin{cases} \lambda\pi_j = (j+1)\cdot\mu\pi_{j+1}, & 0 \leq j \leq N-1 \\ \lambda\pi_j = (N\mu + (j-N+1)\theta)\cdot\pi_{j+1}, & j \geq N. \end{cases} \tag{2.6}$$

Figure 2.4: Erlang-A – Transition-Rate Diagram



The solution of (2.6) is well-known ([29]) and has the following form:

$$\pi_j = \begin{cases} \frac{(\lambda\mu)^j}{j!}\pi_0 & ,0 \le j \le N, \\ \prod_{k=N+1}^{j}\left[\frac{\lambda}{N\mu+(k-N)\theta}\right]\frac{(\lambda/\mu)^N}{N!}\pi_0 & ,j > N, \end{cases} \tag{2.7}$$

$$\pi_0 = \left[\sum_{j=0}^{N}\frac{(\lambda/\mu)j}{j!} + \sum_{j=N+k}^{\infty}\prod_{k=N+1}^{j}\Big(\frac{\lambda}{N\mu+(k-s)\theta}\Big)\frac{(\lambda/\mu)^N}{N!}\right]^{-1}.$$

In terms of the steady-state distribution π, it is possible deduce the Delay Probability. According to PASTA, this probability is determined by the following sum:

$$P(W_q > 0) = \sum_{j\ge N}\pi_j, \tag{2.8}$$

where π_j is the steady-state distribution (2.7).

Formulae (2.7) consist of infinite sums, which may cause some numerical problems. To circumvent them, Palm [29] proposed to use the following special functions:

$$Gamma\ Function: \qquad \Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt, \qquad x > 0;$$

$$Incomplete\ Gamma\ Function: \qquad \gamma(x,y) \triangleq \int_0^y t^{x-1}e^{-t}dt, \qquad x > 0,\ y \ge 0.$$

In addition, let us define the following function:

$$A(x,y) \triangleq \frac{xe^y}{y^x}\gamma(x,y). \tag{2.9}$$

By applying these special functions, Zeltyn in [41] obtained an elegant representation of the steady-state distribution (2.7):

$$\pi_j = \begin{cases} \pi_N \cdot \dfrac{N!}{j!(\lambda/\mu)^{N-j}} & ,0 \le j \le N, \\ \pi_N \cdot \dfrac{(\lambda/\theta)^{j-N}}{\prod_{k=1}^{j-N}(\frac{N\mu}{\theta}+k)} & ,j > N, \end{cases} \tag{2.10}$$

where

$$\pi_N = \frac{E_{1,N}}{1+[A(\frac{N\mu}{\theta},\frac{\lambda}{\theta})-1]\cdot E_{1,N}}. \tag{2.11}$$

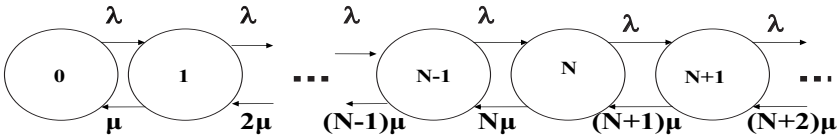
Recall that $E_{1,N}$ is the Erlang-B formula (2.4).

In Section 3.2 of this work we present a technique that enables one to calculate the Delay Probability (2.8) in terms of Idle and Busy periods. It is not hard to show that, when substituted into (2.8), the representation (2.10) yields exactly the same expression for the Delay Probability as our result (3.11).

M/M/∞

Let us now consider an important special case of $M/M/N+M$ queues. Assume that the average service

Figure 2.5: $M/M/\infty$ – Transition-Rate Diagram



rate μ is equal to the average individual abandonment rate θ. The transition-rate diagram of this process is identical to that of an $M/M/\infty$ queue with arrival rate λ and service rate μ (see Figure 2.5). Substituting $\mu = \theta$ into the solution of the balance equations ((2.6) or (2.10)) shows that the steady-state probability that there are exactly j customers in the queueing system is determined by the Poisson distribution with parameter $\frac{\lambda}{\mu}$:

$$\pi_j = \frac{e^{-\lambda/\mu}(\lambda/\mu)^j}{j!} \tag{2.12}$$

Expected Waiting and Abandonment Probability

A very useful property of queues with exponential patience time is the following relation between the expected waiting time and the probability of abandoning:

$$P(Aband) = \theta \cdot E(W_q). \tag{2.13}$$

Proof

The proof of this relation is very simple and is based on a balance equation and on Little's formula. According to the balance equation, the following equality holds:

$$\theta \cdot E(L_q) = \lambda \cdot P(Aband). \tag{2.14}$$

Here $E(L_q)$ is the average number of delayed customers in the system, in steady-state. The left side of this equation represents the abandonment rate from the considered queue and the right side represents the arrival rate of those customers who will eventually abandon.

Little's formula represents $E(L_q)$ in the terms of the arrival rate λ and the expected waiting time:

$$E(L_q) = \lambda E(W_q). \tag{2.15}$$

Substitution of (2.15) into (2.14) yields the relation (2.13). □

2.2 Three Operational Regimes: ED, QD, QED

Organizations have their own preferences in their everyday functioning. Some try to get the most from the available resources, while others see customers' satisfaction as the most important target. Depending on organizational preferences, three different operational regimes arise:

- Efficiency driven (ED)
- Quality driven (QD)
- Quality-Efficiency driven (QED)

As the number of servers increases, which is relevant for moderate to large call centers, these regimes can be formally characterized by simply relating the number of servers to the offered load. This will be now done within the framework of Erlang-A, following [11]. (One could do it also for Erlang-C, following [16]. The resulting regimes would then be somewhat different, notably ED. We chose to focus on Erlang-A as it is more applicable to call centers.)

- **Efficiency Driven (ED) Regime:**

The efficiency driven regime is characterized by very high servers utilization (close to 100%) and relatively high abandonment rate (10% or more). In the ED regime, the offered load $R = \lambda/\mu$ is noticeably larger than the number of agents N. This means that the system would explode unless abandonment take place. The formal characterization of the ED regime is in terms of the following relationship between N and R:

$$N \approx R(1 - \epsilon),$$

where $0 < \epsilon < 1$ is a QoS parameter: a larger value of ϵ implies longer waiting times and more abandonment.

- **Quality Driven (QD) Regime:**

In the quality-driven regime the emphasis is given to customers' service quality. This regime is characterized by relatively low servers utilizations (for large call centers below 90%, and for moderate ones around 80% and perhaps less) and very low abandonment rate. Formally, this regime is characterized by:

$$N \approx R(1 + \epsilon).$$

- **Quality and Efficiency Driven (QED) Regime:**

This regime is the most relevant for call centers operation. It combines a relatively high utilization of servers (around $90 - 95\%$) and low abandonment rate ($1\% - 3\%$). Because of its importance, and for historical perspective, we present here both Erlang-C and Erlang-A in the QED regime.

Erlang-C

The Erlang-C QED regime goes back as early as Erlang [7], where he derived it via marginal analysis of the benefit of adding a server. This regime is characterized by the *square-root safety-staffing rule*, which we now describe. (Erlang indicated that the rule had been practiced actually since 1913.)

Let $R = \lambda/\mu$ denote the *Offered Load*. Then the square root safety-staffing rule states the following: for moderate to large values of R, the appropriate staffing level is of the form

$$N = R + \beta \cdot \sqrt{R}, \tag{2.16}$$

where β is a positive constant that depends on the desired level of service; β will be referred to as the Quality-of-Service (QoS) parameter: the larger the value of β, the higher is the service quality. The second term on the right side of (2.16) is the excess (safety) capacity, beyond the nominal requirement R, which is needed in order to achieve an accepted service level under stochastic variability.

The form of (2.16) carries with it a very important insight. Denote by Δ the safety staffing level (above the minimum $R = \lambda/\mu$.) Then, if β is fixed, an n-fold increase in the offered load R requires that the safety staffing Δ increases by only$\sqrt{n}$ -fold, which constitutes significant economies of scale.

What does (2.16) guarantee as far as *QoS* is concerned? For Erlang-C, this is the subject of the seminal paper by Halfin and Whitt [16], where they provided the following answer:

Theorem 1 *Consider a sequence of $M/M/N$ queues, indexed by $N = 1, 2, \ldots$. Denote the parameters of the N-th system with a subscript N, for example, $R_N = \lambda_N/\mu$, $\rho_N = R_N/N$. Then, as the number of servers N grows to infinity, the square-root safety-staffing rule applies asymptotically if and only if the delay probability converges to a constant α ($0 < \alpha < 1$), in which case the relation between α and β is given by the Halfin-Whitt function:*

$$\alpha = [1 + \frac{\beta}{h(-\beta)}]^{-1}; \quad 0 < \beta < \infty; \tag{2.17}$$

$h(t)$ is the hazard rate of the Standard Normal Distribution.

Note that (2.16) applies if and only if $\sqrt{\rho_N}(1 - \rho_N)$ converges to β ($\beta > 0$). Indeed, formally the Theorem of Halfin-Whitt reads:

$$\text{As } N \uparrow \infty, \quad P_N(W_q > 0) \equiv E_{2,N} \;\; \rightarrow \;\; \alpha, \qquad (0 < \alpha < 1) \tag{2.18}$$

$$\text{iff } \sqrt{N}(1 - \rho_N) \;\; \rightarrow \;\; \beta, \qquad (\beta > 0) \tag{2.19}$$

$$\text{equivalently } N \;\; \approx \;\; R_N + \beta\sqrt{R_N}. \tag{2.20}$$

The square-root staffing-safety rule was thoroughly analyzed in [38], which was based on [16]. In practice, this rule makes the life of a call-center manager easier: he or she can actually specify the desired delay probability and achieve it by following the square-root safety staffing rule (2.16), simply choosing the right β.

Erlang-A

The Erlang-A analogue of Theorem 1 was proved in [11], and it is given as follows:

Theorem 2 *Consider a sequence of $M/M/N + M$ queues, indexed by $N = 1, 2, \ldots$ As the number of servers N grows to infinity, the square-root safety-staffing rule (2.16) applies asymptotically if and only if the delay probability converges to a constant α ($0 < \alpha < 1$), in which case the relation between α and β is given by the Garnett function*

$$\alpha = \Big[1 + \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{h(-\beta)}\Big]^{-1}, \qquad -\infty < \beta < \infty, \tag{2.21}$$

where $\hat{\beta} = \beta\sqrt{\mu/\theta}$.
Moreover, the above conditions apply if and only if $\sqrt{N}P_N(Aband)$ *converges to some positive constant* γ *which is given by*

$$\gamma = \alpha\beta\left[\frac{h(\hat{\beta})}{\hat{\beta}} - 1\right]^{-1}. \tag{2.22}$$

Formally, Theorem 2 reads:

$$\begin{aligned} \textit{As } N \uparrow \infty, \quad P_N(W_q > 0) &\rightarrow \alpha, \qquad (0 < \alpha < 1); && (2.23)\\ \textit{iff } \sqrt{N}(1 - \rho_N) &\rightarrow \beta, \qquad (-\infty < \beta < \infty); && (2.24)\\ \textit{iff } N &= R_N + \beta\sqrt{R_N} + o(\sqrt{R_N}); && (2.25)\\ \textit{iff } \sqrt{N}P_N(Aband) &\rightarrow \gamma, \qquad (0 < \gamma < \infty). && (2.26) \end{aligned}$$

An important feature of Erlang-A is that, unlike Erlang-C, it is always stable whenever the abandonment rate θ is positive.

Theorem 2 demonstrates that the square-root safety-staffing rule prevails for Erlang-A as well. The *QoS* parameter β now depends on both the abandonment rate θ and the delay probability α. It is significant that here β may take also negative values (since Erlang-A is always stable).

The analysis of the $M/M/N + M$ model can be extended to $M/M/N + G$, in which the distribution of customers' patience takes a general form. The exact analysis of $M/M/N + G$ was first performed in [3].

Zeltyn [41] extended some of the results in [3], and then continued with asymptotic analysis, as the number of servers N grows indefinitely. One of the main outcomes of [41] is the analogue of Theorem 2 for $M/M/N + G$. It applies when patience has a positive density at the origin, say $g(0) > 0$. Then it follows that the asymptotic performance measures for $M/M/M+G$ are exactly those for $M/M/N+M$ but with only substituting $g(0)$ for θ.

Remark 1 *The above characterizations of the operating regimes are insightful, yet they are essentially structural in the sense that the precise values of the QoS parameters remain unspecified. For Eralng-C, the specifications of these values were carried out in [4], based on economic considerations that trade off delay costs against servers' salaries. For Erlang-A, this is done in [27].*

Remark 2 *From Theorems 1 and 2 follows that a characterization of the QED regime could be:*

$$\textit{QED:} \qquad \lim_{N\to\infty} P_N(W_q > 0) = \alpha, \quad 0 < \alpha < 1.$$

Along these lines, one can characterize also the ED and QD regimes, namely:

$$\textit{ED:} \qquad \lim_{N\to\infty} P_N(W_q > 0) = 1;$$

$$\textit{QD:} \qquad \lim_{N\to\infty} P_N(W_q > 0) = 0.$$

Interestingly, the performance-measure $P_N(W_q > 0)$*is rarely tracked in practise.*

Chapter 3

Markovian N-Server Queues: Analysis through Excursions

We are beginning this chapter by presenting a technique that makes it possible to find the stationary/limit distribution of any Markov process restricted to some subset of states A, if the stationary/limit distribution of the process defined on the entire set of states S is given. Then we show how this technique applies to reversible processes. In Section 3.1, we show how, by using the distribution of the reversible restricted processes, one can calculate the expected duration of any excursion of the original process. Section 3.2 presents closed-form expressions of the Delay Probability for our Markovian queues in terms of busy and idle periods. Through these expressions of the Delay Probability, we are going to identify Erlang-C and Erlang-B queues as being two extreme forms of Erlang-A. Finally, by using the developed expression, we analyze the behavior of the Delay Probability under ED, QED and QD regimes.

3.1 Restriction to a Set via Time-Change

Here we are presenting the Time-Change technique. This method will be used in the following sections for getting a new expression of the Delay Probability which enables easy analysis.

The idea of the technique is the following. Consider a general Markov process X with some steady-state distribution given: $X(\infty) \stackrel{d}{=} \pi$. Denote by A a subset of states of X. Let us construct a new process $X_A(t)$ in the following way. We observe the original process X only at times when it is within the subset of A. To formalize this, define:

$$\mathcal{L}(t) = \int_0^t 1_{\{X(u)\in A\}} du.$$

Note that $\mathcal{L}(t)$ is the entire time up to t that X spends in the states of A.

Now define *X restricted to A*, which we denote by X_A, as the process $X_A = \{X_A(t), t \geq 0\}$ given by

$$X_A(t) = X\left(\mathcal{L}^{-1}(t)\right) \equiv X\left(\tau_A(t)\right), \quad t \geq 0.$$

Here $\tau_A(t)$ is a right-continuous inverse of $\mathcal{L}(t)$, namely

$$\tau_A(t) \triangleq \inf\{s:\ \mathcal{L}(s) > t\}. \tag{3.1}$$

Theorem 3 *Let X be an ergodic irreducible right-continuous left-limit Markov process on the discrete set of states S, with a limitting (stationary) distribution π. Then, the limitting distribution of the restricted process X_A is the restriction of π to A. Formally,*

$$X_A(\infty) \stackrel{d}{=} X(\infty)|X(\infty) \in A.$$

Proof
Consider $X_A(t) = X(\tau_A(t))$, which is X restricted to A.
Define the following:

- $\pi(B) = P(X(\infty) \in B)$,
- $\pi_A(B) = P(X_A(\infty) \in B.)$

We are to show that for any $B \subset A$,

$$\pi_A(B) = \lim_{T\to\infty} P(X(T) \in B|X(T) \in A) \ \equiv \ \frac{\pi(B)}{\pi(A)}.$$

First, we state that X_A is a Markov process. Indeed,

1. It is given that X is a right-continuous left-limit Markov process;
2. $\mathcal{L}(t) = \int_0^t 1_{\{X(u)\in A\}} du$ is a continuous additive functional.

Consequently, as shown in [34], if τ_A is defined by (3.1), then $X_A(t) = X(\tau_A(t))$ is also a right-continuous left-limit Markov process.

It is known (see [34]), that X is ergodic if and only if

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(u))du = \sum_{j\in\Omega} \pi(j) f(j),$$

for any function f, such that the right-hand-side of the above is well defined.

Thus, the calculation of the limit $\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(\tau_A(u)))du$ will allow us to conclude whether the restricted process X_A is ergodic, and if so, to identify its stationary/limitting distribution.

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(\tau_A(u)))du =$$

$$\text{change of variables:} \qquad s = \tau_A(u) \quad \Rightarrow \quad du = d\mathcal{L}(s).$$

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t} \int_0^{\tau_A(t)} f(X(s)) d\mathcal{L}(s) &= \\
&= \lim_{t\to\infty} \frac{\tau_A(t)}{t} \frac{1}{\tau_A(t)} \int_0^{\tau_A(t)} f(X(s)) \mathbf{1}_A(X(s)) ds;
\end{aligned}$$

$$\lim_{t\to\infty} \frac{\tau_A(t)}{t} = \lim_{t\to\infty} \frac{t}{\mathcal{L}_A(t)} = \frac{1}{\pi(A)},$$

$$\lim_{t\to\infty} \frac{1}{\tau_A(t)} \int_0^{\tau_A(t)} f(X(s)) \mathbf{1}_A(X(s)) ds = \sum_{j\in A} \pi(j) f(j).$$

The last two statements follow from the ergodicity of *X*.

$$\Rightarrow \quad \lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(\tau_A(u))) du = \frac{\sum_{j\in A} \pi(j) f(j)}{\pi(A)}.$$

We have shown that $\lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(\tau_A(u)))du$ exists and it is finite. As a result, the restricted Markov process is ergodic with the limitting/stationary distribution

$$\pi_A(B) = \frac{\pi(B)}{\pi(A)}, \qquad \text{for any} B \subseteq A. \qquad \square \tag{3.2}$$

Theorem 3 can be applied to any ergodic Markov process X. If, in addition, X is reversible, the distribution of any subprocess restricted to a set A can be actually calculated by just omitting all the states that are not included in A from its transition diagram.

Definition 1 *Let $X = \{X(t), t \geq 0\}$ be a Markov process on a discrete state space. Denote its transition rates by $[q_{ij}]$. Let A be a subset of states of X and define the Kelly process X_A^K over A in terms of its transition rates (for $i \neq j$):*

$$q_{i,j}^K = \begin{cases} q_{i,j} & \text{if} \quad i,j \in A \\ 0 & \text{otherwise} \end{cases}.$$

Definition 2 *Consider a Markov process with a stationary distribution $\{\pi_i\}$. Then, this Markov process is called reversible if the transition rates between each pair of states* i *and* j *in the state space obey*

$$q_{i,j}\pi_i = q_{j,i}\pi_j,$$

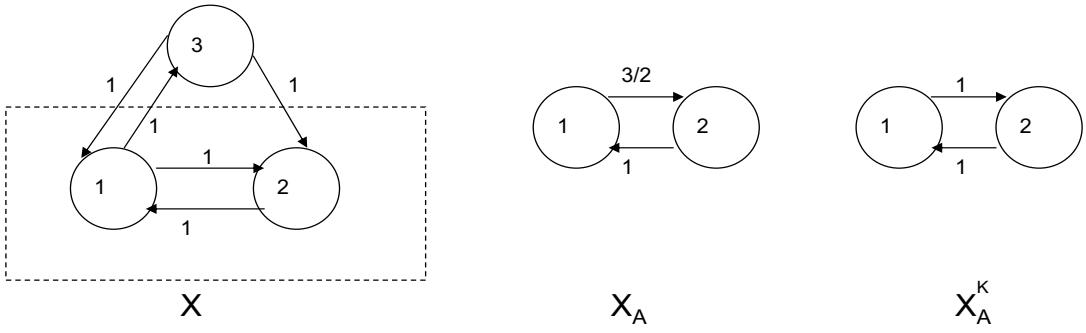
where $q_{i,j}$ is the transition rate from state i *to state* j *and π_i and π_j are the stationary probabilities of being in states* i *and* j*, respectively.*

Theorem 4 *Let $X = \{X(t), t \geq 0\}$ be a reversible Markov process. Then the stationary distribution of the restricted process X_A is equal to the stationary distribution of the Kelly process X_A^K.*

The complete proof of Theorem 4 can be found in [22], hence the terminology for Kelly process. Below we present a proof of this theorem for a special case, which is important for our further research. But before that, we show an example of a non-reversible process and conclude that the distribution of the Kelly process defined over some set of states A differs from that of the process restricted to the same set.

Example 1 *Consider the Markov process $X = \{X(t), t \geq 0\}$ defined over the states $\Omega = \{1, 2, 3\}$. The transition rate diagram of X is presented in Figure 3.1. Let us analyze the Kelly process X_A^K on the set $A = \{1, 2\}$ and X_A restricted to the same set A. We are to show that these two processes have a different distribution.*

Figure 3.1: Example - Non-Reversible Process



Proof

Both X_A and X_A^K are also presented in Figure 3.1. The diagram of the Kelly process X_A^K is obtained by "erasing" all the paths outgoing from A.

In the restricted process X_A the duration of stay in state $\{1\}$ is found as the geometrical sum of exponential times: $\sum_i^N exp(2)$, where $N \stackrel{d}{=} Geo(\frac{3}{4})$.
Hence, the total time that X_A spends in state $\{1\}$ is exponential $exp(2 \cdot \frac{3}{4} = 1.5)$.
The time of the restricted process X_A in state $\{2\}$ is $exp(1)$.

As seen from the transition-rate diagram (Figure 3.1), the restricted process X_A defined by set A, and the Kelly process X_A^K are not equal in distribution. $\square$

We have just shown that a Kelly process is not neccessarily equal in distribution to the corrsonding restricted process. Now let us consider a special case, which is important for our further research, and prove that Theorem 4 does apply for it. The case we are concentrating on is the calculation of the delay probability for $M/M/N + M$ via the Erlang-B formula, after observing that $M/M/N/N$ is the restriction of $M/M/\infty$ to the set of its first N+1 states (from 0 to N).

Example 2 *Consider an $M/M/\infty$ queue with the arrival rate λ and the service rate μ. Let $L(t)$ denote the number of customers in the system at time $t \geq 0$, and let the vector π be the steady-state distribution of L: $L(\infty) \stackrel{d}{=} \pi$.*

Define L_- to be L restricted to states $\{0, 1, \ldots, N\}$.

Then $L_- \stackrel{d}{=} M/M/N/N$, with the arrival rate λ, service rate μ and

$$P(L_-(\infty) = N) = \frac{\pi(N)}{\sum_{i=0}^{N} \pi(i)} = E_{1,N}$$

Proof
For L_-, the duration of a visit in a state $i, \quad i \in \{0, 1, \ldots, N-1\}$, is $\exp(\lambda + i\mu)$, as in the original $M/M/\infty$ queue.

Each time the original queue L reaches state N, it is followed by state $N+1$ with probability $\frac{\lambda}{\lambda+N\mu}$, or by state $N-1$ with probability $\frac{\mu}{\lambda+N\mu}$. Whenever the original process L starts moving to $N+1$, it leaves the restricted set, and the time of L_- then stops advancing. In this case, the only possible way to return to the restricted set is through the state of N.

The duration of stay in state N of L_- depends on the series of visits of L in state N. The duration of each visit is distributed exponentially $\exp(r = \lambda + N\mu)$. The number of such visits is distributed geometrically $Geo(p = \frac{N\mu}{\lambda+N\mu})$. Now, one deduces that the visit time of L_- in state N is distributed exponentially $\exp(p \cdot r = N\mu)$, being a geometric sum of i.i.d. exponentials.

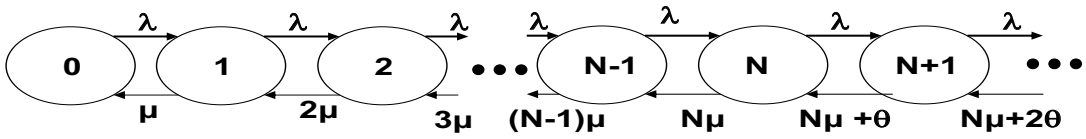
To conclude, the restricted process L_- has the same distribution as $M/M/N/N$:

$$L_- \stackrel{d}{=} M/M/N/N \qquad \square$$

3.2 Up/Down Crossings - The Erlang-A formula

Consider an $M/M/N+M$ queueing system with arrival rate λ, service rate μ, abandonment rate θ and total number of servers N. Let $L(t)$ be the total number of customers at time $t \geq 0$. Then $L = \{L(t), t \geq 0\}$ is a Birth-and-Death process with the transition rate diagram depicted in Figure 3.2.

Figure 3.2: Erlang-A – Transition-Rate Diagram



Let us define the following:

- $L_- = \{L_-(t), t \geq 0\}$ is the number of customers in the system where at least one server is idle. The process L_- is L restricted to states $\{0, 1, \ldots, N-1\}$, so its distribution is identical to an $M/M/N-1/N-1$ queue (Erlang-B) with arrival rate λ and service rate μ.

- $L_+ = \{L_+(t), t \geq 0\}$ is the number of customers in the system where all the servers are busy. The process L_+ is L restricted to states $\{N, N+1, \ldots\}$. Its distribution can be described by a Birth-and-Death process which resembles an $M/M/\infty$ queue with arrival rate λ, with the only difference being that the first server starts a busy period works at rate $N\mu$ and each additional server joins with rate θ, so that the total service rate at each state $i \in \{N+1, \ldots\}$ is $\mu_i = N\mu + (i-N)\theta$.

The total number of customers L alternates between L_+ and L_-.

- Let $T_{N-1,N}$ be the expected duration of an L **idle period**. Formally, given that L starts at state $N-1$, $T_{N-1,N}$ is the expectation of the first hitting time of state N.

- Let $T_{N,N-1}$ be the expected duration of an L **busy period**. Formally, given that L starts at state N, $T_{N,N-1}$ is the expectation of the first hitting time of state $N-1$.

The delay probability can be found via PASTA from the following relation:

$$P(W_q > 0) = \frac{T_{N,N-1}}{T_{N,N-1} + T_{N-1,N}} = \Big[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\Big]^{-1}. \tag{3.3}$$

To find $T_{N,N-1}$ and $T_{N-1,N}$, we use the following relation, observed by Whitt [39] (a proof will be provided momentarily):

$$T_{N-1,N} = \frac{1}{\lambda_{N-1}\pi_-(N-1)} = \frac{1}{\lambda E_{1,N-1}}, \tag{3.4}$$

$$T_{N,N-1} = \frac{1}{\mu_N \pi_+(0)}. \tag{3.5}$$

Here $E_{1,N-1}$ is the Erlang-B Loss Probability given by

$$\begin{aligned} E_{1,N} &= \frac{R^N}{N!} \Big/ \sum_{k=0}^{N} \frac{R^k}{k!}, \quad \text{where } R = \lambda/\mu \text{ is the Offered Load,} \\ &= \frac{P(Y=N)}{P(Y \leq N)}, \quad \text{where } Y \stackrel{d}{=} Pois(R), \\ &= P(Y = N | Y \leq N); \end{aligned}$$

$\pi_+(0)$ denotes the stationary probability that L_+ is at state N. This probability is expressed by (3.6)-(3.8), and its calculation is presented in Appendix 3.5.1:

$$\pi_+(0) = \frac{(\lambda/\theta)^{N\mu/\theta} \Big/ (N\mu/\theta)!}{\sum_{i=0}^{\infty} (\lambda/\theta)^{N\mu/\theta+i}/(N\mu/\theta+i)!} \tag{3.6}$$

$$= \frac{P(X = N\mu/\theta)}{P(X \geq N\mu/\theta)}, \qquad where \quad X \stackrel{d}{=} Poisson(\lambda/\theta), \tag{3.7}$$

$$= P(X = N\mu/\theta | X \geq N\mu/\theta). \tag{3.8}$$

In the above writing, we assume that $N\mu/\theta$ is an integer. However, for further asymptotical analysis, and as will be shown in Appendix 3.5.1, this assumption is unnecessary since $\pi_+(0)$ can be re-expressed in the terms of special functions (see [41]):

$$\pi_+(0) \quad = \quad \frac{(\lambda/\theta)^{N\mu/\theta}}{[N\mu/\theta]e^{\lambda/\theta}\gamma(\frac{N\mu}{\theta},\frac{\lambda}{\theta})}, \tag{3.9}$$

where

$$\gamma(x,y) = \int_0^y t^{x-1}e^{-t}dt, \qquad x>0, \quad y\geq 0.$$

Equation (3.4) can be proved easily as follows:

Note that the expected duration of a single Idle Excursion is $\frac{1}{\pi_-(N-1)\mu_{N-1}}$. (Idle Excursion refers to an excursion from $N-1$ to $N-1$ without leaving the "idle" states *0, 1, ..., N-1*) The number of such excursions, before the process L_- leaves state $N-1$ (to state *N*), has a Geometric distribution starting from zero, with probability of success $\frac{\lambda_{N-1}}{\lambda_{N-1}+\mu_{N-1}}$. Hence, the expected duration of an Idle period is calculated by the Wald formula:

$$T_{N-1,N} = E(\text{Idle Excursion}) \times E(\#\text{ of Idle Excursions}) = \frac{1}{\pi_-(N-1)\mu_{N-1}} \cdot \frac{\mu_{N-1}}{\lambda_{N-1}}.$$

Following the same approach, one can immediately derive (3.5).

After establishing expressions (3.5) and (3.4) for the expected duration of the busy and idle periods of L, we can substitute them into (3.3) to obtain the following result:

$$P(W_q>0) \quad = \quad \left[1+\frac{\pi_+(0)}{\rho\pi_-(N-1)}\right]^{-1} \tag{3.10}$$

$$\quad = \quad \left[1+\frac{1}{\rho}\frac{P(X=N\mu/\theta|X\geq N\mu/\theta)}{P(Y=N-1|Y\leq N-1)}\right]^{-1}, \tag{3.11}$$

where $X \stackrel{d}{=} Pois(\lambda/\theta)$, $Y \stackrel{d}{=} Pois(\lambda/\mu)$ and $\rho=\frac{\lambda}{N\mu}$ is the Offered Load per server.

3.3 Special Cases

3.3.1 $M/M/\infty$

We would like to begin this section with an analysis of the $M/M/\infty$ queue. Alternatively, we analyze an $M/M/N+M$ queue where the individual customer abandonment rate θ is equal to the service rate μ of a single server. Under this assumption, the idle period $T_{N-1,N}$ of the queue does not change, but the

busy period $T_{N,N-1}$ can be presented in a more elegant way.

By substituting $\theta = \mu$ into (3.11), we immediately obtain the following:

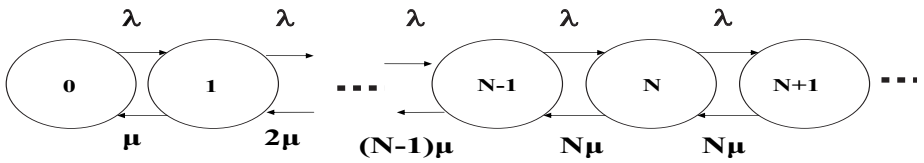
$$P(W_q > 0) = \left[1 + \frac{1}{\rho}\frac{P(Y = N|Y \geq N)}{P(Y = N-1|Y \leq N-1)}\right]^{-1}, \quad Y \stackrel{d}{=} Pois(R). \tag{3.12}$$

$$P(W_q > 0) \approx \left[1 + \frac{P(Y = N|Y \geq N)}{P(Y = N|Y \leq N)}\right]^{-1}, \quad for \;\; N \;\; large \;\; and \;\; \rho \approx 1. \tag{3.13}$$

3.3.2 Erlang-C

Now let us consider an extreme example of the Erlang-A queue. Here we analyze the case with infinite patience, namely the Erlang-C queue. The transition-rate diagram of this queue is presented in Figure 3.3. We will check the limit of $P(W_q > 0)$ in (3.11), as θ converges to 0.

Figure 3.3: Erlang-C - Transition-Rate Diagram



Notice that (3.11) depends on θ only via $\pi_+(0) = P(X = N\mu/\theta | X \geq N\mu/\theta)$. Therefore, we start with checking the convergence of $\lim_{\theta\to 0}\pi_+(0)$.

Lemma 1

$$\lim_{\theta\to 0}\pi_+(0) = 1 - \rho.$$

Proof:

$$\begin{aligned}
\lim_{\theta\to 0}\pi_+(0) &= \lim_{\theta\to 0}\frac{(\lambda/\theta)^{N\mu/\theta}\Big/(N\mu/\theta)!}{\sum_{i=0}^{\infty}(\lambda/\theta)^{N\mu/\theta+i}/(N\mu/\theta+i)!} \\
&= \lim_{\theta\to 0}\frac{1}{\sum_{i=0}^{\infty}(\lambda/\theta)^i/\prod_{j=1}^{i}(\frac{N\mu}{\theta}+j)} \\
&= \lim_{\theta\to 0}\frac{1}{\sum_{i=0}^{\infty}(\lambda/\theta)^i\cdot(\theta/N\mu)^i/\prod_{j=1}^{i}(1+\frac{j\theta}{N\mu})} \\
&= \frac{1}{\sum_{i=0}^{\infty}(\lambda/N\mu)^i}
\end{aligned}$$

$$= \quad 1-\rho. \qquad \square$$

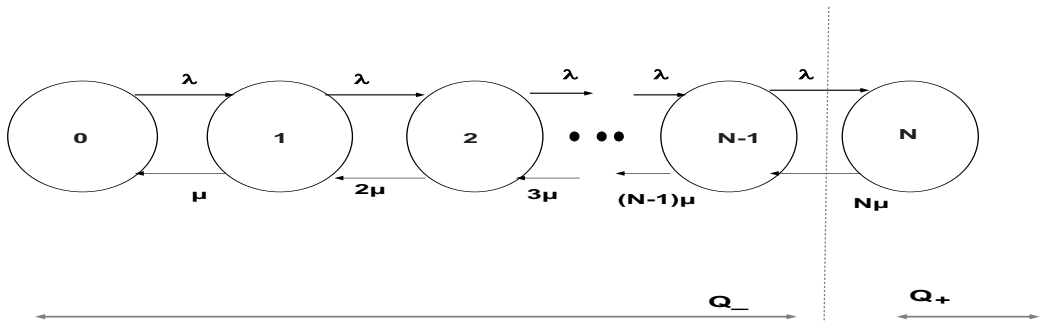
Substituting the result of this last lemma into (3.11) yields the theoretical result, known as the Erlang-C formula, or $E_{2,N}$.

$$\lim_{\theta\to 0} P(W_q > 0) = E_{2,N} = \left[1 + \frac{1-\rho}{\rho\ P(Y = N-1 | Y \leq N-1)}\right]^{-1} = \left[1 + \frac{1-\rho}{\rho\ E_{1,N-1}}\right]^{-1}. \quad (3.14)$$

3.3.3 Erlang-B

Another extreme example of the Erlang-A model is Erlang-B. Here customers do not wish to wait and abandon immediately if there are no servers available upon their arrivals. The transition-rate diagram of this queue is presented in Figure 3.4. Such a queue can be described as an $M/M/N+M$ queue

Figure 3.4: Erlang-B - Transition-Rate Diagram



with infinitely impatient customers ($\theta = \infty$). In the Erlang-B queue no customer waits, so the notation $P(W_q > 0)$ in (3.11) is not meaningful. Hence, we will denote this probability by $P(L(t) \geq N)$, where $L(t)$ is the number of customers in the system at time t.

Now we will check the convergence of (3.11) as θ diverges to ∞. Observe that in (3.11) only $\pi_+(0)$ defined in (3.6) depends on θ, so we begin with calculating the convergence of $\lim_{\theta\to\infty} \pi_+(0)$.

Lemma 2 *Let* $\pi_+(0) = \dfrac{(\lambda/\theta)^{N\mu/\theta}}{[N\mu/\theta]e^{\lambda/\theta}\gamma(\frac{N\mu}{\theta}, \frac{\lambda}{\theta})}$ *(See 3.9). Then*

$$\lim_{\theta\to\infty} \pi_+(0) = 1.$$

Proof

We are interested in evaluating the following limit:

$$\lim_{\theta\to\infty} \frac{(\lambda/\theta)^{N\mu/\theta}\exp(-\lambda/\theta)}{\frac{N\mu}{\theta}\gamma(\frac{N\mu}{\theta}, \frac{\lambda}{\theta})}.$$

The Numerator

The limit of the numerator is found by using the L'Hospital rule:

$$\lim_{\theta\to\infty} (\lambda/\theta)^{N\mu/\theta} \exp(-\lambda/\theta) \quad = \quad 1,$$

The Denominator

To find the limit of the denominator, we use the recursive presentation of the Incomplete Gamma Function (See [1]):

$$\gamma(a+1,x) = a\gamma(a,x) - x^a \exp(-x). \tag{3.15}$$

This recursive representation (3.15) leads to the following conclusion:

$$\lim_{\theta\to\infty} \frac{N\mu}{\theta}\gamma(\frac{N\mu}{\theta},\frac{\lambda}{\theta}) \quad = \quad \lim_{\theta\to\infty}\left[\gamma(\frac{N\mu}{\theta}+1,\frac{\lambda}{\theta}) + (\lambda/\theta)^{N\mu/\theta}\exp(-\lambda/\theta)\right] = 1,$$

since

$$\lim_{\theta\to\infty} \gamma(\frac{N\mu}{\theta}+1,\frac{\lambda}{\theta}) = \lim_{\theta\to\infty}\left[\Gamma(\frac{N\mu}{\theta}+1) - \Gamma(\frac{N\mu}{\theta}+1,\frac{\lambda}{\theta})\right] = 0.$$

The statement of Lemma 2 is now proven. □

By using Lemma 2 it immediately follows from (3.11) that

$$\begin{aligned}
\lim_{\theta\to\infty} P(L(t)\geq N) &= \left[1 + \frac{1}{\lambda/N\mu P(Y=N-1|Y\leq N-1)}\right]^{-1} \\
&= \left[1 + \frac{P(Y\leq N-1)}{\lambda/N\mu P(Y=N-1)}\right]^{-1} \\
&= \left[\frac{P(Y=N)+P(Y\leq N-1)}{P(Y=N)}\right]^{-1} \\
= \frac{P(Y=N)}{P(Y\leq N)} &= P(Y=N|Y\leq N) \\
&= E_{1,N}.
\end{aligned}$$

In the above, we used the relation

$$\frac{\lambda}{N\mu}P(Y=N-1) \quad = \quad \frac{R}{N}\cdot\frac{e^{-R}R^{N-1}}{(N-1)!} = P(Y=N).$$

This result gives rise to the following insight: as the customers' impatience θ grows to infinity, $P(L(t)\geq N)$ converges to $P(Blocked) = E_{1,N}$. This means that $P(L(t) > N)$ converges to 0, i.e. the model never gets to states $i > N$. Consequently, Erlang-B is indeed an extreme example of Erlang-A, as the impatience grows indefinitely.

3.4 Asymptotic Analysis

This section presents an asymptotic analysis of the Delay Probability when the number of servers is large, under the three operational regimes QED, ED and QD.

The convergence of the Delay Probability in these operational regimes is well-known. Yet, the present analysis gives some interesting insight and shows not only the final limiting values but also different components that have their impact both on the final limit and on the rate of convergence to it.

The section is divided into three subsections. Each subsection presents a different regime, for which we evaluate the convergence of the Delay Probability (3.11):

$$\lim_{N\to\infty} P(W_q > 0) = \lim_{N\to\infty} \left[1 + \frac{1}{\rho}\frac{\pi_+(0)}{E_{1,N-1}}\right]^{-1}. \tag{3.16}$$

It has been shown above that $\pi_+(0)$ has a probabilistic representation under the assumption that $N\mu/\theta$ is an integer:

$$\pi_+(0) = P(X = N\mu/\theta | X \geq N\mu/\theta) \equiv \frac{P(X = N\mu/\theta)}{P(X \geq N\mu/\theta)}, \qquad \text{where} \quad X \stackrel{d}{=} Poisson(\lambda/\theta).$$

In addition, note that $\lim_{N\to\infty} E_{1,N-1} = \lim_{N\to\infty} E_{1,N}$, hence in our subsequent asymptotic analysis we calculate $\lim_{N\to\infty} E_{1,N}$ instead of $\lim_{N\to\infty} E_{1,N-1}$.

Recall that $E_{1,N}$ has a probabilistic representation as well:

$$E_{1,N} = P(Y = N | Y \leq N) \equiv \frac{P(Y = N)}{P(Y \leq N)}, \qquad \text{where} \quad Y \stackrel{d}{=} Poisson(R_N \equiv \lambda/\mu).$$

In each of the following three subsections, the analysis of the Delay Probability is conducted along five steps. The first four are devoted to the procedures of limit calculations, while the last step summarizes all the procedures.

- **Step 1** $\lim_{N\to\infty} P(Y \leq N)$;
- **Step 2** $\lim_{N\to\infty} P(Y = N)$;
- **Step 3** $\lim_{N\to\infty} P(X \geq N\mu/\theta)$;
- **Step 4** $\lim_{N\to\infty} P(X = N\mu/\theta)$;
- **Step 5** Merge the results of **Steps 1-4** and calculate the Delay Probability limit (3.16).

3.4.1 QED Regime: The Garnett Function

Let us now analyze the convergence of the Delay Probability (3.16) in the QED regime. Here, the arrival rate λ grows to infinity in such way that the **number of severs** N is expressed by the square-root staffing rule:

$$N \approx R_N + \beta\sqrt{R_N}, \tag{3.17}$$

where $-\infty < \beta < \infty$ is a QoS parameter and $R_N = \frac{\lambda}{\mu}$ is the Offered Load.

Step 1

$$\lim_{N\to\infty} P(Y \le N) = \lim_{N\to\infty} P\left(\frac{Y - R_N}{\sqrt{R_N}} \le \frac{N - R_N}{\sqrt{R_N}}\right)$$

$$\Rightarrow \lim_{N\to\infty} P(Y \le N) = P(Z \le \beta) = \Phi(\beta), \qquad where \qquad Z \stackrel{d}{=} N(0,1),$$

assuming that $\lim_{N\to\infty} \frac{Y-R_N}{\sqrt{R_N}} = \beta$, for $-\infty < \beta < \infty$, which holds in the QED regime. Here we are using the normal approximation to the Poisson distribution: $\frac{(Poisson(\lambda)-\lambda)}{\sqrt{\lambda}}$ converges in distribution to $N(0,1)$, as $\lambda \uparrow \infty$.

Step 2

$$\begin{aligned}
\lim_{N\to\infty} P(Y = N) &= \lim_{N\to\infty} P\left(\frac{N - R_N - 1}{\sqrt{R_N}} < \frac{Y - R_N}{\sqrt{R_N}} \le \frac{N - R_N}{\sqrt{R_N}}\right) \\
\Rightarrow \lim_{N\to\infty} \sqrt{R_N}\, P(Y = N) &= \lim_{N\to\infty} \sqrt{R_N}\, P(\beta - \frac{1}{\sqrt{R_N}} < Z \le \beta) \\
&= \lim_{N\to\infty} \sqrt{R_N}\left(\Phi(\beta) - \Phi(\beta - \frac{1}{\sqrt{R_N}})\right) \\
&= \phi(\beta),
\end{aligned}$$

i.e. $\lim_{N\to\infty} P(Y = N)$ converges to 0 at rate $\Theta(1/\sqrt{R_N})$.

Combining the results of Steps 1 and 2, one concludes that

$$\lim_{N\to\infty} \sqrt{R_N} \cdot E_{1,N} = h(-\beta), \tag{3.18}$$

where $h(\cdot)$ is the hazard rate of the Standard Normal Distribution.

Step 3

$$\lim_{N\to\infty} P\Big(X \ge N\mu/\theta\Big) = \lim_{N\to\infty} P\Big(\frac{X - \lambda/\theta}{\sqrt{\lambda/\theta}} \ge \frac{N\mu/\theta - \lambda/\theta}{\sqrt{\lambda/\theta}}\Big)$$

In order to analyze the convergence of this expression, we need to know what happens to $\lim_{N\to\infty} \dfrac{N\mu/\theta - \lambda/\theta}{\sqrt{\lambda/\theta}}$ in the QED regime.

Lemma 3

$$\lim_{N\to\infty} \frac{(N\mu - \lambda)/\theta}{\sqrt{\lambda/\theta}} = \beta\sqrt{\mu/\theta} \qquad \Longleftrightarrow \qquad \lim_{N\to\infty} \sqrt{N}(1 - \rho_N) = \beta.$$

Proof

$$\begin{aligned}\lim_{N\to\infty}\frac{(N\mu-\lambda)/\theta}{\sqrt{\lambda/\theta}} &= \lim_{N\to\infty}\left(\sqrt{\frac{\theta}{\lambda}}\frac{(N\mu-\lambda)}{\theta}\right)\\ = \lim_{N\to\infty}\left(\frac{N\mu(1-\lambda/N\mu)}{\sqrt{\lambda\theta}}\right) &= \lim_{N\to\infty}\left(\sqrt{N}(1-\lambda/N\mu)\sqrt{\frac{N\mu}{\lambda}}\sqrt{\frac{\mu}{\theta}}\right)\\ &= \beta\sqrt{\mu/\theta}.\end{aligned}$$

□

Lemma 3 allows one to deduce the convergence of $P\Big(X \geq N\mu/\theta\Big)$ in the QED regime:

$$\lim_{N\to\infty} P\Big(X \geq N\mu/\theta\Big) = 1 - P(Z \geq \hat{\beta}) = 1 - \Phi(\hat{\beta}), \quad where \quad Z \stackrel{d}{=} N(0,1), \quad \hat{\beta} = \beta\sqrt{\mu/\theta}.$$

Step 4

$$\begin{aligned}P(X = N\mu/\theta) &= P\left(\frac{N\mu-\lambda_N/\theta-1}{\sqrt{\lambda_N/\theta}} < \frac{X-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} \leq \frac{N\mu-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}}\right)\\ \Rightarrow \lim_{N\to\infty}\sqrt{\lambda_N}\,P(X = N\mu/\theta) &= \lim_{N\to\infty}\sqrt{\lambda_N}\,P(\hat{\beta}-\frac{1}{\sqrt{\lambda_N/\theta}} < Z \leq \hat{\beta})\\ &= \lim_{N\to\infty}\sqrt{\lambda_N}\left(\Phi(\hat{\beta})-\Phi(\hat{\beta}-\frac{1}{\sqrt{\lambda_N/\theta}})\right)\\ &= \sqrt{\theta}\cdot\phi(\hat{\beta}).\end{aligned}$$

From Steps 3 and 4 we conclude that

$$\lim_{N\to\infty}\sqrt{\lambda_N}\cdot\pi_+^N(0) = \sqrt{\theta}\cdot h(\hat{\beta}). \tag{3.19}$$

Step 5

For further investigation of the Delay Probability, we need the following lemma.

Lemma 4 *In the QED regime, the offered load per server converges to 100 %.*
Formally,

$$\textit{if} \quad (N-R_N)/\sqrt{R_N} \to \beta, \quad -\infty < \beta < \infty, \qquad \textit{then} \qquad \rho_N \to 1.$$

Basing the calculation of the limitting Delay Probability on Lemma 4 and results (3.18) and (3.19), we derive the following:

$$\begin{aligned}\lim_{N\to\infty} P(W_q > 0) &= \lim_{N\to\infty}\left[1+\frac{\pi_+(0)}{\rho E_{1,N}}\right]^{-1}\\ &= \lim_{N\to\infty}\left[1+\frac{\sqrt{\frac{\theta}{\lambda_N}}\,h(\hat{\beta})\sqrt{\frac{\lambda_N}{\mu}}}{\rho\,h(-\beta)}\right]^{-1}\end{aligned}$$

$$= \left[1 + \frac{h(\hat{\beta})}{\sqrt{\frac{\mu}{\theta}}\ h(-\beta)}\right]^{-1}$$
$$= \left[1 + \frac{h(\hat{\beta})/\hat{\beta}}{h(-\beta)/\beta}\right]^{-1}.$$

Note that the last expression is exactly the Garnett formula (see Theorem 2) for the Delay Probability [11].

3.4.2 Efficiency-Driven (ED) Regime

In the Efficiency-Driven regime, the number of servers is determined by the following staffing rule:

$$N = \lambda_N/\mu - \epsilon\lambda_N/\mu, \qquad \text{where} \quad \epsilon > 0. \tag{3.20}$$

Let us check the convergence of (3.16) under this regime by performing the steps described at the beginning of this section.

Step 1

$$\begin{aligned}
\lim_{N\to\infty} P(Y \le N) &= \lim_{N\to\infty} P\left(\frac{Y - R_N}{\sqrt{R_N}} \le \frac{N - R_N}{\sqrt{R_N}}\right) \\
&= \lim_{N\to\infty} P\left(Z \le -\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right), \qquad \text{where} \quad Z \stackrel{d}{=} N(0,1) \\
&= \lim_{N\to\infty} P\left(Z \ge \epsilon\sqrt{\frac{\lambda_N}{\mu}}\right), \\
= \lim_{N\to\infty} \frac{\phi(\epsilon\sqrt{\frac{\lambda_N}{\mu}})}{\epsilon\sqrt{\frac{\lambda_N}{\mu}}} &= 0.
\end{aligned}$$

In the last line we use the equality $\lim_{a\to\infty} P(Z \ge a) = \lim_{a\to\infty} \phi(a)/a$, which can be obtained through L'Hospital's rule.

Step 2

$$\begin{aligned}
\lim_{N\to\infty} P(Y = N) &= \lim_{N\to\infty} P\left(\frac{N - R_N}{\sqrt{R_N}} - \sqrt{\frac{\mu}{\lambda_N}} < \frac{Y - R_N}{\sqrt{R_N}} \le \frac{N - R_N}{\sqrt{R_N}}\right) \\
\Rightarrow \lim_{N\to\infty} \sqrt{\frac{\lambda_N}{\mu}} \frac{P(Y = N)}{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right)} &= \lim_{N\to\infty} \frac{1}{\sqrt{\mu}} \frac{1}{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right)} \sqrt{\frac{\mu}{\lambda_N}}\ \phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right) = 1.
\end{aligned}$$

i.e., in the ED regime, $P(Y = N)$ converges to 0 at rate $\Theta(\frac{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right)}{\sqrt{\frac{\lambda_N}{\mu}}})$.

Combining the results of Steps 1 and 2, we conclude that in the ED regime the blocking probability converges to ϵ:

$$\lim_{N\to\infty} E_{1,N} = \epsilon. \tag{3.21}$$

Step 3

$$\begin{aligned}
\lim_{N\to\infty} P(X \geq N\mu/\theta) &= \lim_{N\to\infty} P\left(\frac{X-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} \geq \frac{N\mu/\theta-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}}\right) \\
&= \lim_{N\to\infty} P\left(Z \geq -\epsilon\sqrt{\frac{\lambda_N}{\theta}}\right), \qquad \text{where} \qquad Z \overset{d}{=} N(0,1) \\
&\Rightarrow \lim_{N\to\infty} P(X \geq N\mu/\theta) = 1.
\end{aligned}$$

Step 4

$$\begin{aligned}
\lim_{N\to\infty} P(X = N\mu/\theta) &= \lim_{N\to\infty} P\left(\frac{N\mu/\theta-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} - \sqrt{\frac{\theta}{\lambda_N}} < \frac{X-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} \leq \frac{N\mu/\theta-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}}\right) \\
&\Rightarrow \lim_{N\to\infty} \frac{\lambda_N}{\phi\Big(-\epsilon\sqrt{\frac{\lambda_N}{\theta}}\Big)} P(X = N\mu/\theta) \\
&= \lim_{N\to\infty} \frac{\lambda_N}{\phi\Big(-\epsilon\sqrt{\frac{\lambda_N}{\theta}}\Big)} \sqrt{\frac{\theta}{\lambda_N}}\, \phi\Big(-\epsilon\sqrt{\frac{\lambda_N}{\theta}}\Big) \\
&= \sqrt{\theta},
\end{aligned}$$

i.e., in the ED regime, $P(X = N\mu/\theta)$ converges to 0 at rate $\Theta\left(\sqrt{\frac{\theta}{\lambda_N}}\, \phi(\epsilon\sqrt{\frac{\lambda_N}{\theta}})\right)$.

Steps 3 and 4 allow us to see what happens to $\pi_+^N(0)$ in the ED regime:

$$\lim_{N\to\infty} \frac{\sqrt{\lambda_N/\theta}}{\phi(\epsilon\sqrt{\lambda_N/\theta})} \cdot \pi_+^N(0) \quad = 1. \tag{3.22}$$

Step 5

In the ED regime, the offered load per server is constant and exceeds 100 %:

$$\lim_{N\to\infty} \rho_N = \frac{1}{1-\epsilon}.$$

This observation and the intermediate results (3.21 and 3.22) show that in the ED regime, the Delay Probability (3.16) converges to 1, which is in line with known results.

$$\lim_{N\to\infty} P(W_q > 0) = \lim_{N\to\infty} \left[1 + (1-\epsilon)\, \frac{h(-\epsilon\sqrt{\frac{\lambda_N}{\theta}})}{\epsilon\sqrt{\frac{\lambda_N}{\theta}}}\right]^{-1} = 1. \tag{3.23}$$

Let us define $\gamma_N = \epsilon\sqrt{\frac{\lambda_N}{\theta}} = \epsilon\sqrt{\frac{N(1-\epsilon)}{\theta}}$.
Then the delay probability (3.23) converges to 1 at the following rate:

$$\Theta\left(\frac{\phi(\gamma_N)}{\gamma_N}\right). \tag{3.24}$$

Recall that this last statement means that

$$\lim_{N\to\infty} \frac{P(W_q > 0) - 1}{\phi(\gamma_N)/\gamma_N} = 1.$$

3.4.3 Quality Driven (QD) Regime

In the Quality Driven regime, the number of servers is determined by the following rule:

$$N = \lambda_N/\mu + \epsilon\lambda_N/\mu \qquad \text{where} \quad \epsilon > 0. \tag{3.25}$$

Here we check the convergence of the Delay Probability given by (3.16) in this regime.
Step 1

$$\begin{aligned}\lim_{N\to\infty} P(Y \le N) &= \lim_{N\to\infty} P\left(\frac{Y - R_N}{\sqrt{R_N}} \le \frac{N - R_N}{\sqrt{R_N}}\right) \\ &= \lim_{N\to\infty} P\left(Z \le \epsilon\sqrt{\frac{\lambda_N}{\mu}}\right) = 1.\end{aligned}$$

Step 2

$$\begin{aligned}\lim_{N\to\infty} P(Y = N) &= \lim_{N\to\infty} P\left(\frac{N - R_N}{\sqrt{R_N}} - \sqrt{\frac{\mu}{\lambda_N}} < \frac{Y - R_N}{\sqrt{R_N}} \le \frac{N - R_N}{\sqrt{R_N}}\right) \\ \Rightarrow \lim_{N\to\infty} \frac{\sqrt{\lambda_N/\mu}}{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right)} \cdot P(Y = N) &= \\ &= \lim_{N\to\infty} \frac{\sqrt{\lambda_N/\mu}}{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right)} \sqrt{\frac{\mu}{\lambda_N}}\, \phi\left(\epsilon\sqrt{\frac{\lambda_N}{\mu}}\right) = 1.\end{aligned}$$

Combining the results of Steps 1 and 2, we conclude that in the QD regime the blocking probability converges to 0 at rate $\Theta\left(\frac{\phi(\epsilon\sqrt{\lambda_N/\mu})}{\sqrt{\lambda_N/\mu}}\right)$:

$$\lim_{N\to\infty} E_{1,N} = \lim_{N\to\infty} \sqrt{\mu/\lambda_N}\, \phi(\epsilon\sqrt{\lambda_N/\mu})\,, \tag{3.26}$$

$$\Rightarrow \lim_{N\to\infty} \frac{\sqrt{\lambda_N/\mu}}{\phi(\epsilon\sqrt{\lambda_N/\mu})} \cdot E_{1,N} = 1. \tag{3.27}$$

Step 3

$$\begin{aligned}\lim_{N\to\infty} P(X \geq N\mu/\theta) &= \lim_{N\to\infty} P\left(\frac{X-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} \geq \frac{N\mu/\theta-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}}\right)\\ &= \lim_{N\to\infty} P\left(Z \geq \epsilon\sqrt{\frac{\lambda_N}{\theta}}\right)\\ &= \lim_{N\to\infty} \frac{\phi(\epsilon\sqrt{\frac{\lambda_N}{\theta}})}{\epsilon\sqrt{\frac{\lambda_N}{\theta}}} = 0.\end{aligned}$$

In the last line, we again turned to $\lim_{a\to\infty} P(Z \geq a) = \lim_{a\to\infty} \phi(a)/a$.

Thus $$\lim_{N\to\infty} \frac{\epsilon\sqrt{\frac{\lambda_N}{\theta}}}{\phi(\epsilon\sqrt{\frac{\lambda_N}{\theta}})} \cdot P(X \geq N\mu/\theta) = 1.$$

Step 4

$$\begin{aligned}\lim_{N\to\infty} P(X = N\mu/\theta) &= \lim_{N\to\infty} P\left(\frac{N\mu/\theta-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} - \sqrt{\frac{\theta}{\lambda_N}} < \frac{X-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}} \leq \frac{N\mu/\theta-\lambda_N/\theta}{\sqrt{\lambda_N/\theta}}\right)\\ &\Rightarrow \lim_{N\to\infty} \frac{\sqrt{\lambda_N/\theta}}{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\theta}}\right)} \cdot P(X = N\mu/\theta)\\ &= \lim_{N\to\infty} \frac{\sqrt{\lambda_N/\theta}}{\phi\left(\epsilon\sqrt{\frac{\lambda_N}{\theta}}\right)} \cdot \sqrt{\frac{\theta}{\lambda_N}}\, \phi\left(\epsilon\sqrt{\frac{\lambda_N}{\theta}}\right) = 1.\end{aligned}$$

Steps 3 and 4 show what happens to $\pi_+^N(0)$ in the QD regime.

$$\lim_{N\to\infty} \pi_+^N(0) = \epsilon. \tag{3.28}$$

Step 5

In the QD regime, the offered load per server is constant and does not exceed 100 %:

$$\lim_{N\to\infty} \rho_N = \frac{1}{1+\epsilon}.$$

This observation and the intermediate results (3.26 and 3.28) show that in the QD regime the Delay Probability (3.16) converges to 0, which is again in line with the known results:

$$\lim_{N\to\infty} P(W_q > 0) = \lim_{N\to\infty} \left[1 + (1+\epsilon)\, \frac{\Phi(\epsilon\sqrt{\lambda_N/\mu})}{\sqrt{\mu/\lambda_N}\, \phi(\epsilon\sqrt{\lambda_N/\mu})}\right]^{-1} = 0. \tag{3.29}$$

Let us define $\nu_N = \epsilon\sqrt{\frac{\lambda_N}{\mu}}$. Then the converges rate of the delay probability (3.16) can be presented as follows:

$$\Theta(\frac{\nu_N}{\phi(\nu_N)}). \tag{3.30}$$

We observe a certain symmetry between the convergence rate of the delay probability in the ED regime (3.24) and the one in the QD regime (3.30). In the ED regime, the convergence rate of delay probability depends on $\gamma \equiv \epsilon \sqrt{\frac{\lambda_N}{\theta}}$, while in the QD regime this rate is determined by the term $\omega \equiv \epsilon \sqrt{\frac{\lambda_N}{\mu}}$. The only difference between γ and ω is dictated by the service rate of L_+ and L_- respectively, i.e. by the service rate of the queue part which becomes negligible in the current regime.

3.4.4 Busy and Idle Periods under Different Operational Regimes

The limits and the convergence rates of the delay probability in the three operational regimes can be concluded from the convergence of the busy and idle periods in these regimes. Their convergence rates are summarized in Table 3.1.

We see, for example, that in the QED regime, Busy and Idle periods converge to zero at the same rate, so it makes perfect sense that the delay probability converges to a constant that is neither 0 nor 1.

In the ED regime, the busy period is very long (converges to infinity), while the idle period converges to zero.

Table 3.1: Convergence Rates of Busy and Idle Periods in the Three Operational Regimes

	Busy Period $T_{N,N-1}$		Idle Period $T_{N-1,N}$	
	lim	rate	lim	rate
QED	0	$1/\sqrt{N}$	0	$1/\sqrt{N}$
ED	∞	$\frac{1}{\sqrt{N}\ \ \phi(\epsilon\sqrt{\frac{N(1-\epsilon)}{\mu}})}$	0	$1/N$
QD	0	$1\ /\ N$	∞	$\frac{1}{\sqrt{N}\ \ \phi(\epsilon\sqrt{\frac{N(1+\epsilon)}{\theta}})}$

It is interesting to see that the convergence rates and limits of the busy and idle periods in the ED and QD regimes are almost exactly opposite to each other. This, apparently, can be related to the reversibility of the underlying Birth-and-Death process altough the idle period of a queue is defined over a finite number of states ($L_- = L|L \in \{0, 1, \ldots, N-1\}$), while its busy period is defined over an infinite number of states ($L_+ = L|L \in \{N, N+1, \ldots\}$).

3.5 Appendix

3.5.1 The L_+ Queue: Calculation of the Steady-State Distribution

To find $\pi_+(0)$ we solve the following balance equations:

$$\begin{cases} \lambda\pi_+(i-1) = (N\mu + i\theta)\pi_+(i) & ,1 \le i < \infty \\ \sum_{i=0}^{\infty} \pi_+(i) = 1 \end{cases} \tag{3.31}$$

Assuming that $N\mu/\theta$ is an integer, it follows from Equation (3.31) that $\pi_+(k)$ is given by

$$\pi_+(k) = \frac{\pi_+(0)(\lambda/\theta)^k (N\mu/\theta)!}{(N\mu/\theta + k)!}.$$

Using the fact that $\sum_{k=0}^{\infty} \pi_+(k) = 1$, we obtain that

$$\pi_+(0) \quad = \quad \frac{(\lambda/\theta)^{N\mu/\theta} \Big/ (N\mu/\theta)!}{\sum_{i=0}^{\infty} (\lambda/\theta)^{N\mu/\theta+i}/(N\mu/\theta + i)!}.$$

Note that

$$\begin{aligned} \pi_+(0) &= \frac{P(X = N\mu/\theta)}{P(X \ge N\mu/\theta)}, \qquad where \quad X \stackrel{d}{=} Pois(\lambda/\theta), \\ &= P(X = N\mu/\theta | X \ge N\mu/\theta). \end{aligned}$$

The "integer" assumption allows the probabilistic representation of $\pi_+(0)$ but it is not necessary. We can re-write $\pi_+(0)$ in terms of the incomplete Gamma function $\gamma(\cdot,\cdot)$, using the same approach as in [41]:

$$\pi_+(0) \quad = \quad \frac{(\lambda/\theta)^{N\mu/\theta}}{[N\mu/\theta] e^{\lambda/\theta} \gamma(\frac{N\mu}{\theta}, \frac{\lambda}{\theta})}, \tag{3.32}$$

where

$$\gamma(x,y) = \int_0^y t^{x-1} e^{-t} dt, \qquad x > 0, \quad y \ge 0.$$

Thus we obtain an expression for $\pi_+(0)$ that solves the balance equations (3.31) and relaxes the "integer" assumption.

Chapter 4

Erlang-C with Priorities

In this chapter we begin our introduction to queues with heterogenous customers served in accordance to their importance. These models are very useful since they describe many service environments. Examples include banks, whose customers are differentiated according to their account status; hospitals, where urgent patients do not wait in the common queue; call-centers, where the customers may be differentiated by their requests, languages or value, etc.

The assumption throughout this work is that **all servers are statistically identical**, all customers need the same service and the queue is work-conserving, i.e., no reservations of servers is allowed to guarantee a better quality of service for higher priority customers.

Two priority disciplines, under which these assumptions hold are: **Preemptive** and **Non-Preemptive** priorities.

Non-Preemptive Priority: Under this discipline, a customer of priority i enters service only when there are no waiting customers of higher priorities. Once service started, it cannot be interrupted, even upon the appearance of a delayed higher-priority customer.

Preemptive Priority: Here higher-priority customers are not ”aware” of lower priority ones. That is, if a higher-priority customer arrives when all the servers are busy, and there are lower-priority customers in service, a customer of the lowest-priority in queue is immediately returned to the head of its original queue, and the higher-priority customer enters service (immediately upon arrival).

4.1 Model Description

This section develops a formal description of the priority models studied in this chapter. The same notation is used for $M/M/N$ and $M/M/N+M$ priority queues to emphasize the common structure of their performance, in particular the expected waiting time.

- There are K customer types.
- A customer of type k has priority (preemptive or non-preemptive) over a customer of type j if and

only if $k < j$, $1 \leq k, j \leq K$. In particular, customers of the lowest type *(k=1)* have the highest priority, and of the highest type ($k = K$) have the lowest priority.

- Customers of type k arrive at rate λ_k; arrivals are Poisson, independent among the classes.
- Service rate is μ (the same for all customers);service durations are exponential, independent of the arrivals.
- $\rho = \frac{\lambda}{N\mu}$ is the servers utilization (assuming no abandonment). Here $\lambda = \sum_{i=1}^{K} \lambda_i$ is the total arrival rate.

Notations of Priority Queues

- $E_{pr}(W_q^k)$ $(E_{np}(W_q^k))$ is the expected waiting time of type k under preemptive (non-preemptive) priority discipline,
- $E_{pr}(W_q^{1\to k})$ $(E_{np}(W_q^{1\to k}))$ is the expected waiting time averaged over first k types under preemptive (non-preemptive) priority discipline,
- $E_{pr}(L_q^k)$ $(E_{np}(L_q^k))$ is the expected number of the delayed (in queue) type k customers under preemptive (non-preemptive) priority,
- $E_{pr}(L_q^{1\to k})$ $(E_{np}(L_q^{1\to k}))$ is the expected number of the delayed (in queue) customers of types $1, \ldots, k$ under preemptive (non-preemptive) priority,
- $P_{pr}(W_q^k > 0)$ $(P_{np}(W_q^k > 0))$ is the probability that customers of type k are delayed under preemptive (non-preemptive) priority discipline,
- $P_{pr}(W_q^{1\to k} > 0)$ $(P_{np}(W_q^{1\to k} > 0))$ is the probability that a customer of any type $1, \ldots, k$ is delayed under preemptive (non-preemptive) priority,
- $P_{pr}(Aband^k)$ $(P_{np}(Aband^k))$ is the probability that customers of type k abandon under preemptive (non-preemptive) priority discipline,
- $P_{pr}(Aband^{1\to k})$ $(P_{np}(Aband^{1\to k}))$ is the probability that customers of the first k types abandon under preemptive (non-preemptive) priority,
- $E_{pr}(W_q)$ $(E_{np}(W_q))$ is the expected waiting time of all types under preemptive (non-preemptive) priority discipline,
- $E_{pr}(L_q)$ $(E_{np}(L_q))$ is the expected total number of the delayed (in queue) customers under preemptive (non-preemptive) priority discipline.

Notations of Related Queues without Priorities

- $E_{\lambda_k}(W_q)$ is the expected waiting time in an $M/M/N$ queue with homogeneous customers which arrive at rate λ_k,
- $E(W_q^{1\to k})$ is the expected waiting time in an $M/M/N$ queue with homogeneous customers which arrive at rate $\lambda_{1\to k} = \sum_{i=1}^{k} \lambda_i$,
- $E_{\lambda}(W_q)$ is the expected waiting time in an $M/M/N$ queue with homogeneous customers which arrive at rate $\lambda = \sum_{i=1}^{K} \lambda_i$,
- $P_{\lambda_k}(W_q > 0)$ is the delay probability in an $M/M/N$ queue with homogeneous customers which arrive at rate λ_k (Erlang-C formula),
- $P(W_q^{1\to k} > 0)$ is the delay probability in an $M/M/N$ queue with homogeneous customers which arrive at rate $\lambda_{1\to k} = \sum_{i=1}^{k} \lambda_i$,
- $P_{\lambda}(W_q^{k} > 0)$ is the delay probability in an $M/M/N$ queue with homogeneous customers which arrive at rate $\lambda = \sum_{i=1}^{K} \lambda_i$.

4.2 Exact Results

Consider a general Erlang-C queue with priorities, as described in Section 4.1. Here we present known results for the expected waiting time under both priority disciplines.

4.2.1 Preemptive Priority

The expected waiting time under preemptive priority discipline is found recursively, following **the same** expressions for queues with and without abandonment:

$$E_{pr}(W_q^k) = \left[\lambda_{1\to k} E_{pr}(W_q^{(1\to k)}) - \lambda_{1\to(k-1)} E_{pr}(W_q^{(1\to(k-1))})\right]\lambda_k^{-1}, \qquad k = 1\ldots K. \tag{4.1}$$

Relation (4.1) is a direct consequence of Little's Law. Indeed, a customer of type k "sees" customers of only two kinds: those of a higher priorities (i.e., types $1,\ldots,k-1$) and those of type k. The total number of customers of types $1,\ldots,k$ consists of the total number of delayed customers of the higher priority (types $1,\ldots,k-1$) and the delayed customers of type k. That is,

$$E_{pr}(L_q^{1\to k}) = E_{pr}(L_q^{1\to(k-1)}) + E_{pr}(L_q^k). \tag{4.2}$$

By Little's Law, Equation (4.2) can be restated as follows:

$$\lambda_{1\to k} E_{pr}(W_q^{(1\to k)}) = \lambda_{1\to(k-1)} E_{pr}(W_q^{(1\to(k-1))}) + \lambda_k E_{pr}(W_q^{(k)}). \tag{4.3}$$

Now the recursive relation (4.1) follows.

The total number of customers of the first k types is distributed as in the $M/M/N(+M)$ queue with the arrival rate $\lambda_{1\to k}$ due to the fact that the customers of the higher priorities do not see lower-priorities customers:

$$E_{pr}(L_q^{1\to k}) = E(L_q^{1\to k}). \tag{4.4}$$

After taking into account the fact that higher priorities are not aware of the lower ones (see (4.4)), the following formulation for the expected waiting time of type k under preemptive priority is derived:

$$E_{pr}(W_q^k) = \Big[\lambda_{1\to k}E(W_q^{(1\to k)}) - \lambda_{1\to(k-1)}E(W_q^{(1\to(k-1))})\Big]\lambda_k^{-1}, \qquad k = 1\dots K. \tag{4.5}$$

In the case of Erlang-C, an exact expression for the expected waiting time is not hard to deduce from (4.5):

$$E_{pr}(W_q^k) = \frac{E_{2,N}(\lambda_{1\to k})}{\lambda_k(1-\sigma_k)} - \frac{E_{2,N}(\lambda_{1\to(k-1)})}{\lambda_k(1-\sigma_{k-1})}. \tag{4.6}$$

Here

$$\sigma_k = \sum_{i=1}^{k}\rho_i, \qquad \rho_i = \frac{\lambda_i}{N\mu} \quad \text{is the fraction of time a server spends on customers of type } i$$

and $E_{2,N}(\cdot)$ is the delay probability in the *M/M/N* queue (Erlang-C Formula (2.2)).

4.2.2 Non-Preemptive Priority

Results for the waiting time under non-preemptive priority were presented by Kella and Yechially, [20], who determined Laplace transform of the waiting time distribution for any type k and showed that the expected waiting time is given by the following expression:

$$E_{np}(W_q^k) = E_{2,N}(\lambda)\Big[N\mu(1-\sigma_k)(1-\sigma_{k-1})\Big]^{-1}, \tag{4.7}$$

where $E_{2,N}(\lambda)$ and σ_k are defined as before.

To define these Laplace transforms and to prove (4.7), Kella and Yechially used a vacations approach. Subsection 4.2.3 presents an alternative proof, which provides an important insight on the three basic components of the expected waiting time. Its idea can be used to calculate not only expectations, but also Laplace transforms for waiting time of any type k. The approach is due to Gurvich [17] and it can be skipped without loss of reading-continuity.

4.2.3 Expected Waiting Time under Non-Preemptive Priority: Proof of (4.7)

The proof consists of the two following steps:

- **Step 1**
  The delay probability is the same for any type k and is equal to the delay probability in the $M/M/N$ queue with arrival rate λ and service rate μ (Erlang-C formula (3.3)).
  $P(W_q^k > 0) = E_{2,N}(\lambda)$ for any $k = 1,\dots,K$,

- **Step 2**
  The expected waiting time given waiting is given by the following expression:
  $E_{np}(W_q^k|W_q^k > 0) = \left[N\mu(1-\sigma_k)(1-\sigma_{k-1})\right]^{-1}$

After these two steps the result follows.

- **Step 1**
  The delay probability does not depend on the internal order of customer service. This is the reason why this probability is the same for all work-conserving queues with the arrival rate λ and service rate μ and is given by Erlang-C (3.3).

- **Step 2**
  Now let us study the expected waiting time given waiting of a type k customer.
  If there is waiting, customers of this type are served as under an $M/G/1$ queue where G is the busy period distribution of an $M/M/1$ queue with arrival rate $\lambda_{1\to(k-1)} \equiv \sum_{i=1}^{k-1}\lambda_i$ and service rate $N\mu$. The expectation of the busy period of the $M/M/1$ queue is given by $\dfrac{1}{N\mu(1-\sigma_{k-1})}$ (See, for example, Kleinrock [21], p. 213, Equation 5.141).

  Given waiting, a customer of type k "sees" upon arrival only customers of equal or higher priority. Thus, to determine how many busy periods he needs to wait, we notice that this number is equal to one plus the queue length in an $M/M/1$ model with arrival rate $\lambda_{1\to k} \equiv \sum_{i=1}^{k}\lambda_i$ and service rate $N\mu$. The stationary distribution of this queue is $Geom_0(1-\frac{\lambda_{1\to k}}{N\mu})$, where $Geom_0(\cdot)$ is a geometric distribution starting at zero (See, for example, Kleinrock [21], p. 96, Equation 3.23). This is why the number of busy periods to wait is distributed $Geom(1-\frac{\lambda_{1\to k}}{N\mu})$ with the mean $\dfrac{1}{(1-\sigma_k)}$. In other words, the number of the busy periods to wait is distributed as the total number of customers in an $M/M/1$ model with arrival rate $\lambda_{1\to k} \equiv \sum_{i=1}^{k}\lambda_i$ and service rate $N\mu$.

For future analysis, we observe that the expected waiting time of type *k* consists of the following three components.

1. The **delay probability**, which is found using global system characteristics. It depends on the total arrival rate λ, the service time μ and the number of the servers N.

2. Given waiting, a type-k customer is advanced in his queue only when there are no customers of higher priorities. This is why, customers of type k are exposed to an $M/G/1$ queue where G is the **busy period** of an $M/M/1$ with the arrival rate $\lambda_{1\to(k-1)}$ and service rate $N\mu$.

3. A queue, which a **delayed** type-k customer faces, consists of customers with priority no lower than his, which means that the average **queue length** for him depends on the arrivals of the first k types. That is, if there is waiting, the average number of the busy periods to wait is found similarly

to the average number of customers in an $M/M/1$ queue with arrival rate $\lambda_{1\to k}$ and service rate $N\mu$.

To formally summarize:

$$\begin{aligned} E_{np}(W_q^k) &= P(W_q^k > 0) \times E(W_q^k | W_q^k > 0) \\ &= \underbrace{E_{2,N}(\lambda)}_{1} \times \underbrace{\frac{1}{N\mu(1-\sigma_{k-1})}}_{2} \times \underbrace{\frac{1}{(1-\sigma_k)}}_{3} \\ &= \frac{E_{2,N}(\lambda)}{N\mu(1-\sigma_{k-1})(1-\sigma_k)}. \qquad \square \end{aligned}$$

4.3 An Asymptotic Example with Two Customer Types: QED and ED

This section deals with the analysis of queues with two types of customers, $K = 2$, in the QED and ED regimes. The results obtained for the lowest priority can be applied to queues with any number of customer types, because it is possible to consider any k first types as the highest-priority customers, and the rest $k+1, \ldots, K$ types as customers of the lowest priority.

The analysis of both the QED and ED regimes is organized in the following way. For each operational regime, we start from the preemptive priority discipline, and continue to the non-preemptive discipline. The analysis is conducted using analytical tools for ED, QED and QD regimes developed by Zeltyn in [42], and the exact formulae (4.1) and (4.7) of the expected waiting time for any type k re-stated below for $K = 2$.

For the case with two customer types, equations (4.1) and (4.7) read as follows:

$$E_{pr}(W_q^1) = \frac{E_{2,N}(\lambda_1)}{N\mu(1-\rho_1)}, \tag{4.8}$$

$$E_{pr}(W_q^2) = \frac{\lambda \frac{E_{2,N}(\lambda)}{N\mu(1-\rho)} - \lambda_1 \frac{E_{2,N}(\lambda_1)}{N\mu(1-\rho_1)}}{\lambda_2}, \tag{4.9}$$

$$E_{np}(W_q^1) = \frac{E_{2,N}(\lambda)}{N\mu(1-\rho_1)}, \tag{4.10}$$

$$E_{np}(W_q^2) = \frac{E_{2,N}(\lambda)}{N\mu(1-\rho_1)(1-\rho)}. \tag{4.11}$$

Note that in queues with two types of customers, $\sigma_1 \equiv \rho_1$ and $\sigma_2 \equiv \rho_1 + \rho_2 = \rho$.

4.3.1 QED Regime

The QED operational regime for $M/M/N$ queues was first introduced by Halfin and Whitt in [16]. Under this regime, the service rate μ is constant, and as the number of servers N and the arrival rate λ increase infinitely, the *square root* staffing rule prevails:

$$N \approx R_N + \beta\sqrt{R_N}, \quad \lambda \to \infty, \qquad \beta > 0, \tag{4.12}$$

where $R_N = \frac{\lambda_N}{\mu}$ is the *Offered Load* in an $M/M/N$ queue with arrival rate λ and service rate μ.
In addition, we assume that, as the total arrival rate λ grows to infinity, the fraction of the arrival rate of each type remains a positive constant ($\frac{\lambda_k}{\lambda} = const$ for $k = 1,\ 2$).

Preemptive Discipline

First Type

Under this discipline customers of the first type enjoy the QD regime, because they are not troubled with the presence of the lower priority. That is, they see a system staffed by the following rule:

$$N \approx \frac{\lambda_1}{\mu}(1+\delta), \tag{4.13}$$

as λ_1 and N increase infinitely.
Recall that the fraction of time, in which a single server works with first-type customers, stays constant (ρ_1), and it is similar to the offered load per agent in the QD regime with an arrival rate λ_1. Hence

$$\delta \to \frac{\rho_2}{\rho_1}.$$

In order to find how fast the waiting time of the first type converges to 0, we need to know the convergence of the delay probability in an $M/M/N$ queue with arrival rate λ_1. This convergence rate is calculated using the approximation of the delay probability for the QD regime (see [42], Remark 5.1):

$$P_{\lambda_1}(W_q > 0) \approx \frac{1}{\sqrt{2\pi N}} \cdot \frac{\rho_1^N}{\rho_2} \cdot e^{1-\rho_1}.$$

Based on this approximation, we conclude that the convergence rate of the **first-type** expected waiting time is $\Theta(\frac{\rho_1^N}{N\sqrt{N}})$:

$$\lim_{N\to\infty} \frac{N\sqrt{N}}{\rho_1^N} \cdot E_{pr}(W_q^1) = \lim_{N\to\infty} \frac{N\sqrt{N}}{\rho_1^N} \cdot \frac{\frac{1}{\sqrt{2\pi N}} \cdot \frac{\rho_1^N}{\rho_2} \cdot e^{1-\rho_1}}{N\mu(1-\rho_1)} = \frac{e^{\rho_2}}{\sqrt{2\pi}\mu\rho_2^2}.$$

Second Type

The convergence rate of the **second type** under preemptive priority discipline is $\Theta(\frac{1}{\sqrt{N}})$:

$$\lim_{N\to\infty} \sqrt{N} E_{pr}(W_q^2) = \lim_{N\to\infty} \sqrt{N} \frac{\lambda \frac{E_{2,N}(\lambda)}{N\mu(1-\rho)} - \lambda_1 \frac{E_{2,N}(\lambda_1)}{N\mu(1-\rho_1)}}{\lambda_2} = \frac{\alpha}{\rho_2\mu\beta}.$$

Non-Preemptive Discipline

First Type

The expected waiting time of the **first type** under non-preemptive priority converges at rate $\Theta(1/N)$:

$$\lim_{N\to\infty} N E_{np}(W_q^1) = \lim_{N\to\infty} N \frac{E_{2,N}(\lambda)}{N\mu(1-\rho_1)} = \frac{\alpha}{\mu\rho_2},$$

where $\alpha = \alpha(\beta)$ is the Halfin-Whitt function [16], see (2.17).

Second Type

The expected waiting time of the **second type** under non-preemptive priority converges at rate $\Theta(1/\sqrt{N})$:

$$\lim_{N\to\infty} \sqrt{N} E_{np}(W_q^2) = \lim_{N\to\infty} \sqrt{N} \frac{E_{2,N}(\lambda)}{N\mu(1-\rho_1)(1-\rho)} = \frac{\alpha}{\mu\beta\rho_2}.$$

4.3.2 ED Regime

Now let us consider a sequence of two-type $M/M/N$ queues in the ED operational regime. Assume that the total arrival rate is $\lambda \to \infty$, the service rate μ is constant and the total number of servers N is detrmined by:

$$N(1-\rho_N) = \gamma \quad \text{for some } 0 < \gamma < \infty. \tag{4.14}$$

In addition, we again assume that as the total arrival rate λ grows to infinity, the fraction of the arrival rate of each type remains a positive constant ($\frac{\lambda_k}{\lambda} = const$ for $k = 1,\ 2$).

Preemptive Discipline

First Type

Here we again use the fact that the higher priority customers are not aware of the lower priority and enjoy the QD regime, i.e. they see a system with the staffing level given by (4.13):
Repeating the arguments of the QED regime for the highest priority under the preemptive discipline and using the approximation of the delay probability for the QD regime ([42], Remark 5.1), we obtain:

$$P_{\lambda_1}(W_q > 0) \approx \frac{1}{\sqrt{2\pi N}} \cdot \frac{\rho_1^N}{\rho_2} \cdot e^{1-\rho_1}.$$

The convergence rate of the expected waiting time of the **first type** under preemptive priority discipline is $\Theta(\frac{(\rho_1)^N}{N\sqrt{N}})$:

$$\lim_{N\to\infty} \frac{N\sqrt{N}}{\rho_1^N} \cdot E_{pr}(W_q^1) = \lim_{N\to\infty} \frac{N\sqrt{N}}{(\rho_1)^N} \cdot \frac{\frac{1}{\sqrt{2\pi N}} \cdot \frac{(\rho_1)^N}{\rho_2} \cdot e^{1-\rho_1}}{N\mu(1-\rho_1)} = \frac{1}{\sqrt{2\pi}\mu\rho_2^2}.$$

Second Type

The convergence rate of the expected waiting time of the **second type** under preemptive priority is $\Theta(1)$ (the same rate as that of $E_{np}(W_q^2)$):

$$\lim_{N\to\infty} E_{pr}(W_q^2) = \lim_{N\to\infty} \frac{\lambda \frac{P_\lambda(W_q>0)}{N\mu(1-\rho)} - \lambda_1 \frac{P_{\lambda_1}(W_q>0)}{N\mu(1-\rho_1)}}{\lambda_2} = \frac{1}{\rho_2\mu\gamma}.$$

Non-Preemptive Discipline

First Type

The convergence rate of the expected waiting time of the **first type** under non-preemptive priority is is $\Theta(1/N)$:

$$\lim_{N\to\infty} NE_{np}(W_q^1) = \lim_{N\to\infty} N\frac{P_\lambda(W_q > 0)}{\mu(1-\rho_1)} = \frac{1}{\mu\rho_2}$$

Second Type

The convergence rate of the expected waiting time of the **second type** under non-preemptive priority is $\Theta(1)$:

$$\lim_{N\to\infty} E_{np}(W_q^2) = \lim_{N\to\infty} \frac{P_\lambda(W_q > 0)}{N\mu(1-\rho_1)(1-\rho)} = \frac{1}{\mu\gamma\rho_2}.$$

Remark 3 *Note the asymptotic equivalence of the lowest priority under preemptive and non-preemptive disciplines. This was emphasized by Ashlagi in [2]. We will see that this behavior is preserved in queues with abandonments. A detailed explanation of this phenomenon is given at the beginning of Subsection 5.3. Table 5.1 summarizes the rates of convergence in the QED and ED regimes for queues with and without abandonment.*

Chapter 5

Erlang-A with Priorities

In this chapter we analyze priority queues with impatient customers. We begin with a description of our models, present some new results for non-preemptive priority, and conclude with asymptotic analysis.

5.1 Model Description

This section presents a general description of the models studied in this chapter. Note, that the description of Erlang-A queues is very similar to the description of $M/M/N$ queues in Section 4.1.

- There are K customer types.
- A customer of type k has priority (preemptive or non-preemptive) over a customer of type j if and only if $k < j$, $1 \leq k, j \leq K$. In particular, customers of the first type *(k=1)* have the highest priority.
- Customers of type k arrive at rate λ_k; arrivals are Poisson.
- Service rate is μ (the same for all customers); service durations are exponential.
- $\rho = \frac{\lambda}{N\mu}$ is the offered load per server. Here $\lambda = \sum_{i=1}^{K} \lambda_i$ is the total arrival rate.
- Abandonment rate is θ (the same for all customers); customers' patience is exponential.

Let us define the following:
Notations of Priority Queues

- $E_{pr}(W_q^k)$ $(E_{np}(W_q^k))$ is the expected waiting time of type k under preemptive (non-preemptive) priority discipline,
- $E_{pr}(W_q^{1\to k})$ $(E_{np}(W_q^{1\to k}))$ is the expected waiting time averaged over first k types under preemptive (non-preemptive) priority discipline,

- $E_{pr}(L_q^k)$ $(E_{np}(L_q^k))$ is the expected number of the delayed (in queue) type k customers under preemptive (non-preemptive) priority discipline,
- $E_{pr}(W_q^{1\to k})$ $(E_{np}(W_q^{1\to k}))$ is the expected number of the delayed (in queue) customers of types $1,\ldots,k$ under preemptive (non-preemptive) priority,
- $P_{pr}(W_q^k > 0)$ $(P_{np}(W_q^k > 0))$ is the probability that customers of type k are delayed under preemptive (non-preemptive) priority discipline,
- $P_{pr}(W_q^{1\to k} > 0)$ $(P_{np}(W_q^{1\to k} >))$ is the probability that a customer of any type $1,\ldots,k$ is delayed under preemptive (non-preemptive) priority discipline,
- $P_{pr}(Aband^k)$ $(P_{np}(Aband^k))$ is the probability that customers of type k abandon under preemptive (non-preemptive) priority discipline,
- $P_{pr}(Aband^{1\to k})$ $(P_{np}(Aband^{1\to k}))$ is the probability that customers of the first k types abandon under preemptive (non-preemptive) priority,
- $E_{pr}(W_q)$ $(E_{np}(W_q))$ is the expected waiting time of all types under preemptive (non-preemptive) priority,
- $E_{pr}(L_q)$ $(E_{np}(L_q))$ is the expected total number of the delayed (in queue) customers under preemptive (non-preemptive) priority discipline.

Notations of Related Queues without Priorities

- $E_{\lambda_k}(W_q)$ is the expected waiting time in an $M/M/N+M$ queue with homogeneous customers which arrive at rate λ_k,
- $E(W_q^{1\to k})$ is the expected waiting time in an $M/M/N+M$ queue with homogeneous customers which arrive at rate $\lambda_{1\to k}=\sum_{i=1}^{k}\lambda_i$,
- $E_{\lambda}(W_q)$ is the expected waiting time in an $M/M/N+M$ queue with homogeneous customers which arrive at rate $\lambda=\sum_{i=1}^{K}\lambda_i$,
- $P_{\lambda_k}(W_q>0)$ is the delay probability in an $M/M/N+M$ queue with homogeneous customers which arrive at rate λ_k (Erlang-A formula),
- $P(W_q^{1\to k}>0)$ is the delay probability in an $M/M/N+M$ queue with homogeneous customers which arrive at rate $\lambda_{1\to k}=\sum_{i=1}^{k}\lambda_i$,
- $P_{\lambda}(W_q^k>0)$ is the delay probability in an $M/M/N+M$ queue with homogeneous customers which arrive at rate $\lambda=\sum_{i=1}^{K}\lambda_i$,
- $P_{\lambda_k}(Aband)$ is the abandonment probability in an $M/M/N+M$ queue with homogeneous customers which arrive at rate λ_k,

- $P(Aband^{1\to k})$ is the abandonment probability in an $M/M/N+M$ queue with homogeneous customers which arrive at rate $\lambda_{1\to k}=\sum_{i=1}^{k}\lambda_i$.

5.2 Exact Results

5.2.1 Preemptive Priority

The expected waiting time under preemptive priority is found by a recursive expression, which is **the same** (See (4.1) for comparison) for queues with and without abandonment:

$$E_{pr}(W_q^k)=\left[\lambda_{1\to k}E_{pr}(W_q^{(1\to k)})-\lambda_{1\to(k-1)}E_{pr}(W_q^{(1\to(k-1))})\right]\lambda_k^{-1},\qquad k=1\dots K. \tag{5.1}$$

This relation can be re-stated in terms of measures without priorities as follows:

$$E_{pr}(W_q^k)=\left[\lambda_{1\to k}E(W_q^{(1\to k)})-\lambda_{1\to(k-1)}E(W_q^{(1\to(k-1))})\right]\lambda_k^{-1},\qquad k=1\dots K. \tag{5.2}$$

5.2.2 Non-Preemptive Priority: Expected Waiting Time of First-Type Customers

Let us analyze the expected waiting time of the **highest** priority (first-type customers) under the non-preemptive priority discipline. In the next section we will use the obtained result to calculate the expected waiting time for any type k.

Theorem 5 *The delay probability for any type under the non-preemptive priority discipline is the same as in the $M/M/N+M$ queue without priorities, with arrival rate λ, service rate μ and abandonment rate θ (Erlang-A formula):*

$$P_{np}(W_q^k>0)=P_\lambda(W_q>0)\qquad k=1,\dots,K. \tag{5.3}$$

*The expected waiting time of the **delayed** customers with the highest priority is the same as in an $M/M/N+M$ queue without priorities with arrival rate λ_1 and the rest of the parameters the same:*

$$E_{np}(W_q^1|W_q^1>0)=E_{\lambda_1}(W_q|W_q>0). \tag{5.4}$$

Statements (5.3) and (5.4) yield an expression for the expected waiting time of customers with the highest priority:

$$E_{np}(W_q^1)\;\;=\;\;P_\lambda(Wait>0)\cdot E_{\lambda_1}(W_q|W_q>0) \tag{5.5}$$

Remark 4 *The ratio between the expected waiting times of the highest priority customers under the preemptive and the non-preemptive disciplines is equal to the ratio of the appropriate delay probabilities:*

$$\frac{E_{pr}^1(W_q)}{E_{np}^1(W_q)}=\frac{P_{\lambda_1}(W_q>0)}{P_\lambda(W_q>0)}. \tag{5.6}$$

Remark 5 *Note that the statement of Theorem 5 holds also for Erlang-C queues. Equation (4.1) for the expected waiting time of customers with the highest priority (Kella and Yechially [20]) reads as follows:*

$$E_{np}(W_q^1) = E_{2,N}(\lambda) \cdot \frac{1}{N\mu(1-\rho_1)}, \tag{5.7}$$

where $E_{2,N}(\lambda)$ is the delay probability in an $M/M/N$ queue with arrival rate λ (Erlang-C formula, (3.3)), and $\frac{1}{N\mu(1-\rho_1)} = E_{\lambda_1}(W_q|W_q > 0)$.

Proof

First, let us ascertain when the highest priority customers get delayed. Consider the total number of present customers as a Birth & Death process. Its transition-rate diagram is presented in Figure 5.1.

Customers of any type are delayed if they arrive to states $N, N+1, \ldots$. The transition rates in the

Figure 5.1: Total number of customers with non-preemptive priorities

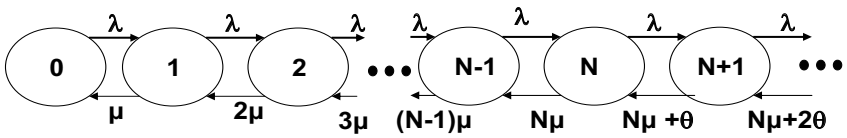


diagram do not depend on the internal discipline of the queue, consequently the delay probability of any customers type, by PASTA, is given by the Erlang-A formula:

$$P_{np}(W_q^k > 0) = P_\lambda(W_q > 0), \qquad i = 1, \ldots, K.$$

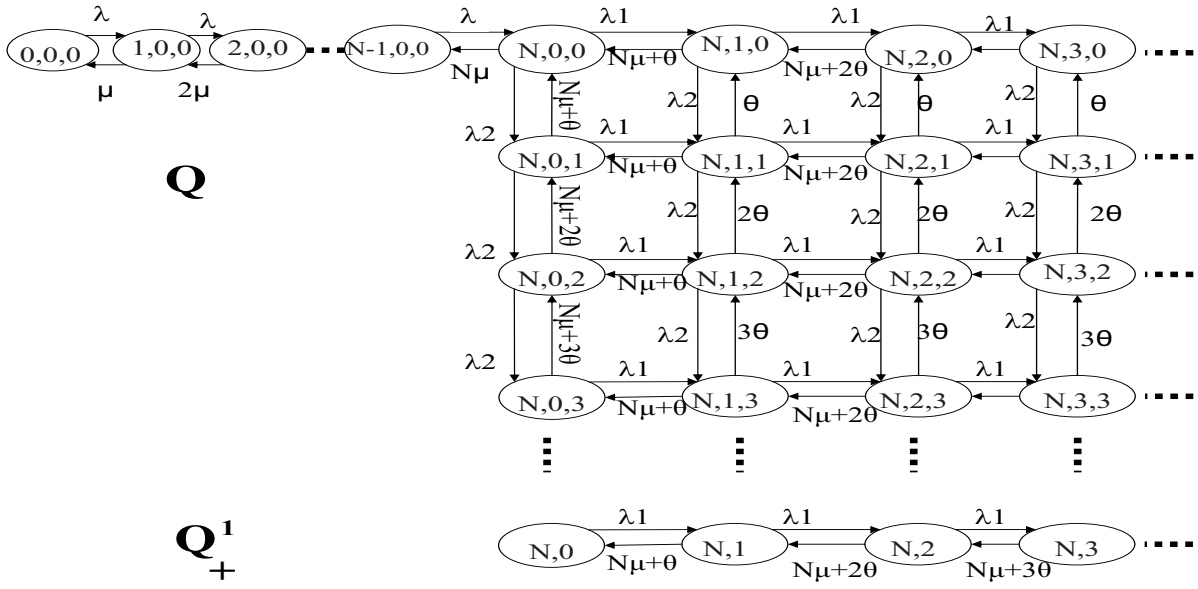
To prove part (5.4) of this theorem, we note that customers of the highest priority can classify all customers into two types: the first type consisting of customers of their own type only, and the second type (lower priority) of all other types of customers. Hence, in order to check the expected waiting time, if there is waiting, of the first type of customers, it is sufficient to analyze an $M/M/N+M$ queue with $K = 2$. Figure 5.2 presents the transition rate diagram of a two-type queue with non-preemptive priorities. Here, the first number in each state is the number of busy servers, the second entry is the total number of the delayed customers of the first type, and the last entry is the number of the delayed customers of the second type.

We define L_+ to be a sub-process of the original queue restricted to the states $k \geq N$, that is, only those states where all the servers are busy. Using the excursions technique presented in the previous chapter, it can be shown that L_+ is distributed like an $M/M/1+M$ queue with two customer types with arrival rates λ_1 and λ_2, service rate $N\mu + \theta$ and abandonment rate θ.[1]

Let us now aggregate the states of L_+ with the same number of the highest priority customers. Define

[1]Note that the transition rates for the L_+ part are similar to those in an $M/M/1+M$ queue under preemptive priority, with $K = 2$, arrival rates λ_1 and λ_2, and service rate $N\mu + \theta$.

Figure 5.2: Non-Preemptive Priority Queue with $K = 2$



L_+^1 to be the total number of the delayed first type customers. The transition rates diagram of L_+^1 is presented at the bottom of Figure 5.2. In this diagram, state (N, i) means that there are N busy servers and i delayed first-type customers. It is a diagram of a single-type $M/M/1 + M$ queue with arrival rate λ_1, service rate $N\mu + \theta$ and abandonment rate θ, which is similar to the L_+ part of a single-type queue with arrival rate λ_1, service rate μ and abandonment rate θ.

The fact that the number of delayed first-type customers has the same distribution as the number of delayed customers in a queue with a single customer type with arrival rate λ_1, allows us to conclude, by Little's Law, that the expected waiting time, if there is waiting, of the first type under non-preemptive priority is equal to the expected waiting time, if there is waiting, in the queue without priorities with total arrival rate λ_1.

Now it remains to conclude the final expression of the expected waiting time for the customers of the highest priority.

$$E_{np}(W_q^1) = P_\lambda(W_q > 0) \cdot E_{\lambda_1}(W_q | W_q > 0). \tag{5.8}$$

□

It is worth mentioning here, that Theorem (5) can be applied for the calculation of the expected waiting time of any first k types:

$$E_{np}(W_q^{1\to k}) = P_\lambda(W_q > 0) \cdot E(W_q^{1\to k} | W_q^{1\to k} > 0). \tag{5.9}$$

We will use this observation together with Little's Law to obtain the expected waiting time of any type k.

5.2.3 Non-Preemptive Priority: Expected Waiting Time of Type-k Customers

We have mentioned in Subsection 4.2.3 that the expected waiting time of type-k customers is comprised of three components. For queues with abandonments these components are the following: (1) The delay probability in a queue with arrival rate λ (Erlang-A), (2) The expected duration of a busy period in an $M/M/1+M$ queue with arrival rate $\lambda_{1\to(k-1)}$ and service rate $N\mu$ and (3) The expected number of customers in an $M/M/1+M$ queue with arrival rate $\lambda_{1\to k}$ and service rate $N\mu$. The problem is that these components, especially the second one, are not easily calculated.

In this section we develop a recursive expression for the expected waiting time of type-k customers. It is based on the same idea as the recursive formula for the preemptive priority (See Equation (4.1) or (5.1) for $M/M/N$ or $M/M/N+M$ accordingly).

- **Step 1**
  Calculate $E_{np}(W_q^1)$ by (5.5):
  $$E_{np}(W_q^1) = P_\lambda(W_q > 0) \cdot E_{\lambda_1}(W_q | W_q > 0).$$
- **Step 2**
  In general: "Merge" the first k types into a single highest-priority type with arrival rate $\lambda_{1\to k}$ and calculate the average waiting time $E_{np}(W_q^{1\to k})$ of these k types, by using (5.5) again:
  $$E_{np}(W_q^{1\to k}) = P_\lambda(W_q > 0) \cdot E(W_q^{1\to k} | W_q^{1\to k} > 0).$$
- **Step 3**:
  Let us use the same logic as in the case of preemptive priority. The total number of customers of types 1 through $(k-1)$ and customers of type k is equal to the number of all customers of the first k types. Applying Little's Law we receive:
  $$E_{np}(L_q^{1\to k}) = E_{np}(L_q^{1\to(k-1)}) + E_{np}(L_q^k) \tag{5.10}$$
  $$E_{np}(W_q^{1\to k}) = \frac{\lambda_{1\to(k-1)}}{\lambda_{1\to k}} E_{np}(W_q^{1\to(k-1)}) + \frac{\lambda_k}{\lambda_{1\to k}} E_{np}(W_q^k) \tag{5.11}$$
  $$\Rightarrow E_{np}(W_k) = \frac{\lambda_{1\to k} E_{np}(W_q^{1\to k}) - \lambda_{1\to(k-1)} E_{np}(W_q^{1\to(k-1)})}{\lambda_k}. \tag{5.12}$$

Note that the recursive relation (5.12) is similar to the recursion (5.1) for the preemptive priority, and the only difference is in the calculation of the expected waiting time of the highest priority, i.e., in **Step 2**.

5.3 Asymptotic Equivalence of the Lowest Priority

It was shown by Ashlagi in [2], that under ED, QED, QD operational regimes and also in the conventional heavy traffic, the expected waiting time of the lowest priority in queues without abandonment converges

to zero under preemptive and non-preemptive disciplines, both at the same rate. These results can be expanded to queues with the same individual abandonment rate for all types of customers. But before we do that, we explain the physics of this asymptotic equivalence.

5.3.1 The Physics of the Asymptotic Equivalence

Consider two $M/M/N+M$ systems with *K=2* customer types. Suppose that in the first system the priority discipline is **non-preemptive**, and in the second system the priority discipline is **preemptive**. Let us assume that the lowest priority is not negligible: $\lim_{N\to\infty} \frac{\lambda_2}{\lambda} = \rho_2,\ \ 0 < \rho_2 \leq 1$. It follows from this assumption that $\lim_{N\to\infty} \frac{\lambda_1}{\lambda} = \rho_1,\ \ 0 \leq \rho_1 < 1$.
This subsection presents an intuitive explanation of the asymptotic equivalence of preemptive and non-preemptive disciplines, as far as the average waiting time of the lowest priority is concerned. The explanation covers the ED, QD and QED operational regimes and also conventional heavy traffic.

Any arriving second-type customer joins a queue that consists of customers of both types. Thus, by PASTA the average length of a queue in front of a lowest-priority customer is equal under both preemptive and non-preemptive disciplines and can be found using an $M/M/N+M$ model with arrival rate λ, service rate μ and abandonment rate θ.

Non-preemptive priority: A delayed second-type customer can advance one position in his queue when there are no waiting customers of the first type. Consequently, the average time it takes him to move forward in his queue is equal to one busy period of an $M/M/1+M$ queue with arrival rate λ_1 and service rate $N\mu$.

Preemptive priority: Due to the possibility of preemptions, the time before the second-type customer is advanced is longer than a single busy period of an $M/M/1+M$ queue with arrival rate λ_1. In order to determine this time, we need to multiply the expected busy-period duration by the average number of times a lowest-priority customer re-starts his service (each time, due to preemption).

Let us consider some specific low-priority customer under the preemptive priority discipline, which is currently starting service. He needs $exp(\mu)$ time to accomplish his service and then leaves the system. We start from the analysis of the QED and ED regimes.

QED and ED Regimes: $N \approx \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}},\ \ -\infty < \beta < \infty;$ $\qquad N \approx \frac{\lambda}{\mu} - \epsilon\frac{\lambda}{\mu},\ \ \epsilon > 0$ **.**
The offered load per server converges to 1 (QED) or exceeds 1. Thus, the servers utilization is close to 100%.

Due to the assumption that the second-type is not negligible, in steady state customers of the first type start their service immediately (asymptotically). Thus, the total number of servers busy with the highest priority is proportional to the fraction of the highest-priority arrivals:

$$N_1 \approx \frac{\lambda_1}{\mu} = \rho_1\, N.$$

The rest of the servers are "free" to serve the second type. That is, the average number of the servers working with the second type is

$$N_2 = N - N_1 \approx \rho_2 N.$$

The number of high-priority customers which arrive during a single service time is $\frac{\lambda_1}{\mu} + O(\sqrt{\frac{\lambda_1}{\mu}})$.

The average number of high-priority customers that leave the system during a single service time is $\frac{\lambda_1}{\mu}$ $(= N_1)$.

This means, that there will be $O(\sqrt{\frac{\lambda_1}{\mu}})$ preemptions during a single service time.

The observation above allows us to conclude that the total number of interruptions of any randomly-chosen second-type customer converges to 0:

$$\lim_{N\to\infty} P(preemption) = \lim_{N\to\infty} \frac{O(\sqrt{\frac{\lambda_1}{\mu}})}{\rho_2 N} \approx \lim_{N\to\infty} \frac{O(\sqrt{N})}{N} = 0.$$

QD Regime: $N \approx \frac{\lambda}{\mu} + \epsilon\frac{\lambda}{\mu},\ \epsilon > 0.$

Under the assumptions of the QD regime, first-type customers are not sensitive to the change of the operational regime. Thus, the average number of servers busy with the highest priority at some moment of time remains $N_1 \approx \dfrac{\lambda_1}{\mu} = \rho_1\, N.$

The first-type customers enter the service with rate $\frac{\lambda_1}{\mu} + O(\sqrt{\frac{\lambda_1}{\mu}})$, and the average number of highest-priority customers who leave the system during a single service does not change either: $\dfrac{\lambda_1}{\mu}$.

The average number of servers needed for the lowest priority at each moment of time is $N_2 \approx \dfrac{\lambda_2}{\mu}$. Note, that $\lim_{N\to\infty}(N_1 + N_2) < \lim_{N\to\infty} N$.

The number of second-type customers, arriving during a single service time is $\frac{\lambda_2}{\mu} + O(\sqrt{\frac{\lambda_2}{\mu}})$.

The average number of lowest-priority customers leaving the system during a single service time is $\frac{\lambda_2}{\mu}$ $(= N_2)$.

It is easy to see that, in steady state, there are ΔN servers, available at each moment of time, where

$$\text{Here} \qquad \Delta N \;=\; N - N_1 - N_2 - O\left(\sqrt{\frac{\lambda_1}{\mu}}\right) - O\left(\sqrt{\frac{\lambda_2}{\mu}}\right) \;\approx\; \epsilon\frac{\lambda}{\mu} - O\left(\sqrt{\frac{\lambda}{\mu}}\right), \qquad \Delta N \;>\; 0.$$

($\lim_{N\to\infty} \Delta N > 0$). This is why, in the QD regime the number of preemptions converges to zero as N grows to infinity.

We have shown that in the three operational regimes the probability of the low-priority service interruption converges to zero, $\lim_{N\to\infty} P(preemption) = 0$. The number of preemptions is distributed $Geom_0(1 - P(preemption))$. This is why,

$$\lim_{N\to\infty} E(\#preemptions) = 0.$$

Thus, we can conclude that the expected waiting time of the lowest priority is asymptotically the same under the preemptive and non-preemptive priority disciplines:

$$\lim_{N\to\infty}\frac{E_{np}(W_q^K)}{E_{pr}(W_q^K)}=\lim_{N\to\infty}\frac{(E_\lambda(L_q)+1)\times E(B^{1\to(K-1)})}{(E_\lambda(L_q)+1)\times E(B^{1\to(K-1)})\times(1+E(\#preemptions))}=1, \quad (5.13)$$

where $E(B^{1\to(K-1)})$ is a Busy Period in an $M/M/1+M$ queue with arrival rate $\lambda^{1\to(K-1)}$, service rate $N\mu$ and abandonment rate θ.

Conventional Heavy Traffic

Now we assume that there is a single server and his utilization converges to 1. For simplicity, the following explanation addresses queues without abandonment. In the case with impatient customers the explanation below applies with minor changes.

Let ρ_N converge to 1 in the following manner: $\lim_{N\to\infty}\sqrt{N}(1-\rho_N)=c$ for some $0<c<\infty$.

Under **Non-Preemptive** discipline, to move forward one position in his queue, a delayed lowest-priority customer waits a single busy period of an $M/M/1$ queue with the service rate $N\mu$ and arrival rate λ_1.

As it was mentioned, both under **Preemptive and Non-Preemptive** disciplines, the queue length upon an arrival of the lowest priority customer, if there is waiting, is distributed like a queue length in an $M/M/1$ queue with the arrival rate $\lambda=\lambda_1+\lambda_2$.

Thus,

$$E_{np}(W_q^K)=\underbrace{\frac{1}{\mu(1-\rho_1)}}_{E(busy\ period),\ \lambda_1,\ \mu,1}\times\underbrace{\frac{1}{(1-\rho)}}_{E(L),\ \lambda,\ \mu,1}$$

Under **Preemptive Priority** discipline, the time until a delayed lowest-priority customer moves forward one position is distributed as a busy period of an $M/M/1$ queue with the service rate $N\mu$ and arrival rate λ_1, similarly to the non-preemptive discipline.

However, there are additional high-priority customers which arrive after the service of the second-type customer has begun making him return to the queue.

The probability of preemption is $P(preemption)=\frac{\lambda_1}{\lambda_1+\mu}\equiv\frac{\rho_1}{1+\rho_1}$, as the competition of two exponential random variables. As a result, the number of preemptions is distributed $Geom_0(1-\frac{\rho_1}{1+\rho_1})$, and the expected number of preemptions is ρ_1.

Thus,

$$E_{pr}(W_q^K)=\underbrace{\frac{1}{\mu(1-\rho_1)}}_{E(busy\ period),\ \lambda_1,\ \mu,1}\times\left[\underbrace{\frac{1}{(1-\rho)}}_{queue\ upon\ arrival}+\underbrace{\rho_1}_{\#preemptions}\right]$$

We notice, that there is a finite number (ρ_1) of the high-priority customers which enter the queue due to preemptions, while the queue length upon a low-priority customer arrival is exactly the same as under non-preemptive priority and diverges to infinity as ρ approaches 1.

As a result, it is possible to conclude that under conventional heavy traffic preemptive and non-preemptive

disciplines result in asymptotically the same expected waiting time for customers of the lowest priority:

$$\lim_{r\to\infty}\frac{E_{np}(W_q^K)}{E_{pr}(W_q^K)}=\lim_{r\to\infty}\frac{\frac{1}{\mu(1-\rho_1)}\times\frac{1}{(1-\rho)}}{\frac{1}{\mu(1-\rho_1)}\times\left(\frac{1}{(1-\rho)}+\rho_1\right)}=1. \tag{5.14}$$

Using the same arguments, it can be shown that for any finite number of types K, the preemptive and the non-preemptive priority disciplines are asymptotically equivalent for the customers of the lowest priority K.

The fact of asymptotic equivalence of the two disciplines is important because in many cases we may apply the known results for one priority discipline to another.

Now, using exact analysis, we are going to obtain convergence rate of the expected waiting time for the highest and for the lowest priorities under each discipline and see that the lowest priority is asymptotically the same under both priority disciplines, as predicted.

Let us consider a pair of $M/M/N+M$ queues under preemptive and non-preemptive priority disciplines. For asymptotic analysis, similarly to the analysis of an $M/M/N$ queue, we use the results of Zeltyn et al. [42] for the ED, QED and QD operational regimes and the exact formulae (5.15-5.18) listed below.

Equations (5.1 and 5.12) for the two customer types read as follows:

$$E_p(W_q^1) = \frac{1}{\theta}P_{\lambda_1}(Aband) \tag{5.15}$$

$$E_p(W_q^2) = \frac{1}{\theta}\frac{\lambda P_\lambda(Aband)-\lambda_1 P_{\lambda_1}(Aband)}{\lambda_2} \tag{5.16}$$

$$E_{np}(W_q^1) = \frac{P_\lambda(W_q>0)}{\theta}P_{\lambda_1}(Aband|W_q>0), \tag{5.17}$$

$$E_{np}(W_q^2) = \frac{P_\lambda(W_q>0)}{\theta}\cdot\frac{\lambda P_\lambda(Aband|W_q>0)-\lambda_1 P_{\lambda_1}(Aband|W_q>0)}{\lambda_2} \tag{5.18}$$

The next two subsections present an asymptotic analysis of queues with two customer types under both priority disciplines in ED and QED operational regimes. For each subsection the convergence is shown in the same way. First, to show, how Equations (5.15)-(5.18) converge as N grows to infinity, we check separately the convergence of their main components:

- $P_\lambda(W_q>0),\ \ P_\lambda(Aband|W_q>0),\ \ P_\lambda(Aband) \qquad (=P_\lambda(W_q>0)P_\lambda(Aband|W_q>0))$;
- $P_{\lambda_1}(W_q>0),\ \ P_{\lambda_1}(Aband|W_q>0),\ P_{\lambda_1}(Aband) \quad (=P_{\lambda_1}(W_q>0)P_{\lambda_1}(Aband|W_q>0))$.

After that we aggregate the results to find the convergence of (5.15-5.18).

5.3.2 QED: An Example with Two Customer Types

We assume that the total arrival rate converges to infinity, $\lambda \to \infty$, and that the total number of servers N is given by:

$$N \approx \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}}, \quad \lambda \to \infty, \qquad\qquad -\infty < \beta < \infty. \tag{5.19}$$

In addition, we assume that as λ grows to infinity, the offered load per server for each type of customers, ρ_i, stays constant:

$$\lim_{N\to\infty} \frac{\lambda_k}{N\mu} = \rho_k, \qquad k = 1,\ 2.$$

Now, to obtain the convergence of expressions (5.15)-(5.18), let us analyze the convergence of their main components.

1. The delay probability in the QED regime converges to α:

   $$\lim_{N\to\infty} P_\lambda(W_q > 0) = \alpha,$$

   where α is given by Garnett function [11]:

   $$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \tag{5.20}$$

   in which $\hat{\beta} \triangleq \beta\sqrt{\frac{\mu}{\theta}}$.

2. Formula (4.6) in [42] gives an approximation for the probability of abandoning, if there is waiting, in the QED regime:

   $$P_\lambda(Aband|W_q > 0) = \frac{1}{\sqrt{N}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot [h(\hat{\beta}) - \hat{\beta}] + o(1/N). \tag{5.21}$$

3. To analyze the convergence of $P_{\lambda_1}(W_q > 0)$ and $P_{\lambda_1}(Aband)$, we notice that in a queue with a single type of customers who arrive at rate λ_1 and the number of servers is determined by (5.19), the customers are served similarly to the QD regime. This is why, the number of the servers in such a queue can be described as follows:

   $$N \approx \frac{\lambda_1}{\mu} + \frac{\rho_2}{\rho_1} \cdot \frac{\lambda_1}{\mu}. \tag{5.22}$$

   The convergence of the delay and the abandonment probabilities in the QD regime is determined by Theorem 5.1 (a-b) in [42]:

   $$P_{\lambda_1}(W_q > 0) \approx \frac{1}{\sqrt{2\pi N}} \cdot \frac{1}{\delta}\left(\frac{1}{1+\delta}\right)^{N-1} \exp(\lambda_1\delta/\mu) \tag{5.23}$$

   $$\approx \frac{1}{\sqrt{2\pi N}} \cdot \frac{\rho_1^N}{\rho_2} \cdot e^{1-\rho_1}, \tag{5.24}$$

   $$P_{\lambda_1}(Aband|W_q > 0) = \frac{1}{N} \cdot \frac{1}{1-\rho_1} \cdot \frac{\theta}{\mu} + o(1/N). \tag{5.25}$$

Now, after the convergence of each component has been determined, it is easy to obtain the convergence of the waiting time of any type under both priority disciplines.

Preemptive discipline

$$E_{pr}(W_q^1) \cdot \frac{N\sqrt{N}}{(\rho_1)^N} = \frac{e^{\rho_2}}{\sqrt{2\pi}\,\rho_2^2\mu} \tag{5.26}$$

$$E_{pr}(W_q^2) \cdot \sqrt{N} = \frac{\alpha}{\rho_2\sqrt{\theta\mu}} \cdot [h(\hat{\beta}) - \hat{\beta}]. \tag{5.27}$$

We see that the waiting time of the first type customers converges to zero at rate $\Theta(\frac{\rho_1^N}{N\sqrt{N}})$, as expected according to the QD regime. The service level of the lowest-priority customers fits the QED regime, and their waiting time converges to 0 at rate $\Theta(\frac{1}{\sqrt{N}})$.

Non-Preemptive Discipline

$$E_{np}(W_q^1) \cdot N = \frac{\alpha}{\rho_2\mu} \tag{5.28}$$

$$E_{np}(W_q^2) \cdot \sqrt{N} = \frac{\alpha}{\rho_2\sqrt{\theta\mu}} \cdot [h(\hat{\beta}) - \hat{\beta}]. \tag{5.29}$$

Under the non-preemptive discipline, the highest priority experience QD service level "conditioned on waiting", i.e. the highest priority enjoy service before the lowest priority, but they cannot interrupt in-process service. This is why their waiting time converges to zero at rate $\Theta(1/N)$. This rate is faster than the convergence under QED in a queue without priorities but not as fast as under the preemptive priority discipline.

The waiting time of the lowest priority converges to zero at rate $\Theta(\frac{1}{\sqrt{N}})$, just like under preemptive priority. Moreover, the ratio between the expected waiting time and $\frac{1}{\sqrt{N}}$ converges to the same limit as under preemptive priority.

5.3.3 ED: An Example with Two Customer Types

We assume that the total arrival rate converges to infinity $\lambda \to \infty$ and that the total number of servers N is given by:

$$N = \frac{\lambda}{\mu}(1-\gamma), \quad \text{for some } 0 < \gamma < \infty. \tag{5.30}$$

Again, the assumption is that as λ grows to infinity, the fraction of time spent with each type of customer, ρ_i, stays constant.

Now, to obtain the convergence of expressions (5.15)-(5.18), let us analyze the convergence of their main components.

1. The delay probability in the ED regime converges to 1:

$$\lim_{N\to\infty} P_\lambda(W_q > 0) = 1.$$

2. Formula (6.5) in [42] gives an approximation of the probability of abandoning, if there is waiting, in the ED regime:

$$P_\lambda(Aband) \approx \gamma. \tag{5.31}$$

3. To analyze the convergence of $P_{\lambda_1}(W_q > 0)$ and $P_{\lambda_1}(Aband)$, we again use QD approximations, using the same arguments as in the case of QED. The convergence of the delay and the abandonment probabilities in the QD regime is determined by Theorem 5.1 (a-b) in [42]:

$$\begin{aligned} P_{\lambda_1}(W_q > 0) &\approx \frac{1}{\sqrt{2\pi N}} \cdot \frac{1}{\delta}\left(\frac{1}{1+\delta}\right)^{N-1} \exp(\lambda_1 \delta/\mu) && (5.32)\\ &\approx \frac{1}{\sqrt{2\pi N}} \cdot \frac{\rho_1^N}{\rho_2} \cdot e^{1-\rho_1}, && (5.33)\\ P_{\lambda_1}(Aband|W_q > 0) &= \frac{1}{N} \cdot \frac{1}{1-\rho_1} \cdot \frac{\theta}{\mu} + o(1/N). && (5.34) \end{aligned}$$

Preemptive discipline

$$\begin{aligned} \lim_{N\to\infty} E_{pr}(W_q^1) \frac{N\sqrt{N}}{\rho_1^N} &= \frac{e^{\rho_2}}{\sqrt{2\pi}\rho_2^2\,\mu}, && (5.35)\\ \lim_{N\to\infty} E_{pr}(W_q^2) &= \lim_{N\to\infty} \frac{\rho\gamma}{\theta\rho_2} = \frac{\gamma}{\theta\rho_2}. && (5.36) \end{aligned}$$

We see that both in Erlang-C and Erlang-A queues in QED and ED regimes, the convergence rate of $E_{pr}(W_q^1)$ is $\Theta(\frac{\rho_1^N}{N\sqrt{N}})$

The convergence rate of $E_{pr}(W_q^2)$ is $\Theta(1)$.

Non-Preemptive Discipline

$$\begin{aligned} \lim_{N\to\infty} N E_{np}(W_q^1) &= \frac{1}{\rho_2\mu}, && (5.37)\\ \lim_{N\to\infty} E_{np}(W_q^2) &= \frac{\gamma}{\theta\rho_2}. && (5.38) \end{aligned}$$

The convergence rate of the expected waiting time of the highest priority under non-preemptive discipline is $\Theta(1/N)$, which is the same for all previously considered examples ($M/M/N$ under QED and ED and $M/M/N+M$ under QED).

The convergence rate of $E_{np}(W_q^2)$ is $\Theta(1)$, which is the same rate as under preemptive priority. Additionally, $E_{np}(W_q^2)$ itself converges to the same limit as under preemptive priority.

5.3.4 Summary of Convergence Rates

To emphasize the similarity of the preemptive and non-preemptive disciplines for the lowest priority, Table 5.1 presents a summary of the waiting-time convergence rates for the two-type examples of Erlang-C

/ A queues.

This table shows that the first-type customers are not sensitive to the changes in the operational regime. The convergence rate of the preemptive priority remains exponential under QED and ED in queues with and without abandonment. And under non-preemptive priority, the convergence rate of the first type is $\Theta(1/N)$ both for QED and ED regimes.

The performance of the lowest priority is influenced by the operational regime and not by the priority discipline. Table 5.1 shows that in the same operational regime, the convergence rate of the lowest priority both under preemptive and non-preemptive disciplines is similar to that in queues without priorities.

Table 5.1: Convergence Rates Under Both Disciplines in Queues with and without Abandonment

	QED $N = R + \beta\sqrt{R}$		ED $N = R - \gamma R$	
	$M/M/N$	$M/M/N+M$	$M/M/N$	$M/M/N+M$
$E_{np}(W_q^1)$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$	$\Theta(1/N)$
$E_{np}(W_q^2)$	$\Theta(1/\sqrt{N})$	$\Theta(1/\sqrt{N})$	$\Theta(1)$	$\Theta(1)$
$E_{pr}(W_q^1)$	$\Theta(\frac{\rho_1^N}{N\sqrt{N}})$	$\Theta(\frac{\rho_1^N}{N\sqrt{N}})$	$\Theta(\frac{\rho_1^N}{N\sqrt{N}})$	$\Theta(\frac{\rho_1^N}{N\sqrt{N}})$
$E_{pr}(W_q^2)$	$\Theta(1/\sqrt{N})$	$\Theta(1/\sqrt{N})$	$\Theta(1)$	$\Theta(1)$

5.4 Higher Priorities: Convergence of the Expected Waiting Time

This section deals with the convergence of the expected waiting time of any priority k higher than the lowest priority K, under the QED and ED operational regimes. We also assume that the lowest priority is not negligible, $\lim_{N\to\infty} \rho_K > 0$, and that the total number of customers of the types $1 \dots k$ is not negligible, $\lim_{N\to\infty} \sigma_k > 0$.

5.4.1 Preemptive Priority

Let us start with the preemptive-priority discipline. The expected waiting time of type k is given by recursion (4.5), which is the same for Erlang-C and Erlang-A queues. This equation is re-stated below for Erlang-A queues using the relation (2.13) between the abandonment probability and the expected

waiting time:

$$E_{pr}(W_q^k) = \frac{1}{\theta \cdot \lambda_k}\Big[\lambda_{1\to k} P(Aband_q^{(1\to k)}) - \lambda_{1\to(k-1)} P(Aband_q^{(1\to(k-1))})\Big], \qquad k = 1 \dots K. \quad (5.39)$$

In the case of Erlang-C queues, the expected waiting time is given by (4.6), which is repeated here for convenience of the reader:

$$E_{pr}(W_q^k) = \frac{E_{2,N}(\lambda_{1\to k})}{\lambda_k(1-\sigma_k)} - \frac{E_{2,N}(\lambda_{1\to(k-1)})}{\lambda_k(1-\sigma_{k-1})}.$$

If the number of servers is determined by the ED or QED staffing rule, then Erlang-C or A queueing systems with the arrival rate $\lambda_{1\to k}$ or $\lambda_{1\to(k-1)}$ experience service under light traffic (QD). This is why we apply Theorem 5.1 (a-b) from [42] to determine the convergence of $E_{2,N}(\lambda_{1\to k}),\ E_{2,N}(\lambda_{1\to(k-1)}),$ $P(Aband^{1\to k}),$ and $P(Aband^{1\to(k-1)})$:

$$E_{2,N}(\lambda_{1\to k}) = P(W_q^{1\to k} > 0) \ \approx \ \frac{1}{\sqrt{2\pi N}} \cdot \frac{\sigma_k^N}{1-\sigma_k} \cdot e^{1-\sigma_k}, \quad (5.40)$$

$$P(Aband^{1\to k} | W_q^{1\to k} > 0) \ = \ \frac{1}{N} \cdot \frac{1}{1-\sigma_k} \cdot \frac{\theta}{\mu} + o(1/N). \quad (5.41)$$

Utilizing results (5.40-5.41), we see that the convergence rate of $E_{pr}(W_q^k)$ is $\Theta(\frac{\sigma_k^N}{\sqrt{N^3}})$, and it is the same for Erlang-C and Erlang-A queues:

$$\lim_{N\to\infty} E_{pr}(W_q^k) \cdot \frac{\sqrt{N^3}}{\sigma_k^N} = \frac{e^{1-\sigma_k}}{\sqrt{2\pi}(1-\sigma_k)^2\mu}, \qquad k = 1, \dots, K-1. \quad (5.42)$$

Moreover, the ratio between $E_{pr}(W_q^k)$ and its convergence rate converges to the same constant (see (5.42)) for both Erlang-C and Erlang-A queues. This means that **the expected waiting time of any priority k, $k < K$, can be approximated by Erlang-C model with the arrival rate $\lambda_{1\to k}$.**

5.4.2 Non-Preemptive Priority Discipline

The general expression for the expected waiting time of any type k is found using recursion (5.12). It is repeated below for convenience:

$$E_{np}(W_q^k) \ = \ \frac{\lambda_{1\to k} \cdot E_{np}(W_q^{1\to k}) - \lambda_{1\to(k-1)} E_{np}(W_q^{1\to(k-1)})}{\lambda_k}.$$

Several algebraic operations on (5.12) and the relation (2.13) lead to the following result:

$$E_{np}(W_q^k) \ = \ \frac{P_\lambda(W_q > 0)}{\lambda_k\ \theta}\Big(\lambda_{1\to k} P(Aband^{1\to k} | W_q > 0) - \lambda_{1\to(k-1)} P(Aband^{1\to(k-1)} | W_q > 0)\Big).$$

Now let us analyze the convergence of the expected waiting time of type k in the QED and ED operational regimes while the number of servers increases indefinitely. The expected waiting time of any type k is a product of the two following elements:

- $P_\lambda(W_q > 0)$,
- $E_{np}(W_q^k|W_q^k > 0) = \frac{1}{\lambda_k}\Big(\lambda_{1\to k}\frac{P(Aband^{1\to k}|W_q > 0)}{\theta} - \lambda_{1\to(k-1)}\frac{P(Aband^{1\to(k-1)}|W_q > 0)}{\theta}\Big)$.

We check separately the convergence of each of them and later combine the results to make a conclusion about $\lim_{N\to\infty} E_{np}(W_q^k)$.

In the QED regime, the delay probability converges to some positive constant α less than 1 (Garnett Function (5.20)):

$$P_\lambda(W_q > 0) = \alpha + o(1/\sqrt{R_N}). \tag{5.43}$$

This can be concluded from our analysis of the delay probability in the QED regime in Subsection 3.4.1. In the ED regime the delay probability converges to 1:

$$P_\lambda(W_q > 0) = 1 + o(1/N). \tag{5.44}$$

This is concluded from the analysis of the delay probability in the ED regime in Subsection 3.4.2.

To analyze the convergence of the expected waiting time, if there is waiting of type k customers we use the approximations of the abandonment probability in the QD regime, developed by Zeltyn (see [42], Theorem 5.1 (b)). This is made possible due to the assumption that the lowest priority is not negligible:

$$\begin{aligned} P_(Aband^{1\to k}|W_q^{1\to k} > 0) &= \frac{1}{N}\cdot\frac{1}{1-\sigma_k}\cdot\frac{\theta}{\mu} + o(1/N) \\ P(Aband^{1\to(k-1)}|W_q^{1\to k} > 0) &= \frac{1}{N}\cdot\frac{1}{1-\sigma_{k-1}}\cdot\frac{\theta}{\mu} + o(1/N) \end{aligned}$$

Substituting these approximations into conditioned waiting time, we obtain the following:

$$\begin{aligned} E_{np}(W_q^k|W_q^k > 0) &= \frac{1}{\lambda_k\ \theta}\Big(\frac{\lambda_{1\to k}\ \theta}{N\mu(1-\sigma_k)} - \frac{\lambda_{1\to(k-1)}\ \theta}{N\mu(1-\sigma_{k-1})}\Big) + o(1/N) \\ &= \frac{1}{\lambda_k}\Big(\frac{\sigma_k}{1-\sigma_k} - \frac{\sigma_{k-1}}{1-\sigma_{k-1}}\Big) + o(1/N) \\ &= \frac{\sigma_k - \sigma_k\sigma_{k-1} - \sigma_{k-1} + \sigma_k\sigma_{k-1}}{\lambda_k(1-\sigma_k)(1-\sigma_{k-1})} + o(1/N) \end{aligned}$$

$$\Rightarrow \quad E_{np}(W_q^k|W_q^k > 0) = \frac{1}{N\mu(1-\sigma_k)(1-\sigma_{k-1})} + o(1/N). \tag{5.45}$$

Let us now combine the convergence results of the delay probability and the expected waiting time given waiting. For the QED regime, the relevant results are equations (5.43) and (5.45), and for the ED regime the expected waiting time is a combination of (5.44) and (5.45). One can easily see, that under QED

or ED, the expected waiting time of type k, $k < K$, which is presented as a limit of a product of two elements, is equal to the product of the limits of these elements:

$$\Rightarrow \quad \lim_{N\to\infty} E_{np}(W_q^k) = \lim_{N\to\infty} P_\lambda(W_q > 0) \cdot \lim_{N\to\infty} \frac{1}{N\mu(1-\sigma_k)(1-\sigma_{k-1})}. \tag{5.46}$$

Note that the expected waiting time of any type k, $k > K$, is given asymptotically by the same expression as for $M/M/N$ queues (see (4.7)) with the only difference that the delay probability is given by the Erlang-A formula, and not by the Erlang-C one.

By using (5.46) we find the convergence rate of the expected waiting time for any type k, $k < K$. It can be seen that under both the **QED** and the **ED** operational regimes, the convergence rate is $\Theta(\frac{1}{N})$, since the delay probability converges to some constant ($0 < \alpha < 1$ for the QED and 1 for the ED).

Chapter 6

Towards Time-Stable Performance of Time-Varying Call Centers

In this chapter, we show how **stationary** models can be used in a **time-varying** environment to help determine an appropriate staffing level. This is done by analyzing via simulation four different call centers. For each center, the staffing level is determined by either a simulation-based algorithm (ISA) or by the square-root safety-staffing rule.[1]

In addition, we check and compare the performance of the two staffing methods, PSA and Lagged PSA, the first of which is widely used in industry.

The chapter is organized as follows.

A detailed explanation of the ISA algorithm, which determines staffing level for a given target delay probability, is provided in Section 6.1. Section 6.2 lists different performance measures that can be calculated using our simulation tool. Section 6.3 gives a short summary of the prevalent staffing methods, PSA and Lagged PSA. Section 6.4 presents simulated performance of the four call centers, the staffing for which is determined with the help of PSA, Lagged-PSA and square-root staffing.
Appendix 6.5 describes the current implementation of the performance measures listed in Section 6.2, as well as alternative ways of their calculation.

6.1 Description of the ISA algorithm

Here we describe the simulation-based Iterative-Staffing Algorithm (ISA). In our implementation, the algorithm determines time-dependent staffing levels aiming to achieve a given constant-over-time delay probability at all time intervals.

For the implementation of the algorithm we assume that we have an $M_t/G/s_t$ system with homo-

[1] For each of these call centers both ISA and the square-root safety-staffing rule were applied, and the differences between these two staffing methods were found to be practically insignificant.

geneous customers. We assume that service times are iid variables of a given general distribution, which are independent of arrivals. The Poisson arrivals are fully specified by their arrival rate function $\{\lambda(t);\ \ 0 \le t \le T\}$.

To start, we fix an arrival-rate function, a service-time distribution and a time horizon $[0, T]$. For any random quantity of interest, let $X_t^{(i)}$ denote its value at time t in the i^{th} iteration, $t \in [0, T]$. Although the algorithm is formulated in continuous time, staffing decisions are made at discrete times. This is achieved by dividing the time-horizon into small intervals of length δ. The number of servers is constant within each interval.

Let $s_t^{(i)}$ be the staffing level at time t in iteration i, for $0 \le t \le T$. Let $L_t^{(i)}$ denote the random number of customers in the system at time t under this staffing function. We estimate the distributions of $L_t^{(i)}$ for each i and t, by performing multiple (5000) independent replications. The algorithm starts with infinitely many servers ($s_t^{(0)} \equiv \infty$). In this implementation we choose a large finite number of servers which guarantees a negligible delay probability.

The algorithm iteratively performs the following steps, until convergence is obtained. (Here, convergence means that the staffing levels do not change much after an iteration. By the algorithm implementation, they are allowed to change by some threshold τ, which we took to be 1.)

1. Given the i^{th} staffing function $\{s_t^{(i)} : 0 \le t \le T\}$, evaluate the distributions of $L_t^{(i)}$ for all t, using the simulation.

2. For each $t,\ \ 0 \le t \le T$, let $s_t^{(i+1)}$ be the least number of servers such that the delay probability constraint is met at time t,[2] i.e. let

$$s_t^{(i+1)} = argmin\{c \in \mathbb{N} : P\Big(L_t^{2,(i)} \ge c\Big) \le \alpha\}. \tag{6.1}$$

3. If there is a negligible change in the staffing from iteration i to iteration $i+1$, then stop. Formally,

$$||s^{(i+1)} - s^{(i)}||_\infty \equiv max\{|s_t^{(i+1)} - s_t^{(i)}| : 0 \le t \le T\} \le \tau, \tag{6.2}$$

   then stop and let $s^{(i+1)}$ be the proposed staffing function. Otherwise, advance to the next iteration, i.e., increase i to $i+1$ and go back to step 1.

Let ∞ denote the index of the last iteration of ISA, so that $s_t^{(\infty)}$ denotes the final staffing level at time t and $L_t^{(\infty)}$ denotes the random number of the customers in system at time t, with the obtained staffing function $s^{(\infty)}$. Then if convergence is reached, the determined staffing function satisfies the following: $P\Big(L_t^{(\infty)} \le s_t^{(\infty)}\Big) \approx \alpha,$ for all $0 \le t \le T$.

The implementation of the algorithm is written in C++ and is an adaptation of an existing software written by Z. Feldman (see [8] or [9]).

[2] We take the event that "all servers are busy at time t" to mean that a (virtual) arrival at time t would be delayed before service.

6.2 Calculation of performance measures

Here we present a list of the performance measures implemented in our simulation software. The implementation of these measures is discussed in Appendix 6.5.

The ISA algorithm ([8] and [9]) allows one to calculate these measures over a pre-determined time horizon, partition intervals and number of replications which may vary depending on the user choice.

In the formulae below:

- superscript *j* indicates the j^{th} replication;
- subscript *t* indicates the t^{th} partition interval; all intervals are of size ΔT;
- *Reps* is the total number of replications.

The measures listed below are calculated for each time interval separately. They are referred to as *Dynamic Performance Measures*, to emphasize their time-dependence.

- R_t: Offered Load;
- β_t: Implied Service Quality;
- ρ_t: Servers Utilization (Fraction of Time Serving Customers);
- $P_t(Aband)$: Abandonment Probability;
- $P_t(W_q > 0)$: Delay Probability;
- Q_t: Average Queue Length;
- W_t: Average Waiting Time;
- $E(W_t|W_t > 0)$: Average Waiting Time Conditioned on Waiting.

The following measures, referred to as *Overall Performance Measures*, are calculated at the end of the simulation run.

- $E(W)$: Average Waiting Time;
- $E(W|W > 0)$: Average Waiting Time Conditioned on Waiting;
- $P(Aband)$: Average Abandonment Probability;
- $P(W_q > 0)$: Average Delay Probability;
- ρ: Average Servers Utilization.

6.3 Short Staffing Intervals: PSA and Lagged PSA Approximations

In the following two subsections, we present two techniques for staffing time-varying queues, PSA and Lagged PSA, as described in the article by Green et al. [14].

6.3.1 PSA and SIPP

There are several approaches for coping with time-varying arrivals. The traditional solution for staffing a queue with short service times and a high quality of service is the *Pointwise Stationary Approximation (PSA)*. This approximation describes a time-dependent queue at each time t *using a stationary model with the arrival rate and other parameters of this time t.*

In practice the number of servers stays constant during time intervals, which we call here staffing intervals. The PSA method can be adapted to such conditions by *Segmented PSA*. The latter determines staffing for each staffing interval as the maximum of all PSA-generated staffing levels in this interval. In general, this method tends to overstaff slightly, but its results can be refined by simulations.

In practice, many commercial software packages use the following approach: The arrival rates are first averaged over the whole staffing interval, and the staffing level in that interval is set according to a corresponding stationary model. This method is referred to as *Stationary Independent Period-by-Period (SIPP)* [13].

Both Segmented PSA and SIPP are based on the same principle and assume that all time periods are independent. The difference is as follows. With Segmented PSA we first find *all* possible staffing levels for all arrival rates during some staffing interval, while with SIPP we average arrival rates for this interval and find the appropriate staffing only once.

6.3.2 The Lagged Pointwise Stationary Approximation

While PSA methods perform well for fast service rates, for medium to low service rates some adjustment may be needed. An intuitive explanation is that each customer stays in the system during his service time, hence the number of customers in system lags behind the arrival rate. In such a case, staffing by the arrival rate at every given moment is not very accurate because during the lag period the number of customers in the system may change.

A very good example of the lag impact is described by Litvak et al. in their report [23] for an emergency department in Massachusetts. Their main finding, obtained exclusively by observations, is that there is a lag between the arrivals of ambulances and the demand for doctors. There was no observations of what would have happened if the number of doctors and nurses lagged behind the arrival rate for six hours, the latter being an average service time in this ED. We believe, that if the observations were continued, the authors would have found out that this lag in staffing lead to a much improved level of performance.

The **Infinite-Server Model** provides insight into the staffing problem. Indeed, we use the $M_t/GI/s_t +$

GI model with medium-to-low service rates and a high-quality-of-service standard to describe our environment. Its corresponding infinite-server model is $M_t/GI/\infty$. This model allows one to find the number of assigned servers with no resource constraints. The distribution of the number of busy servers at each time *t* in the $M_t/GI/\infty$ queue can be found analytically. Although in our original queue the number of servers is not infinite, the associated Infinity-Server model is nevertheless intimately related to it, as will become clear in the sequel.

Solution for the $M_t/GI/\infty$ Model. The number of busy servers at time *t* in the $M_t/GI/\infty$ model has a Poisson distribution with time-varying mean $m_\infty(t)$, which can be expressed in the following 3 ways:

$$m_\infty(t) = E[\lambda(t-S_e)]E[S] = E\Big[\int_{t-S}^{t}\lambda(u)du\Big] = \int_{-\infty}^{t}[1-G(t-u)]\lambda(u)du, \tag{6.3}$$

where
S is the service time with cdf G,
and S_e is a random variable with the residual lifetime cdf associated with S, i.e.

$$P(S_e \le t) \equiv \frac{1}{E[S]}\int_0^t [1-G(u)]du, \quad t \ge 0. \tag{6.4}$$

From the representation (6.3), we see that the number of busy servers depends on the arrivals during the latest service time. This fact provides a theoretical support to the results described in [23] and also explains why applying Lagged-PSA staffing generally leads to better results than PSA, as will be seen from our simulation experiments presented further.

6.4 Empirical Examples

Here we compare the results of square-root staffing with the results of staffing by PSA and Lagged-PSA, applying these methods to four different call-centers. The description of the results for each center is organized in a very similar manner, and, to make the description easier to follow, we repeat in each subsection the same theoretical formulae.

For each empirical example the results are presented in the following way. Each subsection starts with a detailed summary of performance under square-root staffing, which is compared against an appropriate stationary model. Then the results of staffing by PSA and Lagged PSA are presented.

6.4.1 First Empirical Example - Green, Kolesar and Soares [13]

The $M_t/M/s_t + M$ queue presented here was originally studied by Green, Kolesar and Soares in [13]. The simulated environment is as follows:

- The running horizon is 24 hours.

- Performance statistics are calculated over the period of time between 6 a.m. and 17 p.m. to make sure that the arrival rate is large enough, so that QED approximations are applicable. (Our results actually reveal the time-period over which QED approximations are applicable.)
- All empirical values are calculated as an average over 5000 sample paths (simulations).
- Service time is assumed to be distributed exponentially, with mean $1/\mu$ = 0.1 hours or 6 minutes. (This is given in [13].)
- Customers patience is assumed to be distributed exponentially, with mean $1/\theta$ = 0.1 hours or 6 minutes. (There is no account of abandonment in [13].)
- The queue discipline is assumed FCFS.
- The arrival rate function is presented in Figures 6.1 and 6.2 (being adopted from [13]).

Square-Root Staffing

The goal of the experiments is to achieve time-stable performance in the face of time-varying arrivals. We are using staffing by constant value of the Quality-of-Service (QoS) parameter β which determines an appropriate staffing level, fixed over the staffing interval, for each time interval as follows:

$$s_t = R_t + \beta\sqrt{R_t}. \tag{6.5}$$

Here, $\{R_t,\ 0 \le t \le 24\}$ is the time-varying average number of customers (= busy servers) in an $M_t/M/\infty$ queue (5000 sample paths), with the arrival rate as in Figure 6.2 and average service time of 6 minutes.

We tested 11 values of β, from 2 to -2 in step 0.4, focusing on $\beta = 1.2$, 0 and -1.2; the latter 3 values correspond to the QD, QED and ED operational regimes respectively.
The results will now be elaborated on.

Staffing according to (6.5), achieves a time-stable level of the delay probability. Summary of the delay probabilities is presented in Figure 6.3, and its stability from 6 a.m. to 17 p.m. is remarkable.

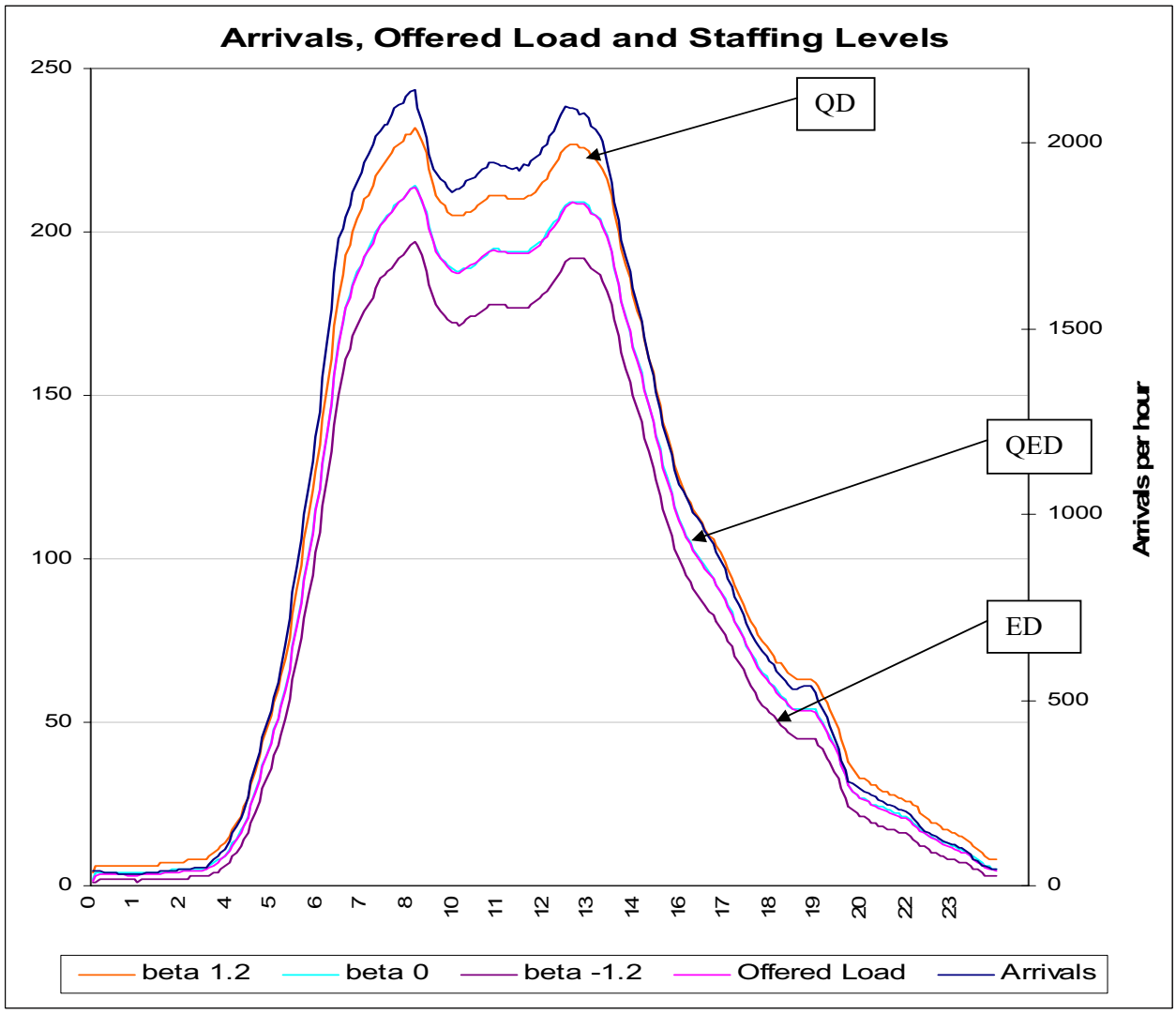
Figure 6.4 shows a comparison of the simulated overall (global) delay probability and the probability of waiting more than 30 seconds, if there is waiting, with the theoretical probabilities found by an appropriate stationary models. The theoretical global delay probability for a constant β is found by the Garnett function [11]:

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty, \tag{6.6}$$

where $\hat{\beta} = \beta\sqrt{\mu/\theta}$. (Note that only β is required on this case, since we have assumed that $\mu = \theta$.)

All the theoretical calculations are based on the stationary $M/M/s+M$ model with constant arrival rate and constant number of servers, both calculated in the following way: for each of the 5000 sample paths,

Figure 6.1: Arrivals, Offered Load and Staffing Levels



we calculated the average arrival rate and the average number of servers over the period of time from 6 a.m. till 17 p.m., and then averaged the 5000 averages.

The theoretical value of the probability of waiting more than 30 seconds, if there is waiting, is based on this stationary model and reads as follows (see [42]):

$$P\left\{W_q > \frac{t}{\sqrt{s}} | W_q > 0\right\} \approx \frac{\bar{\Phi}(\hat{\beta} + \sqrt{\theta\mu}\, t)}{\bar{\Phi}(\hat{\beta})} \quad . \tag{6.7}$$

Here

$\Phi(x)$ is the cumulative distribution function of the standard normal distribution (mean=0, std=1),

Figure 6.2: Arrivals, Offered Load and Staffing Levels

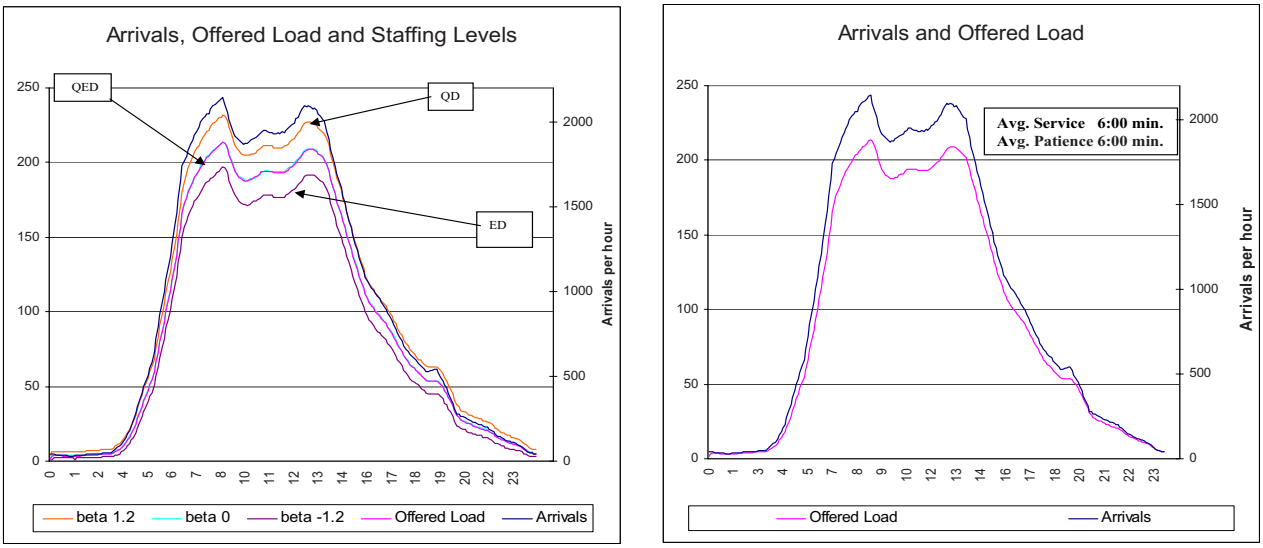
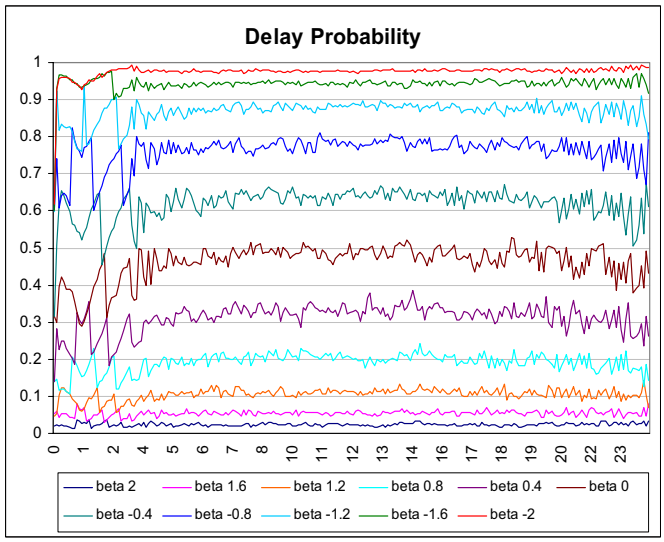


Figure 6.3: Stable Delay Probability



$\bar{\Phi}(x)$ is the survival function $(\Phi(x) = 1 - \bar{\Phi}(x))$,
$\phi(x)$ is the density of the standard normal distribution,
$h(x) \triangleq \phi(x)/\bar{\Phi}(x)$ is the hazard rate of the standard normal distribution.

Note that, in contrast to Equation (6.6), for calculating (6.7) one must specify s, μ, θ separately.

Figure 6.5 presents the plots of the abandonment probability and of the expected waiting time, for all eleven values of the tested β, and compares each simulated measure with its theoretical value. To find

Figure 6.4: Global Performance (6:00-17:00) of (1) Delay Probability; (2) Probability of Waiting More than 30 sec., if there is Waiting

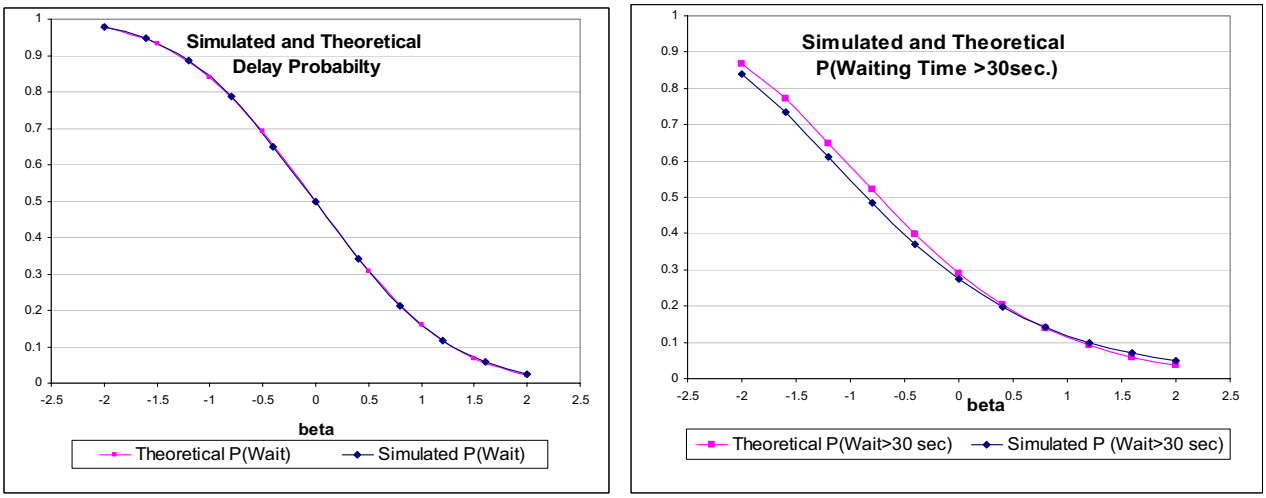
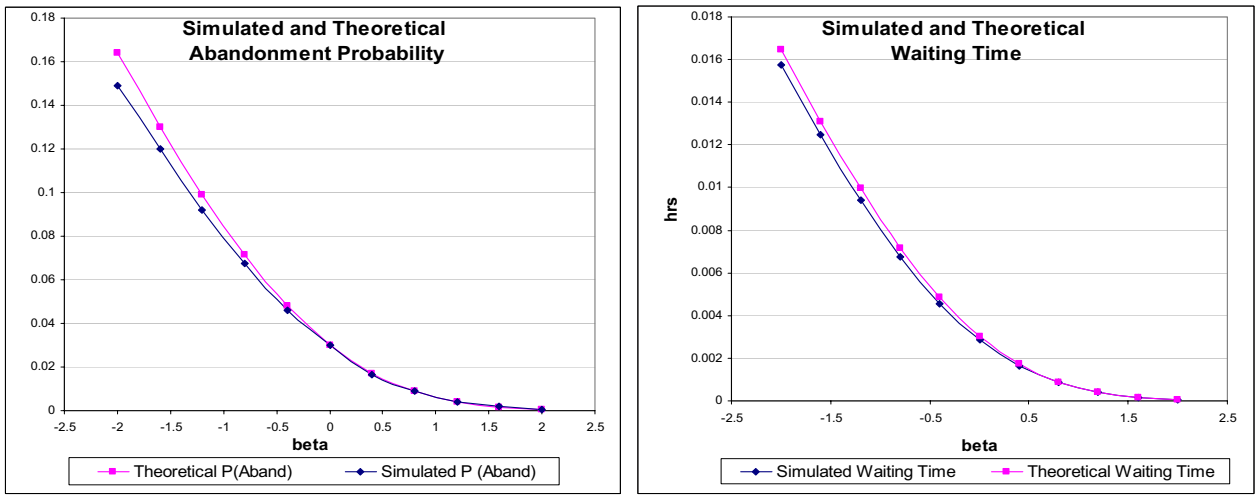


Figure 6.5: Global (6:00-17:00) Performance of (1) Abandonment Probability; (2) Average Waiting Time



the theoretical values we used again the results from [42]:

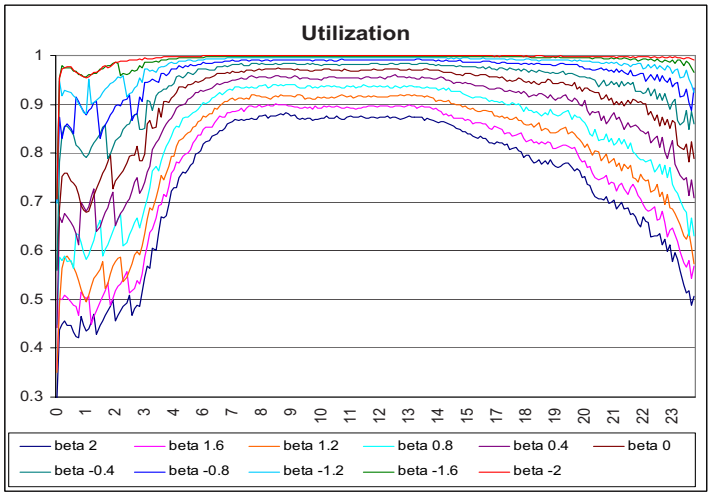
$$P(Aband) \approx \frac{1}{\sqrt{s}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \tag{6.8}$$

$$E(W_q) \approx \frac{1}{\sqrt{s}} \cdot \frac{1}{\theta} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \quad \left(= \frac{1}{\theta} P(Aband)\right) \tag{6.9}$$

The calculation of these theoretical values is based on the same stationary $M/M/s+M$ model with the constant arrival rate and number of servers, averaged as described above.

Server utilization is also relatively constant during the hours from 6:00 till 17:00: see Figure 6.6. In this

Figure 6.6: Average Servers Utilization



example, it takes more time for this measure to stabilize for higher values of β (lower target delay probability). This situation is different from that with the probability of abandoning, which is more stable for small target delay probabilities (compare with Figure 6.8).

Average values of waiting time and queue length under the three operational regimes are presented in Figure 6.7.

The dynamics of the abandonment probability is presented in Figure 6.8. One can see from the zoomed plot that for high values of β this probability is low and stable over the whole simulation period, while for close to 0 and negative values of β it stabilizes only during the period from 6:00 till 17:00.

Histograms of the waiting time, if there is waiting, for the three operational regimes are presented in Figure 6.9. The theoretical graphs were created via (6.7) using the average number of servers during the period from 6:00 till 17:00. The fit between practice and theory is again remarkable.

PSA and Lagged PSA

Figure 6.10 presents delay probabilities obtained by applying PSA and Lagged PSA and by the ISA algorithm. The average service time is rather short (6 minutes), thus significant differences in performance

Figure 6.7: Average Waiting Time and Queue Length for QD, QED and ED Regimes

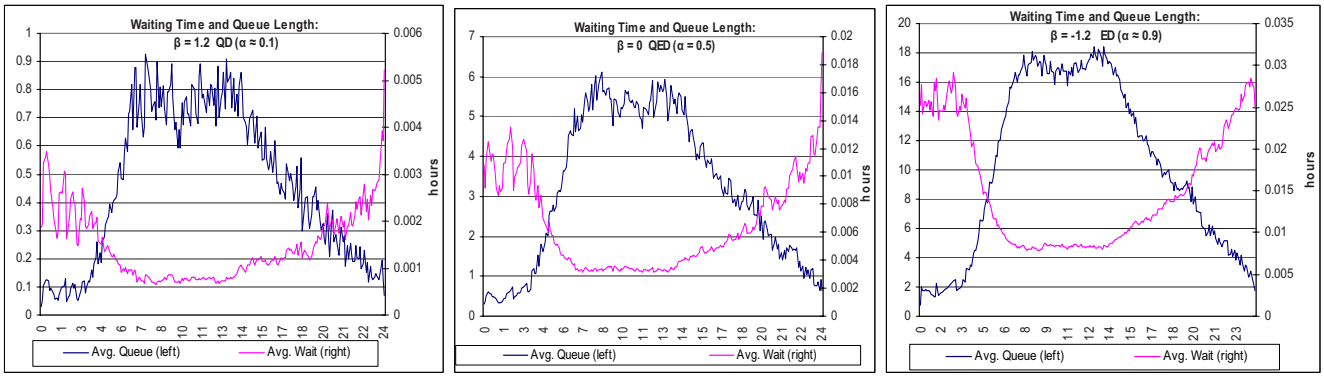
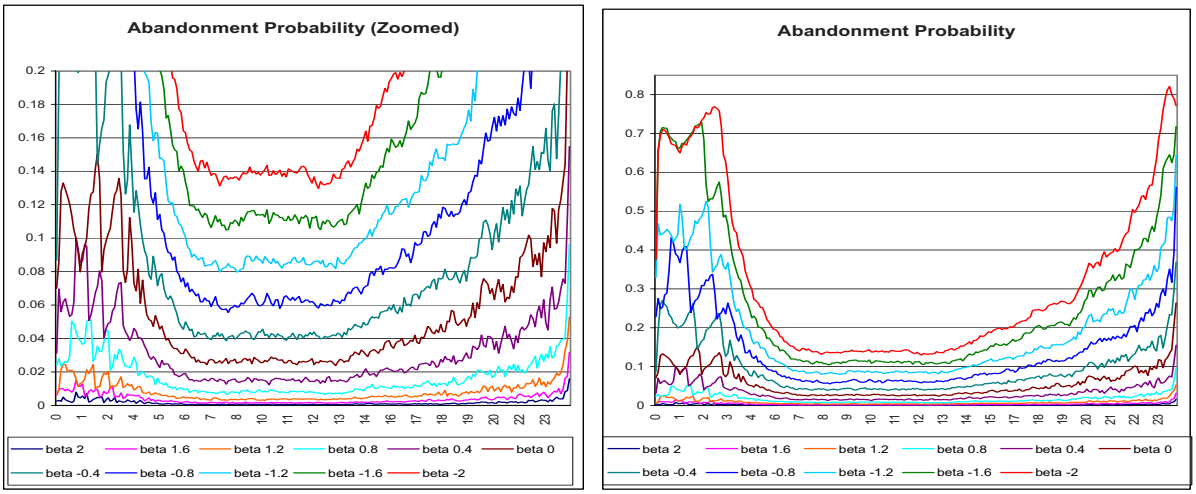


Figure 6.8: Dynamics of Abandonment Probability



between these three methods in the levels of the delay probabilities in the morning hours (from 4 till 6) and in the afternoon (from 14 to 20) are rather unexpected. During different periods of the day, where the arrival rate is changing very fast, the definition of ”short” service time varies, which results in over-staffing in the morning hours, and understaffing in evening hours.

Indeed, let us consider time $t = 4.1$ with target delay probability $\alpha = 0.5$ (See Figure 6.10). The delay probability obtained by the PSA staffing is 0.181 instead of the target 0.5. This happens because of the PSA assumption that the number of customers in the system at time t is given by the arrival rate $\lambda(t) = 240.4$ customers per hour. Thus, the solution suggested by PSA is 25 servers, and it takes in

Figure 6.9: Waiting Time, if there is Waiting: Empirical vs. Theoretical Distribution

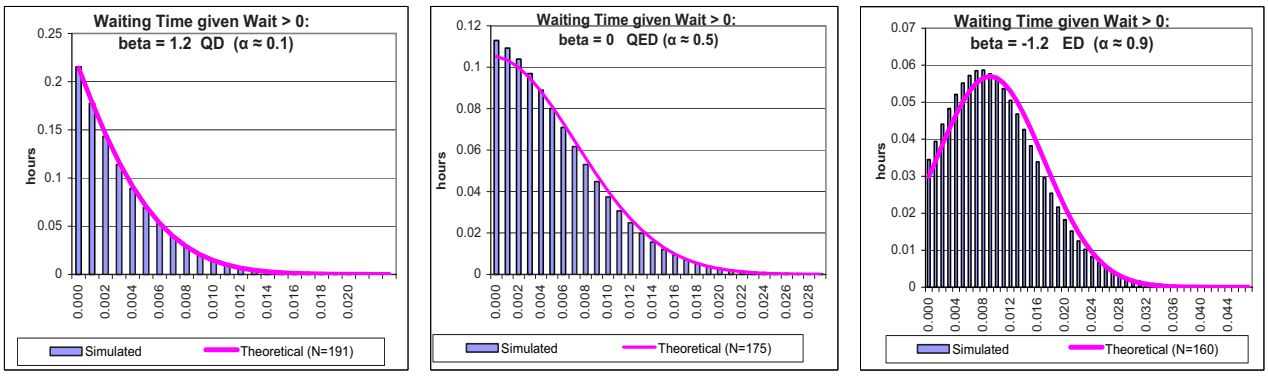
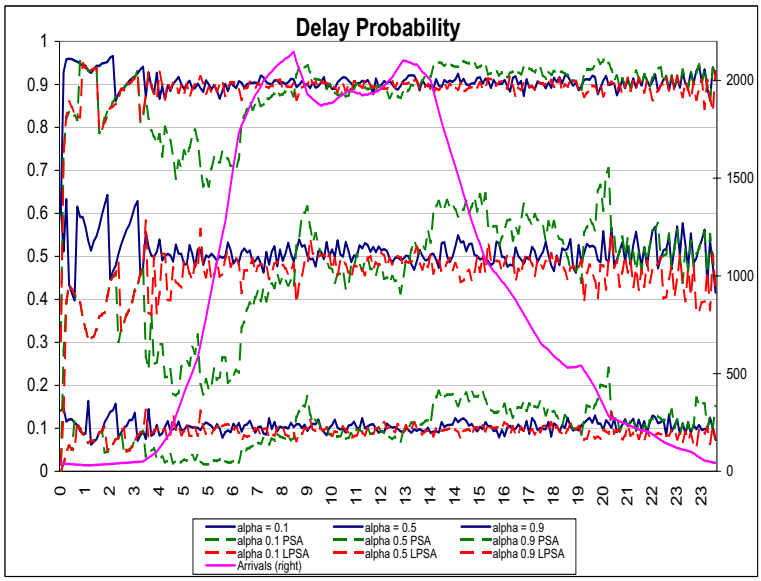


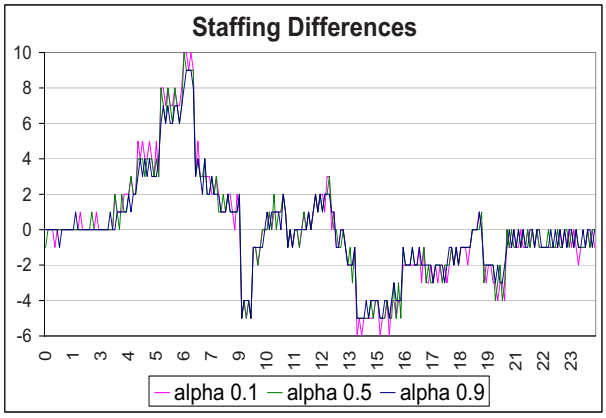
Figure 6.10: Delay Probabilities Obtained by Different Staffing Methods



account 240.4 customers. However, at this moment the servers must handle those customers who arrived in the previous time interval (the expected service time is exactly one time interval in our partition), that is, there are $\lambda(t - 6\ min.) = 199.575$ customers to be served at this moment of time. In a stationary $M/M/25 + M$ model with constant arrival rate 199.575 customers per hour and the rest parameters as defined above, the delay probability is 0.16, which is close to the simulated 0.18 instead of the target 0.5!

During the evening hours, staffing by PSA leads to understaffing. The arrival rate decreases very fast during these hours, so taking into account a current time interval instead of its previous one results in a

Figure 6.11: Staffing Differences between PSA and Lagged PSA Methods



deficit of servers and too high delay probabilities.

We do not present a separate plot for each staffing level because, due to the low resolution, the differences between the curves are barely noticeable. These differences never exceed 10 servers, and for such a large call-center with hundreds of servers (see Fig. 6.3), it will be impossible to recognize them on a single plot. Instead, Figure 6.11 summarizes the differences in the staffing levels obtained by PSA and Lagged PSA. The absolute differences between staffing levels do not depend on the target α and change during the day together with the arrival rate. We observe that fluctuations of the delay probability of the PSA method take place during the time periods where PSA and Lagged PSA determine different staffing levels. The larger these differences, the larger the deviations of PSA from the target delay probability.

6.4.2 Second Empirical Example - A Small Israeli Bank

Here we experiment with the three different staffing methods using the data of a relatively small call-center of a small Israeli bank. In this center, the maximal expected arrival rate does not exceed 120 customers per hour and the call center works only for 16 hours. This call-center was described by Sakov et.al. in [26].

The simulated environment is as follows:

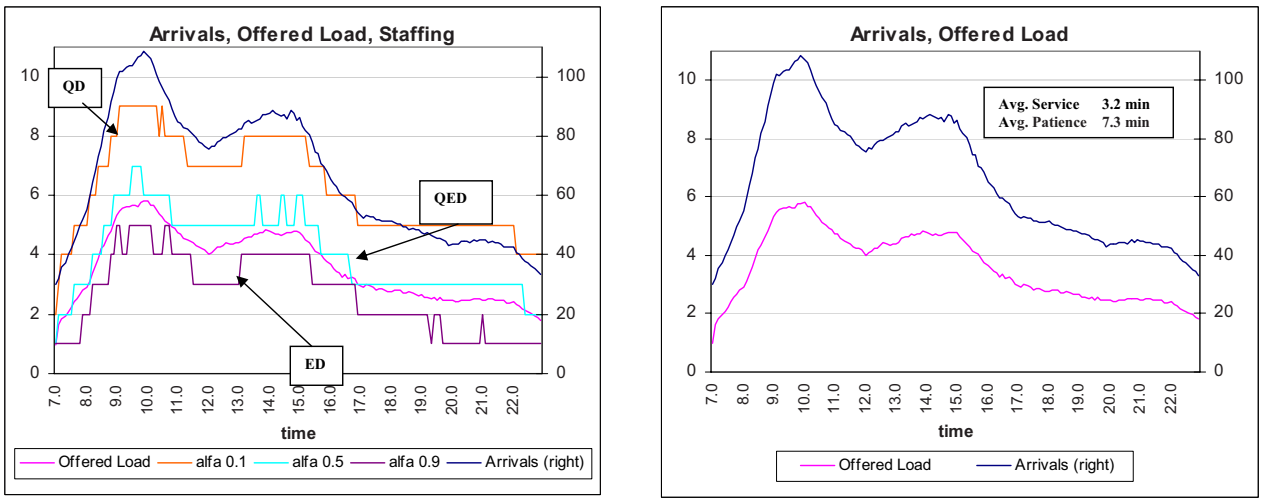
- The running horizon is 16 hours: from 7:00 am till 23:00 pm.
- Performance statistics are calculated over the period of time between $9:00$ a.m. and $23:00$ p.m.
- All empirical values are calculated as an average over 5000 sample paths (simulations).
- Service time is assumed to be distributed exponentially, with mean $1/\mu$ = 3.2 minutes. (This is given in [5])

- Customers patience is assumed to be distributed exponentially, with mean $1/\theta$ = 7.3 minutes.
- The queue discipline is assumed FCFS.
- The arrival rate function is presented in Figure 6.12.

Staffing by ISA

The goal of the experiments is to achieve time-stable performance in the face of time-varying arrivals.

Figure 6.12: Arrivals, Offered Load and Staffing Levels



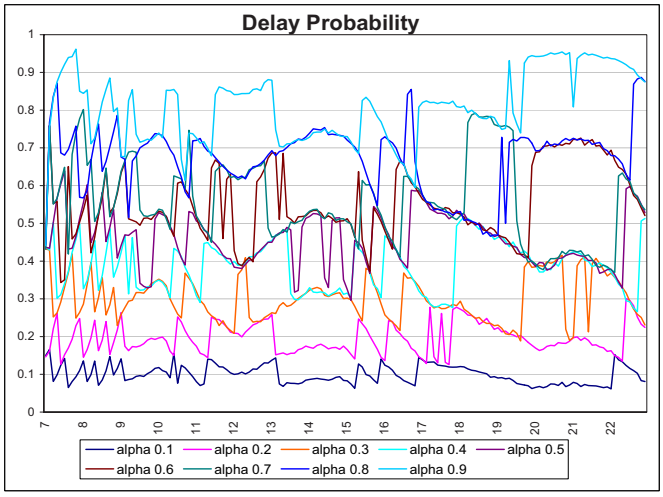
In this example, we present results of staffing by ISA and not by square-root staffing as before. It is important to mention that the differences between these two methods are negligible.

According to the description of ISA (Section 6.1), the algorithm determines, given the target delay probability α, for each time interval (6 minutes) an appropriate staffing level, fixed over the staffing interval. We tested nine values of α, from 0.1 to 0.9 in step 0.1 focusing on α = 0.1, 0.5 and 0.9; the latter 3 values correspond to the QD, QED and ED operational regimes respectively.
The results will now be elaborated on.

Staffing according to ISA achieves relatively stable level of the delay probability even for this small call-center. The fluctuations during the simulation run are the largest for this call-center among all the considered examples (compare Figure 6.3 with Figure 6.13, for example), and this is due to its small size. Summary of the delay probabilities obtained as a result of staffing by ISA is presented in Figure 6.13.

Figure 6.14 shows a comparison of both the simulated global delay probability and probability of waiting more than 30 seconds, if there is waiting, with the theoretical probabilities found by an appropriate stationary model. The theoretical global delay probability for a constant β is found by the Garnett function

Figure 6.13: Delay Probability Summary



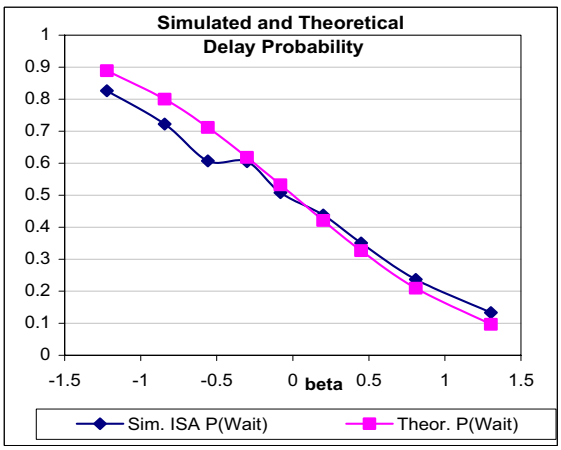
[11]:

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty, \tag{6.10}$$

where $\hat{\beta} = \beta\sqrt{\mu/\theta}$. (Note that in this case $\mu \neq \theta$ so in this case all three parameters β, μ and θ are required.)

All the theoretical calculations are based on the stationary $M/M/s + M$ model with constant arrival

Figure 6.14: Global Performance (10:00-24:00) of (1) Delay Probability; (2) Probability of Waiting More than 30 sec., if there is Waiting



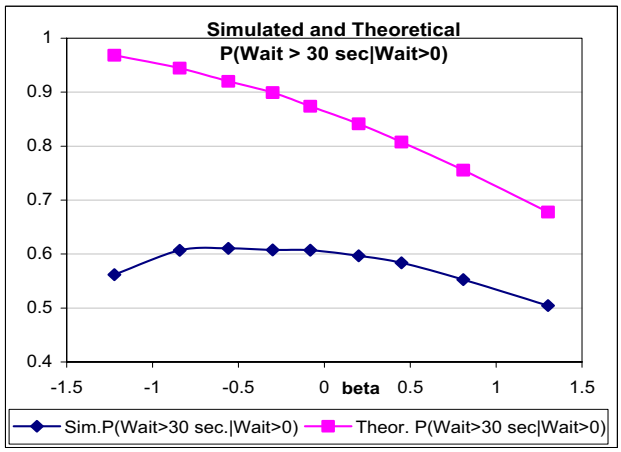
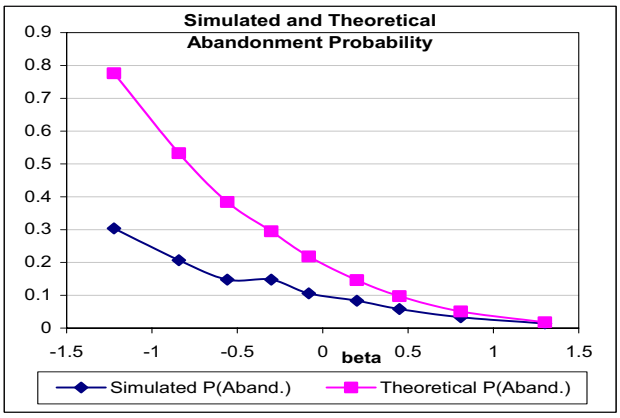
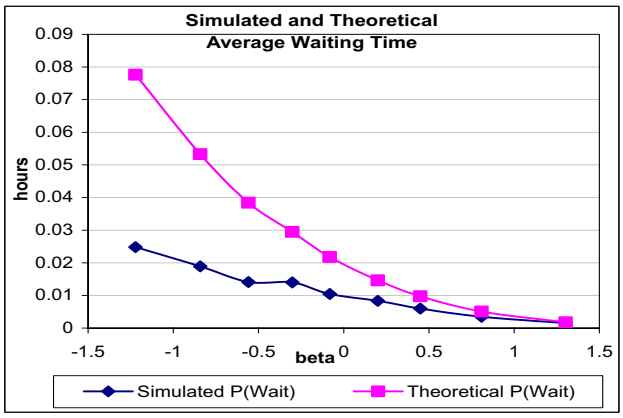


Figure 6.15: Global (10:00-24:00) Performance of (1) Abandonment Probability; (2) Average Waiting Time





rate and number of servers, both calculated in the following way: for each of the 5000 sample paths, we calculated the average arrival rate and the average number of servers over the period of time from 9 a.m. till 23 p.m., and then averaged the 5000 averages.

The theoretical value of the probability of waiting more than 30 seconds, if there is waiting, is based on this stationary model and reads as follows (see [42]):

$$P\left\{W_q > \frac{t}{\sqrt{s}}\middle| W_q > 0\right\} \approx \frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}t)}{\bar{\Phi}(\hat{\beta})}. \tag{6.11}$$

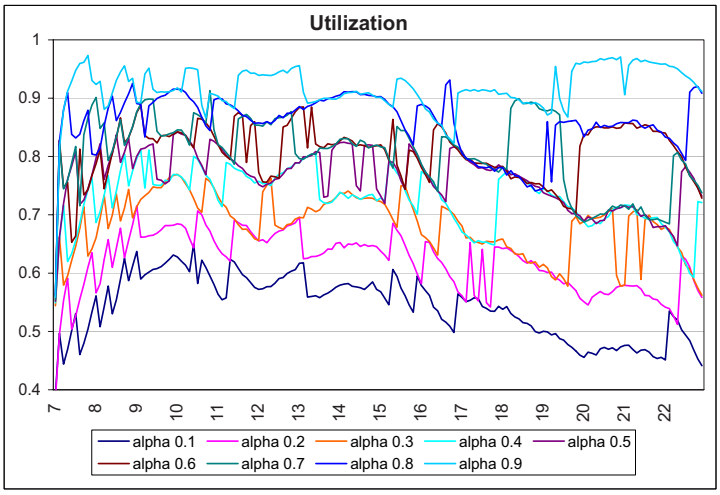
Here
$\Phi(x)$ is the cumulative distribution function of the standard normal distribution (mean=0, std=1),
$\bar{\Phi}(x)$ is the survival function $(\Phi(x) = 1 - \bar{\Phi}(x))$,
$\phi(x)$ is the density of the standard normal distribution,
$h(x) \triangleq \phi(x)/\bar{\Phi}(x)$ is the hazard rate of the standard normal distribution.

Figure 6.14 clearly shows that the approximations for the probability of waiting more than 30 seconds are not applicable for this environment. In addition, when compared to Figure 6.5, the simulated delay probabilities are also relatively different from their theoretical stationary values, especially for the target α greater than 0.6, though the theoretical and the simulated curves of the delay probability seem rather close for $\beta > -0.5$.

Figure 6.15 presents the plots of the abandonment probability and of the expected waiting time, for all nine values of the tested α and compares each simulated measure with its theoretical value. To find the theoretical values we used again the results from [42] for a stationary model:

$$P(Aband) \approx \frac{1}{\sqrt{s}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \tag{6.12}$$

Figure 6.16: Average Servers Utilization



$$E(W_q) \approx \frac{1}{\sqrt{s}} \cdot \frac{1}{\theta} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \quad \left(= \frac{1}{\theta} P(Aband)\right) \tag{6.13}$$

The calculation of these theoretical values is based on the same stationary $M/M/s + M$ model with constant arrival rate and number of servers, averaged as described above.

These stationary approximations do not fit the simulated environment at all. Figure 6.15 shows that the difference between the theoretical and the simulated values is large, and it increases as α grows.

Server utilization is also relatively constant during most of the day (see Figure 6.16). In this example, it takes more time for this measure to stabilize for lower values of the target delay probability α. This situation is different from that with the probability of abandoning, which is more stable for small target delay probabilities (compare with Figure 6.18).

The average values of waiting time and queue length under the three operational regimes are presented in Figure 6.17.

The dynamics of the abandonment probability is presented in Figure 6.18. One can see that for low values of α this probability is low and stable over the whole simulation period.

Histograms of waiting time, if there is waiting, for the three operational regimes are presented in Figure 6.19. The theoretical curves were created via Eq. (6.11) using the average number of servers during the period from 9:00 till 23:00. Note that for such a small average number of servers, this approximation does not work at all, and the theoretical curve does not describe the simulated distribution.

PSA and Lagged PSA

Figure 6.20 presents staffing levels obtained by PSA and Lagged PSA methods. The results of these

Figure 6.17: Average Waiting Time and Queue Length for QD, QED and ED Regimes

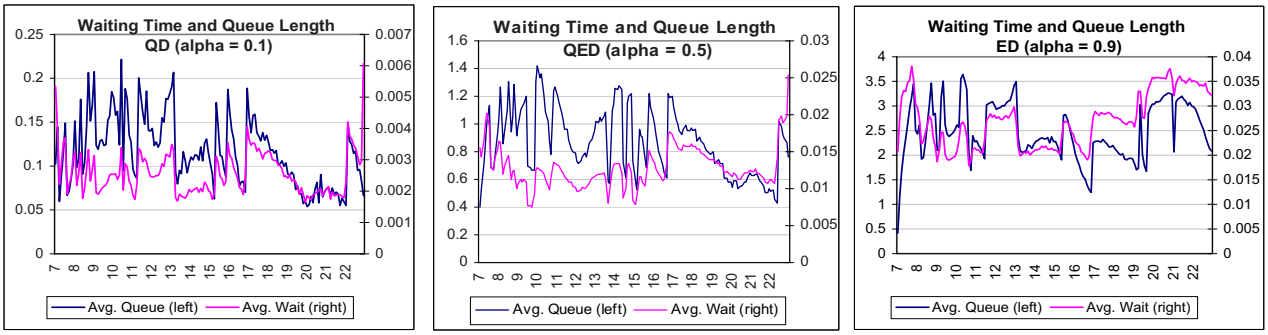
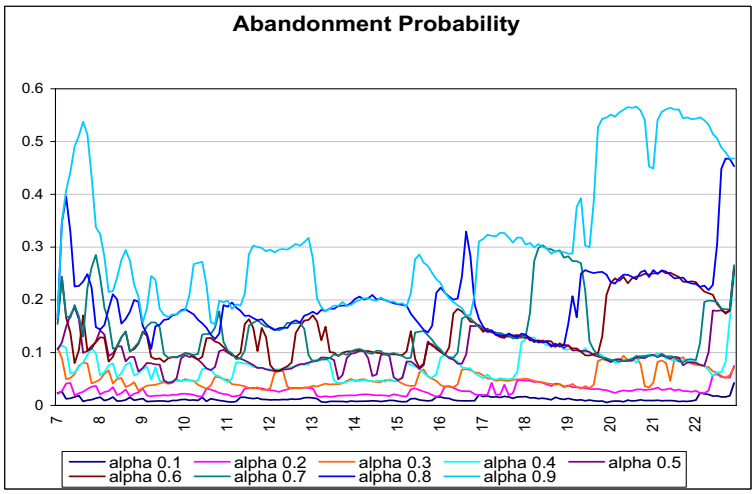


Figure 6.18: Dynamics of Abandonment Probability



two methods are very close. We do not compare ISA with PSA and Lagged PSA staffing levels because in the experiments with ISA the minimal staffing interval was six minutes, while for PSA and Lagged PSA the minimal staffing interval was 3.2 minutes, i.e., one average service time. This was done to facilitate the application of Lagged PSA, which, for any time interval, uses arrival rate that lags exactly one average service time. The difference between PSA and Lagged PSA never exceeds a single server, and the performance of the queue is very similar under both of them. Figure 6.21 presents a summary of the delay and abandonment probabilities. For this measure, these two staffing methods also give a very similar outcome. The performance over the day is not stable, even though the staffing intervals are very short.

6.4.3 Third Empirical Example - An Israeli Cellular Company

Here we analyze a medium call center (maximal arrival rate reaches 500 customers per hour), which also works for 16 hours per day. The simulated environment is as follows:

- The running horizon is 16 hours: from 7:00 am till 23:00 pm.
- Performance statistics are calculated over the period of time between 10:00 a.m. and 23:00 p.m.
- All empirical values are calculated as an average over 5000 sample paths (simulations).
- Service time is assumed to be distributed exponentially, with mean $1/\mu$ = 3.3 minutes (based on data analysis).
- Customers patience is assumed to be distributed exponentially, with mean $1/\theta$ = 0.1 hours or 6 minutes.
- The queue discipline is assumed FCFS.
- The arrival rate function is presented in Figures 6.22 and 6.23.

Square-Root Staffing

The goal of the experiments is to achieve time-stable performance in the face of time-varying arrivals. We are using staffing with a constant value of the Quality-of-Service (QoS) parameter β. This determines an appropriate staffing level, which is fixed over the staffing intervals as follows:

$$s_t = R_t + \beta\sqrt{R_t}. \tag{6.14}$$

Here, $\{R_t,\ 7 \leq t \leq 23\}$ is the time-varying average number of customers (= busy servers) in an $M_t/M/\infty$ queue (5000 sample paths), with the arrival rate as in Figure 6.23 and average service time of

Figure 6.19: Waiting Time, if there is Waiting: Empirical vs. Theoretical Distribution

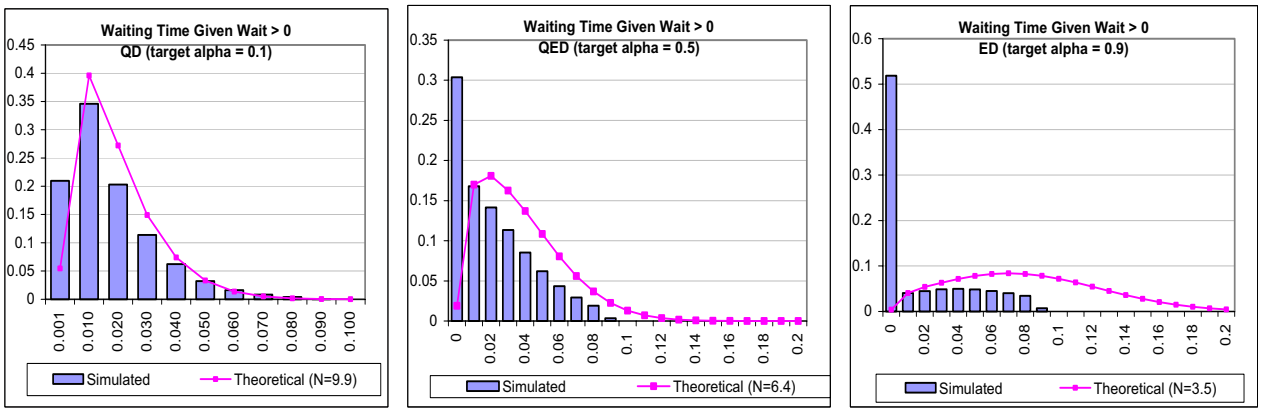


Figure 6.20: Arrivals, Offered Load and Different Staffing Levels

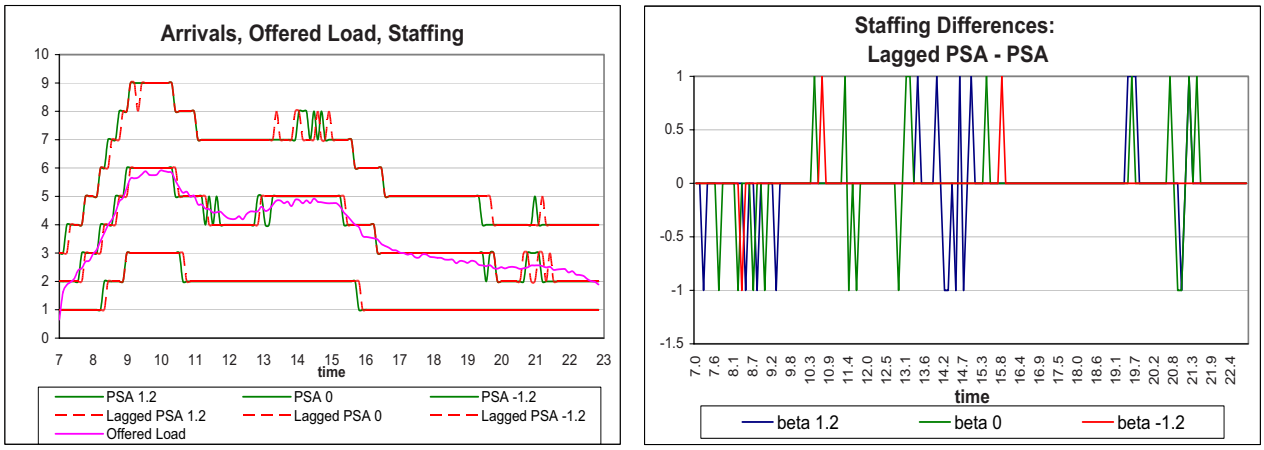
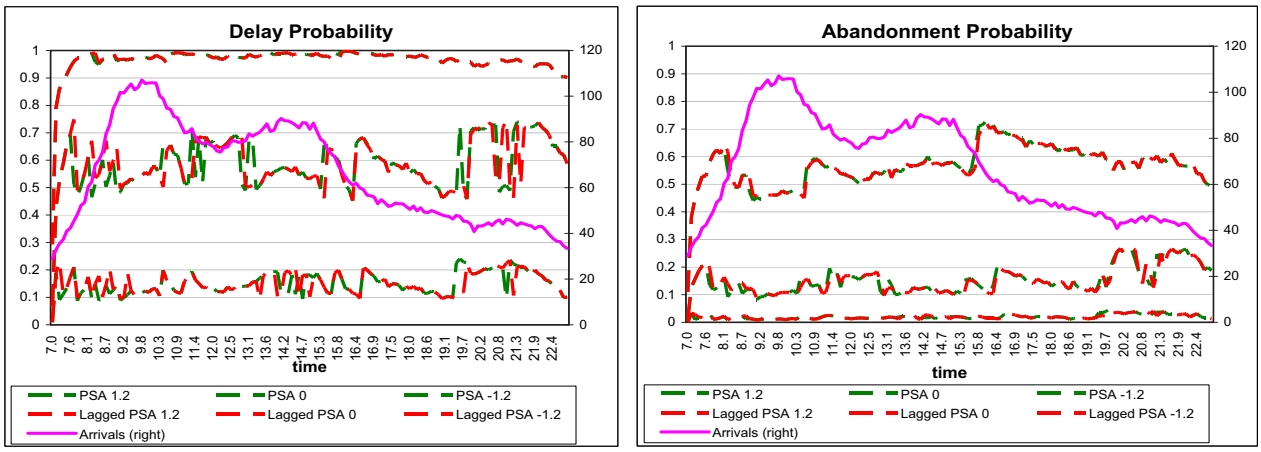


Figure 6.21: Delay and Abandonment Probabilities Obtained by Different Staffing Methods



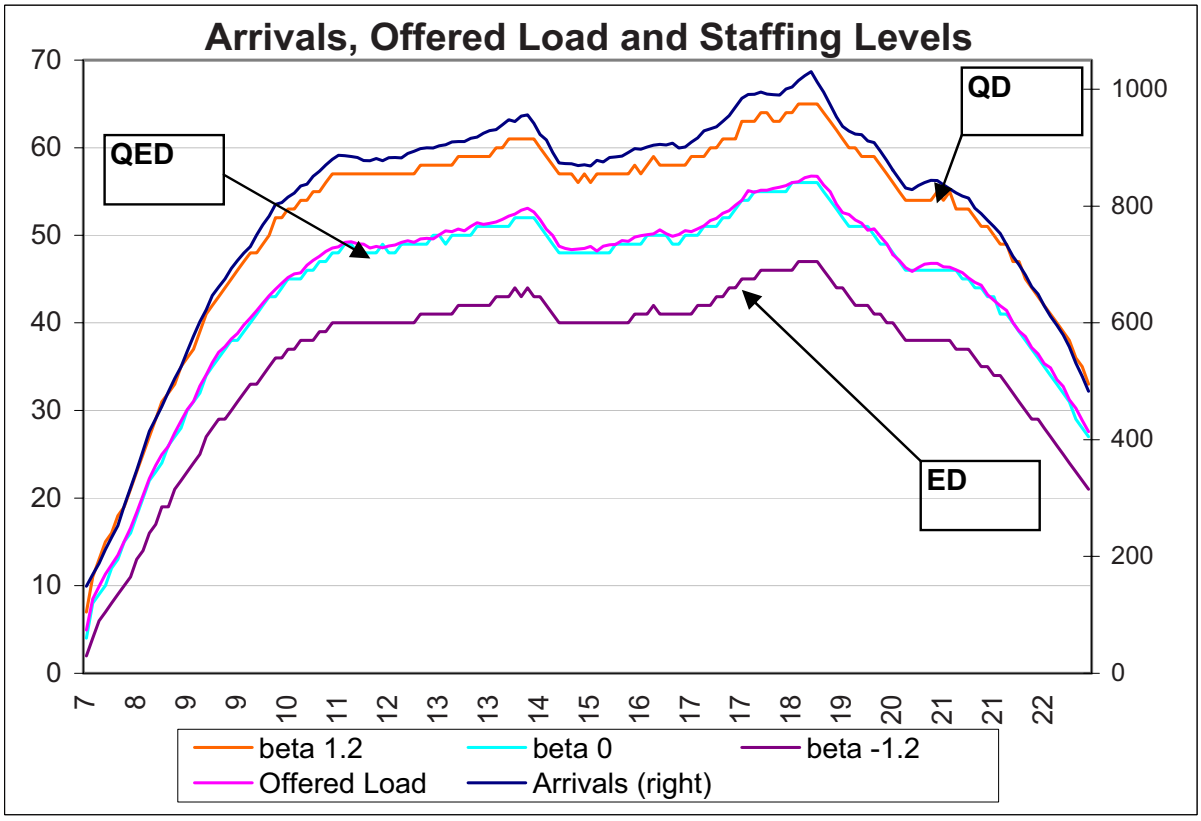
3.3 minutes.

Seven values of β were tested, (from 1.2 to -1.2 with the step -0.4), focusing on $\beta = 1.2$, 0 and -1.2 which correspond to the QD, QED and ED operational regimes respectively.
The results will now be elaborated on.

Staffing according to (6.14), achieves a time-stable level of the delay probability. Summary of the delay probabilities is presented in Figure 6.24.

Figure 6.25 shows a comparison of the simulated overall (global) delay probability and the probability of waiting more than 30 seconds, if there is waiting, with the theoretical probabilities found by an appro-

Figure 6.22: Arrivals, Offered Load and Staffing Levels



priate stationary models. Here, theoretical values are calculated using the average simulated value of β. The theoretical global delay probability for a constant β is found by the Garnett function [11]:

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty, \tag{6.15}$$

where $\hat{\beta} = \beta\sqrt{\mu/\theta}$. (Note that in this case $\mu \neq \theta$ so in this case all three parameters β, μ and θ are required.)

All the theoretical calculations are based on the stationary $M/M/s+M$ model with constant arrival rate and constant number of servers, both calculated in the following way: for each of the 5000 sample paths, we calculated the average arrival rate and the average number of servers over the period of time from 10 a.m. till 23 p.m., and then averaged the 5000 averages.

The theoretical value of the probability of waiting more than 30 seconds, if there is waiting, is based on

Figure 6.23: Arrivals, Offered Load and Staffing Levels

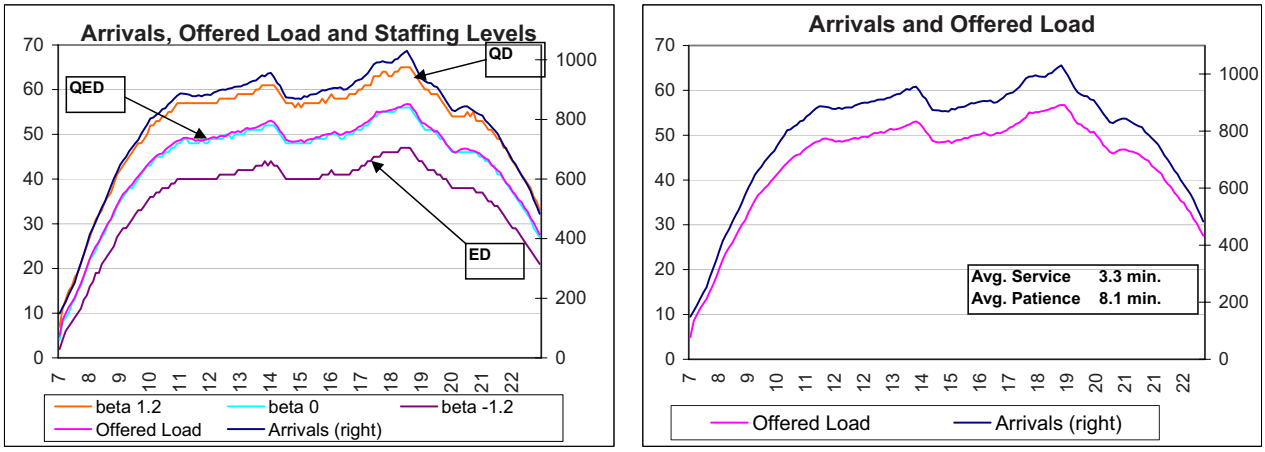
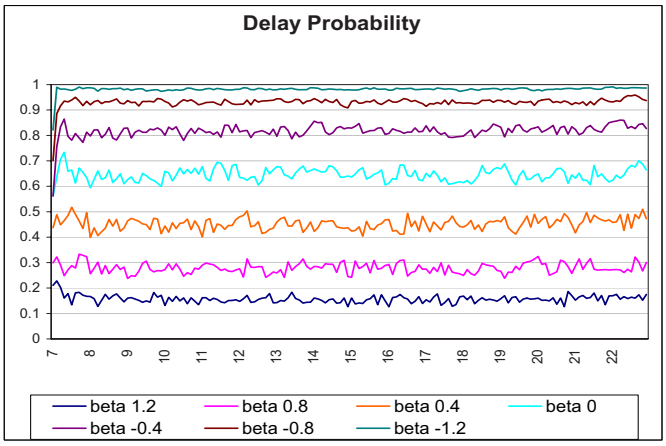


Figure 6.24: Delay Probability Summary



this stationary model and reads as follows (see [42]):

$$P\{W_q > \frac{t}{\sqrt{s}} | W_q > 0\} \approx \frac{\bar{\Phi}(\hat{\beta} + \sqrt{\theta\mu}t)}{\bar{\Phi}(\hat{\beta})} \quad . \tag{6.16}$$

Here

$\Phi(x)$ is the cumulative distribution function of the standard normal distribution (mean=0, std=1),

$\bar{\Phi}(x)$ is the survival function $(\Phi(x) = 1 - \bar{\Phi}(x))$,

$\phi(x)$ is the density of the standard normal distribution,

$h(x) \triangleq \phi(x)/\bar{\Phi}(x)$ is the hazard rate of the standard normal distribution.

We conclude from Figure 6.25 that the delay probability can be approximated by the stationary model

very well, while for the probability of waiting more than 30 seconds, if there is waiting, the difference between theoretical and simulated values grows as the target α increases although these differences are significantly less than in the previous example.

Figure 6.26 presents the plots of the abandonment probability and of the expected waiting time, for all nine values of the tested α, and compares each simulated measure with its theoretical value. To find the

Figure 6.25: Global Performance (10:00-23:00) of (1) Delay Probability; (2) Probability of Waiting More than 30 sec., if there is Waiting

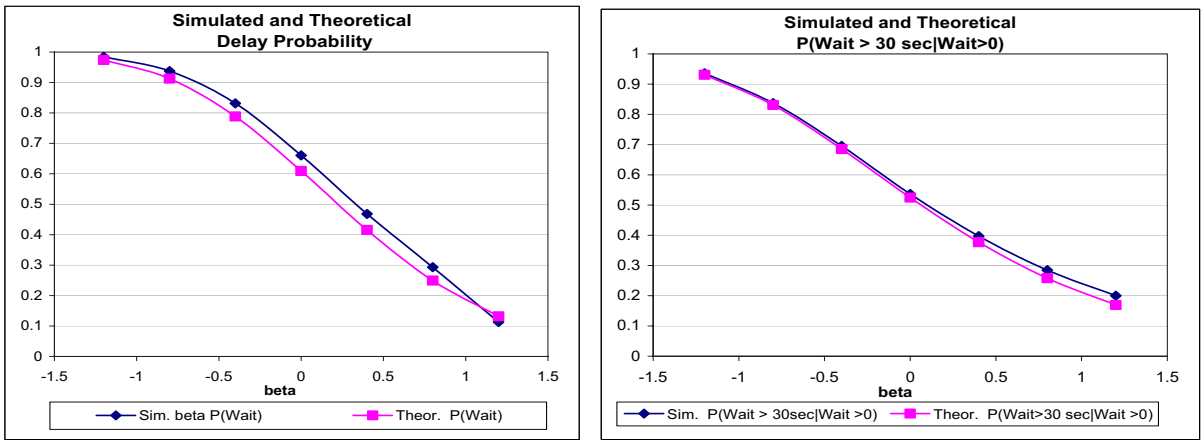


Figure 6.26: Global (10:00-24:00) Performance of (1) Abandonment Probability; (2) Average Waiting Time

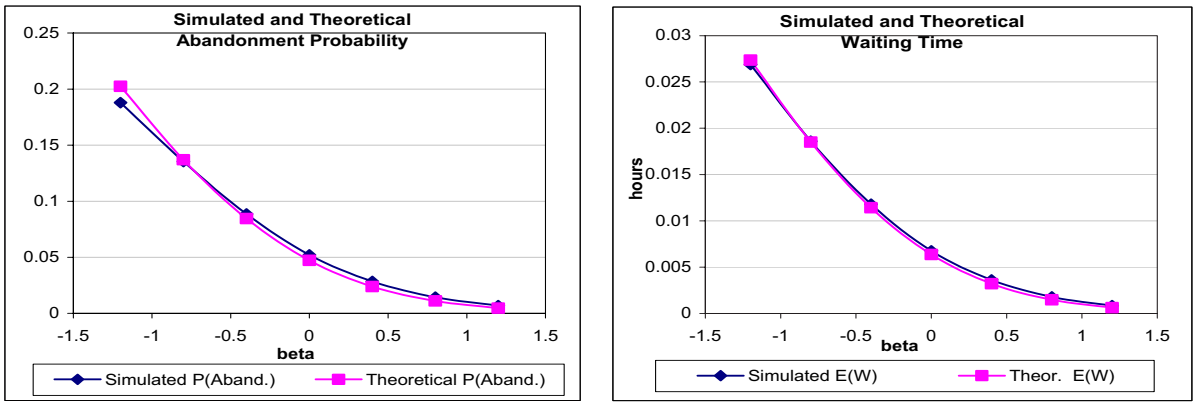
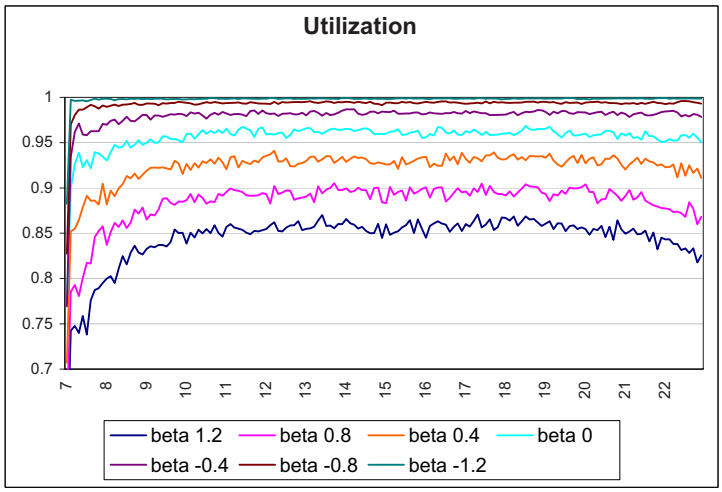


Figure 6.27: Average Servers Utilization



theoretical values we used again the results from [42]:

$$P(Aband) \approx \frac{1}{\sqrt{s}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \tag{6.17}$$

$$E(W_q) \approx \frac{1}{\sqrt{s}} \cdot \frac{1}{\theta} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \quad \left(= \frac{1}{\theta} P(Aband)\right) \tag{6.18}$$

The calculation of these theoretical values is based on the same stationary $M/M/s + M$ model with constant arrival rate and number of servers, averaged as described above.

The approximations are very good when the target delay probability is lower than a half ($0 \leq \beta \leq 1.2$), and are less precise for larger values of the target α.

Server utilization is also relatively constant during most of the day (see Figure 6.27). In this example, it takes more time for this measure to stabilize for lower target delay probability. This situation is different from that with the probability of abandoning, which is more stable for small target delay probabilities (compare with Figure 6.29).

The average values of waiting time and queue length under the three operational regimes are presented in Figure 6.28.

The dynamics of the abandonment probability is presented in Figure 6.29. One sees that for low values of α this probability is low and stable over the whole simulation period.

Histograms of waiting time, if there is waiting, for the three operational regimes are presented in Figure 6.30. The theoretical graphs were created via (6.16) using the average number of servers during the period from 10:00 till 24:00.

Figure 6.28: Average Waiting Time and Queue Length for QD, QED and ED Regimes

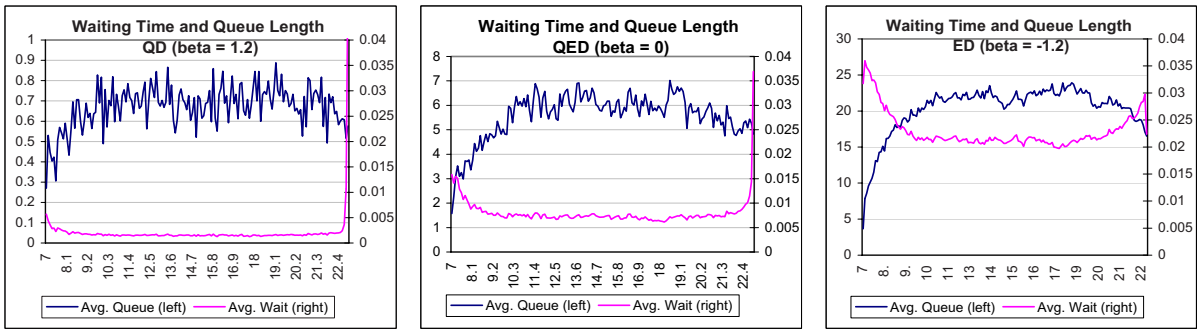
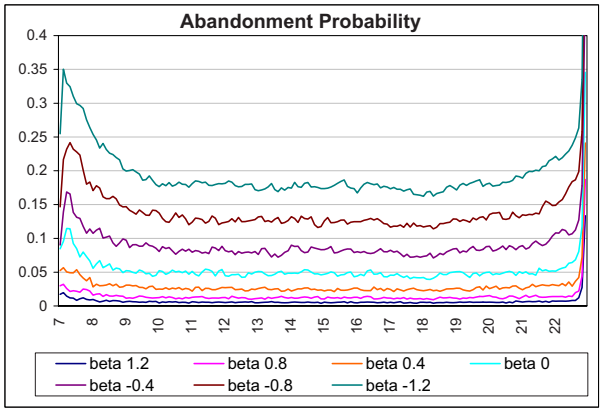


Figure 6.29: Dynamics of Abandonment Probability



PSA and Lagged-PSA

The summary of arrivals, offered load and staffing levels for the three operational regimes (QD,QED and ED), obtained by PSA and Lagged PSA methods, is presented in Figure 6.31. In addition, this figure shows that the staffing differences between PSA and Lagged-PSA methods never exceed a single server.

The delay probability (see Figure 6.32) is more stable than in the previous example, under all staffing methods, especially during the hours of high arrival rate. PSA method does not perform very well during the first three and the last two hours. At the beginning of the day it leads to overstaffing, because it is a period of a very fast growth of the arrival rate. Overstaffing can be identified from the plot of the delay and abandonment probabilities (Figure 6.32). During the last two hours, PSA results in slight understaffing, which follows from the delay probability being higher than the target level.

The abandonment probability obtained by both methods (Figure 6.32) is stable over the main part of the day.

Figure 6.30: Waiting Time, if there is Waiting: Empirical vs. Theoretical Distribution

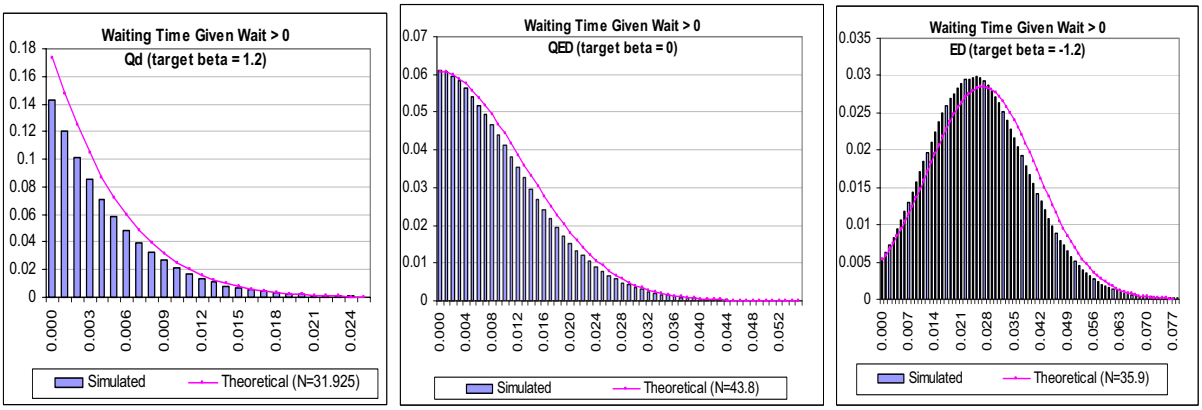
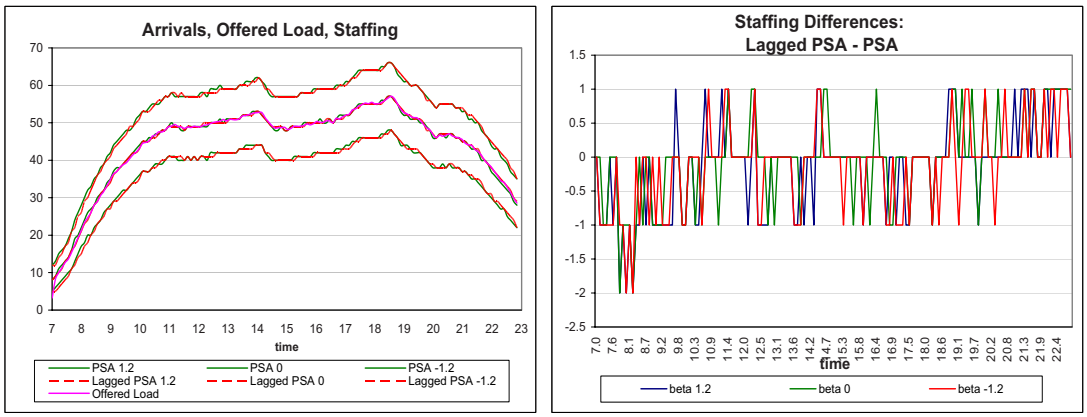


Figure 6.31: Arrivals, Offered Load and Different Staffing Levels



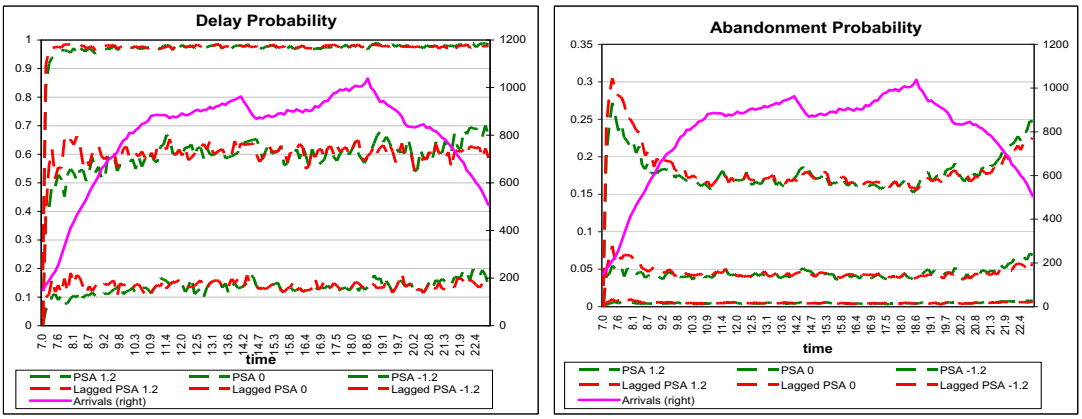
6.4.4 Last Empirical Example - Charlotte Call-Center

Here we analyze the medium call-center which was studied in the course "Service engineering" (096324) [45]. This example is different from the previous ones, because here the customers' patience is to be estimated using the operational ACD report of the call-center. This report is presented in Figure 6.33. In practice, the target of the empirical staffing level was to answer the calls which arrive during the day in average within 30 seconds. This target was achieved. However, the performance during the day was rather unstable. Below we compare the simulated results of staffing according to constant β with the empirical results.

The simulated environment is as follows:

- The running horizon is 10 hours: from 8:00 am till 18:00 pm.

Figure 6.32: Delay and Abandonment Probabilities Obtained by Different Staffing Methods



- All empirical values are calculated as an average over 5000 sample paths (simulations).
- Service time is assumed to be distributed exponentially, with mean $1/\mu$ = 1/12 hours or 5 minutes, as described in [45].)
- Customers patience is assumed to be distributed exponentially. The evaluation of its mean was conducted using the 4CC software ([43]) and is explained later.
- The queue discipline is assumed FCFS.
- The minimal staffing interval in practice is 30 minutes.
- In the simulations, we compared staffing intervals of 5 minutes (a single average service time) and of 30 minutes.
- The arrival rate function, the offered load and the empirical staffing level are presented in Figures 6.34 and 6.35.

The empirical customers' patience during each half-an-hour period was estimated in the following way. The parameters, such as arrival rate, service time and the number of servers, were uploaded to the 4CC software [43] using the option "Advanced Query". Then the empirical abandonment probability of this interval was set to be the goal. On the basis of the uploaded data, 4CC determined the lower and the upper limits of the customers mean patience.

Figure 6.36 presents the estimated abandonment rate during each half an hour which was found as an average of the lower and upper limits. These limits were very close for almost the whole day except of the first interval (8:00-8:30) and the two last ones (17:00-18:00). This is why later we do not use these intervals for the estimation of the average customers patience.

One of the limitations of our simulation software is the underlying assumption that the abandonment rate

Figure 6.33: Example of ACD Report

Asymptotic Operational Regimes

Example of Half-Hour ACD Report

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
Total	20,577	19,860	3.5%	30	307	95.1%	
8:00	332	308	7.2%	27	302	87.1%	59.3
8:30	653	615	5.8%	58	293	96.1%	104.1
9:00	866	796	8.1%	63	308	97.1%	140.4
9:30	1,152	1,138	1.2%	28	303	90.8%	211.1
10:00	1,330	1,286	3.3%	22	307	98.4%	223.1
10:30	1,364	1,338	1.9%	33	296	99.0%	222.5
11:00	1,380	1,280	7.2%	34	306	98.2%	222.0
11:30	1,272	1,247	2.0%	44	298	94.6%	218.0
12:00	1,179	1,177	0.2%	1	306	91.6%	218.3
12:30	1,174	1,160	1.2%	10	302	95.5%	203.8
13:00	1,018	999	1.9%	9	314	95.4%	182.9
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**
14:00	1,173	1,082	7.8%	78	313	99.5%	188.9
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**
15:00	1,137	1,122	1.3%	15	320	96.9%	205.8
15:30	1,169	1,137	2.7%	17	311	97.1%	202.2
16:00	1,107	1,059	4.3%	46	315	99.2%	187.1
16:30	914	892	2.4%	22	307	95.2%	160.0
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**
17:30	420	420	0.0%	0	328	73.8%	103.5
18:00	49	49	0.0%	14	180	84.2%	5.8

is constant during the whole day. As follows from Figure 6.36, this assumption does not hold in the real life. In our further experiments we set the abandonment rate θ to be 10.2 customers per hour, which was found as the average of the estimated abandonment rates weighted by the number of arrivals in each half-hour interval.

In addition, Figure 6.36 presents the empirical quality of service parameter during the whole day. We see that it is very unstable; so that during the same day some customers are served under the ED regime ($\beta < -1$), while others - under the QD regime ($\beta > 2$).

Square-Root Staffing

The goal of our experiments is to achieve time-stable performance in the face of time-varying arrivals.

Figure 6.34: Arrivals, Offered Load and Staffing Levels

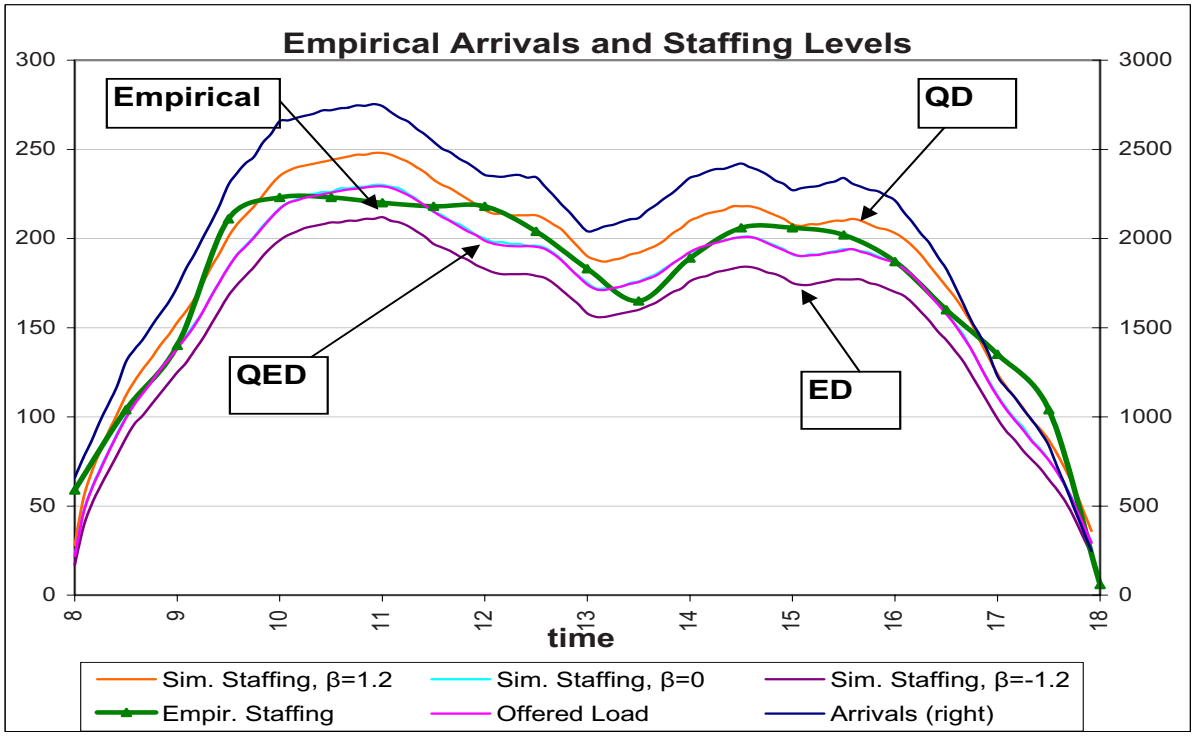
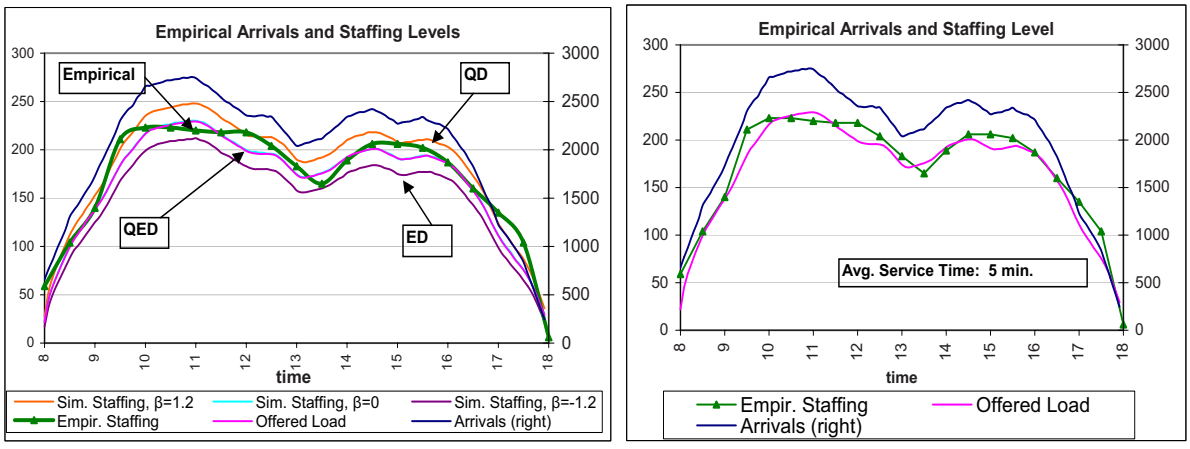


Figure 6.35: Arrivals, Offered Load and Staffing Levels



For both 5-minute and 30-minute staffing intervals, we are using staffing with a constant value of the

Figure 6.36: Estimated Quality of Service and Customer Patience

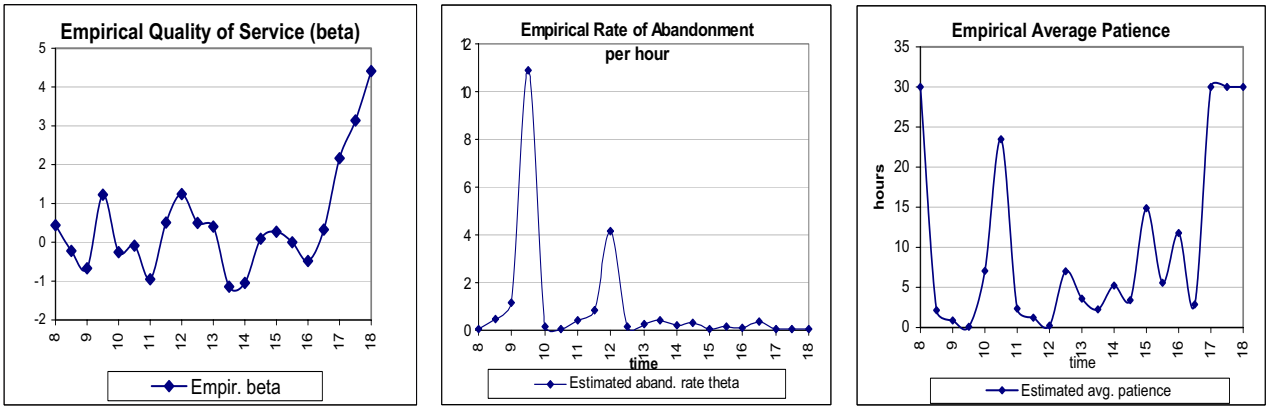
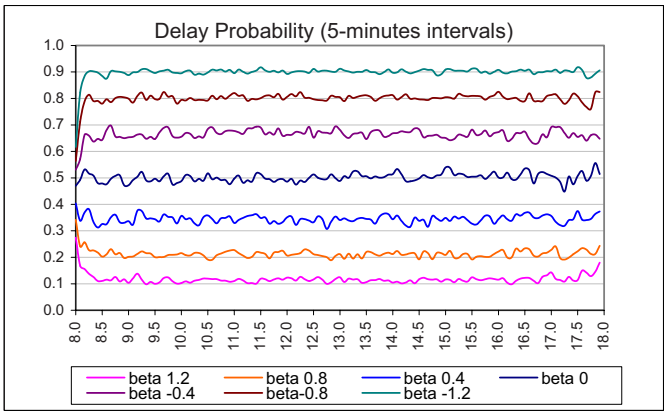


Figure 6.37: Delay Probability Summary



Quality-of-Service (QoS) parameter β. This determines an appropriate staffing level, which is fixed over the staffing interval for each time interval as follows:

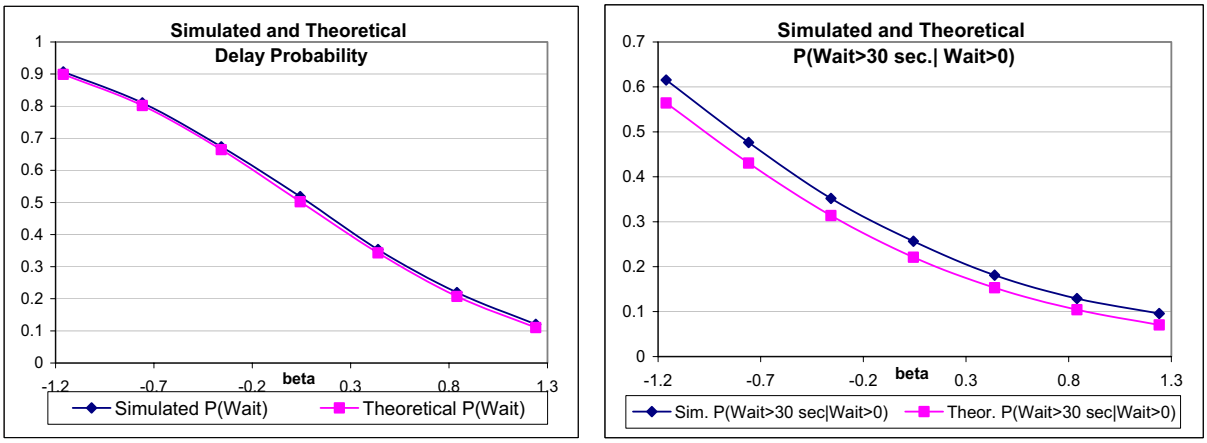
$$s_t = R_t + \beta\sqrt{R_t}. \tag{6.19}$$

Here, $\{R_t,\ 0 \leq t \leq 24\}$ is the time-varying average number of customers (= busy servers) in an $M_t/M/\infty$ queue (5000 sample paths), with the arrival rate as in Figure 6.35 and average service time of 6 minutes.

In both cases, seven values of β were tested, (from 1.2 to -1.2 with the step -0.4), focusing on $\beta = 1.2$, 0 and -1.2, which correspond to the QD, QED and ED operational regimes respectively.

5-minute staffing intervals

Figure 6.38: Global Performance (10:00-16:00) of (1) Delay Probability; (2) Probability to Wait More than 30 sec., if there is Waiting



Staffing according to (6.19) achieves a time-stable level of the delay probability. Summary of the delay probabilities is presented in Figure 6.37.

Figure 6.38 shows a comparison of the simulated overall (global) delay probability and the probability of waiting more than 30 seconds, if there is waiting, with the theoretical probabilities found by an appropriate stationary models. The theoretical global delay probability for a constant β is found by the Garnett function [11]:

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty, \tag{6.20}$$

where $\hat{\beta} = \beta\sqrt{\mu/\theta}$. (Note that in this case $\mu \neq \theta$ so in this case all three parameters β, μ and θ are required.)

All the theoretical calculations are based on the stationary $M/M/s+M$ model with constant arrival rate and constant number of servers, both calculated in the following way: for each of the 5000 sample paths, we calculated the average arrival rate and the average number of servers over the period of time from 10 a.m. till 16 p.m., and then averaged the 5000 averages.

The theoretical value of the probability of waiting more than 30 seconds, if there is waiting, is based on this stationary model and reads as follows (see [42]):

$$P\left\{W_q > \frac{t}{\sqrt{s}} \middle| W_q > 0\right\} \approx \frac{\bar{\Phi}(\hat{\beta} + \sqrt{\theta\mu}t)}{\bar{\Phi}(\hat{\beta})} \quad . \tag{6.21}$$

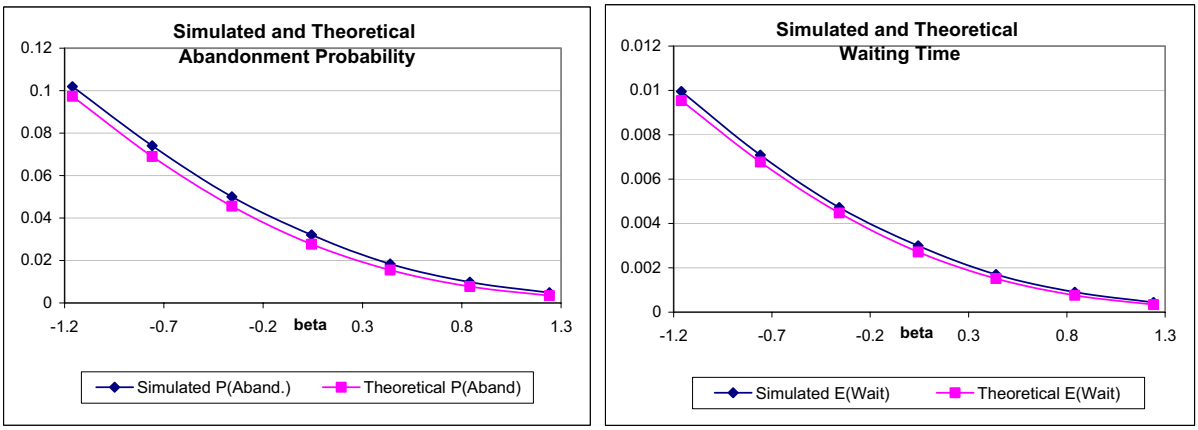
Here

$\Phi(x)$ is the cumulative distribution function of the standard normal distribution (mean=0, std=1),

$\bar{\Phi}(x)$ is the survival function ($\Phi(x) = 1 - \bar{\Phi}(x)$),
$\phi(x)$ is the density of the standard normal distribution,
$h(x) \triangleq \phi(x)/\bar{\Phi}(x)$ is the hazard rate of the standard normal distribution.

Figure 6.39 presents the plots of the abandonment probability and of the expected waiting time, for all

Figure 6.39: Global (10:00-16:00) Performance of (1) Abandonment Probability; (2) Average Waiting Time



seven values of the tested β and compares each simulated measure with its theoretical value. To find the theoretical values we used again the results from [42] for a stationary model:

$$P(Aband) \approx \frac{1}{\sqrt{s}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \tag{6.22}$$

$$E(W_q) \approx \frac{1}{\sqrt{s}} \cdot \frac{1}{\theta} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{\theta}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} \quad \left(= \frac{1}{\theta} P(Aband)\right) \tag{6.23}$$

The calculation of these theoretical values is based on the same stationary $M/M/s + M$ model with the constant arrival rate and number of servers, averaged as described above.

Server utilization is also relatively constant during most of the day (see Figure 6.40). In this example, it takes more time for this measure to stabilize for higher values of β (lower target delay probability). This situation is different from that with the probability of abandoning, which is more stable for small target delay probabilities (compare with Figure 6.42).

Average values of waiting time and queue length under the three operational regimes are presented in Figure 6.41.

The dynamics of the abandonment probability is presented in Figure 6.42. One can see that for low values of α this probability is low and stable over the whole simulation period.

Histograms of the waiting time, if there is waiting, for the three operational regimes are presented in

Figure 6.40: Average Servers Utilization

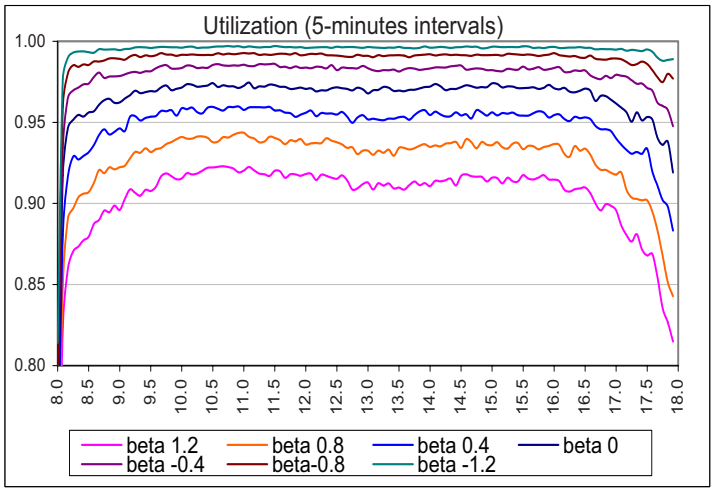


Figure 6.41: Average Waiting Time and Queue Length for QD, QED and ED Regimes

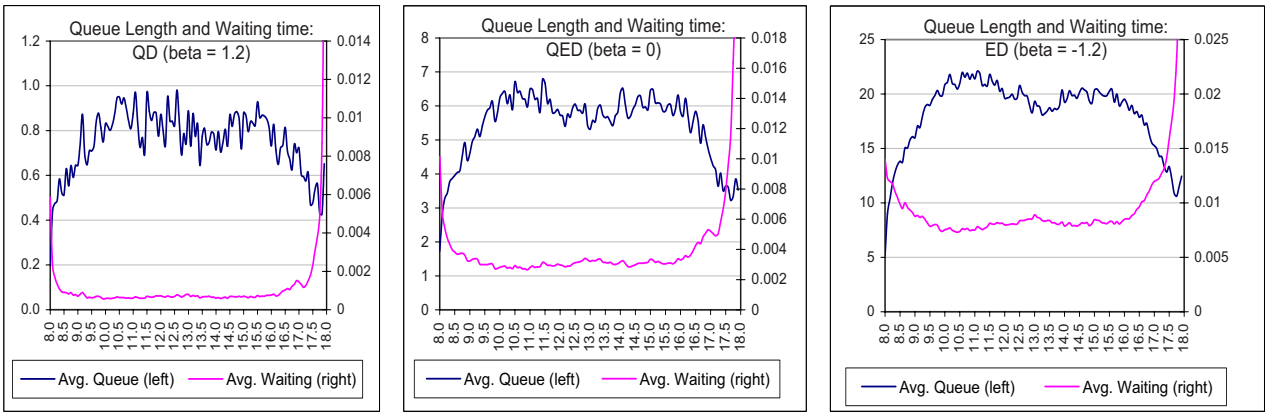


Figure 6.43. The theoretical graphs were created via (6.21) using the average number of servers during the period from 10:00 till 16:00, and the accuracy is excellent.

30-minute staffing intervals

In practice, such short staffing intervals may be an impractical solution, since in reality people often lack flexibility and staffing levels cannot be changed every five minutes. Below are the results of our simulations for staffing using a constant quality-of-service parameter β on 30-minute-long staffing intervals. Our results exhibit performance that is significantly better than that prevailing in industry.

The recommended staffing levels for the three considered operational regimes are presented in Figure

Figure 6.42: Dynamics of Abandonment Probability

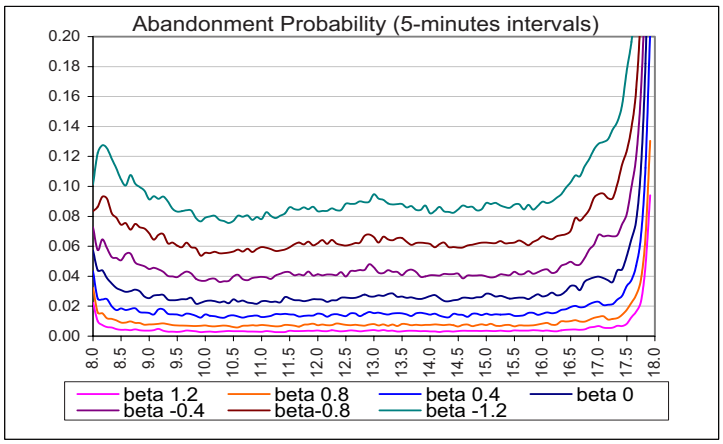
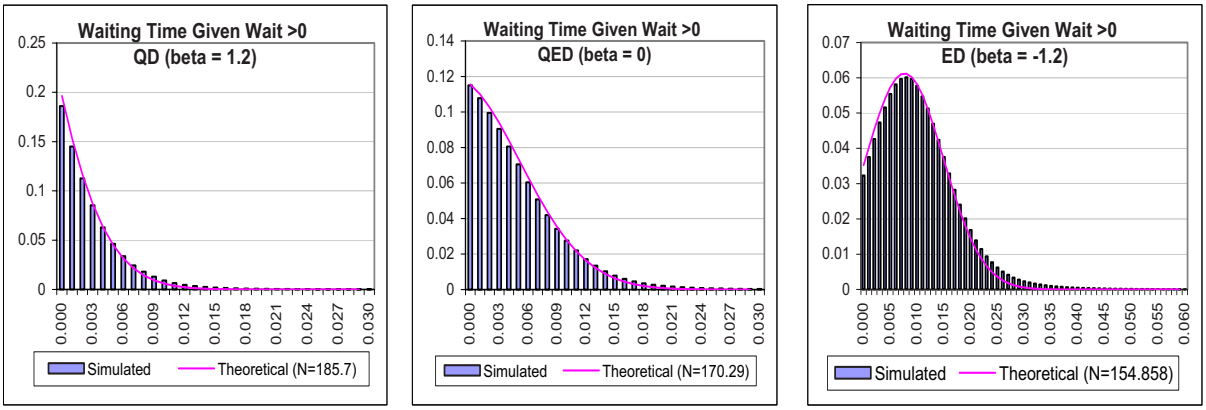


Figure 6.43: Waiting Time, if there is Waiting: Empirical vs. Theoretical Distribution



6.44.

The empirical staffing level, which differs from the simulated one, leads to a very unstable performance during the day. Figure 6.45 presents a comparison between the simulated β under three different operational regimes with the empirical quality-of-service parameter. This figure shows that though the staffing intervals are relatively large, the simulated β is very stable.

Delay and abandonment probabilities are, of course, less stable than in case with the 5-minute staffing intervals, but they are still very predictable. Figure 6.46 presents a summary of these probabilities for all seven values of β.

The abandonment probability during the whole day under any operational regime is much more stable than the empirical one. The comparison between the simulated and the empirical probabilities of aban-

Figure 6.44: Arrivals, Offered Load and Staffing Levels

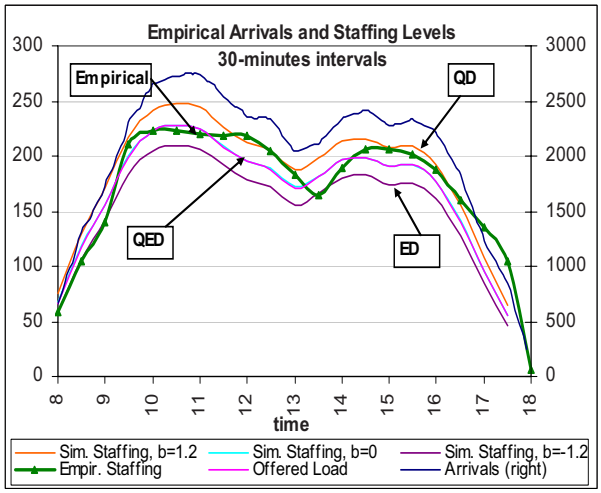
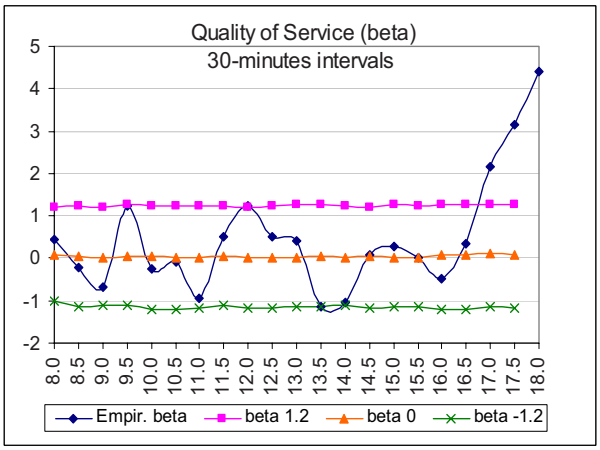


Figure 6.45: Empirical and Simulated Quality of Service



doning are shown in Figure 6.47.

The overall performance of Charlotte call-center is also improvable. Table 6.1 compares some empirical performance measures with those obtained by our simulations. As follows from this table, the overall performance of the call-center could be improved. The objective of answering all the calls within an average of 30 seconds could be achieved using a smaller number of servers. In fact, the present number of servers could have been sufficient to decrease the average waiting time from 30 to 10 seconds.

The utilization of the servers could also be increased: under the ED regime it exceeds 99 % and the average waiting time is not significantly different from the target 30 seconds.

Figure 6.46: Summary of Delay and Abandonment Probabilities

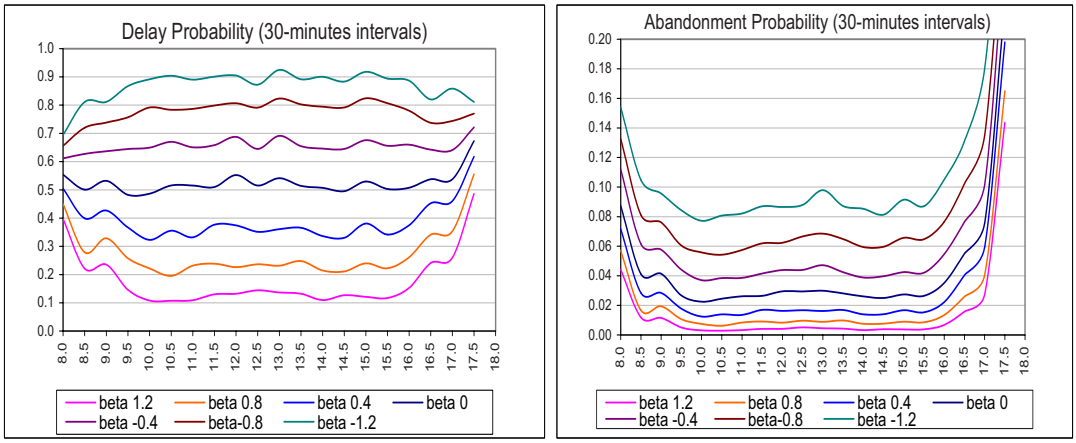
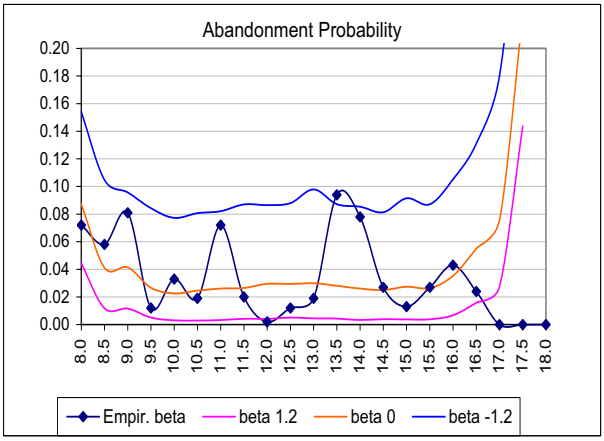


Figure 6.47: Empirical and Simulated Abandonment Probability



PSA and Lagged PSA

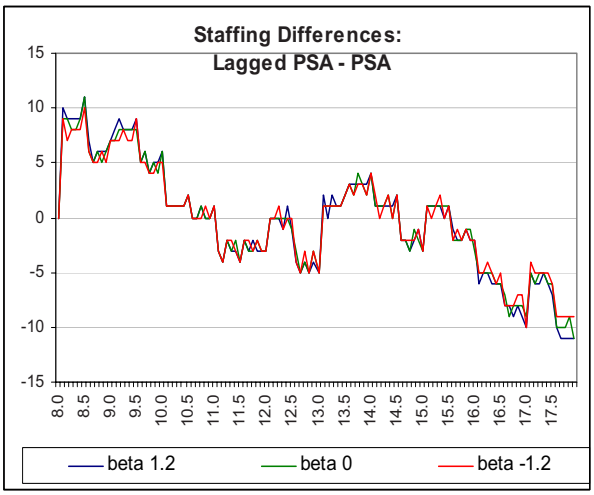
Table 6.1: Charlotte - Comparison of Empirical and Simulated Performance

	Agent hours	P(Aband)	ASA (sec)	Utilization
Empirical	1781.5	3.5 %	30	95.1 %
$\beta = 1.2$ (QD)	1857.4	0.49 %	1.58	90.33 %
$\beta = 0$ (QED)	1702.92	3.2 %	10.78	96.5 %
$\beta - 1.2$ (ED)	1548.58	10.19 %	35.82	99.4 %

In the Charlotte call-center, staffing differences between PSA and Lagged PSA cannot be ignored: for the time intervals in which the arrival rates change fast, the differences in the number of servers can amount up to 12 servers. Figure 6.48 presents staffing differences between Lagged PSA and PSA staffing over the day, under the QD, QED and ED regimes. We do not present a separate plot for each staffing level because, due to the low resolution, the differences between the curves are barely noticeable.

It appears that the time lag has a strong impact on the overall performance of the system. Figure 6.49

Figure 6.48: Staffing Differences between the Lagged PSA and PSA methods



shows that the use of PSA gives unsuitable delay probabilities: too high for the first part of the day with an increasing arrival rate, and too low for the second part of the day with the decreasing arrival rate. In contrast, both Lagged PSA and square-root staffing give very close and stable delay probabilities.

In addition, the abandonment probability (see Figure 6.50) is subject to the changes in the staffing levels: the probability of abandonment is very stable both under the square-root staffing and Lagged PSA, whereas staffing by PSA worsens the abandonment probability which becomes very unstable during the day.

6.4.5 Conclusions

Systems with staffing levels determined by ISA or according to square-root staffing with constant β performed very similarly for all our examples. Hence we describe the performance of only one of these methods.

The examples above differ in the levels of the offered load during the work day. In our second example, the range of offered load is within 1 to 6 hours per hour (Erlangs), while in the first and in the last examples the offered load falls within the range of approximately 50 to more than 200 Erlangs. Thus, we

Figure 6.49: Delay Probability obtained by different staffing methods

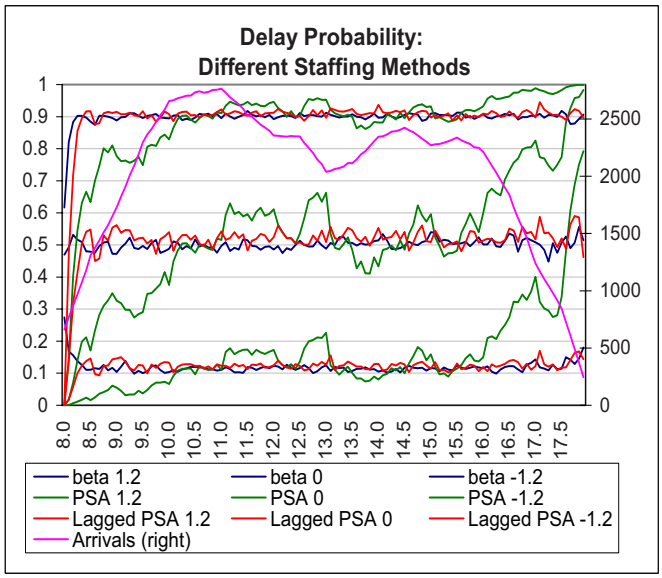
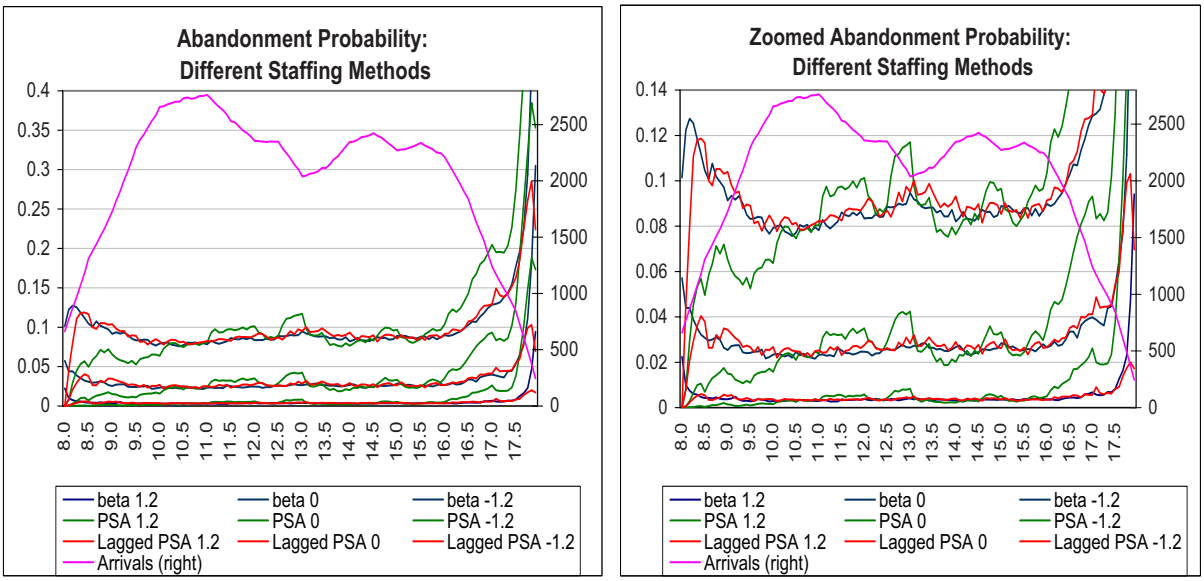


Figure 6.50: Abandonment Probability obtained by different staffing methods



can conclude that most of the QED approximations are inaccurate for small system size, since none of the approximations, except for the Garnett formula (6.6), fits in our second empirical example (a small Israeli bank).

The Garnett Function (6.6): First of all, we wish to emphasize that the Garnett approximation for the delay probability is very robust. As follows from our second example, this approximation can be applied even to a very small call-center, though it is rather surprising that the asymptotic results are so close to the simulated values even when the number of severs is small (5-10 servers).

Delay Probability: Square-root safety staffing enabled a very stable delay probability during the simulation run in all examples, except for the second one, where the deviations of the delay probability were large during the whole simulation run due to the very small size of this system. In the first example, there was a warm-up period with relatively large deviations during the first hours of the run due to the low offered load in this period.

Waiting-Time Histograms: The empirical waiting-time histograms of the delayed customers almost coincide with the theoretical curves in all cases, except for the second example. In the third example, in which the offered load does not exceed 60 servers per hour, the theoretical curve (Figure 6.25) does not precisely describe the empirical distribution of waiting time if there is waiting under the the QD regime, though starting from some point, the two are rather close to each other. QED and ED empirical histograms in this example are very close to their theoretical curves.

In the first and the last examples (Figures 6.9 and 6.30), where the offered load during the day approaches 200 servers per hour, all the empirical histograms can be reliably predicted by their appropriate theoretical curves.

In the second example, the empirical histograms (Figure 6.19) do not resemble the theoretical curves at all, the reason being a very small size of this system.

Different Staffing Methods In our examples, square-root safety staffing gave rise to the most stable performance . From the description of Examples 1 and 4, it follows that Lagged PSA is rather close to square-root staffing, though it is a slightly less stable. Lagged PSA performs better than PSA in all cases except for the very small call-center, where they performed in a very similar mode. In general, staffing by PSA leads to overstaffing during the first part of the day with the increasing offered load, and it results in under-staffing in the second part of a day with the decreasing offered load.

In Examples 2 and 3, ISA or β staffing was not compared with PSA and Lagged-PSA due to different length of the staffing intervals between the simulations with ISA and PSA and Lagged-PSA. The different staffing shifts were simulated due to short average service time (about three minutes). To apply Lagged-PSA, we needed to set the minimal staffing interval to be 3 minutes, while for square-root safety staffing, the staffing interval was six minutes, because shorter intervals are of no practical value. Although in these examples, it was not possible to compare the performance of the three methods, we assume that the performance of Lagged-PSA would be very similar to that of ISA or β staffing with the 3-minutes staffing intervals.

6.5 Appendix

This Appendix presents the explanation of our present implementation of the performance measures listed in Section 6.2, alternative ways of these calculations and a comparison of the implemented and the alternative methods.

6.5.1 Performance Measures Implemented in the ISA Algorithm

Recall that in the formulae below:
superscript *j* indicates the j^{th} replication;
subscript *t* indicates a t^{th} partition interval of size ΔT;
total number of replications is *Reps*.

Dynamic Performance Measures

Offered Load R_t

$$R_t = \frac{\sum_{j=1}^{Reps} L_t^j}{Reps},$$

where
L_t^j is the total number of customers at the end of interval *t* in replication *j*; recall that *Reps* is the number of replications.

Implied Service Quality β_t

$$\beta_t = \frac{s_t - R_t}{\sqrt{R_t}},$$

where
s_t is the number of servers during the interval *t* (which is held fixed by ISA over an interval).

Servers Utilization ρ_t

$$\rho_t = \frac{\sum_{j=1}^{Reps} \rho_t^j}{Reps};$$

here
ρ_t^j is the servers utilization during interval *t* in replication *j*:

$$\rho_t^j = \frac{Busy_t^j}{\Delta T \cdot s_t},$$

where
$Busy_t^j$ is the total time the servers were working during the interval t in the j^{th} replication.

Abandonment Probability $P_t(band)$

$$P_t(Aband) = \frac{\sum_{j=1}^{Reps} P_t^j(Aband)}{Reps};$$

here

$$P_t^j(Aband) = \frac{Abandoned_t^j}{Arrived_t^j},$$

where $Arrived_t^j$ is the number of customers who arrived at time interval *t* in replication *j*, and $Abandoned_t^j$ is the number of customers who arrived at time interval *t* in replication *j* and eventually abandoned.

Delay Probability $P_t(W_q > 0)$

$$P_t(W_q > 0) = \frac{\sum_{j=1}^{Reps} P_t^j(W_q > 0)}{Reps};$$

here

$$P_t^j(W_q > 0) = \frac{Delayed_t^j}{Arrived_t^j},$$

where $Delayed_t^j$ is the number customers who arrived at time interval *t* in replication *j* and did not start their service immediately.

Average Queue Length Q_t

$$Q_t = \frac{\sum_{j=1}^{Reps} Q_t^j}{Reps},$$

where Q_t^j is a queue length in interval *t* in j^{th} replication.

Average Waiting Time W_t

$$W_t = \frac{\sum_{j=1}^{Reps} W_t^j}{Reps},$$

where W_t^j is an average waiting time of customers who **arrived** in interval *t* in j^{th} replication:

$$W_t^j = \frac{(Waiting\,time)_t^j}{Arrived_t^j}.$$

Average Waiting Time, if there is Waiting $E(W_t|W_t > 0)$

$$E(W_t|W_t > 0) = \frac{\sum_{j=1}^{Reps} (Waiting\,time)_t^j}{\sum_{j=1}^{Reps} Delayed_t^j},$$

where $(Waiting\,time)_t^j$ is the total waiting time of customers who arrived at time *t* in the j^{th} replication.

Overall Performance Measures

Average Waiting Time W

$$W = \frac{\sum_{j=1}^{Reps} W^j}{Reps};$$

here

$$W^j = \frac{(Waiting\,Time)^j}{Released^j},$$

where $Released^j$ is the total number of customers who left the queue during the j^{th} replication.

Average Waiting Time, if there is on Waiting $E(W|W > 0)$

$$E(W|W > 0) = \frac{\sum_{j=1}^{Reps} E(W^j|W^j > 0)}{Reps},$$

where

$$E(W^j|W^j > 0) = \frac{Total\,Waiting\,Time^j}{Total\ Delayed^j}.$$

Average Abandon Probability $P(Aband)$

$$P(Aband) = \frac{\sum_{j=1}^{Reps} P^j(Aband)}{Reps},$$

where

$$P^j(Aband) = \frac{Abandoned^j}{Released^j}.$$

Average Delay Probability $P(W_q > 0)$

$$P(W_q > 0) = \frac{\sum_{j=1}^{Reps} P^j(W_q > 0)}{Reps},$$

where

$$P^j(W_q > 0) = \frac{Delayed^j}{Released^j}.$$

Servers Utilization ρ

$$\rho = \frac{\sum_{j=1}^{Reps} \rho^j}{Reps};$$

here ρ^j is the total servers utilization during j^{th} replication.

$$\rho^j = \frac{Busy^j}{\Delta T \sum_{t=1}^{T} s_t},$$

where $Busy^j$ is the total time the servers were busy during j^{th} iteration.

6.5.2 Alternative Ways of Calculations

This section presents **alternative ways** for calculating some performance measures in the simulation software. First, we present an approach, which is based on a different way of averaging. It yields typically approximately the same results as the method implemented in the code, but the outcomes could vary in case of large deviations of the estimated measure (i.e., Delay probability) over time. Later, we discuss different definitions of abandonment probability

Single-Batch Approach

In the ISA implementation, almost all the performance measures are calculated as an average of many ($Reps$) averages. The only exception is Expected Waiting Time, if there is Waiting for the whole simulation horizon $E(W|W > 0)$. For all others, first, we calculated the needed measure for a single replication (batch), and the final result was calculated as an average of all the batches.

The difference in the calculation of $E(W|W > 0)$ is that in its calculation we use a single very long batch of length $Reps \times T$. Other performance measures, for example, Delay Probability, can be also calculated by this approach.

Both implemented and the single-batch methods have their advantages and disadvantages. Finding the average of multiple batches smoothes and neutralizes extreme deviations which take place in some single batch.

On the other hand, taking into account all the results as a single batch provides more data about steady state. Performance measures are calculated basing on a longer period of time so they are more informative. This approach is efficient if simulation runs are long or expensive. Its disadvantage is that it is more sensitive to large deviations of the estimated measure - because here we average only once, and not two times as in the implemented method. The Expected Waiting Time if there is Waiting $E(W|W > 0)$ was calculated following the Single-Batch Approach because according to the simulation results, it converges to its theoretical value much faster if it is implemented this way.

As an example, here is the calculation of Abandonment Probability during time interval t carried out under the Single-Batch Approach.

Abandon Probability of type *i* during interval *t* $P_t^i(Aband)$

$$P_t^i(Aband) = \frac{\sum_{j=1}^{Reps} Abandoned_t^{i,j}}{\sum_{j=1}^{Reps} Arrived_t^{i,j}}$$

Table 6.2 summarizes the existing and the alternative calculation methods.

Abandonment Probability

There is a certain ambiguity in the definition of the Abandonment Probability. In the simulation software

	Existing	Alternative
$P_t^i(Aband)$	$\frac{\sum_{j=1}^{Reps} P_t^{i,j}(Aband)}{Reps}$	$\frac{\sum_{j=1}^{Reps} Abandoned_t^{i,j}}{\sum_{j=1}^{Reps} Arrived_t^{i,j}}$
$P_t^i(W_q > 0)$	$\frac{\sum_{j=1}^{Reps} P_t^{i,j}(W_q>0)}{Reps}$	$\frac{\sum_{j=1}^{Reps} Delayed_t^{i,j}}{\sum_{j=1}^{Reps} Arrived_t^{i,j}}$
W_t^i	$\frac{\sum_{j=1}^{Reps} W_t^{i,j}}{Reps}$	$\frac{\sum_{j=1}^{Reps} (Waiting\,Time)_t^{i,j}}{\sum_{j=1}^{Reps} (Arrived)_t^{i,j}}$
$E(W_t^i \mid W_t^i > 0)$	$\frac{\sum_{j=1}^{Reps} E(W_t^{i,j} \mid W_t^{i,j}>0)}{Reps}$	$\frac{\sum_{j=1}^{Reps} (Waiting\,Time)^{i,j}}{\sum_{j=1}^{Reps} (Delayed)^{i,j}}$
W^i	$\frac{\sum_{j=1}^{Reps} W^{i,j}}{Reps}$	$\frac{\sum_{j=1}^{Reps} (Waiting\,Time)^{i}}{\sum_{j=1}^{Reps} (Released)^{i}}$
$E(W^i \mid W^i > 0)$	$\frac{\sum_{j=1}^{Reps} Total\,Waiting\,Time^{i,j}}{\sum_{j=1}^{Reps} Total\,Delayed^{i,j}}$	$\frac{E(W^{i,j} \mid W^{i,j}>0)}{Reps}$
$P^i(Aband)$	$\frac{\sum_{j=1}^{Reps} P^{i,j}(Aband)}{Reps}$	$\frac{\sum_{j=1}^{Reps} Abandoned_t^{i}}{\sum_{j=1}^{Reps} Released_t^{i}}$
$P^i(W_q > 0)$	$\frac{\sum_{j=1}^{Reps} P_j^{i}(W_q>0)}{Reps}$	$\frac{\sum_{j=1}^{Reps} Delayed_t^{i}}{\sum_{j=1}^{Reps} Released_t^{i}}$

Table 6.2: Existing and Alternative Ways of Performance Measures Calculations

we consider the ratio of those customers who arrived during time interval *t* and abandoned later (not necessary during this time interval) and the total arrivals during the same time interval.

One can also think of Abandonment Probability in the following way, taking into account all the previous intervals: it can be found as a ratio of all the customers who decide to abandon until time interval *t* and the total number of arrived customers till this time interval. These two approaches may lead to different results in their evaluation.

Here we present the second approach. Let us consider the abandonment rate at time *t* as the number of customers who abandon during the time interval *t*; denote it r_t. Theoretically, using balance equations,

this arrival rate is found from the following expression:

$$r_t = \theta * E[Q_t],$$

where Q_t is the queue length at time t. In this case the abandonment probability is calculated via

$$P_t(Aband) = r_t/\lambda_t,$$

where λ_t is the arrival rate at time t.

One sees that, from the customer point of view, the method implemented in the simulation software is more ”informative”, while the alternative method has more managerial insights.

Moreover, under the QED regime, we expect that $r_t/\sqrt{R_t}$ should be approximately constant. So, following this new definition of abandonment probability, we obtain that the ratio $P_t(Aband) * \lambda_t/\sqrt{R_t}$ is also approximately stable. Simulation experiments confirm that. However, this observation is not practically useful for stabilizing $P_t(Aband)$, since $\lambda_t/\sqrt{R_t} \approx \sqrt{R_t}$ is of the order of 10’s.

Chapter 7

Time-Stable Performance of Time-Varying Queues with Static Priorities

This chapter explains how it is possible to identify staffing levels that give rise to a time-stable performance for *all* types. We use the same approach as in [9] and apply its Iterative Staffing Algorithm (ISA) within a time-varying environment. We present detailed results for two $M_t/M/s_t + M$ models (time-varying Erlang-A) and a summary of several additional models, all with two types of customers.

As before, the assumption is that servers are independent but statistically identical; in other words, service times for *all* customers have the *same* exponential distribution. For such models, the main findings of our analysis are as follows:

- Overall success in stabilizing performance: very successful in stabilizing the delay probability α and implied service grade β, and reasonably successful in stabilizing waiting times, queue lengths and abandonment rates, especially for the high priority customers.
- Global performance of our time-varying systems correspond to an appropriate stationary system. The fit is better for systems with low fraction of one of the types and less exact for systems with approximately the same fraction of arrivals for both types.
- Dependence on the total arrival rate only, as opposed to the vector of type arrival-rates. (It is rather clear to us that this is due to the fact that all service times are type-independent and identically distributed.) Consequently, the staffing problem can be reduced to staffing a *single-type* Erlang model.

Abandonments play a crucial role in system performance; indeed, adding abandonment enables the algorithm to converge significantly faster than those without. For instance, the running horizon was 24 time units (instead of 72 time units required for Erlang-C queues). The number of iterations till ISA convergence is also small due to the customers' impatience. The highest number of iterations is 4, after which the algorithm always converges even for $\alpha = 0.8$ and 0.9, while in queues without abandonment,

for these values of α the algorithm did not succeed to converge even after 100 iterations.

If customers do not abandon, the warming-up period is relatively long. To ensure that a system reaches steady state, we take 72 time units as the simulation running horizon. For under-loaded systems (with target $\alpha = 0.1, 0.2, 0.3$) the number of iterations till convergence is small, while for highly-utilized systems ($\alpha = 0.8, 0.9$) it could be very large, hence the highest value of α we consider in the simulation experiments of Erlang-C queues is 0.75. The number of iterations till convergence for this latter α is 12 at the most. If we continue experiments to $\alpha = 0.8$ or more, this number exceeds 100 iterations, which takes about 6 hours in computer time.

7.1 Two-types customers in the QED regime. Simulations results

To check the performance of ISA in an environment with heterogeneous customers, the algorithm was applied to various queueing systems with time-varying arrival rates. The results are described below.

7.1.1 An Example with the Time-Varying Erlang-A Model

Here we present the performance of ISA for the time-varying Erlang-A models ($M_t/M/s_t + M$) with two customer types and sinusoidal arrival rates of each type.

Models Description:

- The running horizon is 24 time units and performance statistics are collected after the 6^{th} time units to make sure that the system reaches to a steady state.
- All the values are calculated as an average of 5000 iterations.
- Service time is distributed exponentially with mean $1/\mu$ = 1 time unit.
- Customer patience is distributed exponentially with mean $1/\theta = 1$ time unit.
- There are two types of customers. Customers of the first type have a non-preemptive priority over the second type customers.
- The queue discipline within each class is FCFS.

First (70-30) System Arrival Rates:

- First-type arrivals are given by a non-homogenous Poisson process with the arrival rate $\lambda_1(t) = 70 + 21 \cdot sin(3t)$. Period is $2\pi/3$.
- Second-type arrivals are given by a non-homogenous Poisson process with the arrival rate $\lambda_2(t) = 30 + 12 \cdot sin(2t)$. Period is π.

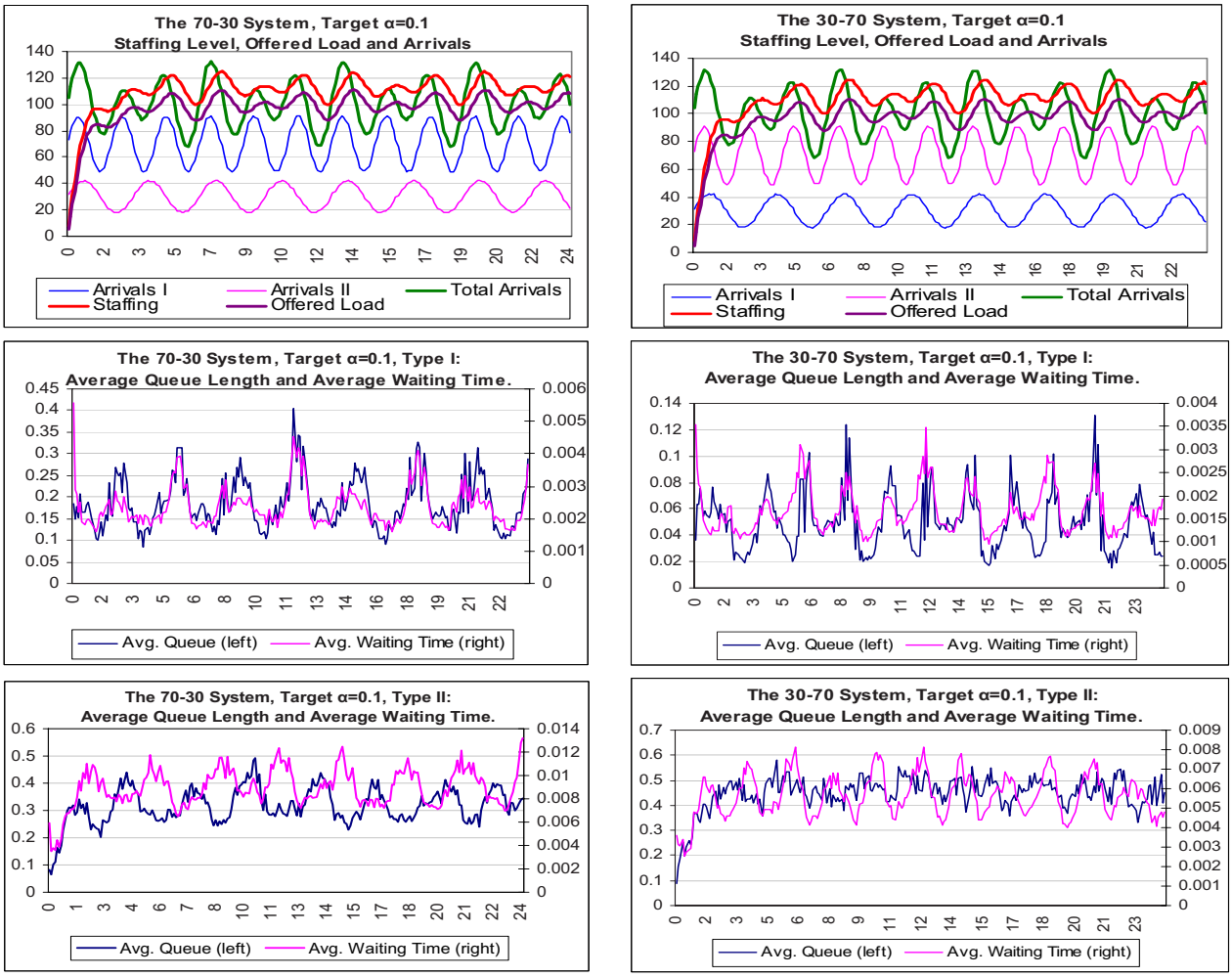
Second (30-70) System Arrival Rates:

- First-type arrivals are given by a non-homogenous Poisson process with the arrival rate $\lambda_1(t) = 30 + 12 \cdot sin(2t)$. Period is π.
- Second-type arrivals are given by a non-homogenous Poisson process with the arrival rate $\lambda_2(t) = 70 + 21 \cdot sin(3t)$. Period is $2\pi/3$.

The total arrival rate in both systems is given by $\lambda(t) = 100 + 12 \cdot \sin(2t) + 21 \cdot \sin(3t)$ with period 2π.

Staffing levels, obtained for both systems for the tested values of α, are shown in Figures 7.1-7.3. The total arrival rate is the same in both systems, hence for the same values of α the determined staffing level is the same. Queue lengths and expected waiting times are presented in the second part of these Figures.

Figure 7.1: Target α=0.1 - (1) Staffing Level, Offered Load and Arrival Function; (2) Waiting Time and Queue Length of Both Classes.



Customers abandonment makes a system stable. As mentioned above, the running horizon was decreased from 72 in Erlang-C to 24 time units because systems with abandonment converge to steady state significantly faster. As expected, the algorithm obtains stable delay probability for both customer types and for both systems throughout its running period. (See Figure 7.4.)

Figure 7.5 presents a summary of abandon probabilities for the highlighted QD, QED and ED regimes. This figure shows that ISA is less successful in stabilizing the abandonment probability for large values of the target delay probability α. The greater target α, the less stable the abandon probability.

In our systems, customers have exponential patience with $\theta = 1$ so in stationary models, by the relation

Figure 7.2: Target α=0.5 - (1) Staffing Level, Offered Load and Arrival Function; (2) Waiting Time and Queue Length of Both Classes.

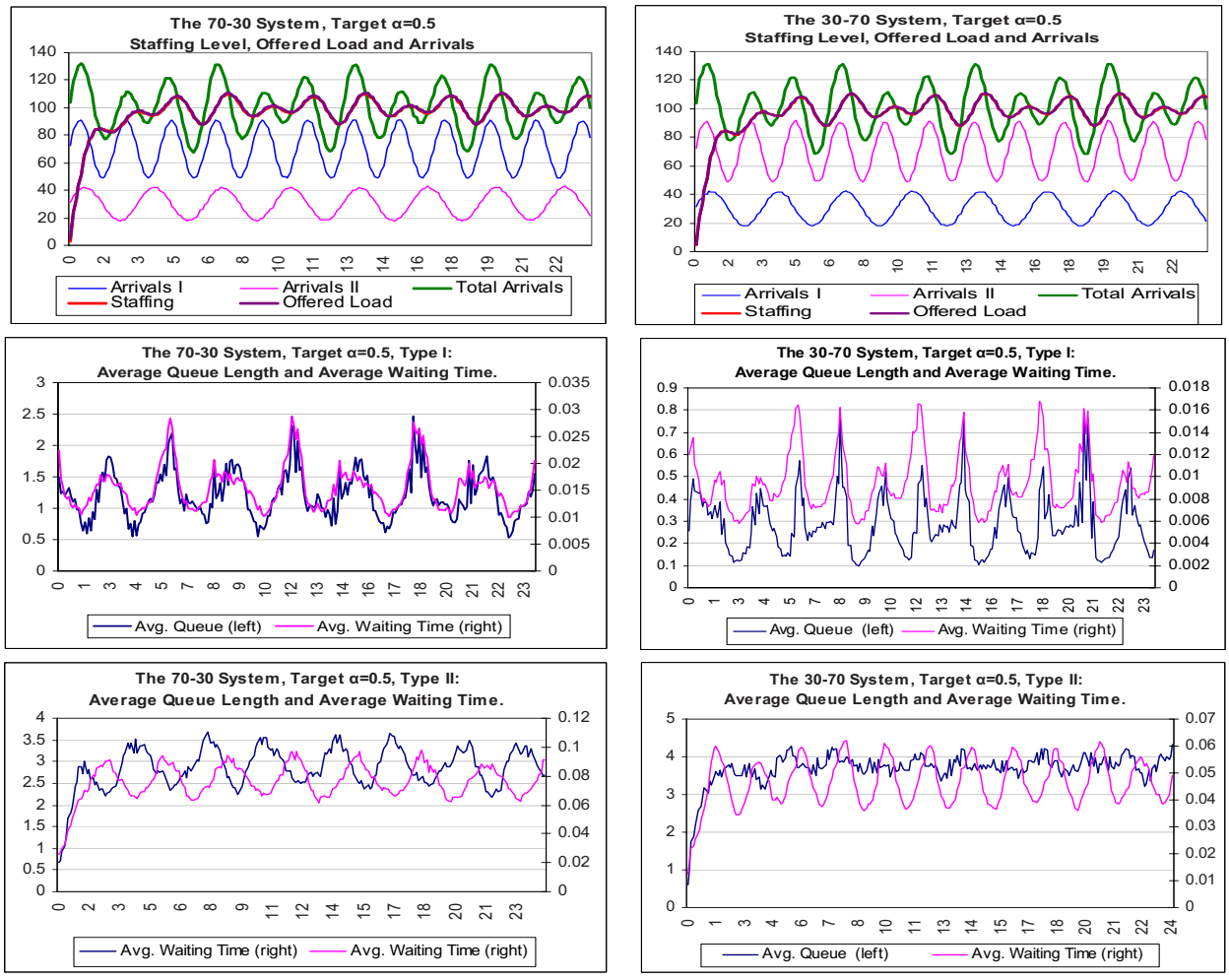
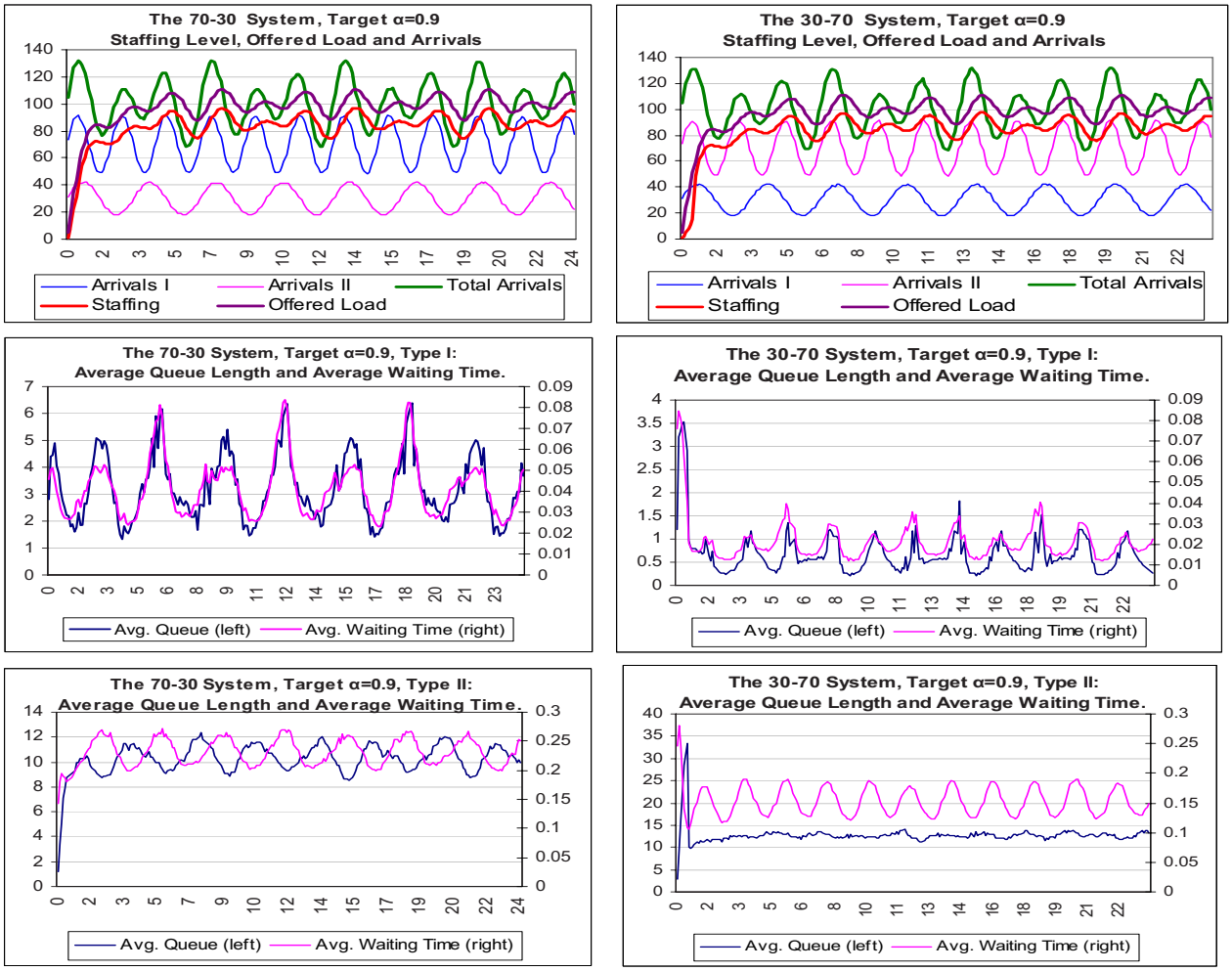


Figure 7.3: Target α=0.9 - (1) Staffing Level, Offered Load and Arrival Function; (2) Waiting Time and Queue Length of Both Classes.



$P(Aband_i) = \theta E[W_i]$ the abandon probability must be equal to the expected waiting time. Figures 7.6 - 7.7 verify this relation for our time-varying models under the three considered regimes. It follows that the abandon probability of the high priority is virtually equal to the expected waiting time for all the three highlighted values of α, as predicted by the theoretical relation between the expected waiting time and the abandonment probability in stationary queues. The abandonment probability of the low priority is always above the expected waiting time, except for the case $\alpha = 0.1$ in the 30-70 system where these two measures are rather close. In all three cases, there clearly exist a dependence between the abandonment probability and the expected waiting time, but this relation is different from from equality of the abandonment probability and the expected waiting time for the lowest priority.

Figure 7.4: Summary of Delay Probability.

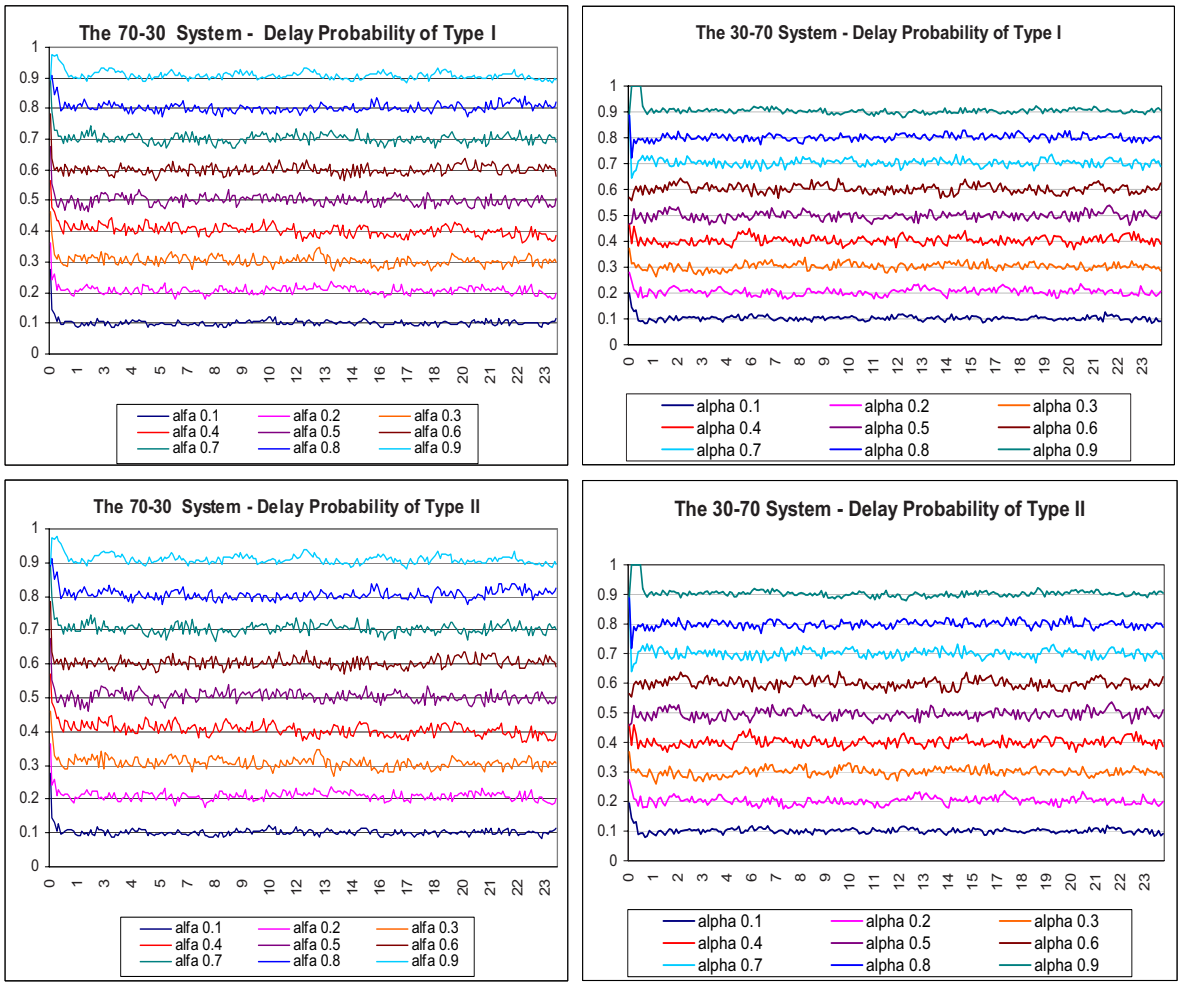


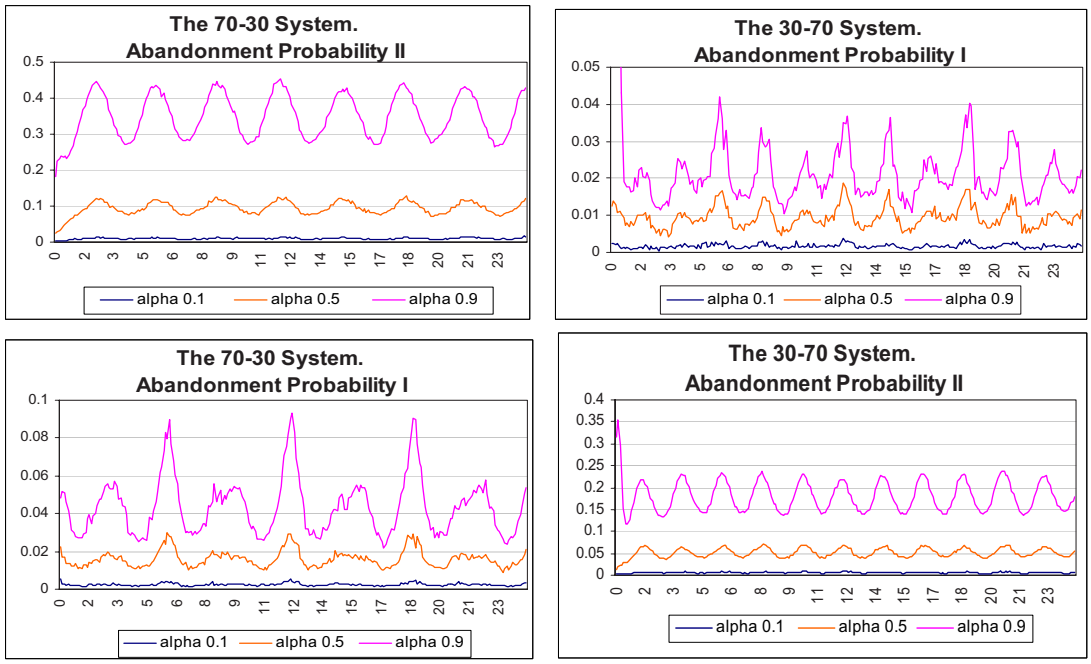
Figure 7.8 is the comparison of the theoretical curve and the empirical pairs (α_i, β_i). The empirical values in both systems are very close to the theoretical curve, even for the highest $\alpha = 0.9$.

Dynamics of the implied grade of service β is presented in Figure 7.9. It is rather stable during simulation runs. In the simulation of 30-70 system with the target $\alpha = 0.9$ it drops down during the first time unit but then stabilizes fast. In the 70-30 system there is no such a sharp drop but there are more fluctuations during the simulation run.

The utilization summary of both systems for different values of the target α is presented in Figure 7.10. This measure stabilizes relatively fast - during the first two time units for all values of the target α.

Waiting-time histograms are presented in Figures 7.11-7.12. It is interesting to observe that waiting time

Figure 7.5: Abandonment Probability.



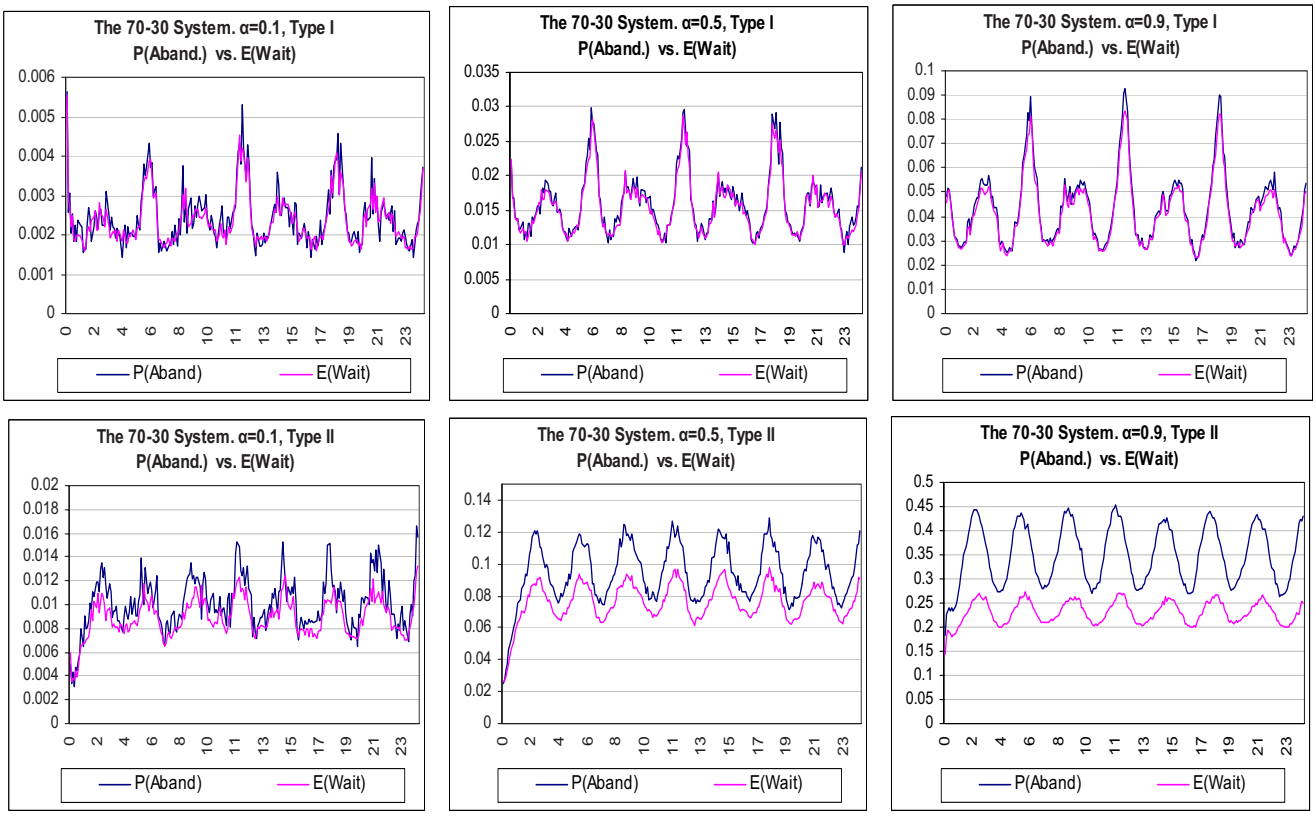
of the **first** class in both systems can be approximated by the exponential distribution similar to stationary $M/M/N$ model without abandonment.

To build the theoretical curves of the waiting-time distribution both in 70-30 and 30-70 models we took the average number of servers during the period from 6^{00} till 24^{00}. The average arrival rate was calculated over the same period ($\lambda_1 = 70$ in the 70-30 queue, or $\lambda_1 = 30$ in the 30-70 queue). The service rate of the stationary model was assumed to be $\mu = 1$. In this case, the theoretical curve was built using the exponential distribution with the mean $\frac{1}{N\mu(1-\rho_1)}$.

To analyze the waiting time distribution of the first type, we actually use the following fact observed in the previous chapter. Given waiting, the highest priority is not aware of the delayed second-type customers. Hence λ_2 is not needed to predict how long the **delayed** customers of the first type will wait. This is why when the staffing level is determined by the QED regime and all the servers are busy, the first-type customers experience QD performance (conditioned on waiting). In the QD regime the probability of abandonment is very small, and hence the approximation by the Erlang-C model works.

The stationary distribution of the lowest priority is more problematic since, in this case, the abandonment probability cannot be ignored. Here we present only the simulated histogram without any theoretical curve.

Figure 7.6: The 70-30 System, $\alpha = 0.1$ - Abandon Probability vs. Waiting Time.



7.1.2 Results and Conclusions

We end with section with some additional results, comparisons and conclusions, based on our experiments.

- **The impact of abandonment: Erlang-C vs. Erlang-A**

Abandonments play a very important role in queue performance. Taking them in account not only makes a model more realistic, it also improves most of the performance measures. As we mentioned earlier, the running horizon in the simulation of Erlang-A queues was decreased from 72 to 24 hours, since queues with abandonments reach steady state much faster than those without. In addition, ISA managed to converge for all values of α from 0.1 to 0.9, while in Erlang-C we had to stop after $\alpha = 0.75$; and the maximal number of the algorithm iterations till convergence was 4 instead of 12 for the Erlang-C.

Abandonments also allow to decrease the staffing level required to obtain the desired delay probability α. Figure 7.13 presents the final staffing for $\alpha = 0.1$, 0.5 and 0.9 for the simulated queues with and

Figure 7.7: The 30-70 System, Abandon Probability vs. Waiting Time.

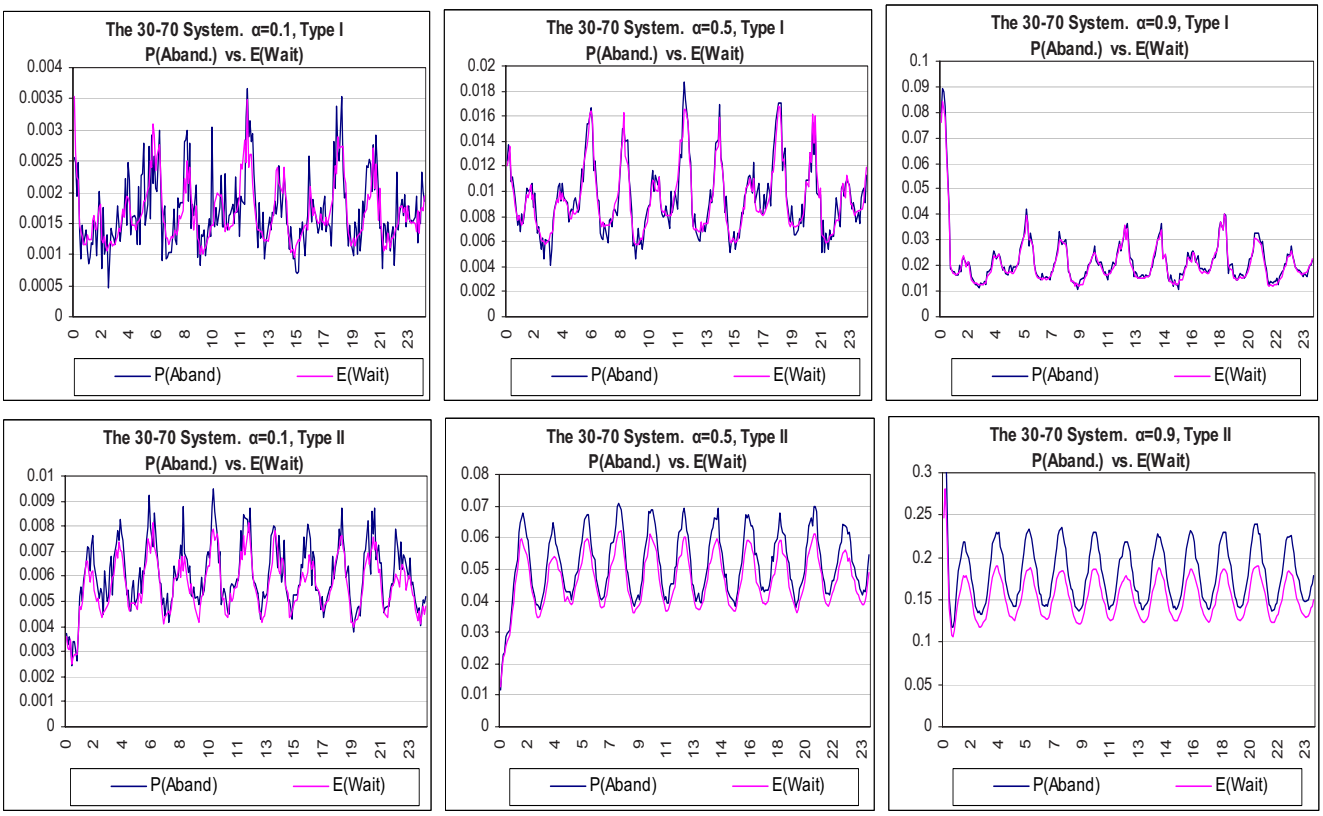


Figure 7.8: Theoretical (Garnett Function) and Empirical Probability of Delay vs. β

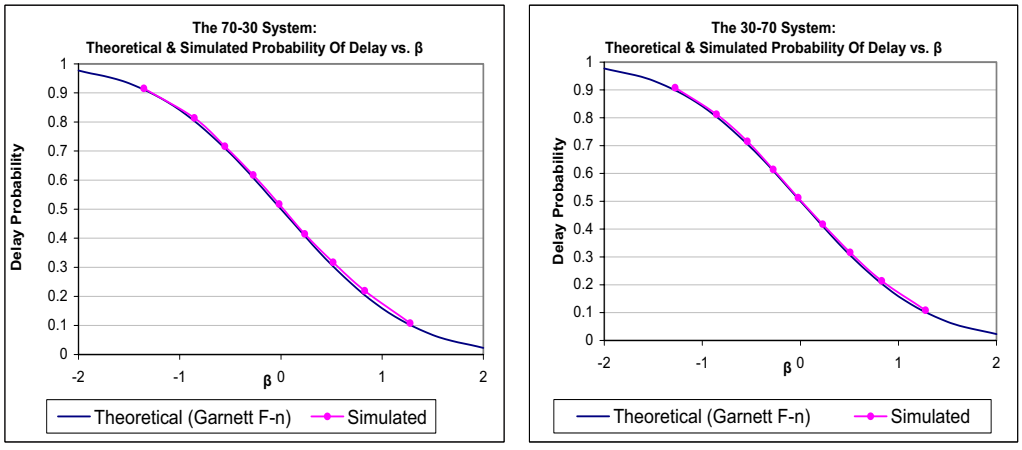


Figure 7.9: Summary of the Implied Service Grade β.

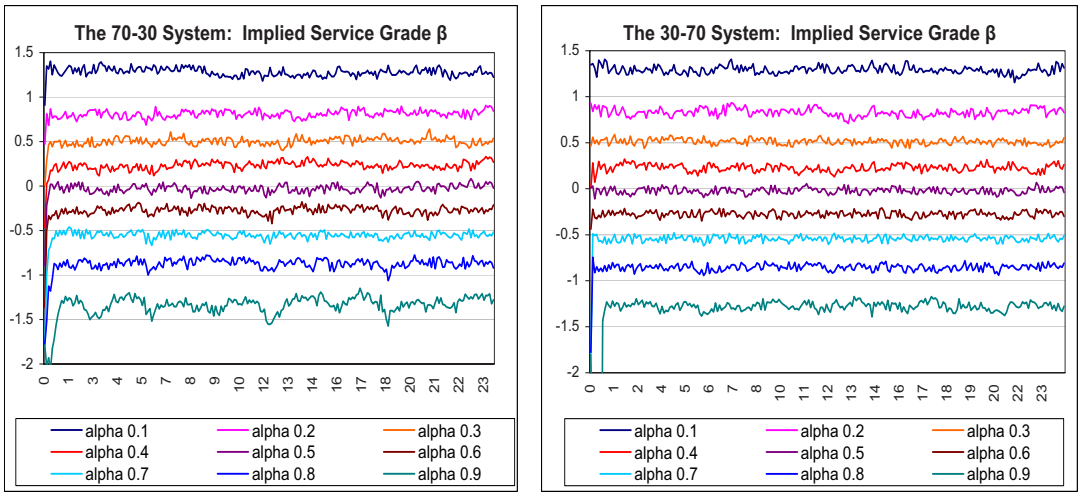
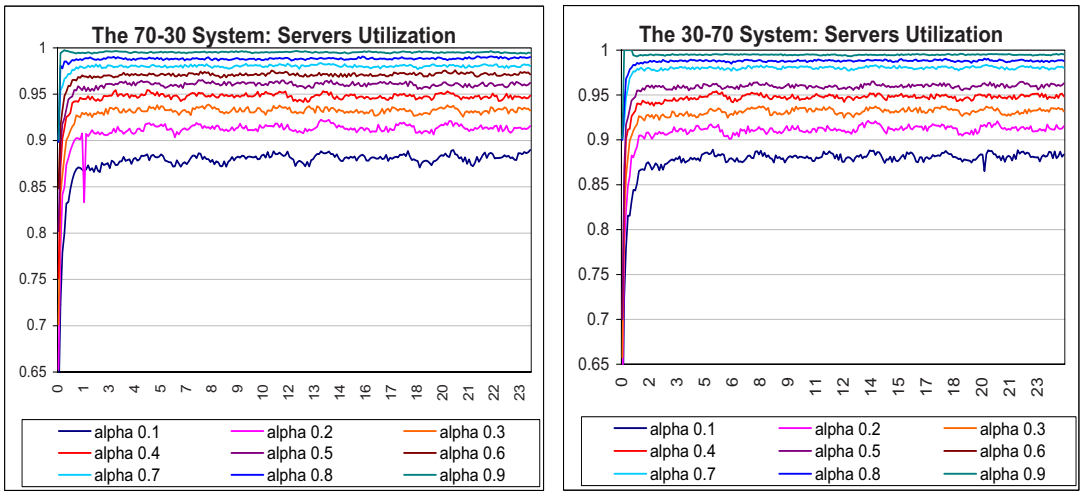


Figure 7.10: Utilization Summary

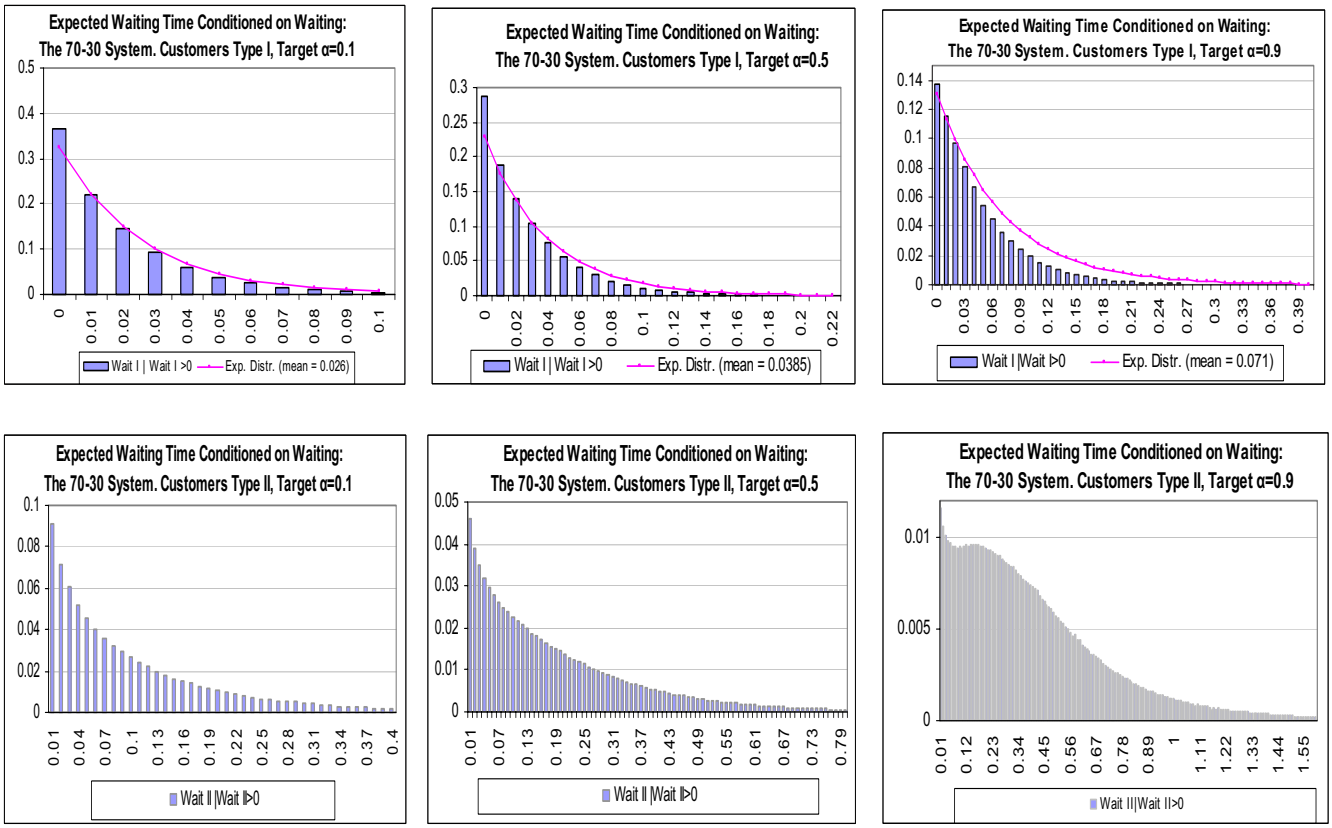


without abandonment. [1]

The savings of labor can be quantified by the area between the staffing curves. It comes out that allowing customers patience with $\theta = 1$ leads to total labor savings of 143.4 time units for $\alpha = 0.1$, 214.8 time units for $\alpha = 0.5$ and 314.92 time units for $\alpha = 0.9$. It may perhaps be better to quantify savings by

[1]The algorithm did not converge for $\alpha = 0.9$ for the Erlang-C model but using the fact that staffing of a multi-type queue is similar to staffing of a single-type queue with the same total arrival rate and relying on the Feldman Z. et al [9] we assume that in the Erlang-C queue in order to obtain $\alpha = 0.9$ (ED) one should staff close to the offered load. This is why in Figure 7.13 (3), the Erlang-C Staffing and the Offered Load curve are similar.

Figure 7.11: The 70-30 System - Waiting Time Histograms.



looking at the savings of labor per shift. Dividing the saving in time-units by the number of time-units they are taken over, we come up with savings of about 6, 14 and 114 servers per shift, for α = 0.1, 0.5 and 0.9 respectively. The labor savings increase as α increases.

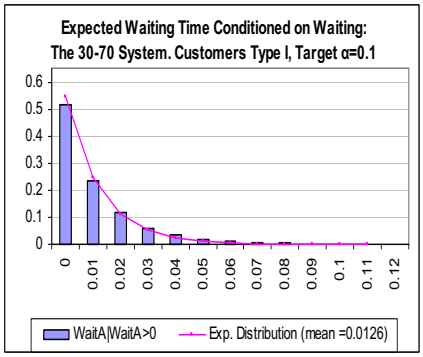
- **Erlang-A with Priorities vs. Erlang-A with Homogenous Customers**

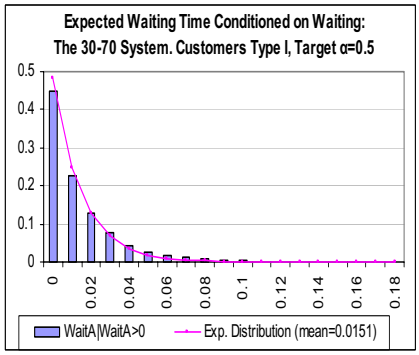
To see how system's performance is influenced by the customers differentiation, we compare the results of 70-30 and 30-70 queues simulations with the results of a single-class $M_t/M/s+M$ queue simulation with arrival rate $\lambda(t) = 100 + 21\sin(3t) + 12\sin(2t)$, that is, equal to the total arrival rate of the 70-30 and 30-70 systems. The results are as follows:

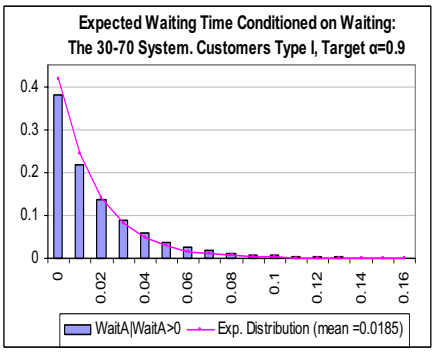
- A single-class queue reaches the steady state faster than a queue with heterogenous customers. This is clear in Figure 7.14, which presents a summary of the implied service grade β for all values of the target α.

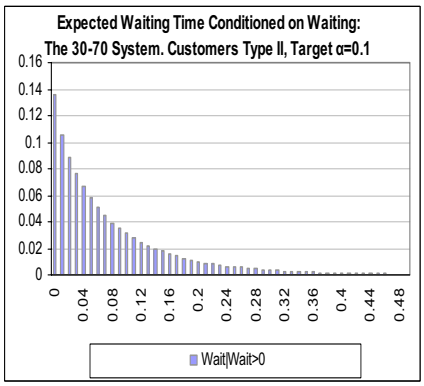
  One can see that β stabilizes immediately for all α, while in the two-types systems for high values of α there was some warmup period.

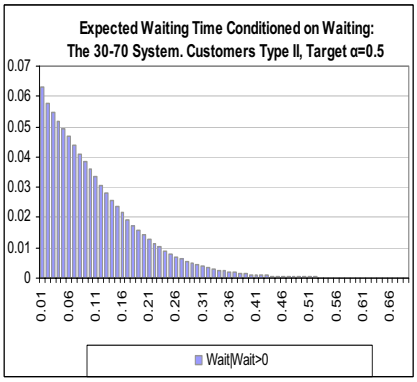
Figure 7.12: The 30-70 System - Waiting Time Histograms.

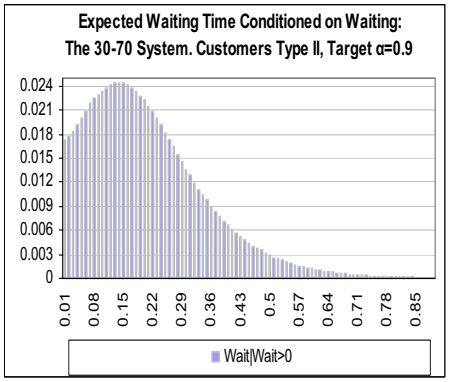












- As expected, customers prioritization **shortens** expected waiting times and queue lengths of the high priority and **increases** these for the low priority. Figure 7.15 presents the waiting times and the queue lengths for the Erlang-A queue with homogenous customers. The following Figure 7.16 compares these measures with the waiting times of the 70-30 and the 30-70 systems with the target $\alpha = 0.9$ (The differences are the largest for this α.) One can see that in the 30-70 system the first-type customers "profit" more from their status: their average waiting time decreases almost 10 fold (!) for $\alpha = 0.9$. Additionally, in the 30-70 system, the waiting time of the low priority grows about 1.5 times, while in the 70-30 system the decrease of the waiting time for the high priority is about 2 times, and the increase for the low priority is about 2 times. This makes sense because the fraction of the high priority is greater in the 70-30 system, hence a first-type customer, although enjoying his priority over the second type, yet needs to wait for other customers of his type which arrived before him.

Figure 7.13: Staffing: Erlang-C vs. Erlang-A. (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED), (3) $\alpha = 0.5$ (QED)

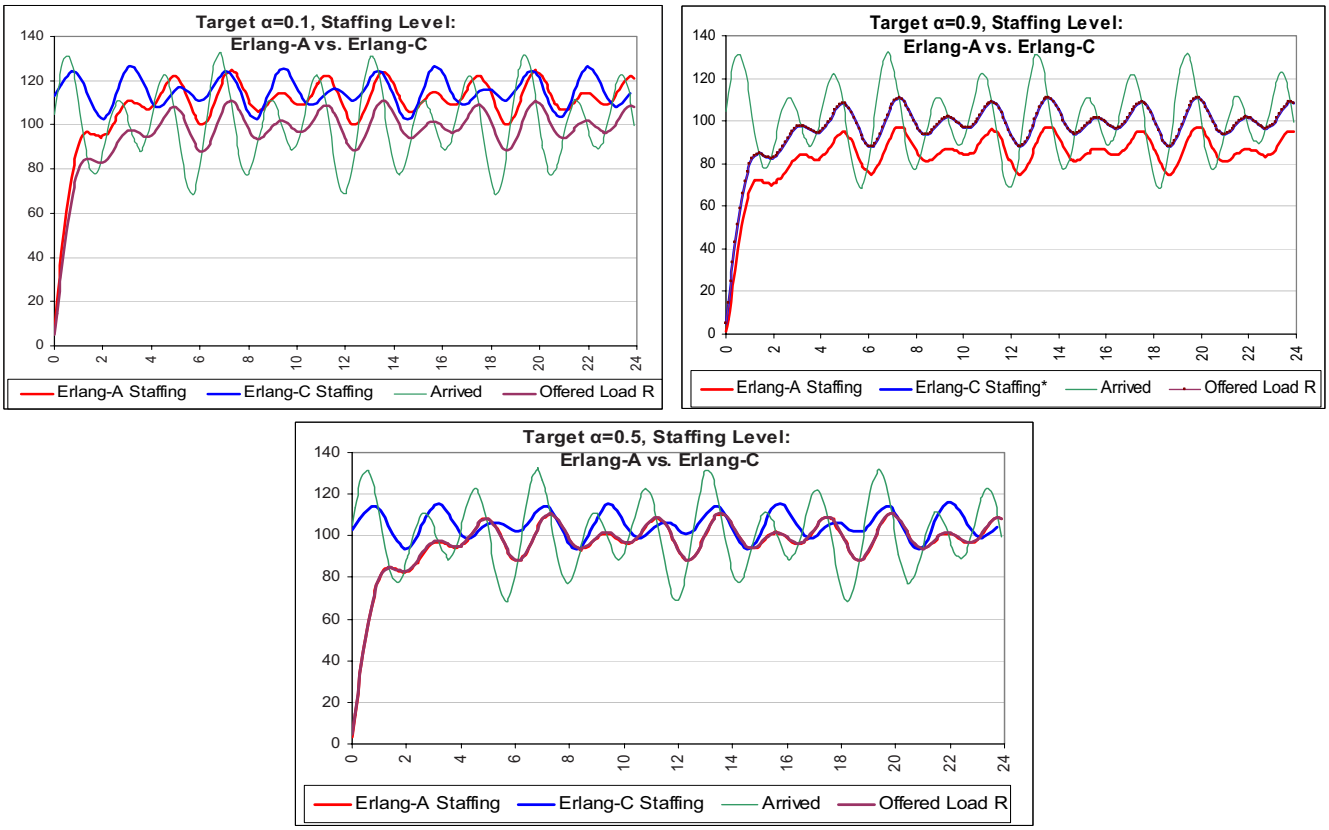


Figure 7.14: Implied Service Grade β for the Single-Type Queue

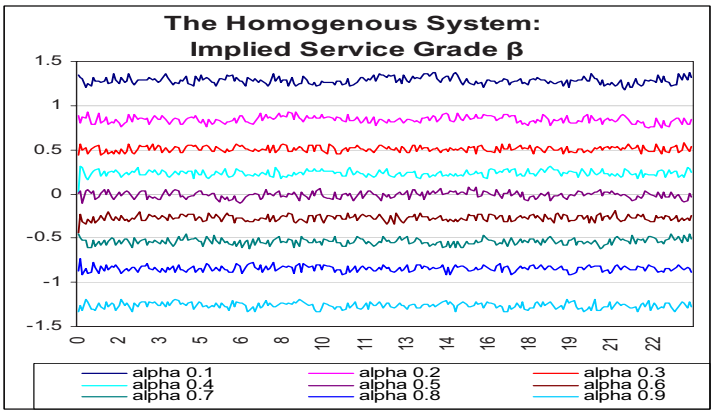


Figure 7.15: Waiting Times and Queue Lengths for the Single-Type Queue

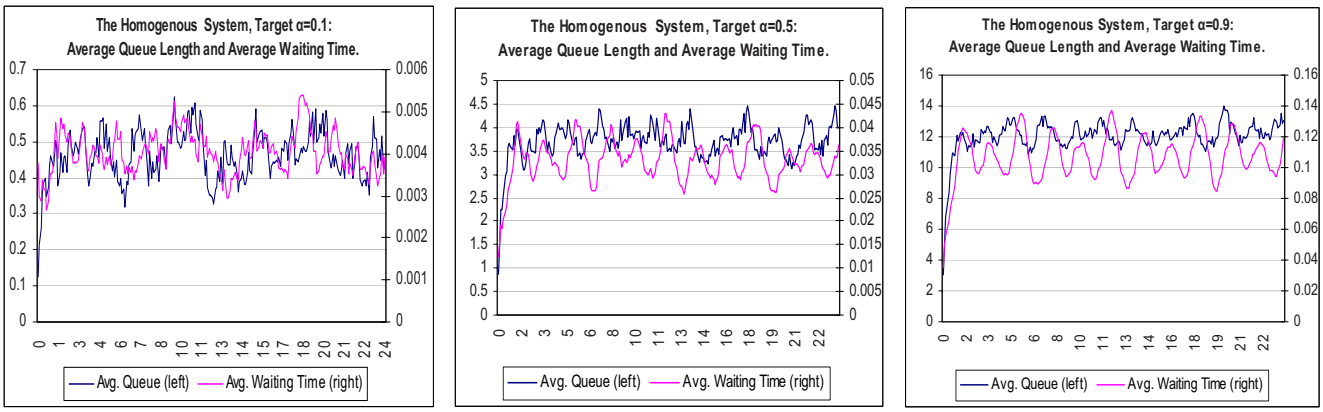
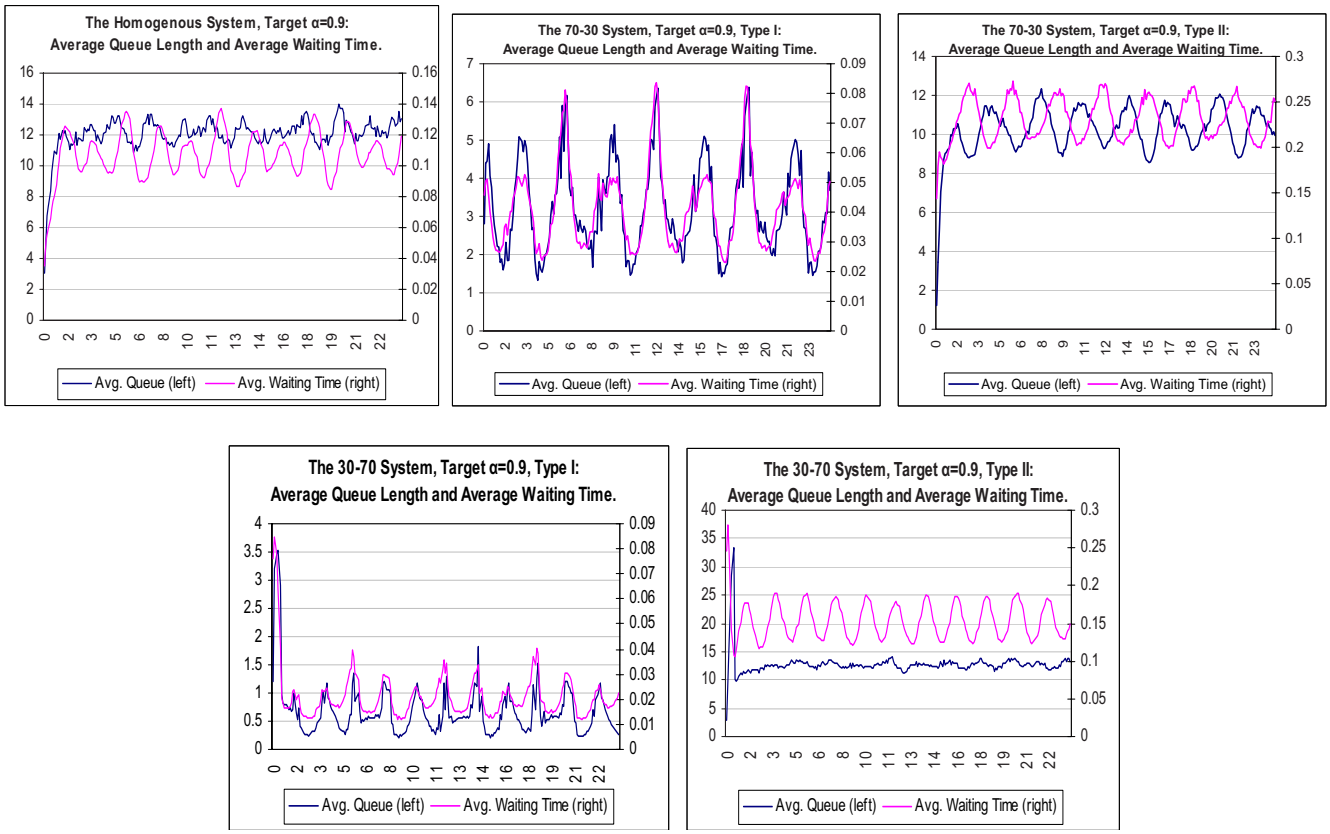


Figure 7.16: Target $\alpha = 0.9$: Waiting Times and Queue Lengths of the Single-Type Queue vs. 70-30 and 30-70 Queues



Chapter 8

Heavy-Traffic Approximations

Conventional/Classical heavy-traffic approximations of queues are typically **two-moment approximations** in the sense that means, variances and covariances of the input parameters determine the approximations. For example, the same heavy-traffic performance is expected for $M/M/N$ and $M/LN/N$ (LN stands for log-normal) if the mean and standard deviation of LN are equal, assuming that the two systems have the same arrival rate, service rate and the same number of servers.

However, this is not what happens in practice for highly utilized systems in which many servers work in parallel. The project of Schwartz [32] contains some simulation results for the delay probability and the expected waiting time conditioned on waiting for three different service time distributions - exponential $(M/M/100)$, log-normal $(M/LN/100)$ and deterministic $(M/D/100)$. The purpose of these simulations was to compare highly utilized systems with a large number of servers with different service-time distribution. The first two moments of the log-normal and the exponential distributions are equal to 1 and, as mentioned above, the number of servers is 100.

Following conventional heavy traffic approximations, the expected waiting time should be equal for the exponential and log-normal cases, but, as follows from [32], log-normal services give rise to significantly lower values of the waiting time. Note that the ordering of the delay probabilities $(D < LN < Exp)$ is consistent with that of the expected waiting time conditioned on waiting.

The purpose of the experiments described in this chapter is to check whether it is possible to predict the Expected Waiting Time by using only the first two moments of the service-rate distribution for queues with impatient customers under the QED regime.

This chapter is organized as follows. Section 8.1 shows a derivation of a closed-form expression for the expected queue length under the ED regime. The analysis in this section is based on the paper of Ward [33], where she obtains the two-moments approximation for the queue length. When the expected queue length is known, the expected waiting time is directly derived from Little's Law. The section closes with the presentation of the simulation results which show that as the offered load increases, the approximation of the expected waiting time becomes more and more precise.

In Section 8.2 we repeat the experiments of Schwartz [32] but for $M/GI/100 + M$ queues under the

QED regime. The simulation results expand the main findings of [32] to queues with abandonment and show that, in the QED regime, the order among average waiting times that arose in [32], due to varying service-time distributions with equal first moments, is preserved.

Lastly, a short comparison of queues with 100 servers with and without abandonment is given in Section 8.3.

8.1 Heavy-Traffic Approximations

Following the approach in [33], in this section we develop a two-moment approximation for the average queue length and waiting time in the $GI/GI/N+GI$ queue under heavy traffic.

We start with a single-server $GI/GI/1+GI$ queue with arrival rate λ and service rate μ. Let u_1 be the distribution of an inter-arrival time, v_1 be the distribution of a service time and F be a distribution of customers patience.

As shown in [33], the steady-state mean of the $GI/GI/1+GI$ queue is approximately given by

$$E(L_q) \approx E(N(m,\ b^2)|N(m,b^2) \geq 0), \tag{8.1}$$

where $N(m,\ b^2)$ is a Normal variable with a mean m and variance b^2. The approximation is asymptotically valid in heavy traffic, namely while ρ_N converges to 1 as $\lim_{N\to\infty}\sqrt{N}(1-\rho_N) = c$ for some $c, \quad -\infty < c < \infty$.

Here

$$m = \frac{\mu(\rho-1)}{F'(0)}; \qquad\qquad b^2 = \frac{\mu^3[\rho\, var(u_1) + (\rho \wedge 1)var(v_1)]}{2F'(0)} \tag{8.2}$$

(Note that for notational simplicity, in [33] the service rate μ is equal to 1.)

Let us find a closed-form expression for the conditional expectation (8.1).

$$E(N(m,\ b^2)|N(m,b^2) \geq 0) \quad =$$

$$\frac{1}{\Phi(m/b)} \int_0^\infty \frac{x}{b\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2b^2}\right) dx$$

$$= \frac{1}{\Phi(m/b)} \left[\int_0^\infty \frac{x-m}{b\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2b^2}\right) dx \quad + \right.$$

$$\left. \int_0^\infty \frac{m}{b\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2b^2}\right) dx \right]$$

$$= \frac{1}{\Phi(m/b)} \left[\int_0^\infty \frac{b}{\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2b^2}\right)\ d\frac{(x-m)^2}{2b^2} \quad + \quad m\Phi(m/b) \right]$$

$$= \frac{b \ \exp\left(-\frac{m^2}{2b^2}\right)}{\sqrt{2\pi}\Phi(m/b)} + m.$$

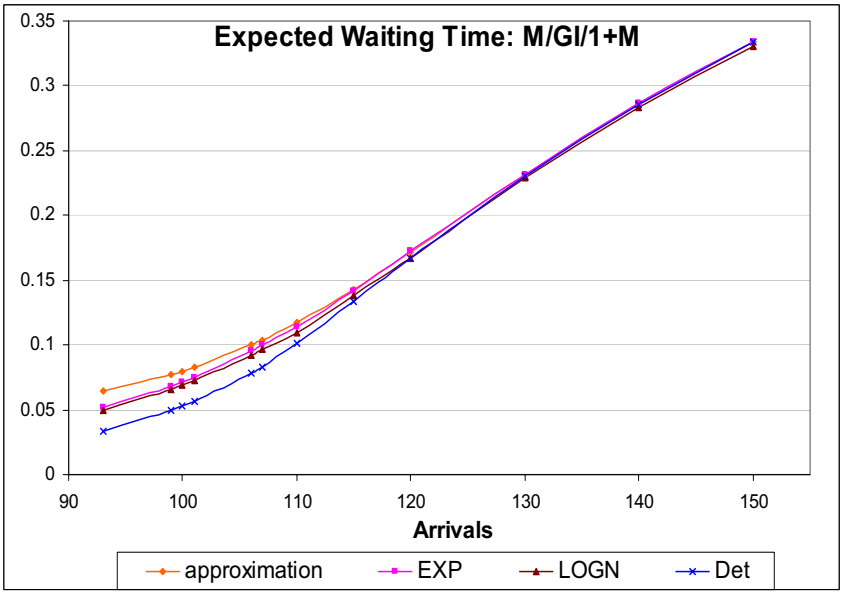
The average queue length is thus given by:

$$E(L_q) \approx m + \frac{b \ \exp\left(-\frac{m^2}{2b^2}\right)}{\sqrt{2\pi}\Phi(m/b)}. \tag{8.3}$$

Then the approximated expected waiting time is a direct consequence of (8.3) and Little's Law:

$$E(W_q) = \frac{L_q}{\lambda} \approx \frac{m}{\lambda} + \frac{b \ \exp\left(-\frac{m^2}{2b^2}\right)}{\sqrt{2\pi}\ \lambda\ \Phi(m/b)} \tag{8.4}$$

Figure 8.1: The $M/GI/1+M$ queue - Empirical Results vs. Approximations



Now, we will consider the single-server Erlang-A ($M/M/1+M$) as an important special case. Let us assume that the arrival rate is λ, the service rate is μ and the abandonment rate is θ.
Here

$$u_1 \stackrel{d}{=} exp(\lambda), \qquad v_1 \stackrel{d}{=} exp(\mu) \qquad \text{and} \qquad F'(0) = \theta.$$

By substituting the distributions of u_1 and v_1 into (8.2), we obtain the following:

$$m = \frac{\lambda - \mu}{\theta}, \qquad b^2 = \frac{\mu^2 + \lambda(\lambda \wedge \mu)}{2\lambda\theta}. \tag{8.5}$$

To check the performance of the approximation (8.4), we compared it with the simulated average waiting time in $M/GI/1+M$ queues with three different service-time distributions. The considered service-time distributions were: Exponential, Log-Normal with CV=1 and Deterministic.

For each simulated environment, the rest of the parameters were as follows:

- The arrival rate changes from 93 to 150 customers per hour;
- The service rate is $\mu = 100$ customers per hour;
- The individual abandonment rate is $\theta = 1$.

Figure 8.1 presents a summary of the average waiting time values obtained under different service-time distributions and arrival rates. The comparison of the theoretical waiting time with the empirical results shows that the approximation is excellent for highly utilized systems ($\rho \geq 1.2$).

Some additional results are found in the summary of simulations, conducted by Reed [30]. This work considers $M/M/1+GI$ queues with equal arrival and service rates ($\lambda = \mu = 2500$ customers per hour) and the mean of individual abandonment 1 hour. The queues differ in the patience distributions.

The report [30] includes results of Deterministic abandonment distribution and Gamma distribution $G(p)$, where p is the shape parameter, for $p = 5, \ldots, 0.2$. The variance of $G(p)$ is $\frac{1}{p}$, and it increases as p decreases.

The first conclusion of these experiments is that the relative error in the queue length and in the abandonment probability increases as the variance increases (or as the shape parameter p decreases). The error in the Queue-Length prediction for Deterministic abandonment distribution was only 3.38 %, while for $G(0.2)$ it was as big as 23.56%. The abandonment probability is influenced in the same fashion: The error in the case of the patience distribution $G(5)$ was only 1.41%, and under $G(0.2)$ it grew up to 49.36%.

The accuracy of the approximation is also a matter of scale,which follows from the second sequence of simulations in [30]. This second experiment tested the $M/M/1/+GI$ queues with $GI = G(0.2)$. Arrival rate λ is again equal to the service rate μ, $\lambda = \mu = n, \quad \ldots \quad n = 1000, \ldots, 100,000,000$. The error in queue length decreases from 16.38% for $n = 1000$ to 2.92 % for $n = 100,000,000$; and the error for abandonment probability decreases from 31.57% for $n = 1000$ to 3.19% for $n = 100,000,000$.

This observation is in line with our results presented in Figure 8.1. We conduct experiments with service rate $\mu = 100$, i.e. one order less than the minimal in [30]. Relying on the results of [30], we can thus conclude that the error of the queue-length approximation (8.1) will be relatively large. This conclusion is supported by the plot presented in Figure 8.1.

8.2 $M/G/100+M$ Queues

As already mentioned at the beginning of this chapter, according to the heavy-traffic approximations in Erlang-C queues, the expected waiting time depends on the service-time distribution only through its

first two moments and can be approximated by *Kingman's Law*. Schwartz in [32] shows that under the QED regime this approximation is not very good. The comparison of the exponential and the log-normal distributions with the same first two moments in [32] shows that the exponential distribution consistently leads to greater delays. (The mean of the service distribution was always 1, and the CV of the log-normal distribution was also 1.)

In this section we check the impact of the service-time distribution on the expected waiting time in queues with abandonment. Here is the description of the simulated queues.

- Customers arrivals are given by Poisson process with the arrival rates ranging from 93 to 107 customers per hour which approximately corresponds to β from 0.7 to -0.7;
- Customers patience is exponential with the rate $\theta = 1$;
- The service rate is the same for all the distributions, $\mu = 1$; equivalently, $E(S) = 1$.

The service time distributions in the experiments are

- Exponential: $M/M/100 + M$;
- Deterministic: $M/Det/100 + M$;
- Log-Normal with $CV = 1$: $M/LN(CV = 1)/100 + M$;
- User-defined Special service-time distribution: service time is a random variable which can get only one of two values:

$$P(S = 1/k) = \begin{cases} 0.999897, & \text{if } 1/k = 1/0.989897, \\ 0.000103, & \text{if } 1/k = 1/100, \\ 0 & \text{otherwise .} \end{cases}$$

  We denote this distribution as $Special(100)$ to emphasize the longest possible service time.

- User-defined Special service-time distribution: service time is a random variable which can get only one of two values:

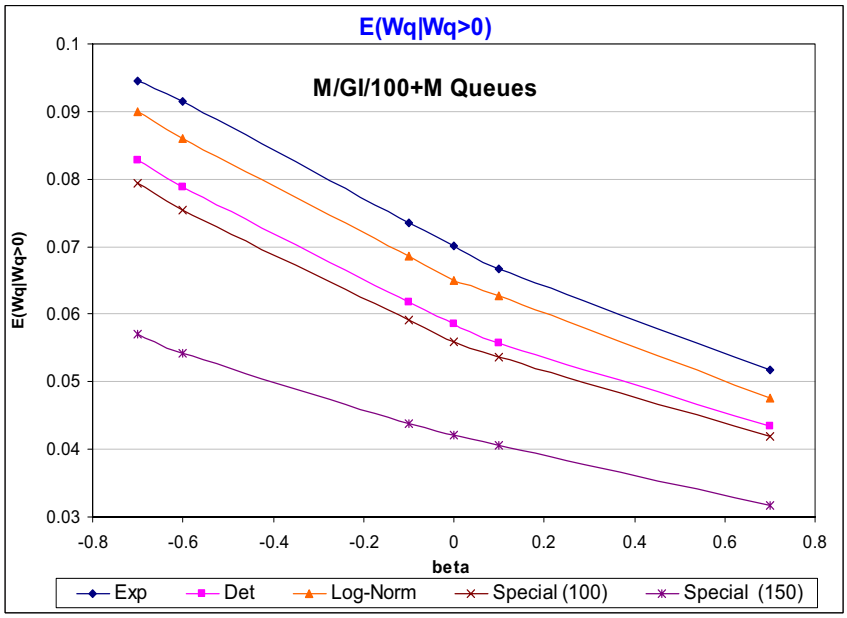
$$P(S = 1/k) = \begin{cases} 0.9999549, & \text{if } 1/k = 1/0.99338, \\ 0.0000451, & \text{if } 1/k = 1/150, \\ 0 & \text{otherwise .} \end{cases}$$

  We denote this distribution as $Special(150)$ to emphasize the longest possible service time.

The mean and the variance of both user-defined distributions are equal to 1.

To check the impact of service-time distribution on system performance, we compare the expected waiting time, given there is waiting, the delay probability and the abandonment probability under the five service-time distributions described above.

Figure 8.2: Expected Waiting Time, if there is Waiting, in Queues with 100 servers



Figures 8.2 and 8.3 summarize these data for all tested $\beta = 0.7, \ldots, -0.7$.

It can be seen from these figures that the $Special(150)$ service-time distribution always leads to the lowest expected waiting time, if there is waiting, and the abandonment probability, while the delay probability is always the highest under the deterministic distribution (See Figure 8.3).

Under the exponential distribution of service time, delayed customers on average experience the longest delay (See Figure 8.2). In addition, the abandonment probability is also the highest for this distribution.

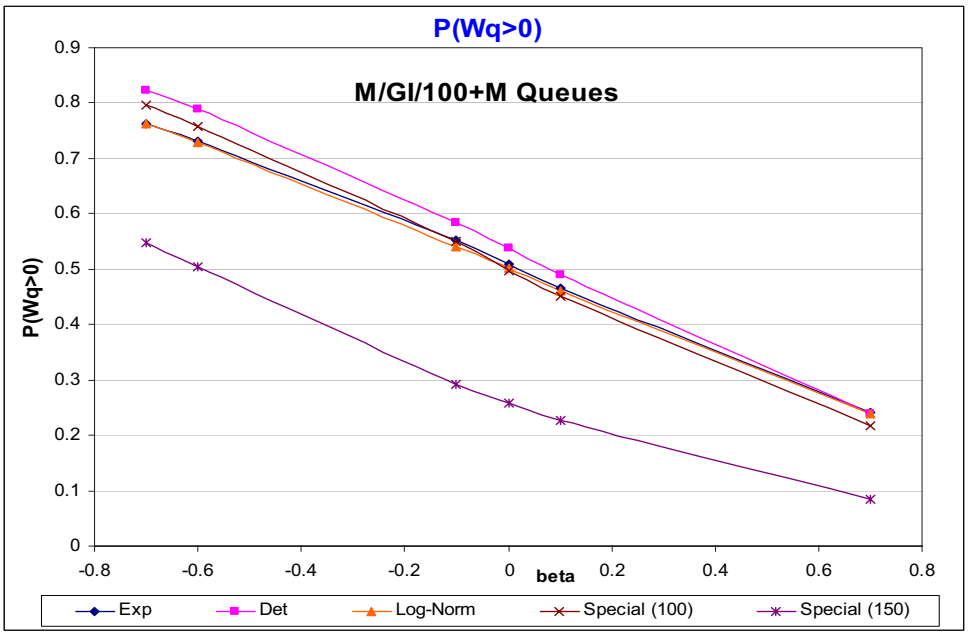
For all tested values of β, the log-normal distribution gives rise to the expected waiting time and abandonment probability somewhere between the exponential and deterministic distributions. (This resembles the results of Schwartz [32], though the location between the distributions is different).

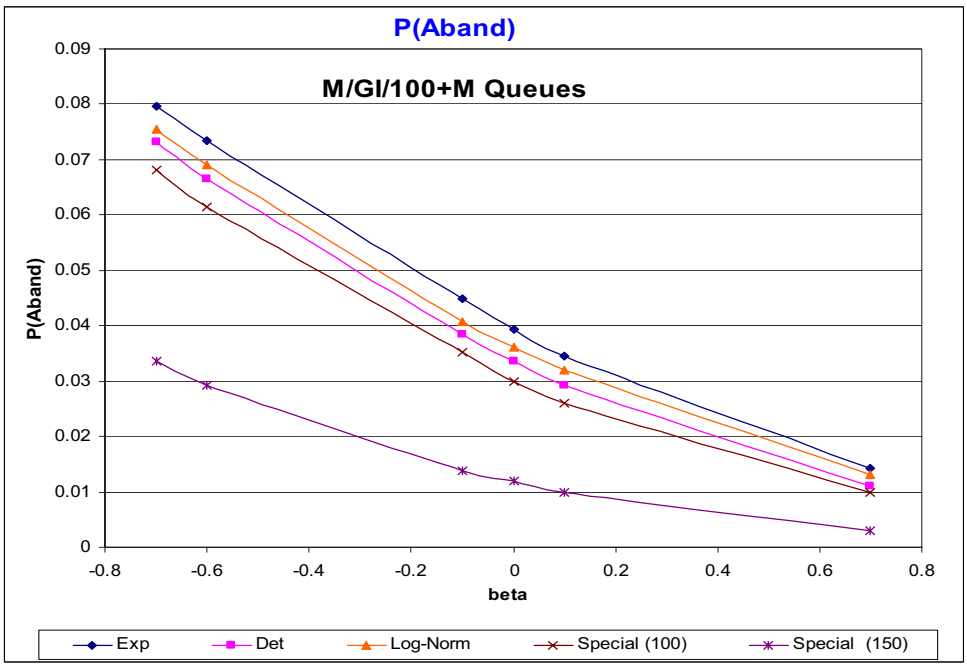
From Figure 8.3 we notice that under the $Special(100)$ distribution, the delay probability grows faster than under the exponential and log-normal. Yet, it does not reach the deterministic distribution, which consistently leads to the highest delay probability.

Special(150) distribution: Let us concentrate on the second Special distribution with values 0.9338 and 150. As the probability of the low value is very close to 1 (p=0.999897) and the number of servers is large, we can say that customers hardly feel those servers who work 150 time units. If a single server happened to work 150 hours (and such a probability is very low), the number of active servers decreases by 1. In such case, customers see an $M/Det/(N-1)+M$ queue with service time of 0.99338 till the next very long service event. Because the probability of such a long service duration is very low, it

is reasonable to assume that the probability of a single, two or more simultaneous long service events is negligible, and the performance of a queue with $Special(150)$ service-time distribution will be very close to that of an $M/Det/(N-1)+M$ queue with the service time 0.99338 customers per hour.

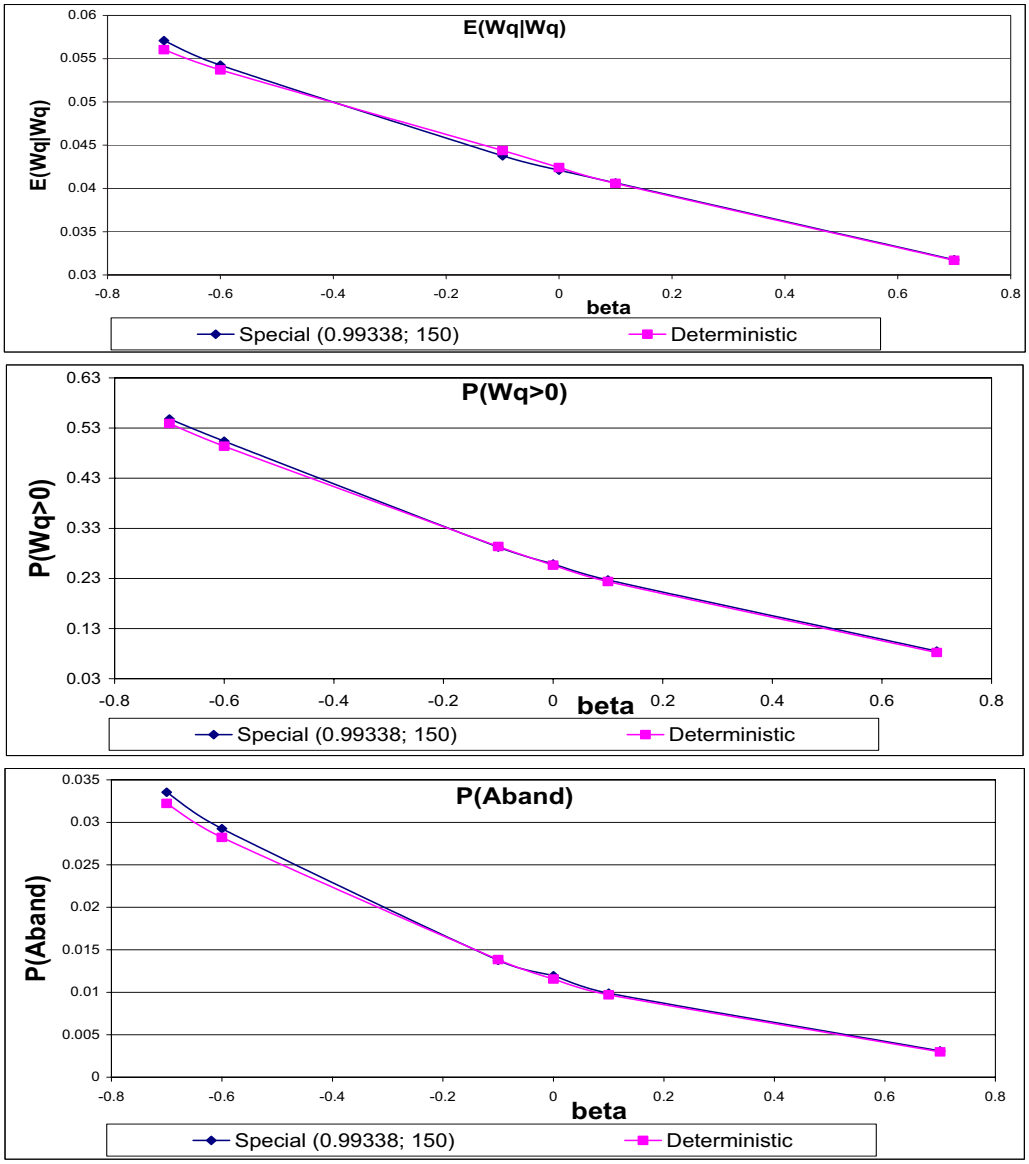
Figure 8.3: Delay and Abandonment in Queues with 100 servers





This similarity of these two systems can be easily checked by simulations. The results of these simula-

Figure 8.4: $M/Special(150)/100$ vs. $M/Det(0.99338)/99$



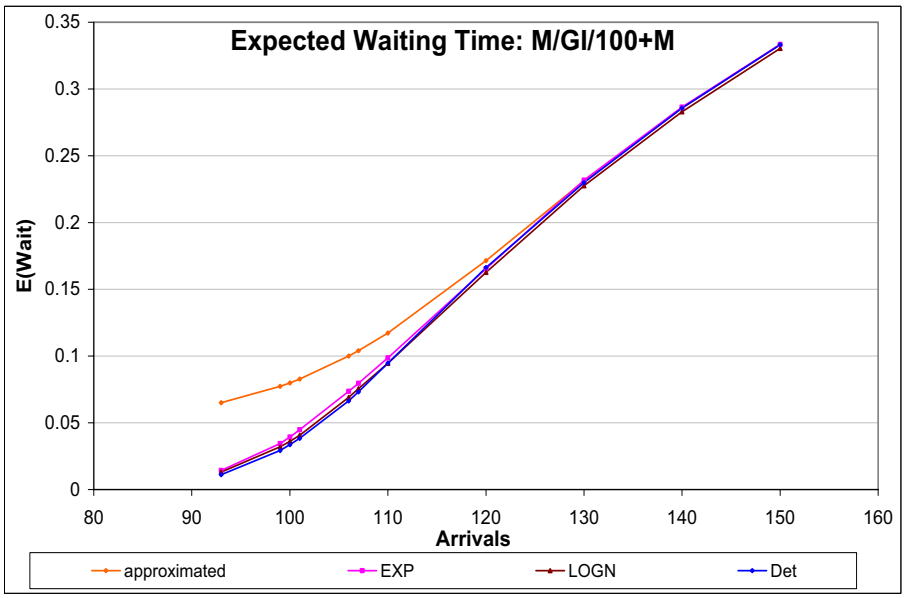
tions for different arrival rates $\lambda = 93, \ldots, 107$ customers per hour (or $\beta = 0.7, \ldots - 0.7$ accordingly) are resented in Figure 8.4. We can see that, as expected, the three tested performance measures are very similar. Yet, queues with Deterministic service time and 99 servers perform slightly better for high arrival rates

Heavy-Traffic Approximation for Many Servers: When the servers of any $M/GI/N+M$ queue work under the ED regime, the delay probability is close to 1, and the customers' experience is not different

from that of $M/M/1 + M$ queue with the service rate $N\mu$. So we may suggest that the heavy-traffic approximations developed in [33] will work for this case, too.

We check this hypothesis by simulations, comparing their results to the theoretical values. Figure 8.5

Figure 8.5: The $M/GI/100 + M$ queue - Empirical Results vs. Approximations



presents these results. As expected, for highly utilized servers the simulated values are close to the approximations. But for for the QED regime, and for the early stages of ED regime the differences between the simulated and the approximated values are significant. So the approximation performs here worse than in case of a single server (compare to Figure 8.1). In this case, the approximated values of the waiting time approaches the simulated ones when the offered load per server exceeds 1.2, the same way as in the case of a single server. However, for lower values of the offered load the error of the approximation used for 100 servers is greater than in case of the single server.

8.3 M/G/100 vs. M/G/100+M

It is worth comparing the simulation results of queues without abandonment, obtained by Schwartz [32], with our simulation results of queues with abandonment.

Comparing Figure 8.2 with Figure 7 in [32], we can see the differences in expected waiting time, if there is waiting, as a function of decreasing β. The order of the distributions does not change after adding the abandonment. In both cases the exponential service-time distribution results in the highest values of the expected waiting time, and the deterministic one leads to the lowest values. The difference is that in the Erlang-C queues the waiting time, if there is waiting, under the exponential distribution is twice as

large as this measure under the deterministic distribution. This coefficient is explained theoretically by the Khintchine-Pollazchek Formula. As a result, when the value of β decreases, the absolute difference between the deterministic and exponential distributions increases.

In the case of Erlang-A queues, there is less difference between the distributions, though the order of the distributions does not change. Figure 8.2 shows that under the QED, the absolute difference between these distributions virtually does not change. Moreover, using the theoretical result of A. Ward [33] we can predict that further increase of the arrival rate will lead to smaller differences between the service distributions with the same first two moments.

The delay probability behaves differently in queues with and without abandonment. In Erlang-C queues the highest delay probability is achieved by the exponential distribution, the lowest is achieved by the deterministic one and the log-normal distribution is situated between the two of them, very close to the deterministic one. This order is preserved for any arrival rate tested in [32]. Yet, in queues with abandonment, the deterministic service-time distribution yields the highest delay probability, while the plots of this probability under the log-normal and exponential distributions almost coincide.

Chapter 9

Future Research

Priority Queues: In our analysis of priority queues (Chapters 4 and 5), we used a model with the same service rate μ and with the same abandonment rate θ. This assumption allowed us to conduct the exact analysis of any priority but made the model less applicable.

General Service and Abandonment Rates: One of the possible future research directions is analyzing models where different customer types are served with different rate μ_k and have different abandonment rate θ_k. The exact analysis of such models is very complicated, whether the customer types differ only by their patience, or both by service and abandonment rates.

In both cases, it is reasonable to conjecture that under the ED or QED regimes as the number of servers grows to infinity and the lowest priority is not negligible, the only type that continues to abandon is type K (the lowest priority). This intuition is based on our result presented in Section 5.4. There we showed that under the ED and QED regimes the expected waiting time conditioned on waiting of higher priorities in Erlang-A queues converges to that in Erlang-C queues, i.e. abandonments of types $1, \ldots, K-1$ become negligible.

In the models with equal service rates and equal abandonment rates the delay probability is equal to that in an $M/M/N(+M)$ queue with the arrival rate $\lambda = \sum_{i=1}^{K}(\lambda_i)$ for any type k. For models with different abandonment rates (and/or different service rates) under non-preemptive priority, the delay probability is the same for any type k but it is hard to find a closed-form expression for this probability.

To analyze models with different service rates μ_k, one might need more advanced mathematical tools. The exact solution of such model is probably impossible, but yet some asymptotic conclusions might be feasible. For example, we can expect that in the case of non-preemptive priority under the ED and QED regimes, when the lowest priority is not negligible, the convergence rate of the highest priority remains $\Theta(1/N)$. Under the preemptive priority this rate will be exponential.

Waiting-Time Distribution: Another challenge in the analysis of priority queues is a formal determination of the waiting-time distribution of type k customers under the non-preemptive priority. Here, the distribution of the delayed first-type customers waiting time is exponential, similarly to the waiting time

conditioned on waiting in an $M/M/1$ queue with the arrival rate λ_1 and service rate $N\mu$. The distribution of other types is more complicated. On this stage, we can only observe that other non-lowest **delayed** types face an $M/G/1$ queue under light traffic.

Time-Varying Queues: Our simulation results presented in Chapters 6 and 7 and results of Feldman [9] clearly show that many of the overall performance measures of time-varying queues can be found by using an appropriate stationary model. However, the theoretical explanation of this fact is beyond our present understanding.

Heavy-Traffic Two-Moments Approximations: The experiments with different service-time distributions described in Chapter 8 support the main finding of Schwartz, [32]. Our experiments show that under the QED regime, the general performance of an $M/G/N+M$ queue depends not only on the first two moments of the service-time distribution, but also on the distribution itself. We found that different service-time distributions with the same two first moments lead to different performance, which contradicts the conventional heavy-traffic approximations.

A proper understanding of the impact of the service-time distribution under the QED regime on the overall performance is an interesting and important research problem. Progress in this field has been achieved in the recent work of Mandelbaum and Momcilovic [25] for finite-support services and Reed for general service times [31].

Bibliography

[1] Abramowitz, M.; Stegun, I.A. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. *U.S. Department of Commerce.* Online version available at: $www.knovel.com/knovel2/Toc.jsp?BookID = 528\&VerticalID = 0$.

[2] Ashlagi I., 2005. Some Results for the M/M/N Queue with Multi-Type Customers. *Technical Report, Technion*.

[3] Baccelli F., Hebuterne G., 1981. On Queues with Impatient Customers. *F.J. Kylstra (Ed.)*, Performance. North-Holland Publishing Company, pp.159-179.

[4] Borst S., Mandelbaum A., Reiman M. 2004. Dimensioning Large Call Centers. *Operations Research*, 52(1), pp. 17-34.

[5] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Zeltyn, S., Zhao, L. and Haipeng, S. 2005. Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association*, Vol 100: 36-50.

[6] Erlang, A.K., 1909. The Theory of Probabilities and Telephone Conversations. Nyt Tidsskrift Mat. D 20, pp. 33-39.

[7] Erlang A.K., 1948. On the rational determination of the number of circuits. In *The life and works of A.K.Erlang*. Brockmeyer E., Halstrom H.L. and Jensen A., eds. Copenhagen: The Copenhagen Telephone Company.

[8] Feldman Z., June 2004. Staffing of Time-Varying Queues To Achieve Time-Stable Performance. Technion. Downloadable from *http://iew3.technion.ac.il/serveng/References*.

[9] Feldman Z., Mandelbaum A., Massey W.A. and Whitt W. 2005. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Submitted to Management Science*.

[10] Gans, N., Koole, G., Mandelbaum, A., 2003. Telephone Call Centers: Tutorial, Review and Research Prospects. Invited review paper by *Manufacturing and Service Operations Management* (M&SOM), 5 (2).

[11] Garnett O., Mandelbaum A. and Reiman M. 2002. Designing a Call Center with Impatient Customers. *Manufacturing and Service Operations Management*, 4(3), pp. 208-227.

[12] Garnett O. and Mandelbaum, A., 2000. An Introduction to Skill-Based Routing and its Operational Complexities. *Teaching note, Technion,* Downloadable from: http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf.

[13] Green L.V, Kolesar P. J., Soares J. 2001. Improving the SIPP approach for for staffing service service systems that have cyclic demands. *Operations Research*,49, 549-564.

[14] Green L.V, Kolesar P. J., Whitt W. 2007. Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management (POMS), vol. 16, No. 1* pp. 13-39.

[15] Gurvich I., Armony M., Mandelbaum A., Staffing and Control of Large-Scale Service Systems with Multiple Customer Classes and Fully Flexible servers. Submitted to *Management Science*.

[16] Halfin, S., Whitt, W., 1981. Heavy-traffic Limits for Queues with many Exponential Servers. *Operations research* 29, pp. 567-587.

[17] Gurvich I., Priority Queues (CRM). Teaching Note. Downloadable from *http://iew3.technion.ac.il/serveng/References*.

[18] Henderson S., Atlason J. and Epelman M., 2004. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research* 127, pp. 333-358.

[19] Jelenkovic P., Mandelbaum A. and Momcilovic P., 2004. Heavy Traffic Limits for Queues with Many Deterministic Servers. *QUESTA* 47, pp. 53-69.

[20] Kella O. and Yechiali U. 1985. Waiting Times in the Non-Preemptive Priority M/M/c Queue. *Stochastic Models*, 1, 257-262.

[21] Kleinrock L. 1975. Queueing Systems. *John Wiley & Sons*.

[22] Kelly, F.P. 1979. Markov Processes and Reversibility. *Wiley, New York.*

[23] Litvak. 2002. Root Cause Analysis of Emergency Department Crowding and Ambulance Diversion in Massachusets. A report submitted by Boston University *Program for the Management of Variability in Health Care Delivery, http://www.bu.edu/mvp/*. Last accessed on November, 14, 2007.

[24] Mandelbaum A., 2004. Call Centers. Research Bibliography with Abstracts. Version 6, Technion, Downloadable from: http://iew3.technion.ac.il/serveng/References/references.html.

[25] Mandelbaum, A. and Momcilovic P. 2007. Queues with Many Servers: The Virtual Waiting-Time Process in the QED Regime. Downloadable from: *http://ie.technion.ac.il/serveng/References/references.html* Last accessed April, 2008.

[26] Mandelbaum A., Sakov A. and Zeltyn S.2000. Empirical Analysis of a Call Center. *Technical Report*.

[27] Mandelbaum, A., Zeltyn, S., 2006, Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. Under Revision in *Operations Research.*

[28] Nakibly E., 2002. Predicting Waiting Times in Telephone Service Systems. *M.Sc. Thesis, Technion.*

[29] Palm, C., 1957. Research on Telephone Traffic Carried by Full Availability Groups. *Tele*, vol. 1, 107, (English translation of results first published in 1946 in Swedish in the same journal which was entitled *Teknisska Meddelanden fran Kungl. Telegrafstyrelsen.*).

[30] Reed J., 2006. Heavy-Traffic Approximations. *Private Communication.*

[31] Reed, J. E., 2007. The G/GI/N Queue in the Halfin-Whitt Regime. Available at *http://pages.stern.nyu.edu/ jreed/Papers/ReedHalfinWhitt101507.pdf.*

[32] Schwartz R., 2002. Simulation Experiments with M/G/100 Queues in the Halfin-Whitt (Q.E.D.) Regime. *Technion.*

[33] Ward A. R., Glynn, P. W.' 2005. A diffusion approximation for a $GI/GI/1$ queue with balking and reneging. *Queueing Systems*, vol. 50, Number 4, pp. 371-400.

[34] Sharpe M.J., General Theory of Markov Processes, *Academic, San-Diego.*

[35] Wallace R, Whitt W., 2004. A Staffing Algorithm for Call Centers with Skill-Based Routing. working paper, Submitted to *Manufacturing and Service Operations Management* (M&SOM).

[36] Whitt W., 1999. Dynamic Staffing in a Telephone Call Center Aiming to Immediately Answer All Calls. *Operations Research Letters*, vol. 24, pp. 205-212.

[37] Whitt W., 1999. Predicting Queueing Delays. *Management Science*, vol. 45, No. 6, pp. 870-888.

[38] Whitt W., 1992. Understanding of Efficiency of Multi-Server Service Systems. *Management Science* Vol. 38, No. 5, 708-723.

[39] Whitt W., Private communication.

[40] Winter, N., 2000. Waiting: Integrating Social and Psychological Perspectives in Operations Management. *Omega*, 28, 611-629.

[41] Zeltyn S., 2004. Call Centers with Impatient Customers: Exact Analysis and Many-Server Asymptotics of the M/M/n+G queue. *PhD Thesis, Technion.*

[42] Zeltyn S., 2005. Call Centers with Impatient Customers: Exact Analysis and Many-Server Asymptotics of the M/M/n+G queue.

[43] 4CC - For Call Centers Software.
Downloadable from *http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm*

[44] Service Engineering (096324) course site: *http://iew3.technion.ac.il/serveng*

[45] Service Engineering (096324), Lecture 13: QED Q's Part II Staffing. Available at: *http://ie.technion.ac.il/serveng/Lectures* . Last accessed on November, 14, 2007.